THE UNIVERSITY
of EDINBURGH

Distributions of RNA polymerase and transcript numbers in models of gene expression describing the mRNA life-cycle

THE UNIVERSITY
of EDINBURGH

Tatiana Filatova

Thesis submitted for the degree of
Doctor of Philosophy

The University of Edinburgh
School of Biological Sciences & School of Mathematics
College of Science and Engineering

2022

Abstract

Transcription, the production of RNA from a gene, is an inherently stochastic process, as recent experiments have firmly established. This stochasticity makes the modelling of genetic networks highly challenging. Recent decades have seen a rise in the development of new mathematical models of gene regulatory networks that aim to extract relevant biological information from experimental data. The telegraph model of gene expression, where the gene switches between active and inactive states, is the most widely used in the literature. However, it has been shown that it cannot explain several experimental observations, as it does not capture many biological details such as transcription factor and polymerase binding to the gene, RNA nuclear retention, multi-step elongation, RNA maturation, etc.

The chemical master equation (CME) describes stochastic chemical reaction networks and, hence, is a commonly used tool in the mathematical modelling of such networks. Specifically, it describes how the joint probability distribution of the copy number of different chemical species evolves in time under spatially homogeneous conditions. Unfortunately, this equation can be solved analytically only in a few cases, while on the other hand, stochastic simulations can be computationally expensive and slow. For these reasons, various approximation techniques have been developed lately to approximate solutions to hitherto unsolved complex master equations. For example, the geometric singular perturbation theory serves as a very useful tool for finding approximate solutions to CMEs of biological models which feature processes on different time scales.

In this thesis, we study the formulation and detailed analysis of three different analytically tractable stochastic models that capture the main features of gene expression under various additional assumptions and that can potentially provide means to infer parameter values from experimental data. We quantify which and how different approximation methods can be applied to systems of interest in order to obtain closed-form analytical solutions.

The first model presented in this thesis is a stochastic model of gene expression with polymerase recruitment and pause release, two steps necessary for messenger RNA (mRNA) production. For this model, which captures the bursty production of mRNA molecules, we derive the exact steady-state distribution of mRNA numbers. Additionally, this model includes the translation process – synthesis of protein from mRNA – and we apply perturbation techniques in order to obtain an approximate steady-state distribution of protein numbers.

The second model that we are studying in this work is a stochastic model of RNA transcription, which focuses on capturing the processes of transcriptional initiation, elongation, premature detachment, pausing, and termination. In this model, the gene is divided into an arbitrary number of segments. The results from our analysis uncover the explicit dependence of the statistics of nascent (actively transcribed) and mature (cellular) RNA on transcriptional parameters. By performing mathematical analysis, we derive exact closed-form expressions for the mean and variance of nascent RNA fluctuations on each gene segment, as well as for the total nascent RNA on a gene. Additionally, we obtain the exact expressions for the first two moments of mature RNA fluctuations while we present an approximation approach for deriving distributions for the total numbers of nascent and mature RNA in various parameter regimes.

The third model that we study in this thesis is a stochastic model that describes the dynamics of signal-dependent gene expression and its propagation downstream of transcription. In this model, the activation of the gene promoter is time-dependent due to the temporal variation in transcription factor (protein) numbers; after transcription initiation, the produced mRNA undergoes an arbitrary number of stages of its life cycle. For any time-dependent stimulus and in the case of bursty gene expression, we developed a novel procedure that allows us to obtain approximate time-dependent distributions of mRNA numbers at all stages of its life cycle. We derive an expression for the error in the approximation and verify its accuracy via stochastic simulation. We show that, depending on the frequency of oscillation and the time of measurement, a stimulus can lead to cytoplasmic amplification or attenuation of transcriptional noise.

To summarize, this thesis presents a detailed explanation of the construction of three families of stochastic models of gene expression and demonstrates how to perform mathematical analysis of the complex CMEs that represent these models. A number of novel approximation methods that

address some difficulties in solving the CME are included in this study, while one of the main goals of this work is to show that extracting biological information from mathematical models can provide us with a better understanding of cells' functions.

Lay summary

A eukaryotic cell consists of three parts: the cell membrane, the nucleus, and the cytoplasm. The cytoplasm fills the space between the membrane and the nucleus. The DNA, which consists of genes (sections of DNA) and contains the genetic information responsible for the development and function of an organism, is housed inside the nucleus. The nucleus is also where the information that is stored in a gene, is copied into a new molecule of messenger RNA (mRNA); this is the process of mRNA production, and it is called transcription. After being produced, the mRNA molecule carries the copied message to the ribosomes, which are cellular machines that use the encoded information in order to perform biological protein synthesis in the cytoplasm. The mechanism of protein production is called translation. This fundamental process, which enables cells to convert encoded information in DNA to synthesise proteins, is called gene expression. Due to technological advancements, several experimental techniques can be used to measure the number of mRNA and protein molecules in single cells. Experimental data show that measured numbers vary randomly over time and that this variation is different from cell to cell. Consequently, there are random fluctuations in the gene expression process, and they have a profound effect on cellular functions.

In the last few decades, scientists have tried to shed light on the underlying mechanisms of gene expression by using mathematical models that can help extract information from experimental observations. The most well-known and widely used model, that describes mRNA dynamics, is the so-called telegraph model. This model consists of three sets of events: (i) Gene switches between active and inactive states; i.e. the RNA transcription process can or can not begin. The simple biological explanation for this switching is that for transcription to be initiated, certain protein molecules, called transcription factors, that participate in this process must be present and located near the gene. (ii) When the gene is active, the transcription of the mRNA may begin, and (iii) when an mRNA molecule is produced, it may decay. In other words, in the telegraph model, all these events are modelled as a system of four chemical reactions: gene activation, gene inactivation, mRNA production, and mRNA degradation; all reactions are random. Now, mathematically speaking, this system of chemical reactions can be described by an equation, the solution of which is the probability distribution of mRNA molecule numbers; specifically, this distribution provides us with the information of what is the probability of finding a certain number of mRNA molecules in a cell at a certain time. The analytical expression of this distribution paired with experimental data can be used to estimate the rates at which reactions occur in the model, which may provide us with a better understanding of which processes happen faster/slower than others during gene expression. The telegraph model is a simplified representation of transcription and unfortunately can not explain several biological observations because gene expression is a much more complicated process in real life. In this thesis, we present the construction and mathematical analysis of three more biologically realistic models of gene expression.

The first model of interest in this thesis is a model of gene expression that takes into account the significant role of a polymerase molecule in the process of transcription; polymerase is an enzyme that helps to assemble the mRNA molecule by attaching to and moving along the DNA. This model is an extended version of the telegraph model, with the difference that the gene fluctuates between three different states. There are two permissive states of the gene, on and off (transcription factor is bound, and the gene activity is depending on the binding state of the polymerase molecule), and one non-permissive state of the gene (neither transcription factor nor polymerase is bound). This change in gene states is not commonly modelled, but our work shows the importance of including this biological detail in our model. The second model, that we study, focuses on the complex process of mRNA transcription. Some well known biochemical steps of transcription are not often modelled in detail and here we develop a model that takes into account these steps e.g. transcriptional initiation, polymerase movement along DNA, polymerase detachment from DNA, polymerase pausing on DNA, and transcription termination. By performing mathematical analysis, we try to understand how the common telegraph model emerges from this more complex model while showing also how fluctuations in the number of RNA molecules depend on model parameters. Our third model considers oscillatory signal-dependent dynamics that affect the mRNA life cycle. It is well known that internal and external signals affect cell fate and hence, this model can provide us with some insights into the

underlying mechanism of how gene expression responds to stimuli. We show that, depending on the frequency of signal oscillations and the measurement time, the signal can increase or decrease the fluctuation in the mRNA number in the cytoplasm compared to those in the nucleus.

To summarize, in this thesis, we show the formulation of three models of gene expression that incorporate more biological details than the simple telegraph model. These models can potentially provide means to estimate parameter values from experimental data, and hence make it possible for us to get a better understanding of cells' functions. Our work demonstrates how to perform mathematical analysis of these complex models, and how to extract biological information from our results. Specifically, we approximate the probability distributions of molecule numbers of species of interest in all three models by applying a variety of mathematical techniques, as these distributions cannot typically be found in closed-form.

Declaration of Authorship

I, **Tatiana Filatova**, declare that this thesis entitled *"Distributions of RNA polymerase and transcript numbers in models of gene expression describing the mRNA life-cycle"* has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work that has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

The work presented in Chapter 2 contains parts of the publication in Biophysical Journal (2020), "A Stochastic Model of Gene Expression with Polymerase Recruitment and Pause Release" by Zhixing Cao, **Tatiana Filatova**, Diego A. Oyarzún and Ramon Grima. This study was conceived by all the authors. The author's contribution to this paper can be found at the end of the published manuscript. I note that R.G., who is co-author of the manuscript, is also my PhD supervisor. I also note that, that the Sections 2.2-2.4 from Chapter 2 of this thesis include results from the joint work of all the authors, while the Section 2.5 is part of my contribution to the published manuscript and has been supervised by the rest of co-authors. The notation used in Chapter 2 of this thesis is not the same as in the published manuscript, while the figures in this chapter have been produced by myself for this thesis and are not included in the published manuscript.

The work presented in Chapter 3 was previously published in the Bulletin of Mathematical Biology (2021) as "Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination" by **Tatiana Filatova**, Popović Nikola and Ramon Grima. Additionally, the work presented in Chapter 4 has been accepted for publication in Mathematical Biosciences (2022) as "Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions", by the same authors. The major part of the research and the major part of the writing of these two studies has been conducted by the candidate. The two co-authors contributed by guiding the research and the publication writing process, as well as giving feedback and helping with the corrections for the publication.

Tatiana Filatova

July 5, 2022

Publications

- **Filatova, T.**, Popović, N., & Grima, R. (2022). Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions. *Mathematical Biosciences, 347: 108828.*
- **Filatova, T.**, Popović, N., & Grima, R. (2021). Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination. *Bulletin of Mathematical Biology, 83(1), 1–62.*
- Cao, Z., **Filatova, T.**, Oyarzún, D. A., & Grima, R. (2020). A stochastic model of gene expression with polymerase recruitment and pause release. *Biophysical Journal, 119(5), 1002–1014.*

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Ramon Grima. I deeply thank him for giving me this opportunity, and for being so supportive during my PhD. I am thankful to Ramon for leading my studies and for being always patient and understanding when I was reaching my dead ends. I am beyond grateful to you for being my supervisor, my boss, my mentor, my friend, and most importantly, for being my teacher in academic research and this life.

An especially heartfelt thanks goes to my second supervisor, Nikola Popović for his always invaluable advice and wonderful support. I sincerely appreciate his continuous encouragement that helped me progress and dare to take steps I wouldn't have made without it. I can't thank you enough for your time and your kindness that made my learning and working experience so enjoyable during these years.

I would like to express my sincere appreciation to Peter Swain for joining my PhD committee and providing me with useful feedback.

I would like to thank the University of Edinburgh for the studentship that allowed me to conduct this thesis and for the wonderful environment at the campus that has made my working days very pleasant.

I owe particular thanks to all the past and present members of the Grima Group. It was nice to observe how we have grown together over these years. Thank you all for our moments, events, parties, and trips that have made my studies and life in Edinburgh a wonderful time.

I am extremely grateful to my friend Samuel Casasola Zamora. I deeply thank you for always having me covered. Your kind help and support mean the world to me.

Special thanks go to my friend Panagiotis Kaklamanos, with whom we started the PhD journey together. Thank you, Panos, for your support and invitations to workshops and events.

Finally, I would like to express my gratitude to my family and friends for always supporting me and believing in me. I'm forever indebted to my grandmother for showing me love like no other.

For my mum Elena Filatova

Contents

Appendix

List of Abbreviations

TF	Transcription Factor
Pol II	RNA polymerase II
GFP	Green Fluorescent Protein
FISH	Fluorescence In *Situ* Hybridization
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
CRN	Chemical Reactions Network
LNA	Linear Noise Approximation
CME	Chemical Master Equation
SSE	System Size Expansion
LFPE	Liner Fokker-Planck Equation
CFPE	Chemical Fokker-Planck Equation
CLE	Chemical Langevin equation
SSA	Stochastic Simulation Algorithm
GSPT	Geometric Singular Perturbation Theory
NB	Negative Binomial
FF	Fano Factor
CV	Coefficient of Variation
RE	Relative Error
HD	Hellinger Distance
TASEP	Totally Asymmetric Simple Exclusion Process
FM	Full Model
RM	Reduced Model
EM	Effective Model
GRN	Gene Regulatory Network

Notation

- ◇ Integer numbers: $\mathbb{Z}$
- ◇ Natural numbers: $\mathbb{N}$
- ◇ Real numbers: $\mathbb{R}$
- ◇ Complex numbers: $\mathbb{C}$
- ◇ Real part of the complex number, z: $\Re[z]$
- ◇ Imaginary part of the complex number, z: $\Im[z]$
- ◇ Empty set that denotes sources and sinks of molecules: $\varnothing$
- ◇ Real small perturbation parameter: $\varepsilon \ll 1$
- ◇ N-dimensional array of integer numbers: $\vec{n} = (n_2, n_2, \ldots, n_N)$, where $n_i \in \mathbb{Z}$
- ◇ Step operator: $\mathbb{E}_{n_i}^k[f(\vec{n})] = f(n_1, n_2, \ldots, n_i + k, \ldots, n_N)$, where $k \in \mathbb{Z}$
- ◇ Probability function: $P(\vec{n}; t)$ denotes that there are n_i the number of molecules of species i $(i = 1, 2, \ldots, N)$ in the system at time, t
- ◇ Probability generating function: $F(\vec{z}; t) = \sum_{n_1,\ldots,n_N=0}^{\infty} P(\vec{n}; t) z_1^{n_1} \ldots z_N^{n_N}$, where $z_i \in [0, 1]$
- ◇ Normalization condition: $F(\vec{z}; t)|_{(z_1=1,\ldots,z_N=1)} = \sum_{n_1,\ldots,n_N=0}^{\infty} P(\vec{n}; t) = 1$
- ◇ Marginal probability function: $P(n; t) = \dfrac{1}{n!} \dfrac{d^n}{dz^n} F(z; t)|_{(z=0)}$
- ◇ Mean value of a distribution: $\langle n \rangle = \partial_z F(z)|_{(z=1)}$
- ◇ Variance of a distribution: $Var(n) = [\partial_z^2 F(z) + \partial_z F(z) - (\partial_z F(z))^2]|_{(z=1)}$
- ◇ Coefficient of variance: $CV = \sqrt{Var(n)}/\langle n \rangle$
- ◇ Fano factor: $FF = Var(n)/\langle n \rangle$

Chapter 1

Introduction

1.1 Life cycle of RNA in eukaryotic cells

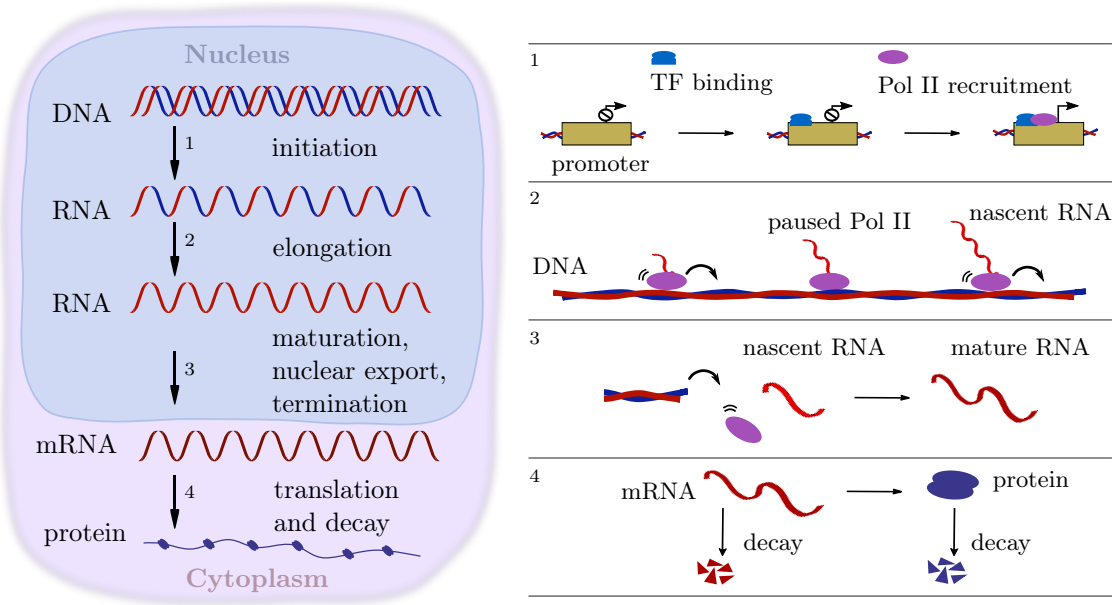


Figure 1.1: *Life cycle of RNA in eukaryotic cells.* Production of RNA is a cyclic process that consists of three principal sets of transcription events: initiation, elongation, and termination. In eukaryotes, RNA is transcribed in the nucleus part of the cell. Transcriptional factors (TF) bind to the inactive promoter region of DNA and enable RNA polymerase II (Pol II) recruitment to the transcriptional start site of DNA and switch the promoter to a transcriptionally active state. After transcription initiation, the polymerase moves along the gene, resulting in elongation of the nascent RNA. During elongation, the polymerase can pause at certain sequences of the DNA, by regulating in this way the transcription. The nascent RNA passes through multiple maturation processes before becoming mature messages RNA (mRNA), while at the same time it proceeds through the final step in the transcription cycle, the termination. During the termination process, the polymerase detaches from the gene and is free to begin a new search for a promoter, while the now mature mRNA exports from the nucleus into the cytoplasm and participates in translation (protein synthesis). Both mRNA and protein degrade in the cytoplasm.

Cells come in two types, eukaryotes and prokaryotes; these two types have structural and functional differences between them. Eukaryotic cells have a membrane-bound nucleus and prokaryotic cells do not. In this thesis, our discussion and studies are based on eukaryotic cells. In eukaryotes, the messenger RNA (mRNA) carries genetic information from DNA, which is located in the nucleus, to the sites of protein synthesis in the cytoplasm (the ribosomes). Transcription of RNA happens in three main steps: initiation, elongation, and termination. Each of these steps consists of multiple individual events. The production of RNA happens inside the nucleus from where it gets exported into the cytoplasm leading to protein synthesis, whereas protein must be transported into the nucleus from the cytoplasm to regulate gene expression. The life cycle of mRNA ends with the degradation process in the cytoplasm. Please see a simple schematic representation of the mRNA life cycle in Fig. 1.1. Transcription and translation processes have been intensively studied in *vitro*, and more recently, directly in living cells; we are going to briefly discuss these processes in the following paragraphs.

The first stage of the RNA life cycle is termed **transcription initiation**, and it is promoted by a set of protein complexes, generally referred to as transcription factors (TF). TFs together with other proteins lead to the formation of a preinitiation complex (approximately 100 proteins), which promotes (as an activator), or blocks (as a repressor) the recruitment of an RNA polymerase II (Pol II) to the target genes [1–3]. The recruitment process happens when the Pol II is properly placed at the transcriptional start site of the promoter of the gene [1, 4]. The establishment of the polymerase-promoter complex results in the formation of an open complex in which the DNA duplex is unpaired, allowing Pol II to access the nucleotide bases and start copying the message. Once the Pol II has formed a phosphodiester bond between the first ribonucleotides, it escapes the promoter by translocating one base and repeating the process of phosphodiester bond formation; this results in **elongation** of the nascent RNA, whose stand gets longer thanks to the addition of new nucleotides [5]. The elongation reaction continues at an average rate of 20 to 30 nucleotides per second until the complete gene has been transcribed [6, 7]. During elongation, the Pol II pauses at certain sequences, which allows appropriate RNA editing factors to bind [8, 9]; Pol II pausing is an important mechanism for transcription regulation. After short or long pauses, the Pol II may move back several bases if certain conditions for elongation are not appropriate. This backtracking movement may affect the length of the transcribed nascent RNA. Usually, after long pauses, an actively transcribing Pol II molecule may detach from the DNA template; this is known as polymerase premature termination, and it is an important mechanism verifying that only the intended gene is transcribed [6]. The transcription cycle ends with **termination**, which happens when the Pol II reaches a certain sequence of DNA known as a terminator. At this point, after an extended pause in elongation, the nascent RNA dissociates from the transcribing Pol II and the DNA template returns to the base-paired conformation; this leads to the Pol II detachment from the DNA template, which becomes free to search again for a promoter [6].

A transcribed nascent RNA molecule has to undergo various **maturation** processes before it gets exported from the nucleus into the cytoplasm and becomes a functionally active mature RNA molecule ready for translation of proteins. Some main steps in nascent RNA maturation are capping, polyadenylation, and splicing. Capping is the mechanism for the addition of 7-methyl-guanosine caps to the 5' end of the nascent RNA. The 5' cap protects the nascent mRNA from degradation and assists in ribosome binding during translation [10]. Polyadenylation is the mechanism of adding poly-A tails at the 3' end of the RNA. The poly-A tail is important for the stability of the mRNA [11]. RNA splicing is the mechanism by which introns are removed, while exons are retained in mature mRNA molecules. Introns are not expressed in proteins, while exons are the coding regions of DNA sequences that correspond to proteins; hence, RNA splicing is an important step that prepares the mRNA to be translationally functional [12]. After this processing, the mature mRNA molecule undergoes **nuclear export** and becomes cytoplasmic mRNA. Nuclear retention duration times stem from multiple biological processes including capping, polyadenylation, splicing, chromatic dissociation, nuclear diffusion, RNA binding to proteins and export factors, and RNA successful transport across the nuclear pore [13, 14]. In the cytoplasm, the mRNA participates in the synthesis of proteins and also degrades.

The mRNA **decay** process begins much earlier than the mRNA is exported into the cytoplasm [15] (and references therein). When the mRNA processing in the nucleus is complete, the mRNA bears a 5′ cap structure and a 3′ poly-A tail that protects the message from decay. mRNA decay mechanisms consist of multiple steps, where the first step is the removal of the poly-A tail by a deadenylase enzyme in the nucleus before export; this is the initiation of the deadenylation process. Once poly-A shortening is complete, the 5′ 7-methyl-guanosine cap is rapidly removed (decapping) and the rest of the mRNA is attacked by 5′ and 3′ exonucleases [15–20]. There is experimental evidence showing that an mRNA molecule can be degraded in the nucleus [19, 21]; however, generally the mRNA gets exported into the cytoplasm where it produces proteins, and at the same time it approaches its death gradually, passing through a series of states (e.g. terminal deadenylation, decapping) before reaching the last phase called final degradation [15, 16, 18].

While the mRNA is in the cytoplasm, it participates in **translation**: the process of translation of the sequences of nucleotides in an mRNA into the sequence of amino acids in a polypeptide chain. During translation, mRNAs along with transfer RNAs (tRNAs; carry specific amino acids) and ribosomes co-function to produce a specific polypeptide, which later folds into an active protein. Alike transcription, translation also proceeds in three main steps: initiation, elongation, and termination. During initiation, the ribosome assembles around the target mRNA, which is followed by elongation, where the ribosome moves from one mRNA codon to the next and creates in this way an amino acid chain. Translation termination occurs when the ribosome reaches a stop codon on the mRNA; then, it releases the synthesized polypeptide, while the ribosomal complex remains intact and moves on to the next mRNA to be translated. It has been found that proteins typically exist for at least several mRNA lifetimes [22–24]. Please see [6, 25] for details about translation mechanisms.

1.2 Counting RNAs and proteins

By counting mRNAs and visualizing proteins in cells, we can obtain a better understanding of the rules that govern gene expression. There are numerous experimental techniques that measure mRNA and protein abundance within a population of cells, while in recent years it had become possible to analyse gene expression in single cells on a single-molecule level [26]. For example, the **green fluorescent protein (GFP)** has been used as the main tool in cell biology to infer the mechanisms of gene expression; GFP can attach to and mark another protein with fluorescence, enabling the detection of a particular protein in an organic structure, i.e. GFP can be used to localize proteins, to follow their movement or to study the dynamics of the subcellular compartments to which these proteins are targeted [27]. Additionally, a study in [28] shows that GFP is a reliable reporter of gene expression in individual eukaryotic cells when fluorescence is measured by flow cytometry. Although the GFP approach enables the study of proteins in cells, it can not be used to study other biomolecules of interest such as DNA or mRNA.

A widely used experimental method that enables the measurement of mRNA in cells is the so-called **single-molecule fluorescence in *situ* hybridization (smFISH)** [26, 29]. FISH was initially developed to detect and localize the presence or absence of specific DNA sequences on chromosomes, while now FISH can also be used to detect and localize specific RNA target sequences. The basic principles of FISH are the following. Particular DNA or RNA sequences can be identified within cells by taking advantage of the ability of nucleic acids to anneal to each other under the proper conditions to form a duplex DNA:DNA or RNA:RNA or DNA:RNA, known as a hybrid. Hybrids between natural and artificial nucleic acids are possible; hence, there are various techniques that use either DNA or RNA probes to bind to DNA or RNA targets within a biological sample - a method known as in *situ* hybridization (ISH). The earliest ISH methods were using radioactive probes that were characterised by several disadvantages, including harm to human health. Fluorescent probes were later developed, and methods that employed these probes became known as FISH. By applying FISH and using fluorescence microscopy, the target DNA or RNA can be reliably imaged by applying probes. The binding of fluorescent probes (DNA or RNA) to a single molecule of mRNA (or DNA sequence) provides sufficient fluorescence to accurately detect

and localize each target mRNA (or DNA sequence) in a wide-field fluorescent microscopy image. Probes that do not bind to the intended sequence do not achieve sufficient localized fluorescence to be distinguished from the background. More recent developments in imaging technology allow for the visualization and semi-automated quantification of individual mRNA molecules while using FISH; this method is known as RNA-FISH or single-molecule FISH. Studies using smFISH have revealed sites of RNA processing, transport, and cytoplasmic localization [30] and have quantified the number of polymerases (actively transcribed nascent RNA) in single cells [31–34]. However, smFISH can only provide a snapshot of mRNA abundance and gene activity inside a cell because the cell has to be fixed for the application of this technique.

Another method for counting actively transcribing polymerase molecules in a single cell is termed **electron microscopy**. Electron microscopes use a beam of accelerated electrons as a source of illumination and can reveal the structure of single cells by producing high-resolution images. These microscopes cannot be used to image living cells because the electrons destroy the samples; however, electron micrographs can reveal the number of RNA polymerases engaged in transcribing a single gene in a single cell at a specific instant in time [35–39].

After quantifying RNA and protein molecules in single cells, experimental evidence reveals that the expression of individual genes is highly variable from cell to cell, even within a population of identical cells. The variations are thought to arise from a typically small number of molecules involved in gene expression, with protein numbers often on the order of hundreds of molecules, mRNA on the order of tens of molecules, and the genes themselves often present in just one or two copies per cell. The underlying reasons for large fluctuations in the number of gene products present in each cell constitute the subject of numerous research works [31,40–42]. Studies conclude that the variation in gene expression is due to stochastic production and destruction of gene products, and this phenomenon has been termed as **noise in gene expression**. For example, a study performed by A. Raj et al. in [31] shows that the mRNA production in mammalian cells is a result of various stochastic events; specifically, they present direct evidence that genes transition randomly between transcriptionally active and inactive states, leading to bursty transcription of mRNA. They show that this gene fluctuation results in cell-to-cell variations in gene expression in clonal cells.

To understand the stochasticity in gene expression and the regulation mechanisms of transcription and translation, it is important to know: (i) the number of mRNA and proteins per cell, (ii) the number of mRNAs transcribed per gene, and (iii) the number of proteins translated per mRNA. The measurements of single molecules can be used to obtain probability distributions of the number of gene product molecules. The experimentally observed distributions can be compared with mathematical models and provide one with important information about gene expression. Both deterministic and stochastic types of mathematical models have been used to infer kinetic parameters from the experimental data; however, **stochastic modelling of gene expression** has seen higher interest since stochasticity plays an important role in gene expression. Mathematical or computational analysis of a stochastic model can provide one with a probability distribution of molecule numbers for species of interest (RNA or protein), which can be compared to the distribution obtained experimentally. This thesis is based on studies of stochastic models and in the next section we discuss how stochastic models help us identify the key mechanisms for transcription regulation.

1.3 Stochastic modelling of gene expression

Gene expression is a fundamentally stochastic process, with randomness in transcription, translation and degradation, leading to significant cell-to-cell variations (noise) in mRNA and protein levels. Stochastic models of gene expression that incorporate this randomness of biological processes have been developed to shed light on the underlying mechanisms of gene expression. In many cases there has been an excellent agreement between the models and the experiments, enabling us to detect the key processes that control noise on transcription or translation levels. In this section, we are going to discuss the most used stochastic models.

The simplest model of stochastic mRNA and protein dynamics is the so-called **two-stage model**, which describes the transcription of mRNAs from a gene and translation of proteins from the produced mRNAs [43]. Both mRNAs and proteins may degrade after their production. The processes of transcription, translation and degradation are modelled as first-order chemical reactions:

$$\begin{aligned} Gene &\longrightarrow Gene + mRNA && \text{mRNA transcription,} \\ mRNA &\longrightarrow mRNA + Protein && \text{protein translation,} \\ mRNA &\longrightarrow \varnothing && \text{mRNA degradation,} \\ Protein &\longrightarrow \varnothing && \text{protein degradation.} \end{aligned}$$

Here, the empty set $\varnothing$ denotes a sink of molecules. These first-order reactions are effective since each encapsulates the effect of many underlying biochemical reactions; for networks of chemical reactions please see Section 1.5.1. If the protein species is neglected in this model, then the mRNA dynamics are described by a simple birth-death process and the exact time-dependent mRNA distribution is a Poisson distribution [43]. The time-dependent distribution for protein numbers has been obtained [43] under the assumption that the protein lifetime is greater than the mRNA lifetime; This distribution is given in terms of a Gaussian hypergeometric function [44], while it simplifies to a negative binomial distribution in steady-state. There are numerous experimental shreds of evidence that the distribution of the number of mRNA molecules is often non-Poissonian [31,45–48], which means that this simple two-stage model does not capture important biological processes from transcription. A modified stochastic model for mRNA dynamics was then proposed; this model intends to capture the mRNA processing mechanism and it is described by the following reactions:

$$\begin{aligned} Gene &\longrightarrow Gene + unprocessed\ mRNA && \text{mRNA transcription,} \\ unprocessed\ mRNA &\longrightarrow processed\ mRNA && \text{mRNA processing,} \\ processed\ mRNA &\longrightarrow \varnothing && \text{mRNA degradation.} \end{aligned}$$

The unprocessed/processed mRNA species can be interpreted in many ways, e.g. nuclear/cytoplasmic mRNA or nascent/mature mRNA. A study performed in [49] has defined these mRNA species as unspliced/spliced. The distribution for both mRNA species in this model is Poissonian, hence, the development of other stochastic models remained desired.

In eukaryotic cells, it is believed that a major source of noise lies in the dynamics of TF binding to the promoter of a gene, which may trigger random promoter switching between transcriptionally active and inactive states [50, 51]. It is known that TFs control the rate of transcription of genetic information from DNA to mRNA. The function of TFs is to regulate the activation and inactivation of genes and to make sure that they are expressed at the right time and in the right amount throughout the life of the cell. The fluctuations of the promoter between inactive-active states result in bursty production of RNA; this phenomenon is termed "burst-like transcription", in which a large number of mRNA molecules are transcribed in short periods during the gene's active times, followed by long transcriptionally inactive times of the gene [31, 45, 47, 52, 53]. For this reason, promoter activation (due to TF binding) happens at the so-called "bursty initiation rate", while its inactivation (due to TF detachment) happens at the so-called "burst-termination rate". An interesting note is that the analysis of experimental data of the steady-state distributions of cytoplasmic mRNA in yeast for a number of different genes has suggested that yeast genes fall into two different classes: those that are transcribed in random uncorrelated events clearly separated in time and without any transcriptional memory (Poissonian transcription), and those that are transcribed in bursts [26, 50, 54]. Single-cell experiments verify the transcriptional bursting for many organisms [50, 54] (plus references therein). The size and frequency of transcriptional bursts affect the magnitude of temporal fluctuations in mRNA and protein content of a cell and thus constitute an important source of intracellular noise [48]. Additionally, studies have shown that bursty gene expression can be satisfactorily described by a so-called **three-stage model**, where the promoter switches between active and inactive states [31, 45, 55–60]. In this model transcription of mRNA can only occur when the gene is active, and as before, all processes are modelled as

first-order chemical reactions:

$inactive\ Gene \rightleftharpoons active\ Gene$	promoter switching - TF binding/unbinding,
$active\ Gene \longrightarrow active\ Gene + mRNA$	mRNA transcription,
$mRNA \longrightarrow mRNA + Protein$	protein translation,
$mRNA \longrightarrow \varnothing$	mRNA degradation,
$Protein \longrightarrow \varnothing$	protein degradation.

If the protein species is neglected, the model is called **telegraph model**; this model is widely adopted in the literature and has seen extensive use. The steady-state mRNA distribution of the three-stage model has been obtained in [31, 61], while the time-dependent solution has been reported in [61,62]; both, steady-state and time-dependent expressions are given in terms of confluent hypergeometric functions of the first kind [44, 63]. The steady-state protein distribution for the same model, has been obtained in [43] and it is valid under the assumption that the proteins have longer life times than the mRNAs; the expression of this distribution is in terms of a Gaussian hypergeometric function, while it simplifies to negative binomial distribution under certain conditions for the kinetic parameters of the model.

While the distribution obtained from the telegraph model can typically fit cellular RNA abundance data, there are innate difficulties with the interpretation of that fit: fluctuations in cellular RNA numbers and, hence, the shape of the experimental RNA distribution do not only reflect transcription, but also many processes downstream thereof, such as splicing, RNA degradation, and partitioning during cell division. A recent review in [64] provides the readers with a thorough discussion about the fact that despite the widespread use of the telegraph model it cannot accurately describe transcription kinetics for all genes, in all systems. This is mainly because the assumptions of the telegraph model – constant rates for initiation, degradation, and switching between active and inactive states – are unrealistic in many biological systems, since the transcription changes in response to a multitude of signals, but the model does not easily account for this. Several studies have shown that the telegraph model can be inconsistent with experimental data for mRNA dynamics and here, we are going to mention some examples. Recently, Bartman et al. have studied the RNA polymerase II dynamics in mammalian cells and have shown that the telegraph model can not explain their obtained experimental data; the telegraph model does not include an independently regulated pause release step and hence cannot distinguish the effects of changing polymerase pause release versus polymerase recruitment rates [65] (for details see supplementary figure 3 in [65]). In a different study, Bothma et al. have examined the dynamic regulation of even-skipped (*eve*) stripe 2 expression in the *Drosophila* embryo and have examined the fluctuations of *eve* transcription [66]. Their findings reveal that the occurrence of multiple rates of polymerase II loading argues against the telegraph model. Instead, the data are consistent with a "multistate mode" where the promoter switches between several discrete transcriptional states (for details see figure 4 in [66]). Additionally, Suter et al. in their work in [53] have identified gene-specific "on" and "off" switching rates in transcriptional activity in mammalian cells. They have shown that the "on" intervals followed exponential distributions, suggesting a first-order inactivation of gene transcription, while in contrast, the "off" intervals showed a local maximum that was best described by assuming two sequential exponential processes, indicating a refractory period in the "off" state before the gene can be activated again. This appears to be in contrast to the telegraph model, according to which the gene inactivation times follow an exponential distribution. Last to mention here, Neuert et al. have identified a single quantitative four-state model to understand and predict gene expression dynamics in yeast cells in response to various environmental and genetic perturbations [67]. Their observations indicate that the telegraph and three-state models are too simple to explain their experimental data.

Further regulation mechanisms and stochastic models of mRNA transcription are discussed in the next section.

1.4 Regulation of gene expression

All steps in transcription are subject to some degree of regulation of gene expression. In this section, we are going to discuss some of them and specify those which inspired our research work in this thesis.

Recent experimental studies suggest that the gene promoter displays stochastic fluctuations on different time scales, a phenomenon termed as **multi-scale transcriptional bursting** [51,65]. In order to get a better understanding of this mechanism of gene expression, research in [65] suggests a stochastic model, which considers the TF and Pol II dynamics. In this model, the promoter fluctuates between three states. Two of these states are related to the TF binding/unbinding, but alike in the three-stage model, here, even when the TF is bound to the gene, the gene is not transcriptionally active without the Pol II. Only after TF binding, the Pol II requires the promoter, leading it to a new permissive state where the Pol II pauses. Shortly after, Pol II releases from the pause state leading to the production of an mRNA molecule and then it unbinds the promoter. Change in the rates of burst initiation/termination, as well as in rates of Pol II recruitment/pause release provide means to regulate transcription [65, 68]. Further discussion about this model and its detailed mathematical analysis is presented in Chapter 2 of this thesis.

Usually, the time scales of promoter switching are much slower (hours) than the time-scale of **transcription initiation** (minutes) [48, 69], and hence the latter is considered one of the main regulatory step of transcription. Transcription initiation involves several distinct and complicated steps, and specifically studies in [5,70] show the existence of several transcription initiation steps in *vivo*. These studies also suggest that multiple initiation steps of similar duration lead to a reduction in fluctuations in the number of mRNAs in a cell when compared to those produced from single-step initiation. The models described in the previous section consider transcription initiation as a one-step process, however, an interesting example of a stochastic model where transcription initiation happens in two sequential steps can be found in [71]; in this model, the first step represents the formation of the preinitiation complex at the promoter, while the second step represents the RNA polymerase escape from the promoter leading to an initiation event.

Following initiation, the **elongation** of a nascent RNA molecule (or equivalently RNA polymerase that is actively engaged in the transcription of a gene) is also a highly stochastic process. Measurements of nascent RNA are not affected by post-transcriptional processes and are more direct readouts of transcriptional dynamics. This means that distributions of nascent RNAs can reveal some rules about transcription regulation via tuning elongation rates. Examples of models that describe elongation as a continuous process with a constant speed have been studied in [71,72], while a study in [70] presents a model where elongation is modelled as a multi-step stochastic process. The latter inspired our work on a similar model, where we study and compare the fluctuations of nascent and mature RNAs; we present our complete work on this model in Chapter 3 of this thesis.

In addition, experimental studies show that **nuclear retention** and transport of transcripts between the nucleus and the cytoplasm is an efficient mechanism for buffering random fluctuations in the number of mRNA molecules arising from bursts in transcription in mammalian cells [13,14, 73–76]. It has been estimated that in mammalian cells the nuclear retention times range from a few minutes up to one hour and a half; Battich et al. have used data obtained from high-quality RNA-seq dataset for 282 genes in mouse bone marrow-derived macrophages and approximated the nuclear retention time of newly synthesized transcripts to be between 5–90 mins with a mean value to be approximately 20 min [13]. Halpern et al. also have estimated nuclear lifetimes of a few minutes for the majority of the studied genes from mouse liver cells [74]. Mor et al. have obtained an estimation for nuclear retention time of approximately 5-40 mins by studying genes constructs containing different forms of human DNA [77]. On the other hand, Schwanhausser et al. have shown that the median estimated mRNA half-life for mammalian cells is 9 hours by studying more than 5000 genes from mouse fibroblasts [78]. The non-negligible period that the mRNA spends in the nucleus suggests that nuclear retention processes can alter significantly the mRNA dynamics. Also, it has been found that in some cases the export rates can be slow and

comparable to cytoplasmic mRNA degradation rates but in most mammalian cells they are higher than the latter [13, 74]. Some studies have revealed that decreasing the value of the nuclear export rates can lead to a dramatic reduction of variability of cytoplasmic mRNA without changing the average cytoplasmic mRNA level [13, 73, 74].

Studies in [79–81] show that genes actively use different regulatory mechanisms to reduce stochastic fluctuations not only in mRNA levels but in protein levels as well. However, nuclear retention is not an efficient mechanism for decreasing variability at the protein level because protein noise level seems to be invariant of the export rates. One reason for this is that nuclear mRNA transport can dramatically increase mRNA auto-correlation times, which enhances variability in protein levels by making it difficult for protein molecules to average out fluctuations in the underlying mRNA population. Study in [73] shows that for mRNA export to significantly buffer the fluctuations on the protein level, the export rate will have to be comparable to the protein degradation rate, but in general, the protein half-lives are much longer than nuclear retention times [78].

Nuclear retention, as well as mRNA maturation processes, have been modelled in various studies as one-step processes, despite experimental evidence of these processes being a chain of multiple random events [13, 14, 70–76, 82]. In Chapter 4 we propose and study a novel stochastic model of gene expression where we divide the mRNA life cycle into multiple stages to incorporate all possible mRNA states during its lifetime.

1.5 Methods

The most common steps and methods that are used for studying stochastic models of gene expression, and also implemented in our research for this thesis, are as follows.

1. A general approach for developing stochastic models of gene expression is to postulate a **Network of Chemical Reactions (CRN)** [83] based on observed species and plausible reaction pathways.

2. The first step for studying any system of biochemical reactions is to write down the **Chemical Master Equation (CME)** for the system of interest [83–85]. Mathematically speaking, the CME is a (finite or infinite-dimensional) system of linear ordinary differential equations (ODEs) that describes the time evolution of the probabilities of observing at a certain time a specific state in the system, given some initial conditions. Solution of the CME requires the application of a combination of various techniques from the theory of differential equations and dynamical systems. Generally, the CME can be solved analytically only in a few cases (see and [86] references therein). Common methods that are used to obtain an exact or an asymptotic solution to the CME are the following.

3. The exact solution for the CME can be obtained numerically by using the Gillespie **Stochastic Simulation Algorithm (SSA)** [87–89]. However, numerical results are not always enough for understanding the behaviour of stochastic systems in different biological limits, while an analytical solution to the CME can provide one with a better understanding.

4. The **Linear Noise Approximation (LNA)** is a commonly used tool for approximating the moments of a probability distribution. It has been shown that the LNA and the CME exactly agree up to second-order moments for a class of chemical systems [90]. However, this method is not sufficient for obtaining an exact closed-form solution for the CME, as it does not in general yield full probability distribution.

5. The **method of probability-generating function** [91] is often used to convert the large system of ODEs which represents the CME, into a system of partial differential equations (PDEs) for the probability-generating functions. Unfortunately, in most cases, solving the obtained system of PDEs is a hard task and hence, other methods are required for a full mathematical analysis.

6. The **method of characteristics** [92] is usually applied to transform the system of PDEs for the probability-generating functions into a dynamical system of a few ODEs. In some cases, at this step, it is possible to obtain an exact closed-form solution of the probability-generating function; however, in most cases, it is not plausible and the application of approximation techniques is needed.

7. A widely used perturbative technique to obtain an asymptotic solution of the system of ODEs after the application of the method of characteristics is the **Geometric Singular Perturbation Theory (GSPT)** [93]. GSPT is a useful tool to study biological systems with a clear separation of time scales [94]. Since it is well known that some processes involved in gene expression occur on a fast time scale compared to other processes, GSPT has seen its application in studies of multi-scale models of gene expression [95].

We include a detailed description of some of these methods in the following part of this section. Additionally, we present an example of mathematical analysis of a simple gene expression model at the end of this section.

1.5.1 Chemical reaction network (CRN)

The underlying mechanisms of stochastic processes are generally complicated because they involve several types of chemical species and chemical reactions; we model these processes as random events that occur at a random point in time. A simple example is the birth-death process of a species, M that we can describe by the following system of reactions:

$$\varnothing \longrightarrow M, \qquad M \longrightarrow \varnothing, \tag{1.1}$$

where the empty set, $\varnothing$ denotes sources and sinks of molecules of species, M. The first reaction models the production (birth) of species, M, while the second reaction models the degradation (death) of M. Another simple example is the underlying mechanism of protein synthesis, which we can describe as,

$$G \xrightarrow{k_1} G + mRNA, \qquad mRNA \xrightarrow{k_2} mRNA + protein. \tag{1.2}$$

In this example, the gene (denoted by G) produces an mRNA molecule with a rate k_1 (transcription) and the mRNA produces a protein molecule with a rate k_2 (translation).

The set of chemical species and chemical reaction constitutes the CRN [83], that is typically represented as,

$$\sum_{i=1}^{N} s_{ij} X_i \xrightarrow{k_j} \sum_{i=1}^{N} r_{ij} X_i \quad \text{for} \quad j = 1, \ldots, R. \tag{1.3}$$

In the above network, X_i for $i = 1, \ldots, N$, denote N different chemical species, which interact via R chemical reactions; k_j is the macroscopic reaction rate constant of the j^{th} reaction where $j = 1, \ldots, R$. Additionally, s_{ij} and r_{ij} (non-negative integers) are stoichiometric coefficients that denote the number of reactant and product molecules, respectively. We define the non-negative integer,

$$m_j = \sum_{i=1}^{N} s_{ij} \tag{1.4}$$

being the order of j^{th} reaction. For example, we call the j^{th} reaction 'unimolecular' if $m_j = 1$ (or first-order reaction) and 'bimolecular' if $m_j = 2$ (or second-order). Our system is called linear if $m_j \leq 1$ for all the reactions in the system [83].

1.5.2 Chemical master equation (CME)

We consider the general CRN given by Eq. 1.3, where the chemical reactions happen with certain probabilities under certain rules and in a well-stirred compartment of volume Ω. We assume dilute conditions in the compartment; i.e., the summed volume of all the molecules in the system is much smaller than the total volume of the compartment. This means that we do not have to take into account the volume of the molecules. Now, we define the vector, $\vec{n} = (n_1, \ldots, n_N)$, where n_i is the number of molecules of species, X_i. The state vector, $\vec{n}$ fully determines the state of the system at any time, t [85]. In a very small time interval Δt, when only one reaction can take place – let us assume that j^{th} reaction occurs – the system changes state from $\vec{n}$ to $\vec{n} + \vec{S_j}$, where $\vec{S_j}$ is a vector whose entries correspond to the j^{th} column of the stoichiometry matrix, $\mathbf{S}$ whose elements are defined as $S_{ij} = r_{ij} - s_{ij}$. The probability of the system being at state, $\vec{n}$ at time, $t + \Delta t$ is then given by:

$$P(\vec{n}; t + \Delta t) = P(\vec{n}; t) + \Delta t \sum_{j=1}^{R} f_j(\vec{n} - \vec{S_j}) P(\vec{n} - \vec{S_j}; t) - \Delta t \sum_{j=1}^{R} f_j(\vec{n}) P(\vec{n}; t). \tag{1.5}$$

In the above equation, we have the following description for the right-hand terms. The term, $P(\vec{n}; t)$ denotes the probability of the system being already in-state, $\vec{n}$ at the time, t and no change in the number of molecules happens, while the rest two terms denote the probability of the system changing from the state, $\vec{n} - \vec{S_j}$ to $\vec{n}$ during the time interval, Δt. In Eq. (1.5), $f_j(\vec{n})$ are the microscopic propensity functions [83, 96] and in the case of mass-action kinetics they are defined as:

$$f_j(\vec{n}) = k_j \Omega \prod_{i=1}^{N} \frac{n_i!}{(n_i - s_{ij})! \Omega^{s_{ij}}}. \tag{1.6}$$

We note that, here, we have assumed that the propensity functions do not depend on time. In the limit of small, Δt and $\Delta t \mapsto dt$, we have that $f_j(\vec{n})dt$ is the probability for the j^{th} reaction to happen in a time interval, dt. In this case, Eq. (1.5) transform to its simple form, which is called the Chemical Master Equation (CME) and it is given by

$$\frac{dP(\vec{n}; t)}{dt} = \sum_{j=1}^{R} f_j(\vec{n} - \vec{S_j}) P(\vec{n} - \vec{S_j}; t) - \sum_{j=1}^{R} f_j(\vec{n}) P(\vec{n}; t). \tag{1.7}$$

The CME describes the stochastic dynamics of the system; specifically, it describes how the joint probability distribution of the copy number of different chemical species evolves in time under spatially homogeneous conditions. We have implicitly assumed the initial condition, $P(\vec{n}_0; t_0)$ which is the probability of the system being at some initial state, $\vec{n}_0$ at some initial time, t_0. The CME was firstly introduced by Donald A McQuarrie [84] and later on, was derived by Daniel T Gillespie [85].

1.5.3 Stochastic simulation algorithm (SSA)

The SSA in the context of chemical kinetics was firstly proposed by Daniel T Gillespie in 1976 [87–89]. By using a Monte-Carlo approach, the SSA generates a statistically correct trajectory (possible solution) of the stochastic process described by the CME in Eq. (1.7). The main steps of the Gillespie algorithm for generating one trajectory (in the case of time-independent propensity functions) are as follows:

1. Initialize the time $t = 0$ and the number of molecules of each species, $\vec{n} = \vec{n}(t = 0)$, and perform the following steps as long as $t \leq t_{\max}$; t_{max} is the time of interest.

2. Generate two independent uniform random numbers on the interval $(0, 1)$: r_1, which will be used in the definition for the time that passes until the next reaction occurs; and r_2, which will be used for the definition of the next reaction that occurs.

3. The time that passes until the next reaction occurs, Δ, is exponentially distributed with parameter,

$$f_0 = \sum_{j=1}^{R} f_j(\vec{n}); \tag{1.8}$$

f_0 is the sum of all the propensity functions evaluated at the current state, $\vec{n}$. Hence, the time interval, Δ is then given by the expression,

$$\Delta = \frac{\ln(1/r_1)}{f_0}. \tag{1.9}$$

4. Identify which reaction is going to occur next by picking the reaction index j_0 to satisfy the inequality

$$\sum_{j=1}^{j_0-1} \frac{f_j(\vec{n})}{f_0} < r_2 \leq \sum_{j=1}^{j_0} \frac{f_j(\vec{n})}{f_0}. \tag{1.10}$$

5. According to which reaction has occurred, update the species vector, $\vec{n}$.

6. Update the time by replacing t with $t + \Delta$.

7. If $t \leq t_{\max}$, then go back to step 2; otherwise, end.

A detailed description of the modified SSA for a model with time-dependent propensity functions can be found in Appendix Section B.5.

As computers have become faster, the SSA has been used to simulate increasingly complex systems. Additionally, an alternative algorithm for the numerical solution of the CME is the finite state projection algorithm proposed by Brian Munsky and Mustafa Khammash [97]. Both algorithms are particularly useful for simulating reactions within cells when the number of reactants is low. However, for complex systems, these algorithms are computationally expensive [97,98] and hence, analytical exact or approximate solutions of the CME appear to be more attractive compared to simulations.

1.5.4 Linear noise approximation (LNA)

As has been mentioned before, LNA is a tool that is widely used to approximate the moments of the probability distribution. LNA is obtained through volume expansion of the CME – a method called system size expansion (SSE) developed by van Kampen [83,98,99]. SSE is a widely accepted approximation of the CME, and we are going to describe here the basic steps for its derivation. Foremost, the following assumption for the SSE application must be held: the microscopic propensity functions in the CME can be expanded as

$$f_j(\vec{n}) = \Omega \sum_{q=0}^{\infty} \Omega^{-q} f_j^{(q)}(\frac{\vec{n}}{\Omega}) \quad \text{for} \quad j = 1, \ldots, R, \tag{1.11}$$

where we remind that Ω is the volume of the system and $f_j^{(0)}$ are the so-called macroscopic propensity functions. Such expansion always exist when the microscopic propensity functions are of the form as in Eq. (1.6) (case of mass-action kinetics). The idea behind the SSE method, is to separate the species concentrations into a deterministic part and a fluctuating part as:

$$\frac{\vec{n}}{\Omega} = \vec{\nu} + \frac{\vec{\epsilon}}{\sqrt{\Omega}}, \tag{1.12}$$

where we have the vectors of species concentrations, $\vec{\nu} = (\nu_1, \ldots, \nu_N)$ and $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$. The terms $\Omega\vec{\nu}$ and $\sqrt{\Omega}\vec{\epsilon}$ can be seen as deterministic macroscopic and stochastic mesoscopic contributions to the system state $\vec{n}$, respectively. Thus, as $\Omega \to \infty$ leads to $\vec{n}/\Omega \to \vec{\nu}$. We have that, ν_i is the concentration of species X_i and is defined as the solution of the deterministic rate equation,

$$\frac{d\nu_i}{dt} = \sum_{j=1}^{R} S_{ij} f_j^{(0)}(\vec{\nu}) \quad \text{for} \quad i = 1, \ldots, N, \tag{1.13}$$

where the macroscopic propensity function from Eq. (1.11), $f_j^{(0)}$ is defined as

$$f_j^{(0)}(\vec{\nu}) = \lim_{\Omega\to\infty} \frac{f_j(\Omega\vec{\nu})}{\Omega} \quad \text{for} \quad j = 1, \ldots, R. \tag{1.14}$$

The deterministic rate equation in Eq. (1.13) can be written in matrix form as,

$$\frac{d\vec{\nu}}{dt} = \mathbf{S} \cdot \vec{f}^{(0)}(\vec{\nu}), \tag{1.15}$$

where $\mathbf{S}$ is the stoichiometry matrix and $\vec{f}^{(0)}(\vec{\nu}) = (f_1^{(0)}(\vec{\nu}), \ldots, f_R^{(0)}(\vec{\nu}))$ is the vector of macroscopic propensity functions.

Now, by taking into account Eq. (1.12), we express the vector of molecule numbers, $\vec{n}$ in terms of the vector $\vec{\epsilon}$ and rewrite the probability function as:

$$\Pi(\vec{\epsilon}; t) = \Omega^{\frac{N}{2}} P(\Omega\vec{\nu} + \sqrt{\Omega}\vec{\epsilon}; t). \tag{1.16}$$

For the SSE, we plug the Eq. (1.11), Eq. (1.12) and Eq. (1.16) into the CME in Eq. (1.7), and perform Taylor expansion around $\vec{\epsilon} = \vec{0}$ ($\vec{0}$ is a N - dimensional zero vector here). In this way, we obtain an expansion of the CME in powers of $\Omega^{-1/2}$ and by truncating this expansion to zeroth order (Ω^0) we derive the LNA for the CME, which is given by:

$$\frac{d\Pi(\vec{\epsilon}; t)}{dt} = \Big[-\sum_{i=1}^{N} \frac{\partial}{\partial\epsilon_i} \sum_{q=1}^{N} J_{iq}\epsilon_q + \frac{1}{2} \sum_{i=1}^{N}\sum_{q=1}^{N} D_{iq} \frac{\partial}{\partial\epsilon_i}\frac{\partial}{\partial\epsilon_q} \Big] \Pi(\vec{\epsilon}; t) + \mathcal{O}(\Omega^{-1/2}); \tag{1.17}$$

this is a linear Fokker-Planck equation (LFPE) [98], where we have defined:

$$J_{iq} = \frac{\partial}{\partial\nu_q} \sum_{j=1}^{R} S_{ij} f_j^{(0)}(\vec{\nu}) \quad \text{and} \quad D_{iq} = \sum_{j=1}^{R} S_{ij} S_{qj} f_j^{(0)}(\vec{\nu}). \tag{1.18}$$

If follows from the definitions that the $(N \times N)$ - dimensional matrix $\mathbf{J}$ is the Jacobian matrix of the rate equations given in Eq. (1.13), while the $(N \times N)$ - dimensional matrix $\mathbf{D} = \mathbf{S} \cdot \mathbf{Diag}(\vec{f}^{(0)}) \cdot \mathbf{S}^T$ is the Diffusion matrix of the system, where $\mathbf{Diag}(\vec{f}^{(0)})$ is a diagonal matrix whose elements are the entries in the vector $\vec{f}^{(0)}$. Note that the elements of both matrices are dependent on the solution of the rate equation, $\vec{\nu}$ and are thus generally time-dependent.

The linear Fokker-Planck equation Eq. (1.17) can be solved by a multivariate normal distribution under certain initial conditions. By multiplying Eq. (1.17) with ϵ_i and integrating over $\vec{\epsilon}$, one can obtain an ODE for the first moments of the distribution, $\langle\epsilon_i\rangle$ for $i = 1, \ldots, N$. By assuming that the mean concentration is zero for zero time, one can easily find that $\langle\epsilon_i\rangle = 0$ for all times. Hence, the solution of Eq. (1.17) is a multivariate normal distribution with zero mean. The vectors $\vec{\epsilon}$ and $\vec{n}$ are related through the linear relationship specified in Eq. (1.12); this means that the distribution of $\vec{n}$ is also a multivariate normal distribution, the mean of which can be derived from the rate equation given in Eq. (1.15) for $\vec{\nu} = \langle\vec{n}\rangle$. Additionally, one can also show that the time-dependent covariance matrix, $\mathbf{C}$ of the distribution over $\vec{n}$ satisfies the Lyapunov equation [98, 100]:

$$\frac{d\mathbf{C}}{dt} = \mathbf{J} \cdot \mathbf{C} + \mathbf{C} \cdot \mathbf{J}^T + \mathbf{D}. \tag{1.19}$$

LNA describes the lowest order fluctuations in the deterministic mean. For a system with linear reactions, the first two moments of the distributions predicted by the LNA agree exactly with the first two moments predicted by the CME. However, this is not generally the case for systems with non-linear reactions [101, 102]. It is therefore appealing to consider other approximations of the CME apart from the SSE.

The chemical Fokker-Planck equation (CFPE) and the corresponding chemical Langevin equation (CLE) are commonly used approximations of the CME. Kramers and Moyal developed a Taylor series expansion of the CME and obtained the CFPE by assuming that all terms with derivatives greater than two are negligible [103]. A major and important difference between the CME and the CFPE is that the molecule number is a positive integer for the CME while it is a real number for the CFPE. Later on, Gillespie revived the question of the validity of the CFPE by deriving the chemical Langevin equation (CLE) without invoking truncation of the Kramers-Moyal expansion of the CME [104]; The CLE is exactly equivalent to the CFPE in the sense that its solution generates exact sample paths of the CFPE. Grima et al. have shown that the CFPE is generally more accurate than the LFPE (or equivalently the LNA) [105]. The LFPE, which is obtained by considering only the lowest order (limit of large volumes) in the perturbative expansion of the CME in powers of the inverse square root of the system volume (the system-size expansion), is different from the CFPE; of particular concern is that the LFPE is linear, whereas the CFPE is nonlinear. Taking into account higher-order terms in the system-size expansion does not lead to the CFPE as well. However, interestingly, in the limit of large volumes, the CFPE does reduce to LFPE [105]. This means that in the limit of large volumes, the Lyapunov equation takes the form as in Eq. (1.19), either it is derived from the LFPE or the CFPE.

1.5.5 Geometric Singular Perturbation Theory (GSPT)

Biological systems are often characterised by processes that occur on different time scales. Systems that describe stochastic models of gene expression frequently exhibit reaction dynamics with these kinds of features; e.g. mRNA is short-lived compared to protein. This results in the appearance of small parameters in the CME describing the gene regulatory networks, and thus gives rise to the application of perturbation techniques. GSPT was originally developed by N. Fenichel [93,106] and has been used as a useful tool for the analysis of 'fast-slow' systems [107, 108]; GSPT allows us to study the fast and the slow dynamics of the systems separately. Some examples of its application to models of gene expression can be found in [43, 95, 109, 110]. Here, we present a brief overview of the GSPT.

We consider a system of first-order autonomous ordinary differential equations (ODEs) in the general standard form also referred to as the 'slow system'

$$\varepsilon \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{y}, \varepsilon), \tag{1.20a}$$

$$\dot{\mathbf{y}} = \mathbf{g}(\mathbf{x}, \mathbf{y}, \varepsilon), \tag{1.20b}$$

where $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^l$, with $m, l \in \mathbb{N}$. For simplicity, the functions $\mathbf{f} : \mathbb{R}^m \times \mathbb{R}^l \times \mathbb{R}^+ \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^m \times \mathbb{R}^l \times \mathbb{R}^+ \to \mathbb{R}^l$ are assumed to be of class $\mathcal{C}^\infty$ (also called smooth; i.e., it has derivatives of all orders) in all their arguments. Also, $0 < \varepsilon \ll 1$ is a real small perturbation parameter, and the overdot denotes differentiation with respect to the 'slow time' τ. Now, we introduce the new 'fast time' $t = \tau/\varepsilon$, which we substitute into Eq. (1.20) to find the 'fast system'

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}, \mathbf{y}, \varepsilon), \tag{1.21a}$$

$$\mathbf{y}' = \varepsilon \mathbf{g}(\mathbf{x}, \mathbf{y}, \varepsilon), \tag{1.21b}$$

where the prime denotes the derivative with respect to t. For positive ε, the systems in Eq. (1.20) and Eq. (1.21) are equivalent; however, in the singular limit of $\varepsilon \to 0$, we obtain two different systems: setting $\varepsilon = 0$ in Eq. (1.20), we have the 'reduced problem'

$$\mathbf{0} = \mathbf{f}(\mathbf{x}, \mathbf{y}, 0), \tag{1.22a}$$

$$\dot{\mathbf{y}} = \mathbf{g}(\mathbf{x}, \mathbf{y}, 0), \tag{1.22b}$$

while by setting $\varepsilon = 0$ in Eq. (1.21), we obtain the 'layer problem'

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}, \mathbf{y}, 0), \tag{1.23a}$$
$$\mathbf{y}' = \mathbf{0}. \tag{1.23b}$$

The 'reduced problem' in Eq. (1.22) implies that the flow of $\mathbf{y}$ is constrained to lie on the l - dimensional 'critical manifold' $\mathcal{S}_0$ that is defined by $\mathbf{f} = \mathbf{0}$, while $\mathbf{x}$ is assumed to vary in an appropriately chosen subset of $\mathbb{R}^m$. The 'layer problem' in Eq. (1.23) implies that $\mathbf{y}$ is merely a parameter which parametrizes the m - dimensional flow of $\mathbf{x}$, the equilibria of which are located on $\mathcal{S}_0$. The variable $\mathbf{x}$ is referred to as the 'fast variable', while $\mathbf{y}$ is referred to as the 'slow variable'.

The aim of GSPT is to infer the flow of the 'slow system' and the 'fast system' from Eq. (1.20) and Eq. (1.21), respectively, from the simplified dynamics of the corresponding 'reduced problem' and 'layer problem' from Eq. (1.22) and Eq. (1.23). One of the major assumptions of GSPT made for the critical manifold $\mathcal{S}_0$ is the following: $\mathcal{S}_0$ is compact and 'normally hyperbolic', i.e., the eigenvalues of the Jacobian matrix evaluated on $\mathcal{S}_0$, $D_{\mathbf{x}}\mathbf{f}(\mathbf{x}, \mathbf{y}, 0)$, are uniformly bounded away from the imaginary axis (please refer to [107] for more details). Now, we assume that the Jacobian matrix has m_s eigenvalues with negative real part and m_u eigenvalues with positive real part, where we have that $m_u + m_s = m$. Any point of the critical manifold $\mathcal{S}_0$ is an equilibrium point of the layer problem, and hence, each such point admits a m_u - dimensional unstable manifold and m_s - dimensional stable manifold. By taking the union of these manifolds with the space of the critical manifold, we can present the following definitions: $\mathcal{W}^{\mathrm{u}}(\mathcal{S}_0)$ and $\mathcal{W}^{\mathrm{s}}(\mathcal{S}_0)$ are the corresponding $(m_u + l)$ - dimensional unstable and $(m_s + l)$ - dimensional stable manifolds for $\mathcal{S}_0$, respectively. By having these, Fenichel's theorems imply that for $\varepsilon > 0$ and sufficiently small, the $\mathcal{S}_0$, $\mathcal{W}^{\mathrm{u}}(\mathcal{S}_0)$ and $\mathcal{W}^{\mathrm{s}}(\mathcal{S}_0)$ manifolds will persist (i.e. they are invariant under the flow of the 'slow system'); please refer to [107] for more details.

GSPT appears to be very successful in the case of normally hyperbolic critical manifolds, however, there are limitations of GSPT when normal hyperbolicity breaks down; e.g. one cause for the loss of hyperbolicity can be due to a zero eigenvalue of the Jacobian. The breakdown of normal hyperbolicity often gives rise to interesting dynamics, such as the formation of periodic solutions, which can be treated with 'blow-up' techniques [94, 111] (and references therein).

1.5.6 Example

In this paragraph, we are going to present an example of a detailed mathematical analysis of a stochastic model of gene expression, where we implement all the methods described in this section. We will study a simple extension of the telegraph model to incorporate the process of nuclear retention. A schematic representation of the model is given in Fig. 1.2.

Detailed description of the model

The stochastic model of gene expression in Fig.1.2 models in one step the mechanism of nuclear retention of RNA; i.e. this model explicitly takes into account the process of RNA export from the nucleus into the cytoplasm. In this model, the promoter of the gene can be in two states: active, G_{on} and inactive, G_{off}; the promoter activation happens at rate s_u, while the inverse reaction occurs at rate s_b. Transcription initiation of nuclear RNA (denoted by M_N) can happen with rate r only when the promoter is active. When a molecule of M_N is produced, it gets exported into the cytoplasm at rate k, and it becomes cytoplasmic mRNA (denoted by M_C). Finally, we have that, M_N degrades in the nucleus with rate d_N, while M_C degrades in the cytoplasm with rate d_m.

Here, we note that if we ignore the cytoplasmic species in this model by setting $k = 0$, then our model simplifies to the well-known telegraph model.

Network of chemical reactions

In this model we have four species (G_{on}, G_{off}, M_N, M_C) and six reactions. However, there is only one promoter in the system hence, the species G_{off} can be eliminated from the system since it can

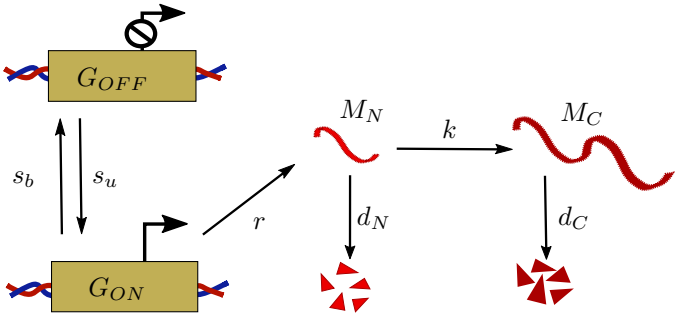


Figure 1.2: *Stochastic model of gene expression with nuclear retention.* In this model, the promoter of the gene switches between active and inactive states, G_{on} and G_{off}, respectively; the switching happens with rates s_u and s_b. While the promoter is active, the transcription initiation of the nuclear RNA (denoted by M_N) may occur with rate r, which is followed by the process of RNA export into the cytoplasm with rate k; the cytoplasmic RNA is denoted by M_C. Finally, the nuclear and cytoplasmic mRNA decay with rates d_N and d_C, respectively.

be always expressed in terms of G_{on} (see below). The system describing the network of chemical reactions according to Eq. (1.3) is given by,

$$\begin{aligned}
&G_{off} \underset{s_b}{\overset{s_u}{\rightleftharpoons}} G_{on} && \text{(promoter switching)},\\
&G_{on} \xrightarrow{r} G_{on} + M_N && \text{(nuclear RNA transcription)},\\
&M_N \xrightarrow{k} M_C && \text{(RNA export into the cytoplasm)},\\
&M_N \xrightarrow{d_N} \varnothing && \text{(nuclear RNA degradation)},\\
&M_C \xrightarrow{d_C} \varnothing && \text{(cytoplasmic mRNA degradation)}.
\end{aligned} \tag{1.24}$$

We note that this is a linear system since all the reactions are of first-order. The stoichiometry matrix of the system is given by,

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}. \tag{1.25}$$

Chemical master equation

Here, we begin with the following definition; we define g^* and g being the number of inactive and active promoters in the system, respectively. Since we have only one promoter and it can be in two states, it means that $g^* + g = 1$ ($g^* = 1 - g$) and hence, g is a binary number that denotes the promoter state; i.e. $g = 1$ when the promoter is active and $g = 0$ when it is inactive. Now, we define the vector of species numbers $\vec{n} = (g,\ n_N,\ n_C)$, where n_N denotes the number of M_N molecules and n_C denotes the number of M_C molecules. Therefore, the vector of propensity functions for the system in Eq. (1.24) according to Eq. (1.6) is given by

$$\vec{f} = (s_u(1-g),\ s_b g,\ rg,\ kn_N,\ d_N n_N,\ d_C n_C). \tag{1.26}$$

Now, we define $P(\vec{n};t)$ being the probability of finding the system in state $\vec{n}$ at time t. Since, the promoter can be in two states, we additionally define $P_0(\vec{n};t)$ and $P_1(\vec{n};t)$ being the probability of finding the system in state $\vec{n}$ at time t when the promoter is inactive and active, respectively; these probabilities satisfy the relationship: $P_0(\vec{n};t) + P_1(\vec{n};t) = P(\vec{n};t)$. In this way, we can eliminate the dependence of the probability functions on g. According to Eq. (1.7) the CMEs for our model

are given by the following two ODEs:

$$\begin{aligned}
\frac{dP_0(n_N,n_C;t)}{dt} =& s_b P_1(n_N,n_C;t) - s_u P_0(n_N,n_C;t) + k(n_N+1)P_0(n_N+1,n_C-1;t) \\
&- k n_N P_0(n_N,n_C;t) + d_N(n_N+1)P_0(n_N+1,n_C;t) - d_N n_N P_0(n_N,n_C;t) \\
&+ d_C(n_C+1)P_0(n_N,n_C+1;t) - d_C n_C P_0(n_N,n_C;t), \\
\frac{dP_1(n_N,n_C;t)}{dt} =& s_u P_0(n_N,n_C;t) - s_b P_1(n_N,n_C;t) + k(n_N+1)P_1(n_N+1,n_C-1;t) \\
&- k n_N P_1(n_N,n_C;t) + d_N(n_N+1)P_1(n_N+1,n_C;t) - d_N n_N P_1(n_N,n_C;t) \\
&+ d_C(n_C+1)P_1(n_N,n_C+1;t) - d_C n_C P_1(n_N,n_C;t) \\
&+ r P_1(n_N-1,n_C;t) - r P_1(n_N,n_C;t).
\end{aligned} \tag{1.27}$$

The above equation can be rewritten in a compact form as

$$\begin{aligned}
\partial_t P_0 =& s_b P_1 - s_u P_0 + k(\mathbb{E}_{n_N}\mathbb{E}_{n_C}^{-1} - 1)n_N P_0 + d_N(\mathbb{E}_{n_N} - 1)n_N P_0 + d_C(\mathbb{E}_{n_C} - 1)n_C P_0, \\
\partial_t P_1 =& s_u P_0 - s_b P_1 + k(\mathbb{E}_{n_N}\mathbb{E}_{n_C}^{-1} - 1)n_N P_1 + d_N(\mathbb{E}_{n_N} - 1)n_N P_1 + d_C(\mathbb{E}_{n_C} - 1)n_C P_1 \\
&+ r(\mathbb{E}_{n_N}^{-1} - 1)P_1,
\end{aligned} \tag{1.28}$$

where $\mathbb{E}_{n_i}^{c} f(n_1,\ldots,n_N) = f(n_1,\ldots,n_i+c,\ldots,n_N)$ with $c \in \mathbb{Z}$, denotes the standard step operator acting on function f [83].

Linear noise approximation

Now, by using the LNA, we can find analytical expressions for the first two moments of the RNA distributions. Note that our system is linear because it is described by first-order reactions; hence, the first two moments that can be obtained from LNA are exactly the same as the ones that can be obtained by using the CME. Firstly, to find the mean number of RNA molecules, we can use Eq. (1.15), which for this model reads as

$$\partial_t\langle\vec{n}\rangle = \mathbf{S}\cdot\vec{f}(\langle\vec{n}\rangle) \quad \Longrightarrow \quad \begin{aligned} \partial_t\langle g\rangle &= f_1 - f_2, \\ \partial_t\langle n_N\rangle &= f_3 - f_4 - f_5, \\ \partial_t\langle n_C\rangle &= f_5 - f_6, \end{aligned} \tag{1.29}$$

where $\langle\cdot\rangle$ denotes the mean and f_j for $j = 1,\ldots,6$ are the entries of the vector $\vec{f}$. For a steady-state solution of the system in Eq. (1.29), we set the time-derivatives to zero and obtain the following solution:

$$\begin{aligned}
\langle g\rangle &= \frac{s_u}{s_u+s_b}, \\
\langle n_N\rangle &= \frac{s_u}{s_u+s_b}\frac{r}{k+d_N}, \\
\langle n_C\rangle &= \frac{s_u}{s_u+s_b}\frac{r}{k+d_N}\frac{k}{d_C},
\end{aligned} \tag{1.30}$$

where $\langle g\rangle$ also represents the probability of the promoter being in its active state. For the second moments of RNA distributions, we need to solve the Lyapunov equation given in Eq. (1.19); for finding the solution of the covariance matrix $\mathbf{C}$, we again set the time-derivative in this equation to zero, and now it reads as

$$\begin{aligned}
&\mathbf{J}\cdot\mathbf{C} + \mathbf{C}\cdot\mathbf{J}^T + \mathbf{D} = \mathbf{0} \quad \Longrightarrow \\
&\sum_{q=1}^{3}(J_{iq}C_{qj} + J_{jq}C_{iq}) + D_{ij} = 0 \quad \text{for } \; i,j = 1,2,3,
\end{aligned} \tag{1.31}$$

where $\mathbf{0}$ is a zero matrix. The (3×3) - dimensional Jacobian, $\mathbf{J}$ and the diffusion, $\mathbf{D}$ matrices are defined in Eq. (1.18) and for our system, in this example, they read as

$$\mathbf{J} = \begin{pmatrix} -s_u - s_b & 0 & 0 \\ r & -k - d_N & 0 \\ 0 & k & -d_C \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} f_1 + f_2 & 0 & 0 \\ 0 & f_3 + f_4 + f_5 & -f_4 \\ 0 & -f_4 & f_4 + f_6 \end{pmatrix}. \tag{1.32}$$

At this moment, one has all the ingredients to find the solution for all the elements of the covariance matrix, C_{ij} with $i, j = 1, 2, 3$. Here, we present only the form of the variance expressions:

$$\begin{aligned} C_{22} &= Var(n_N) = \langle n_N \rangle + \langle n_N \rangle^2 \frac{s_b}{s_u} \frac{k + d_N}{s_u + s_b + k + d_N}, \\ C_{33} &= Var(n_C) = \langle n_C \rangle + \langle n_C \rangle^2 \frac{s_b}{s_u} \frac{k + d_N}{s_u + s_b + k + d_N} \frac{s_u + s_b + k + d_N + d_C}{s_u + s_b + d_C} \frac{d_C}{k + d_N + d_C}. \end{aligned} \tag{1.33}$$

There are some useful results from our up to point mathematical analysis of this model. For example, from Eq. (1.30) we obtain the following relationship:

$$\frac{\langle n_C \rangle}{\langle n_N \rangle} = \frac{k}{d_C}. \tag{1.34}$$

This means that the ratio of the nuclear export rate over the cytoplasmic degradation can be estimated experimentally only by measuring the mean number of nuclear and cytoplasmic RNA or their ratio. Note that is is valid in steady-state, since Eq. (1.34) was obtained in the limit of long times. Another potentially useful result that comes from this analysis is related to transcriptional noise. The coefficient of variation squared, which is defined as,

$$CV^2 = \frac{Variance}{Mean^2}, \tag{1.35}$$

is frequently used as a measurement of noise. Hence, by defining the ratio,

$$\frac{CV_C^2}{CV_N^2} = \frac{\langle n_N \rangle}{\langle n_C \rangle} \frac{Var(n_C)}{Var(n_N)}, \tag{1.36}$$

one can determine conditions under which the cytoplasmic RNA noise is less than the nuclear mRNA noise ($CV_{n_C}^2 < CV_{n_N}^2$) when the system is in steady-state.

Probability-generating function

Solving the system of ODEs that compose the CMEs in Eq. (1.28) is a hard task. The step in this paragraph involves transforming this system of ODEs into a system of PDEs, which can be potentially solved. First, we define the probability generating functions for our system as,

$$\begin{aligned} F_q(z_N, z_C; t) &= \sum_{n_N=0}^{\infty} \sum_{n_C=0}^{\infty} P_q(n_N, n_C; t) z_N^{n_N} z_C^{n_C} \quad \text{for} \quad q \in \{0, 1\} \\ F(z_N, z_C; t) &= F_0(z_N, z_C; t) + F_1(z_N, z_C; t), \end{aligned} \tag{1.37}$$

where $z_N,\ z_C \in [0, 1]$ are real variables. Now, by using these definitions, we rewrite our CMEs as

$$\begin{aligned} \partial_t F_0 =& s_b F_1 - s_u F_0 - k(z_N - z_C)\partial_{z_N} F_0 - d_N(z_N - 1)\partial_{z_N} F_0 - d_C(z_C - 1)\partial_{z_C} F_0, \\ \partial_t F_1 =& s_u F_0 - s_b F_1 - k(z_N - z_C)\partial_{z_N} F_1 - d_N(z_N - 1)\partial_{z_N} F_1 - d_C(z_C - 1)\partial_{z_C} F_1 \\ &+ r(z_N - 1)F_1. \end{aligned} \tag{1.38}$$

The above system of PDEs is still impossible to solve analytically and hence, we use the mathematical tool that follows.

Method of characteristics

We define a real parameter s being our characteristic variable, and now we convert the system of PDEs given in Eq. (1.38) into the following system of ODEs:

$$\begin{aligned} \partial_s t &= 1, \\ \partial_s z_N &= k(z_N - z_C) + d_N(z_N - 1), \\ \partial_s z_C &= d_C(z_C - 1), \\ \partial_s F_0 &= s_b F_1 - s_u F_0, \\ \partial_s F_1 &= s_u F_0 - s_b F_1 + r(z_N - 1)F_1. \end{aligned} \tag{1.39}$$

Note that the first equation in the above system gives us $s = t$; hence, the time variable can be used as our characteristic variable. The existence of an integral-form solution to the system above follows from the fact that the reaction scheme for our model contains first-order reactions only; however, it is hard to obtain a closed-form solution. Here, we note that the exact solution for the marginal probability distribution for the molecule number of the nuclear RNA is known, and it is the solution of the telegraph model with the degradation rate being $k + d_N$. Let us assume that we are interested in finding an analytical closed-form solution for the marginal distribution of the molecule numbers of the final product of the gene in our model, which is the cytoplasmic mRNA. In this case, we can obtain an approximate solution by using GSPT.

Geometric singular perturbation theory

For convenience in our derivations that follow, we introduce new definitions: $u_N = z_N - 1$, $u_C = z_C - 1$ and we rescale the model parameters as, $\kappa = k/d_C$, $\delta_k = d_N/k$, $\sigma_b = s_b/d_C$, $\sigma_u = s_u/d_C$, $\rho = r/d_C$ and $\tau = d_C t$. Now, we rewrite the system in Eq. (1.39) as

$$\begin{aligned} \partial_\tau u_N &= \kappa(u_N - u_C) + \kappa\delta_k u_N, \\ \partial_\tau u_C &= u_C, \\ \partial_\tau F_0 &= \sigma_b F_1 - \sigma_u F_0, \\ \partial_\tau F_1 &= \sigma_u F_0 - \sigma_b F_1 + \rho u_N F_1. \end{aligned} \tag{1.40}$$

Experimental data show that the nuclear export process is generally faster than the cytoplasmic degradation [13, 74]; this means that $k \gg d_C$ and thus we can define a real small parameter $\varepsilon = 1/\kappa \ll 1$ and rewrite the system in Eq. (1.40) as

$$\begin{aligned} \varepsilon \cdot \partial_\tau u_N &= u_N - u_C + \delta_k u_N, \\ \partial_\tau u_C &= u_C, \\ \partial_\tau F_0 &= \sigma_b F_1 - \sigma_u F_0, \\ \partial_\tau F_1 &= \sigma_u F_0 - \sigma_b F_1 + \rho u_N F_1. \end{aligned} \tag{1.41}$$

By comparing the systems in Eq. (1.41) and Eq. (1.20), we have that according to GSPT the one in Eq. (1.41) is the 'slow system', τ is our slow timescale, u_N is the fast variable, while u_C, F_0 and F_1 are the slow variables of the system. Additionally, we have that $m = 1$ and $l = 3$. Now, we introduce a new fast time $\tau_f = \tau/\varepsilon$, which we substitute into Eq. (1.41) to find the 'fast system'

$$\begin{aligned} \partial_{\tau_f} u_N &= u_N - u_C + \delta_k u_N, \\ \partial_{\tau_f} u_C &= \varepsilon \cdot u_C, \\ \partial_{\tau_f} F_0 &= \varepsilon \cdot (\sigma_b F_1 - \sigma_u F_0), \\ \partial_{\tau_f} F_1 &= \varepsilon \cdot (\sigma_u F_0 - \sigma_b F_1 + \rho u_N F_1); \end{aligned} \tag{1.42}$$

this system can be compared to the one given in Eq. (1.21). The reduced problem (see Eq. (1.22)) for the system in Eq. (1.41) implies that the flow of $(u_C,\ F_0,\ F_1)$ is constrained to lie on the $(l = 3)$ - dimensional 'critical manifold' $\mathcal{S}_0$ that is defined by $\mathbf{f} = \mathbf{0}$:

$$u_N - u_C + \delta_k u_N = 0 \quad \Longrightarrow \quad u_N = u_C \mu, \tag{1.43}$$

where $\mu = k/(k + d_N)$ represents the surviving probability of the nuclear RNA molecule. The parameters u_C, F_0 and F_1 are assumed to vary in an appropriately chosen subset of $\mathbb{R}$. From the layer problem (see Eq. (1.23)) of the system in Eq. (1.42), we conclude that $\mathbf{y} = (u_C,\ F_0,\ F_1)$ is a parameter which parametrizes the $(m = 1)$ - dimensional flow of $\partial_{\tau_f} u_N$, the equilibria of which are located on $\mathcal{S}_0$. The Jacobian matrix layer problem has positive eigenvalue $\lambda = 1 + \delta_k$; hence, the critical manifold $\mathcal{S}_0$ is 'normally hyperbolic' – and, in fact, normally repelling – with an $(m+l = 4)$ - dimensional unstable manifold $\mathcal{W}^u(\mathcal{S}_0)$.

The GSPT thus implies that $\mathcal{S}_0$ will persist, for ε positive and sufficiently small, as a slow manifold $\mathcal{S}_\varepsilon$ that is (locally) invariant, smooth, and $\mathcal{O}(\varepsilon)$-close to $\mathcal{S}_0$. As the unstable manifold $\mathcal{W}^u(\mathcal{S}_0)$ equals the entire phase space of Eq. (1.41), it trivially persists as the unstable manifold $\mathcal{W}^u(\mathcal{S}_\varepsilon)$ for $\mathcal{S}_\varepsilon$.

Probability distribution for molecule numbers of cytoplasmic RNA

By using the GSPT, we have managed to separate the system from Eq. (1.40) into a slow system as it is given by Eq. (1.41) and a fast system as it is given by Eq. (1.42). Basically, the assumption that the nuclear export rate is fast, makes the variable u_N of the system being fast, and it can be neglected while studying the dynamics of the slow system by using Eq. (1.43). Hence, for $\varepsilon = 0$, we rewrite the slow system as

$$\begin{aligned} \partial_\tau u_C &= u_C, \\ \partial_\tau F_0 &= \sigma_b F_1 - \sigma_u F_0, \\ \partial_\tau F_1 &= \sigma_u F_0 - \sigma_b F_1 + \rho\mu u_C F_1. \end{aligned} \tag{1.44}$$

The first equation in the above systems provides us with the following chain rule: $\partial_\tau = u_C \partial_{u_C}$, which we use to rewrite the rest two equations as

$$\begin{aligned} u_C \partial_{u_C} F_0 &= \sigma_b F_1 - \sigma_u F_0, \\ u_C \partial_{u_C} F_1 &= \sigma_u F_0 - \sigma_b F_1 + \rho\mu u_C F_1. \end{aligned} \tag{1.45}$$

By summing the two equations in the above system, we obtain the relation

$$\partial_{u_C} F = \rho\mu F_1. \tag{1.46}$$

The first equation in Eq. (1.45), the Eq. (1.46) and the relation $F = F_0 + F_1$, all together give us the following confluent hypergeometric differential equation (also known as Kummer's equation) [44]:

$$u_C \partial_{u_C}^2 F + (\sigma_u + \sigma_b - \rho\mu u_C)\partial_{u_C} F - \sigma_u \rho\mu F = 0, \tag{1.47}$$

which admits the solution

$$F(u_C) = C \cdot {}_1F_1\Big(\sigma_u; \sigma_u + \sigma_b; \rho\mu u_C\Big). \tag{1.48}$$

where ${}_1F_1$ denotes the confluent hypergeometric function; here, we consider only one of two independent fundamental solutions of Kummer's differential equation, as we are seeking a solution in steady-state where the variable u_C is bounded. The constant C in Eq. (1.48) is a constant of integration that is determined from the normalisation condition on the full generating function, $F|_{u_C=0} = 1$, which gives us $C = 1$. The probability distribution $P(n_C)$ of cytoplasmic mRNA can thus be found from the formula,

$$P(n_C) = \frac{1}{n_C!} \frac{d^{n_C}}{du_C{}^{n_C}} F(u_C)|_{u_C=-1}, \tag{1.49}$$

which yields the analytical expression,

$$P(n_C) = \frac{1}{n_C!} \frac{(s_u)_n}{(s_b + s_u)_n} (\rho\mu)^{n_C} {}_1F_1\big(\sigma_u + n_C; \sigma_u + \sigma_b + n_C; -\rho\mu\big), \tag{1.50}$$

where $(a)_s = \Gamma(a+s)/\Gamma(a)$ is the Pochhammer symbol.

In this section, we have explained that mathematical analysis of even very simple stochastic models of gene expression is demanding and requires the application of several mathematical tools. In this example, we have shown all the steps in our derivations taken in order to obtain a steady-state distribution of the molecule numbers of cytoplasmic mRNA from the model described in Eq. (1.24). This example could be potentially extended and present a way of obtaining a time-dependent distribution for cytoplasmic mRNA, but it is outside our interests in this section since the reader can find our work on obtaining time-dependent distributions in Chapter 2 and Chapter 4.

1.6 The layout of the thesis

We include a lay summary at the beginning of each main chapter of this thesis to help communicate research to readers who may be outside the specific research area.

Chapter 2 presents a mathematical analysis of a stochastic model of gene expression with polymerase recruitment and polymerase pause release. We begin with an introduction in Section ,2.1 and continue with a detailed description of the model in Section 2.2. We derive an exact steady-state and an approximate time-dependent distributions of mRNA molecule numbers in Section 2.3 and Section 2.4, respectively. Additionally, we obtain a steady-state distribution for protein molecule numbers in Section 2.5 and finish this chapter with a discussion in Section 2.6.

Chapter 3 focuses on mathematical analysis of novel detailed stochastic models of transcription; we introduce two models, one without and one with a polymerase pausing mechanism. We include an introduction in Section 3.1, which is followed by construction and analysis of the model without pausing in Sections 3.2-3.4. We present our study of the extended model with polymerase pausing in Section 3.5. We conclude this chapter with a summary/discussion and proposed extensions of the studied models in Section 3.6 and Section 3.7, respectively.

In Chapter 4 we present our work on a stochastic model of gene expression that considers time-dependent stimuli. We start with an introduction in Section 4.1, then move to model description and its reduction in Section 4.2. Eventually, we present the main mathematical analysis of the model and the obtained results in Section 4.3. In Section 4.4 we show our analysis of the model for a general case of stimuli functions, and finally, we summarize our conclusions in Section 4.5. We finish the chapter with Section 4.6, where we discuss how a modified version of our model can be used to study a gap gene of a fruit fly embryo.

Chapter 5 is devoted to our conclusions and discussion about possible future research directions. In order not to disrupt the reading, some technical derivations are presented in the Appendix.

Chapter 2

Stochastic model of gene expression with polymerase recruitment and pause release

This chapter contains published work. Please see the Declaration of Authorship for details.

Lay summary

RNA transcription is a process that starts with the binding of general transcription factors (proteins that participate in transcription) to the promoter region of a gene. The action of these factors results in the attraction and recruitment of a polymerase molecule (an enzyme that is important for transcription) to the promoter, which subsequently initiates transcription of the corresponding gene. Experimental observations show that the transcription initiation process can start when the promoter of the gene is transcriptionally active, which is the case when a transcription initiation complex (complex of proteins, including the RNA polymerase enzyme and its various accessory proteins) has been properly formed. The formation of the complex occurs in multiple steps, which indicates that the promoter can actually take many states (transcriptionally active and inactive) during this formation process; a number of studies have built models by taking this evidence into account.

A recent work by C. R. Bartman et al. [65] has suggested a stochastic model where the promoter appears in three different states. In this study, the experiments were performed on single cells by using a combination of two experimental methods – so-called RNA fluorescence in situ hybridization (FISH) and chromatin immunoprecipitation sequencing (ChIP-seq) – to quantify the polymerase molecules and nascent RNA. Additionally, Bartman et al. constructed four possible models of transcription, where two of them are characterised by a promoter that can switch between two states and the other two are characterised by a promoter that can switch between three states; all the models specify different networks of reactions. Then, computation simulations were used for all the models in order to identify the one that predicted best the experimental outcome. It was observed that the model that is consistent with experimental data is featured a promoter that fluctuates between three different states, while it includes *polymerase dynamics*; we refer to it as the "multi-scale" model.

In the multi-scale model, the promoter is assumed to be in a non-permissive state when it is free of transcription factors and polymerases. When a transcription factor is bound, the promoter obtains its permissive state and can fluctuate between being transcriptionally active or inactive, depending on the binding state of the polymerase molecule; i.e., the promoter is in its permissive-inactive state if no polymerase is present, and it switches to its permissive-active state after *polymerase recruitment*. In the case of transcription factor detachment from the gene, the promoter

switches back to the non-permissive state. In this model, only one polymerase molecule is allowed to bind each promoter-proximal region at a time. The permissive-active promoter state represents a state in which a polymerase is bound and paused; however, the *polymerase pause release* process must take place before a second polymerase can be recruited to the promoter. While the promoter is in the permissive-active state, the polymerase can be released from the paused state, leading to the production of an RNA molecule; when the polymerase is released, it unbinds the promoter and the promoter switches back to its permissive-inactive state. After an RNA molecule is produced, it can degrade with a certain probability per unit of time.

The findings by Bartman et al. inspired us to perform mathematical analysis of the multi-scale model, after extending it and including protein production and degradation. Similar models have been studied before, but unlike the multi-scale model, they assume that the promoter state does not change after the synthesis of an RNA molecule. The idea is that analytical expressions for distributions of the number of RNA and proteins can be used for future studies instead of simulations, which are usually time-consuming. We followed the mathematical steps as described in our example in Subsection 1.5.6 and obtained the desired distributions for both RNA and protein numbers. One of the main results from our mathematical analysis shows that if polymerase binding or unbinding processes are much faster than the rest of the processes in the model, then the multi-scale model can be reduced to the simple telegraph model.

2.1 Introduction

As it has been already mentioned, the telegraph model of gene expression is the most adopted in the literature; however, it lacks many important biological details. Recent studies have extended the telegraph model in various directions (see [64] for a recent review). Mammalian cells have been shown to display complex promoter dynamics during the switch from transcriptionally inactive to active states. Such dynamics cannot be described by a single reaction step whose time is exponentially distributed [53], as assumed by the telegraph model. In [112] this complexity is accounted for by deriving analytical expressions linking the Fano factor of mRNA distributions to the general waiting-time distribution of the time to switch from inactive to active states. In contrast, other works [71, 113–115] have sought to describe promoter dynamics with transitions between a number of discrete promoter states, only some of which are active; in special cases of such models, the steady-state distribution of mRNA fluctuations can be derived analytically. Moreover, dynamic regulation of eve stripe 2 expression in living Drosophila [66] suggests the occurrence of multiple rates of Pol II loading, which argues in favour of the multistate model rather than the simpler telegraph model. Another study, based on live-cell imaging of the amoeba Dictyostelium, postulates a continuum of transcriptional states [116] rather than discrete states. All these models share a common property with the telegraph model, namely, that when a transcript is produced, the gene state is unchanged.

Bartman et al. [65] recently argued that it is unclear how polymerase recruitment and pause release, two well-known steps in mRNA production, map onto the active and inactive states assumed by the telegraph model. This argument also applies to the various multistate variants of the telegraph model. In particular, in these models, one cannot tell whether the initiation of a burst permits polymerase recruitment to occur or whether it permits release from the paused state. In [65], the telegraph model and several possible models of transcription were considered that incorporated bursting (burst initiation and termination steps) together with polymerase recruitment and pause release steps. Using stochastic simulations in conjunction with RNA FISH and Pol II ChIP-seq measurements, they showed that the only model compatible with the data is one in which (i) polymerase recruitment follows after burst initiation and (ii) only one polymerase is permitted to bind each promoter-proximal region at a time, and this bound polymerase has to undergo pause release before a second polymerase can be recruited to a gene copy (in line with the findings in [117, 118]). While this model has three effective gene states, it is not a special case of the multistate gene models studied in [113–115]. These models assume that the gene state does not change upon production of mRNA because they model the production of a mature transcript

without detailed modelling of the steps between transcriptional initiation and termination. However, the model expounded in [65] models transcription at a finer level of detail, which requires that the production of nascent mRNA results in a change of gene state, a property that is crucial to capture property (ii) above. Note the number of nascent mRNA molecules, irrespective of their length, is equal to the number of polymerases currently transcribing the gene [119]. An interesting recent review discussing the assumptions behind common gene expression models, including those with polymerase dynamics, can be found in [120].

Protein dynamics are as important as mRNA dynamics for understanding the mechanisms of gene expression. For many organisms, data from experiments performed on single-cells, has shown that gene expression can be well described by a three-stage model [31, 45, 59, 60, 121]. This model is basically an extension of the telegraph model to incorporate protein synthesis; i.e., the system of reaction of this model consists of gene switching between active and inactive state, mRNA production when the gene is active, mRNA degradation, protein synthesis from mRNA and protein degradation. All these reactions are modelled as first-order chemical reactions and the steady-state solution for protein distribution has been proposed in [43], which is valid when the protein lifetime is much greater than the mRNA lifetime. It has been found that typically proteins exist for at least several mRNA lifetimes, and this assumption has been used in various studies [22–24, 43, 122, 123].

In this chapter, we study a detailed analysis of the "multi-scale model"; this is the model proposed by Bartman et al. [65] (model consistent with the experimental data for mRNA dynamics) with an extension to incorporate protein dynamics, please see Fig. 2.1. The conventional three-stage model does not include an independently regulated pause release step and hence cannot differentiate the effects of changing polymerase pause release versus polymerase recruitment rates, whereas the multi-scale model studied here can distinguish these effects. However, based on the assumptions of time-scale separation of various rates of the model and by using perturbation techniques, we show that the multi-scale model can be reduced to the three-stage model.

The chapter is organized as follows. A detailed description of the model of interest is presented in Section 2.2. Detailed derivations of the exact closed-form expression for the distribution of mRNA molecules are presented in Section 2.3, while a simple closed-form expression for the approximate time-dependent mRNA distribution is derived in Section 2.4. Additionally, an approximate steady-state distribution of protein molecule number, under the assumption of short-lived mRNA, is obtained in Section 2.5. Finally, the chapter concludes with a summary and discussion in Section 2.6.

2.2 Model Setup

We consider a stochastic multi-scale transcriptional bursting model, recently introduced in [65] and henceforth referred to as the multi-scale model, as is shown in Fig. 2.1. The system of chemical reactions describing this model is given by

$$\begin{aligned}
G_0 \underset{s_b}{\overset{s_u}{\rightleftharpoons}} G_{10} &\qquad \text{promoter switching - TF binding/unbinding,}\\
G_{10} \xrightarrow{s_p} G_{11} &\qquad \text{promoter switching - Pol II binding,}\\
G_{11} \xrightarrow{s_b} G_0 &\qquad \text{promoter switching - TF and Pol II unbinding,}\\
G_{11} \xrightarrow{r} G_{10} + M &\qquad \text{mRNA transcription,}\\
M \xrightarrow{v} M + P &\qquad \text{protein translation,}\\
M \xrightarrow{d_m} \varnothing &\qquad \text{mRNA degradation,}\\
P \xrightarrow{d_p} \varnothing &\qquad \text{protein degradation,}
\end{aligned} \tag{2.1}$$

where $\varnothing$ denotes sinks of molecules. In this model, the promoter of the gene fluctuates between three states depending on if transcription factor (TF) and polymerase II (Pol II) are bound or not to the gene. Hence, we have: two permissive states (G_{10} and G_{11}; TF is bound to the gene) and

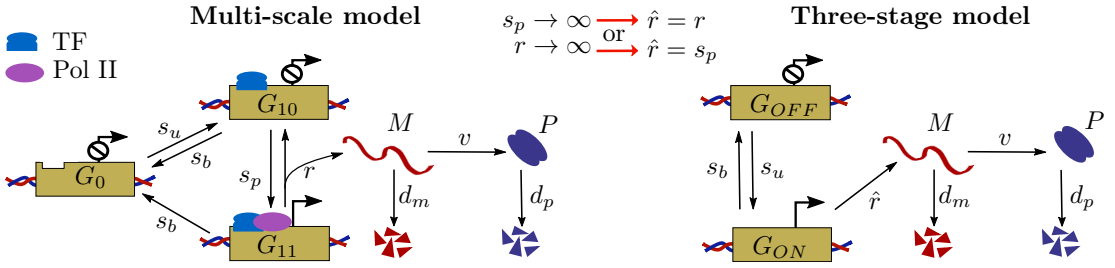


Figure 2.1: *Schematic of the stochastic multi-scale transcriptional bursting model (left) and three-stage model (right).* Please see the main text in Section 2.2 for a detailed description of the models. Mathematical analysis in this chapter shows that the three-stage model, with transcription rate, $\hat{r}$ is a special case of the multi-scale model in some biological limits. Specifically, for large parameter, s_p (polymerase binding rate) or r (polymerase pause release rate), the multi-scale model simplifies to the three-stage model with transcription rate being $\hat{r} = r$ or $\hat{r} = s_p$, respectively.

a non-permissive state (G_0; TF is not bound to the gene). The transition from G_0 to G_{10} (burst initiation) is mediated TF binding with rate constant, s_u which is reversible with rate constant s_b (this transition may alternatively represent other processes such as nucleosome remodelling). Subsequently, the binding of Pol II to G_{10} with rate constant s_p (which is proportional to Pol II abundance) leads to G_{11}. This represents a state in which Pol II is paused and models the experimental observation that Pol II pauses downstream of the transcription initiation site preceding productive elongation [118]. The polymerase is released from this state with rate constant, r leading to the production of an mRNA molecule (denoted as M) and the unbinding of polymerase, which returns the promoter to state G_{10}. In the paused state G_{11}, both the polymerase and the transcription factor can unbind from the gene and lead to the non-permissive state G_0 (burst termination). Both reversible switches operate at different timescales (hours versus minutes) with $\max\{s_b, s_u\} \ll \min\{r, s_p\}$, leading to multi-scale transcriptional bursting [51, 65]. The produced mRNA is further translated into protein (defined as P) with translation rate v and decays with rate d_m, while the protein molecules have degradation rate d_p. We assume all reactions to be first-order, characterized by exponentially distributed waiting times between successive reactions.

Here we note that, the reaction $G_{11} \to G_{10} + M$ can be an effective description of the reactions: $G_{11} \to G_{10} + N$, $N \to M$ where N is nascent mRNA (pre-spliced mRNA). The reaction $N \to M$ is often modelled with a deterministic time delay [72] but theory shows that it can be modelled with a stochastic exponential time delay provided the timescale of nascent mRNA production $(1/r)$ is much smaller than the timescale governing the transitions between permissive and non-permissive states, $(1/(s_u + s_b))$ [72]. Also, no explicit nascent mRNA description is needed, provided that it is short-lived compared to mature mRNA. Since these two conditions are physiologically realistic in many cases, we choose to ignore detailed modelling of nascent mRNA dynamics and model the direct production of mature mRNA with an exponentially distributed time delay.

Also, we note that, although this multi-scale model has three effective gene states (one of which regulates polymerase pause release), it is not a special case of existing multi-state models because in this model, the gene state changes upon production of new mRNAs. This is based on the experimental observation that unless the polymerase is unpaused (and nascent mRNA starts being actively transcribed by this polymerase), there can be no binding of new Pol II. In contrast, current models assume the gene state does not change upon the production of mRNA because they model the production of a mature transcript without detailed modelling of the steps between transcriptional initiation and termination.

In the next sections, we show how to derive closed-form solutions for mRNA and protein distributions. Additionally, we include derivations showing that in the limit of large parameters r or s_p, the multi-scale model can be well described by the simple three-stage model.

2.3 Exact solution for the steady-state probability distribution of mRNA numbers

For simplicity in our derivations, we rescale all the kinetic parameters with mRNA decay rate and obtain the following non-dimentional parameters: $\sigma_u = s_u/d_m$, $\sigma_b = s_b/d_m$, $\sigma_p = s_p/d_m$, $\rho = r/d_m$ and the time variable as $\tau = t \cdot d_m$. We define functions $P_j(n;\tau)$ ($j \in \{0, 10, 11\}$) as the probability of the promoter being in state G_j with n representing the number of mRNA molecules in the system at time τ, while $P(n;\tau) = \sum_j P(n,\tau)_j$ is the total probability function. The time-evolution of the probabilities P_j is described by a set of non-dimensional coupled master equations as:

$$\partial_\tau P_0 = (\mathbb{E}_n^1 - 1)nP_0 - \sigma_u P_0 + \sigma_b P_{10} + \sigma_b P_{11}, \tag{2.2a}$$

$$\partial_\tau P_{10} = (\mathbb{E}_n^1 - 1)nP_{10} + \sigma_u P_0 - (\sigma_b + \sigma_p)P_{10} + \rho\mathbb{E}_n^{-1}P_{11}, \tag{2.2b}$$

$$\partial_\tau P_{11} = (\mathbb{E}_n^1 - 1)nP_{11} - (\sigma_b + \rho)P_{11} + \sigma_p P_{10}, \tag{2.2c}$$

where $\mathbb{E}_n^c[f(n)] = f(n+c)$, with $c \in \mathbb{Z}$, denotes the standard step operator [83] acting on function f. In order to solve the above equations, we use the method of generating functions; we define the functions $F_j(z;\tau) = \sum_{n=0}^{\infty} P_j(n;\tau)z^n$ for every promoter state, G_j ($j \in \{0, 10, 11\}$), while the total generating function is given by $F = \sum_j F_j$. Now, we rewrite the system in Eq. (2.2) as a set of coupled partial differential equations (PDEs):

$$\partial_\tau F_0 + (z-1)\partial_z F_0 = -\sigma_u F_0 + \sigma_b F_{10} + \sigma_b F_{11}, \tag{2.3a}$$

$$\partial_\tau F_{10} + (z-1)\partial_z F_{10} = \sigma_u F_0 - (\sigma_b + \sigma_p)F_{10} + \rho z F_{11}, \tag{2.3b}$$

$$\partial_\tau F_{11} + (z-1)\partial_z F_{11} = -(\sigma_b + \rho)F_{11} + \sigma_p F_{10}. \tag{2.3c}$$

We introduce the new variable, $u = z - 1$ and for the steady-state solution we set the terms $\partial_\tau F_j$ ($j = 0, 10, 11$) in the above equations equal to zero. We rewrite the system in Eq. (2.3) as:

$$u(\partial_u F_0) = -\sigma_u F_0 + \sigma_b F_{10} + \sigma_b F_{11}, \tag{2.4a}$$

$$u(\partial_u F_{10}) = \sigma_u F_0 - (\sigma_b + \sigma_p)F_{10} + \rho(u+1)F_{11}, \tag{2.4b}$$

$$u(\partial_u F_{11}) = -(\sigma_b + \rho)F_{11} + \sigma_p F_{10}. \tag{2.4c}$$

Solving Eq. (2.4c) for F_{10} as function of F_{11}, substituting the result in Eq. (2.4b) and solving F_0 as function of F_{11} we get Eq. (2.4a) being a third order ordinary differential equation (ODE) for $F_{11}(u)$:

$$u^2\partial_u^3 F_{11} + (1 + b_1 + b_2)u\partial_u^2 F_{11} + (b_1 b_2 - \sigma_p\rho u)\partial_u F_{11} - \sigma_p\rho a F_{11} = 0, \tag{2.5}$$

where $a = 1 + \sigma_u$, $b_1 = 1 + \sigma_b + \sigma_p + \rho$ and $b_2 = 1 + \sigma_b + \sigma_u$. Now we define a new variable as $x = \sigma_p\rho u$ and hence Eq. (2.5) converts to a new ODE,

$$x^2\partial x^3 F_{11} + (1 + b_1 + b_2)x\partial_x^2 F_{11} + (b_1 b_2 - x)\partial_x F_{11} - a F_{11} = 0. \tag{2.6}$$

Eq. (2.6) is a canonical form of differential equation for a generalized hypergeometric function, which admits a solution of form $F_{11}(x) = C \cdot {}_1F_2(a; b_1, b_2; x)$; here, ${}_1F_2$ is a generalized hypergeometric function [44, 63] and C is a constant of integration. Summing the equations in Eq. (2.4) and using the definition for the full generating function, $F(u)$ we have that $\partial_u F = \rho F_{11}$, which leads to an exact solution for F, $F(u) = \tilde{C} \cdot {}_1F_2(a-1; b_1-1, b_2-1; \sigma_p\rho u)$. $\tilde{C}$ is the new constant of integration and one can easily show that $\tilde{C} = 1$ by using the normalization condition, $F|_{(u=0)} = \sum_{n=0}^{\infty} P(n) = 1$. The final expression for the generating function of the steady-state mRNA distribution is then given by

$$F(u) = {}_1F_2(\sigma_u; \sigma_p + \rho + \sigma_b, \sigma_b + \sigma_u; \sigma_p\rho u), \tag{2.7}$$

The probability function is then given by the general formula,

$$P(n;t) = \frac{1}{n!}\frac{d^n}{du^n}F(u;t)\Big|_{(u=-1)}, \tag{2.8}$$

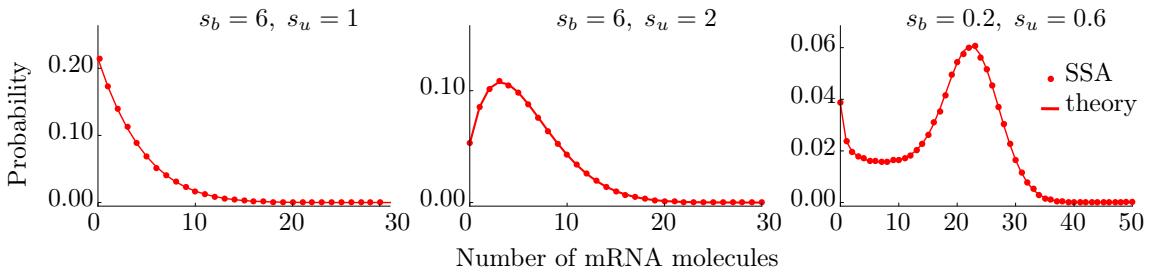


Figure 2.2: *Steady-state mRNA distribution from the multi-scale model.* Stochastic simulations (SSA; dots) verify our analytical exact closed-form solution for mRNA distribution (Eq. (2.9); lines). The parameter values that have been used are: $s_p = 40$ min^{-1}, $r = 60$ min^{-1}, $d_m = 1$ min^{-1}, while s_b and s_u are specified in the plots.

and it follows that the analytical expression for the steady-state marginal probability of finding n mRNAs in a cell is given by:

$$P(n) = \frac{(\sigma_p\rho)^n}{n!}\frac{\Gamma(\sigma_u+n)}{\Gamma(\sigma_u)}\frac{\Gamma(\sigma_b+\rho+\sigma_p)}{\Gamma(\sigma_b+\rho_u+\sigma_p+n)}\frac{\Gamma(\sigma_b+\sigma_u)}{\Gamma(\sigma_b+\sigma_u+n)}\times \\ \times {}_1F_2(\sigma_u+n;\sigma_p+\rho+\sigma_b+n,\sigma_b+\sigma_u+n;-\sigma_p\rho). \tag{2.9}$$

In Fig. 2.2 we verify our analytical solution given in Eq. (2.9) by comparing it with the numerical solution obtained from SSA; we plot the distribution for three different sets of values for parameters, s_b and s_u (parameters related to the promoter switching) to show unimodal and bimodal cases.

2.4 Approximate solution for the time-dependent probability of mRNA numbers in case of large parameter r or s_p

One can easily see that the solution in Eq. (2.9) is symmetric for parameters r and s_p. Here, we claim that for large parameter r (or s_p), the stochastic multi-scale model without protein dynamics (i.e. $v = 0$), reduces to the simple telegraph model (please see Fig. 2.1) for which the time-dependent solution of mRNA distribution is known. We prove our claim by using perturbation techniques in our mathematical analysis. The exact steady-state solution of the generating function of mRNA distribution from the telegraph model has been previously reported in [31,43,61] and its analytical expression is given by:

$$F(u)_{tel} = {}_1F_1(\sigma_u;\sigma_u+\sigma_b;\hat{\rho}u), \tag{2.10}$$

where ${}_1F_1$ is the confluent hypergeometric function of the first kind [44,63] and $\hat{\rho} = \hat{r}/d_m$ is the rescaled mRNA transcription rate in the telegraph model ($\hat{r}$ is the transcription rate). The closed-form time-dependent solution of the generating function of mRNA distribution from the telegraph model has been previously reported in [61,62] and its analytical expression is given by:

$$F(u;\tau)_{tel} = f_s(u;\tau){}_1F_1(\sigma_u;\sigma_u+\sigma_b;\hat{\rho}u) + f_{ns}(u;\tau){}_1F_1(1-\sigma_b;2-\sigma_u-\sigma_b;\hat{\rho}u) \tag{2.11}$$

where

$$f_s(u;\tau) = {}_1F_1(-\sigma_u;1-\sigma_u-\sigma_b;-\hat{\rho}e^{-\tau}u)$$
$$f_{ns}(u;\tau) = \frac{\sigma_u}{(\sigma_u+\sigma_b)(1-\sigma_u-\sigma_b)}\hat{\rho}ue^{-(\sigma_u+\sigma_b)\tau}{}_1F_1(\sigma_b;1+\sigma_u+\sigma_b;-\hat{\rho}e^{-\tau}u).$$

Note that, in steady-state (i.e. for $t\to\infty$) we have the limits $f_s(u;\tau)\to 1$ and $f_{ns}(u;\tau)\to 0$ and the above expression simplifies to the one in Eq. (2.10). This solution was obtained by assuming

that at time zero, the promoter is in its inactive state and there are zero mRNA molecules in the system.

The idea of multi-scale model reduction to the telegraph model (i.e. three-stage model without translation) is the following. The expression for the mean number of mRNA molecules can be found by taking into account the solution of the probability generating function and applying the formula, $\langle n\rangle = \partial_u F(u)|_{(u=0)}$ where $\langle\cdot\rangle$ denotes the mean. By considering Eq. (2.7) and Eq. (2.10), we obtain the following expressions for the mean value of mRNA numbers from the multi-scale and the telegraph models, respectively:

$$\langle n\rangle = \frac{\sigma_u\sigma_p\rho}{(\sigma_p+\rho+\sigma_b)(\sigma_b+\sigma_u)} \quad \text{and} \quad \langle n\rangle_{tel} = \frac{\sigma_u\hat{\rho}}{\sigma_b+\sigma_u}. \tag{2.12}$$

By equating the above expressions and solving for $\hat{r}$, we get that $\hat{r} = s_p r/(s_p+r+s_b)$. This indicates that in the limit $r\to\infty$ or $s_p\to\infty$, the multi-scale model can be reduced to the telegraph model with transcription rate being $\hat{r}=s_p$ or $\hat{r}=r$, respectively. Thereafter, this means that the time-dependent distribution of mRNA numbers from the multi-scale bursting model can be well approximated by the time-dependent solution of the telegraph model given in Eq. (2.11). In the rest of this section, we show our detailed analysis for the reduction of the multi-scale model to the telegraph model in case of large parameters r and s_p. The reduction of the multi-scale model into an effective telegraph model, without making the aforementioned assumptions for model parameters, can be found in our published manuscript [122].

2.4.1 Large parameter r

For large parameter r we have that the stochastic multi-scale model reduces to the telegraph model with transcription rate, $\hat{r}=s_p$; i.e. for large r, the solution in Eq. (2.7) simplifies to the one in Eq. (2.10) with $\hat{r}=s_p$ ($\hat{\rho}=\sigma_p$). We are going to show this by using perturbation techniques; i.e. for large parameter r, we use the parametrization $r\mapsto r/\varepsilon$ ($\rho\mapsto\rho/\varepsilon$), where ε is a small perturbation parameter. Then the system in Eq. (2.4) transforms into the slow system:

$$u(\partial_u F_0)+\sigma_u F_0-\sigma_b F_{10}-\sigma_b F_{11}=0, \tag{2.13a}$$

$$\varepsilon[u(\partial_u F_{10})-\sigma_u F_0+(\sigma_b+\sigma_p)F_{10}]=\rho(u+1)F_{11}, \tag{2.13b}$$

$$\varepsilon[u(\partial_u F_{11})+\sigma_b F_{11}-\sigma_p F_{10}]=-\rho F_{11}. \tag{2.13c}$$

Now, we use an asymptotic expansion of the generating functions over ε as:

$$F_j = F_j^{(0)}+\varepsilon F_j^{(1)}+\mathcal{O}(\varepsilon^2), \tag{2.14}$$

and we substitute these expressions into Eq. (2.13). By collecting the leading-order terms for ε^0, we get that $F_{11}^{(0)}=0$; this indicates that for large r the promoter state, G_{11} in the model can be considered negligible. Using this result and collecting the first-order terms for ε^1, we obtain the following system of ODEs:

$$u(\partial_u F_0^{(0)}) = -\sigma_u F_0^{(0)}+\sigma_b F_{10}^{(0)}, \tag{2.15a}$$

$$u(\partial_u F_{10}^{(0)}) = \sigma_u F_0^{(0)}-(\sigma_b+\sigma_p)F_{10}^{(0)}+\rho(u+1)F_{11}^{(1)}, \tag{2.15b}$$

$$0 = \sigma_p F_{10}^{(0)}-\rho F_{11}^{(1)}. \tag{2.15c}$$

Eq. (2.15c) gives us $\sigma_p F_{10}^{(0)}=\rho F_{11}^{(1)}$, while summing up the equations in (2.15) we have that $\partial_u F^{(0)}=\rho F_{11}^{(1)}$. Using these relations and the fact that $F_0^{(0)}=F^{(0)}-F_{10}^{(0)}$, from Eq. (2.15a) we get the Kummer's differential equation,

$$\partial_{u^2}F^{(0)}u+(\sigma_u+\sigma_b-\sigma_p u)\partial_u F^{(0)}-\sigma_p\sigma_u F^{(0)}=0, \tag{2.16}$$

which admits as solution a confluent hypergeometric function of the first kind,

$$F^{(0)}(u;\tau) = C \cdot {}_1F_1(\sigma_u, \sigma_u + \sigma_b, \sigma_p u), \tag{2.17}$$

where C some constant of integration, and one can easily find that $C = 1$ by applying the normalization condition, $F|_{(u=0)} = 1$. The expression in Eq. (2.17) is the same as the one in given by Eq. (2.10) for $\hat{r} = s_p$; hence, it is the solution of the telegraph model with transcription rate $\hat{r} = s_p$ and this proves our claim. This indicates that the time-dependent solution of the multi-scale model for large parameter r can be well approximated by Eq. (2.11) with $\hat{r} = s_p$. We present the verification of these results in Fig. 2.3; In plot (A) we show the time-evolution of the mRNA probability, while in the plot (B) we show that for increasing parameter r, the exact mRNA distribution of the multi-scale model approaches the mRNA distribution of the telegraph model with $\hat{r} = s_p$.

The explanation of the model reduction in the case of large parameter r is the following. When the polymerase pause release rate, r is very fast it means that the Pol II is almost never bounded to gene (state of the gene with bounded Pol II is $G_{11} \equiv 0$), and the model reduces to the telegraph model with active promoter state being G_{10} and inactive promoter state being G_0. The production of an mRNA molecule involves the slow reaction step from G_{10} to G_{11} with rate, s_p followed by a very fast reverse step with rate r. Hence, the rate of mRNA production in the reduced model is determined by the reaction rate of the slowest reaction, i.e., it is equal to s_p.

2.4.2 Large parameter s_p

By similar reasoning as before, we can deduce that in case of large parameter s_p, the solution in Eq. (2.7) simplifies to the one in Eq. (2.10) with $\hat{r} = r$ ($\hat{\rho} = \rho$). In the limit of large s_p, we use the parametrization $s_p \mapsto s_p/\varepsilon$ ($\sigma_p \mapsto \sigma_p/\varepsilon$); hence, from Eq. (2.4) we obtain the slow system:

$$u(\partial_u F_0) + \sigma_u F_0 - \sigma_b F_{10} - \sigma_b F_{11} = 0, \tag{2.18a}$$

$$\varepsilon[u(\partial_u F_{10}) - \sigma_u F_0 + \sigma_b F_{10} - \rho(u+1)F_{11}] = -\sigma_p F_{10}, \tag{2.18b}$$

$$\varepsilon[u(\partial_u F_{11}) + \sigma_b F_{11} + \rho F_{11}] = \sigma_p F_{10}. \tag{2.18c}$$

Using again asymptotic expansion of the generating functions and collecting the terms for ε^0, we have that $F_{10}^{(0)} = 0$; this indicates that for large s_p the promoter state, G_{10} in the model can be considered negligible. Taking into account this equation and collecting the terms of order ε^1, we obtain the system:

$$u(\partial_u F_0^{(0)}) = -\sigma_u F_0^{(0)} + \sigma_b F_{11}^{(0)}, \tag{2.19a}$$

$$0 = \sigma_u F_0^{(0)} + \rho(u+1)F_{11}^{(0)} - \sigma_p F_{10}^{(1)}, \tag{2.19b}$$

$$u(\partial_u F_{11}^{(0)}) = -(\sigma_b + \rho)F_{11}^{(0)} + \sigma_p F_{10}^{(1)}. \tag{2.19c}$$

Summing up the equations in (2.19) we get that $\partial_u F = \rho F_{11}^{(0)}$, which together with $F_0^{(0)} = F^{(0)} - F_{11}^{(0)}$ can be substituted into equations (2.19b), (2.19c) and give the Kummer's differential equation for $F^{(0)}$,

$$\partial_{u^2} F^{(0)} u + (\sigma_u + \sigma_b - \rho u)\partial_u F^{(0)} - \rho\sigma_u F^{(0)} = 0, \tag{2.20}$$

which after applying the normalization condition we can see that admits as a solution the confluent hypergeometric function,

$$F^{(0)}(u;\tau) = {}_1F_1(\sigma_u, \sigma_u + \sigma_b, \rho u), \tag{2.21}$$

Eq. (2.21) is also the steady-state generating function for mRNA distribution from the telegraph model with transcription rate $\hat{r} = r$; this means that for large parameter s_p, the time-dependent generating function of mRNA distribution from the multi-scale model can be well approximated by Eq. (2.11) with $\hat{r} = r$. Since our system is symmetric for parameters r and s_p, it means that the plots in Fig. 2.3 will be exactly the same for $r \mapsto s_p$ and $\hat{r} = r$.

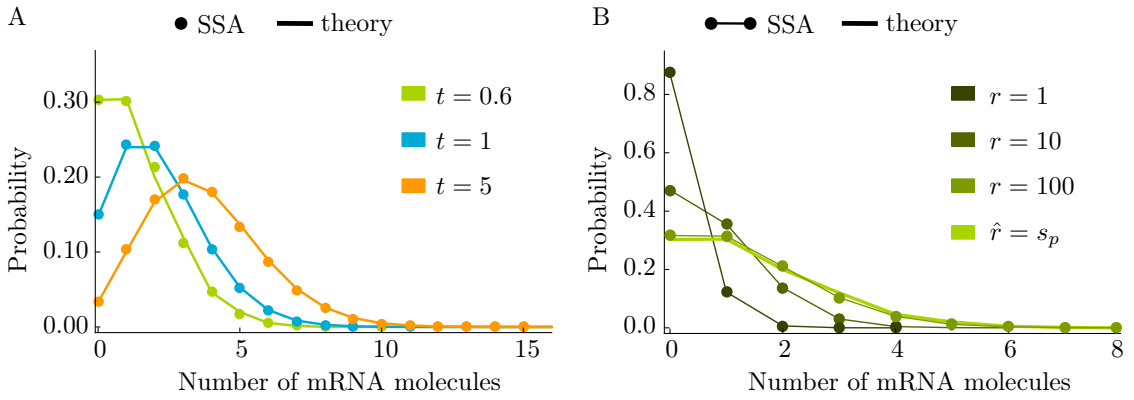


Figure 2.3: *Time-dependent mRNA distribution from the multi-scale model and accuracy of its approximation.* A) Time-dependent mRNA distribution for three different time values as it has been predicted by stochastic simulations (SSA; dots) and the theory (Eq. (2.8) together with Eq. (2.11) and $\hat{r} = s_p$; solid lines). B) Time-dependent mRNA distribution for time, $t = 0.6$ min and for three different values of parameter, r; the plotted results are as predicted by stochastic simulations (SSA; dots with thin lines) and approximate theory (Eq. (2.8) together with Eq. (2.11) and $\hat{r} = s_p$; solid line). Note that the approximate solution is obtained for the limit of large, r hence, it is independent of this parameter, while the time-dependent solution obtained from SSA approaches the approximate solution for increasing values of the parameter, r as the theory predicts. A-B) The parameter values that have been used are: $s_p = 5\ \text{min}^{-1}$, $r = 600\ \text{min}^{-1}$, $d_m = 1\ \text{min}^{-1}$, $s_b = 2\ \text{min}^{-1}$ and $s_u = 6\ \text{min}^{-1}$.

The explanation of the model reduction in the case of large parameter s_p is the following. When the polymerase recruitment rate, s_p is very fast it means that the Pol II is almost always bound to the gene (state of the gene without Pol II is $G_{10} \equiv 0$), and the model reduces to the telegraph model with active promoter state being G_{11} and inactive promoter state being G_0. The synthesis of an mRNA molecule happens due to a slow reaction rate, r; hence, the mRNA transcription rate in the reduced model is determined by $\hat{r} = r$.

2.5 Analytical solution for the approximate steady-state probability distribution of protein numbers

In this section, we are going to perform detailed derivation in order to obtain a closed-form expression for the protein's number distribution, under the assumption that the mRNA is short-lived compared to the protein ($d_m \gg d_p$). For simplicity in our analysis, we rescale all kinetic parameters by the protein decay rate, d_p as: $\tilde{\sigma}_b = s_b/d_p$, $\tilde{\sigma}_u = s_u/d_p$, $\tilde{\rho} = r/d_p$, $\tilde{\lambda} = v/d_p$, $\tilde{\sigma}_p = s_p/d_p$ and $\tilde{\delta}_m = d_m/d_p$, while the time variable is rescaled as $\tilde{\tau} = t \cdot d_p$. We define $P_j(n, m; \tilde{\tau})$ as the probability of the promoter being at state, G_j ($j = 0, 10, 11$) with n and m being the numbers of mRNA molecules and protein molecules, respectively, in the cell at time $\tilde{\tau}$. The master equations for the time-evolution of these probabilities are given by:

$$\begin{aligned}
\partial_\tau P_0 &= \tilde{\lambda}(\mathbb{E}_m^{-1} - 1)nP_0 + \tilde{\delta}_m(\mathbb{E}_n - 1)nP_0 + (\mathbb{E}_m - 1)mP_0 - \tilde{\sigma}_u P_0 + \tilde{\sigma}_b P_{10} + \tilde{\sigma}_b P_{11}, \\
\partial_\tau P_{11} &= \tilde{\lambda}(\mathbb{E}_m^{-1} - 1)nP_{11} + \tilde{\delta}_m(\mathbb{E}_n - 1)nP_{11} + (\mathbb{E}_m - 1)mP_{11} + \tilde{\sigma}_p P_{10} - (\tilde{\sigma}_b + \tilde{\rho})P_{11}, \\
\partial_\tau P_{10} &= \tilde{\lambda}(\mathbb{E}_m^{-1} - 1)nP_{10} + \tilde{\delta}_m(\mathbb{E}_n - 1)nP_{10} + (\mathbb{E}_m - 1)mP_{10} + \tilde{\sigma}_u P_0 - (\tilde{\sigma}_p + \tilde{\sigma}_b)P_{10} \\
&\quad + \tilde{\rho}\mathbb{E}_n^{-1} P_{11}.
\end{aligned} \tag{2.22}$$

Now, we define the full generating function as $F(u,w;\tilde{\tau}) = \sum_j F_j(u,w;\tilde{\tau})$, where

$$F_j(u,w;\tilde{\tau}) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} P(n,m;\tilde{\tau})(1+u)^n(1+w)^m \quad \text{for} \quad j = 0, 10, 11. \tag{2.23}$$

We apply the method of generating functions on the system of CMEs in Eq. (2.22) and we obtain the following system of PDEs:

$$\begin{aligned}
\mathbb{L}F_0 &= -\tilde{\sigma}_u F_0 + \tilde{\sigma}_b F_{10} + \tilde{\sigma}_b F_{11},\\
\mathbb{L}F_{10} &= \tilde{\sigma}_u F_0 - (\tilde{\sigma}_p + \tilde{\sigma}_b)F_{10} + \tilde{\rho}(u+1)F_{11},\\
\mathbb{L}F_{11} &= \tilde{\sigma}_p F_{10} - (\tilde{\sigma}_b + \tilde{\rho})F_{11},
\end{aligned} \tag{2.24}$$

where $\mathbb{L}$ is a differential operator defined as $\mathbb{L} = \partial_{\tilde{\tau}} + (\tilde{\delta}_m u - \tilde{\lambda} w(u+1))\partial_u + w\partial_w$. By using the method of characteristics, we get that $\partial_s \tilde{\tau} = 1$, which implies that $\tilde{\tau} \equiv s$ can be used as an independent variable; the rest of the characteristic equations are given by:

$$\partial_{\tilde{\tau}} u = \tilde{\delta}_m[u - \lambda w(u+1)], \tag{2.25a}$$

$$\partial_{\tilde{\tau}} w = w, \tag{2.25b}$$

$$\partial_{\tilde{\tau}} F_0 = -\tilde{\sigma}_u F_0 + \tilde{\sigma}_b F_{10} + \tilde{\sigma}_b F_{11}, \tag{2.25c}$$

$$\partial_{\tilde{\tau}} F_{10} = \tilde{\sigma}_u F_0 - (\tilde{\sigma}_p + \tilde{\sigma}_b)F_{10} + \tilde{\rho}(u+1)F_{11}, \tag{2.25d}$$

$$\partial_{\tilde{\tau}} F_{11} = \tilde{\sigma}_p F_{10} - (\tilde{\sigma}_b + \tilde{\rho})F_{11}. \tag{2.25e}$$

where $\lambda = v/d_m$ is the mean translational burst size. Under the assumption that the mRNA decays much faster than the protein ($d_m \gg d_p$), we can make the substitution: $\tilde{\delta}_m \mapsto \tilde{\delta}_m/\varepsilon$, where $\varepsilon \ll 1$ is a small perturbation parameter. Then, the Eq. (2.25a) can be rewritten as $\varepsilon \cdot \partial_{\tilde{\tau}} u = \tilde{\delta}_m[u - \lambda w(u+1)]$, which means that for $\varepsilon \to 0$ (mRNA degradation rate is much larger than the protein degradation rate) we have that $u = \lambda w/(1-\lambda w)$. Also, from Eq. (2.25b) we obtain the chain rule, $\partial_{\tilde{\tau}} \equiv w\partial_w$. By using these two obtained results, we transform equations (2.25c)-(2.25e) into the following system of ODEs for the generating functions:

$$w\partial_w F_0 = -\tilde{\sigma}_u F_0 + \tilde{\sigma}_b F_{10} + \tilde{\sigma}_b F_{11}, \tag{2.26a}$$

$$w\partial_w F_{10} = \tilde{\sigma}_u F_0 - (\tilde{\sigma}_p + \tilde{\sigma}_b)F_{10} + \tilde{\rho}\frac{1}{1-\lambda w}F_{11}, \tag{2.26b}$$

$$w\partial_w F_{11} = \tilde{\sigma}_p F_{10} - (\tilde{\sigma}_b + \tilde{\rho})F_{11}. \tag{2.26c}$$

By summing up the equations in (2.26), we get that $\partial_w F = \tilde{\rho}\lambda F_{11}/(1-\lambda w)$. By using this result, the equations (2.26c)-(2.26b) and the fact that $F = F_0 + F_{10} + F_{11}$, it follows that the function $F(w)$ satisfies the following third-order ODE:

$$\begin{aligned}
&(1-\lambda w)w^2\partial_w^3 F + [1 + b_1 + b_2 - \lambda w(3 + b_1 + b_2)]w\partial_w^2 F +\\
&[b_1 b_2 - \lambda w(1 + b_1 + b_2 + b_1 b_2 + \tilde{\rho}\tilde{\sigma}_p)]\partial_w F - \lambda\tilde{\rho}\tilde{\sigma}_p\tilde{\sigma}_u F = 0,
\end{aligned} \tag{2.27}$$

which admits the solution

$$F(w) = \tilde{C}\cdot{}_3F_2(a_1, a_2, a_3; b_1, b_2; \lambda w), \tag{2.28}$$

where $b_1 = \tilde{\sigma}_b + \tilde{\sigma}_u$, $b_2 = \tilde{\sigma}_b + \tilde{\sigma}_p + \tilde{\rho}$ and the constants a_1, a_2,a_3 are roots of the equations:

$$\begin{aligned}
a_1 a_2 a_3 &= \tilde{\rho}\tilde{\sigma}_p\tilde{\sigma}_u,\\
a_1 + a_2 + a_3 &= b_1 + b_2,\\
a_1 a_2 + a_1 a_3 + a_2 a_3 &= b_1 b_2 + \tilde{\rho}\tilde{\sigma}_p.
\end{aligned} \tag{2.29}$$

In Eq. (2.28), the constant, $\tilde{C}$ is constant of integration, which one can easily show that $\tilde{C} = 1$ by applying the normalization condition, $F|_{(w=0)} = 1$. In order to obtain the expression for the

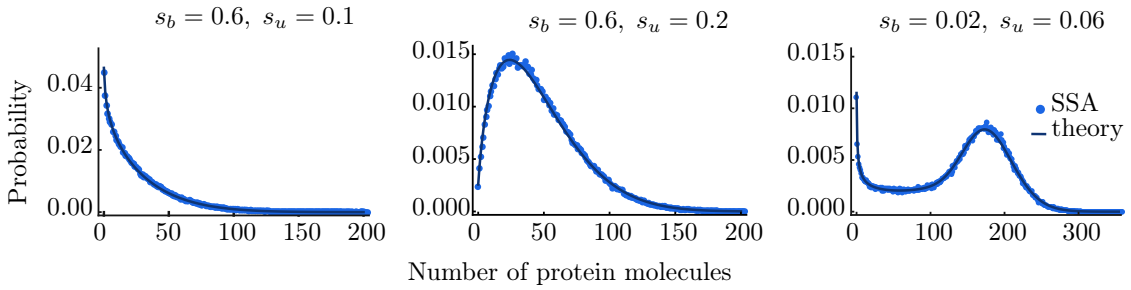


Figure 2.4: *steady-state protein distribution from the multi-scale model.* Stochastic simulations (SSA; dots) verify our analytical exact closed-form solution for mRNA distribution (Eq. (2.30); lines). The parameter values that have been used are: $s_p = 4$ min^{-1}, $r = 60$ min^{-1}, $d_m = 1$ min^{-1}, $v = 5$ min^{-1}, $d_p = 0.1$ min^{-1} while s_b and s_u are specified in the plots.

steady-state protein distribution we use a formula similar to the one given by Eq. (2.8) and we get that,

$$P(m) = \frac{\lambda^m}{m!}\frac{(a_1)_m(a_2)_m(a_3)_m}{(b_1)_m(b_2)_m}\,{}_3F_2(a_1+m, a_2+m, a_3+m; b_1+m, b_2+m; -\lambda), \tag{2.30}$$

where $(\cdot)_m$ is the Pochhammer symbol. We verify the obtained solution by performing simulations, and we present our results in Fig. 2.4.

Here, we note that the solution in Eq. (2.28) is symmetric for parameters r and s_p ($\tilde{\rho}$ and $\tilde{\sigma}_p$). In the rest of this section, we show that for large parameter r (or s_p), the solution given in Eq. (2.28) reduces to the Gaussian hypergeometric function (${}_2F_1$), which was reported in [43], for the classical three-stage model of gene expression in the limit of fast mRNA decay.

2.5.1 Large parameter r

In order to simplify the solution in Eq. (2.28) for large parameter r, we use the parametrization $r \mapsto r/\varepsilon$ ($\tilde{\rho} \mapsto \tilde{\rho}/\varepsilon$) and rewrite the system of ODEs from Eq. (2.26) as:

$$w\partial_w F_0 = -\tilde{\sigma}_u F_0 + \tilde{\sigma}_b F_{10} + \tilde{\sigma}_b F_{11}, \tag{2.31a}$$

$$\varepsilon[w\partial_w F_{10} - \tilde{\sigma}_u F_0 + (\tilde{\sigma}_p + \tilde{\sigma}_b)F_{10}] = \tilde{\rho}\frac{1}{1-\lambda w}F_{11}, \tag{2.31b}$$

$$\varepsilon[w\partial_w F_{11} - \tilde{\sigma}_p F_{10} + \tilde{\sigma}_b F_{11}] = -\tilde{\rho}F_{11}. \tag{2.31c}$$

After the asymptotic expansion of the generating functions over the small perturbation parameter, ε in the same way as in Eq. (2.14) and substituting them in the above equation, we collect the leading-order terms (ε^0) and get that $F_{11}^{(0)} = 0$ (indicated negligible G_{11} promoter state in the model for large r). By collecting the first-order terms (ε^1) we obtain the following system;

$$w\partial_w F_0^{(0)} = -\tilde{\sigma}_u F_0^{(0)} + \tilde{\sigma}_b F_{10}^{(0)}, \tag{2.32a}$$

$$w\partial_w F_{10}^{(0)} = \tilde{\sigma}_u F_0^{(0)} - (\tilde{\sigma}_b + \tilde{\sigma}_p)F_{10}^{(0)} + \frac{\tilde{\rho}}{1-\lambda w}F_{11}^{(1)}, \tag{2.32b}$$

$$0 = \tilde{\sigma}_p F_{10}^{(0)} - \tilde{\rho}F_{11}^{(1)}. \tag{2.32c}$$

By summing the equations in (2.32) one can find that $\partial_w F^{(0)} = \tilde{\rho}\lambda F_{11}^{(1)}/(1-\lambda w)$; this result paired with the equation $\tilde{\rho}F_{11}^{(1)} = \tilde{\sigma}_p F_{10}^{(0)}$ from (2.32c) gives us the following second-order ODE for the leading-order full generating function, $F^{(0)}(x)$ where $x = \lambda w$;

$$x(1-x)\partial_x^2 F^{(0)} + [b_1 - (1 + \tilde{\sigma}_p + \tilde{\sigma}_u + \tilde{\sigma}_b)x]\partial_x F^{(0)} - \tilde{\sigma}_p\tilde{\sigma}_u F^{(0)} = 0. \tag{2.33}$$

where $b_1 = \tilde{\sigma}_u + \tilde{\sigma}_b$, as defined earlier. Eq. (2.33) is a hypergeometric differential equation, which admits as a solution the Gaussian hypergeometric function, $F^{(0)}(w) = C \cdot {}_2F_1(a_1, a_2; b_1; \lambda w)$. By applying the normalization condition, $F^{(0)}|_{(w=0)} = 1$ to this function, we obtain that $C = 1$; hence, the leading-order steady-state solution of the full generating function is given by:

$$F^{(0)}(w) = {}_2F_1(\alpha_1, \alpha_2; b_1; \lambda w), \tag{2.34}$$

where

$$\alpha_1 = \frac{1}{2}\Big(\tilde{\sigma}_p + \tilde{\sigma}_u + \tilde{\sigma}_b + \sqrt{(\tilde{\sigma}_p + \tilde{\sigma}_b + \tilde{\sigma}_u)^2 - 4\tilde{\sigma}_p\tilde{\sigma}_u}\Big),$$
$$\alpha_2 = \frac{1}{2}\Big(\tilde{\sigma}_p + \tilde{\sigma}_u + \tilde{\sigma}_b - \sqrt{(\tilde{\sigma}_p + \tilde{\sigma}_b + \tilde{\sigma}_u)^2 - 4\tilde{\sigma}_p\tilde{\sigma}_u}\Big).$$

Eq. (2.34) represents the solution of the generating function of protein distribution for the three-stage model; hence, in the limit of large transcriptional rate r, the multi-scale bursting models converges to the three-stage model with transcriptional rate being $\hat{r} = s_p$.

2.5.2 Large parameter s_p

In order to simplify the solution in Eq. (2.28) for large parameter s_p, we use the parametrization $s_p \mapsto s_p/\varepsilon$ ($\tilde{\sigma}_p \mapsto \tilde{\sigma}_p/\varepsilon$) and rewrite the system of ODEs from Eq. (2.26). By following exactly the same steps as in the case for large r, one can easily derive the following second-order ODE for the leading-order full generating function, $F^{(0)}(x)$ where $x = \lambda w$ as before;

$$x(1-x)\partial_x^2 F^{(0)} + [b_1 - (1 + \tilde{\rho} + \tilde{\sigma}_u + \tilde{\sigma}_b)x]\partial_x F^{(0)} - \tilde{\rho}\tilde{\sigma}_u F^{(0)} = 0. \tag{2.35}$$

The Eq. (2.35) has exactly the same expression as the one in Eq. (2.33) but with substitution of parameter $\tilde{\rho}$ with parameter $\tilde{\sigma}_p$. This means that in the limit of large parameter s_p, the solution in Eq. (2.28) simplifies to the same expression as in Eq. (2.34), but with a change of parameters, $s_p \mapsto r$ ($\tilde{\sigma}_p \mapsto \tilde{\rho}$), which represents the solution of the three-stage model with transcription rate being $\hat{r} = r$.

2.6 Summary and discussion

In this chapter, we have presented a detailed analytical study of a multi-scale model of bursty gene expression based on recent experimental data from mammalian cells [65]. In general, it is very hard or even impossible to obtain analytical solutions for chemical master equations describing gene expression models; however, the multi-scale model is analytically tractable, and we have obtained analytical expressions for mRNA and protein distributions. Specifically, we have shown derivations of: (i) an exact closed-form expression for the steady-state distribution of mRNA molecules, (ii) simple closed-form expressions for the approximate time-dependent mRNA distribution, and (iii) an approximate steady-state distribution for protein numbers in the limit of short-lived mRNA. All the solutions for the probability distributions are in terms of hypergeometric functions, which means that for certain parameter values of the model, these distributions can present bimodality.

Additionally, we have shown that when it is impossible to solve exactly the CME of the model, we can use perturbation techniques in order to obtain approximate solutions. For example, we have found that, for fast polymerase pause release rate (r) or for fast polymerase recruitment rate (s_p), the time-dependent mRNA distribution from the three-stage model, serves as a very good approximation for the time-dependent mRNA distributions from the multi-scale model. In general, the results from the mathematical analysis show that the three-stage model is a special case of the multi-scale model; the multi-scale model takes explicitly into account the Pol II dynamics (hence, three gene states), while the three-stage model lacks this biological detail (hence, two gene states).

An extensive analysis of the multi-scale model has been performed by Zhixing Cao and has been published in [122]; the results show that the time-dependent mRNA distribution of the multi-scale

model with polymerase dynamics (without the protein part in the model; $v = 0$) can be accurately approximated by the telegraph model, modified with a Michaelis-Menten-like dependence of the effective transcription rate on polymerase abundance. Specifically, in this two-state telegraph model, the transcription rate of a gene locus is $\hat{r} = rs_p/(r + s_p)$, where s_p is the binding rate of Pol II, which is proportional to the local number of Pol II molecules at the gene locus with active transcription [124]. This equation implies that the transcription rate is proportional to the local number of Pol II molecules if s_p is approximately less than r; i.e., if the Pol II binding rate is less than or equal to the rate at which Pol II is unpaused. In contrast, if unpausing is the rate-limiting step ($r \ll s_p$), then the transcription rate is practically independent of the local Pol II number. Generally, this analysis shows that the multi-scale model supports the observation that there are differences in transcriptional activity between different stages of the cell cycle that cannot be explained by the conventional telegraph model [122] (also see discussion on page 7).

To summarize, although the multi-scale model captures two biological details (Pol II recruitment and Pol II pause release) that the simple telegraph model does not, and it is consistent with experimental data for mRNA dynamics from the study performed in [65], this model still lacks a number of significant steps in gene expression, such as Pol II elongation, Pol II pausing, Pol II premature detachment etc. Consequently, a detailed model of RNA transcription, which includes these processes, is presented in the next chapter.

Chapter 3

Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination

This chapter contains published work. Please see the Declaration of Authorship for details. The Sections 3.1-3.6 contain the published article by T.Filatova et al. [123]. The Lay summary and Section 3.7 are supplementary to the publication and are written for the purposes of this thesis.

Lay summary

In Chapter 2 we studied the "multi-scale" model of gene expression, which is characterised by three effective gene promoter states depending on the binding state of transcription factor and polymerase molecules to the gene. The switching between these gene-states constitutes the modelling of polymerase recruitment and pause release processes which are two essential steps for mRNA production. Although, C. R. Bartman et al. show in [65] that the multi-scale model is useful for investigating if transcription can be regulated by changing the rates of the two aforementioned biochemical steps, one can argue that it is inadequate for studying transcription regulation. The main reasons are the following: (i) There is experimental evidence suggesting that the time intervals of transitions between the transcriptionally inactive and active gene-states (hours) are generally longer than the time that it takes for polymerase recruitment and pause release (tens of minutes). We have shown that when the polymerase recruitment or pause release processes are fast compared to other processes in the model, then the multi-scale model simplifies to the simple telegraph model, which is inconsistent with experimental data for mRNA dynamics obtained from various studies (see discussion on page 7). This means that other steps in transcription may be the key to regulating transcriptional activity by forming a slower layer of transcriptional regulation compared to polymerase recruitment and pause release. (ii) Even though the multi-scale model is a good attempt to incorporate details into the modelling of transcription initiation, it is missing important biological details related to polymerase elongation and transcription termination. It is for these reasons that we construct and present here a new stochastic model of gene expression, where polymerase recruitment and pause release are assumed to be fast processes and the focus falls on the elongation and termination steps of transcription. Henceforth, we will refer to this model as the "detailed" model;

A thorough description of the detailed model is the following. In this model, the promoter can switch between two states, where it can be either transcriptionally active or inactive. Transcription

initiation can occur only when the promoter is active, and when it happens, we assume that an actively transcribing polymerase is loaded at the beginning of the transcriptional site of the gene. We assume that the gene has a certain length and that it is divided into an arbitrary number of segments. When a polymerase loads on the beginning of the gene, it occupies the first gene segment, and it is ready to proceed to the elongation process. We model *elongation as a random multi-step process*, meaning that the polymerase can randomly hop to the next gene segment with a constant probability per unit of time, which represents the elongation rate. The benefit of modelling elongation as a multi-step process and dividing the gene into segments is that it allows us to study the fluctuations of the number of polymerases on each gene segment separately. Additionally, we can study the fluctuations of the total number of polymerases on the gene, which we define as the sum of polymerases from all the gene segments. During elongation, a polymerase molecule actively transcribes the gene and produces nascent RNA, and the length of the *nascent RNA grows as the polymerase moves along the gene*; hence, the number of polymerases is closely related to the number of nascent RNA molecules in our model. Experimental studies have indicated that *polymerase pausing* is a common regulatory step in the transcription of many genes; hence, we include this biological process in our model as well. As elongation proceeds in our detailed model, the polymerase can switch randomly to a paused state, whereas it can again randomly switch back to the actively moving state. *Polymerase premature detachment* is rare and usually is not considered in stochastic models of gene expression; however, other studies have inspired us to incorporate this biological step in our detailed model. This means that a polymerase molecule (paused or actively moving) can detach from the gene, independently of its location on it. Elongation is complete when the polymerase molecule reaches the end of the gene. The polymerase can fall off the last gene segment, which models the *transcription termination process*; if this happens, the nascent RNA produced from this polymerase becomes mature RNA. Finally, after a molecule of mature RNA is produced, it can degrade with a certain probability per unit of time.

As one can see, our detailed model does capture more biological details than the telegraph model; however, this model has certain restrictions: (i) The number of gene segments must be sufficiently large for the dynamics to be described at a fine spatial resolution, while at the same time the length of a gene segment must be larger than the length of a polymerase if the polymerase is treated as a solid object. (ii) The transcription initiation must be a sufficiently slower process than the elongation in order to prevent polymerase interactions with each other while they move along the gene (otherwise, polymerase traffic can occur).

The five species of interest in our detail model are the following: (1) polymerases on each gene segment; from now on we will call them the local polymerases, (2) total polymerases on the gene, which is defined as the sum of local polymerases, (3) nascent RNAs on each gene segment or the so-called local nascent RNA, (4) total nascent RNAs on the gene, which is defined as well as the sum of local nascent RNAs, and finally (5) mature RNA. We are interested in understanding how the number of molecules of these species fluctuates when our system has reached a steady-state. This means that we have performed our mathematical analysis only in the limit of large time. By using all the methods described in Section 1.5, we have obtained analytical expressions for the mean and the variance of the number of molecules for all the species. Additionally, we have obtained analytical expressions for the total polymerase and mature RNA distributions by applying some approximation techniques.

Here, it is worth noting that the rates of: promoter switching, transcription initiation, polymerase pausing and activation, premature detachment, and mature mRNA degradation are all experimentally measurable. The length of polymerases and genes, the elongation time, and the polymerase elongation speed can also be measured experimentally. The value of the polymerase hopping rate from one gene segment to the next can also be estimated from experimental data since it is defined as the ratio of the number of gene segments over total elongation time in our model. This means that the analytical expressions for the moments and distributions of the species in our model can provide means to estimate the values of transcriptional parameters involved in the model.

Abstract

Recent advances in fluorescence microscopy have made it possible to measure the fluctuations of nascent (actively transcribed) RNA. These closely reflect transcription kinetics, as opposed to conventional measurements of mature (cellular) RNA, whose kinetics is affected by additional processes downstream of transcription. Here, we formulate a stochastic model which describes promoter switching, initiation, elongation, premature detachment, pausing, and termination while being analytically tractable. We derive exact closed-form expressions for the mean and variance of nascent RNA fluctuations on gene segments, as well as of total nascent RNA on a gene. We also obtain exact expressions for the first two moments of mature RNA fluctuations, and approximate distributions for total numbers of nascent and mature RNA. Our results, which are verified by stochastic simulation, uncover the explicit dependence of the statistics of both types of RNA on transcriptional parameters and potentially provide a means to estimate parameter values from experimental data.

3.1 Introduction

Transcription, the production of RNA from a gene, is an inherently stochastic process. Specifically, the interval of time between two successive transcription events is a random variable whose statistics depend on multiple single-molecule events behind transcription [50]. When the distribution of this random variable is exponential, we say that expression is constitutive; in that case, the number of transcripts produced in a certain interval of time follows a Poisson distribution. On the other hand, when the distribution of times between two successive transcripts is non-exponential, then the number of transcripts is non-Poissonian. A special case of such non-constitutive behaviour is the bursty expression, whereby transcripts are produced in short bursts that are separated by long silent intervals [52, 53]. In yeast, genes whose expression is constitutive include MDN1, KAP104, and DOA1, whereas PDR5 is an example of a gene whose expression is bursty [48].

For two decades, mathematical models of gene expression have been developed to predict the distribution of RNA abundance. By matching the theoretical distribution with experimental measurements from microscopy-based methods [125], one hopes to obtain insight into the underlying kinetics of transcription and estimate transcriptional parameters. The standard model of gene expression which has been used for these analyses is the telegraph model [61], whereby a gene can be in two states. Transcription occurs in one of the states, whereupon RNA degrades; first-order kinetics is assumed for all processes. Even though the distribution from the telegraph model generally fits well cellular RNA data, there are innate difficulties with the interpretation of that fit (see discussion on page 7).

To counteract these difficulties, in the past few years, mathematical models [70–72, 126] have been developed to predict the statistics of nascent RNA, i.e. of RNA in the process of being synthesised by the RNA polymerase molecule (RNAP), which can be visualised and quantified due to recent advances in fluorescence microscopy [5, 32, 127–129]. In contrast to cellular RNA, the statistics of nascent RNA are a direct reflection of the transcription process; hence, these models can potentially give more insight than the simpler, but cruder telegraph model. Choubey and collaborators [70, 71] have developed a stochastic model with the following properties: (i) a gene can be in two states (active or inactive); (ii) from the active state, transcription initiation occurs in two sequential steps: the pre-initiation complex is formed, after which the RNA polymerase escapes the promoter; (iii) once on the gene, the polymerase moves from one base pair to the next (with some probability) until the end of the gene is reached, when transcription is terminated and polymerase detaches. Queuing theory is used to derive analytical expressions for the transient and steady-state means and variances of numbers of RNAP that are attached to the gene in the long-gene limit when the elongation time is practically deterministic. Xu et al. [72] have considered a coarse-grained version of that model, whereby the movement of RNAP from one base pair to the next is not explicitly modelled, obtaining an analytical expression for the total RNAP distribution in steady-state conditions. More recently, Cao and Grima [126] have studied a model of eukaryotic

gene expression that yields approximate time-dependent distributions of both nascent and cellular RNA abundance as a function of the parameters controlling gene switching, DNA duplication, partitioning at cell division, gene dosage compensation, and RNA degradation; in their coarse-grained model, the movement of RNAP is not explicitly modelled, while the elongation time is assumed to be exponentially distributed, which simplifies the requisite analysis.

The complexity of nascent RNA models has thus far not allowed the same detailed level of analysis as has been possible with the much simpler telegraph model. A few shortcomings of current models can be summarised as follows: (i) distributions of nascent RNA have been derived from models that do not explicitly model the movement of RNAP along a gene [72, 126], resulting in a disconnect between theoretical description and the microscopic processes underlying transcription; (ii) while the analysis of single-cell sequencing data and electron micrograph data yields the positions of individual polymerases along the gene, allowing for the calculation of statistics (means and variances) of the numbers of RNAP on gene segments that are obtained after binning, detailed models of RNAP elongation [70, 71] provide analytical results only for total RNAP on a gene and hence cannot be used to understand gene segment data; (iii) analytical calculations of the statistics of nascent RNA ignore important details of the transcription process such as pausing, traffic jams, backtracking, and premature termination, some of which have to-date been explored via stochastic simulation [70, 114, 130–132].

In this study, we overcome some of the aforementioned shortcomings of analytically tractable models for the transcription process. In Section 3.2, we study a stochastic model for promoter switching and the stochastic movement of RNAP along a gene, allowing for premature termination. We derive exact closed-form expressions for the first and second moments (means and variances) of local RNAP fluctuations on gene segments of arbitrary length, which allows us to study how these statistics vary along a gene as a function of transcriptional parameters; we also obtain expressions for the mean and variance of the total RNAP on the gene which generalize previous work by Choubey et al. [70]. In Section 3.3, we investigate approximations for the distributions of total RNAP and mature RNA, showing in particular that Negative Binomial distributions can provide an accurate approximation in certain biologically meaningful limits. In Section 3.4, we illustrate the difference between the statistics of local and total RNAP fluctuations and those of light fluorescence due to tagged nascent RNA. In Section 3.5, we extend our model to include pausing by deriving approximate expressions for the mean, variance, and distribution of observables. We conclude with a discussion of our results in Section 3.6 and possible extensions of our model in Section 3.7.

3.2 Detailed stochastic model of transcription: setup and analysis

In this section, we specify the stochastic model studied here; then, we derive closed-form expressions for the moments of mature RNA and of local and total RNAP fluctuations in various parameter regimes.

3.2.1 Setup of model

We consider a stochastic model of transcription that includes the processes of initiation, elongation, and termination, as illustrated in Fig. 3.1. For simplicity, we divide the gene into L segments; the RNAP on gene segment i is then denoted by P_i. The promoter can be either in the inactive state (G_{off}) or the active state (G_{on}), switching from the inactive state to the active one with rate s_u and from the active state to the inactive one with rate s_b. When the promoter is active, initiation commences via the binding of an RNAP with rate r, denoted by P_1. Subsequently, the RNAP either moves from a gene segment to the neighbouring segment with rate k, or it prematurely detaches with rate d. Note that here we have made two assumptions: (i) the movement of RNAP is unidirectional, away from the promoter site and hence left to right, with no pausing or backtracking allowed; (ii) the detachment and elongation rates are independent of the position of RNAP on the gene. Each RNAP has associated with it a nascent RNA tail that grows longer as the RNAP transcribes more of the gene. When the RNAP reaches the last gene segment, termination occurs,

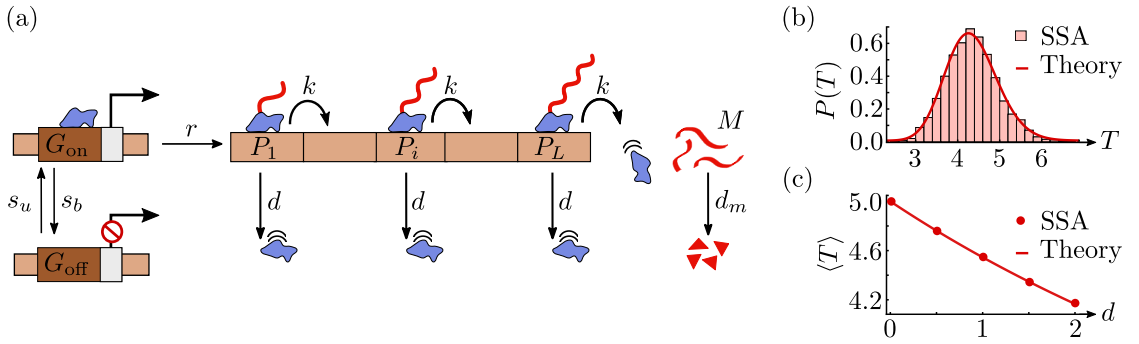


Figure 3.1: Model of transcription. (a) The gene is arbitrarily divided into L segments, with RNAP (blue) on gene segment i denoted by P_i. The promoter switches from the active state $G_{\rm on}$ to the inactive state $G_{\rm off}$ with rate s_b, while the reverse switching occurs with rate s_u. When the promoter is active, initiation of RNAP occurs with rate r. Initiation is followed by elongation, which is modelled as RNAP 'hopping' from gene segment i to the neighbouring segment $i+1$ with rate k, i.e. as the transformation of species P_i to P_{i+1}. RNAP prematurely detaches from the gene with rate d. A nascent RNA tail (red), attached to the RNAP, grows as elongation proceeds. Termination is modelled by the change of P_L with rate k to mature RNA (M), which subsequently degrades with rate d_m. In panel (b), we show the probability distribution $P(T)$ of the total elongation time T – the time between initiation and termination – as predicted by the stochastic simulation algorithm (SSA; histogram) and our theory (Erlang distribution with shape parameter L and rate $k+d$; solid line). The parameter values used are $L=50$, $k=10/\text{min}$, and $d=1.5/\text{min}$. In panel (c), we show the dependence of the mean of the distribution $P(T)$ on the RNAP detachment rate (d), as predicted by SSA (dots) and our theory ($\langle T\rangle = L/(k+d)$; solid line). The relevant parameter values are $L=50$ and $k=10/\text{min}$.

i.e. the RNAP-nascent RNA complex gets dissociated from the gene leading to a mature RNA (M) which degrades with rate d_m. Note that for simplicity, we have not considered excluded-volume interaction between adjacent RNAPs here; hence, we make the implicit assumption of low 'traffic', which is plausible when the initiation rate is sufficiently low. (We test the validity of this assumption through simulations below.)

Since several polymerase molecules are usually located on a gene during transcription, together with the fact that the electron micrographs can reveal the number of polymerases engaged in transcribing a single gene and their distance from the gene promoter, it makes sense for the purpose of mathematical modelling to divide the gene into L segments. The proper definition for the non-dimensional, non-negative integer parameter, L is the ratio of the gene length over the length of a gene segment. Note that, while the choice of L is arbitrary, it should be kept in mind that L needs to be sufficiently large for the dynamics to be described at a fine spatial resolution. However, L also has to be small enough for the length of each gene segment to be much larger than the footprint of an RNAP; the latter is needed to ensure the validity of the low-traffic assumption. The main reason for dividing the gene into segments is to consider a case other than the simple exponentially distributed elongation time. Here, the elongation time which is the total time T from initiation to termination, that is, conditioning on those realisations for which the RNAP does not prematurely detach, is Erlang distributed with mean $\langle T\rangle = L/(k+d)$ and coefficient of variation $1/\sqrt{L}$; see Appendix A.1 for a derivation and Figs. 3.1(b) and (c) for verification through stochastic simulation (SSA). To summarise, if $L=1$ then we have exponentially distributed elongation time, while the larger L is, the narrower is the distribution of T and the more deterministic is elongation itself.

Note that the total number of RNAPs transcribing the gene is equal to the number of nascent RNA molecules present, *irrespective of their lengths*; to shed light on the fluctuations of nascent

RNA, in this section we, therefore, focus on the calculation of statistics of local and total RNAP fluctuations. We define the vector of molecule numbers $\vec{m} = (n_0, n_1, \ldots, n_L, n)$, and we write $\langle n_0 \rangle$, $\langle n_i \rangle$ $(i = 1, 2, \ldots, L)$, and $\langle n \rangle$ for the average numbers of molecules of active gene, RNAP, and mature RNA, respectively. The above model can then be conveniently described by $L + 2$ species interacting via a set of $2L + 4$ reactions with the following rate functions:

Species	Molecule numbers	Position (in $\vec{m}$)
G_{on}	n_0	1
$P_i, \quad i \in \{1, \ldots, L\}$	n_i	$i + 1$
M	n	$L + 2$

Reaction	Rate function f_j
$G_{\text{on}} \xrightarrow{s_b} G_{\text{off}}$	$f_1 = s_b \langle n_0 \rangle$
$G_{\text{off}} \xrightarrow{s_u} G_{\text{on}}$	$f_2 = s_u(1 - \langle n_0 \rangle)$
$G_{\text{on}} \xrightarrow{r} G_{\text{on}} + P_1$	$f_3 = r \langle n_0 \rangle$
$P_i \xrightarrow{k} P_{i+1}, \quad i \in \{1, \ldots, L-1\}$	$f_{i+3} = k \langle n_i \rangle$
$P_L \xrightarrow{k} M$	$f_{L+3} = k \langle n_L \rangle$
$P_i \xrightarrow{d} \varnothing, \quad i \in \{1, \ldots, L\}$	$f_{i+L+3} = d \langle n_i \rangle$
$M \xrightarrow{d_m} \varnothing$	$f_{2L+4} = d_m \langle n \rangle$

Note that G_{off} is not an independent species; the reason is that the binary state of the gene implies a conservation law, with the sum of the numbers of G_{on} and G_{off} equalling 1. Hence, the number of independent species in the model is $L + 2$. The rate functions f_j are the averaged propensities from the underlying chemical master equation (CME); note that, because our reaction network is composed of first-order reactions, these rate functions also equal the reaction rates in the corresponding deterministic rate equations. The description of our model is completed by the $(L + 2) \times (2L + 4)$-dimensional stoichiometric matrix $\mathbf{S}$; the element $\mathbf{S}_{ij}$ of $\mathbf{S}$ gives the net change in the number of molecules of the i-th species when the j-th reaction occurs. Given the ordering of species and reactions as described in the Tables above, it follows that the matrix $\mathbf{S}$ has the simple form

$$\begin{aligned} \mathbf{S}_{11} &= -1, & \mathbf{S}_{12} &= 1, & & \\ \mathbf{S}_{i,i+1} &= 1, & \mathbf{S}_{i,i+2} &= -1, & \mathbf{S}_{i,i+L+2} &= -1, \\ \mathbf{S}_{L+2,L+3} &= 1, & \mathbf{S}_{L+2,2L+4} &= -1, & & \end{aligned} \tag{3.1}$$

where $i = 2, \ldots, L + 1$.

3.2.2 Closed-form expressions for moments of mature RNA and local RNAP

In this subsection, we outline the derivation of the steady-state means and variances of local RNAP fluctuations (on each gene segment), as well as of mature RNA. Our results are summarised in the following two propositions.

Proposition 1. *Let $\eta = s_u/(s_u + s_b)$ be the fraction of time the gene spends in the active state, let $\rho_k = r/k$ be the mean number of RNAPs binding to the promoter site in the time it takes for a single RNAP to move from one gene segment to the next, let $\rho = r/d_m$ be the mean number of RNAPs binding to the promoter site in the time it takes for a mature RNA to decay, and let $\mu = k/(k + d)$ be the probability that an RNAP molecule moves to the next gene segment rather than detaching prematurely. Then, the steady-state mean numbers of molecules of the active gene, RNAP, and mature RNA are given by*

$$\langle n_0 \rangle = \eta, \tag{3.2a}$$

$$\langle n_i \rangle = \eta \rho_k \mu^i \qquad \text{for } i = 1, \ldots, L, \tag{3.2b}$$

$$\langle n \rangle = \eta \rho \mu^L, \tag{3.2c}$$

respectively.

Prop. 1 can be proved in a straightforward fashion, as follows. Using the underlying CME, one can show from the corresponding moment equations [100] that the time evolution of the vector $\langle\vec{m}\rangle$ of mean molecule numbers in a system of zeroth-order or first-order reactions, i.e. with propensities that are linear in the number of molecules, is given by the time derivative $d\langle\vec{m}\rangle/dt = \mathbf{S}\cdot\vec{f}$. Given the form of the stoichiometric matrix $\mathbf{S}$ and of the rate functions f_j, as described in Section 3.2.1, it follows that the mean numbers of all species in steady-state can be obtained by solving the following system of $L+2$ algebraic equations:

$$\begin{aligned} 0 &= s_u(1-\langle n_0\rangle) - s_b\langle n_0\rangle, \\ 0 &= r\langle n_0\rangle - (k+d)\langle n_1\rangle, \\ 0 &= k\langle n_{i-1}\rangle - (k+d)\langle n_i\rangle \qquad \text{for } i = 2,\dots,L, \\ 0 &= k\langle n_L\rangle - d_m\langle n\rangle. \end{aligned} \tag{3.3}$$

These equations can easily be solved simultaneously to yield the steady-state value of $\langle\vec{m}\rangle$, as given in Eq. (3.2).

Proposition 2. *Let $\tau_p = 1/(d+k)$, $\tau_g = 1/(s_u+s_b)$, and $\tau_m = 1/d_m$ be the timescales of fluctuations of RNAP, gene, and mature RNA, respectively, and define the three new parameters*

$$\alpha = \frac{1}{1+\tau_p/\tau_g}, \quad \gamma = \frac{1}{1+\tau_p/\tau_m}, \quad \text{and} \quad \theta = \frac{1}{1+\tau_m/\tau_g}.$$

Furthermore, let $\beta = s_b/s_u$ denote the ratio of gene inactivation and activation rates. Then, the variances and covariances of molecule number fluctuations of the active gene, RNAP, and mature RNA are given by

$$\mathrm{Var}(n_0) = \langle n_0\rangle^2\beta, \tag{3.4a}$$

$$\mathrm{Cov}(n_0,n_i) = \langle n_0\rangle\langle n_i\rangle\alpha\beta\cdot f_{1i}, \qquad \text{where } f_{1i} = \alpha^{i-1}; \tag{3.4b}$$

$$\mathrm{Cov}(n_0,n) = \langle n_0\rangle\langle n\rangle\alpha\beta\cdot f_{1M}, \qquad \text{where } f_{1M} = \theta\alpha^{L-1}, \tag{3.4c}$$

$$\mathrm{Cov}(n_i,n_j) = \delta_{ij}\langle n_i\rangle + \langle n_i\rangle\langle n_j\rangle\alpha\beta\cdot f_{ij}, \qquad \text{where } f_{ij} = f(i,j)+f(j,i), \tag{3.4d}$$

$$\mathrm{Cov}(n_i,n) = \langle n_i\rangle\langle n\rangle\alpha\beta\cdot f_{iM}, \qquad \text{where } f_{iM} = \gamma^i\theta\alpha^{L-1} + (1-\gamma)\sum_{q=1}^{i}\gamma^{i-q}f_{qL}, \tag{3.4e}$$

$$\mathrm{Var}(n,n) = \langle n\rangle + \langle n\rangle^2\alpha\beta\cdot f_{MM}, \qquad \text{where } f_{MM} = f_{LM}, \tag{3.4f}$$

and where $i,j = 1,\dots,L$. Here, δ_{ij} is the Kronecker delta; moreover,

$$f(i,j) = \frac{\alpha^{i+j-1}}{(2\alpha-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\alpha-1}{2\alpha}\,{}_2F_1\big(1,i+j;j;\tfrac{1}{2\alpha}\big)\Big],$$

where ${}_2F_1$ denotes the generalised hypergeometric function of the second kind [44], which is defined as

$${}_2F_1(a_1,a_2;b_1;z) = \sum_{s=0}^{\infty}\frac{(a_1)_s(a_2)_s}{(b_1)_s}\frac{z^s}{s!},$$

with $(a)_s = \Gamma(a+s)/\Gamma(a)$ the Pochhammer symbol.

Here, we note that an alternative representation of the functions f_{ij} in Eq. (3.4d), in terms of finite sums, is given in Eq. (A.2.33) of Appendix A.2.

As above, since the underlying propensities are linear in the number of molecules, the CME implies [100] that the corresponding second moments in steady-state are exactly given by a Lyapunov equation. That equation, which is precisely the same as the one that is obtained from the linear-noise approximation (LNA) [98], takes the form

$$\mathbf{J}\cdot\mathbf{C}+\mathbf{C}\cdot\mathbf{J}^T+\mathbf{D}=\mathbf{0}. \tag{3.5}$$

Here, $\mathbf{C}$, $\mathbf{J}$, and $\mathbf{D}$ are $(L+2)\times(L+2)$-dimensional matrices; $\mathbf{C}$ is a variance-covariance matrix that is symmetric ($\mathbf{C}_{ij}=\mathbf{C}_{ji}$), $\mathbf{J}$ is the Jacobian matrix with elements $\mathbf{J}_{ij}=\partial(\mathbf{S}\cdot\vec{f})_i/\partial\langle n_j\rangle$, and $\mathbf{D}=\mathbf{S}\cdot\mathbf{Diag}(\vec{f})\cdot\mathbf{S}^T$ is a diffusion matrix, where $\mathbf{Diag}(\vec{f})$ is a diagonal matrix whose elements are the entries in the rate function vector $\vec{f}$. The non-zero elements of $\mathbf{J}$ are given by

$$\begin{aligned}
\mathbf{J}_{11}&=-(s_u+s_b), & & & \\
\mathbf{J}_{21}&=r, & \mathbf{J}_{22}&=-(k+d), & \\
\mathbf{J}_{i,i-1}&=k, & \mathbf{J}_{ii}&=-(k+d) & \text{for } i=3,\dots,L+1,\\
\mathbf{J}_{L+2,L+1}&=k, & \mathbf{J}_{L+2,L+2}&=-d_m, &
\end{aligned} \tag{3.6}$$

while the non-zero elements $\mathbf{D}_i$ read

$$\begin{aligned}
\mathbf{D}_{11}&=s_b\langle n_0\rangle+s_u(1-\langle n_0\rangle), & & & \\
\mathbf{D}_{22}&=r\langle n_0\rangle+(k+d)\langle n_1\rangle, & \mathbf{D}_{23}&=-k\langle n_1\rangle, & \\
\mathbf{D}_{i,i-1}&=-k\langle n_{i-2}\rangle, & \mathbf{D}_{ii}&=k\langle n_{i-1}\rangle+(k+d)\langle n_i\rangle & \text{for } i=3,\dots,L+1,\\
\mathbf{D}_{i,i+1}&=-k\langle n_{i-1}\rangle & & & \text{for } i=3,\dots,L,\\
\mathbf{D}_{L+2,L+1}&=-k\langle n_L\rangle, & \mathbf{D}_{L+2,L+2}&=k\langle n_L\rangle+d_m\langle n\rangle. &
\end{aligned} \tag{3.7}$$

Given the structure of the matrices $\mathbf{J}$ and $\mathbf{D}$ above, the Lyapunov Eq. (3.5) can be solved explicitly for the covariance matrix $\mathbf{C}$ whose elements are given by Eq. (3.4). The solution by induction is involved and can be found in Appendix A.2, which proves Prop. 2.

Simplification in bursty and constitutive limits

Bursty limit. We now consider a particular parameter regime – the limit of large initiation rate r and large gene inactivation rate, s_b such that $b=r/s_b$ is constant. Since the fraction of time spent in the active state is η, it follows that the gene is mostly in the inactive state in that limit. During the short periods of time when it transitions to the active state, a burst of initiation events occurs; in particular, a mean number b of RNAPs bind to the promoter during activation. Hence, such genes are often termed bursty, since transcription proceeds via sporadic bursts of activity and b is called the mean transcriptional burst size. For r and s_b large with b constant, the expressions for the first two moments of RNAP at every gene segment and of mature RNA from Eqs. (3.2) and (3.4), respectively, simplify to

$$\langle n_i\rangle_b=b\upsilon_k\mu^i, \tag{3.8a}$$

$$\langle n\rangle_b=b\upsilon_m\mu^L, \tag{3.8b}$$

$$\mathrm{Cov}(n_i,n_j)_b=\delta_{ij}\langle n_i\rangle_b+\langle n_i\rangle_b\langle n_j\rangle_b(\upsilon_k\mu)^{-1}\cdot h_{ij},\qquad \text{where } h_{ij}=\frac{1}{2^{i+j-2}}\frac{\Gamma(i+j-1)}{\Gamma(i)\Gamma(j)}, \tag{3.8c}$$

$$\mathrm{Cov}(n_i,n)_b=\langle n_i\rangle_b\langle n\rangle_b(\upsilon_k\mu)^{-1}\cdot h_{iM},\qquad \text{where } h_{iM}=(1-\gamma)\sum_{q=1}^{i}\gamma^{i-q}\cdot h_{qL} \tag{3.8d}$$

$$\mathrm{Var}(n)_b=\langle n\rangle_b+\langle n\rangle_b^2(\upsilon_k\mu)^{-1}\cdot h_{MM},\qquad \text{where } h_{MM}=h_{LM}; \tag{3.8e}$$

here, the subscript b denotes the moments in the bursty limit. Moreover, $\upsilon_k=s_u/k$, $\upsilon_m=s_u/d_m$, and $h_{ij}=f_{ij}|_{\alpha\to 0}$ denotes the simplified function f_{ij} in the limit of $\alpha\longrightarrow 0$, which is achieved when

$s_b \to \infty$. We note that the above expressions for the functions h_{ij} are derived from the expressions for f_{ij} that are given in Eq. (A.2.33), rather than from those in Eq. (3.4d). The reason is that, in the bursty limit, we have that $\frac{1}{2\alpha} \to \infty$, in which case the identity in Eq. (A.2.36) does not hold. The bursty limit in Eq. (A.2.33) is simply taken by collecting terms that are not dependent on α, since $\alpha \longrightarrow 0$ in that limit.

To test the accuracy of our theory, in Fig. 3.2 we compare our analytical expressions for the mean of local RNAP numbers, as well as for various measures of local RNAP fluctuations – the coefficient of variation CV, the Fano factor FF, and the Pearson correlation coefficient CC – with those calculated from stochastic simulation using Gillespie's algorithm (SSA) [89]. Simulations are performed for two different scenarios: (i) without volume exclusion, where the footprint of RNAPs is not taken into account; and (ii) with volume exclusion, where RNAPs are treated as solid objects with a footprint of 35bp, which is the value reported in [132]. For our simulations in Fig. 3.2, we use parameter values characteristic for the gene PDR5 of length 3070bp, as reported in [48]. Our choice of $L = 30$ implies that the length of each gene segment is about 100bp and, hence, that at most 3 RNAPs can fit in each segment when volume exclusion is taken into account. In this case, the Gillespie's algorithm is modified such that the initiation and RNAP 'hopping' rates are proportional to the available volume in the gene segment which the RNAP is moving to. That is achieved by rescaling the transcription initiation rate as $r \mapsto r(1 - n_1/3)$ and the RNAP hopping rate from the i-th to the $(i + 1)$-th gene segment as $k \mapsto k(1 - n_{i+1}/3)$. Since we use parameters measured for a gene that demonstrates bursty expression (PDR5) [48], we test the accuracy of both the exact theory from Eqs. (3.2) and (3.4) and the approximate expressions given in Eq. (3.8).

The perfect agreement between our exact theory (solid lines) and simulation without volume exclusion (dots) provides a numerical validation of that theory. Our approximate theory (dashed lines) also yields a reasonably good approximation; the mismatch can be decreased if the degree of burstiness is increased, i.e. by increasing the parameters r and s_b relative to the other rates in the model. We also note that the theory is in good agreement with simulation with volume exclusion (open circles), which shows that the 'low traffic' assumption upon which our theory is based is valid.

The following interesting observations can be made from these figures: (i) if the rate of premature detachment is greater than zero, then the mean of local RNAP decreases monotonically with the distance i from the promoter according to a power law, whereas that mean is constant along the gene if there is no premature detachment, as expected; (ii) the size of RNAP fluctuations, as measured by CV, decreases with i for small premature detachment rates, but increases with i for sufficiently large values of the detachment rate; (iii) the Fano factor approaches 1 – the value of FF for a Poissonian distribution – as i increases, which is due to the dispersal of the burst as stochastic elongation proceeds; (iv) the correlation coefficient between the local RNAP on two neighbouring gene segments decreases monotonically with i, which is exacerbated by premature detachment and is a direct result of the stochasticity inherent in the elongation process.

The observation in (iii) can be explained in detail as follows. When the detachment rate is zero, a burst of RNAPs rapidly bind to the promoter, leading to large fluctuations near that site; however, thereafter each RNAP moves distinctly from all others due to stochastic elongation. Hence, the burst is gradually dispersed as elongation proceeds, which implies a decrease in the variance of fluctuations with increasing i. When the detachment rate is non-zero, then the same effect is at play; however, the increase in the variance of fluctuations along the gene is now counteracted by the decrease of mean RNAP numbers, which leads to two types of behaviour: for small i, CV decreases with i, since the variance dominates over the mean, while for large i, the opposite occurs and CV increases with i.

Constitutive limit. The other common parameter regime is that of constitutive gene expression, where the gene spends most of its time in the active state and transcription is continuous, which corresponds to the limit of very small s_b. In that limit, the expressions from Eqs. (3.2) and (3.4) simplify to

$$\langle n_i \rangle_c = \mathrm{Var}(n_i)_c = \rho_k \mu^i \qquad \text{and} \qquad \langle n \rangle_c = \mathrm{Var}(n)_c = b\rho\mu^L, \tag{3.9}$$

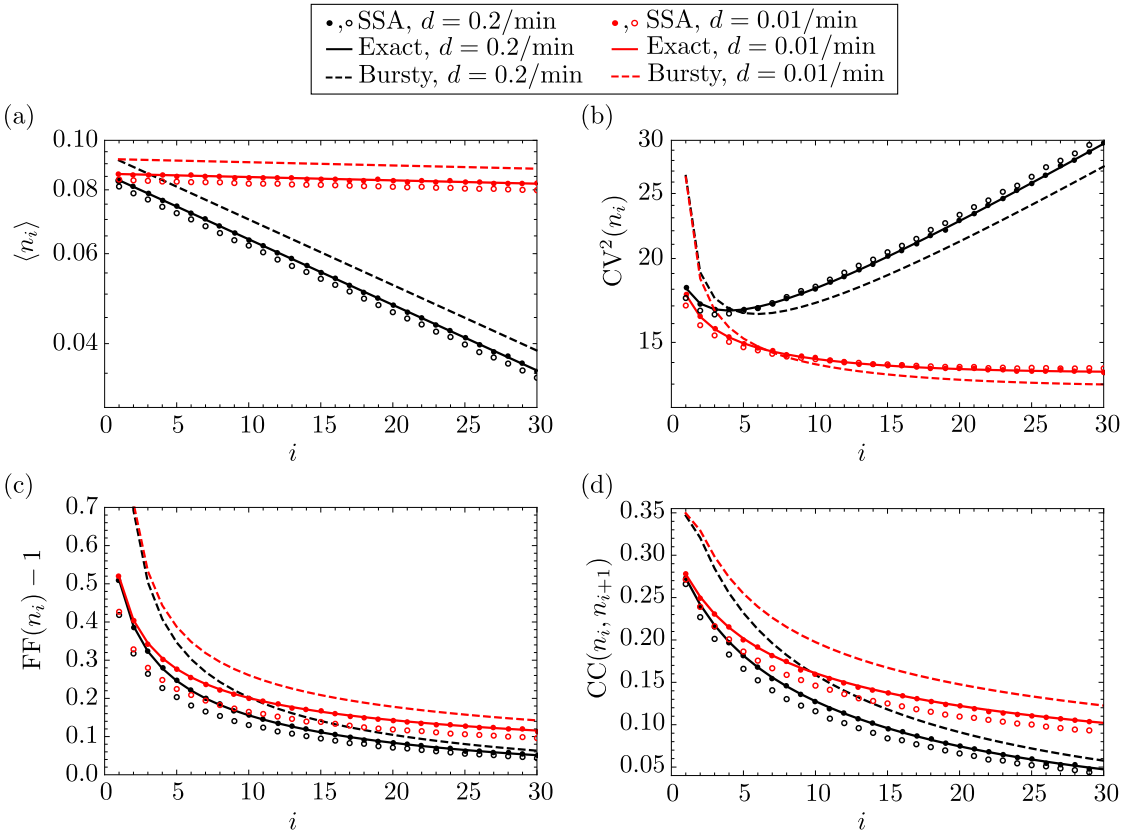


Figure 3.2: First and second moments of the distribution of local RNAP for the PDR5 gene in yeast, which demonstrates bursty expression. In panels (a), (b), (c), and (d), we show the dependence of the mean, coefficient of variation squared, Fano factor, and Pearson correlation coefficient, respectively, of local RNAP fluctuations on gene segment i, as predicted by our exact theory (Eqs. (3.2) and (3.4); solid lines), the approximate theory in the bursty limit (Eq. (3.8); dashed lines), and simulation via Gillespie's stochastic simulation algorithm (SSA), respectively. We performed simulations for two different cases: without volume exclusion (dots) and with volume exclusion (open circles). The parameters are fixed to $s_u = 0.44$/min, $s_b = 4.7$/min, and $r = 6.7$/min, which are characteristic of the PDR5 gene in yeast, as reported in Supplemental Table 2 of [48]. The number of gene segments is arbitrarily chosen to be $L = 30$. The total elongation time $\langle T \rangle = 4.5$ min is also reported for PDR5, described as the synthesis time and denoted by τ in [48]. The elongation rate by definition takes the value of the ratio $k = L/\langle T \rangle - d \approx L/\langle T \rangle$, since $d \ll k$. The detachment rate d is arbitrarily chosen to be $d = 0.01$/min (red lines and dots) or $d = 0.2$/min (black lines and dots). Note that, for the SSA, moments are calculated from one long trajectory with a few million time points, sampled at unit intervals.

while the covariances $\mathrm{Cov}(n_i, n_j)_c$ and $\mathrm{Cov}(n_i, n)_c$ between the species are zero; here, the subscript c denotes the constitutive limit. This drastic simplification reflects the fact that, in the constitutive limit, the distributions of mature RNA and local RNAP are Poissonian: as the regulatory network is effectively given by $\emptyset \to P_1 \to P_2 \to ... \to P_L \to M \to \emptyset$ then, the result follows directly from the exact solution provided in [133].

To further test the accuracy of our theory, in Fig. 3.3 we compare our analytical expressions for the mean of local RNAP numbers, as well as for various measures of local RNAP fluctuations, with those calculated from stochastic simulation using Gillespie's algorithm, where we use parameters measured for a gene that demonstrates constitutive expression (DOA1) [48]. As before, we test

the accuracy of both the exact theory given by Eqs. (3.2) and (3.4) and the approximate expressions from Eq. (3.9). Unsurprisingly, we observe agreement between exact theory (solid lines) and simulation (dots); the mismatch between our approximate theory and simulation is due to the fact that the gene does not spend 100% of its time in the active state – the true constitutive limit – but, rather, $s_u/(s_u + s_b) \approx 85\%$. The local mean RNAP number decreases with distance from the promoter, as was the case for bursty expression in the previous subsubsection, which is to be expected. The various measures which depend on the second moments are, however, considerably different: CV increases monotonically with i, independently of the rate of premature detachment, while FF and CC are very close to 1 and zero, respectively; moreover, the latter two measures practically show very little variation along the gene. The lack of transcriptional bursting explains all these effects in a straightforward fashion.

Finally, we remark that the accuracy of our expressions for the mean and variance of mature RNA, as given in Eq. (3.2) and (3.4), is verified by simulation (SSA) in Figs. 3.4(a) and (b) for parameters typical of the bursty PDR5 gene. The meaning of the dependence of descriptive statistics on L is discussed in the next section.

3.2.3 Closed-form expressions for moments of total RNAP

While local RNAP fluctuations are measurable in experiment, as discussed in the Introduction, measurements of total RNAP on a gene are typically reported. Hence, in this section, we briefly discuss descriptive statistics of total RNAP fluctuations.

Recalling that n_i is the number of RNAP molecules on the i-th gene segment, the total number of RNAPs on the gene – arbitrarily divided into L segments – is given by $n_{tot} = \sum_{i=1}^{L} n_i$. Given Eq. (3.2) and (3.4), the steady-state mean $\langle n_{tot} \rangle = \sum_{i=1}^{L} \langle n_i \rangle$ and the steady-state variance $\mathrm{Var}(n_{tot}) = \sum_{i,j=1}^{L} \mathrm{Cov}(n_i, n_j)$ of the total RNAP distribution are given by

$$\langle n_{tot} \rangle = \eta \rho_k \mu \frac{\mu^L - 1}{\mu - 1} \quad \text{and} \quad \mathrm{Var}(n_{tot}) = \langle n_{tot} \rangle + \alpha\beta(\eta\rho_k)^2 \sum_{i,j=1}^{L} \mu^{i+j} \cdot f_{ij}. \tag{3.10}$$

For a detailed derivation of the variance in Eq. (3.10), we refer to Appendix A.3. These expressions for the mean and variance of the total RNAP distribution simplify in the bursty and constitutive limits, as can be seen in Appendix A.4. The accuracy of Eq. (3.10) is tested by comparing against stochastic simulation with SSA in Figs. 3.4(c) and (d). Both mean and variance are seen to increase monotonically with the number of gene segments L, as we keep the mean elongation time constant; the mean shows very little dependence on L, while the dependence of the variance is more pronounced. We recall that, while the parameter L is arbitrary in principle, it actually determines the size of fluctuations in the elongation time. Since that time is the sum of L independent exponential variables with mean $1/(k + d)$ each, it follows that the distribution of the elongation time T is Erlang with mean $\langle T \rangle = L/(k + d)$ and coefficient of variation squared equal to $1/L$. Hence, the larger L is, the narrower is the distribution of T and the more deterministic is elongation itself. Thus, Figs. 3.4(c) and (d) predict that the mean and variance of total RNAP increase rapidly with decreasing fluctuations in the elongation time T. It hence follows that models in which the elongation rate is assumed to be exponentially distributed [126], which correspond to the case where $L = 1$ in our model, underestimate the size of nascent RNA fluctuations.

3.2.4 Special case of deterministic elongation

Next, we derive expressions for the descriptive statistics of total RNAP and mature RNA in the limit of large L taken at constant mean elongation time, which corresponds to deterministic elongation. As is shown in Fig. 3.4, these statistics converge quickly to the ones obtained in the large-L limit; hence, the resulting limiting expressions are likely to be useful across a variety of genes.

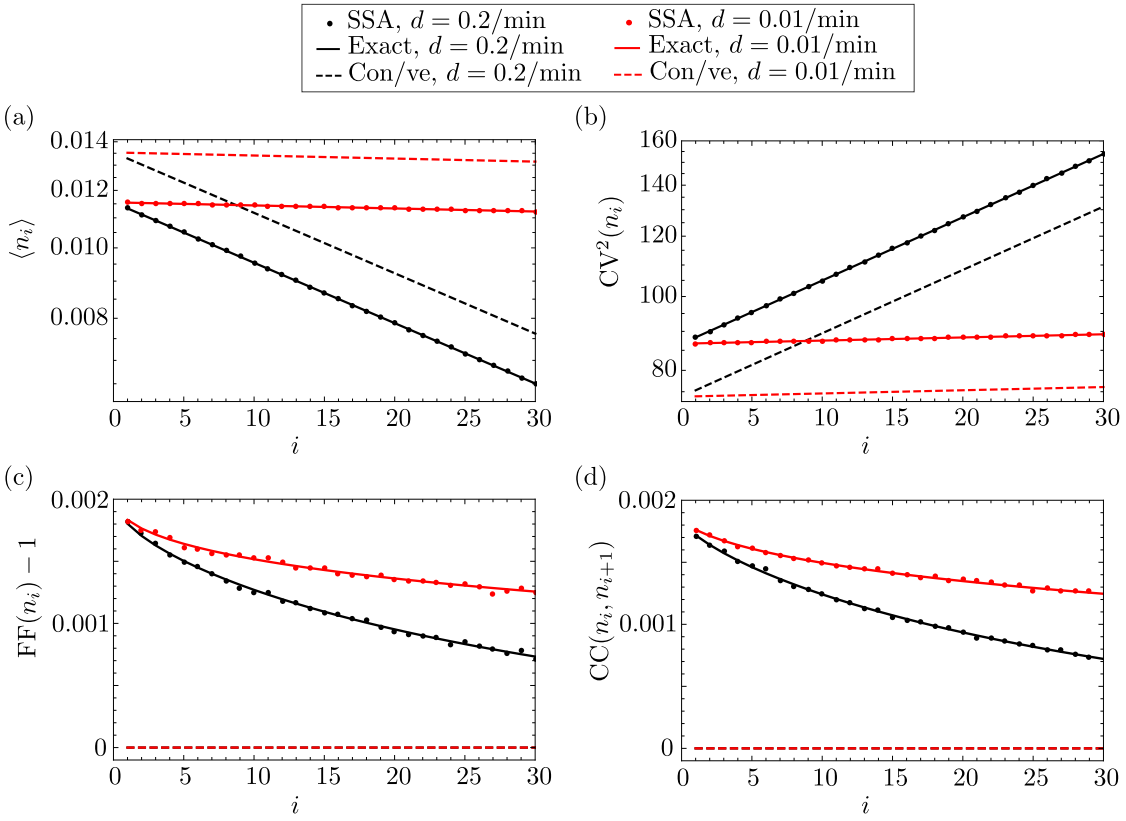


Figure 3.3: First and second moments of the distribution of local RNAP for the DOA1 gene in yeast, which demonstrates constitutive expression. In panels (a), (b), (c), and (d), we show the dependence of the mean, coefficient of variation squared, Fano factor, and Pearson correlation coefficient, respectively, of local RNAP fluctuations on gene segment i, as predicted by our exact theory (Eqs. (3.2) and (3.4); solid lines), the approximate theory in the constitutive limit (Eq. (3.9); dashed lines), and simulation via Gillespie's stochastic simulation algorithm (SSA; dots), respectively. The parameters are fixed to $s_u = 0.7$/min, $s_b = 0.12$/min and $r = 0.14$/min, which are characteristic of the DOA1 gene in yeast, as reported in Supplemental Table 2 of [48]. The number of gene segments is arbitrarily chosen to be $L = 30$. The total elongation time $\langle T \rangle = 2.9$ min is also reported for DOA1, described as the synthesis time and denoted by τ in [48]. The elongation rate by definition takes the value of the ratio $k = L/\langle T \rangle - d \approx L/\langle T \rangle$, since $d \ll k$. The detachment rate d is arbitrarily chosen to be $d = 0.01$/min (red lines and dots) or $d = 0.2$/min (black lines and dots). Note that, for the SSA, moments are calculated from one long trajectory with a few billion time points, sampled at unit intervals.

Moments of total RNAP distribution

We define the non-dimensional parameters $\delta_g = \tau_g/\tau_d$, $T_g = \langle T \rangle/\tau_g$, and $T_d = \langle T \rangle/\tau_d$, which correspond to the ratio of the gene timescale and the polymerase detachment timescale, the ratio of the mean elongation time and the gene timescale, and the ratio of the mean elongation time and the polymerase detachment timescale, respectively; here, $\tau_d = 1/d$, as before. Substituting $k \mapsto L/\langle T \rangle - d$ into Eq. (3.10) and taking the limit of deterministic elongation, i.e. letting $L \to \infty$ at constant $\langle T \rangle$, we obtain the following expressions for the mean, variance, and CV^2 of total

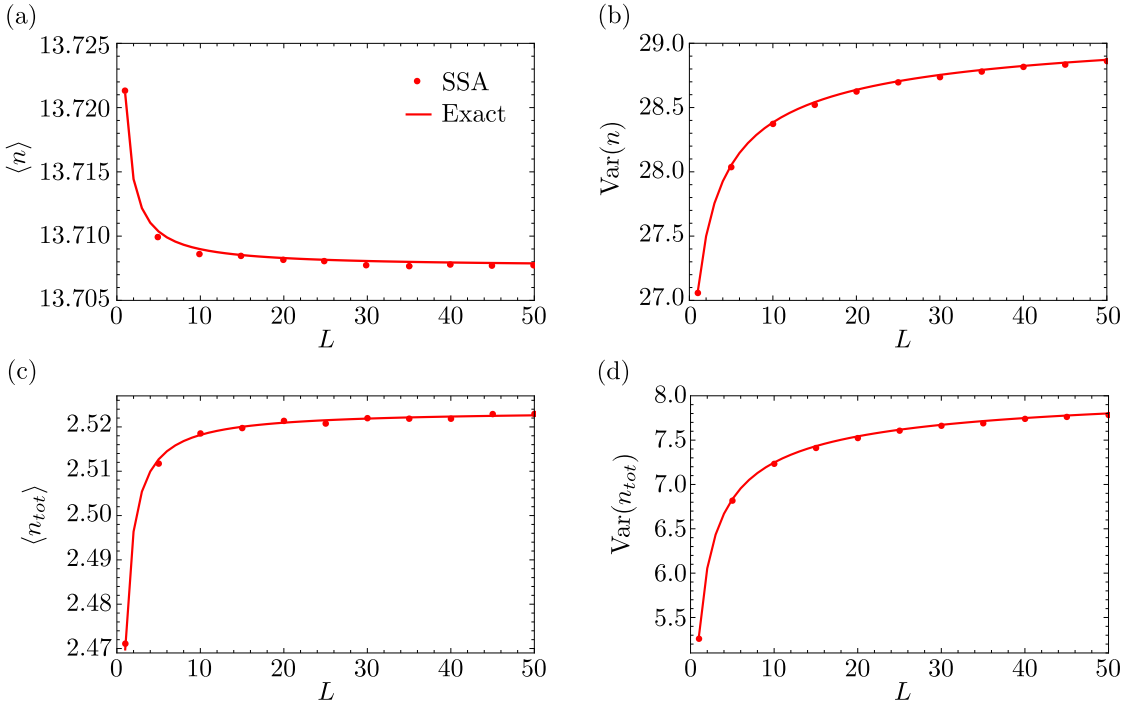


Figure 3.4: Mean and variance of the distributions of mature RNA and total RNAP for the PDR5 gene in yeast. In panels (a) and (b), we show the dependence of the moments of mature RNA fluctuations on the number of gene segments L, as predicted by our theory (Eqs. (3.2) and (3.4); solid lines) and SSA (dots). In panels (c) and (d), we show the dependence of the moments of total RNAP on L, as predicted by our exact theory (Eq. (3.10); solid lines) and SSA (dots). The parameters s_u, s_b, r, and $\langle T \rangle$ are characteristic of the PDR5 gene, and are the same as in Fig. 3.2. The premature detachment rate is chosen to be $d = 0.01/\text{min}$; the elongation rate is then given by $k \approx L/\langle T \rangle$. The degradation rate of mature RNA is $d_m = 0.04/\text{min}$, which is chosen such that the mean mature RNA is roughly consistent with that reported in Fig. 6(b) of [48]. Note that, for the SSA, moments are calculated from one long trajectory with a few billion time points, sampled at unit intervals.

RNAP:

$$\begin{aligned}
\langle n_{tot} \rangle_\infty &= \eta \frac{r}{d}(1 - e^{-T_d}), \\
\text{Var}(n_{tot})_\infty &= \langle n_{tot} \rangle_\infty + \langle n_{tot} \rangle_\infty^2 \cdot \beta \delta_g \frac{(\delta_g - 1) + (\delta_g + 1)e^{-2T_d} - 2\delta_g e^{-T_g} e^{-T_d}}{(\delta_g - 1)(\delta_g + 1)(1 - e^{-T_d})^2}, \\
\text{CV}^2(n_{tot})_\infty &= \langle n_{tot} \rangle_\infty^{-1} + \beta \delta_g \frac{(\delta_g - 1) + (\delta_g + 1)e^{-2T_d} - 2\delta_g e^{-T_g} e^{-T_d}}{(\delta_g - 1)(\delta_g + 1)(1 - e^{-T_d})^2}.
\end{aligned} \tag{3.11}$$

Here, the subscript ∞ denotes the limit of $L \to \infty$. A detailed derivation of the variance in Eq. (3.11) can be found in Lemma A.3.1 of Appendix A.3.

In the special case when RNAP does not prematurely detach from the gene, i.e. for $d = 0$, the expressions in Eq. (3.11) simplify to

$$\begin{aligned}
\langle n_{tot} \rangle_{(\infty;0)} &= \eta r \langle T \rangle, \\
\text{Var}(n_{tot})_{(\infty;0)} &= \langle n_{tot} \rangle_{(\infty;0)} + \langle n_{tot} \rangle_{(\infty;0)}^2 \cdot 2\beta T_g^{-1}\big(1 - T_g^{-1} + T_g^{-1} e^{-T_g}\big), \\
\text{CV}^2_{(\infty;0)} &= \langle n_{tot} \rangle_{(\infty;0)}^{-1} + 2\beta T_g^{-1}\big(1 - T_g^{-1} + T_g^{-1} e^{-T_g}\big),
\end{aligned} \tag{3.12}$$

where the subscript $(\infty;0)$ denotes the limit of $(L,d) \to (\infty,0)$. The expressions in Eq. (3.12) have been previously reported in [70], where they were derived using queuing theory. Hence, our expressions in Eq. (3.11) constitute a generalisation of known results, by further taking into account premature detachment of RNAP from the gene.

Eq. (3.12) shows that the coefficient of variation squared of total RNAP, denoted by $\mathrm{CV}^2_{(\infty;0)}$, can be written as the sum of two terms: (i) the inverse of the mean which is expected if the distribution of total RNAP is Poissonian, and (ii) a term that increases with increasing β and decreasing T_g. Hence, the latter term provides a measure for the deviation of the total RNAP distribution from a Poissonian. In particular, it shows that the deviation is significant in genes for which (i) the fraction of time spent in the inactive state is large (large β), and (ii) the elongation time is much shorter than the switching time between the active and inactive states (small T_g).

Moments of mature RNA distribution

Similarly, in the limit of deterministic elongation, it is straightforward to show that the expressions for the mean and variance of the distribution of mature RNA given by Eqs. (3.2) and (3.4) reduce to

$$\langle n \rangle_\infty = \eta\rho e^{-T_d} \qquad \text{and} \quad \mathrm{Var}(n)_\infty = \langle n \rangle_\infty + \langle n \rangle^2_\infty \cdot \beta\theta. \tag{3.13}$$

These expressions can be further simplified in the special case of no premature detachment to read

$$\langle n \rangle_{(\infty;0)} = \eta\rho \quad \text{and} \quad \mathrm{Var}(n)_{(\infty;0)} = \langle n \rangle_{(\infty;0)} + \langle n \rangle^2_{(\infty;0)} \cdot \beta\theta. \tag{3.14}$$

Note that the mean and variance are precisely the same as would be obtained from the telegraph model, for which the corresponding Fano factor in the bursty limit is given by Eq. (3.16) below. Hence, we anticipate that, in the limit of no premature detachment and deterministic elongation, the distribution of mature RNA from our transcription model is the same as the distribution obtained from the coarser telegraph model. A formal proof of that claim will be given in Section 3.3.

Relationship between Fano factors of total RNAP and mature RNA

Specifying to the case of no premature detachment, it is interesting to note that in the bursty limit, i.e. for $r, s_b \to \infty$ at constant mean burst size $b = r/s_b$ in Eq. (3.12), the Fano factor of total RNAP is given by

$$\mathrm{FF}_{\mathrm{n}(b;\infty;0)} = 1 + 2b; \tag{3.15}$$

see also Eq. (A.4.3) in Appendix A.4. Here, the subscript n denotes nascent RNA (total RNAP). Eq. (3.15) is in contrast to the Fano factor of mature RNA in the same bursty limit:

$$\mathrm{FF}_{\mathrm{m}(b;\infty;0)} = 1 + b, \tag{3.16}$$

see Eq. (A.4.8) in Appendix A.4, where the subscript m denotes mature RNA. (Note that $\mathrm{FF}_{\mathrm{m}(b;\infty;0)}$ also equals the Fano factor of the telegraph model in the same bursty limit [31].) Hence, by comparing Eqs. (3.15) and (3.16), we can deduce the following for bursty expression: (i) if the telegraph model is used to estimate the mean transcriptional burst size from total RNAP data where the elongation time is deterministic, then the mean burst size will be overestimated by a factor of two – in other words, the implicit assumption that the elongation time is exponentially distributed is inadequate; (ii) fluctuations in total RNAP (nascent RNA) deviate more from Poisson statistics, for which the Fano factor equals one, than fluctuations in mature RNA.

More generally, if we do not enforce the bursty limit, then we find the following relationship between the Fano factors of total RNAP and mature RNA, which are calculated from Eqs. (3.12) and (3.14), respectively:

$$\frac{\mathrm{FF}_{\mathrm{n}(\infty;0)}}{\mathrm{FF}_{\mathrm{m}(\infty;0)}} = 1 + \frac{e^{-T_g} T_r T_{s_b} \Xi}{T_g^2 \left[T_r T_{s_b} + T_g (T_g + T_m) \right]}. \tag{3.17}$$

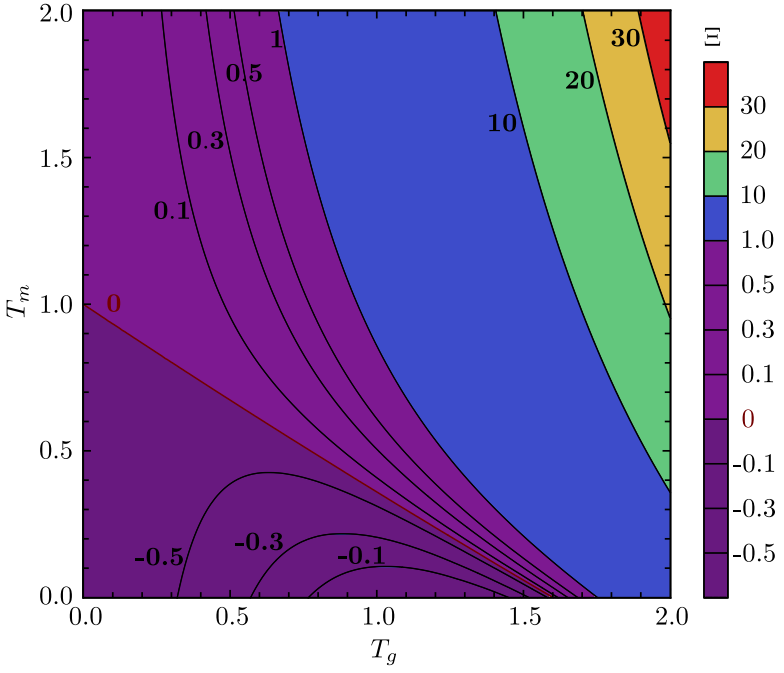


Figure 3.5: Comparison between the Fano factors of nascent and mature RNA. Contour plot showing the variation of Ξ – a measure of the difference between the two Fano factors which is defined in Eq. (3.18) – with the non-dimensional parameters T_g and T_m which denote the ratio of the mean elongation to that of the timescales of promoter switching and of mature RNA decay, respectively. As can be appreciated from Eq. (3.17), Ξ is positive if the Fano factor of nascent RNA is larger than that of mature RNA and negative if the reverse is true. The line $T_m \approx 1 - \frac{5}{8}T_g$, where $\Xi = 0$, shows where the two Fano factors are identical.

Here,

$$\Xi = 2(T_g + T_m) + e^{T_g}[2(T_g - 1)T_m + (T_g - 2)T_g], \tag{3.18}$$

while $T_g = (s_u + s_b)\langle T\rangle$, $T_r = r\langle T\rangle$, $T_m = d_m\langle T\rangle$, and $T_{s_b} = s_b\langle T\rangle$ are non-dimensional parameters representing the ratio of the mean elongation time to the timescales of promoter switching, initiation, decay of mature RNA, and gene deactivation, respectively. From Eq. (3.17), we deduce that $\mathrm{FF}_{\mathrm{n}(\infty;0)} > \mathrm{FF}_{\mathrm{m}(\infty;0)}$ if and only if $\Xi > 0$. From the Eq. (3.18), one can easily show that,

$$\Xi \geq 0 \quad \text{if and only if} \quad T_m \geq 1 - T_g \frac{2(1+T_g) + e^{T_g}(T_g^2 - 2)}{2 + 2e^{T_g}(T_g - 1)} = 1 - T_g f(T_g). \tag{3.19}$$

It is clear from the contour plot of Ξ in Fig. 3.5 that the curve $\Xi = 0$ can be well approximated by a line for $0 \leq T_g \leq 2$. This means that for the same interval of T_g, the function $f(T_g)$ in Eq. (3.19) can be well appreciated by a constant as $f(T_g) \approx [f(T_g = 0) + f(T_g = 2)]/2 \approx 0.64 \approx 5/8$.

Hence, the Fano factor of nascent RNA is larger than that of mature RNA if and only if the above (approximate) condition is satisfied. In the bursty limit, $T_g \to \infty$ due to $s_b \to \infty$ which, together with $T_m > 0$, implies that Eq. (3.19) holds; the condition is also satisfied if promoter switching is very fast compared to elongation. By contrast, if $T_m < 1$ and $T_g < 1$, then it is possible to have the opposite scenario where the Fano factor of mature RNA is larger than that of nascent RNA, which occurs for example if promoter switching and mature RNA decay are very slow compared to elongation.

Sensitivity of coefficient of variation of total RNAP and mature RNA

Since we have found explicit expressions for the first two moments of the distributions of total RNAP and of mature RNA, we can now estimate the sensitivity of the noise in each of those to small perturbations in the transcriptional parameters. Specifically, we calculate the logarithmic sensitivity (LS), which is also known as the relativity sensitivity, of the coefficient of variation to a parameter s, which is defined as $\Lambda_s = (s/\mathrm{CV})(\partial \mathrm{CV}/\partial s)$. (That definition implies that a 1% change in the value of the parameter s results in a change of Λ_s% in CV.)

In Table 3.1, we report the logarithmic sensitivity of the coefficient of variation (CV) of total RNAP fluctuations, which is obtained from Eq. (3.12), to perturbations in the parameters s_u, s_b, r, and $\langle T \rangle$. Similarly, in Table 3.1, we report the logarithmic sensitivity of the coefficient of variation of mature RNA fluctuations from Eq. (3.14) to perturbations in the parameters s_u, s_b, r, and d_m. In both cases, these sensitivities are calculated for parameter values estimated for five genes in yeast, as reported in [48]; see Table 3.1.

The following observations can be made regarding the sensitivity of the noise in total RNAP fluctuations: (i) for the two genes PDR5 and POL1 which spend most of their time in the inactive state due to $s_b \gg s_u$, CV is most sensitive to changes in the parameters s_u and $\langle T \rangle$; (ii) for the genes DOA1, MDN1, and KAP104 which spend most of their time in the active state due to $s_u \gg s_b$, CV is most sensitive to changes in the parameters r and $\langle T \rangle$; (iii) the size of mature RNA fluctuations is found to be most sensitive to perturbations in s_u and d_m for PDR5 and POL1, and to perturbations in r and d_m for the other three genes. We furthermore note that for both total RNAP and mature RNA, r is the least sensitive parameter for the genes which are mostly inactive, whereas it is among the most sensitive parameters for genes that are mostly active.

Table 3.1: Logarithmic sensitivity (LS) of the coefficient of variation CV of total RNAP and mature RNA fluctuations for five genes in yeast; see Section 3.2.4 for a discussion. (a) Parameter values from Supplemental Tables 2 and 4 in [48]. The degradation rate d_m of mature mRNA is estimated from the reported mean number of mature RNA, the parameters s_u, s_b, r, and Eq. (3.14) for the mean. (b) Logarithmic sensitivity of CV of total RNAP fluctuations. (c) Logarithmic sensitivity of CV of mature mRNA fluctuations. The most sensitive parameter and the next most sensitive one are marked in dark bold and italic, respectively.

	PDR5	POL1	DOA1	MDN1	KAP104
(a)					
Mean mature RNA #	13.40	3.13	2.59	6.12	4.93
$\langle T \rangle$(min)	4.50	3.75	2.90	16.75	3.50
$s_u(\mathrm{min}^{-1})$	0.44	0.07	0.70	0.70	0.70
$s_b(\mathrm{min}^{-1})$	4.70	0.68	0.12	0.12	0.12
$r(\mathrm{min}^{-1})$	6.70	2.00	0.14	0.19	0.27
$d_m(\mathrm{min}^{-1})$	0.04	0.06	0.05	0.03	0.05
LS	PDR5	POL1	DOA1	MDN1	KAP104
(b)					
Λ_{s_u}	**-0.52**	**-0.51**	-0.09	-0.12	-0.11
Λ_{s_b}	0.18	0.29	0.09	0.09	0.10
Λ_r	-0.15	-0.12	**-0.49**	*-0.47*	*-0.47*
$\Lambda_{\langle T \rangle}$	*-0.48*	*-0.34*	**-0.49**	**-0.50**	**-0.49**
LS	PDR5	POL1	DOA1	MDN1	KAP104
(c)					
Λ_{s_u}	**-0.50**	**-0.52**	-0.09	-0.10	-0.11
Λ_{s_b}	0.23	0.20	0.08	0.08	0.09
Λ_r	-0.23	-0.15	*-0.49*	*-0.48*	*-0.48*
Λ_{d_m}	**0.50**	*0.47*	**0.50**	**0.50**	**0.50**

3.3 Approximate distributions of total RNAP and mature RNA

Thus far, we have derived expressions for the first two moments of the distributions of total RNAP and mature RNA. Naturally, it would also be useful to derive closed-form expressions for the distributions themselves; such a derivation is, however, analytically intractable in general [133] due to the presence of the catalytic reaction $G_{on} \to G_{on} + P_1$, which models initiation of the transcription process. Still, there are two special cases where analytical distributions are known: (i) when the elongation time is considered to be fixed, which corresponds to our model with $L \to \infty$ at constant $\langle T \rangle$ [72]; (ii) when the elongation time is exponentially distributed, corresponding to our model with $L = 1$, in which case the distribution of total RNAP is identical to the one which is derived from the telegraph model [31, 61]. While one may argue that the analytical distribution of RNAP for deterministic elongation times may well approximate the stochastic (finite-L) case, the issue remains that the exact solution is not given in terms of simple functions unless promoter switching is slow compared to initiation, elongation, and termination, in which case the solution reduces to a weighted sum of two Poisson distributions [72]. Hence, it is generally very difficult to apply in practice, such as to infer parameters from data using a Bayesian approach. Moreover, to our knowledge, no exact solutions are known for the distribution of mature RNA in our model. In this section, we aim to devise a simple approximation for the distribution of total RNAP numbers in terms of the Negative Binomial (NB) distribution; these simple distributions have shown great flexibility in describing complex gene expression models with a large number of parameters [126]. Finally, by means of singular perturbation theory, we will obtain the distribution of mature RNA under the assumption that RNA polymerase elongation is faster than degradation of mature RNA.

3.3.1 Approximation of total RNAP distribution

We approximate the distribution of total RNAP transcribing the gene via a Negative Binomial distribution, as follows. The mean and variance of the Negative Binomial distribution $\mathrm{NB}(q, p)$ are given by $pq/(1-p)$ and $pq/(1-p)^2$, respectively. By assuming that these are equal to the exact mean and variance, respectively, of the total RNAP distribution, see Eq. (3.10), we obtain effective values for the parameters p and q:

$$n_{tot} \sim \mathrm{NB}(q, p) \equiv \mathrm{NB}\left(\frac{\langle n_{tot} \rangle^2}{\mathrm{Var}(n_{tot}) - \langle n_{tot} \rangle}, \frac{\mathrm{Var}(n_{tot}) - \langle n_{tot} \rangle}{\mathrm{Var}(n_{tot})} \right). \tag{3.20}$$

In Fig. 3.6, we show a comparison between the distributions of total RNAP obtained from SSA (dots) and the Negative Binomial approximation in Eq. (3.20 (solid lines). Our results are presented for two different values of the number of gene segments: $L = 1$ (exponentially distributed elongation time; left column) and $L = 50$ (quasi-deterministic elongation time; right column). Additionally, we rescale our gene inactivation rate as $s_b \mapsto s_b \epsilon$, and we present results for three different values of the parameter ϵ: 10^{-3}, the constitutive limit of the gene being mostly in the active state (top row); 10^{-1}, where the gene spends almost equal amounts of time in the active and inactive states, with $s_b \approx s_u$ (middle row); and 1, the bursty limit, where the gene spends most of its time in the inactive state (bottom row).

We can make several observations, as follows. For both $L = 1$ and $L = 50$, the Negative Binomial approximation performs well for bursting and constitutive expression (top and bottom rows), whereas it is appreciably poor when expression is in between those two limits (middle row). Intuitively, this observation can be explained via the following reasoning. In the limits of the gene being mostly in the active state (constitutive expression) or the inactive state (bursty expression), the distribution of total RNAP is necessarily unimodal. However, when the gene spends a considerable amount of time in each state, the distribution is the sum of two conditional distributions which can manifest either as bimodality or as a wide unimodal distribution, neither of which can be captured by a Negative Binomial distribution. Assuming bursty expression, the Negative Binomial distribution is a more accurate approximation to the distribution obtained from SSA for $L = 1$ than it is for $L = 50$; the reason is that $L = 1$ corresponds to the telegraph

model [31], in which case it can be proven analytically that the distribution reduces to a Negative Binomial in the limit of bursty expression. For constitutive expression, the Negative Binomial approximation is equally good for $L = 1$ and $L = 50$, as the distribution is necessarily Poissonian then and as it is well known that a Negative Binomial distribution can approximate a Poissonian to a high degree of accuracy. In summary, our results hence indicate that Eq. (3.20) yields a good approximation for the total RNAP distribution of bursty and constitutively expressed genes.

We also note from Fig. 3.6 that the comparison between the SSA distributions for $L = 1$ and $L = 50$, with equal mean elongation times, highlights the importance of modelling elongation with the correct distribution of elongation times for genes that are non-constitutive, i.e. for $\epsilon = 10^{-1}$ or $\epsilon = 1$. In particular, if the elongation time is quasi-deterministic ($L = 50$), there appears to be a significant increase in the probability of observing zero total RNAP transcribing the gene compared to models with an exponentially distributed elongation time ($L = 1$).

3.3.2 Approximation of mature RNA distribution

Next, we apply singular perturbation theory to formally derive the distribution of mature RNA when the elongation rate is much larger than the degradation rate of mature RNA.

We start by defining $P_j(\vec{n};t)$ $(j = 0, 1)$ as the probability of the state $\vec{n} = (n_1, \ldots, n_L, n)$ at time t while the gene is either active (0) or inactive (1). Note that n_i is the number of RNAPs on gene segment i for $i = 1, \ldots, L$, while n is the number of mature RNAs. The time evolution of the probabilities $P_j(\vec{n};t)$ can be described by a system of coupled CMEs:

$$\begin{aligned}
\partial_t P_0 &= s_u P_1 - s_b P_0 + r(\mathbb{E}_{n_1}^{-1} - 1)P_0 + k\sum_{i=1}^{L-1}\big(\mathbb{E}_{n_i}\mathbb{E}_{n_{i+1}}^{-1} - 1\big)n_i P_0 + k\big(\mathbb{E}_{n_L}\mathbb{E}_{n}^{-1} - 1\big)n_L P_0 \\
&\quad + d\sum_{i=1}^{L}(\mathbb{E}_{n_i} - 1)n_i P_0 + d_m(\mathbb{E}_n - 1)nP_0, \\
\partial_t P_1 &= s_b P_0 - s_u P_1 + k\sum_{i=1}^{L-1}\big(\mathbb{E}_{n_i}\mathbb{E}_{n_{i+1}}^{-1} - 1\big)n_i P_1 + k\big(\mathbb{E}_{n_L}\mathbb{E}_{n}^{-1} - 1\big)n_L P_1 \\
&\quad + d\sum_{i=1}^{L}(\mathbb{E}_{n_i} - 1)n_i P_1 + d_m(\mathbb{E}_n - 1)nP_1,
\end{aligned} \tag{3.21}$$

where $\mathbb{E}_{n_i}^{c}[f(\vec{n})] = f(n_1, n_2, \ldots, n_i + c, \ldots, n_L, n)$, with $c \in \mathbb{Z}$, denotes the standard step operator. We assume that the elongation rate k is faster than the degradation rate d_m of mature RNA, i.e. that $k/d_m \gg 1$. Since $k = L/\langle T\rangle - d$, it follows that in the limit of deterministic elongation ($k \to \infty$), i.e. for $L \to \infty$ at constant mean elongation time $\langle T\rangle$, the condition $k/d_m \gg 1$ is naturally satisfied.

In order to find an analytical expression for the propagator probabilities $P(\vec{n};t)$ which satisfies the system of CMEs in Eq. (3.21), we define the probability-generating function as $F = \sum_j F_j$, with $F_j(\vec{z};t) = \sum_{\vec{n}=\vec{0}}^{\infty} P_j(\vec{n};t)\vec{z}^{\vec{n}}$; here, $\vec{z} = (z_1, \ldots, z_L, z)$ is a vector of variables corresponding to the state $\vec{n}$. Given the equations for $P_j(\vec{n};t)$ from Eq. (3.21), we obtain the following systems of PDEs for the corresponding generating functions $F_j(\vec{z};t)$:

$$\begin{aligned}
\mathbb{L}[F_0] &= s_u F_1 - s_b F_0 + r(z_1 - 1)F_0, \\
\mathbb{L}[F_1] &= s_b F_0 - s_u F_1,
\end{aligned} \tag{3.22}$$

where

$$\mathbb{L} = \partial_t + k\sum_{i=1}^{L-1}(z_i - z_{i+1})\partial_{z_i} + k(z_L - z)\partial_{z_L} + d\sum_{i=1}^{L}(z_i - 1)\partial_{z_i} + d_m(z - 1)\partial_z \tag{3.23}$$

is a differential operator acting on the generating functions F_0 and F_1. Eq. (3.22) represents a system of coupled, linear, first-order partial differential equations (PDEs). Now, we introduce the

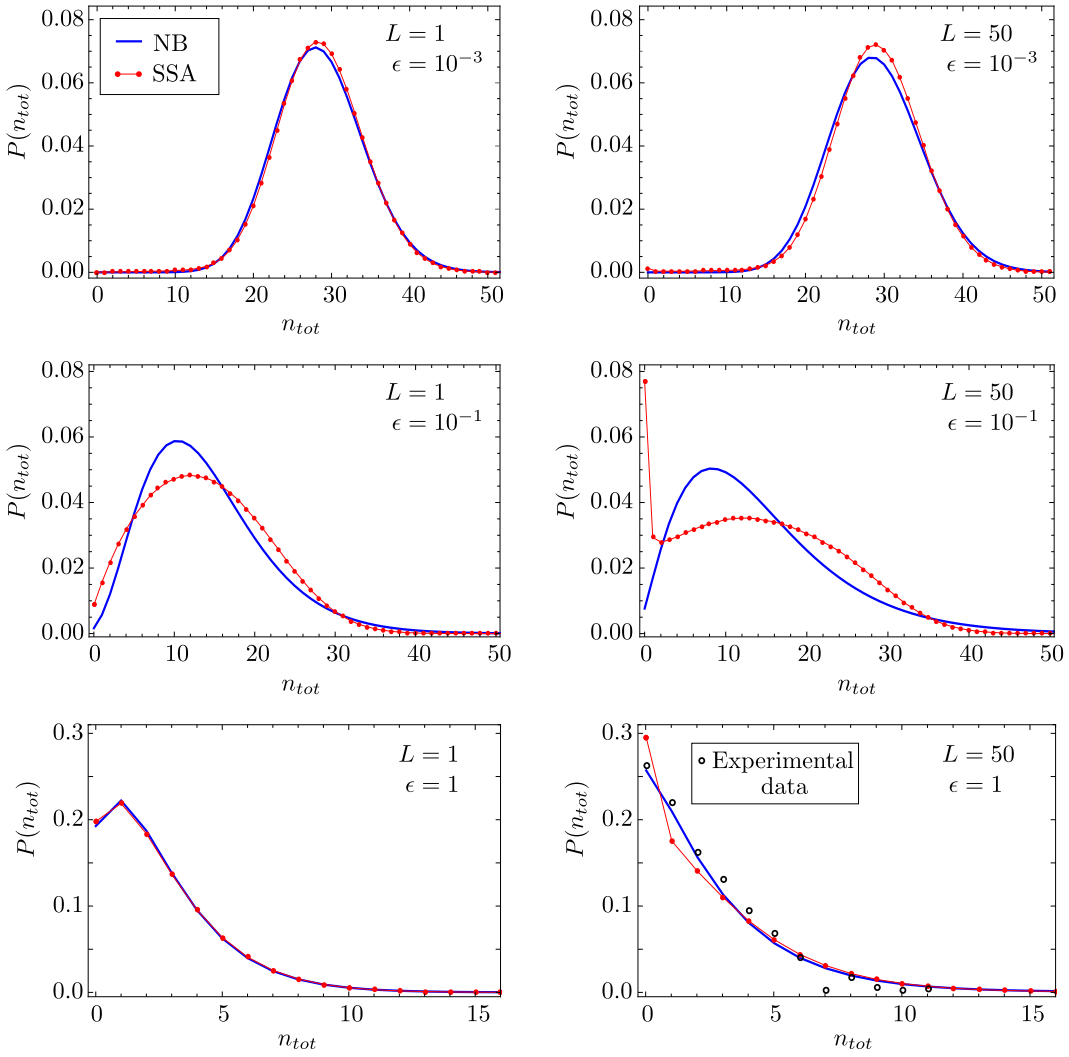


Figure 3.6: Steady-state distribution of total RNAP and its approximation by a Negative Binomial distribution. We compare the approximation from Eq. (3.20) (blue lines) with the distribution of total RNAP obtained from stochastic simulation (SSA; red dots). With the exception of s_b, the parameters are for the PDR5 gene in yeast, and are hence the same as in Fig. 3.2, with $d = 0.01/\text{min}$. Results are presented for two different values of L, corresponding to an exponentially distributed elongation time ($L = 1$) and a quasi-deterministic elongation time ($L = 50$); k is rescaled such that the two have the same mean elongation time. Additionally, we rescale the gene inactivation rate via $s_b \mapsto s_b\epsilon$, where $\epsilon = 10^{-3}, 10^{-1}, 1$, corresponding to constitutive, general, and bursty expression, respectively. (Here, general expression is neither clearly constitutive nor bursty, since the gene spends roughly equal amounts of time in the inactive and active states.) Note that $\epsilon = 1$ results in a distribution of nascent RNA that is consistent with that measured for PDR5; the experimental data from Fig. 6(b) of [48] is plotted for comparison. The Negative Binomial approximation is found to be accurate in the limits of constitutive and bursty expression (top and bottom rows), independently of L.

new variables $u_i = z_i - 1$ $(i = 1, \ldots, L)$ and $u = z - 1$ to rewrite Eq. (3.22) as

$$\begin{aligned} \mathbb{L}[F_0] &= s_u F_1 - s_b F_0 + r u_1 F_0, \\ \mathbb{L}[F_1] &= s_b F_0 - s_u F_1; \end{aligned} \tag{3.24}$$

here, the operator in Eq. (3.23) now takes the form

$$\mathbb{L} = \partial_t + k\sum_{i=1}^{L-1}(u_i - u_{i+1})\partial_{u_i} + k(u_L - u)\partial_{u_L} + d\sum_{i=1}^{L} u_i\partial_{u_i} + d_m u\partial_u. \tag{3.25}$$

In order to find an analytical solution to Eq. (3.24), we rescale all rates and the time variable by the decay rate of mature RNA; then, we apply the method of characteristics, with s being the characteristic variable. The first characteristic equation gives $d_m(dt/ds) = 1$, with solution $s \equiv t' = d_m t$; hence, we can use the variable t' as the independent variable and thus convert the system of PDEs in Eq. (3.24) into a characteristic system of ordinary differential equations (ODEs),

$$\dot{u}_i = (k/d_m)[u_i - u_{i+1} + (d/k)u_i] \qquad \text{for } i = 1, \ldots, L-1, \tag{3.26a}$$
$$\dot{u}_L = (k/d_m)[u_L - u + (d/k)u_L], \tag{3.26b}$$
$$\dot{u} = u, \tag{3.26c}$$
$$\dot{F}_0 = (s_u/d_m)F_1 - (s_b/d_m)F_0 + (r/d_m)u_1 F_0, \tag{3.26d}$$
$$\dot{F}_1 = (s_b/d_m)F_0 - (s_u/d_m)F_1, \tag{3.26e}$$

where the overdot denotes differentiation with respect to t'. The existence of an integral-form solution to Eq. (3.26) follows from the fact that the reaction scheme in Fig. 3.1 contains first-order reactions only. Under the assumption that $k \gg d_m$, we define $\varepsilon = d_m/k$; then, we apply Geometric Singular Perturbation Theory (GSPT) [93, 107], with $0 < \varepsilon \ll 1$ as the (small) singular perturbation parameter. We hence separate the system in Eq. (3.26) into fast and slow dynamics, which will allow us to find an asymptotic approximation for F_0 and F_1 in steady-state. Given the above definition of ε, Eqs. (3.26a) and (3.26b), the governing equations for u_i in the 'slow system', become

$$\begin{aligned} \varepsilon\dot{u}_i &= u_i - u_{i+1} + (d/k)u_i \qquad \text{for } i = 1, \ldots, L-1, \\ \varepsilon\dot{u}_L &= u_L - u + (d/k)u_L, \end{aligned} \tag{3.27}$$

where u_i $(i, \ldots, L)$ are the fast variables and u, F_0, and F_1 are the slow ones. Setting $\varepsilon = 0$ in Eq. (3.27), we can express the variables u_i as $u_i = \mu \cdot u_{i+1}$, with $\mu = k/(k+d)$ for $i = 1, \ldots, L$. Finally, we write the variable u_1 as $u_1 = \mu^L \cdot u$. Next, given Eq. (3.26c), we apply the chain rule, with $dt' \equiv du \cdot u$, to rewrite Eqs. (3.26d) and (3.26e) as

$$F_0' d_m u = s_u F_1 - s_b F_0 + r\mu^L u F_0, \tag{3.28a}$$
$$F_1' d_m u = s_b F_0 - s_u F_1, \tag{3.28b}$$

where the prime now denotes differentiation with respect to u. Solving Eq. (3.28a) for F_1 and substituting the result into Eq. (3.28b), we obtain the second-order ODE

$$d_m^2 u F_0'' + d_m(d_m + s_b + s_u - r\mu^L u)F_0' - r\mu^L(d_m + s_u)F_0 = 0 \tag{3.29}$$

for $F_0(u)$. Eq. (3.29) is a confluent hypergeometric differential equation (Kummer's equation) [44] which admits the solution

$$F_0(u) = C \cdot {}_1F_1\Big(\frac{d_m + s_u}{d_m}; \frac{d_m + s_b + s_u}{d_m}; \frac{r}{d_m}\mu^L u\Big), \tag{3.30}$$

where ${}_1F_1$ denotes the confluent hypergeometric function; here, we consider only one of two independent fundamental solutions of Kummer's differential equation, as we are seeking a solution in steady-state where the variable u is bounded. The constant C in Eq. (3.30) is a constant of integration that is determined from the normalisation condition on the full generating function: $F = F_0 + F_1$. From Eq. (3.28), one finds that F satisfies

$$F' = \frac{r}{d_m}\mu^L F_0. \tag{3.31}$$

Making use of Eq. (3.31) and applying the normalisation condition $F|_{u=0} = 1$, we find that the generating function in steady-state reads

$$F(z) = {}_1F_1\Big(\frac{s_u}{d_m}; \frac{s_b+s_u}{d_m}; \frac{r}{d_m}\mu^L(z-1)\Big). \tag{3.32}$$

The probability distribution $P(n)$ of mature RNA can be found from the formula

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

which yields the analytical expression

$$P(n) = \frac{1}{n!}\frac{(s_u)_n}{(s_b+s_u)_n}\Big(\frac{r}{d_m}\Big)^n(\mu^L)^n\,{}_1F_1\big(\tfrac{s_u}{d_m}+n; \tfrac{s_b+s_u}{d_m}+n; -\tfrac{r}{d_m}\mu^L\big), \tag{3.33}$$

where $(\cdot)_n$ is the Pochhammer symbol, as before. Note that the mean and variance of mature mRNA, as calculated from the distribution in Eq. (3.33), agree exactly with Eqs. (3.2c) and (3.4f) in the limit of fast elongation rate ($k \to \infty$). Note also that the solution in Eq. (3.33) depends on the parameter μ^L, which represents the survival probability of an RNAP molecule, i.e. the probability that RNAP will not prematurely detach from the gene. Finally, we take the limit of deterministic elongation, letting $L \to \infty$ at constant $\langle T\rangle$, which leads to

$$P(n) = \frac{1}{n!}\frac{(s_u)_n}{(s_b+s_u)_n}\Big(\frac{r}{d_m}\Big)^n e^{-nd\langle T\rangle}\,{}_1F_1\big(\tfrac{s_u}{d_m}+n; \tfrac{s_b+s_u}{d_m}+n; -\tfrac{r}{d_m}e^{-d\langle T\rangle}\big). \tag{3.34}$$

Note that in the limit of no premature detachment ($d = 0$), Eq. (3.34) is precisely equal to the distribution of mature RNA predicted by the telegraph model, which is in wide use in the literature [31]. Hence, our perturbative approach can be seen as a means to formally derive the conventional telegraph model of gene expression starting from a more fundamental and microscopic model. In Fig. 3.7, we verify our analytical solution with stochastic simulation for two different genes in yeast. We also note that, for non-zero premature detachment rates ($d \neq 0$), Eq. (3.34) is the steady-state solution predicted by the telegraph model, with parameter r renormalised to $re^{-d\langle T\rangle}$; that is to be expected, as the latter is the rate at which RNAPs undergo termination, leading to mature RNAs.

3.4 Statistics of fluorescent nascent RNA signal

Thus far, we have determined the statistics of the total number of RNAP transcribing the given gene; these are also the statistics of the number of nascent RNA molecules. However, in experiments using single-molecule fluorescence in situ hybridisation (smFISH [72]), molecule numbers of nascent RNA cannot be directly determined. Rather, the experimentally measured RNA 'abundance' is the fluorescent signal emitted by oligonucleotide probes bound to the RNA. Since the length of the nascent RNA grows as RNAP moves away from the promoter, it follows that we must account for the increase in the fluorescent signal as elongation proceeds.

In this section, we take into account these experimental details to obtain closed-form expressions for the mean and variance of the fluorescent signal of local and total nascent RNA. We assume that the signal from nascent RNA on the i-th gene segment is given by $r_i = (\nu/L)in_i$ for $i = 1, \ldots, L$, where ν is some experimental constant; the value of the parameter $(\nu/L)i$ is increasing with i, which models the fact that the fluorescent signal becomes stronger as RNAP moves along the gene. The formula for the mean fluorescent signal at the gene segment i is then given by $\langle r_i\rangle = (\nu/L)i\langle n_i\rangle$, where $\langle n_i\rangle$ follows from Eq. (3.2b); the covariance of two fluorescent signals along the gene, r_i and r_j $(i, j = 1, \ldots, L)$, is given by $\mathrm{Cov}(r_i, r_j) = (\nu/L)^2 ij\mathrm{Cov}(n_i, n_j)$, where $\mathrm{Cov}(n_i, n_j)$ is obtained from Eq. (3.4d). In Figs. 3.8(a) and (b), we plot the mean and Fano factor of the local signal as a function of the gene segment i; note the contrast between the statistics of the fluorescent signal and the corresponding statistics of local RNAP – which is the statistics of nascent RNA – shown in Figs. 3.2(a) and (c).

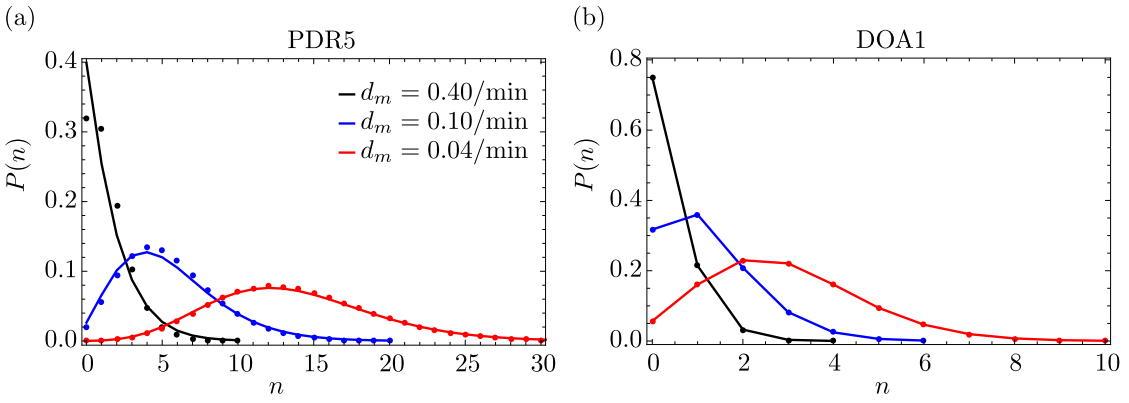


Figure 3.7: Steady-state distribution of mature RNA for two different genes in yeast. We compare the distribution obtained from SSA (dots) to the perturbative approximation in Eq. (3.33) (solid lines) for two different genes. In panel (a), we consider the PDR5 gene, fixing the parameters as in Fig. 3.2: $s_u = 0.44$/min, $s_b = 4.7$/min, $r = 6.7$/min, $d = 0.01$/min, and $\langle T \rangle = 4.5$ min. The degradation rate of mature RNA takes the values $d_m = 0.04, 0.10, 0.40$/min; note that the experimental value is $d_m = 0.04$/min. In panel (b), we consider the DOA1 gene, fixing the parameters as in Fig. 3.3: $s_u = 0.7$/min, $s_b = 0.12$/min, $r = 0.14$/min, $d = 0.01$/min, and $\langle T \rangle = 2.9$ min. The degradation rate of mature RNA again takes the values $d_m = 0.04, 0.10, 0.40$/min; the experimental value is $d_m = 0.05$/min. For both genes, the agreement between SSA and our perturbative approximation increases with k/d_m, as expected, since Eq. (3.33) is derived under the assumption that $k \gg d_m$. Note that the distribution is practically independent of L, since Eq. (3.33) depends on L only through μ^L, which for small premature detachment rates d implies $\mu^L \approx 1$ for any L.

Similarly, denoting by $r_{tot} = \sum_{i=1}^{L} r_i$ the total fluorescent signal across the gene, we find the following expressions for the steady-state mean $\langle r_{tot} \rangle = \sum_{i=1}^{L} \langle r_i \rangle$ and the steady-state variance $\mathrm{Var}(r_{tot}) = \sum_{i,j=1}^{L} \mathrm{Cov}(r_i, r_j)$:

$$\begin{aligned} \langle r_{tot} \rangle &= \nu\eta\rho_k\mu \frac{\mu^L[L\mu - (L+1)] + 1}{L(\mu-1)^2}, \\ \mathrm{Var}(r_{tot}) &= \left(\frac{\nu}{L}\right)^2 \eta\rho_k \sum_{i=1}^{L} i^2\mu^i + \left(\frac{\nu}{L}\right)^2 \alpha\beta(\eta\rho_k)^2 \sum_{i,j=1}^{L} ij \cdot \mu^{i+j} \cdot f_{ij}. \end{aligned} \tag{3.35}$$

For a detailed derivation of the variance in Eq. (3.35), see Eq. (A.5.1) in Appendix A.5; see also Appendix A.6 for the corresponding expressions in the bursty, constitutive, and deterministic elongation limits. In Figs. 3.8(c) and (d), we show the mean and Fano factor of the total signal as a function of the number of gene segments (L); as above, note the contrasting difference between the statistics of the fluorescent signal and the corresponding statistics of total RNAP – which is the statistics of total nascent RNA – shown in Figs. 3.4(c) and (d).

Hence, the calculation of the statistics of the number of nascent RNAs from the raw signal intensity presents a challenge and has to be approached carefully. The expressions presented above allow for the inference of transcriptional parameters from the first two moments of the fluorescent signal by means of moment-based inference techniques [134]. Quantitative information about nascent RNA can also be obtained from electron micrograph images [37], which avoids the challenges presented by smFISH.

3.5 Model extension with pausing of RNAP

Thus far, we have studied a model where RNAPs do not pause as they move along the gene. A natural extension is provided by a modified model in which RNAPs pause along the gene at random

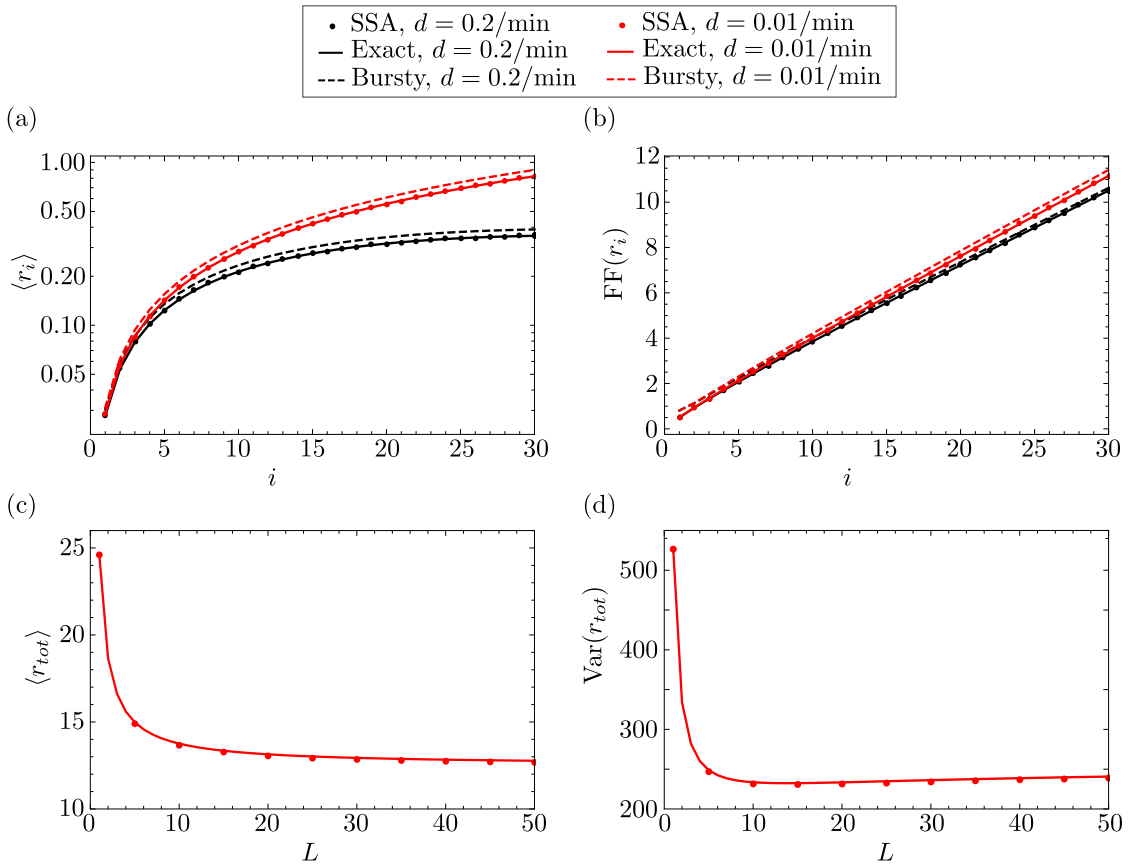


Figure 3.8: First and second moments of the local and total fluorescent signal for the bursty gene PDR5 in yeast. In panels (a) and (b), we show the dependence of the mean and the Fano factor of local fluorescent signal fluctuations on the gene segment i, as predicted by our exact theory (solid lines) and SSA (dots), respectively. The plots for $CV^2(r_i)$ and $CC(r_i, r_{i+1})$ are identical to those of $CV^2(n_i)$ and $CC(n_i, n_{i+1})$ in Fig. 3.2. The number of gene segments is arbitrarily chosen to be $L = 30$. In panels (c) and (d), we show the dependence of the mean and variance of total fluorescent signal fluctuations on the number of gene segments L, as predicted by our exact theory (Eq. (3.35); solid lines) and SSA (dots). The parameters s_u, s_b, r, and $\langle T \rangle$ are characteristic of the PDR5 gene and take the same values as in Fig. 3.2, as do the rates of elongation and RNAP detachment. The value of the parameter ν is arbitrarily chosen to be $\nu = 10$.

sites and elongation is characterised by three processes: forward hopping, pausing, and unpausing of RNAP. The motivation for studying this extended model, which has recently been considered via stochastic simulation in [132], is that experiments have revealed that RNAP exhibits pauses of varying duration, typically on the timescale of few seconds [135, 136].

3.5.1 Closed-form expressions for moments of local RNAP fluctuations

We extend the model described in Fig. 3.1 by assuming that the RNAP on the gene segment i can switch between a non-paused (actively moving) state P_i and a paused state $\bar{P}_i$. The actively moving state P_i switches to $\bar{P}_i$ with rate r_p, while the reverse reaction occurs with rate r_a. Premature detachment from the actively moving RNAP occurs with rate d_a, whereas it occurs with rate d_p from the paused RNAP. The resulting extended model is illustrated in Fig. 3.9(a). In Appendix A.1, we derive the mean and variance of the corresponding elongation time, which is not Erlang distributed now, as was the case for the model without pausing. Furthermore, we find two

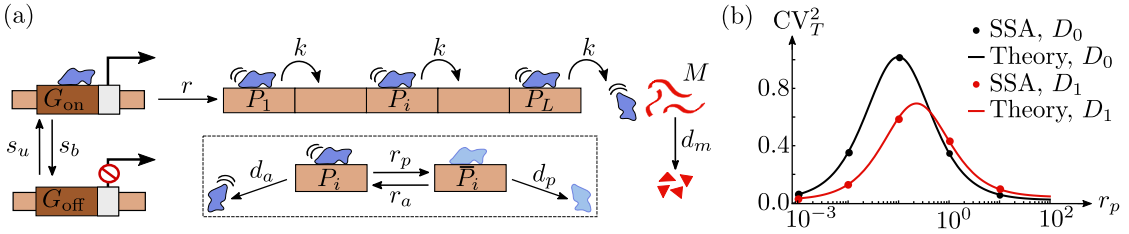


Figure 3.9: Model of transcription that includes RNAP pausing. In panel (a), we extend the model in Fig. 3.1 so that it takes into account pausing of RNAP at random segments on the gene. Pausing on gene segment i is modelled by the transition from the active state P_i to the paused state $\bar{P}_i$ with rate r_p, while the reverse ('unpausing') transition occurs with rate r_a. Premature termination of RNAP occurs with rate d_a from the actively moving state, and with rate d_p from the paused state. In panel (b), we show the dependence of the coefficient of variation squared (CV_T^2) of the elongation time distribution on the pausing rate (r_p), as predicted from SSA (dots) and theory (Eq. (A.1.7); solid lines). Results are shown for two different parameter regimes: $D_0 \equiv \{d_a = 0/\text{min} = d_p\}$ (no premature polymerase detachment) and $D_1 \equiv \{d_a = 0.05/\text{min} = d_p\}$ (premature polymerase detachment). The remaining parameters are fixed to $L = 50$, $k = 10/\text{min}$, and $r_a = 0.1/\text{min}$.

interesting properties of the coefficient of variation CV_T^2 of the elongation time: (i) in the limit of large L at constant mean elongation time, CV_T^2 does not tend to zero, which implies that elongation is not deterministic; (ii) for small rates of premature detachment, CV_T^2 is at its maximum when $r_p \approx r_a$, i.e. when RNAP spends roughly half of its time in the paused state. See Appendix A.1 for details and Fig. 3.9(b) for a confirmation through stochastic simulation.

Proposition 3. *Let the number of RNAP molecules in the active state P_i be denoted by n_i^a, let the number of molecules in the paused state $\bar{P}_i$ be n_i^p, and let the number of molecules of mature RNA be denoted by n. Let $\sigma = r_p/r_a$ be the ratio of the pausing and activation rates, let $\pi_{r_a} = r_a/(r_a + d_p)$ be the probability of RNAP switching to the actively moving state from the paused state, and let $\pi_{d_p} = d_p/(r_a + d_p)$ be the probability of premature RNAP detachment from the paused state. Furthermore, define the new parameters $\tilde{\mu} = k/(k + d_a + r_p\pi_{d_p})$ and $\lambda = \sigma\pi_{r_a}$.*

Then, it follows that the steady-state mean number of RNAP molecules in the active and paused states on gene segment i ($i = 1, \dots L$) is given by

$$\langle n_i^a \rangle = \eta\rho_k\tilde{\mu}^i \quad \textit{and} \quad \langle n_i^p \rangle = \langle n_i^a \rangle\lambda. \tag{3.36}$$

Hence, the total mean number of RNAP molecules on each gene segment i reads

$$\langle n_i \rangle = \langle n_i^a \rangle + \langle n_i^p \rangle = \langle n_i^a \rangle(1 + \lambda). \tag{3.37}$$

The proof of Prop. 3 can be found in Appendix A.7. Note that in the limit of no pausing, i.e. for $r_p = 0$, Eq. (3.37) reduces to the expression for the mean of RNAP reported in Eq. (3.2b).

Proposition 4. *Let $\tau_{r_a} = 1/r_a$ be the timescale of RNAP activation from the paused state, let $\tau_{d_p} = 1/d_p$ be the timescale of premature termination of paused RNAP, let $\tau_p = 1/(k + d_a)$ be the typical time that an actively moving RNAP spends on a gene segment, and let $\tau_{pp} = 1/(r_a + d_p)$ be the typical time spent in the paused state. Furthermore, define the new parameters $\lambda_{r_p} = \pi_{r_p}/(1 - \pi_{r_p})$, where $\pi_{r_p} = r_p/(r_p + k + d_a)$ is the probability of the actively moving RNAP switching to the paused state, as well as*

$$\omega_{r_a} = \frac{\pi_{r_a}\tau_g}{\pi_{r_a}\tau_{r_a} + \tau_g}, \quad \tilde{\alpha} = \frac{\tau_g + \lambda_{r_p}\pi_{d_p}\tau_g}{\tau_g + \tau_p + \lambda_{r_p}\tau_g(1 - \omega_{r_a})}, \quad \textit{and} \quad \omega = \frac{\tau_g}{\tau_{pp} + \tau_g}. \tag{3.38}$$

Assume that the elongation rate is faster than the rates of RNAP pausing, activation, and premature termination, i.e. that $k \gg r_a, r_p, d_a, d_p$. Then, it follows that to leading order in $1/k$, asymptotic

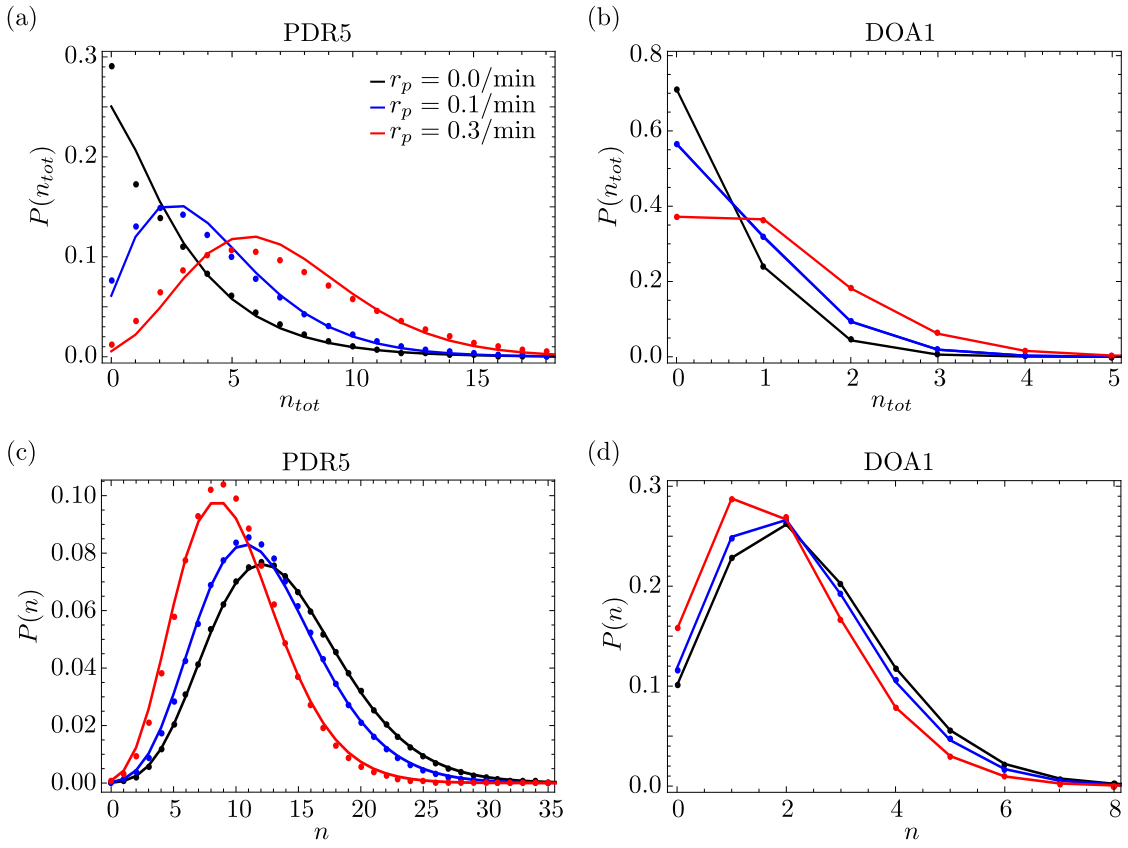


Figure 3.10: Dependence of the steady-state probability distributions of total RNAP and mature RNA on the RNAP pausing rate r_p for two different genes in yeast. In panels (a) and (b), we compare the distribution $P(n_{tot})$ of the total number of RNAP molecules, as predicted by our model (solid lines), with that obtained from SSA (dots) for yeast genes PDR5 and DOA1, respectively. The model prediction involves fitting a Negative Binomial distribution with a mean and variance given by the closed-form expressions in Eqs. (3.41) and (3.42). In panels (c) and (d), we compare the distribution $P(n)$ of mature RNA, as obtained from singular perturbation theory (Eq. (3.43); solid lines) with the SSA (dots) for yeast genes PDR5 and DOA1, respectively. Note that for both genes, we keep all parameters fixed (including the elongation rate k) while varying the pausing rate r_p to simulate an experiment where the pausing rate can be perturbed directly. The parameters for each gene can be found in Table 3.1; we furthermore used $L = 50$ and fixed k to $L/\langle T \rangle$, where $\langle T \rangle$ is the mean elongation time measured experimentally and reported in Table 3.1. Note that the actual mean elongation time is not fixed, as it depends on the pausing rate (r_p) via Eq. (3.40). The remaining parameters are fixed to $r_a = 0.1$/min, $d_a = 0.01$/min, and $d_p = 0.03$/min. The value of d_a is taken from Table 1 in [131], where it is reported as the premature termination rate of polymerase in *E. coli*; the value of d_p was chosen to be larger than that of d_a to simulate a scenario where premature detachment is enhanced in the paused state. Note that our theory is less accurate for PDR5 than it is for DOA1, as all parameters are very small compared to the elongation rate in the latter case, hence satisfying better the assumptions behind the theory.

expressions for the variances and covariances of molecule number fluctuations of active and paused

RNAP are given by

$$\begin{aligned}
&\mathrm{Cov}(n_i^a, n_j^a) = \delta_{ij}\langle n_i^a\rangle + \langle n_i^a\rangle\langle n_j^a\rangle\tilde{\alpha}\beta \cdot g_{ij}^{aa}, && \text{where } g_{ij}^{aa} = g^{aa}(i,j) + g^{aa}(j,i),\\
&\mathrm{Cov}(n_i^a, n_j^p) = \langle n_i^a\rangle\langle n_j^p\rangle\tilde{\alpha}\beta \cdot g_{ij}^{ap}, && \text{where } g_{ij}^{ap} = \omega\tilde{\alpha}^{j-1},\\
&\mathrm{Cov}(n_i^p, n_j^a) = \langle n_i^p\rangle\langle n_j^a\rangle\tilde{\alpha}\beta \cdot g_{ij}^{pa}, && \text{where } g_{ij}^{pa} = \omega\tilde{\alpha}^{i-1},\\
&\mathrm{Cov}(n_i^p, n_j^p) = \delta_{ij}\langle n_i^p\rangle + \langle n_i^p\rangle\langle n_j^p\rangle\tilde{\alpha}\beta \cdot g_{ij}^{pp}, && \text{where } g_{ij}^{pp} = (g_{ij}^{ap} + g_{ij}^{pa})/2;
\end{aligned} \tag{3.39}$$

here, $i, j = 1, 2, \ldots, L$ *and*

$$g^{aa}(i,j) = \frac{\tilde{\alpha}^{i+j-1}}{(2\tilde{\alpha}-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\tilde{\alpha}-1}{2\tilde{\alpha}}\,{}_2F_1\big(1, i+j; j; \tfrac{1}{2\tilde{\alpha}}\big)\Big].$$

These results are proved in full in Appendix A.7. From Appendix A.1, we also have that the mean elongation time in the pausing model is given by

$$\langle T\rangle = L\frac{(r_a + d_p)^2 + r_a r_p}{(r_a + d_p)[(k + d_a)(r_a + d_p) + d_p r_p]}. \tag{3.40}$$

Solving Eq. (3.40) for the elongation rate k, we find that in the limit of $L \to \infty$ taken at constant mean elongation time, k tends to infinity and hence is much larger than r_a, r_p, d_a, and d_p, which implies that the results of Prop. 4 hold naturally in that limit.

3.5.2 Approximate distributions of total RNAP and mature RNA

Negative Binomial approximation of total RNAP distribution

We define the total number of RNAP molecules as $n_{tot} = \sum_{i=1}^{L} n_i$. It then immediately follows from Eq. (3.37) that the mean of the total RNAP distribution in the pausing model is given by

$$\langle n_{tot}\rangle = \eta\rho_k(1+\lambda)\tilde{\mu}\frac{\tilde{\mu}^L - 1}{\tilde{\mu} - 1}. \tag{3.41}$$

It can also be shown that the variance of total RNAP fluctuations reads,

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + (\eta\rho_k)^2\tilde{\alpha}\beta\Big[2\sum_{i,j=1}^{L} g^{aa}(i,j) + \lambda(2+\lambda)\omega L\frac{\tilde{\alpha}^L - 1}{\tilde{\alpha} - 1}\Big]; \tag{3.42}$$

see Appendix A.7. Next, we approximate the distribution of total RNAP by a Negative Binomial distribution whose mean and variance match those just derived, i.e. we consider Eq. (3.20) with the mean and variance of the total RNAP distribution given by Eqs. (3.41) and (3.42) now, respectively. The resulting approximate Negative Binomial distribution is compared with the distribution obtained from SSA in Figs. 3.10(a) and (b) for two different yeast genes, PDR5 and DOA1. The results verify that our approximation is accurate provided the elongation rate k is significantly larger than the other parameters, as assumed in Prop. 4.

Perturbative approximation of mature RNA distribution

We can apply singular perturbation theory to formally derive the distribution of mature RNA, assuming that $k/d_m \gg 1$ and $r_a/d_m \gg 1$. Following the derivation in Section 3.3.2, we find the following analytical expression for the steady-state probability distribution of mature RNA:

$$P(n) = \frac{1}{n!}\frac{(s_u)_n}{(s_b + s_u)_n}\Big(\frac{r}{d_m}\Big)^n\big(\tilde{\mu}^L\big)^n\,{}_1F_1\Big(\frac{s_u}{d_m} + n; \frac{s_b + s_u}{d_m} + n; -\frac{r}{d_m}\tilde{\mu}^L\Big); \tag{3.43}$$

see Appendix A.8 for details. Note that the solution in Eq. (3.43) is dependent on the parameter $\tilde{\mu}^L$, which gives the probability that an RNAP molecule does not prematurely detach before termination; see Appendix A.1. Also, note that in the limit of zero premature termination, i.e. for $d_a = 0 = d_p$, Eq. (3.43) is identical to the distribution of mature RNA predicted by the telegraph model. Finally, by solving Eq. (3.40) for k, then substituting the resulting expression into Eq. (3.43) and taking the long-gene limit of $L \to \infty$ at constant $\langle T \rangle$, we obtain that the probability distribution of mature RNA has the same functional form as in Eq. (3.43), albeit with

$$\lim_{L\to\infty} \tilde{\mu}^L = e^{-\psi\langle T\rangle}, \quad \text{where } \psi = \frac{d_a + r_p \pi_{d_p}}{1 + \sigma \pi_{r_a}}. \tag{3.44}$$

Note that Eqs. (3.43) and (3.44) equal the steady-state solution predicted by the telegraph model, with the initiation rate r renormalised to $r\tilde{\mu}^L$ or $re^{-\psi\langle T\rangle}$, respectively. In Figs. 3.10(c) and (d), we verify the accuracy of our analytical solution using stochastic simulation for two different genes in yeast. Note that a change in the pausing rate r_p has relatively little effect on the distribution of mature RNA, as compared to the effect on the distribution of total RNAP; cf. panels (a) and (b) of Fig. 3.10 in comparison with panels (c) and (d), respectively.

3.6 Summary and discussion

In this work, we have analysed a detailed stochastic model of transcription. Our model extends previous analytical work [70, 72] by (i) taking into account salient processes, such as premature detachment and pausing of RNAP, that were previously not considered analytically; (ii) deriving explicit expressions for the mean and variance of RNAP numbers (nascent RNA) on gene segments as well as on the entire gene; (iii) deriving explicit expressions for the mean and variance of the fluorescent nascent RNA signal obtained from smFISH, and identifying differences between the statistics thereof and those of direct measurements of nascent RNA; (iv) finding approximate distributions of total nascent RNA fluctuations on a gene, without assuming slow promoter switching. A number of interesting observations from our work include the following.

- ◇ When the premature detachment rate of RNAP is non-zero and gene expression is bursty, the coefficient of variation of local RNAP fluctuations can either decrease or increase with distance from the promoter. By contrast, when the expression is constitutive, the coefficient of variation increases monotonically with distance from the promoter. Other statistical measures such as the mean, Fano factor, and correlation coefficient of local RNAP numbers decrease monotonically with distance from the promoter.

- ◇ In the limits of bursty expression, deterministic elongation, and no premature detachment or pausing, the Fano factor of total nascent RNA equals $1 + 2b$, whereas that of mature RNA is $1 + b$, where b denotes the mean burst size. An implication is that the telegraph model will result in an overestimate of the mean burst size from nascent RNA data by a factor of 2. Another implication is that deviations from Poisson fluctuations are more apparent in data for nascent RNA than they are for mature RNA. One can further state the following relationship: the Fano factor of nascent RNA equals twice the Fano factor of mature RNA, minus 1. If the expression is non-bursty, then the Fano factor of nascent RNA can be larger or smaller than that of mature RNA, as determined by the condition in Eq. (3.19).

- ◇ For genes characterised by bursty expression, the sensitivity of the noise in total RNAP fluctuations is highest to perturbations in the gene activation rate and the mean elongation time; for constitutive genes, the most sensitive parameters are the initiation rate and the mean elongation time.

- ◇ A Negative Binomial distribution, parametrised with the expressions for the mean and variance of total nascent RNA derived here, provides a good approximation to the true distribution of total nascent RNA fluctuations on a gene when the expression is either bursty or

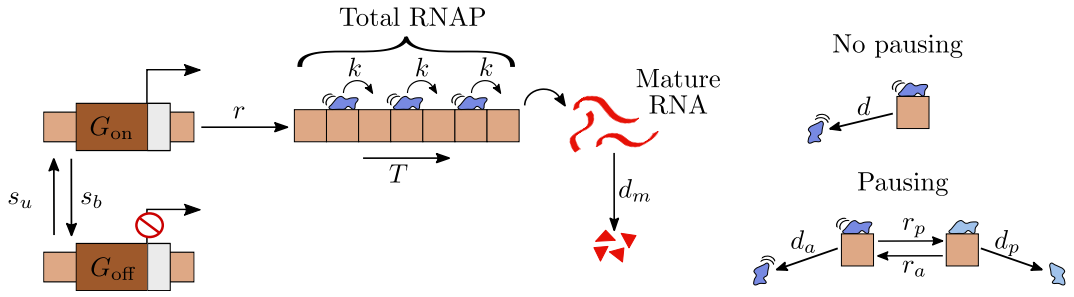
constitutive; the approximation is not accurate when the gene spends roughly equal amounts of time in the active and inactive states. We show that the distribution of nascent RNA is highly sensitive to the distribution of elongation times. In particular, if the elongation time is assumed to be exponentially distributed, as is implicitly assumed by telegraph models of nascent RNA, then the probability of observing zero RNA is much lower than if the elongation time is assumed to be fixed.

- ◇ Using geometric singular perturbation theory (GSPT), we have rigorously proven that, in the limit of deterministic elongation (or fast elongation), no pausing and premature detachment, the steady-state distribution of mature RNA in our model is identical to that in the telegraph model [31]. Consideration of pausing and premature detachment leads to a distribution that can also be obtained from a telegraph model with appropriately renormalised parameters.

A summary of the main theoretical results can be found in the table of main results on the next page, with all requisite parameters and functions defined in the table that follows. The main limiting assumption of our theoretical approach is that the initiation rate is slow enough such that RNAP molecules do not frequently collide with each other while moving along the gene. Hence, the expressions we have derived are reasonable for all but the strongest promoters, which are characterised by very fast initiation rates. We anticipate that approximate closed-form expressions for the corresponding moments can also be derived when volume exclusion between RNAPs is taken into account by a modification of methods previously devised to understand molecular movement and kinetics in crowded conditions [137, 138]. It is also possible to extend our model by including the translation of mature RNA to protein; one can then again apply GSPT to derive distributions for protein numbers in the limit of RNA decaying much faster than protein; however, given the last bullet point above, we anticipate that the resulting protein distribution will be very similar to those derived from models that do not explicitly take into account nascent RNA [43,95]. Further research is required to develop simple approximations of the nascent RNA distribution that are accurate independently of the ratio of gene switching rates. Finally, given the strong recent interest in the development of statistical inference techniques in molecular biology [134, 139, 140], we expect that our closed-form expressions for the moments and distributions of nascent and mature RNA will be useful for developing computationally efficient and accurate methods for estimating transcriptional parameters.

Table of main results

The cartoon represents our model in various limits: no pausing ($r_p = 0$), pausing ($r_p \neq 0$), stochastic elongation (T Erlang distributed), deterministic elongation (T fixed), bursty limit ($r, s_b \to \infty$), and premature RNAP detachment ($d, d_a, d_p \neq 0$). We summarize our analytical expressions for the approximate distributions and moments of total RNAP and mature RNA.



Approximate total RNAP distribution for stochastic elongation

$$n_{tot} \sim \mathrm{NB}\left(\frac{\langle n_{tot}\rangle^2}{\mathrm{Var}(n_{tot}) - \langle n_{tot}\rangle}, \frac{\mathrm{Var}(n_{tot}) - \langle n_{tot}\rangle}{\mathrm{Var}(n_{tot})}\right)$$

No pausing

$$\langle n_{tot}\rangle = \eta\rho_k\mu\frac{\mu^L - 1}{\mu - 1}$$

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + \alpha\beta(\eta\rho_k)^2 \textstyle\sum_{i,j=1}^{L} \mu^{i+j}\cdot[f(i,j) + f(j,i)]$$

Pausing

$$\langle n_{tot}\rangle = \eta\rho_k(1+\lambda)\tilde{\mu}\frac{\tilde{\mu}^L - 1}{\tilde{\mu} - 1}$$

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + \tilde{\alpha}\beta(\eta\rho_k)^2\left[2\textstyle\sum_{i,j=1}^{L} g^{aa}(i,j) + \lambda(2+\lambda)\omega L\frac{\tilde{\alpha}^L - 1}{\tilde{\alpha} - 1}\right]$$

Approximate mature RNA distribution ($k/d_m \gg 1$) as the solution of the telegraph model with renormalised transcription rate rP_μ and RNAP survival probability P_μ

$$n \sim P(n) = \frac{1}{n!}\frac{(s_u)_n}{(s_b + s_u)_n}\left(\frac{r}{d_m}\right)^n (P_\mu)^n\, {}_1F_1\left(\frac{s_u}{d_m} + n; \frac{s_b + s_u}{d_m} + n; -\frac{r}{d_m}P_\mu\right)$$

	Stochastic elongation		Deterministic elongation	
	Detachment	No detachment	Detachment	No detachment
No pausing Pausing	$d \neq 0$ $d_a, d_p \neq 0$	$d = 0$ $d_a = d_p = 0$	$d \neq 0$ $d_a, d_p \neq 0$	$d = 0$ $d_a = d_p = 0$
No pausing Pausing	$P_\mu = \mu^L$ $P_\mu = \tilde{\mu}^L$	$P_\mu = 1$ $P_\mu = 1$	$P_\mu = e^{-d\langle T\rangle}$ $P_\mu = e^{-\psi\langle T\rangle}$	$P_\mu = 1$ $P_\mu = 1$

Ratio of Fano factors (FF) of total RNAP and mature RNA for deterministic elongation without detachment ($d = 0$)

$$R_{FF} = \mathrm{FF_n}/\mathrm{FF_m}$$

	General	Bursty ($s_b, r \to \infty$)
No pausing	$R_{FF} = 1 + \frac{e^{-T_g}T_r T_{s_b}\Xi}{T_g^2\left[T_r T_{s_b} + T_g(T_g + T_m)\right]}$	$R_{FF} = 1 + \frac{b}{1+b}$

Table of definitions of parameters and functions

$$f(i,j) = \frac{\alpha^{i+j-1}}{(2\alpha-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\alpha-1}{2\alpha}{}_2F_1\big(1, i+j; j; \tfrac{1}{2\alpha}\big)\Big],$$

$$g^{aa}(i,j) = \frac{\tilde{\alpha}^{i+j-1}}{(2\tilde{\alpha}-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\tilde{\alpha}-1}{2\tilde{\alpha}}{}_2F_1\big(1, i+j; j; \tfrac{1}{2\tilde{\alpha}}\big)\Big],$$

$$\Xi = 2(T_g + T_m) + e^{T_g}[2(T_g-1)T_m + (T_g-2)T_g]$$

$\eta = s_u/(s_u+s_b)$	Fraction of time the gene spends in the active state.
$\rho_k = r/k$	Mean number of bound RNAPs in the time $1/k$.
$\rho = r/d_m$	Mean number of bound RNAPs in the time $1/d_m$.
$\mu = k/(k+d)$	Local RNAP survival probability (no-pausing case).
$\tau_p = 1/(d+k)$	Timescale of fluctuations of RNAP.
$\tau_g = 1/(s_u+s_b)$	Timescale of fluctuations of gene.
$\tau_d = 1/d$	Timescale of RNAP detachment.
$\tau_m = 1/d_m$	Timescales of fluctuations of mature RNA.
$\alpha = 1/(1+\tau_p/\tau_g)$	Non-dimensional parameter.
$\gamma = 1/(1+\tau_p/\tau_m)$	Non-dimensional parameter.
$\theta = 1/(1+\tau_m/\tau_g)$	Non-dimensional parameter.
$\beta = s_b/s_u$	Ratio of gene inactivation and activation rates.
$b = r/s_b$	Mean burst size.
$\upsilon_k = s_u/k$	Ratio of gene activation and RNAP elongation rates.
$\upsilon_m = s_u/d_m$	Ratio of gene activation and mature RNA degradation rates.
$\delta_g = \tau_g/\tau_d$	Ratio of gene timescale and RNAP detachment timescale.
$T_g = \langle T\rangle/\tau_g$	Ratio of elongation timescale and gene timescale.
$T_d = \langle T\rangle/\tau_d$	Ratio of elongation timescale and RNAP detachment timescale.
$T_r = r\langle T\rangle$	Ratio of the mean elongation time to the timescale of initiation.
$T_m = d_m\langle T\rangle$	Ratio of the mean elongation time to the timescale of decay of mRNA.
$T_{s_b} = s_b\langle T\rangle$	Ratio of the mean elongation time to the timescale of gene deactivation.
$\sigma = r_p/r_a$	Ratio of the pausing and activation rates.
$\pi_{r_a} = r_a/(r_a+d_p)$	Probability of RNAP activation.
$\pi_{d_p} = d_p/(r_a+d_p)$	Probability of premature RNAP detachment from the paused state.
$\lambda = \sigma\pi_{r_a}$	Probability of RNAP pausing from active state.
$\tilde{\mu} = k/(k+d_a+r_p\pi_{d_p})$	Local RNAP survival probability (in pausing case).
$\tau_{r_a} = 1/r_a$	Timescale of RNAP activation from the paused state.
$\tau_{d_p} = 1/d_p$	Timescale of premature termination of paused RNAP.
$\tau_p = 1/(k+d_a)$	Typical time that an actively moving RNAP spends on a gene segment.
$\tau_{pp} = 1/(r_a+d_p)$	Typical time spent in the paused state.
$\lambda_{r_p} = \pi_{r_p}/(1-\pi_{r_p})$	Ratio of active RNAP timescale over RNAP pausing timescale.
$\pi_{r_p} = r_p/(r_p+k+d_a)$	Probability of the actively moving RNAP switching to the paused state.
$\omega_{r_a} = \pi_{r_a}\tau_g/(\pi_{r_a}\tau_{r_a}+\tau_g)$	Non-dimensional parameter.
$\tilde{\alpha} = (\tau_g + \lambda_{r_p}\pi_{d_p}\tau_g)/(\tau_g+\tau_p+\lambda_{r_p}\tau_g(1-\omega_{r_a}))$	Non-dimensional parameter.

3.7 Proposed extensions of the detailed model

The model discussed in this chapter captures a number of transcription mechanisms, though there is still ample scope for potential extensions of this model in order to make it more biologically realistic. Obviously, biologically realistic models are desirable because they can be used to extract important information from experimental data; however, very complex models are frequently not analytically tractable and need to satisfy a number of assumptions in order to be suitable for mathematical analysis. Some small modifications to our detailed model of transcription can still leave the model simple enough for mathematical analysis. We discuss various possible model extensions in the following bullet points.

- ◇ **Number of promoter states.** The first modification to the detailed model that one can think of is related to promoter activation/inactivation. The model can be extended to have three gene states – as in the multi-scale model from Chapter 2 – which can feature the model with processes of polymerase recruitment and pause release. Another modification could be to assume that the promoter can switch between an arbitrary integer number of states; different states are related to different transcription factors that bind to the promoter region and modulate the rate of transcription initiation. Examples of such models can be found in [71, 113–115].

- ◇ **Multi-step transcription initiation** has been observed in experiments [141]. In our model, transcription initiation is modelled as a one-step process. This can be easily extended to a case of a process with more steps. We have mentioned before, the two-step transcription initiation models from [70, 71], while there is a more recent study of a model with multiple steps of initiation that has been performed by J. Szavits-Nossan et al. in [142].

- ◇ **Polymerase backtrack pausing.** For our detailed model, polymerase movement along the gene during elongation happens only in a forward way, and no backtracking is possible. The concepts of backtrack pausing and nonbacktrack pausing have been discussed [143, 144] in and could be as well adapted for our detailed model.

- ◇ **The elongation, pausing, and polymerase detachment rates** are assumed to be independent of the position of the polymerase on the gene in our model. This is not always the case in reality and can be an idea for more modifications to the model. For example, multiple factors can modulate elongation rates and there are observations of an increase in the polymerase elongation rate as the polymerase moves along the gene [68].

- ◇ **Multi-step transcription termination.** In our detailed model, we assume that when a polymerase has reached the end of the gene, it falls off at the same rate as it was hopping on the gene, and at the same time, the transcribed nascent RNA detaches from the polymerase and becomes a mature RNA molecule. In reality, this process does not happen in one-step because there is some delay time before a nascent RNA becomes mature. Introducing a deterministic delay of RNA maturation time in the model would make it more realistic; an example of such a model can be found in [72].

- ◇ **Excluded-volume interaction between adjacent polymerases.** We have discussed before, that our detailed model is reasonable for genes with promoters that are characterised by very slow initiation rates in order to prevent traffic of polymerases on the gene; if polymerases were produced faster than they move along the gene, then they would collide with each other. In order to make the model valid for strong promoters (fast initiation rate) as well, one can treat polymerases as solid objects with a certain length and take into account the volume exclusion between them. This means that the transcription elongation process can actually be seen as a totally asymmetric simple exclusion process (TASEP).

  TASEP is a fundamental dynamical model from nonequilibrium statistical mechanics that was first introduced by MacDonald, Gibbs and Pipkin in 1968 [145, 146]. TASEP describes

particles randomly hopping along a one-directional and one-dimensional chain of sites, where each site can either be empty or contain a single particle. This simple exclusion principle generates an indirect coupling between the particles, as a particle cannot hop to a site that is already occupied by another particle. TASEP was first introduced as a model for the motion of ribosomes along the mRNA strand during the process of translation (see examples in [147, 148]), while it has found numerous other applications in the past decades. TASEP has been successfully used for studying transcriptional dynamics by modelling elongation and considering particles' pausing [130, 149–151]; however, the common factor of all these studies is that they assume that the promoter of the gene can only be in one state (always active).

- ◇ **Gene expression response to signals.** Individual cells detect and respond to diverse internal and external molecular and physical signals. Temporal variation in environmental stimuli usually leads to changes in gene expression. The detailed model can be modified in order to take into account the time-dependent dynamics of stimuli.

After our work in the chapter, our research interest lies in understanding how a stimulus affects the transcript numbers measured at various sub-cellular locations. Hence, we construct a stochastic model describing the dynamics of signal-dependent gene expression and its propagation downstream of transcription and perform an analytical study of this model, which we present in the next chapter.

Chapter 4

Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions

This chapter contains published work. Please see the Declaration of Authorship for details. The Sections 4.1-4.5 contain the published article by T.Filatova et al. [152]. The Lay summary and Section 4.6 are supplementary to the publication and are written for the purposes of this thesis.

Lay summary

The health of organisms and cells depends on appropriate responses to different internal and external signals; e.g. change in the hormone levels or exposure to various pathogens. One of the ways that cells respond to these signals is by regulating the expression of target genes. Specifically, certain signalling molecules – which are usually transcription factors – encode information about the properties and intensity of the signals, and later on, the gene promoters decode the stored information and determine whether a specific gene is activated and by how much. These encoding and decoding mechanisms are still not well understood. Experimental evidence suggests that there are molecular mechanisms that emit time-dependent signals to certain transcription factors and change their concentration. The target genes respond accordingly to different concentrations of transcription factors by activating or inactivating transcription initiation; this is one of the signal-transition mechanisms, and the signalling pathways obtain the key to regulation of gene expression. Since the time-dependent environmental signals lead to changes in gene expression and since many biological steps occur inside a cell between the birth and death of an mRNA molecule, we desire to understand how these signals affect the numbers of RNA molecules that can be measured at various sub-cellular locations (nucleus or cytoplasm). The "multi-scale" model studied in Chapter 2 and the "detailed" model studies in Chapter 3 are not featured signal dependence; hence, in this chapter, we construct and study a stochastic model of gene expression that takes into account time-dependent signal dynamics.

In the model studied here, the promoter can switch between two states, on and off. The promoter activates when transcription factors bind to it. We assume that the information of some *oscillatory time-dependent signal* is encoded in the concentration of transcription factors – i.e. there is *oscillatory time-dependent variation in transcription factor numbers* – and hence, we allow the binding rate of transcription factors (which is also the promoter activation rate) to be time-dependent. The active promoter can switch back to its inactive state with a certain probability per

unit of time. Transcription initiation can occur only when the promoter is active, leading to the synthesis of mRNA. Now, we *divide the life-cycle (production to degradation) of the mRNA into an arbitrary number of stages*, and we assume that the produced mRNA undergoes all of these stages (by randomly changing from one stage to the next with a certain probability per unit of time) before finally decaying. This modelling gives us the means to investigate *how the signal affects the mRNA at different stages of its life-cycle*; e.g. if we define certain initial stages of mRNA as nuclear mRNA and the rest of the stages as cytoplasmic mRNA, then we can study how the numbers of these nuclear and cytoplasmic mRNA species fluctuate due to signal. It is worth noting here that, nuclear–cytoplasmic mRNA is not a unique interpretation of the different life stages of mRNA, e.g. they could be as well seen as unspliced and spliced mRNA or nascent and mature mRNA; hence, the definition and the interpretation of the different mRNA life-cycle stages depend on the user.

Mathematical analysis of this complex model appeared to be a hard task; however, we were able to overcome certain difficulties by making the assumption that this model is valid for genes that are characterised by bursty promoters (i.e., the promoters that spend most of their time being inactive). Due to this assumption, we were able to reduce our model to a more simple version, which was easier to mathematically manipulate. Then, we developed a new approximation technique that allowed us to obtain analytical expressions for time-dependent distributions of mRNA numbers at all stages of its life-cycle. We also derived an expression for the error in the approximation whose accuracy was verified via stochastic simulation. We found that, depending on the frequency of oscillation, a time-dependent signal can lead to an increase or decrease in the fluctuation of the number of cytoplasmic mRNA molecules.

Abstract

Temporal variation of environmental stimuli leads to changes in gene expression. Since the latter is noisy and since many reaction events occur between the birth and death of an mRNA molecule, it is of interest to understand how a stimulus affects the transcript numbers measured at various sub-cellular locations. Here, we construct a stochastic model describing the dynamics of signal-dependent gene expression and its propagation downstream of transcription. For any time-dependent stimulus and assuming bursty gene expression, we devise a procedure that allows us to obtain time-dependent distributions of mRNA numbers at various stages of its life-cycle, e.g. in its nascent form at the transcription site, post-splicing in the nucleus, and after it is exported to the cytoplasm. We also derive an expression for the error in the approximation whose accuracy is verified via stochastic simulation. We find that, depending on the frequency of oscillation and the time of measurement, a stimulus can lead to cytoplasmic amplification or attenuation of transcriptional noise.

4.1 Introduction

Many genes are transcribed in a bursty fashion [53], which is due to the fact that they spend most of their time in the "off" state, switching on for a relatively short time period during which a burst of mRNA molecules is rapidly produced. Furthermore, both the size of the burst in transcript numbers and the time between successive bursts are random [64,153]. Noise can be due to intrinsic and extrinsic sources. Intrinsic noise stems from uncertainty in the timing of individual reaction events leading to transcription, whereas extrinsic noise arises independently of the gene but acts on it, e.g. through the number of RNA polymerases [154]. The mechanisms shaping transcriptional bursting are still not clearly understood and represent a topic of active research [155, 156].

The above has inspired the construction of stochastic models of gene expression with the aim of understanding how the distribution of mRNA numbers varies with transcriptional parameters. The simplest model that is in widespread use is the two-state telegraph model; considering exclusively intrinsic noise, its stochastic dynamics are described by the chemical master equation (CME) [83] which can be solved exactly [26, 61], yielding an explicit analytical solution for the distribution of transcript numbers as a function of the initiation rate, the switching rates between the active ("on") and inactive ("off") states of the gene, and the mRNA degradation rate. Within that model, the burst frequency is the rate at which the gene switches on, while the burst size is the initiation rate divided by the rate of switching off. Modifications of the telegraph model have been proposed to take into account noise in transcript numbers due to a wide variety of biological processes, such as the doubling of the gene copy number during DNA replication, partitioning of molecules between daughter cells during cell division, variability in the cell cycle duration time, coupling of gene expression to cell size or cell cycle phase, multiple off states, proximal-promoter pausing, RNA polymerase fluctuations, export from the nucleus to the cytoplasm, post-transcriptional modifications, and cell-to-cell variation in transcriptional parameters [113, 123, 126, 157–165].

A common property of the bulk of published, analytically solvable models is their lack of description of the coupling of gene expression to an extracellular time-dependent signal. It is known that the identities and intensities of different stresses are transmitted by modulation of certain transcription factors (TFs) in the cytoplasm which exerts an influence on gene expression upon their translocation to the nucleus [166–170]. This modulation is of particular importance in developmental biology whereby spatio-temporally varying distributions of TFs (morphogens) play a key role in establishing the body plan [171–176]. While TFs can exert influence on gene expression via modulation of the burst size and the burst frequency [46], modelling has shown that regulation through the latter is advantageous because weak TF binding is sufficient to elicit strong transcriptional responses [177]. In fact, the changes in distributions of nascent mRNA with TF concentration are very well captured by a telegraph model, modified so that the switching from the "off" state to the "on" state is an increasing function of the concentration [82, 175].

There are very few studies that have attempted to analytically solve the telegraph model (or

similar models) for the distribution of transcript numbers in response to a time-dependent stimulus. In [178], an exact solution of the telegraph model with signal-dependent initiation rate is presented; because the initiation rate controls the burst size and not the frequency, that model does not capture the commonest way by which stimuli affect gene expression. In [179], an approximate solution of an auto-regulatory genetic feedback loop with a signal-dependent initiation rate is presented, which has the same disadvantage as mentioned in the previous study. By contrast, in [180], the stimulus is assumed to affect any one of the parameters in the telegraph model, which is hence compatible with the notion that stimuli are transmitted principally via modulation of the burst frequency; the solution of the resulting stochastic model is approximate, and most accurate when the stimuli are slowly varying. In [181], a model where the burst frequency is modulated by an external signal is studied using a continuum approximation of the master equation. However, in these studies, there is no description of how the signal affects mRNA at different stages in its life-cycle, e.g. through differences between the temporal variation of the transcript numbers at the transcription site, in the nucleus, and in the cytoplasm. That is important, as there are significant measured differences in the distributions and moments of nuclear and cytoplasmic mRNA [74, 182].

In this chapter, we consider a stochastic model of gene expression in which a deterministic, temporally variable TF abundance modulates the burst frequency of a gene. By means of a novel approximation, we obtain closed-form analytical expressions for the time-dependent distribution of mRNA transcript numbers at any stage in the life-cycle, which is often correlated with sub-cellular localization. The chapter is organized as follows. In Section 4.2, we introduce a model of signal-dependent bursty gene expression where changes in some extracellular signal are reflected in the rate at which a gene switches on; for simplicity, we choose this rate to vary sinusoidally in time – an assumption that we relax later on. As the resulting model is multi-variable, the analytical time-dependent solution of its CME is highly challenging. The high dimensionality of the model here stems from the presence of L mRNA species, each describing mRNA abundance at a different stage in the life-cycle. We circumvent this challenge by postulating that the marginal time-dependent solution of mRNA at a particular life-cycle stage is described by an analytically tractable effective one-variable stochastic model with some unknown effective parameters. In Section 4.3, we describe a procedure by which the latter parameters can be found as a function of the parameters of the larger multi-variable model. We then show that for a wide range of parameters, the analytical solution of the one-variable model provides an excellent approximation to the marginal time-dependent distributions in the multi-variable model, as obtained via stochastic simulation. Finally, in Section 4.4, we extend the above procedure to models where the switching-on rate varies in a complex non-sinusoidal manner with time, which reflects the complexity of *in vivo* extracellular stimuli and the intricate molecular details of TF binding. We conclude with a discussion in Section 4.5.

4.2 Model description

In this section, we introduce a number of stochastic gene expression models of the mRNA life-cycle which incorporate a temporally variable TF abundance due to an extracellular stimulus; see Fig. 4.1 for an illustration.

Full model (FM). This model assumes that the gene can be either inactive (off), G_{OFF} or active (on), G_{ON}, and that the mRNA life-cycle is divided into L stages, where the species M_j ($j = 1, 2, \ldots, L$) denotes the mRNA in its j-th life-cycle stage. We assume that TFs can bind to enhancer or promoter sequences with some rate σ' which leads to gene activation, at which point transcription is initiated. Furthermore, we assume that TF numbers vary periodically as $A(1 + \varepsilon \cos(\omega t + \varphi))$ where A is the amplitude, $\omega \geq 0$ denotes the frequency, and $\varphi \in [-\pi, \pi)$ denotes the phase. Note that we choose the constant $|\varepsilon| \leq 1$ such that the TF signal is always a positive-valued function. Note also that the choice of the cosine over a sine does not affect the periodicity of the signal, since $\cos(x) = \sin(x + \pi/2)$ for all $x \in \mathbb{R}$. It then follows by the law of

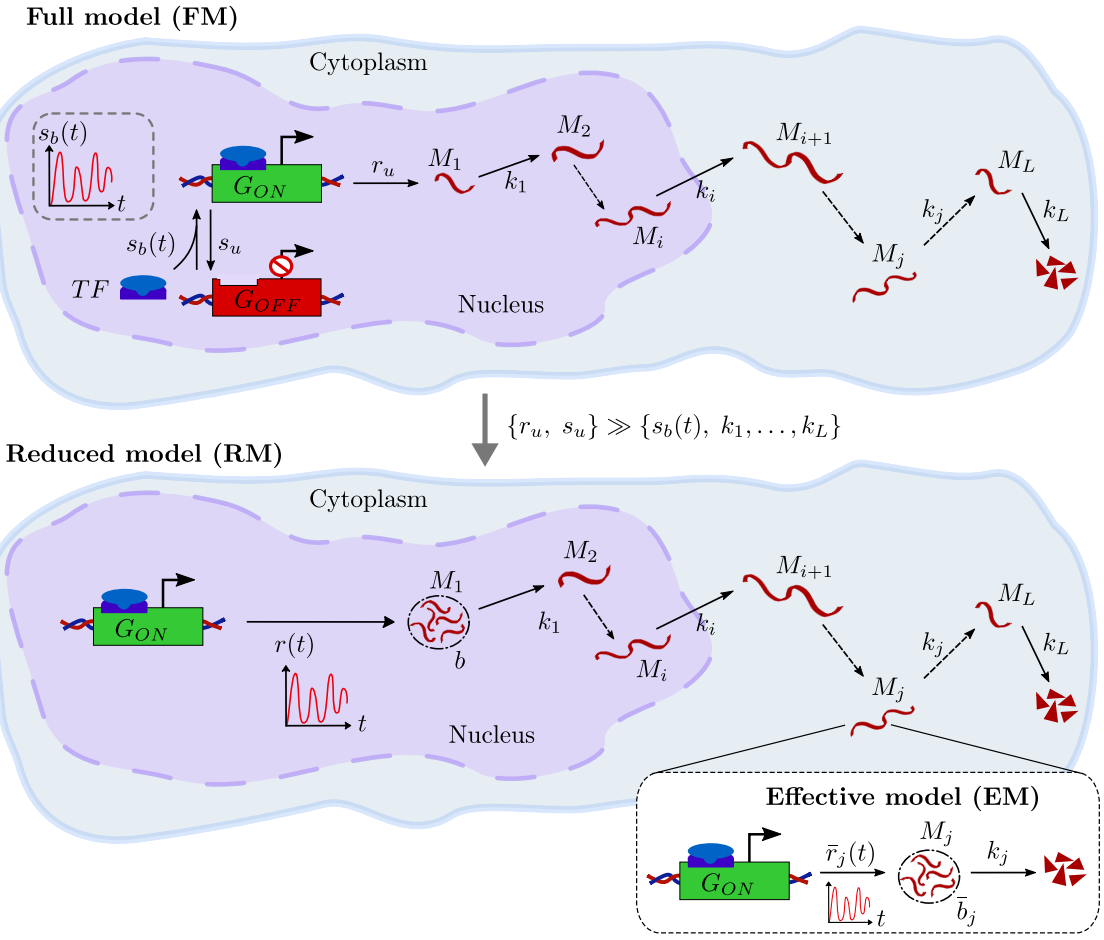


Figure 4.1: *Illustration of three different stochastic models of the mRNA life-cycle.* In the full model (FM), the gene can be in two states, active (G_{ON}) and inactive (G_{OFF}). The binding of TFs causes a transition from the inactive to the active state. The binding rate $s_b(t)$ is time-dependent due to the temporal variation in TF numbers. The gene can switch back to its inactive state with rate s_u. While the gene is active, transcriptional initiation occurs with constant rate r_u, leading to synthesis of mRNA (M_1). After the produced mRNA undergoes L stages of its life-cycle (M_j) with rates k_j ($j = 1, \ldots, L-1$), it finally decays with rate k_L. When transcription is bursty, the FM is well approximated by the reduced model (RM) which assumes that transcriptional initiation and gene inactivation rates (r_u, s_u) are much larger than the remaining kinetic rates. Here, mRNA synthesis occurs at a rate $r(t) = s_b(t)$ in bursts with mean size $b = r_u/s_u$. As before, the produced mRNA undergoes L stages of its life-cycle and eventually decays. In this chapter, we show that the distribution of mRNA numbers in each life-cycle stage in the RM is well approximated by the distribution in an effective model (EM), which incorporates two rates: time-dependent bursty production of mRNA and degradation thereof. The advantage of the EM over the other two models is that it can be solved analytically, yielding time-dependent distributions of mRNA in each life-cycle stage. See the main text for a more detailed description of these models.

mass action that the activation rate from the inactive to the active state is given by

$$s_b(t) = \sigma(1 + \varepsilon \cos(\omega t + \varphi)), \tag{4.1}$$

where $\sigma = A\sigma'$. Note that we have assumed the binding rate to be a linear function of TF numbers here, which is a simplification, as TF binding kinetics is often cooperative [175], a case that we

will discuss later on. Note also that, in reality, TF signals are to some degree noisy; however, in our case, we consider the signal to be deterministic for simplicity. For theoretical approaches developed for the study of stochastic kinetic rates, see e.g. [183–185]. The activation of the gene is a reversible reaction, i.e. the active gene can switch back to the inactive state with rate s_u. Once the gene is activated, transcription initiation starts and with rate r_u leads to mRNA in stage one, denoted as M_1. Subsequently, the synthesized mRNA progresses through its life-cycle by changing from stage M_j to stage M_{j+1} $(j = 1, \ldots, L-1)$ with hopping rate k_j. At the end of its life-cycle, the final mRNA state M_L decays with rate k_L. The system of chemical reactions describing the full model (FM) is given by

$$G_{OFF} \underset{s_u}{\overset{s_b(t)}{\rightleftharpoons}} G_{ON}, \quad G_{ON} \xrightarrow{r_u} G_{ON} + M_1, \quad M_j \xrightarrow{k_j} M_{j+1} \quad (j = 1, \ldots, L-1), \quad M_L \xrightarrow{k_L} \varnothing. \tag{4.2}$$

(Here, the empty set $\varnothing$ denotes a sink of molecules.) Note that the mRNA life-cycle stages can represent any of the following processes: transcription initiation, splicing, elongation, maturation, and degradation [6]. As each of these can be modeled as one-step or multi-step processes, the parameter L is user-defined; hence, we present our analysis for the case of general L here. Note that if we define stages 1 to R to be nuclear, where R is some integer less than L, it follows that the time between initiation and export to the cytoplasm is a sum of R exponential random variables, each with mean $1/k_j$. Hence, the distribution of the nuclear retention time is a hypoexponential distribution; similarly, one can argue that the same distribution describes the lifetime of the cytoplasmic mRNA. Two special cases of this model have been previously studied: (i) the case $L = 1$ with pulse-like (non-sinusoidal) activation rate $s_b(t)$ [186], and (ii) the case of constant (non-time dependent) activation rate for general L [163].

Reduced model (RM). The analytical derivation of the time-dependent mRNA number distribution in a given stage j in the FM is a challenging task. A simplification is achieved from the observation that mRNA expression is often bursty, i.e. that the gene spends most of its time in the off state, producing a short-lived burst of molecules while in the on state [53,64]. That observation leads us to introduce a simpler version of the full model, which we refer to as the reduced model (RM) and which is described by the reaction scheme

$$\varnothing \xrightarrow{r(t)} mM_1, \quad M_j \xrightarrow{k_j} M_{j+1} \quad (j = 1, \ldots, L-1), \quad M_L \xrightarrow{k_L} \varnothing. \tag{4.3}$$

(Here, the empty set $\varnothing$ denotes sources and sinks of molecules.) While there is no explicit gene switching in the RM, it is effectively taken into account by mRNA production occurring in bursts of size m, where $m = 0, 1, \ldots$ is a random variable chosen from a geometric distribution:

$$P(m) = \frac{b^m}{(1+b)^{m+1}}, \quad \text{where } b = \frac{r_u}{s_u}. \tag{4.4}$$

Note that the geometric distribution has a solid experimental and theoretical basis in the context of bursty expression [43,73,126,187–189]. The parameter b is the mean burst size of mRNA molecules produced while the gene is active; values of this parameter for different genes have been reported in various studies [53]. In order to incorporate the dependence of transcription on TF numbers, we assume that burst production occurs with a time-dependent rate, which is exactly the gene activation rate from our FM: $r(t) = s_b(t)$. As for the full model, M_1 undergoes L life-cycle stages after it has been synthesized, followed by final mRNA degradation.

In Appendix B.2, we prove the equivalence of the two models when the transcriptional initiation rate and the gene inactivation rate of the FM are much larger than the remaining kinetic rates, i.e. when $\{r_u,\ s_u\} \gg \{s_b(t),\ k_1, \ldots, k_L\}$; therefore, in the remainder of the chapter, our mathematical analysis is solely based on the RM.

To complete the setup of our RM, we introduce the following definitions. We define the vector of molecule numbers $\vec{n} = (n_1, \ldots, n_L)$, and we write $\langle n_j \rangle$ $(j = 1, \ldots, L)$ for the average number of molecules of species M_j. The RM can then be conveniently described by L species interacting

via a set of $L+1$ reactions with a rate function vector $\vec{f} = (f_1, \ldots, f_{L+1})$ which has the following entries:

$$f_1 = r(t) \quad \text{and} \quad f_j = k_{j-1}\langle n_{j-1}\rangle \quad \text{for } j = 2, \ldots, L+1. \tag{4.5}$$

The rate functions f_j are the averaged propensities of the underlying chemical master equation (CME) [88]. The description of our model is completed by the $L\times(L+1)$-dimensional stoichiometric matrix $\mathbf{S}$; the element S_{ij} of $\mathbf{S}$ gives the net change in the number of molecules of the i-th species when the j-th reaction occurs. Given the ordering of species and reactions as described in Eq. (4.3), it follows that the matrix $\mathbf{S}$ has the simple form

$$S_{11} = m, \quad S_{jj} = 1 \text{ for } j = 2, \ldots, L, \quad \text{and} \quad S_{j,j+1} = -1 \text{ for } j = 1, \ldots, L, \tag{4.6}$$

with the remaining elements being equal to zero.

Effective model (EM). Finding the exact closed-form time-dependent mRNA distributions for each life-cycle stage in the RM is still very difficult. In this chapter, we will show that we can well approximate the distribution of mRNA species in the j-th life-cycle stage in the RM – and, hence, in the FM – by the distribution in a simpler effective model (EM). The latter is defined by the following reaction scheme,

$$\varnothing \xrightarrow{\bar{r}_j(t)} \bar{m}M_j \quad M_j \xrightarrow{k_j} \varnothing, \quad \text{where } \bar{r}_j(t) = \bar{\sigma}_j(1 + \bar{\varepsilon}_j \cos(\omega t + \bar{\varphi}_j)). \tag{4.7}$$

In the EM, $\bar{m}$ is a random variable chosen from the geometric distribution

$$P_j(\bar{m}) = \frac{\bar{b}_j^{\bar{m}}}{(1+\bar{b}_j)^{\bar{m}+1}}, \tag{4.8}$$

where $\bar{b}_j$ is the mean burst size for this model. Note that the subscript j in the parameters $\bar{b}_j$, $\bar{\sigma}_j$, $\bar{\varepsilon}_j$, and $\bar{\varphi}_j$ denotes their dependence on the life-cycle stage j in the RM; these are to be determined later. However, we assume that the signal frequency ω does not depend on the stage j, and that it is the same as in the RM. Finally, the degradation rate of mRNA in the EM is k_j, which is its hopping rate to the next stage in the RM – or its degradation rate if $j = L$.

4.3 Approximation of the distribution of mRNA numbers in the RM

Because of its high dimensionality due to the presence of L species, it is difficult to solve the CME for the RM and, hence, to obtain a time-dependent probability distribution of mRNA numbers in every stage in the life-cycle. We therefore take a different approach to obtain that distribution. In Section 4.3.1, we derive exact closed-form expressions for the first two moments of mRNA distributions in the RM. Subsequently, in Section 4.3.2, we find formulae for the effective parameters of the EM such that the mean number of mRNA molecules in each life-cycle stage matches exactly that in the RM, while the variance in number fluctuations in the two models is matched approximately. Since the EM has the benefit that its CME can be solved exactly in time – as it has only one effective species – we finally obtain an analytical time-dependent distribution of mRNA numbers that is a good approximation of the distribution in the RM.

4.3.1 Exact closed-form expressions for mean and variance of mRNA distributions for the RM in the cyclo-stationary limit

In this section, we obtain analytical closed-form expressions for the first two moments of mRNA distributions in each life-cycle stage, in the limit of long times ($t \to \infty$). Henceforth, we will refer to this limit as the "cyclo-stationary limit" [190], since for $t \to \infty$, our solutions are still functions of time due to the periodic time-dependent TF signal. Note that in our analysis, we ignore fluctuations due to binomial partitioning at cell division; this approximation is valid as long as the mRNA lifetime is much shorter than the mean cell-cycle duration [162].

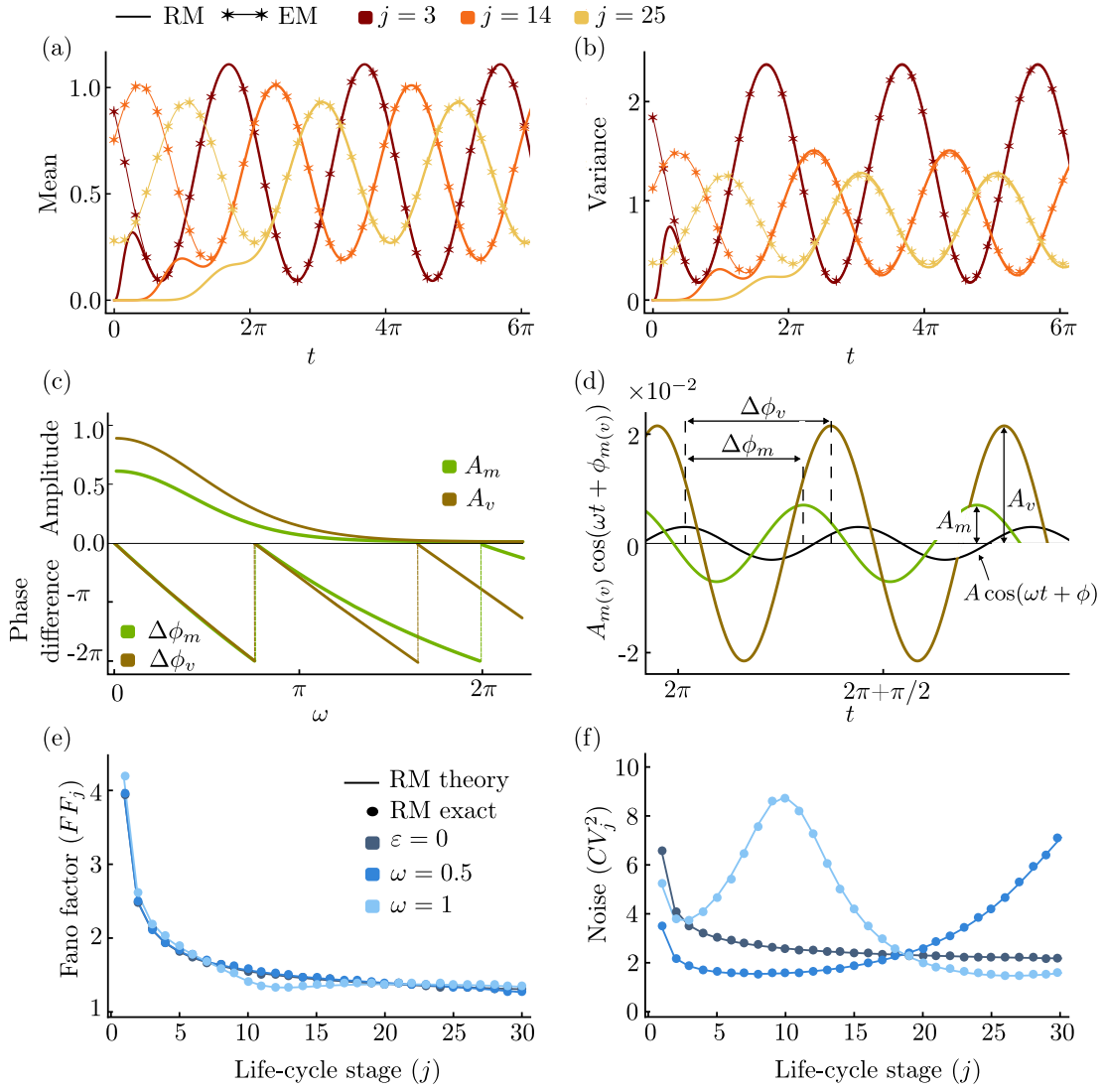


Figure 4.2: *First two moments of the mRNA distributions in the RM.* In panels (a) and (b), we illustrate the time evolution of the mean and the variance of mRNA distributions for three different stages of the mRNA life-cycle; thick solid lines show the direct numerical solution of the moment equations of the RM, given by Eq. (4.9) and Eq. (4.12), while thin solid lines with asterisks correspond to the approximation provided by the EM, Eq. (4.24). In panel (c), we show that the amplitudes of the oscillations in the moments decrease monotonically with signal frequency ω; their phase differences with the signal reach zero for some value of ω at which the moment waves are in phase with the signal wave. In (d), we illustrate the time evolution of the time-dependent terms in the moments and the rescaled signal, $3 \cdot 10^{-3} \cos(\omega t + \varphi)$, for $\omega = 3\pi/2$. For panels (c) and (d), we used Eq. (4.10) and Eq. (4.13); we note that, while the decreasing behavior of the amplitudes and phase differences holds for all stages j ($j = 0, \ldots, L$), we chose to present it for $j = 14$. In panels (e) and (f), we show the variation of the Fano factor (FF_j) and the noise (CV_j^2) over the mRNA life-cycle for the RM. We present the case of constant signal, with $\varepsilon = 0$, and two examples of a time-dependent signal with different frequencies, as predicted by our theory (Eq. (4.18); solid lines) and simulation (SSA; dots) for time point $t = 10$ min. For the parameter values, see Appendix B.1.

Given the CME describing the stochastic dynamics of the RM, it is straightforward to show from the corresponding moment equations [100] that the time evolution of the vector $\langle\vec{n}\rangle$ of mean molecule numbers is given by the set of ordinary differential equations (ODEs)

$$\frac{d\langle\vec{n}\rangle}{dt} = \mathbf{S}\cdot\vec{f}(\langle\vec{n}\rangle), \tag{4.9}$$

where $\vec{f}$ and $\mathbf{S}$ are defined in Eq. (4.5) and Eq. (4.6), respectively. Solving the above system of ODEs and taking the cyclo-stationary limit, we find that the solution can be written as

$$\begin{aligned}\langle n_j\rangle &= bk_j^{-1}\sigma(1+\varepsilon K_j\cos(\omega t+\varphi+\Theta_j))\\ &= bk_j^{-1}\sigma + A_j^m\cos(\omega t+\varphi+\Delta\varphi_j^m),\end{aligned} \tag{4.10}$$

where $A_j^m = bk_j^{-1}\sigma\varepsilon K_j$ is the amplitude of the oscillation in the mean and $\Delta\varphi_j^m = \Theta_j$ is the phase difference between this oscillation and the signal; the superscript m refers to the mean. The remaining parameters are defined as

$$K_j = \prod_{q=1}^{j}\frac{k_q}{\sqrt{k_q^2+\omega^2}} \quad \text{and} \quad \Theta_j = -\sum_{q=1}^{j}\tan^{-1}\left(\frac{\omega}{k_q}\right). \tag{4.11}$$

See Appendix B.3 for a detailed derivation of these results. Since the propensities are linear in the number of molecules, the corresponding second moments at steady state are exactly given by a Lyapunov equation [100]. That equation, which is precisely the same as the one that is obtained from the linear noise approximation (LNA) [98], takes the form

$$\dot{\mathbf{C}} = \mathbf{J}\cdot\mathbf{C} + \mathbf{C}\cdot\mathbf{J}^T + \mathbf{D}, \tag{4.12}$$

where the overdot denotes a time derivative. Here, $\mathbf{C}$, $\mathbf{J}$, and $\mathbf{D}$ are $L\times L$-dimensional matrices: $\mathbf{C}$ is a covariance matrix that is symmetric ($C_{ij} = C_{ji}$), $\mathbf{J}$ is the Jacobian matrix with elements $J_{ij} = \partial(\mathbf{S}\cdot\vec{f})_i/\partial\langle n_j\rangle$, and $\mathbf{D} = \mathbf{S}\cdot\mathbf{Diag}(\vec{f})\cdot\mathbf{S}^T$ is a diffusion matrix, where $\mathbf{Diag}(\vec{f})$ is a diagonal matrix whose elements are the entries in the rate function vector $\vec{f}$. Eq. (4.12) can be solved explicitly for the covariance matrix $\mathbf{C}$, the diagonal elements of which correspond to the variance of mRNA distributions in each life-cycle stage. In the cyclo-stationary limit, the latter are given by

$$\begin{aligned}Var(n_j) &= \langle n_j\rangle + b^2\sigma k_j^{-1}(G_j^0+\varepsilon G_j\cos(\omega t+\varphi+\Phi_j))\\ &= bk_j^{-1}\sigma(1+bG_j^0) + A_j^v\cos(\omega t+\varphi+\Delta\varphi_j^v).\end{aligned} \tag{4.13}$$

Here, A_j^v is the amplitude of oscillations in the variance, while $\Delta\varphi_j^v$ is the phase difference between these oscillations and the signal, which are obtained by solution of the equation

$$A_j^v e^{i\Delta\varphi_j^v} = bk_j^{-1}\sigma\varepsilon K_j e^{i\Theta_j} + b^2k_j^{-1}\sigma\varepsilon G_j e^{i\Phi_j}. \tag{4.14}$$

We also define $G_j e^{i\Phi_j} = 2k_j g_{jj}$ and $G_j^0 = 2k_j g_{jj}|_{\{\omega=0\}}$, where the superscript 0 indicates that the expression is independent of the parameter ω and that it is real. Note that in the above expression, $G_j\cos(\omega t+\varphi+\Phi_j) = G_j\Re[e^{i(\omega t+\varphi+\Phi_j)}] = 2k_j\Re[e^{i(\omega t+\varphi)}g_{jj}]$, where $\Re[z]$ denotes the real part of the complex number z. The functions g_{ij} are given by the solution of the recurrence relation

$$g_{ij} = g_{i-1,j}\frac{k_{i-1}}{k_i+k_j+i\omega} + g_{i,j-1}\frac{k_{j-1}}{k_i+k_j+i\omega} \quad \text{for } i,j = 2,\ldots,L, \tag{4.15}$$

with initial conditions g_{1j} for $j = 1,\ldots,L$ defined as

$$g_{1j} = \prod_{q=1}^{j}\frac{k_{q-1}}{k_1+k_q+i\omega}, \quad \text{where } k_0 = 1. \tag{4.16}$$

For detailed derivations of Eq. (4.13) and the solution of the recurrence relation in Eq. (4.15), we refer the reader to Appendix B.4. One can easily see that the mean stated in Eq. (4.10) and the variance given in Eq. (4.13) are periodic functions in time with the same period, $\tau = 2\pi/\omega$, as the signal function.

In panels (a) and (b) of Fig. 4.2, we show the temporal evolution of the first two moments. In Fig. 4.2(c), we illustrate that the amplitudes of the moments are monotonically decreasing functions of the signal frequency ω. Also, we show that, for some values of the frequency, the moment waves are in phase with the signal, as the phase differences become zero. While these results hold for every stage j in the mRNA life-cycle, we chose to present our plots in (c) for $j = 14$. In addition, it is clear that, while the amplitude of oscillations in the mean is lower than that of oscillations in the variance, the opposite is true for the phase difference of oscillations in the moments with respect to that of the signal, i.e. the variance wave always lags behind the mean wave which itself lags behind the signal. This interesting observation is clarified by a direct comparison of the two waves in Fig. 4.2(d).

One can easily show that if the TF signal frequency is much larger than the hopping rates, i.e. if $\omega \gg k_q$ $(q = 1, \ldots, j)$, then the first two moments of the mRNA distributions are the same as in the case of a time-independent signal ($\varepsilon = 0$),

$$\begin{aligned} \langle n_j \rangle_{\{\varepsilon=0\}} &= \langle n_j \rangle_{\{\omega \gg k_q | q=1,\ldots,j\}} = b k_j^{-1} \sigma, \\ Var(n_j)_{\{\varepsilon=0\}} &= Var(n_j)_{\{\omega \gg k_q | q=1,\ldots,j\}} = b k_j^{-1} \sigma (1 + b G_j^0), \end{aligned} \tag{4.17}$$

which is due to the fact that the amplitudes of the oscillations in the moments are decreasing functions of ω; see Fig. 4.2(c). Note that the time-averaged TF signal, $\int_0^\tau s_b(t)dt/\tau = \sigma$, is the same as the TF signal when $\varepsilon = 0$. Hence, Eq. (4.17) implies that for high signal frequency, the mRNA only senses the constant time-averaged TF signal, which is in agreement with intuition.

We can also compute the Fano factor, which is defined as the ratio of the variance over the mean, and the coefficient of variation squared, defined as the ratio of variance over mean squared. The former is an indicator of how far distributions are from a Poissonian for which the Fano factor is 1, while the latter is a measure of the magnitude of the noise. Analytical expressions for these quantities are obtained from the moment expressions in Eq. (4.10) and Eq. (4.13), and are given by

$$\begin{aligned} FF_j &= 1 + b G_j^0 \frac{1 + \varepsilon \frac{G_j}{G_j^0} \cos(\omega t + \varphi + \Phi_j)}{1 + \varepsilon K_j \cos(\omega t + \varphi + \Theta_j)} \quad \text{and} \\ CV_j^2 &= \frac{k_j}{b\sigma} \frac{1}{1 + \varepsilon K_j \cos(\omega t + \varphi + \Theta_j)} FF_j. \end{aligned} \tag{4.18}$$

In Fig. 4.2(e), we show that for the case of identical hopping rates, with k_j independent of j, the Fano factor has an overall tendency to decrease as the mRNA progresses through its life-cycle, independent of the frequency and amplitude of the signal and of the measurement time, which is due to the factor in front of G_j^0 in Eq. (4.18) being approximately equal to one across parameter space: $FF_j \approx 1 + bG_j^0$. By contrast, in Fig. 4.2(f), we show that while the coefficient of variation squared is monotonically decreasing with life-cycle stage (j) in the absence of an oscillatory signal, in its presence it can increase or decrease with j depending on the signal frequency and the time at which the measurement is taken. This means that the peak position oscillates as the mRNA progresses through its life cycle. All results were verified using the stochastic simulation algorithm (SSA) for the RM – here, we applied a modified version of the Gillespie algorithm, which is described in Appendix B.5. If we associate the cytoplasm with stages j greater than some value, then this observation implies that a signal can either lead to cytoplasmic amplification of transcriptional noise (in the nucleus) or to its attenuation. While the results shown in panels (e) and (f) of Fig. 4.2 assume identical hopping rates, they also qualitatively hold when the hopping rates are non-identical.

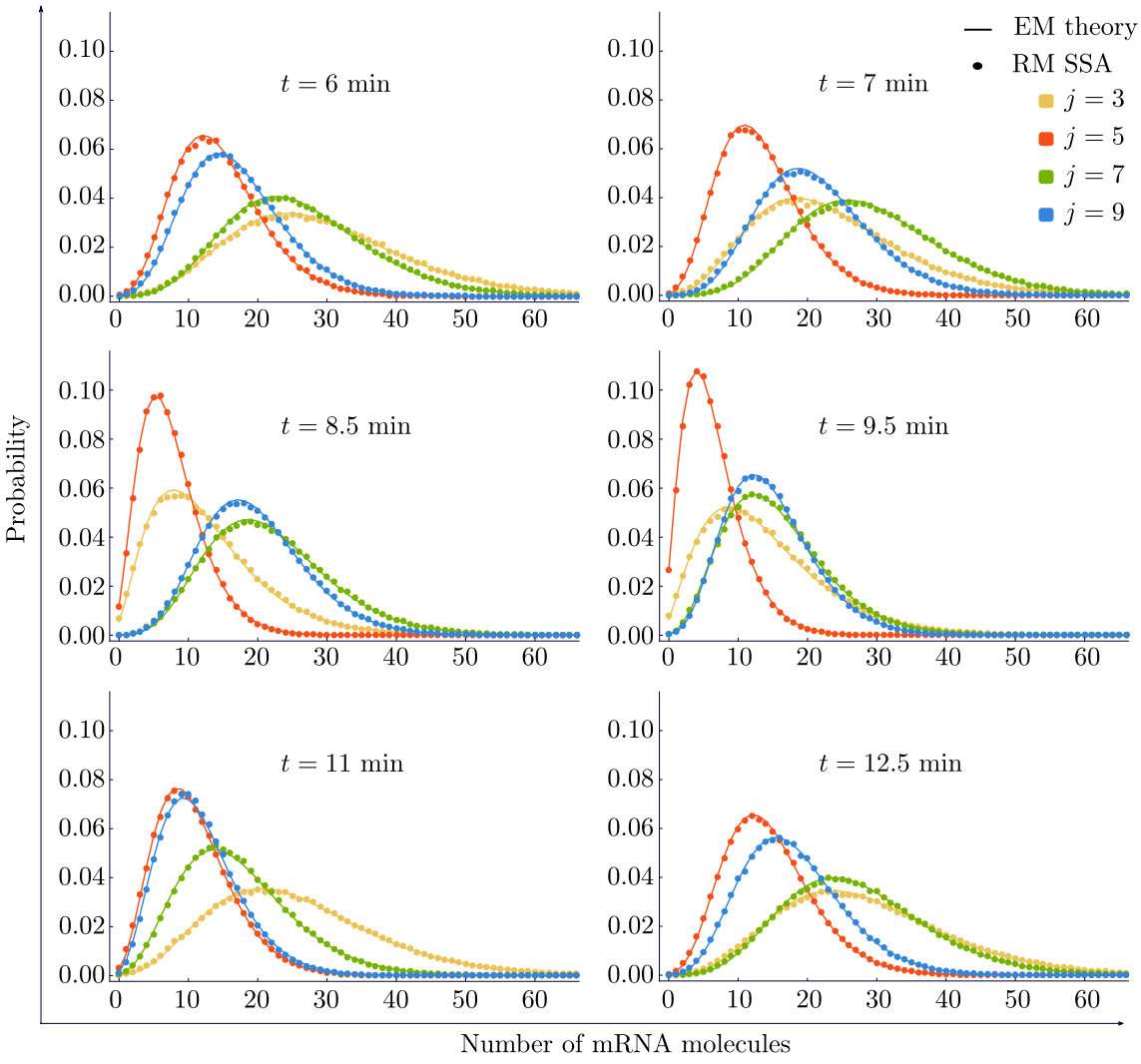


Figure 4.3: *Comparison of the mRNA distributions in the RM with those in the EM.* The mRNA distributions in the RM are not known analytically, and are hence computed from stochastic simulation (SSA; points). The mRNA distributions in the EM are given by Eq. (4.23), together with Eq. (4.21), evaluated in the cyclo-stationary limit and with the constants given in Eq. (4.26) and Eq. (4.27) (solid lines). We show the distributions for four different mRNA life-cycle stages (j) and six time points (t), as stated in the corresponding legends. These time points cover one period of the signal, $\tau = 2\pi$. Note that for $j = 1$, our analytical distribution is exact (not shown), while for states with $j > 1$, the analytical distribution in the EM is a very good approximation to the distribution obtained from simulations of the RM. For the parameter values, see Appendix B.1.

4.3.2 Approximate mRNA distributions for the RM from the EM

While the moments to any order can be derived exactly for the RM, since all propensities are linear, the derivation of an expression for the marginal probability distribution of mRNA in each life-cycle stage proves to be a difficult challenge. Inspired by recent work which approximates the steady state solution of complex models of gene expression by that of simpler models [163], we seek to approximate the time-dependent solution of the multi-variable RM by the solution of the much simpler, one-variable EM. We proceed by finding the exact closed-form time-dependent solution for the mRNA distribution in the EM. In what follows, we assume that j has some fixed value between

1 and L, which is chosen by the user, according to which mRNA life-cycle stage one is interested in. Furthermore, we define n as the number of mRNA molecules of species M_j and $P(n;t)$ as the probability of finding n molecules in the system at time t. Given the reaction scheme of the EM from Eq. (4.7), it follows that the CME of our system is given by

$$\partial_t P(n;t) = \sum_{\bar{m}=0}^{\infty} P_j(\bar{m})\bar{r}_j(t)(\mathbb{E}^{-\bar{m}} - 1)P(n;t) + k_j(\mathbb{E} - 1)nP(n;t), \tag{4.19}$$

where $\mathbb{E}^c[f(n)] = f(n+c)$, with $c \in \mathbb{Z}$, denotes the standard step operator [83]. We define the generating function, $F(u;t) = \sum_{n=0}^{\infty} P(n;t)(u+1)^n$ with $u \in [-1, 0]$, to convert the above equation into the following partial differential equation (PDE):

$$\partial_t F(u;t) + k_j u \partial_u F(u;t) = \bar{r}_j(t)\frac{\bar{b}_j u}{1 - \bar{b}_j u}F(u;t). \tag{4.20}$$

For $\omega \neq 0$, Eq. (4.20) admits the solution

$$\begin{aligned} F(u;t) &= \Big(\frac{1 - \bar{b}_j u\xi_j}{1 - \bar{b}_j u}\Big)^{\frac{\bar{\sigma}_j}{k_j}} \exp\left[\frac{\bar{\sigma}_j\bar{\varepsilon}_j}{\omega}(f_1(t) + f_2(u,t))\right], \quad \text{with} \\ f_1(t) &= \sin(\bar{\varphi}_j) - \sin(\omega t + \bar{\varphi}_j) \quad \text{and} \\ f_2(u,t) &= \Im[e^{i(\omega t + \bar{\varphi}_j)}{}_2F_1(1, i\omega k_j^{-1}, 1 + i\omega k_j^{-1}, \bar{b}_j u)] - \Im[e^{i\bar{\varphi}_j}{}_2F_1(1, i\omega k_j^{-1}, 1 + i\omega k_j^{-1}, \bar{b}_j u\xi_j)], \end{aligned} \tag{4.21}$$

where ${}_2F_1$ is a hypergeometric function of the second kind [44, 63] and we have defined $\xi_j = e^{-k_j t}$; moreover, $\Im[z]$ denotes the imaginary part of a complex number z. For the case of $\omega = 0$, the solution of Eq. (4.20) is given by

$$F(u;t) = \Big(\frac{1 - \bar{b}_j u\xi_j}{1 - \bar{b}_j u}\Big)^{\frac{\bar{\sigma}_j(1 + \bar{\varepsilon}_j\cos(\bar{\varphi}_j))}{k_j}}. \tag{4.22}$$

See Appendix B.6 for a detailed derivation of these solutions. We note that the expressions in Eq. (4.21) and Eq. (4.22) are both well defined: first, $(1 - \bar{b}_j u) > 0$ due to $u \in [-1, 0]$, while $\bar{b}_j$ is some positive parameter; also, $(1 + \bar{\varepsilon}_j\cos(\bar{\varphi}_j)) > 0$ for $\bar{\varepsilon}_j < 1$, which follows from the definition in Eq. (4.26) below. The time-dependent marginal distribution is then found by using the formula

$$P(n;t) = \frac{1}{n!}\frac{d^n}{du^n}F(u;t)\Big|_{\{u=-1\}}. \tag{4.23}$$

Here, we note that to obtain the solution in the cyclo-stationary limit, we merely need to set $\xi_j = 0$ – in that limit, the solution (for $\omega > 0$) will still be time-dependent. Note also that if the production rate is constant, i.e. if $\bar{\varepsilon}_j = 0$ or $\omega = 0$, then the solution reduces to a simple Negative Binomial (NB) distribution, which has been previously reported in [31, 188]. By contrast, when the production rate is time-dependent, the distribution of the mRNA species is not NB and can even be bimodal for some parameter values; see the supplementary Fig. B.6.1.

Having closed-form expressions for the mRNA distributions in the EM is not sufficient to approximate the mRNA distributions in the RM, since the parameters $\bar{b}_j$, $\bar{\sigma}_j$, $\bar{\varepsilon}_j$, and $\bar{\varphi}_j$ are unknown. We now seek to obtain analytical expressions for these unknown parameters by matching our expressions for the mean and the variance in the EM with the RM. From the solution for the generating function given in Eq. (4.21), it is straightforward to show that the first two moments from the EM in the cyclo-stationary limit can be written as

$$\begin{aligned} \langle n_j\rangle_E &= \bar{b}_j\bar{\sigma}_j k_j^{-1}(1 + \bar{\varepsilon}_j K_j^*\cos(\omega t + \bar{\varphi}_j + \Theta_j^*)) \quad \text{and} \\ Var(n_j)_E &= \langle n_j\rangle_E + \bar{b}_j^2\bar{\sigma}_j k_j^{-1}(1 + \bar{\varepsilon}_j G_j^*\cos(\omega t + \bar{\varphi}_j + \Phi_j^*)), \end{aligned} \tag{4.24}$$

where the subscript E refers to the EM and the newly defined parameters are given by

$$K_j^* = \frac{k_j}{\sqrt{k_j^2+\omega^2}}, \quad \Theta_j^* = -\tan^{-1}\left(\frac{\omega}{k_j}\right), \quad G_j^* = \frac{2k_j}{\sqrt{(2k_j)^2+\omega^2}}, \quad \text{and} \quad \Phi_j^* = -\tan^{-1}\left(\frac{\omega}{2k_j}\right). \tag{4.25}$$

Our goal is to match the moments of mRNA distributions from the RM, as stated in Eq. (4.10) and Eq. (4.13), with the moments given in Eq. (4.24). Because of the complicated form of these analytical expressions, there might be more than one way of matching the moments; here, we present the most straightforward one that serves our purpose. First, we exactly match the means by setting $\langle n_j \rangle = \langle n_j \rangle_E$; by inspection, it is easy to verify that matching can be achieved by taking the constants in the EM to read

$$\bar{b}_j \bar{\sigma}_j = b\sigma, \quad \bar{\varepsilon}_j = \varepsilon \frac{K_j}{K_j^*} = \varepsilon Y_j, \quad \text{and} \quad \bar{\varphi}_j = \varphi + \Theta_j - \Theta_j^* = \varphi + \Omega_j. \tag{4.26}$$

Given these parameters, it is not possible to match exactly the variances of the RM and the EM, except when $j = 1$ – the latter being obvious from an inspection of the reaction schemes of both models – or else when the TF signal is not time-dependent, i.e. when $\varepsilon = 0$ or $\omega = 0$. (That case was studied in [163].) Note that in the limit of very large frequency ω – taken to be much larger than the hopping rates – one can also exactly match the second moments in the two models, which is due to the mRNA distributions in the RM being a function of the time-averaged TF signal only in that limit, as noted already. For the general case where $j > 1$, $\varepsilon > 0$, and $\omega > 0$, one can match the time-independent terms in the expressions for the variance in the RM and the EM, which leads to the following additional constraints:

$$\bar{b}_j = bG_j^0 \quad \text{and} \quad \bar{\sigma}_j = \sigma (G_j^0)^{-1}. \tag{4.27}$$

In summary, our approximate solution of the RM is given by Eq. (4.21), with the constants as defined in Eq. (4.26) and Eq. (4.27).

In panels (a) and (b) of Fig. 4.2, we compare the time evolution of the exact mean and variance in the RM (numerical solution of Eq. (4.9) and Eq. (4.12)) with the time evolution of the mean and variance, as computed from the EM model in the cyclo-stationary limit, Eq. (4.24). We observe excellent agreement between the two for various life-cycle stages. In Fig. 4.3, we verify that the distribution in the RM, as computed from stochastic simulation, is also in good agreement with the approximate distribution in the EM that is computed from the generating function given in Eq. (4.21), evaluated in the cyclo-stationary limit of $\xi_j = 0$.

4.3.3 Accuracy of the EM approximation

Next, we seek to investigate in detail the accuracy of the approximation to the RM that is provided by the EM. For each mRNA life-cycle stage j, we define the vector $\vec{p} = (p_1, \ldots, p_k)$ whose i-th entry p_i is the probability of observing i mRNA molecules according to the EM; that probability can be determined from the generating function, Eq. (4.21). (Note that k is some integer which is assumed large enough such that p_k is very small.) Similarly, we define a vector $\vec{q} = (q_1, \ldots, q_k)$ whose i-th entry q_i is the probability of observing i mRNA molecules according to the RM. The Hellinger distance (HD) between the probability distributions in the EM and the RM is then defined as

$$HD = \left[\frac{1}{2}\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2\right]^{\frac{1}{2}}. \tag{4.28}$$

This measure of discrepancy between models, while ideal due to being based on the probability distributions, cannot be calculated analytically, since we do not have the exact analytical distribution for the RM. Hence, it can only be computed from stochastic simulation.

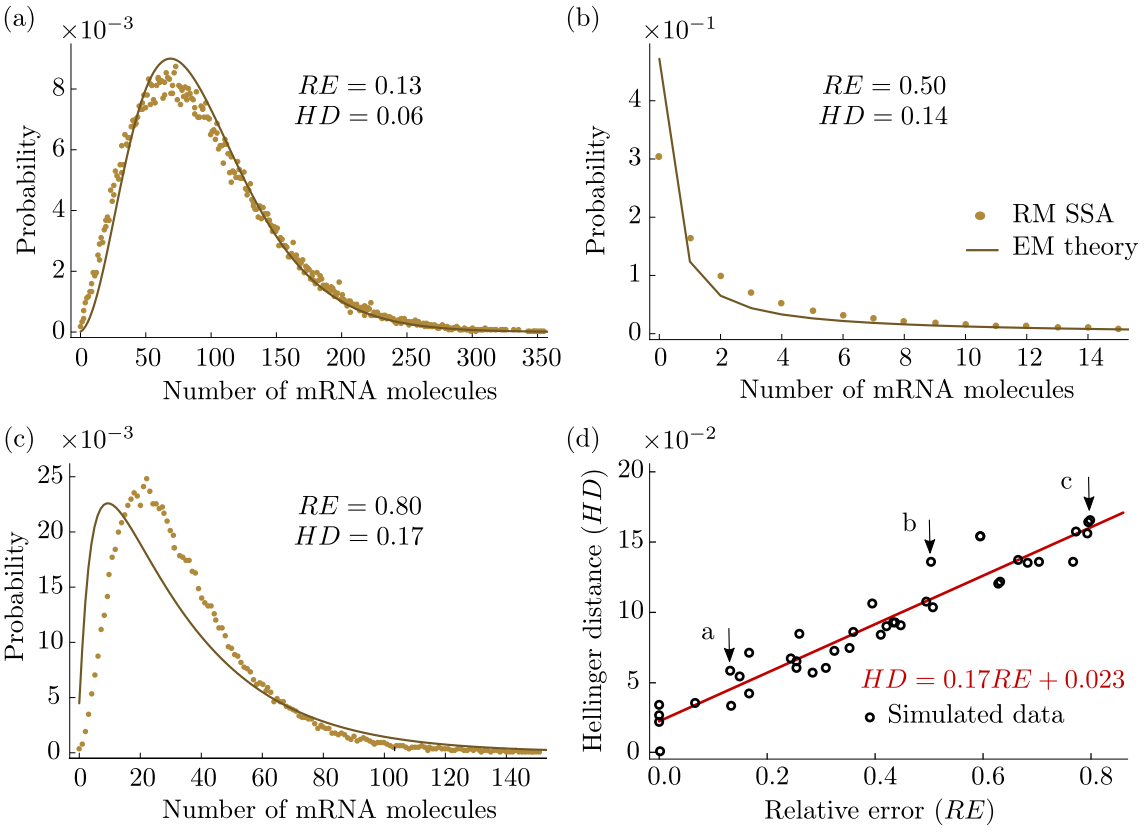


Figure 4.4: *Accuracy of the EM approximation.* In panels (a) through (c), we compare the distribution of mRNA numbers in the EM (Eq. (4.23) together with Eq. (4.21) and constants given in Eq. (4.26) and Eq. (4.27), computed in the cyclo-stationary limit; solid lines) and the RM (from stochastic simulation via the SSA; points). We also compute the HD between the two distributions and the RE of the variance of mRNA numbers in the EM. Comparison of (a) through (c) suggests that the HD increases linearly with the RE; this relationship is confirmed in panel (d) for 40 points and a fitted line $HD = 0.17RE + 0.023$ that is obtained by linear regression. We used Eq. (4.28) to calculate the HD and Eq. (4.29) to calculate the RE. For the parameter values, see Appendix B.1.

A different, analytical measure of the discrepancy between the RM and the EM is given by the relative error (RE) between the cyclo-stationary variance of the mRNA species predicted by both models. Using Eqs. (4.13) and (4.24), for each stage j we define the RE as

$$RE = \frac{|Var(n_j) - Var(n_j)_E|}{|Var(n_j)|} = \frac{b\varepsilon|G_j \cos(\omega t + \varphi + \Phi_j)) - G_j^0 Y_j G_j^* \cos(\omega t + \varphi + \Omega_j + \Phi_j^*))|}{1 + bG_j^0 + \varepsilon[K_j \cos(\omega t + \varphi + \Theta_j)) + bG_j \cos(\omega t + \varphi + \Phi_j)]}. \tag{4.29}$$

In Fig. 4.4, we investigate the relationship between the HD and the RE for different stages in the mRNA life-cycle across a wide range of parameter values. Three different points in parameter space, shown in panels (a) through (c) of Fig. 4.4, suggest that there is a linear relationship between the HD and the RE. This relationship between the two measures is confirmed in Fig. 4.4(d). Since we have an expression for the RE, it is hence easy to say when the EM provides a useful and accurate approximation of the distribution in the RM. We remark that the simple relationship between the HD and the RE is particular to the model under investigation since one would generally expect the HD to depend on moments of order higher than two.

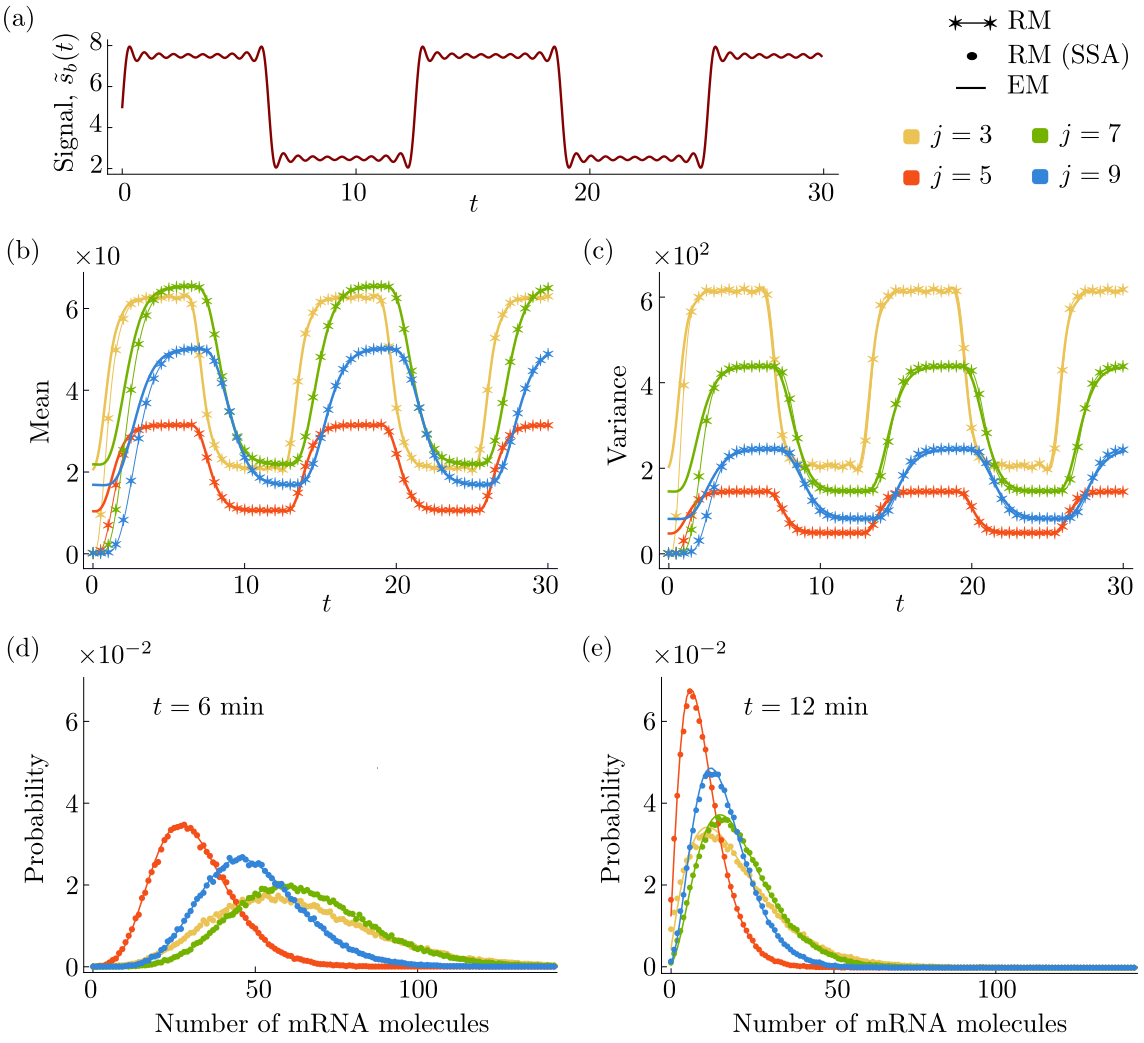


Figure 4.5: *Generalization to the case of a general time-dependent activation signal.* (a) A square wave-like time-dependent signal given by Eq. (4.30), where we specify $A_n = -2/(\pi n)$ and $N = 9$. In panels (b) and (c), we compare the time evolution of the mean and the variance of the mRNA distributions in the RM (numerical solution of Eq. (4.9) and Eq. (4.12) with $r(t) = \tilde{s}_b(t)$ given by Eq. (4.30); thin lines with asterisks) with those in the EM as given by Eq. (4.34) (thick solid lines). In (d) and (e), we compare the mRNA distributions at time points $t = 6$ min and $t = 12$ min, respectively. The distributions are obtained from the SSA for the RM (points) and by Eq. (4.23) together with Eq. (4.32) and Eq. (4.36) for the EM (solid lines). The results in panels (b) through (e) are shown for four different stages in the mRNA life-cycle (j), as stated in the legends. For the parameter values, see Appendix B.1.

4.4 Generalization to the case of an arbitrary activation signal

Thus far, we have considered the approximation of the RM by the EM for the case when the activation rate from the inactive to the active state is given by $s_b(t) = \sigma(1 + \varepsilon \cos(\omega t + \varphi))$. That approximation can be justified for the case of nuclear TF numbers varying in a sinusoidal manner assuming that the rate of switching is directly proportional to TF numbers, i.e. that there is no

cooperativity. Of course, generally one expects the activation rate to have a much more complex time dependence, which is principally due to two factors: (i) environmental stimuli are coupled to gene expression via modulation of the number of TFs in the nucleus [168] – such stimuli will generally change in a complex time-varying manner, which will be reflected in TF numbers; (ii) binding of TFs to DNA is often cooperative [175], which implies that the rate of gene activation can be highly nonlinear in TF numbers via a Hill function dependence. To incorporate both of these factors, in this section we extend our results to the case of a general time-dependent activation rate that can be represented as a truncated Fourier series:

$$\tilde{s}_b(t) = \sigma\Big(1 + \sum_{n=1}^{N} A_n \cos(\omega n t + \varphi_n)\Big); \tag{4.30}$$

here, $N \in \mathbb{N}_+$ and $A_n \in \mathbb{R}$ for all $n \in \{1, \ldots, N\}$, where A_n is such that $\tilde{s}_b(t) \geq 0$. The RM is now given by Eq. (4.3) with $r(t) = \tilde{s}_b(t)$. We hypothesize that the distribution of each state M_j in the RM can be well approximated by a distribution in the EM defined in Eq. (4.7), where the production rate is now given by

$$\bar{r}_j(t) = \bar{\sigma}_j\Big(1 + \sum_{n=1}^{N} \bar{A}_{n,j} \cos(\omega n t + \bar{\varphi}_{n,j})\Big). \tag{4.31}$$

Note that the unknown parameters in this case are $\bar{b}_j$, $\bar{\sigma}_j$, $\bar{A}_{n,j}$, and $\bar{\varphi}_{n,j}$, which we will determine below. All our derivations are performed in the cyclo-stationary limit of $t \to \infty$. The exact closed-form expression for the probability-generating function in the EM when $\omega \neq 0$ is given by

$$\begin{aligned}
F(u;t) &= \Big(\frac{1-\bar{b}_j u\xi_j}{1-\bar{b}_j u}\Big)^{\frac{\bar{\sigma}_j}{k_j}} \exp\Big[\bar{\sigma}_j \sum_{n=1}^{N} \frac{\bar{A}_{n,j}}{\omega n}(f_1(t) + f_2(u,t))\Big], \quad \text{with} \\
f_1(t) &= \sin(\bar{\varphi}_{n,j}) - \sin(\omega n t + \bar{\varphi}_{n,j}) \quad \text{and} \\
f_2(u,t) &= \Im[e^{i(\omega n t + \bar{\varphi}_{n,j})}{}_2F_1(1, i\omega n k_j^{-1}, 1 + i\omega n k_j^{-1}, \bar{b}_j u)] \\
&\quad - \Im[e^{i\bar{\varphi}_{n,j}}{}_2F_1(1, i\omega n k_j^{-1}, 1 + i\omega n k_j^{-1}, \bar{b}_j u\xi_j)],
\end{aligned} \tag{4.32}$$

while for $\omega = 0$, the solution reads

$$F(u;t) = \Big(\frac{1-\bar{b}_j u\xi_j}{1-\bar{b}_j u}\Big)^{\frac{\bar{\sigma}_j}{k_j}\left(1+\sum_{n=1}^{N} \bar{A}_{n,j}\cos(\bar{\varphi}_{n,j})\right)}. \tag{4.33}$$

Here, ${}_2F_1$ is the hypergeometric function of the second kind, as before. The expressions in Eq. (4.32) and Eq. (4.33) are again well defined due to $u \in [-1, 0]$, $\bar{b}_j > 0$, and $\big(1+\sum_{n=1}^{N} \bar{A}_{n,j}\cos(\bar{\varphi}_{n,j})\big)$ being positive for suitably chosen $\bar{A}_{n,j}$, which follows from the definition in Eq. (4.36) below. In order to derive analytical expressions for the unknown parameters, we follow the exact same steps in our mathematical analysis as in Section 4.3. First, we find the moments of the mRNA distributions in each life-cycle stage j for the RM and the EM. These are given by

$$\begin{aligned}
\langle n_j\rangle &= b\sigma k_j^{-1}\Big(1 + \sum_{n=1}^{N} A_n K_{n,j} \cos(\omega n t + \varphi_n + \Theta_{n,j})\Big), \\
\langle n_j\rangle_E &= \bar{b}_j\bar{\sigma}_j k_j^{-1}\Big(1 + \sum_{n=1}^{N} \bar{A}_{n,j} K^*_{n,j} \cos(\omega n t + \bar{\varphi}_{n,j} + \Theta^*_{n,j})\Big), \\
Var(n_j) &= \langle n_j\rangle + b^2\sigma k_j^{-1}\Big(G_j^0 + \sum_{n=1}^{N} A_n G_{n,j} \cos(\omega n t + \varphi_n + \Phi_{n,j})\Big), \quad \text{and} \\
Var(n_j)_E &= \langle n_j\rangle_E + \bar{b}_j^2\bar{\sigma}_j k_j^{-1}\Big(1 + \sum_{n=1}^{N} \bar{A}_{n,j} G^*_{n,j} \cos(\omega n t + \bar{\varphi}_{n,j} + \Phi^*_{n,j})\Big),
\end{aligned} \tag{4.34}$$

where the subscript E refers to the EM and the definition of the new parameters is as follows:

$$\begin{aligned} K_{n,j} &= \prod_{q=1}^{j} \frac{k_q}{\sqrt{k_q^2 + n^2\omega^2}}, \qquad K^*_{n,j} = \frac{k_j}{\sqrt{k_j^2 + n^2\omega^2}}, \qquad G^*_{n,j} = \frac{2k_j}{\sqrt{(2k_j)^2 + n^2\omega^2}}, \\ \Theta_{n,j} &= -\sum_{q=1}^{j} \tan^{-1}\Big(\frac{\omega n}{k_q}\Big), \quad \Theta^*_{n,j} = -\tan^{-1}\Big(\frac{\omega n}{k_j}\Big), \quad \text{and} \quad \Phi^*_{n,j} = -\tan^{-1}\Big(\frac{\omega n}{2k_j}\Big). \end{aligned} \tag{4.35}$$

Also, we define $G_{n,j}e^{i\Phi_{n,j}} = 2k_j g_{n,jj}$ with $g_{n,jj} = g_{jj}|_{\omega\mapsto\omega n}$, where g_{ij} is the solution of the recurrence relation in Eq. (4.15). By matching the moments in Eq. (4.34) in the same manner as in Section 4.3, we find the unknown parameters to be given by

$$\bar{b}_j = bG_j^0,\ \bar{\sigma}_j = \sigma(G_j^0)^{-1},\ \bar{A}_{n,j} = A_n\frac{K_{n,j}}{K^*_{n,j}} = A_nY_{n,j},\ \text{and}\ \bar{\varphi}_{n,j} = \varphi_n + \Theta_{n,j} - \Theta^*_{n,j} = \varphi_n + \Omega_{n,j}. \tag{4.36}$$

The approximate mRNA distribution in each mRNA life-cycle in the cyclo-stationary limit can be obtained by using $P(n;t) = \frac{1}{n!}\frac{d^n}{du^n}F(u;t)\big|_{\{u=-1\}}$ and Eq. (4.32), with parameters given as in Eq. (4.36).

In Fig. 4.5, we provide verification of the accuracy of the resulting generalized version of the EM by means of stochastic simulation. Here, we have used an approximately square wave for the time-dependent activation rate, as shown in Fig. 4.5(a), to model sharp TF pulses as considered in earlier work [186]. In panels (b) and (c) of Fig. 4.5, we show that the moments of mRNA distributions in the RM for four different stages in the mRNA life-cycle are well approximated by the moments of the EM. In panels (d) and (e) of Fig. 4.5, we verify that the approximate mRNA distributions in the EM are in excellent agreement with the mRNA distributions in the RM, which were computed using the SSA for two different time points.

4.5 Summary and discussion

In this study, we have considered a model for bursty transcription that is coupled to a time-varying extracellular stimulus, and we have applied a novel approximation to obtain the time-dependent distributions of mRNA in any life-cycle stage of interest. These stages often correspond to particular sub-cellular localization; hence, the model predicts, for example, the mRNA distribution at the transcription site, elsewhere in the nucleus, and in the cytoplasm – data that is accessible using experimental techniques [48, 182]. We have shown that the resulting approximate distributions are in excellent agreement with stochastic simulation. In addition, we have found that the relative error between the true and the approximate variance of mRNA fluctuations – which can be calculated analytically – is directly proportional to the Hellinger distance between the distributions obtained from simulations and the theoretical ones. (The Hellinger distance is not accessible analytically, but only via simulation.) That relationship provides a convenient means to assess the accuracy of our theory without simulation.

Further, we have shown that apparent bimodality in the mRNA distributions can be generated by a time-varying stimulus when the expression is bursty, which is interesting, considering that the solution of models of bursty expression without a stimulus gives a unimodal Negative Binomial distribution [126]. The intuition behind this phenomenon is clear, however: if the switching rate to the active transcription state is controlled by an oscillatory signal, then there are periods of intense transcription when the signal is very strong, whereas transcription almost switches off when the signal is weak. Our theory shows that if the stimulus is periodic, then (i) the oscillations in the variance of mRNA fluctuations lag behind those in the mean; (ii) the amplitude of oscillations in the first two moments decreases monotonically with the frequency; (iii) the Fano factor of mRNA fluctuations tends to decrease with life-cycle stage; (iv) the noise in mRNA fluctuations, as quantified by the coefficient of variation squared, can increase or decrease with life-cycle stage depending on the time of measurement and the frequency of the stimulus. The latter implies that

the stimulus can either lead to apparent amplification of the noise in the cytoplasm compared to that in the nucleus, or to the opposite case of attenuation. We are not aware of experimental data that can verify these predictions, since observations reported in the literature were made in the absence of a time-varying stimulus [74, 182].

We note that other theoretical studies have sought to derive closed-form time-dependent mRNA distributions for various models of gene expression. These can be classified as follows: (i) those which do not consider a time-varying stimulus [71,109,126,191–193], in that they study how gene expression approaches a steady state given a perturbation that is applied at a point in time, e.g. with the initial condition given by mRNA numbers following cell division – in that case, the kinetic rates do not vary with time; (ii) those which consider a stimulus that varies with time [178–181]. The major difference between our work and the latter is that we derive time-dependent analytical distributions for the mRNA at any stage of its life-cycle, which often correspond to specific subcellular localization. For example, if we choose the simplest case of $L = 2$ in our model, then the distributions of M_1 and M_2 can be interpreted as being for nuclear and cytoplasmic mRNA, which would be under the assumption that the time from initiation to a mature mRNA appearing in the nucleus is exponentially distributed, as is the time for export from the nucleus to the cytoplasm. Deviations from the exponential assumption can also be easily incorporated into our framework. For example, if the distribution of the export time is Erlang with shape parameter k, then one could apply our model with $L = k + 3$, where M_1 is nuclear mRNA, M_{k+3} is the cytoplasmic mRNA, and $M_2, \ldots, M_{k+2}$ are dummy species introduced to capture the Erlang distributed delay. Another interesting application of our model would be to predict the distribution of bound RNA polymerase (RNAP) along the gene, in response to a time-varying stimulus; in that case, under the assumption that volume exclusion is not significant, the species M_i can be interpreted as the RNAP on gene segment i [123]. Besides the forward predictive power of our theory, the practical use thereof might lie in the resulting theoretical distributions, together with likelihood-based inference methods [86, 194], as a reliable means for estimating kinetic parameters from experimental population snapshot data of nascent, nuclear, and cytoplasmic mRNA measured for time-varying extracellular stimuli.

Code Availability SSA code for simulating the reduced model (RM) is available at `https://github.com/TatianaFil`.

4.6 Using modified RM to study the regulation of *hb* gene by Bcd TF

How transcription factors quantitatively control gene expression has been a central question in molecular biology in the past decades. There is a specific class of TFs called morphogens, which are molecules that act as dose-dependent regulators of cell signalling and gene expression. In general, morphogens are substances participating in the morphogenesis process of organisms. For instance, during the first two hours of *Drosophila* melanogaster (often called "small fruit flies") embryo development, a spatially uniform arrangement of identical cells is patterned by graded distributions of morphogens, which play a key role in the establishment of the body plan of the adult fly [174, 195]. For example, the anterior-posterior concentration gradient of Bicoid (Bcd) TF is essential for specifying the anterior segments of the body [195]. The gradient of Bcd is established rapidly (approximately 1 hr after fertilization) and it has been shown that the Bcd concentration decays approximately exponentially with the distance from the anterior pole of the embryo; this exponential shape is consistent with the gradient being formed by a balance of localized synthesis, diffusion, and spatially uniform degradation [174, 196], please see Fig. 4.6(A). There are a number of studies that have attempted to model Bcd gradient, some of which can be found in [171–176,197]. The origin of Bcd pattern formation can be traced to the localized spatial pattern of bcd mRNA, which is deposited during oogenesis at the anterior pole of the egg and is translated soon after fertilization [196]. When the egg is fertilized, there is one single nucleus which has to undergo cleavage. After 14 nuclear division cycles (around 4 hours after fertilization), the embryo consists of approximately 6000 cells. Typically, both nuclear and cytoplasmic Bcd concentrations develop in

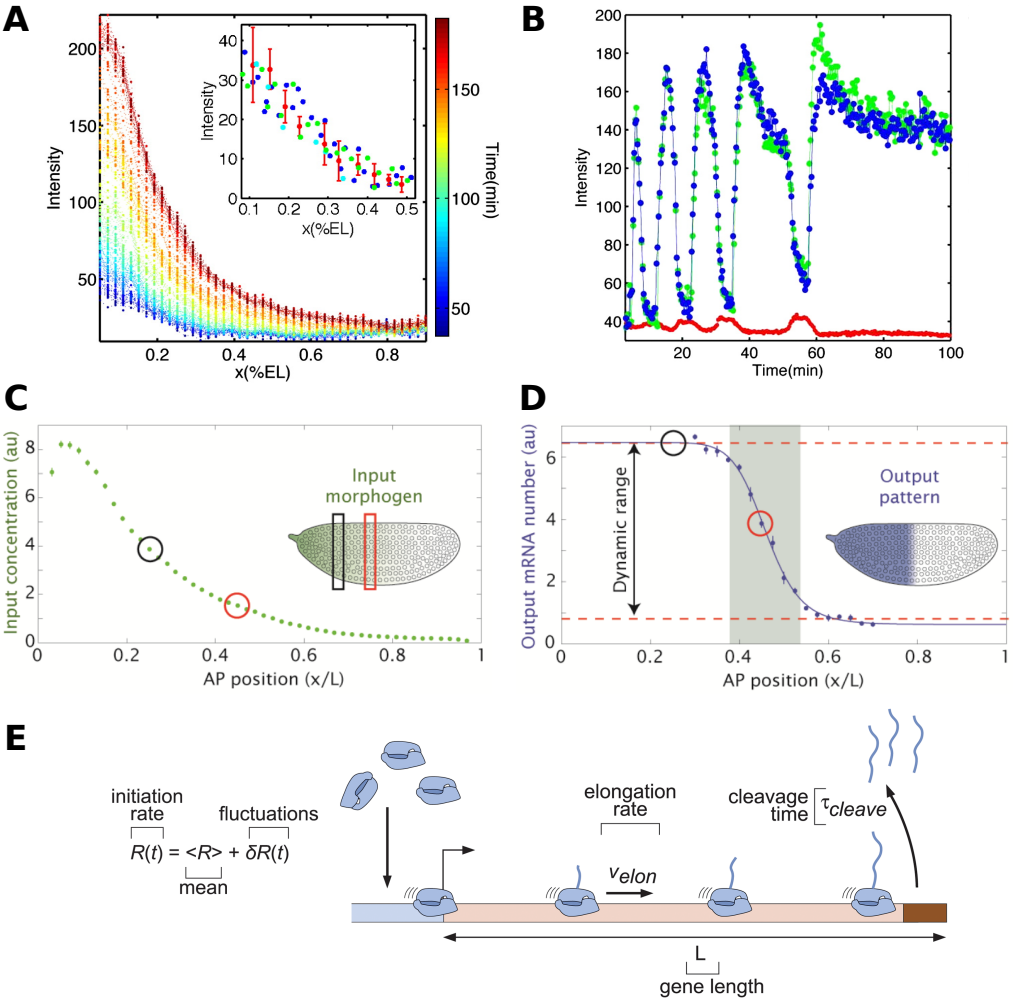


Figure 4.6: **Regulation of *hunchback* gene by the Bicoid transcription factor**. (**A**) Bcd-GFP fluorescence profiles (y-axis; Intensity) projected on the egg's anterior–posterior axis (x-axis; normalised embryo length – EL). Inset shows nuclear Bcd gradients in nuclear cycles 11 (cyan), 12 (red), 13 (green), and 14 (blue) projected on the anterior-posterior axis in the anterior half of the embryo (red error bars for nuclear cycle 12 are over five consecutive time points). Figure is taken from [196]. (**B**) Typical nuclear and cytoplasmic development of Bcd-GFP concentration. Each data point corresponds to the concentration of a single nucleus at a given time point. Blue and green traces follow two individual nuclei, and the red curve corresponds to the average concentration in the interstitial space between the nuclei (cytoplasm). Figure is taken from [196]. (**C-D**) A space-varying input morphogen (C, data shown for Bcd-GFP) determines an output pattern of accumulation of its target gene (D, measured by counting accumulated transcripts). Figure is taken from [199]. (**E**) A simple model of the transcription cycle, incorporating nascent RNA initiation, elongation, and cleavage. Figure is taken from [198].

an oscillatory way during the time of nuclear cycles 9 to 14, please see Fig. 4.6(B). Approximately after 90 mins of egg fertilization ($\approx$ 10 nuclear cycles), the Bcd determines an output pattern of its target genes (also known as gap genes), by binding to their promoter and inducing their expression in a concentration-dependent manner. One of the Bcd gap genes is *hunchback* (*hb*); the space-patterned input Bcd and output *hb* is presented in Fig. 4.6(C-D). Some of the studies that investigate how Bcd regulates *hb* in the early *Drosophila* embryo can be found in [175, 197–199].

Specifically, the work performed by J. Liu et al. in [198] inspired us with the idea that a slightly modified version of our stochastic reduced model (RM) presented in this chapter can also be used to study the regulation of *hb* by Bcd transcription factor. In this work, J. Liu and collaborators developed a method to analyse live-imaging data from the MS2 and PP7 experimental techniques in order to dynamically characterize the steps of initiation, elongation and cleavage of the full transcription cycle of the *hb* reporter gene at single-cell resolution. Also, they used a theoretical model of gene expression where initiation occurs at a time-dependent rate. After initiation, the polymerase molecules move along the gene at a constant speed and after reaching the end of the gene, there follows a deterministic cleavage time, after which the nascent transcript is cleaved; please see Fig. 4.6(E). Finally, the authors used a novel application of the Bayesian inference technique of Markov Chain Monte Carlo to simultaneously infer the effective parameters of average initiation rate, elongation speed, and cleavage time as functions of the position along the embryo. They validated their approach by comparing the inferred average initiation and elongation rates with previously reported results. Hence, our idea is the following. We propose a stochastic model described by the following reactions:

$$\varnothing \xrightarrow{r(t)} M_1, \quad M_j \xrightarrow{k} M_{j+1} \quad \text{for} \quad j = 1, \ldots, L-1, \qquad \text{and} \quad M_L \xrightarrow{k_{cl}} N.$$

This model is very similar to our RM described in Eq. 4.6 (and also to our detailed model from Fig. 3.1 with certain differences: (i) the transcription initiation happens with rate $r(t) = s_b(t)$ and it is not bursty because there is evidence indicating that the *hb* promoter spends almost 53% in its active state [175], (ii) M_j represents polymerases on gene segment, j for $j = 1, \ldots, L$ and k is stochastic elongation rate; then elongation time T is Erlang distributed with shape parameter $L-1$ and rate k, (iii) cleavage of nascent RNA (denoted by N) occurs with exponentially distributed rate, k_{cl}. This model could be used in a same way as in [198] for inferring the parameters $r(t)$, k, and k_{cl} as functions of the position along the embryo. The advantage of this model is that the elongation and cleavage times are stochastic variables, which is more realistic than the deterministic version in [198]. Most importantly, this model is analytically tractable, mathematical analysis of which can be performed by following the same steps as for the RM. One can easily obtain analytical expressions for time-dependent polymerase distributions at each gene segment, and these can serve as a valuable tool for inferring the kinetic parameters of the model from experimental data.

Chapter 5

Conclusions and Outlook

Gene expression is a fundamentally stochastic process, with randomness in transcription, translation and degradation, leading to significant cell-to-cell variations (noise) in mRNA and protein levels. There are a large number of research studies based on mathematical models that intend to shed light on the mechanisms that cause fluctuations in the number of molecules of gene products during their life-cycle. This thesis aims to contribute to this research field and presents work on the construction and the analysis of novel stochastic models of gene expression. The development and application of new approximation techniques for obtaining analytical distributions of molecule numbers are also important aspects of this thesis.

In Chapter 1 we presented an introduction to mRNA life-cycle, and we gave a brief overview of GFP, FISH, and electron microscopy experimental methods that can estimate the number of gene products. Afterwards, we included an extended discussion about the mechanisms that control gene expression; we noted that the stochastic processes of transcriptional bursting, transcription initiation, elongation, and nuclear retention are important steps in transcription that regulate noise in gene expression. Additionally, we presented some widely used mathematical models of stochastic gene expression (e.g., two-stage and three-stage models) and the general steps of mathematical analysis for solving their CME. We gave a thorough description of mathematical tools and methods, such as LNA and GSPT, and explained their important role in studying stochastic models. Finally, we gave an example of a detailed analysis of a stochastic model of transcription with nuclear retention; this serves as a good preliminary task before proceeding to the analysis of more complex models in the subsequent chapters.

In Chapter 2 we demonstrated our study of a stochastic multi-scale transcriptional bursting model, which was found to be consistent with experimental data for mRNA dynamics. This model is characterised by a promoter that fluctuates between three states, while it incorporates details about polymerase recruitment and polymerase pause release cellular processes. We performed our mathematical analysis of this model by applying all the methods described in the introductory chapter, and we obtained analytical expressions for both mRNA and protein distributions. We showed, under which certain biological conditions, our multi-scale models can be reduced to the conventional three-stage model. Generally, there is evidence that the timescales of promoter switching between active and inactive states are slower than the timescales of polymerase recruitment and pause release for many genes, and hence, we eliminated the latter and included other biological details such as elongation and polymerase pausing in the model we studied next.

In Chapter 3 we constructed a new model of stochastic multi-step elongation process. We divided our study in this chapter into two parts; in the first part, we analysed a model without polymerase pausing, while in the second part, we extended our study to a model that incorporates stochastic pausing and unpausing processes. Both models are characterised by a promoter that can randomly switch between two states, transcriptionally active and inactive, and for both models, we studied the fluctuations in the number of nascent (actively transcribing polymerase) and mature

RNA molecules. We obtained approximate steady-state distributions for both species of interest. We showed that the distribution of the total number of polymerases can be well described by a negative binomial distribution (which does not present bimodality) in certain biological limits, while the distribution of mature mRNA is given in terms of a hypergeometric function (can present bimodality). Eventually, we discussed various possible extensions and improvements to this model, where the dependence of promoter activation on the concentration of transcription factors is one of them. The effects of temporal variation in transcription factor abundance on the mRNA life-cycle are the subject of our work that followed.

In Chapter 4 we developed a novel model of gene expression and a novel theoretical approach for obtaining analytical expressions for mRNA distributions. In our new model, we divided the mRNA life-cycle into multiple stages, where the initial stages represent nuclear mRNA and the remaining stages represent cytoplasmic mRNA. The promoter of the gene is characterised by two states, where the activation rate is dependent on a time-dependent signal due to transcription factors present in the nucleus. Mathematical analysis of the described model appeared to be a difficult task and hence, we studied a reduced model, which is equivalent to the full model when the mRNA transcription is bursty. In the reduced version, the transcription initiation rate is time-dependent and incorporates signal dynamics. We obtained approximate distributions for mRNA at each life-cycle stage by using an effective telegraph model. This approximation required matching of mean and variance of the mRNA distributions obtained for the reduced model with the ones obtained from the effective model.

Although we aimed to progressively introduce more biological details in our sequential studies in this thesis, there are various important biological mechanisms that we have not considered; however, we acknowledge the tremendous amount of computational, theoretical, and experimental work that has been performed in the past decades for understanding the effects of these mechanisms on gene expression. Next, we are going to list some studies on biological properties that extend beyond the work in this thesis.

Even though we were unconcerned about developing a model with **feedback loop** for this thesis, we consider it important to note the significance of the autoregulatory feedback loops. It has been shown that autoregulation is the most basic kind of feedback loop, where the stochastic models of autoregulation are based on the following mechanism: a protein synthesized from a gene activates or suppresses its own production. These lead to positive or negative feedback loops, respectively. A detailed review of stochastic modelling of autoregulatory genetic feedback loops has been conducted by J. Holehouse and collaborators in [200]. An interesting fact that is also stated in this review is that many biological systems utilize a combination of positive and negative feedback loops; it has been estimated that 40% of all transcription factors in *Escherichia coli* self-regulate, while most of them participate in autorepression.

The other two important biological mechanisms which have not been considered in our work are the **cell-cycle** and **gene-replication**. During gene replication in eukaryotic cells, the copy number of each gene doubles from two to four and at the end of the cell-cycle, mature mRNA molecules are partitioned between the two daughter cells. Skinner et al. state in [32] the following fact: *"To infer transcription kinetics for a gene of interest, researchers commonly compare the distribution of mRNA copy-number to the prediction of a theoretical model. However, the reliability of this procedure is limited because the measured mRNA numbers represent integration over the mRNA lifetime, contribution from multiple gene copies, and mixing of cells from different cell-cycle phases."* In this study, Skinner et al. simultaneously quantified nascent and mature mRNA in single cells by using smFISH, and the obtained data were fitted to numerical model predictions to analyse mRNA statistics. Their stochastic model incorporates cell-cycle and gene copy-number effects, where each gene copy is described by a stochastic 2-state model (please see [32] for details). In this model, gene copies are independent, and at the end of the cell-cycle, mRNA molecules are binomially partitioned between the two daughter cells. Gene **dosage compensation** – a decrease in the rate of gene activation following gene replication – is also included. By using the available tools, the authors estimated numerically the gene replication time, while an analytical solution for the model

was not presented. Z. Cao et al. extended this model and performed a detailed mathematical analysis of the new version in [126]. This novel model includes mRNA maturation, cell division, gene replication, dosage compensation, and growth-dependent transcription. Even with all these biological details, their model is analytically tractable and Cao et al. derived expressions for the time-dependent distributions of nascent mRNA and mature mRNA numbers.

Another concept that we have not taken into account in the work of this thesis is cellular size. Experiments have shown that cellular size can directly and globally affect gene expression by modulating transcription. For instance, an increase in cell volume can result in an increase in transcriptional burst size [201]. Additionally, there is evidence that the transcriptional and translational outputs scale with cell size at a genome-wide level, maintaining this way **cellular homeostasis**. This mechanism plays an important role in a number of biological contexts; e.g. during embryogenesis, there is rapid cell division that leads to an exponential decrease in individual cell volume, however, the organism must maintain the concentration of most proteins that enable transcription to occur [201, 202] (and references therein). O. Padovan-Merhar et al. performed a study in [201] where they tried to investigate the transcriptional mechanisms that determine the relationship between cellular volume changes and transcript abundance. To do so, they measured transcript abundance and cellular volume simultaneously in individual human cells. Their results show that RNA concentration scales linearly with volume, while they also identified two transcriptional mechanisms that allow cells to maintain RNA concentration homeostasis. A number of independent studies have also shown that the homeostasis of both mRNA and protein concentrations is maintained in an exponentially growing cell volume with variable genome copy numbers [201, 203, 204] (and references therein). However, models of stochastic gene expression often assume a constant transcription rate per gene and constant translation rate per mRNA, which are incompatible with the just mentioned experimental findings. An interesting theoretical study that tried to approach this problem has been performed by J. Lin et al. in [205], where a coarse-grained "growing cell model" (for continuously proliferating cells) that takes into account cell volume growth and cell division was used, and the dynamics of both mRNA and proteins were studied. In this model, the genes are transcribed at different rates, which are dependent on promoter strength and the total number of available polymerases in the cell. Additionally, translation rates of mRNA depend on the number of active ribosomes and mRNA abundance. The results of this study show that: (i) the limiting nature of RNA polymerase and its exponential growth lead to the exponential growth of mRNA numbers, i.e. homeostasis of mRNA concentration comes from the resulting bounded concentration of polymerases, and (ii) the limiting nature of ribosomes in the translation process leads to the exponential growth of protein numbers, i.e. homeostasis of protein concentrations originates from the fact that ribosomes make all proteins.

It is worth mentioning that the advances in microfluidic devices and live-cell imaging have made possible the measurement of gene expression in single cells over cell generations – **cell lineage** [206]. The data in these experiments are sampled at a rate that is much higher than the frequency of cell division, thus providing us with a means to study and understand the temporal variation of gene expression as a cell progresses through its life-cycle. However, the stochastic models studied in this thesis, as well as the standard stochastic models in the literature that are based on the two-stage or three-stage representation of gene expression, lack a description of fluctuations in gene product numbers within a cell lineage. Recently developed models that deal with the aforementioned limitation and provide steady-state distributions of mRNA and protein numbers calculated across a cell lineage can be found in studies performed in [126,207]. A different mathematical approach for studying the fluctuations of mRNA and protein copy numbers within a cell lineage was presented by C. Jia et al. in [208]. In this study, a novel model of gene expression that includes a description of transcription, translation, degradation, bursting, promoter switching, DNA replication, gene dosage compensation, and symmetric or asymmetric partitioning at cell division, was developed, and analytical closed-form expressions for the power spectrum of fluctuations across a lineage were obtained. Unlike distributions, the power spectrum provides an understanding of the correlations between molecule numbers at two time points and the frequency composition of fluctuations in molecule numbers.

Another important factor that influences gene expression, and has not been yet discussed here, is **macromolecular crowding** [209]; intracellular environments are characterised by a high total macromolecular content, and it has been estimated that between 5% and 40% of the total cell volume is physically occupied by various molecules [210]. However, the CME of stochastic models of gene expression is derived by assuming that the chemical reactions happen under dilute conditions in a well-stirred compartment of certain volume (i.e. no volume exclusion); these are evidently simplifications, as it is well known that chemical reactions involve discrete and random collisions between individual molecules. Theoretical approaches that aim to extend the CME and account for the behaviour arising from the interaction between the reactants and the macromolecules in the reaction media of a cell have been developed in [210–212] (and references therein).

Although it has not been in our interest to incorporate the aforementioned biological mechanisms in our models presented in this thesis, we recognize their importance and keep them in mind for our future studies. We hope that the models and the approximation techniques developed in this work can serve as a useful tool in future research in gene expression and serviceably contribute to the field of molecular biology. The models we studied here have not been subject to a theoretical approach before, and hence, our obtained analytical distributions of gene product numbers can provide means for inferring kinetic parameters of our models from experimental data without performing time-consuming stochastic simulations.

Future perspectives

All the discussed and studied models in this thesis are focused on the expression of a single gene, which is a simpler problem than a coordinated expression of many genes. A substantial fraction of research in the fields of molecular biology and mathematical biology centres around the topic of Gene Regulatory Networks (GRNs), which describe the connectivity between a large number of interacting genes in a cell; please see Fig. 5.1 for an example of a simple GRN. Understanding

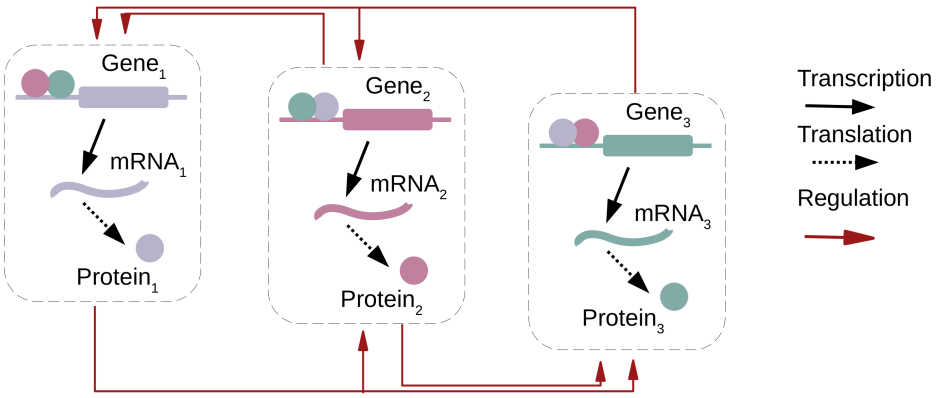


Figure 5.1: **Schematic of a simple three-genes regulatory network.**

their dynamics is important because gene interactions play an important role in a cell's response to the external environment and hence to its fate. However, these stochastic networks appear to be difficult to study because interconnections between genes lead to a noisy system with highly non-linear dynamics. Some studies have introduced simple models of GRNs – where only two or three genes are connected through few interactions – and were able to perform mathematical analysis in these cases [179, 205, 213, 214]. Other studies have considered GRNs with many interacting genes but have used stochastic simulations to study the dynamics in a small portion of parameter space [215]. Therefore, understanding the dynamics of realistic GRNs (that involve thousands of

interacting genes) remains an interesting and open research question. Our future research interest lies in obtaining approximate analytical solutions to a general stochastic model of an arbitrary number of connected genes by means of singular perturbation theory. The aim would be to obtain expressions for the approximate joint probability distribution function of gene products, which is the solution of the CME of the system. Additionally, an important research task would be to fit the analytical solution to distributions obtained from experimental data and infer kinetic parameters of the GRN in various cell types and under various growth conditions.

Appendix A

Supplementary Information for Chapter 3

A.1 Distribution of elongation time

In this section, we answer the following question: what is the distribution of the elongation time, i.e. the time between initiation and termination? In other words, with reference to Fig. 3.9 – which includes the non-pausing model in Fig. 3.1 as a special case – we want to find the distribution of the time at which RNAP leaves gene segment L (termination) if it was in the active state on gene segment 1 at time $t = 0$ (initiation).

Let $z_i(t)$ be the probability of an RNAP to be on gene segment i in the active state at time t, let $\tilde{z}_i(t)$ be the probability of the RNAP to be on gene segment i in the paused state at time t, and let $z_i^*(t)$ be the probability of the RNAP moving to gene segment $i+1$ at time t; note that $z_L^*(t)$ is the probability of the RNAP falling off the gene and forming a mature RNA, since for $i = L$, gene segment $L+1$ does not exist. Then, it follows from the reaction scheme illustrated in Fig. 3.9 that the master equations describing the Markovian dynamics on gene segment i are given by

$$\partial_t z_i(t) = -(r_p + k + d_a)z_i(t) + r_a\tilde{z}_i(t), \tag{A.1.1a}$$
$$\partial_t \tilde{z}_i(t) = -(d_p + r_a)\tilde{z}_i(t) + r_p z_i(t), \tag{A.1.1b}$$
$$\partial_t z_i^*(t) = k z_i(t). \tag{A.1.1c}$$

Now, we use these equations to find the distribution of the time when RNAP jumps to gene segment $i+1$, given that it is on gene segment i in the active state at $t = 0$, i.e. that $z_i(0) = 1$ and $\tilde{z}_i(0) = 0$. Taking the Laplace transform of Eqs. (A.1.1a) and (A.1.1b), we find

$$s\hat{z}_i(s) - 1 = -(r_p + k + d_a)\hat{z}_i(s) + r_a\hat{\tilde{z}}_i(s), \tag{A.1.2a}$$
$$s\hat{\tilde{z}}_i(s) = -(d_p + r_a)\hat{\tilde{z}}_i(s) + r_p\hat{z}_i(s), \tag{A.1.2b}$$

where $\hat{f}(s) = \int_0^\infty e^{-st} f(t)\,dt$. Solving these equations simultaneously, we obtain

$$\hat{z}_i(s) = \frac{s + d_p + r_a}{(s + k + d_a)(s + d_p + r_a) + r_p(s + d_p)} \tag{A.1.3}$$

Let $w(t)dt$ be the probability that the RNAP moves from segment i to $i+1$ in the time interval $(t, t+dt)$. Then, it follows from Eq. (A.1.1c) that $w(t) = \partial_t z_i^*(t) = k z_i(t)$. Integrating $w(t)$ over all times gives us the probability that the RNAP ultimately moves to the next segment $i+1$,

$$\int_0^\infty w(t)\,dt = \hat{w}(0) = k\hat{z}_i(0) = \frac{k\,(r_a + d_p)}{(d_a + k)\,(r_a + d_p) + d_p r_p}. \tag{A.1.4}$$

Note that $\hat{w}(0)$ is identical to the parameter $\tilde{\mu}$, as defined in Prop. 3. Let $y(t)dt$ be the probability that the RNAP moves from gene segment i to segment $i+1$ in the time interval $(t, t+dt)$, *conditioned* on those realisations that lead to an RNAP moving to the next gene segment $i+1$. (In other words, we exclude those realisations that lead to premature detachment.) Then, it follows by the definition of conditional probabilities that $y(t) = w(t)/\hat{w}(0)$, which implies

$$\hat{y}(s) = \frac{\hat{w}(s)}{\hat{w}(0)} = \frac{[(d_a+k)(r_a+d_p)+d_p r_p](r_a+d_p+s)}{(r_a+d_p)[(d_a+k+s)(r_a+d_p+s)+r_p(d_p+s)]}. \tag{A.1.5}$$

It follows that the mean $\langle t\rangle$ and variance $\mathrm{Var}(t)$ of the time t it takes RNAP to move to the next gene segment are given by

$$\langle t\rangle = -\left.\frac{d\hat{y}(s)}{ds}\right|_{s=0} = \frac{(r_a+d_p)^2+r_a r_p}{(r_a+d_p)[(d_a+k)(r_a+d_p)+d_p r_p]}, \tag{A.1.6a}$$

$$\begin{aligned}\mathrm{Var}(t) &= \left.\frac{d^2\hat{y}(s)}{ds^2}\right|_{s=0} - \left(\left.\frac{d\hat{y}(s)}{ds}\right|_{s=0}\right)^2 \\ &= \frac{2r_a r_p(r_a+d_p)(d_a+r_a+d_p+k)+(r_a+d_p)^4+r_a r_p^2(r_a+2d_p)}{(r_a+d_p)^2[(d_a+k)(r_a+d_p)+d_p r_p]^2},\end{aligned} \tag{A.1.6b}$$

respectively. Since RNAP can only move forwards in our model (irreversible motion), it follows that the time it takes an RNAP to move from the i-th to the $(i+1)$-th gene segment is independent of the time taken to move from another, j-th segment to the $(j+1)$-th segment. Hence, the time required for an RNAP to move across the entire gene from the first to the L-th segment, i.e. the 'elongation' time T from initiation to termination, is a sum of L independent and identical random variables. Thus, we can immediately state that the mean elongation time is $\langle T\rangle = L\langle t\rangle$, whereas the variance of the elongation time is $\mathrm{Var}(T) = L\mathrm{Var}(t)$. The coefficient of variation squared takes the form

$$\mathrm{CV}_T^2 = \frac{\mathrm{Var}(T)}{\langle T\rangle^2} = 1 + \frac{2r_a r_p[(d_a+k)(r_a+d_p)+d_p r_p]}{[(r_a+d_p)^2+r_a r_p]^2}. \tag{A.1.7}$$

From Eq. (A.1.7), it can be shown that for small premature detachment rates, the coefficient of variation of the elongation time is maximised when $r_p \approx r_a$. Taking the limit of infinitely many gene segments at constant mean elongation time, i.e. solving for k from the expression for the mean elongation time in Eq. (A.1.6), substituting into Eq. (A.1.7), and taking the limit of $L \to \infty$, we obtain

$$\lim_{L\to\infty} \mathrm{CV}_T^2 = \frac{2r_a r_p}{\langle T\rangle(r_a+d_p)[(r_a+d_p)^2+r_a r_p]}. \tag{A.1.8}$$

For the non-pausing model shown in Fig. 3.1, the above results simplify considerably due to $r_p = 0 = d_p$ and $d_a = d$; in that case, the inverse Laplace transform of Eq. (A.1.5) implies that $y(t)$ is an exponential distribution with parameter $k+d$. Hence, the total time it takes an RNAP to move across the entire gene is the sum of L independent and identically distributed exponential random variables, i.e. an Erlang distribution with shape parameter L and rate $k+d$, which implies that the mean elongation time is $L/(k+d)$, with coefficient of variation $1/\sqrt{L}$. It can be seen from Eq. (A.1.8) that deterministic elongation can only be observed when there is no pausing, i.e. when $r_p = 0$.

A.2 Solution of Lyapunov equation

Proof of Proposition 2. We start by defining the symmetric functions $f_{ij} = f_{ji}$ for $i, j = 1, \ldots, L$ as

$$\begin{aligned} &f_{00} = 1, && f_{0j} = \alpha^{j-1}, && f_{0M} = \theta\alpha^{L-1}, \\ &f_{ij} = (f_{i-1,j} + f_{i,j-1})/2, && f_{iM} = \gamma f_{i-1,M} + (1-\gamma) f_{iL}, && f_{MM} = f_{LM}, \end{aligned} \tag{A.2.1}$$

where the non-dimensional parameters α, γ, and θ are defined in Prop. 2. The elements of the Lyapunov equation given by Eq. (3.5) can be written explicitly as a set of simultaneous equations:

$$\mathbf{C}_{11} \cdot 2\mathbf{J}_{11} = -\mathbf{D}_{11}, \tag{A.2.2a}$$
$$\mathbf{C}_{12} \cdot (\mathbf{J}_{11} + \mathbf{J}_{22}) = -\mathbf{J}_{21}\mathbf{C}_{11}, \tag{A.2.2b}$$
$$\mathbf{C}_{1j} \cdot (\mathbf{J}_{11} + \mathbf{J}_{jj}) = -\mathbf{J}_{j,j-1}\mathbf{C}_{1,j-1} \quad \text{for } j = 3, \ldots, L+1, \tag{A.2.2c}$$
$$\mathbf{C}_{1,L+2} \cdot (\mathbf{J}_{11} + \mathbf{J}_{L+2,L+2}) = -\mathbf{J}_{L+2,L+1}\mathbf{C}_{1,L+1}, \tag{A.2.2d}$$

$$\mathbf{C}_{22} \cdot 2\mathbf{J}_{22} = -2\mathbf{J}_{21}\mathbf{C}_{12} - \mathbf{D}_{22}, \tag{A.2.2e}$$
$$\mathbf{C}_{23} \cdot (\mathbf{J}_{22} + \mathbf{J}_{33}) = -\mathbf{J}_{21}\mathbf{C}_{13} - \mathbf{J}_{32}\mathbf{C}_{22} - \mathbf{D}_{23}, \tag{A.2.2f}$$
$$\mathbf{C}_{2j} \cdot (\mathbf{J}_{22} + \mathbf{J}_{jj}) = -\mathbf{J}_{21}\mathbf{C}_{1j} - \mathbf{J}_{j,j-1}\mathbf{C}_{2,j-1} \quad \text{for } j = 4, \ldots, L+1, \tag{A.2.2g}$$
$$\mathbf{C}_{2,L+2} \cdot (\mathbf{J}_{22} + \mathbf{J}_{L+2,L+2}) = -\mathbf{J}_{21}\mathbf{C}_{1,L+2} - \mathbf{J}_{L+2,L+1}\mathbf{C}_{2,L+1}, \tag{A.2.2h}$$

$$\mathbf{C}_{ii} \cdot 2\mathbf{J}_{ii} = -2\mathbf{J}_{i,i-1}\mathbf{C}_{i-1,i} - \mathbf{D}_{ii} \quad \text{for } i = 3, \ldots, L+1, \tag{A.2.2i}$$
$$\mathbf{C}_{i,i+1} \cdot (\mathbf{J}_{ii} + \mathbf{J}_{i+1,i+1}) = -\mathbf{J}_{i,i-1}\mathbf{C}_{i-1,i+1} - \mathbf{J}_{i+1,i}\mathbf{C}_{ii} - \mathbf{D}_{i,i+1} \quad \text{for } i = 3, \ldots, L, \tag{A.2.2j}$$
$$\mathbf{C}_{ij} \cdot (\mathbf{J}_{ii} + \mathbf{J}_{jj}) = -\mathbf{J}_{i,i-1}\mathbf{C}_{i-1,j} - \mathbf{J}_{j,j-1}\mathbf{C}_{i,j-1} \quad \text{for } i = 3, \ldots, L+1 \text{ and } j = i+2, \ldots, L+1, \tag{A.2.2k}$$
$$\mathbf{C}_{i,L+2} \cdot (\mathbf{J}_{ii} + \mathbf{J}_{L+2,L+2}) = -\mathbf{J}_{i,i-1}\mathbf{C}_{i-1,L+2} - \mathbf{J}_{L+2,L+1}\mathbf{C}_{i,L+1} \quad \text{for } i = 3, \ldots, L+1, \tag{A.2.2l}$$

$$\mathbf{C}_{L+2,L+2} \cdot 2\mathbf{J}_{L+2,L+2} = -2\mathbf{J}_{L+2,L+1}\mathbf{C}_{L+1,L+2} - \mathbf{D}_{L+2,L+2}. \tag{A.2.2m}$$

Now, we substitute the elements of the Jacobian matrix $\mathbf{J}$ and the diffusion matrix $\mathbf{D}$ from Eqs. (3.6) and (3.7), respectively, into the above system of algebraic equations, which we then solve to find the elements of the covariance matrix $\mathbf{C}$. Note that, for the following mathematical derivation, we take into account the expressions for the steady-state mean numbers of species given in Eq. (3.2), as well as the definition of the functions f_{ij} in Eq. (A.2.1).
From Eq. (A.2.2a), one easily obtains $\mathbf{C}_{11} = \eta^2\beta$. Then, it follows from Eq. (A.2.2b) that

$$\mathbf{C}_{12} = \frac{r}{s_u + s_b + k + d}\mathbf{C}_{11} = \rho_k\mu\alpha(\eta^2\beta) = \eta(\eta\rho_k\mu)\alpha\beta = \eta\langle n_1\rangle\alpha\beta \cdot f_{01}. \tag{A.2.3}$$

Eq. (A.2.2c) implies that, for $j = 3, \ldots, L+1$:

$$\mathbf{C}_{1j} = \frac{k}{s_u + s_b + k + d}\mathbf{C}_{1,j-1} = \mu\alpha \cdot \mathbf{C}_{1,j-1} = (\mu\alpha)^{j-2}\mathbf{C}_{12} = (\mu\alpha)^{j-2}(\eta\langle n_1\rangle\alpha\beta) = \eta\langle n_{j-1}\rangle\alpha\beta \cdot f_{0,j-1}. \tag{A.2.4}$$

From Eq. (A.2.2d), we have that

$$\mathbf{C}_{1,L+2} = \frac{k}{s_u + s_b + d_m}\mathbf{C}_{1,L+1} = \frac{k}{d_m}\theta(\langle n_L\rangle\alpha\beta \cdot f_{0L}) = \eta(\frac{k}{d_m}\langle n_L\rangle)(\alpha\beta)(\theta \cdot f_{0L}) = \eta\langle n\rangle \cdot f_{0M}; \tag{A.2.5}$$

from Eq. (A.2.2e), we find

$$\mathbf{C}_{22} = \frac{r\langle n_0\rangle + (k+d)\langle n_1\rangle}{2(k+d)} + \frac{r}{k+d}\mathbf{C}_{12} = \frac{\rho_k\mu\eta + \langle n_1\rangle}{2} + (\rho_k\mu)(\eta\langle n_1\rangle\alpha\beta \cdot f_{01}) = \langle n_1\rangle + \langle n_1\rangle^2\alpha\beta \cdot f_{11}, \tag{A.2.6}$$

since $f_{11} = (f_{01} + f_{10})/2 = f_{01}$ from the definition in Eq. (A.2.1).

From Eq. (A.2.2f), we obtain

$$\begin{aligned}
\mathbf{C}_{23} &= -\frac{k}{2(d+k)}\langle n_1\rangle + \frac{r}{2(k+d)}\mathbf{C}_{13} + \frac{k}{2(d+k)}\mathbf{C}_{22} \\
&= -\frac{\langle n_2\rangle}{2} + \frac{1}{2}(\rho_k\mu\eta)\langle n_2\rangle\alpha\beta\cdot f_{02} + \frac{1}{2}\mu[\langle n_1\rangle + \langle n_1\rangle^2\alpha\beta\cdot f_{11}] \\
&= -\frac{\langle n_2\rangle}{2} + \frac{1}{2}\langle n_1\rangle\langle n_2\rangle\alpha\beta\cdot f_{02} + \frac{\langle n_2\rangle}{2} + \frac{1}{2}(\mu\langle n_1\rangle)\langle n_1\rangle\alpha\beta\cdot f_{11} \\
&= \frac{1}{2}\langle n_1\rangle\langle n_2\rangle\alpha\beta\cdot f_{02} + \frac{1}{2}\langle n_2\rangle\langle n_1\rangle\alpha\beta\cdot f_{11} = \langle n_1\rangle\langle n_2\rangle\alpha\beta\frac{1}{2}(f_{02}+f_{11}) = \langle n_1\rangle\langle n_2\rangle\alpha\beta\cdot f_{12},
\end{aligned} \tag{A.2.7}$$

since $f_{12} = (f_{02} + f_{11})/2$ from the definition in Eq. (A.2.1).

From Eq. (A.2.2g), we have that, for $j = 4, \ldots, L+1$,

$$\mathbf{C}_{2j} = \frac{r}{2(k+d)}\mathbf{C}_{1j} + \frac{k}{2(k+d)}\mathbf{C}_{2,j-1} = \frac{\rho_k\mu}{2}\mathbf{C}_{1j} + \frac{\mu}{2}\mathbf{C}_{2,j-1} = \frac{\rho_k\mu}{2}\sum_{q=0}^{j-4}\left(\frac{\mu}{2}\right)^q\mathbf{C}_{1,j-q} + \left(\frac{\mu}{2}\right)^{j-3}\mathbf{C}_{23}. \tag{A.2.8}$$

The proof of Eq. (A.2.8) is given in Lemma A.2.1. The above expression for $\mathbf{C}_{2j}$ can be further simplified to

$$\begin{aligned}
\mathbf{C}_{2j} &= \frac{\rho_k\mu}{2}\sum_{q=0}^{j-4}\left(\frac{\mu}{2}\right)^q\eta\langle n_{j-q-1}\rangle\alpha\beta\cdot f_{0,j-q-1} + \left(\frac{\mu}{2}\right)^{j-3}\langle n_1\rangle\langle n_2\rangle\alpha\beta\cdot f_{12} \\
&= \sum_{q=0}^{j-4}\left(\frac{1}{2}\right)^{q+1}(\rho_k\mu\eta)(\mu^q\langle n_{j-q-1}\rangle)\alpha\beta\cdot f_{0,j-q-1} + \left(\frac{1}{2}\right)^{j-3}\langle n_1\rangle(\mu^{j-3}\langle n_2\rangle)\alpha\beta\cdot f_{12} \\
&= \sum_{q=0}^{j-4}\left(\frac{1}{2}\right)^{q+1}\langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta\cdot f_{1,j-q-1} + \left(\frac{1}{2}\right)^{j-3}\langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta\cdot f_{12} \\
&= \langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta\left[\sum_{q=0}^{j-4}\left(\frac{1}{2}\right)^{q+1}f_{1,j-q-1} + \left(\frac{1}{2}\right)^{j-3}f_{12}\right] = \langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta\cdot f_{1,j-1}.
\end{aligned} \tag{A.2.9}$$

For the proof of the last equality in Eq. (A.2.9), see Lemma A.2.2.

From Eq. (A.2.2h), we have that

$$\begin{aligned}
\mathbf{C}_{2,L+2} &= \frac{r}{k+d+d_m}\mathbf{C}_{1,L+2} + \frac{k}{k+d+d_m}\mathbf{C}_{2,L+1} = \rho_k\mu\gamma\mathbf{C}_{1,L+2} + \mu\gamma\mathbf{C}_{2,L+1} \\
&= (\rho_k\mu\gamma)(\eta\langle n\rangle\alpha\beta\cdot f_{0M}) + (\mu\gamma)(\langle n_1\rangle\langle n_L\rangle\alpha\beta\cdot f_{1L}) \\
&= (\rho_k\eta\mu)\langle n\rangle\alpha\beta\cdot\gamma f_{0M} + \mu\frac{d_m}{k}\langle n_1\rangle\frac{k}{d_m}\langle n_L\rangle\alpha\beta\cdot\gamma f_{1L} = \langle n_1\rangle\langle n\rangle\alpha\beta\cdot\left[\gamma f_{0M} + \mu\frac{d_m}{k}\gamma f_{1L}\right] \\
&= \langle n_1\rangle\langle n\rangle\alpha\beta\cdot[\gamma f_{0M} + (1-\gamma)\cdot f_{1L}] = \langle n_1\rangle\langle n\rangle\alpha\beta\cdot f_{1M},
\end{aligned} \tag{A.2.10}$$

where f_{1M} is defined in Eq. (A.2.1).

Eqs. (A.2.2i) through (A.2.2k) yield the system

$$\begin{aligned}\mathbf{C}_{ii} &= \frac{k\langle n_{i-2}\rangle + (k+d)\langle n_{i-1}\rangle}{2(k+d)} + \frac{k}{k+d}\mathbf{C}_{i-1,i} = \langle n_{i-1}\rangle + \mu\mathbf{C}_{i-1,i},\\ \mathbf{C}_{i,i+1} &= \frac{\mu}{2}\mathbf{C}_{i-1,i+1} + \frac{\mu}{2}\mathbf{C}_{ii} - \frac{\mu}{2}\langle n_{i-1}\rangle = \frac{\mu}{2}(\mathbf{C}_{i-1,i+1} + \mu\mathbf{C}_{i-1,i}),\\ \mathbf{C}_{ij} &= \frac{\mu}{2}(\mathbf{C}_{i-1,j} + \mathbf{C}_{i,j-1}),\end{aligned} \tag{A.2.11}$$

which can be rewritten more compactly as

$$\mathbf{C}_{ij} = \delta_{ij}\langle n_{i-1}\rangle + \langle n_{i-1}\rangle\langle n_{j-1}\rangle\alpha\beta\cdot f_{i-1,j-1} \qquad \text{for } i,j = 3,\dots,L+1, \tag{A.2.12}$$

where δ_{ij} is the Kronecker delta. A detailed derivation is given in Lemma A.2.3.

From Eq. (A.2.2l), we have that for $i = 3,\dots,L+1$,

$$\begin{aligned}\mathbf{C}_{i,L+2} &= \frac{k}{k+d+d_m}\mathbf{C}_{i-1,L+2} + \frac{k}{k+d+d_m}\mathbf{C}_{i,L+2} = \mu\gamma\mathbf{C}_{i-1,L+2} + (k/d_m)(1-\gamma)\mathbf{C}_{i,L+1}\\ &= \gamma(\mu\langle n_{i-2}\rangle)\langle n\rangle\alpha\beta\cdot f_{i-2,M} + (1-\gamma)\langle n_{i-1}\rangle(k/d_m\langle n_L\rangle)\alpha\beta\cdot f_{i-1,L}\\ &= \langle n_{i-1}\rangle\langle n\rangle\alpha\beta\cdot[\gamma f_{i-2,M} + (1-\gamma)f_{i-1,L}] = \langle n_{i-1}\rangle\langle n\rangle\alpha\beta\cdot f_{i-1,M},\end{aligned} \tag{A.2.13}$$

where f_{iM} is defined in Eq. (A.2.1).

Finally, Eq. (A.2.2m) yields

$$\mathbf{C}_{L+2,L+2} = \frac{k\langle n_L\rangle + d_m\langle n\rangle}{2d_m} + \frac{k}{d_m}\mathbf{C}_{L+1,L+2} = \langle n\rangle + (k/d_m)\langle n_L\rangle\langle n\rangle\alpha\beta\cdot f_{LM} = \langle n\rangle + \langle n\rangle^2\alpha\beta\cdot f_{MM}, \tag{A.2.14}$$

where $f_{MM} = f_{LM}$ is defined in Eq. (A.2.1).

Summarising the above results, we conclude that the solution for the symmetric covariance matrix $\mathbf{C}$ is given by the system in Eq. (3.4), where we have that $\mathrm{Cov}(n_i,n_j) = \mathbf{C}_{i+1,j+1}$, $\mathrm{Cov}(n_i,n) = \mathbf{C}_{i+1,L+2}$ for $i,j = 0,\dots,L$, and $\mathrm{Var}(n,n) = \mathbf{C}_{L+2,L+2}$. Here, the functions f_{ij} are defined as in Eq. (A.2.1). Now, the recurrence relation $f_{ij} = (f_{i-1,j} + f_{i,j-1})/2$ in Eq. (A.2.1) can be solved for $i,j = 1,2,\dots,L$ via the method of generating functions, which gives the following analytical expression:

$$f_{ij} = f(i,j) + f(j,i), \tag{A.2.15}$$

where

$$f(i,j) = \frac{\alpha^{i+j-1}}{(2\alpha-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\alpha-1}{2\alpha}{}_2F_1\big(1,i+j;j;\tfrac{1}{2\alpha}\big)\Big];$$

see Lemma A.2.5 for a detailed derivation. Additionally, we can easily prove that the function f_{iM} in Eq. (A.2.1) can be rewritten as

$$f_{iM} = \gamma^i f_{0M} + (1-\gamma)\sum_{q=1}^{i}\gamma^{i-q}f_{qL}, \tag{A.2.16}$$

as shown in Lemma A.2.4. ■

Lemma A.2.1. *For $j = 4,\dots,L+1$, we have the identity*

$$\mathbf{C}_{2j} = \frac{\rho_k\mu}{2}\mathbf{C}_{1j} + \frac{\mu}{2}\mathbf{C}_{2,j-1} = \frac{\rho_k\mu}{2}\sum_{q=0}^{j-4}\Big(\frac{\mu}{2}\Big)^q\mathbf{C}_{1,j-q} + \Big(\frac{\mu}{2}\Big)^{j-3}\mathbf{C}_{23}, \tag{A.2.17}$$

as stated in Eq. (A.2.8).

Proof. The identity in Eq. (A.2.17) will be proved by induction: one can easily show that it holds for $j = 4$. Now, we assume that Eq. (A.2.17) is true for some $j \geq 5$; hence, for $j+1$, we have

$$\begin{aligned}
\mathbf{C}_{2,j+1} &= \sum_{q=0}^{j-3} \left(\frac{\mu}{2}\right)^q \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1-q} + \left(\frac{\mu}{2}\right)^{j-2} \mathbf{C}_{23} \\
&= \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1} + \sum_{q=1}^{j-3} \left(\frac{\mu}{2}\right)^q \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1-q} + \left(\frac{\mu}{2}\right)^{j-2} \mathbf{C}_{23} \\
&= \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1} + \frac{\mu}{2} \left[\sum_{q=1}^{j-3} \left(\frac{\mu}{2}\right)^{q-1} \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1-q} + \left(\frac{\mu}{2}\right)^{j-3} \mathbf{C}_{23} \right] \\
&= \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1} + \frac{\mu}{2} \left[\sum_{q=0}^{j-4} \left(\frac{\mu}{2}\right)^q \frac{\rho_k \mu}{2} \mathbf{C}_{1,j-q} + \left(\frac{\mu}{2}\right)^{j-3} \mathbf{C}_{23} \right] \\
&= \frac{\rho_k \mu}{2} \mathbf{C}_{1,j+1} + \frac{\mu}{2} \mathbf{C}_{2j},
\end{aligned} \tag{A.2.18}$$

as claimed, which implies that the identity in Eq. (A.2.17) holds for all $j = 4, \ldots, L+1$. ■

Lemma A.2.2. *The function f_{1j}, which is defined by the recurrence relation $f_{1j} = (f_{0j} + f_{1,j-1})/2$ in Eq. (A.2.1), satisfies the identity*

$$f_{1j} = \sum_{q=0}^{j-3} \left(\frac{1}{2}\right)^{q+1} f_{0,j-q} + \left(\frac{1}{2}\right)^{j-2} f_{12} \qquad \text{for } j = 3, \ldots, L, \tag{A.2.19}$$

as stated in Eq. (A.2.9).

Proof. We will again prove Eq. (A.2.19) by induction. For $j = 3$, we have from Eq. (A.2.19) that $f_{13} = (f_{03} + f_{12})/2$, which is true by the definition of f_{13}. We assume that the identity in Eq. (A.2.19) is correct for some $j \geq 4$; then, for $j+1$, the definition of $f_{1,j+1}$, in combination with our assumption, implies

$$\begin{aligned}
f_{1,j+1} &= \frac{1}{2} f_{0,j+1} + \frac{1}{2} f_{1j} = \frac{1}{2} f_{0,j+1} + \frac{1}{2} \left[\sum_{q=0}^{j-3} \left(\frac{1}{2}\right)^{q+1} f_{0,j-q} + \left(\frac{1}{2}\right)^{j-2} f_{12} \right] \\
&= \frac{1}{2} f_{0,j+1} + \frac{1}{2} \left[\sum_{q=1}^{j-2} \left(\frac{1}{2}\right)^{q} f_{0,j+1-q} + \left(\frac{1}{2}\right)^{j-2} f_{12} \right] \\
&= \sum_{q=0}^{j-2} \left(\frac{1}{2}\right)^{q+1} f_{1,j+1-q} + \left(\frac{1}{2}\right)^{j-1} f_{12},
\end{aligned}$$

as claimed. Hence, the equality in Eq. (A.2.19) is true for all $j = 3, \ldots, L$. ■

Lemma A.2.3. *The system in Eq. (A.2.11), which is given by*

$$\begin{aligned}
\mathbf{C}_{ii} &= \langle n_{i-1} \rangle + \mu \mathbf{C}_{i-1,i} && \text{for } i = 3, \ldots, L, \\
\mathbf{C}_{i,i+1} &= \frac{\mu}{2} (\mathbf{C}_{i-1,i+1} + \mu \mathbf{C}_{i-1,i}) && \text{for } i = 3, \ldots, L, \\
\mathbf{C}_{ij} &= \frac{\mu}{2} (\mathbf{C}_{i-1,j} + \mathbf{C}_{i,j-1}) && \text{for } i = 3, \ldots, L+1 \text{ and } j = i+1, \ldots, L+1,
\end{aligned} \tag{A.2.20}$$

is equivalent to the system

$$\mathbf{C}_{ij} = \delta_{ij}\langle n_{i-1}\rangle + \langle n_{i-1}\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i-1,j-1} \qquad \text{for } i,j = 3,\dots,L+1, \tag{A.2.21}$$

as stated in Eq. (A.2.12). Here, the functions f_{ij} are defined as in Eq. (A.2.1).

Proof. We again use the method of induction. For $i = 3$, we have

$$\begin{aligned}
\mathbf{C}_{33} &= \langle n_2\rangle + \langle n_2\rangle^2\alpha\beta \cdot f_{22} = \langle n_2\rangle + \mu\langle n_1\rangle\langle n_2\rangle\alpha\beta \cdot f_{12} = \langle n_2\rangle + \mu\mathbf{C}_{23},\\
\mathbf{C}_{34} &= \langle n_2\rangle\langle n_3\rangle(\beta\alpha) \cdot f_{34} = \langle n_2\rangle\langle n_3\rangle\alpha\beta(f_{22}+f_{13})/2 = [\langle n_2\rangle\langle n_3\rangle\alpha\beta \cdot f_{22} + \langle n_2\rangle\langle n_3\rangle\alpha\beta \cdot f_{13}]/2\\
&= [\mu\langle n_1\rangle\langle n_3\rangle\alpha\beta \cdot f_{13} + \mu^2\langle n_1\rangle\langle n_2\rangle\alpha\beta \cdot f_{12}]/2 = \frac{\mu}{2}[\langle n_1\rangle\langle n_3\rangle\alpha\beta \cdot f_{13} + \mu\langle n_1\rangle\langle n_2\rangle\alpha\beta \cdot f_{12}]\\
&= \frac{\mu}{2}(\mathbf{C}_{24} + \mu\mathbf{C}_{23}),\\
\mathbf{C}_{3j} &= \langle n_2\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{2,j-1} = \langle n_2\rangle\langle n_{j-1}\rangle\alpha\beta(f_{2,j-2}+f_{1,j-1})/2\\
&= [\langle n_2\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{2,j-2} + \langle n_2\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{1,j-1}]/2\\
&= [\langle n_2\rangle\mu\langle n_{j-2}\rangle\alpha\beta \cdot f_{2,j-2} + \mu\langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{1,j-1}]/2\\
&= \frac{\mu}{2}[\langle n_2\rangle\langle n_{j-2}\rangle\alpha\beta \cdot f_{2,j-2} + \langle n_1\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{1,j-1}]\\
&= \frac{\mu}{2}(\mathbf{C}_{3,j-1} + \mathbf{C}_{2j}).
\end{aligned} \tag{A.2.22}$$

Now, we assume that the statement is true for some $i \geq 4$; then, for $i+1$, we have

$$\begin{aligned}
\mathbf{C}_{i+1,i+1} &= \langle n_i\rangle + \langle n_i\rangle^2\alpha\beta \cdot f_{ii} = \langle n_i\rangle + \mu\langle n_{i-1}\rangle\langle n_i\rangle\alpha\beta \cdot f_{i-1,i} = \langle n_i\rangle + \mu\mathbf{C}_{i,i+1},\\
\mathbf{C}_{i+1,i+2} &= \langle n_i\rangle\langle n_{i+1}\rangle\alpha\beta \cdot f_{i,i+1} = \langle n_i\rangle\langle n_{i+1}\rangle\alpha\beta(f_{i-1,i+1}+f_{ii})/2\\
&= [\langle n_i\rangle\langle n_{i+1}\rangle\alpha\beta \cdot f_{i-1,i+1} + \langle n_i\rangle\langle n_{i+1}\rangle\alpha\beta \cdot f_{ii}]/2\\
&= [\mu\langle n_{i-1}\rangle\langle n_{i+1}\rangle\alpha\beta \cdot f_{i-1,i+1} + \mu^2\langle n_{i-1}\rangle\langle n_i\rangle\alpha\beta \cdot f_{i-1,i}]/2\\
&= \frac{\mu}{2}[\langle n_{i-1}\rangle\langle n_{i+1}\rangle\alpha\beta \cdot f_{i-1,i+1} + \mu\langle n_{i-1}\rangle\langle n_i\rangle\alpha\beta \cdot f_{i-1,i}]\\
&= \frac{\mu}{2}(\mathbf{C}_{i,i+2} + \mu\mathbf{C}_{i,i+1}),
\end{aligned} \tag{A.2.23}$$

$$\begin{aligned}
\mathbf{C}_{i+1,j} &= \langle n_i\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i,j-1} = \langle n_i\rangle\langle n_{j-1}\rangle\alpha\beta(f_{i-1,j-1}+f_{i,j-2})/2\\
&= [\langle n_i\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i-1,j-1} + \langle n_i\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i,j-2}]/2\\
&= [\mu\langle n_{i-1}\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i-1,j-1} + \mu\langle n_i\rangle\langle n_{j-2}\rangle\alpha\beta \cdot f_{i,j-2}]/2\\
&= \frac{\mu}{2}[\langle n_{i-1}\rangle\langle n_{j-1}\rangle\alpha\beta \cdot f_{i-1,j-1} + \langle n_i\rangle\langle n_{j-2}\rangle\alpha\beta \cdot f_{i,j-2}]\\
&= \frac{\mu}{2}(\mathbf{C}_{ij} + \mathbf{C}_{i+1,j-1}),
\end{aligned}$$

which is also correct. Hence, the statement of the lemma is true for all i and j, as stated. ■

Lemma A.2.4. *For $i = 1,\dots,L$, the function f_{iM} defined in Eq. (A.2.1) can be simplified as in Eq. (A.2.16); specifically, we have the identity*

$$f_{iM} = \gamma f_{i-1,M} + (1-\gamma)f_{i,L} = \gamma^i \cdot f_{0M} + (1-\gamma)\sum_{q=1}^{i}\gamma^{i-q} \cdot f_{qL}. \tag{A.2.24}$$

Proof. The proof is by induction: for $i = 1$, the identity is obvious. We now suppose that Eq. (A.2.24) is true for some $i \geq 2$; hence, for $i+1$, we have

$$\begin{aligned} f_{i+1,M} &= \gamma^{i+1} \cdot f_{0M} + (1-\gamma) \sum_{q=1}^{i+1} \gamma^{i+1-q} \cdot f_{qL} \\ &= \gamma \left[\gamma^{i} \cdot f_{0M} + (1-\gamma) \sum_{q=1}^{i} \gamma^{i-q} \cdot f_{qL} \right] + (1-\gamma) f_{i+1,L} = \gamma f_{iM} + (1-\gamma) f_{i+1,L}, \end{aligned} \tag{A.2.25}$$

which is correct. Hence, Eq. (A.2.24) is true for all i, as stated. ■

Lemma A.2.5. *For $i, j = 1, \ldots, L$, the solution of the recurrence relation $f_{ij} = (f_{i,j-1} + f_{i-1,j})/2$ in Eq. (A.2.1) is given by $f_{ij} = f(i,j) + f(j,i)$, where*

$$f(i,j) = \frac{\alpha^{i+j-1}}{(2\alpha-1)^{i}} + \frac{1}{2^{i+j-1}} \binom{i+j-1}{i} \Big[1 - \frac{2\alpha-1}{2\alpha} {}_2F_1\big(1, i+j; j; \tfrac{1}{2\alpha}\big) \Big]. \tag{A.2.26}$$

Proof. In order to solve the recurrence relation for the function f_{ij}, we take into account the initial conditions $f_{00} = 1$ and $f_{0j} = f_{j0} = \alpha^{j-1}$. Then, we define a generating function $g(x,y)$ via

$$g(x,y) = \sum_{i,j \geq 0} f_{ij} x^i y^j = f_{00} + \sum_{j \geq 1} f_{0j} y^j + \sum_{i \geq 1} f_{i0} x^i + \sum_{i,j \geq 1} f_{ij} x^i y^j, \tag{A.2.27}$$

where the last term can be rewritten as

$$\begin{aligned} \sum_{i,j \geq 1} f_{ij} x^i y^j &= \sum_{i,j \geq 1} \frac{1}{2} (f_{i-1,j} + f_{i,j-1}) x^i y^j \\ &= \frac{1}{2} x \sum_{i,j \geq 1} f_{i-1,j} x^{i-1} y^j + \frac{1}{2} y \sum_{i,j \geq 1} f_{i,j-1} x^i y^{j-1} \\ &= \frac{1}{2} x \sum_{i \geq 0} \sum_{j \geq 1} f_{ij} x^i y^j + \frac{1}{2} y \sum_{i \geq 1} \sum_{j \geq 0} f_{ij} x^i y^j \\ &= \frac{1}{2} x \Big(\sum_{i,j \geq 0} f_{ij} x^i y^j - \sum_{i \geq 0} f_{i0} x^i \Big) + \frac{1}{2} y \Big(\sum_{i,j \geq 0} f_{ij} x^i y^j - \sum_{j \geq 0} f_{0j} y^j \Big) \\ &= \frac{1}{2} x \Big(g(x,y) - \sum_{i \geq 0} f_{i0} x^i \Big) + \frac{1}{2} y \Big(g(x,y) - \sum_{j \geq 0} f_{0j} y^j \Big). \end{aligned} \tag{A.2.28}$$

Hence, Eq. (A.2.27) becomes

$$g(x,y) = f_{00} + \sum_{j \geq 1} f_{0j} y^j + \sum_{i \geq 1} f_{i0} x^i + \frac{1}{2} x \Big(g(x,y) - \sum_{i \geq 0} f_{i0} x^i \Big) + \frac{1}{2} y \Big(g(x,y) - \sum_{j \geq 0} f_{0j} y^j \Big),$$

which is equivalent to

$$g(x,y) \Big(1 - \frac{1}{2}x - \frac{1}{2}y \Big) = f_{00} \Big(1 - \frac{1}{2}x - \frac{1}{2}y \Big) + \Big(1 - \frac{1}{2}y \Big) \sum_{j \geq 1} f_{0j} y^j + \Big(1 - \frac{1}{2}x \Big) \sum_{i \geq 1} f_{i0} x^i$$

or

$$g(x,y) = f_{00} + \Big(1 - \frac{1}{2}y \Big) \frac{1}{1 - \frac{1}{2}x - \frac{1}{2}y} \sum_{j \geq 1} f_{0j} y^j + \Big(1 - \frac{1}{2}x \Big) \frac{}{1 - \frac{1}{2}x - \frac{1}{2}y} \sum_{i \geq 1} f_{i0} x^i. \tag{A.2.29}$$

Taking into account the initial conditions, we find that

$$\sum_{j\geq 1} f_{0j}y^j = \sum_{j\geq 1}\alpha^{j-1}y^j = \frac{1}{\alpha}\sum_{j\geq 1}(\alpha y)^j \quad \text{and} \quad \sum_{i\geq 1} f_{i0}x^i = \frac{1}{\alpha}\sum_{i\geq 1}(\alpha x)^i, \tag{A.2.30}$$

which we substitute into Eq. (A.2.29) to obtain

$$g(x,y) = 1 + \Big(1-\frac{1}{2}y\Big)\frac{1}{1-\frac{1}{2}x-\frac{1}{2}y}\frac{1}{\alpha}\sum_{j\geq 1}(\alpha y)^j + \Big(1-\frac{1}{2}x\Big)\frac{1}{1-\frac{1}{2}x-\frac{1}{2}y}\frac{1}{\alpha}\sum_{i\geq 1}(\alpha x)^i. \tag{A.2.31}$$

Making use of the well-known symmetric, bivariate generating function of the binomial coefficients

$$\frac{1}{1-s-t} = \sum_{i,j\geq 0}\binom{i+j}{i}s^i t^j, \tag{A.2.32}$$

we can rewrite Eq. (A.2.31) as

$$\begin{aligned} g(x,y) &= 1 + \Big(1-\frac{1}{2}y\Big)\frac{1}{\alpha}\sum_{j\geq 1}(\alpha y)^j \sum_{i,j\geq 0}\binom{i+j}{i}\frac{x^i y^j}{2^{i+j}} + \Big(1-\frac{1}{2}x\Big)\frac{1}{\alpha}\sum_{i\geq 1}(\alpha x)^i \sum_{i,j\geq 0}\binom{i+j}{i}\frac{x^i y^j}{2^{i+j}} \\ &= \Big(1-\frac{1}{2}y\Big)\sum_{i,j\geq 0}\sum_{q=0}^{j-1}\binom{i+q}{i}\frac{\alpha^{j-q-1}}{2^{i+q}}x^i y^j + \Big(1-\frac{1}{2}x\Big)\sum_{i,j\geq 0}\sum_{q=0}^{i-1}\binom{j+q}{q}\frac{\alpha^{i-q-1}}{2^{j+q}}x^i y^j. \end{aligned}$$

Rearranging sums in the above expression, we find

$$\begin{aligned} g(x,y) = \sum_{i,j\geq 0}\Bigg[&\sum_{q=0}^{j-1}\binom{i+q}{i}\frac{\alpha^{j-q-1}}{2^{i+q}} - \sum_{q=0}^{j-2}\binom{i+q}{i}\frac{\alpha^{j-q-2}}{2^{i+q+1}} \\ &+ \sum_{q=0}^{i-1}\binom{j+q}{q}\frac{\alpha^{i-q-1}}{2^{j+q}} - \sum_{q=0}^{i-2}\binom{j+q}{q}\frac{\alpha^{i-q-2}}{2^{j+q+1}}\Bigg]x^i y^j. \end{aligned}$$

Hence, we obtain the following exact expression for the function f_{ij},

$$f_{ij} = \sum_{q=0}^{j-1}\binom{i+q}{q}\frac{\alpha^{j-q-1}}{2^{i+q}} - \sum_{q=0}^{j-2}\binom{i+q}{q}\frac{\alpha^{j-q-2}}{2^{i+q+1}} + \sum_{q=0}^{i-1}\binom{j+q}{q}\frac{\alpha^{i-q-1}}{2^{j+q}} - \sum_{q=0}^{i-2}\binom{j+q}{q}\frac{\alpha^{i-q-2}}{2^{j+q+1}}. \tag{A.2.33}$$

The expression in Eq. (A.2.33) can be simplified further due to its symmetry with respect to the indices i and j: we write $f_{ij} = f(i,j) + f(j,i)$, where $f(i,j)$ is defined as

$$f(i,j) = \sum_{q=0}^{j-1}\binom{i+q}{q}\frac{\alpha^{j-q-1}}{2^{i+q}} - \sum_{q=0}^{j-2}\binom{i+q}{q}\frac{\alpha^{j-q-2}}{2^{i+q+1}}. \tag{A.2.34}$$

The function $f(i,j)$ can be further simplified as

$$\begin{aligned} f(i,j) &= \binom{i+j-1}{j-1}\frac{1}{2^{i+j-1}} + 2\alpha\sum_{q=0}^{j-2}\binom{i+q}{q}\frac{\alpha^{j-q-2}}{2^{i+q+1}} - \sum_{q=0}^{j-2}\binom{i+q}{q}\frac{\alpha^{j-q-2}}{2^{i+q+1}} \\ &= \binom{i+j-1}{j-1}\frac{1}{2^{i+j-1}} + (2\alpha-1)\frac{\alpha^{j-2}}{2^{i+1}}\sum_{q=0}^{j-2}\binom{i+q}{q}\Big(\frac{1}{2\alpha}\Big)^q; \end{aligned} \tag{A.2.35}$$

next, we use the identity

$$\sum_{q=0}^{j}\binom{i+q}{i}x^q = \frac{1}{(1-x)^{i+1}} - x^{j+1}\binom{j+1+i}{j+1}{}_2F_1(1, j+i+2; j+2; x), \tag{A.2.36}$$

where ${}_2F_1$ is again the generalised hypergeometric function of the second kind [44]. Note that the above identity can be used only when $|x| < 1$, as the hypergeometric function ${}_2F_1$ is not defined otherwise.

Hence, Eq. (A.2.35) becomes

$$\begin{aligned} f(i,j) &= \binom{i+j-1}{j-1}\frac{1}{2^{i+j-1}} + (2\alpha-1)\frac{\alpha^{j-2}}{2^{i+1}}\left[\left(\frac{2\alpha}{2\alpha-1}\right)^{i+1} - \frac{1}{(2\alpha)^{j-1}}\binom{j+i-1}{j-1}{}_2F_1\big(1, j+i; j; \tfrac{1}{2\alpha}\big)\right] \\ &= \frac{\alpha^{i+j-1}}{(2\alpha-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\alpha-1}{2\alpha}{}_2F_1\big(1, j+i; j; \tfrac{1}{2\alpha}\big)\Big] \end{aligned} \tag{A.2.37}$$

Given the expression for $f(i,j)$ in Eq. (A.2.37), one can find the corresponding expression for $f(j,i)$ by exchanging the indexes $i \leftrightarrow j$. ■

A.3 Variance of total RNAP distribution

In this section, we derive the exact expression for the variance of the total RNAP distribution, as stated in Eq. (3.10), which is given by the sum over the covariances $\mathrm{Cov}(x_i, x_j)$ $(i, j = 1, \ldots, L)$, as defined in Eq. (3.4d). Hence, we have

$$\begin{aligned} \mathrm{Var}(n_{tot}) &= \sum_{i,j=1}^{L}\mathrm{Cov}(n_i, n_j) = \sum_{i=1}^{L}\mathrm{Var}(n_i) + \sum_{i\neq j}\mathrm{Cov}(n_i, n_j) \\ &= \sum_{i=1}^{L}\big[\langle n_i\rangle + \langle n_i\rangle^2\alpha\beta\cdot f_{ii}\big] + \sum_{i\neq j}\langle n_i\rangle\langle n_j\rangle\alpha\beta\cdot f_{ij} \\ &= \sum_{i=1}^{L}\langle n_i\rangle + \alpha\beta\Big(\sum_{i=1}^{L}\langle n_i\rangle^2\cdot f_{ii} + \sum_{i\neq j}\langle n_i\rangle\langle n_j\rangle\cdot f_{ij}\Big) \\ &= \sum_{i=1}^{L}\langle n_i\rangle + \alpha\beta\sum_{i,j=1}^{L}\langle n_i\rangle\langle n_j\rangle\cdot f_{ij}, \end{aligned} \tag{A.3.1}$$

where the function $\tilde{f}_{ij}$ is given in Eq. (3.10). The first term in Eq. (A.3.1) equals $\langle n_{tot}\rangle$, the mean of the total RNAP distribution, as stated in Eq. (3.10); substituting in the expressions for the means $\langle n_i\rangle$ from Eq. (3.2b), as well, we obtain

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + \alpha\beta(\eta\rho_k)^2\sum_{i,j=1}^{L}\mu^{i+j}\cdot f_{ij}. \tag{A.3.2}$$

Lemma A.3.1. *In the limit of deterministic elongation, i.e. for $L \to \infty$, the expression for $\mathrm{Var}(n_{tot})$ in Eq. (3.10) simplifies to*

$$\mathrm{Var}(n_{tot})_\infty = \langle n_{tot}\rangle_\infty + \beta(\eta r)^2\frac{(s_b+s_u-d) - (s_b+s_u+d)e^{-2d\langle T\rangle} + 2de^{-(s_b+s_u+d)\langle T\rangle}}{d(s_b+s_u+d)(s_b+s_u-d)}, \tag{A.3.3}$$

which can be further simplified to the expression in Eq. (3.11).

Proof. In order to find the limit of $L \to \infty$ in Eq. (3.10) (or Eq. (A.3.2)), we have to evaluate the term $\sum_{i,j=1}^{L} \mu^{i+j} \cdot f_{ij}$ in that limit. For the following derivation, we consider the function $f_{ij} = f(i,j) + f(j,i)$, where $f(i,j)$ is defined in in terms of sums in Eq. (A.2.34). Hence, we have

$$\begin{aligned}\sum_{i,j=1}^{L} \mu^{i+j} \cdot f_{ij} &= \sum_{i,j=1}^{L} \mu^{i+j} f(i,j) + \sum_{i,j=1}^{L} \mu^{i+j} f(j,i) = 2 \sum_{i,j=1}^{L} \mu^{i+j} f(i,j) \\ &= 2\Big[\sum_{i,j=1}^{L} \sum_{q=0}^{j-1} \mu^{i+j} \binom{i+q}{q} \frac{\alpha^{j-q-1}}{2^{i+q}} - \sum_{i,j=1}^{L} \sum_{q=0}^{j-2} \mu^{i+j} \binom{i+q}{q} \frac{\alpha^{j-q-2}}{2^{i+q+1}} \Big] \\ &= \underbrace{2 \sum_{i,j=1}^{L} \mu^{i+j} \binom{i+j-1}{i} \frac{1}{2^{i+j-1}}}_{G_1} + \underbrace{2(2\alpha-1) \sum_{i,j=1}^{L} \sum_{q=0}^{j-2} \mu^{i+j} \binom{i+q}{q} \frac{\alpha^{j-q-2}}{2^{i+q+1}}}_{G_2}.\end{aligned} \tag{A.3.4}$$

Substituting $k \to L/\langle T \rangle - d$ in Eq. (A.3.4) and taking the limit of $L \to \infty$, we have that $G_1 \underset{L\to\infty}{\longrightarrow} 0$; hence, $\mathrm{Var}(n_{tot})$ evaluates to

$$\mathrm{Var}(n_{tot})_\infty = \langle n_{tot} \rangle_\infty + \lim_{L\to\infty} [\alpha\beta(\eta\rho_k)^2 G_2] \tag{A.3.5a}$$

in that limit, which yields the expression in Eq. (A.3.3), as can easily be verified with the computer algebra package Mathematica. Hence, in the limit of deterministic elongation, the expression for the variance of the RNAP distribution in Eq. (3.10) reduces to the one in Eq. (3.11), as claimed. ■

A.4 Moments of total RNAP and mature RNA in bursty and constitutive limits

Moments of total RNAP in the bursty limit. In the bursty limit, the expressions for the mean and variance of the total RNAP distribution given in Eq. (3.10) simplify to

$$\langle n_{tot} \rangle_b = b \frac{s_u}{k} \mu \frac{\mu^L - 1}{\mu - 1}, \quad \text{and} \quad \mathrm{Var}(n_{tot})_b = \langle n_{tot} \rangle_b + b^2 \frac{s_u}{k} \sum_{i,j=1}^{L} \mu^{i+j} \cdot h_{ij}. \tag{A.4.1}$$

If, furthermore, we take the limit of deterministic elongation, with $L \to \infty$ at constant $\langle T \rangle$, Eq. (A.4.1) simplifies to

$$\langle n_{tot} \rangle_{(b;\infty)} = b \frac{s_u}{d} (1 - e^{-T_d}) \quad \text{and} \quad \mathrm{Var}(n_{tot})_{(b;\infty)} = \langle n_{tot} \rangle_{(b;\infty)} + \langle n_{tot} \rangle^2_{(b;\infty)} \frac{d}{s_u} \frac{1 + e^{-T_d}}{1 - e^{-T_d}}, \tag{A.4.2}$$

where the subscript $(b;\infty)$ denotes the bursty limit with infinite L. In the limit of zero RNAP detachment, Eq. (A.4.2) further simplifies to

$$\langle n_{tot} \rangle_{(b;\infty;0)} = b s_u \langle T \rangle \quad \text{and} \quad \mathrm{Var}(n_{tot})_{(b;\infty;0)} = \langle n_{tot} \rangle_{(b;\infty;0)} (1 + 2b), \tag{A.4.3}$$

where the subscript $(b;\infty;0)$ denotes the bursty limit, with $L \to \infty$ and $d \to 0$.

Moments of total RNAP in the constitutive limit. In the constitutive limit, Eq. (3.10) simplifies to

$$\langle n_{tot} \rangle_c = \frac{r}{k} \mu \frac{\mu^L - 1}{\mu - 1} = \mathrm{Var}(n_{tot})_c. \tag{A.4.4}$$

If, furthermore, we take the limit of deterministic elongation, i.e. $L \to \infty$ at constant $\langle T \rangle$, Eq. (A.4.4) simplifies to

$$\langle n_{tot} \rangle_{(c;\infty)} = \frac{r}{d} (1 - e^{-T_d}) = \mathrm{Var}(n_{tot})_{(c;\infty)}; \tag{A.4.5}$$

finally, in the limit of zero RNAP detachment, Eq. (A.4.5) further simplifies to

$$\langle n_{tot}\rangle_{(c;\infty;0)} = r\langle T\rangle = \mathrm{Var}(n_{tot})_{(c;\infty;0)}. \tag{A.4.6}$$

Moments of mature RNA distribution in the bursty limit. In that limit, the closed-form expressions in Eq. (3.8) are given by

$$\langle n\rangle_b = b\upsilon_m\mu^L \quad \text{and} \quad \mathrm{Var}(n)_b = \langle n\rangle_b + \langle n\rangle_b^2(\upsilon_k\mu)^{-1}\cdot\tilde{h}_{MM}, \tag{A.4.7}$$

which in the limit of deterministic elongation simplify to

$$\langle n\rangle_{(b;\infty)} = b\upsilon_m e^{-T_d} \quad \text{and} \quad \mathrm{Var}(n)_{(b;\infty)} = \langle n\rangle_{(b;\infty)} + \langle n\rangle_{(b;\infty)}^2\upsilon_m^{-1}. \tag{A.4.8}$$

In the limit of zero RNAP detachment, these expressions further simplify to

$$\langle n\rangle_{(b;\infty;0)} = b\upsilon_m \quad \text{and} \quad \mathrm{Var}(n)_{(b;\infty;0)} = \langle n\rangle_{(b;\infty;0)} + \langle n\rangle_{(b;\infty;0)}^2\upsilon_m^{-1}. \tag{A.4.9}$$

A.5 Variance of fluctuating total fluorescent signal

By definition, the variance of the total fluorescent signal is given by the sum over all elements $\mathrm{Cov}(r_i, r_j)$ for $i, j = 1, \ldots, L$, where $r_i = (\nu/L)in_i$; the corresponding definitions can be found in Section 3.4 of the main text. Hence, we have that

$$\begin{aligned}
\mathrm{Var}(r_{tot}) &= \sum_{i,j=1}^{L}\mathrm{Cov}(r_i,r_j) = \sum_{i,j=1}^{L}\mathrm{Cov}\Big(\frac{\nu}{L}in_i,\frac{\nu}{L}jn_j\Big) = \Big(\frac{\nu}{L}\Big)^2\sum_{i,j=1}^{L}ij\cdot\mathrm{Cov}(n_i,n_j)\\
&= \Big(\frac{\nu}{L}\Big)^2\Big(\sum_{i=1}^{L}i^2\mathrm{Var}(n_i) + \sum_{i\neq j}ij\cdot\mathrm{Cov}(n_i,n_j)\Big)\\
&= \Big(\frac{\nu}{L}\Big)^2\Big(\sum_{i=1}^{L}i^2[\langle n_i\rangle + \langle n_i\rangle^2\alpha\beta\cdot f_{ii}] + \sum_{i\neq j}ij\langle n_i\rangle\langle n_j\rangle\alpha\beta\cdot f_{ij}\Big)\\
&= \Big(\frac{\nu}{L}\Big)^2\sum_{i=1}^{L}i^2\langle n_i\rangle + \Big(\frac{\nu}{L}\Big)^2\alpha\beta\Big(\sum_{i=1}^{L}i^2\langle n_i\rangle^2\cdot f_{ii} + \sum_{i\neq j}ij\langle n_i\rangle\langle n_j\rangle\cdot f_{ij}\Big)\\
&= \Big(\frac{\nu}{L}\Big)^2\sum_{i=1}^{L}i^2\langle n_i\rangle + \Big(\frac{\nu}{L}\Big)^2\alpha\beta\sum_{i,j=1}^{L}ij\langle n_i\rangle\langle n_j\rangle\cdot f_{ij}.
\end{aligned} \tag{A.5.1}$$

Substituting the expressions for the means $\langle n_i\rangle$ from Eq. (3.2b) into Eq. (A.5.1), we obtain

$$\mathrm{Var}(r_{tot}) = \Big(\frac{\nu}{L}\Big)^2\eta\rho_k\sum_{i=1}^{L}i^2\mu^i + \Big(\frac{\nu}{L}\Big)^2\alpha\beta(\eta\rho_k)^2\sum_{i,j=1}^{L}ij\cdot\mu^{i+j}\cdot f_{ij}, \tag{A.5.2}$$

which is the expression stated in Eq. (3.35).

A.6 Moments of fluctuations in total fluorescent signal in various limits

Deterministic elongation. Substituting $k \mapsto L/\langle T\rangle - d$ and taking the long-gene limit of $L \to \infty$ in Eq. (3.35), we obtain the simplified expressions

$$\begin{aligned}
\langle r_{tot}\rangle_\infty &= \frac{\nu\eta r}{dT_d}[1-(1+T_d)e^{-T_d}],\\
\mathrm{Var}(r_{tot})_\infty &= \langle r_{tot}\rangle_\infty\cdot\mathcal{F}_0 + \langle r_{tot}\rangle_\infty^2\cdot\beta\delta_g\frac{\mathcal{F}_1+\mathcal{F}_2+\mathcal{F}_3}{2(\delta_g-1)^2(\delta_g+1)^2[1-(1+T_d)e^{-T_d}]^2},
\end{aligned} \tag{A.6.1}$$

where

$$\begin{aligned}
\mathcal{F}_0 &= \nu\Big[\frac{2}{T_d} - \frac{T_d e^{-T_d}}{1-(1+T_d)e^{-T_d}}\Big],\\
\mathcal{F}_1 &= (\delta_g-1)^2(2\delta_g+1),\\
\mathcal{F}_2 &= (\delta_g+1)^2[2\delta_g(1+T_d)(1+T_d-T_g)-1]e^{-2T_d},\\
\mathcal{F}_3 &= -4\delta_g^3(1+T_d+T_g)e^{-T_g}e^{-T_d};
\end{aligned} \tag{A.6.2}$$

the expression for the variance in Eq. (A.6.1) is found via the same method as is used in Lemma A.3.1 of Appendix A.3. When there is no detachment of RNAP from the gene, i.e. when $d = 0$, Eq. (A.6.1) simplifies to

$$\begin{aligned}
\langle r_{tot}\rangle_{(\infty;0)} &= \frac{1}{2}\nu\eta r\langle T\rangle,\\
\mathrm{Var}(r_{tot})_{(\infty;0)} &= \langle r_{tot}\rangle_{(\infty;0)}\frac{2\nu}{3} + \langle r_{tot}\rangle^2_{(\infty;0)}\cdot 8\beta T_g^{-1}\Big[\frac{1}{3}-\frac{1}{2}T_g^{-1}+T_g^{-3}-T_g^{-3}\big(1+T_g\big)e^{-T_g}\Big].
\end{aligned} \tag{A.6.3}$$

Bursty limit. In the limit when the rates s_b and r are large, the expressions for the mean and variance of the total fluorescent signal given in Eq. (3.35) become

$$\begin{aligned}
\langle r_{tot}\rangle_b &= \nu b\frac{s_u}{d}\Big(\frac{k}{d}\frac{(1-\mu^L)}{\mu L}-\mu^L\Big),\\
\mathrm{Var}(r_{tot})_b &= \Big(\frac{\nu}{L}\Big)^2 b\frac{s_u}{k}\sum_{i=1}^{L} i^2\mu^i + \Big(\frac{\nu}{L}\Big)^2 b^2\frac{s_u}{k}\sum_{i,j=1}^{L} ij\cdot\mu^{i+j}\cdot\tilde{f}_{ij}.
\end{aligned} \tag{A.6.4}$$

Constitutive limit. When the gene spends most of its time in the active state, Eq. (3.35) simplifies to

$$\begin{aligned}
\langle r_{tot}\rangle_c &= \frac{\nu}{L}\rho_k\mu\frac{1+\mu^L[L(\mu-1)-1]}{(\mu-1)^2},\\
\mathrm{Var}(r_{tot})_c &= \Big(\frac{\nu}{L}\Big)^2\rho_k\mu\frac{1+\mu-\mu^L[L^2\mu^2+(1+L)^2\mu-(2L^2+2L-1)]}{(1-\mu)^3}.
\end{aligned} \tag{A.6.5}$$

Bursty expression with deterministic elongation. In this case, Eq. (A.6.4) simplifies to

$$\begin{aligned}
\langle r_{tot}\rangle_{(b;\infty)} &= \frac{\nu b s_u}{dT_d}\big[1-(1+T_d)e^{-T_d}\big],\\
\mathrm{Var}(r_{tot})_{(b;\infty)} &= \langle r_{tot}\rangle_{(b;\infty)}\cdot\mathcal{F}_0 + \langle r_{tot}\rangle^2_{(b;\infty)}\cdot\frac{d}{2s_u}\frac{1-(1+2T_d+2T_d^2)e^{-2T_d}}{\big[1-(1+T_d)e^{-T_d}\big]^2},
\end{aligned} \tag{A.6.6}$$

where $\mathcal{F}_0$ is given by Eq. (A.6.2). In the special case of no premature RNAP detachment from the gene ($d \to 0$), Eq. (A.6.6) can be further simplified to

$$\begin{aligned}
\langle r_{tot}\rangle_{(b;\infty;0)} &= \frac{1}{2}\nu b s_u\langle T\rangle,\\
\mathrm{Var}(r_{tot})_{(b;\infty;0)} &= \langle r_{tot}\rangle_{(b;\infty;0)}\cdot\frac{2\nu}{3} + \langle r_{tot}\rangle^2_{(b;\infty;0)}\cdot\frac{8}{3s_u\langle T\rangle}.
\end{aligned} \tag{A.6.7}$$

Constitutive expression with deterministic elongation. In this case, Eq. (A.6.5) simplifies to

$$\langle r_{tot}\rangle_{(c;\infty)} = \frac{\nu}{T_d}\frac{r}{d}\big[1-(1+T_d)e^{-T_d}\big] \quad\text{and}\quad \mathrm{Var}(r_{tot})_{(c;\infty)} = \frac{\nu^2}{T_d^2}\frac{r}{d}\big[2-(2+2T_d+T_d^2)e^{-T_d}\big], \tag{A.6.8}$$

which reduces to

$$\langle r_{tot}\rangle_{(c;\infty;0)} = \frac{1}{2}\nu r\langle T\rangle \quad\text{and}\quad \mathrm{Var}(r_{tot})_{(c;\infty;0)} = \frac{1}{3}\nu^2 r\langle T\rangle \tag{A.6.9}$$

for the special case of zero RNAP detachment from the gene.

A.7 Extended model with RNAP pausing

Proof of Proposition 3. The new pausing model presented in Fig. 3.9 can be conveniently described by $2L+2$ species interacting via an effective set of $5L+4$ reactions. The vector $\vec{m}$ of the number of molecules of the respective species is given by $\vec{m} = (n_0, n_1^a, \ldots, n_L^a, n_1^p, \ldots, n_L^p, n)$; in the table below, we summarize the respective positions of each entry in $\vec{m}$, as well as the definition of the rate functions f_j, for $j = 1, \ldots, 5L+4$.

Species	Molecule numbers	Position (in $\vec{m}$)
G_{on}	n_0	1
$P_i, \quad i \in \{1, \ldots, L\}$	n_i^a	$i+1$
$\bar{P}_i, \quad i \in \{1, \ldots, L\}$	n_i^p	$i+L+1$
M	n	$2L+2$

Reaction	Rate function f_j
$G_{\text{on}} \xrightarrow{s_b} G_{\text{off}}$	$f_1 = s_b\langle n_0\rangle$
$G_{\text{off}} \xrightarrow{s_u} G_{\text{on}}$	$f_2 = s_u(1 - \langle n_0\rangle)$
$G_{\text{on}} \xrightarrow{r} G_{\text{on}} + P_1$	$f_3 = r\langle n_0\rangle$
$P_i \xrightarrow{k} P_{i+1}, \quad i \in \{1, \ldots, L-1\}$	$f_{i+3} = k\langle n_i^a\rangle$
$P_L \xrightarrow{k} M$	$f_{L+3} = k\langle n_L^a\rangle$
$P_i \xrightarrow{d_a} \varnothing, \quad i \in \{1, \ldots, L\}$	$f_{i+L+3} = d_a\langle n_i^a\rangle$
$P_i \xrightarrow{r_p} \bar{P}_i, \quad i \in \{1, \ldots, L\}$	$f_{i+2L+3} = r_p\langle n_i^a\rangle$
$\bar{P}_i \xrightarrow{r_a} P_i, \quad i \in \{1, \ldots, L\}$	$f_{i+3L+3} = r_a\langle n_i^p\rangle$
$\bar{P}_i \xrightarrow{d_p} \varnothing, \quad i \in \{1, \ldots, L\}$	$f_{i+4L+3} = d_p\langle n_i^p\rangle$
$M \xrightarrow{d_m} \varnothing$	$f_{5L+4} = d_m\langle n\rangle$

Note that we do not consider G_{off} as an independent species, as a conservation law implies $\langle G_{\text{off}}\rangle = 1 - \langle n_0\rangle$. Given the ordering of species and reactions as described in above tables, we can define the $(2L+2) \times (5L+4)$-dimensional stoichiometry matrix $\mathbf{S}$, with non-zero elements given by

$$\begin{aligned}
\mathbf{S}_{11} &= -1, & \mathbf{S}_{12} &= 1, & & \\
\mathbf{S}_{i,i+1} &= 1, & \mathbf{S}_{i,i+2} &= -1, & \mathbf{S}_{i,i+L+2} &= -1, \\
\mathbf{S}_{i,i+2L+2} &= -1, & \mathbf{S}_{i,i+3L+2} &= 1, & & \\
\mathbf{S}_{i+L,i+2L+2} &= 1, & \mathbf{S}_{i+L,i+3L+2} &= -1, & \mathbf{S}_{i+L,i+4L+2} &= -1, \\
\mathbf{S}_{2L+2,L+3} &= 1, & \mathbf{S}_{2L+2,5L+4} &= -1, & &
\end{aligned} \tag{A.7.1}$$

where $i = 2, \ldots, L+1$. From the associated CME, it can be shown via the moment equations that the time evolution of the vector $\langle\vec{m}\rangle$ of mean molecule numbers in a system of reactions with propensities that are linear in the number of molecules is determined by $d\langle\vec{m}\rangle/dt = \mathbf{S}\cdot\vec{f}$. Given the form of the stoichiometric matrix $\mathbf{S}$ and of the rate functions f_j, it follows that the mean numbers of molecules of active gene, active and paused RNAP, and mature RNA in steady-state can be obtained by solving the following system of $2L+2$ algebraic equations:

$$\begin{aligned}
0 &= s_u(1 - \langle n_0\rangle) - s_b\langle n_0\rangle, & & \\
0 &= r\langle n_0\rangle - (k + d_a + r_p)\langle n_1^a\rangle + r_a\langle n_1^p\rangle, & & \\
0 &= k\langle n_{i-1}^a\rangle - (k + d_a + r_p)\langle n_i^a\rangle + r_a\langle n_i^p\rangle & & \text{for } i = 2, \ldots, L, \\
0 &= r_p\langle n_i^a\rangle - (r_a + d_p)\langle n_i^p\rangle & & \text{for } i = 1, \ldots, L, \\
0 &= k\langle n_L^a\rangle - d_m\langle n\rangle. & &
\end{aligned} \tag{A.7.2}$$

Here, we recall the definition of the following parameters from the main text: $\eta = s_u\tau_g$, where $\tau_g = 1/(s_u + s_b)$ is the gene switching timescale, $\rho_k = r/k$, and $\rho = r/d_m$. Also, we define several

new parameters: $\sigma = r_p/r_a$ as the ratio of the pausing and activation rates; $\pi_{r_a} = r_a/(r_a + d_p)$, which is the probability of RNAP switching to the active state; $\pi_{d_p} = d_p/(r_a + d_p)$, which is the probability of premature termination from the paused RNAP state; $\tilde{\mu} = k/(k + d_a + r_p\pi_{d_p})$; and $\lambda = \sigma\pi_{r_a}$. It follows that the solution of Eq. (A.7.2) can be written as

$$\langle n_0\rangle = \eta, \quad \langle n_i^a\rangle = \eta\rho_k\tilde{\mu}^i, \quad \langle n_i^p\rangle = \langle n_i^a\rangle\lambda, \quad \text{and} \quad \langle n\rangle = \eta\rho\tilde{\mu}^L. \tag{A.7.3}$$

■

Proof of Proposition 4. In order to solve the Lyapunov equation $\mathbf{J}\cdot\mathbf{C} + \mathbf{C}\cdot\mathbf{J}^T + \mathbf{D} = \mathbf{0}$ for the symmetric elements $\mathbf{C}_{ij} = \mathbf{C}_{ji}$ of the $(2L+2)\times(2L+2)$-dimensional covariance matrix $\mathbf{C}$, we will follow the same approach as in Appendix A.2. First, we define the $(2L+2)\times(2L+2)$-dimensional Jacobian and diffusion matrices for our system. The Jacobian matrix $\mathbf{J}$ has the following non-zero elements,

$$\begin{aligned}
&\mathbf{J}_{11} = -(s_u + s_b), && && && \\
&\mathbf{J}_{21} = r, && \mathbf{J}_{22} = -(k + d_a + r_p), && \mathbf{J}_{2,2+L} = r_a, && \\
&\mathbf{J}_{i,i-1} = k, && \mathbf{J}_{ii} = -(k + d_a + r_p), && \mathbf{J}_{i,i+L} = r_a && \text{for } i = 3,\dots,L+1,\\
&\mathbf{J}_{i+L,i} = r_p, && \mathbf{J}_{i+L,i+L} = -(r_a + d_p) && && \text{for } i = 2,\dots,L+1,\\
&\mathbf{J}_{2L+2,L+1} = k, && \mathbf{J}_{2L+2,2L+2} = -d_m, && &&
\end{aligned} \tag{A.7.4}$$

while the non-zero elements of the symmetric diffusion matrix $\mathbf{D}$ are given by

$$\begin{aligned}
&\mathbf{D}_{11} = s_u(1 - \langle n_0\rangle) + s_b\langle n_0\rangle,\\
&\mathbf{D}_{22} = r\langle n_0\rangle + (k + d_a + r_p)\langle n_1^a\rangle + r_a\langle n_1^p\rangle, \qquad \mathbf{D}_{23} = -k\langle n_1^a\rangle, \qquad \mathbf{D}_{2,2+L} = -r_p\langle n_1^a\rangle - r_a\langle n_1^p\rangle;\\
&\text{for } \ i = 3,\dots,L+1:\\
&\mathbf{D}_{ii} = k\langle n_{i-2}^a\rangle + (k + d_a + r_p)\langle n_{i-1}^a\rangle + r_a\langle n_{i-1}^p\rangle, \qquad \mathbf{D}_{i,i+1[i\le L]} = -k\langle n_{i-1}^a\rangle, \quad \mathbf{D}_{L+1,2L+2} = -k\langle n_L^a\rangle,\\
&\mathbf{D}_{i,i+L} = -r_p\langle n_{i-1}^a\rangle - r_a\langle n_{i-1}^p\rangle;\\
&\text{for } \ i = 2,\dots,L+1:\\
&\mathbf{D}_{i+L,i+L} = r_p\langle n_{i-1}^a\rangle + (r_a + d_p)\langle n_{i-1}^p\rangle,\\
&\mathbf{D}_{2L+2,2L+2} = k\langle n_L^a\rangle + d_m\langle n\rangle.
\end{aligned} \tag{A.7.5}$$

Next, using the definition of $\mathbf{J}$ and $\mathbf{D}$ from Eqs. (A.7.4) and(A.7.5), respectively, we solve the Lyapunov equation. Here, we note that we are only interested in expressions for the covariances of fluctuations in active and paused RNAP, but not of mature RNA fluctuations; hence, we require closed-form expressions for the elements $\mathbf{C}_{ij}$ with $i, j \neq 2L+2$, which we derive by following the same procedure as in Appendix A.2.

Now, we recall that $\beta = s_b/s_u$ is the ratio of gene deactivation and activation rates, while $\tau_p = 1/(k + d_a)$ is the typical time that an actively moving RNAP spends on a gene segment. Additionally, let $\tau_{r_a} = 1/r_a$ be the timescale of RNAP activation from the paused state, let $\tau_{d_p} = 1/d_p$ be the timescale of premature termination of paused RNAP, and let $\tau_{pp} = 1/(r_a + d_p)$ be the typical time spent in the paused state. Finally, we define the following new parameters: $\lambda_{r_p} = \pi_{r_p}/(1 - \pi_{r_p})$, where $\pi_{r_p} = r_p/(r_p + k + d_a)$ is the probability of actively moving RNAP switching to the paused state, as well as

$$\omega_{r_a} = \frac{\pi_{r_a}\tau_g}{\pi_{r_a}\tau_{r_a} + \tau_g}, \quad \tilde{\alpha} = \frac{\tau_g + \lambda_{r_p}\pi_{d_p}\tau_g}{\tau_g + \tau_p + \lambda_{r_p}\tau_g(1 - \omega_{r_a})}, \quad \text{and} \quad \omega = \frac{\tau_g}{\tau_{pp} + \tau_g}; \tag{A.7.6}$$

then, closed-form expressions for the covariances of the active gene with itself and the remaining

species are given by

$$\begin{aligned}
&\mathrm{Var}(n_0) = \eta^2\beta \cdot g_{00}^{aa}, && \text{where } g_{00}^{aa} = 1,\\
&\mathrm{Cov}(n_0, n_j^a) = \eta\langle n_j^a\rangle\tilde{\alpha}\beta \cdot g_{0j}^{aa}, && \text{where } g_{0j}^{aa} = \tilde{\alpha}^{j-1},\\
&\mathrm{Cov}(n_0, n_j^p) = \eta\langle n_j^p\rangle\tilde{\alpha}\beta \cdot g_{0j}^{ap}, && \text{where } g_{0j}^{ap} = \omega\tilde{\alpha}^{j-1}.
\end{aligned} \tag{A.7.7}$$

Similarly, closed-form expressions for the covariances between all RNAP species read

$$\begin{aligned}
\mathrm{Cov}(n_i^a, n_j^a) &= \delta_{ij}\langle n_i^a\rangle + \langle n_i^a\rangle\langle n_j^a\rangle\tilde{\alpha}\beta \cdot g_{ij}^{aa},\\
\mathrm{Cov}(n_i^a, n_j^p) &= \langle n_i^a\rangle\langle n_j^p\rangle\tilde{\alpha}\beta \cdot g_{ij}^{ap},\\
\mathrm{Cov}(n_i^p, n_j^a) &= \langle n_i^p\rangle\langle n_j^a\rangle\tilde{\alpha}\beta \cdot g_{ij}^{pa},\\
\mathrm{Cov}(n_i^p, n_j^p) &= \delta_{ij}\langle n_i^p\rangle + \langle n_i^p\rangle\langle n_j^p\rangle\tilde{\alpha}\beta \cdot g_{ij}^{pp},
\end{aligned} \tag{A.7.8}$$

where the functions $g_{ij}^{aa} = g_{ji}^{aa}$, $g_{ij}^{ap} = g_{ji}^{pa}$, and $g_{ij}^{pp} = g_{ji}^{pp}$ satisfy the following recurrence relations:

$$\begin{aligned}
g_{ij}^{aa} &= \frac{[(k+d_a)(r_a+d_p)+r_pd_p](g_{i-1,j}^{aa}+g_{i,j-1}^{aa})+r_ar_p(g_{ij}^{ap}+g_{ij}^{pa})}{2(k+d_a+r_p)(r_a+d_p)},\\
g_{ij}^{ap} &= \frac{[(k+d_a)(r_a+d_p)+r_pd_p]g_{i-1,j}^{ap}+(r_a+d_p)^2g_{ij}^{aa}+r_ar_pg_{ij}^{pp}}{(k+d_a+r_a+r_p+d_p)(r_a+d_p)},\\
g_{ij}^{pp} &= \frac{g_{ij}^{ap}+g_{ij}^{pa}}{2}.
\end{aligned} \tag{A.7.9}$$

Now, we assume that the elongation rate is faster than the rates of RNAP pausing, activation, and premature termination, i.e. that $k \gg r_a, r_p, d_a, d_p$ in Eq. (A.7.9). Taking the limit of $k \to \infty$, we find that the expressions in Eqs. (A.7.7) and (A.7.8) remain unchanged, while Eq. (A.7.9) simplifies to

$$g_{ij}^{aa} = (g_{i-1,j}^{aa} + g_{i,j-1}^{aa})/2, \tag{A.7.10a}$$

$$g_{ij}^{ap} = g_{i-1,j}^{ap}, \tag{A.7.10b}$$

$$g_{ij}^{pp} = (g_{ij}^{ap} + g_{ij}^{pa})/2; \tag{A.7.10c}$$

in particular, to leading order in $1/k$, the functions g_{ij}^{aa}, g_{ij}^{ap}, g_{ij}^{pa}, and g_{ij}^{pp} hence do not depend on k. Eq. (A.7.10a) defines a recurrence relation for the symmetric function $g_{ij}^{aa} = g_{ji}^{aa}$ with initial conditions g_{00}^{aa} and g_{0j}^{aa} from Eq. (A.7.7). Using the same mathematical technique as in Lemma A.2.5, we find that the solution for the function g_{ij}^{aa} is given by $g_{ij}^{aa} = g^{aa}(i,j) + g^{aa}(j,i)$, where

$$g^{aa}(i,j) = \frac{\tilde{\alpha}^{i+j-1}}{(2\tilde{\alpha}-1)^i} + \frac{1}{2^{i+j-1}}\binom{i+j-1}{i}\Big[1 - \frac{2\tilde{\alpha}-1}{2\tilde{\alpha}}{}_2F_1\big(1, i+j; j; \tfrac{1}{2\tilde{\alpha}}\big)\Big]; \tag{A.7.11}$$

Eq. (A.7.10b) is a recurrence relation for the function g_{ij}^{ap} with initial conditions g_{0j}^{ap} from Eq. (A.7.7); the corresponding solution is then given by $g_{ij}^{ap} = \omega\tilde{\alpha}^{j-1}$. Finally, the solution of the recurrence relation in Eq. (A.7.10c) for g_{ij}^{pp} is given by $g_{ij}^{pp} = \omega(\tilde{\alpha}^{j-1} + \tilde{\alpha}^{i-1})/2$. In sum, the leading-order asymptotics (in $1/k$) of the covariances between the various RNAP species for k large is hence given by Eq. (A.7.8), with g_{ij}^{aa}, $g_{ij}^{ap} = g_{ij}^{pa}$, and g_{ij}^{pp} as stated above. ■

Asymptotics of variance of total RNAP distribution. The variance of the total RNAP distribution for the pausing model is given by

$$\mathrm{Var}(n_{tot}) = \sum_{i,j=1}^{L}\big(\mathrm{Cov}(n_i^a, n_j^a) + \mathrm{Cov}(n_i^a, n_j^p) + \mathrm{Cov}(n_i^p, n_j^a) + \mathrm{Cov}(n_i^p, n_j^p)\big), \tag{A.7.12}$$

where the expressions for the corresponding covariances are given in Eq. (3.39). In order to simplify the above expression, we consider each term on the right-hand side in Eq. (A.7.12) separately, as follows:

$$\begin{aligned}
\sum_{i,j=1}^{L} \mathrm{Cov}(n_i^a, n_j^a) &= \sum_{i,j=1}^{L} \delta_{ij}\langle n_i^a\rangle + (\eta\rho_k)^2\tilde{\alpha}\beta \sum_{i,j=1}^{L} g_{ij}^{aa},\\
\sum_{i,j=1}^{L} \mathrm{Cov}(n_i^a, n_j^p) &= (\eta\rho_k)^2\tilde{\alpha}\beta\lambda \sum_{i,j=1}^{L} g_{ij}^{ap},\\
\sum_{i,j=1}^{L} \mathrm{Cov}(n_i^p, n_j^a) &= (\eta\rho_k)^2\tilde{\alpha}\beta\lambda \sum_{i,j=1}^{L} g_{ij}^{pa},\\
\sum_{i,j=1}^{L} \mathrm{Cov}(n_i^p, n_j^p) &= \sum_{i,j=1}^{L} \delta_{ij}\langle n_i^p\rangle + (\eta\rho_k)^2\tilde{\alpha}\beta\lambda^2 \sum_{i,j=1}^{L} g_{ij}^{pp}.
\end{aligned} \tag{A.7.13}$$

Since $\sum_{i,j=1}^{L}\left(\delta_{ij}\langle n_i^a\rangle + \delta_{ij}\langle n_i^p\rangle\right) = \sum_{i=1}^{L}\langle n_i\rangle = \langle n_{tot}\rangle$, Eq. (A.7.12) becomes

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + (\eta\rho_k)^2\tilde{\alpha}\beta \sum_{i,j=1}^{L}\left(g_{ij}^{aa} + \lambda g_{ij}^{ap} + \lambda g_{ij}^{pa} + \lambda^2 g_{ij}^{pp}\right). \tag{A.7.14}$$

Using the expressions for the functions g_{ij}^{aa}, g_{ij}^{ap}, g_{ij}^{pa}, and g_{ij}^{pp} from Eq. (3.39), we conclude that Eq. (A.7.14) further simplifies to

$$\mathrm{Var}(n_{tot}) = \langle n_{tot}\rangle + (\eta\rho_k)^2\tilde{\alpha}\beta\left[2\sum_{i,j=1}^{L} g^{aa}(i,j) + \lambda(2+\lambda)\omega L\frac{\tilde{\alpha}^L - 1}{\tilde{\alpha} - 1}\right]. \tag{A.7.15}$$

A.8 Approximation of mature RNA distribution in extended model

Similarly to Section 3.3.2, we apply geometric singular perturbation theory (GSPT) to formally derive the distribution of mature RNA for the extended pausing model. As was done there, we define $P_j(\vec{n};t)$ $(j = 0, 1)$ as the probability of the state $\vec{n} = (n_1^a, \ldots, n_L^a, n_1^p, \ldots, n_L^p, n)$ at time t while the gene is either active (0) or inactive (1); then, the time evolution of these probabilities can be described by a system of coupled CMEs:

$$\begin{aligned}
\partial_t P_0 = {} & s_u P_1 - s_b P_0 + r(\mathbb{E}_{n_1}^{-1} - 1)P_0 + k\sum_{i=1}^{L-1}(\mathbb{E}_{n_i^a}\mathbb{E}_{n_{i+1}^a}^{-1} - 1)n_i^a P_0 + k(\mathbb{E}_{n_L^a}\mathbb{E}_{n}^{-1} - 1)n_L^a P_0\\
& + d_a\sum_{i=1}^{L}(\mathbb{E}_{n_i^a} - 1)n_i^a P_0 + r_p\sum_{i=1}^{L}(\mathbb{E}_{n_i^a}\mathbb{E}_{n_i^p}^{-1} - 1)n_i^a P_0 + r_a\sum_{i=1}^{L}(\mathbb{E}_{n_i^p}\mathbb{E}_{n_i^a}^{-1} - 1)n_i^p P_0\\
& + d_p\sum_{i=1}^{L}(\mathbb{E}_{n_i^p} - 1)n_i^p P_0 + d_m(\mathbb{E}_n - 1)nP_0,\\
\partial_t P_1 = {} & s_b P_0 - s_u P_1 + k\sum_{i=1}^{L-1}(\mathbb{E}_{n_i^a}\mathbb{E}_{n_{i+1}^a}^{-1} - 1)n_i^a P_1 + k(\mathbb{E}_{n_L^a}\mathbb{E}_{n}^{-1} - 1)n_L^a P_1\\
& + d_a\sum_{i=1}^{L}(\mathbb{E}_{n_i^a} - 1)n_i^a P_1 + r_p\sum_{i=1}^{L}(\mathbb{E}_{n_i^a}\mathbb{E}_{n_i^p}^{-1} - 1)n_i^a P_1 + r_a\sum_{i=1}^{L}(\mathbb{E}_{n_i^p}\mathbb{E}_{n_i^a}^{-1} - 1)n_i^p P_1\\
& + d_p\sum_{i=1}^{L}(\mathbb{E}_{n_i^p} - 1)n_i^p P_1 + d_m(\mathbb{E}_n - 1)nP_1.
\end{aligned} \tag{A.8.1}$$

In order to find analytical expressions for the propagator probabilities $P(\vec{n};t)$ which satisfy the system of CMEs in Eq. (A.8.1), we define the probability-generating functions $F_j(\vec{z};t)$, where $\vec{z} = (z_1^a, \ldots, z_L^a, z_1^p, \ldots, z_L^p, z)$ is a vector of variables corresponding to the state $\vec{n}$. Given the equations for $P_j(\vec{n};t)$ from Eq. (A.8.1), we obtain the following system of PDEs for the corresponding generating functions $F_j(\vec{z};t)$:

$$\begin{aligned} \mathbb{L}[F_0] &= s_u F_1 - s_b F_0 + r(z_1^a - 1)F_0, \\ \mathbb{L}[F_1] &= s_b F_0 - s_u F_1; \end{aligned} \tag{A.8.2}$$

here,

$$\begin{aligned} \mathbb{L} = \partial_t + d_m(z-1)\partial_z + k\sum_{i=1}^{L-1}(z_i^a - z_{i+1}^a)\partial_{z_i^a} + k(z_L^a - z)\partial_{z_L^a} + d_a\sum_{i=1}^{L}(z_i^a - 1)\partial_{z_i^a} \\ + r_p\sum_{i=1}^{L}(z_i^a - z_i^p)\partial_{z_i^a} + r_a\sum_{i=1}^{L}(z_i^p - z_i^a)\partial_{z_i^p} + d_p\sum_{i=1}^{L}(z_i^p - 1)\partial_{z_i^p} \end{aligned} \tag{A.8.3}$$

is a differential operator acting on the functions F_0 and F_1. Eq. (A.8.2) represents a system of coupled, linear, first-order PDEs. Now, we introduce new variables $u_i^a = z_i^a - 1$, $u_i^p = z_i^p - 1$, and $u = z - 1$; we also rescale all rates and the time variable with the degradation rate d_m of mature RNA. Next, we apply the method of characteristics, with s being the characteristic variable. The first characteristic equation will give us $d_m(dt/ds) = 1$, with solution $s \equiv d_m t$; hence, we can use the variable $t' = d_m t$ as the independent characteristic variable and thus convert the system of PDEs in Eq. (A.8.2) into a characteristic system of ODEs:

$$\dot{u}_i^a = (k/d_m)[(u_i^a - u_{i+1}^a) + (d_a/k)u_i^a + (r_p/k)(u_i^a - u_i^p)] \qquad \text{for } i = 1, \ldots, L-1, \tag{A.8.4a}$$

$$\dot{u}_L^a = (k/d_m)[(u_L^a - u) + (d_a/k)u_L^a + (r_p/k)(u_L^a - u_L^p)], \tag{A.8.4b}$$

$$\dot{u}_i^p = (r_a/d_m)[(u_i^p - u_i^a) + (d_p/r_a)u_i^p] \qquad \text{for } i = 1, \ldots, L, \tag{A.8.4c}$$

$$\dot{u} = u, \tag{A.8.4d}$$

$$\dot{F}_0 = (s_u/d_m)F_1 - (s_b/d_m)F_0 + (r/d_m)u_1^a F_0, \tag{A.8.4e}$$

$$\dot{F}_1 = (s_b/d_m)F_0 - (s_u/d_m)F_1, \tag{A.8.4f}$$

where the overdot denotes differentiation with respect to t. Here, we assume that $k/d_m \gg 1$ and $r_a/d_m \gg 1$; hence, we define $\varepsilon = d_m/k$ as the singular perturbation parameter, and we write $d_m/r_a = \varepsilon\delta$, where $\delta = k/r_a = \mathcal{O}(1)$ by assumption. Since $0 < \varepsilon \ll 1$ is small, we can apply GSPT in order to separate the system in Eq. (A.8.4) into fast and slow dynamics, which will allow us to find an asymptotic approximation for F_0 and F_1 in steady-state. With the above definitions, the governing equations for u_i^a and u_i^p in the 'slow system' in Eqs. (A.8.4a) through (A.8.4c) become

$$\varepsilon\dot{u}_i^a = (u_i^a - u_{i+1}^a) + (d_a/k)u_i^a + (r_p/k)(u_i^a - u_i^p) \qquad \text{for } i = 1, \ldots, L-1, \tag{A.8.5a}$$

$$\varepsilon\dot{u}_L^a = (u_L^a - u) + (d_a/k)u_L^a + (r_p/k)(u_L^a - u_L^p), \tag{A.8.5b}$$

$$\varepsilon\dot{u}_i^p = [(u_i^p - u_i^a) + (d_p/r_a)u_i^p]/\delta \qquad \text{for } i = 1, \ldots, L. \tag{A.8.5c}$$

It follows that u_i^a and u_i^p $(i = 1, \ldots, L)$ are the fast variables in our system, while u, F_0, and F_1 are the slow ones. Setting $\varepsilon = 0$ and solving the system in Eq. (A.8.5), we find $u_1^a = \tilde{\mu}^L \cdot u$, where $\tilde{\mu} = k/(k + d_a + r_p\pi_{d_p})$ has previously been defined in Prop. 3. Now, given Eq. (A.8.4d), we apply the chain rule, $dt' \equiv du \cdot u$, to rewrite Eqs. (A.8.4e) and (A.8.4f) as:

$$F_0' d_m u = s_u F_1 - s_b F_0 + r\tilde{\mu}^L u F_0, \tag{A.8.6a}$$

$$F_1' d_m u = s_b F_0 - s_u F_1, \tag{A.8.6b}$$

where the prime now denotes differentiation with respect to u. The system in Eq. (A.8.6) is the same as that in Eq. (3.28), with the substitution $\mu \mapsto \tilde{\mu}$; hence, following the same derivation

as in Section 3.3.2, we conclude that the steady-state analytical expression for the probability distribution of mature RNA is given by

$$P(n) = \frac{1}{n!}\frac{(s_u)_n}{(s_b+s_u)_n}\Big(\frac{r}{d_m}\Big)^n(\tilde{\mu}^L)^n\,{}_1F_1\Big(\frac{s_u}{d_m}+n;\frac{s_b+s_u}{d_m}+n;-\frac{r}{d_m}\tilde{\mu}^L\Big). \tag{A.8.7}$$

Appendix B

Supplementary Information for Chapter 4

B.1 Parameter values and other details of the figures

Fig. 4.2For all panels, we have arbitrarily chosen the number of mRNA stages to be $L = 30$. In panels (c) and (d), we have used the parameter $\sigma = 1/\varepsilon$, with $\varepsilon = 0.9$. In panels (e) and (f), we present our results for the time point $t = 10$ min. We also note that for the case of $\varepsilon = 0$, which indicates a constant signal, both the Fano factor and the noise are independent of time. The parameter values that have been used in all panels are $\sigma = 1\ \text{min}^{-1}$, $\varepsilon = 0.9$, $b = 3$, $\varphi = \pi/2$, $\omega = 1\ \text{min}^{-1}$, and $k_j = 5\ \text{min}^{-1}$ $(j = 0, \ldots, L)$, unless otherwise stated.

Fig. 4.3The parameter values that have been used in all panels are $b = 10$, $\varepsilon = 0.5$, $\sigma = 5\ \text{min}^{-1}$, $\omega = 1\ \text{min}^{-1}$, and $\varphi = \pi/2$, as well as $k_1 = 4.2,\ k_2 = 2.6,\ k_3 = 2.4,\ k_4 = 3.7,\ k_5 = 4.8,\ k_6 = 2.7,\ k_7 = 2.3,\ k_8 = 3.7$, and $k_9 = 3.0$, all of which have units of min^{-1}.

Fig. 4.4The parameter values that have been used in panel (a) are $b = 50$, $\omega = 1\ \text{min}^{-1}$, $\varepsilon = 0.99$, $t = 9.2$ min, and $j = 2$. The parameter values that have been used in panel (b) are $b = 150$, $\omega = 0.22\ \text{min}^{-1}$, $\varepsilon = 0.995$, $t = 9.9$ min, and $j = 8$. The parameter values that have been used in panel (c) are $b = 150$, $\omega = 1\ \text{min}^{-1}$, $\varepsilon = 0.99$, $t = 9.2$ min, and $j = 5$. In order to obtain the simulated data (points) in panel (d), we performed stochastic simulations over a range of parameter space $(b, \omega, \varepsilon) \in [1, 150] \times [0, 5] \times [0, 1]$, for $L = 10$ mRNA life-cycle stages and for two time points, $t = 9.2$ min and $t = 9.9$ min. Then, we randomly chose 40 points to present from the resulting data. The remaining parameters that have been used for panels (a) through (d) are the same, and are given by $\sigma = 5\ \text{min}^{-1}$ and $\varphi = \pi/2$, as well as by $k_1 = 4.2,\ k_2 = 2.6,\ k_3 = 2.4,\ k_4 = 3.7,\ k_5 = 4.8,\ k_6 = 2.7,\ k_7 = 2.3,\ k_8 = 3.7$, and $k_9 = 3.0$, all of which have units of min^{-1}.

Fig. 4.5The parameter values in all panels are as in Fig. 4.3, with the exception of $b = 20$ and $\omega = 0.5\ \text{min}^{-1}$.

B.2 Equivalence of the full model and the reduced model under timescale separation

In this section, we will show that in the limit of $\{r_u,\ s_u\} \gg \{s_b(t),\ k_1, \ldots, k_L\}$, the mRNA distributions obtained from the RM are exactly the same as those in the FM. For the FM, we consider the system of chemical reactions given in Eq. (4.2). We define $P_0(\vec{n}; t)$ to be the probability of finding $\vec{n}$ molecules in the system at time t when promoter is active and $P_1(\vec{n}; t)$ to be the

probability when it is inactive. The vector of the number of mRNA molecules in each life-cycle stage is defined as $\vec{n} = (n_1, \ldots, n_L)$. The CME is then given by the set of coupled equations

$$\begin{aligned}
\partial_t P_0 &= s_b(t)P_1 - s_u P_0 + r_u(\mathbb{E}_{n_1}^{-1} - 1)P_0 + \sum_{j=1}^{L-1} k_j(\mathbb{E}_{n_j}\mathbb{E}_{n_{j+1}}^{-1} - 1)n_j P_0 + k_L(\mathbb{E}_{n_L} - 1)n_L P_0, \\
\partial_t P_1 &= s_u P_0 - s_b(t)P_1 + \sum_{j=1}^{L-1} k_j(\mathbb{E}_{n_j}\mathbb{E}_{n_{j+1}}^{-1} - 1)n_j P_1 + k_L(\mathbb{E}_{n_L} - 1)n_L P_1,
\end{aligned} \tag{B.2.1}$$

where $\mathbb{E}_{n_i}^{c}[f(\vec{n})] = f(n_1, n_2, \ldots, n_i + c, \ldots, n_L)$, with $c \in \mathbb{Z}$, denotes the standard step operator. Now, we define the corresponding probability-generating functions as

$$F_q(\vec{u}; t) = \sum_{n_1,\ldots,n_L=0}^{\infty} P_j(\vec{n}; t)(u_1 + 1)^{n_1} \ldots (u_L + 1)^{n_L}, \tag{B.2.2}$$

for $q = 0, 1$; here, $\vec{u} = (u_1, \ldots, u_L)$ is a vector of real variables corresponding to the state $\vec{n}$, with $u_q \in [-1, 0]$ for $q = 1, \ldots, L$. Hence, we can rewrite the above CME as the system of PDEs

$$\begin{aligned}
\partial_t F_0 + \sum_{j=1}^{L-1} k_j(u_j - u_{j+1})\partial_{u_j} F_0 + k_L u_L \partial_{u_L} F_0 &= s_b(t)F_1 - s_u F_0 + r_u u_1 F_0, \\
\partial_t F_1 + \sum_{j=1}^{L-1} k_j(u_j - u_{j+1})\partial_{u_j} F_1 + k_L u_L \partial_{u_L} F_1 &= s_u F_0 - s_b(t)F_1.
\end{aligned} \tag{B.2.3}$$

Next, using the method of characteristics, we convert the above PDEs into the following system of ordinary differential equations (ODEs):

$$\begin{aligned}
\partial_s t &= 1, \quad \text{which implies } \; t = s, \\
\partial_s u_j &= k_j(u_j - u_{j+1}) \quad \text{for } \; j = 1, \ldots, L-1, \\
\partial_s u_L &= k_L u_L, \quad \text{which implies } \; u_L = u_L^0 e^{k_L s}, \; \text{ with } \; u_L^0 = u_L(0), \\
\partial_s F_0 &= s_b(s)F_1 - s_u F_0 + r_u u_1 F_0, \\
\partial_s F_1 &= s_u F_0 - s_b(s)F_1,
\end{aligned} \tag{B.2.4}$$

where $s \in \mathbb{R}$ is the characteristic variable. Using $x(s) = 1 + \varepsilon\cos(\omega s + \varphi)$, we rewrite the above ODEs for the generating functions as

$$\begin{aligned}
\tfrac{\delta}{\sigma}\partial_s F_0 &= \delta x(s)F_1 - F_0 + b u_1 F_0, \\
\tfrac{\delta}{\sigma}\partial_s F_1 &= F_0 - \delta x(s)F_1.
\end{aligned} \tag{B.2.5}$$

where $b = r_u/s_u$ and $\delta = \sigma/s_u$. We assume that $s_u \gg 2\sigma$, which implies that the promoter spends most of its time in the inactive state, since $s_u \gg 2\sigma \geq s_b(t)$. It follows that $\delta \ll 1$ can be taken as a small perturbation parameter. Also, we assume that the parameter r_u is of the same order of magnitude as s_u such that b remains constant as δ becomes very small. Here, we note that by assuming $r, s_u \gg s_b(t)$, we automatically also assume that the parameters k_j $(j = 1, \ldots, L)$ are of the same order of magnitude as the parameter σ: $r, s_u \gg k_j$. We may then take F_q $(q = 0, 1)$ to have a series expansion in δ:

$$F_q = F_q^{(0)} + \delta F_q^{(1)} + \mathcal{O}(\delta^2). \tag{B.2.6}$$

Substituting the above expansion into Eq. (B.2.5) and collecting leading-order terms in δ, i.e. terms of the order δ^0, we obtain $F_0^{(0)} = 0$. Similarly, collecting first-order terms in δ, we find the system

$$\begin{aligned}
0 &= x(s)F_1^{(0)} - F_0^{(1)} + b u_1 F_0^{(1)}, \\
\tfrac{1}{\sigma}\partial_s F_1^{(0)} &= F_0^{(1)} - x(s)F_1^{(0)}.
\end{aligned} \tag{B.2.7}$$

Using $F = F_0 + F_1$ and Eq. (B.2.7) we obtain the following ODE for $F^{(0)}$:

$$\partial_s F^{(0)} = s_b(s)\frac{bu_1}{1-bu_1}F^{(0)}. \tag{B.2.8}$$

Eq. (B.2.8), together with Eq. (B.2.4), then gives us the following system:

$$\begin{aligned}
\partial_s t &= 1 \quad \text{which implies } \; t = s, \\
\partial_s u_j &= k_j(u_j - u_{j+1}) \quad \text{for } \; j = 1, \dots, L-1, \\
\partial_s u_L &= k_L u_L \quad \text{which implies } \; u_L = u_L^0 e^{k_L s}, \; \text{ with } \; u_L^0 = u_L(0), \\
\partial_s F^{(0)} &= s_b(s)\frac{bu_1}{1-bu_1}F^{(0)}.
\end{aligned} \tag{B.2.9}$$

Now, for the RM, we consider the system of chemical reactions given in Eq. (4.3). We define $P(\vec{n};t)$ to be the corresponding new probability. Then, the CME for our system is given by

$$\partial_t P = \sum_{m=0}^{\infty} P(m)s_b(t)(\mathbb{E}_{n_1}^{-m} - 1)P + \sum_{j=1}^{L-1} k_j(\mathbb{E}_{n_j}\mathbb{E}_{n_{j+1}}^{-1} - 1)n_j P + k_L(\mathbb{E}_{n_L} - 1)n_L P \tag{B.2.10}$$

where $P(m) = b^m/(1+b)^{m+1}$ $(m = 0, 1, \dots)$ is a geometric distribution with mean, b. Then, by using the method of generating functions, we obtain the following PDE;

$$\partial_t F + \sum_{j=1}^{L-1} k_j(u_j - u_{j+1})\partial_{u_j} F + k_L u_L \partial_{u_L} F = s_b(t)\frac{bu_1}{1-bu_1}F. \tag{B.2.11}$$

Since the solution for $F^{(0)}$ from Eq. (B.2.9) and the solution for F from Eq. (B.2.11) are identical, we conclude that the mRNA distributions obtained from the FM and the RM are the same under the timescale separation assumed here.

B.3 Closed-form expressions for the mean number of mRNA molecules

In this section, we present a detailed derivation of the solution to Eq. (4.9) in the cyclo-stationary limit. The governing system of differential equations can be written as

$$\begin{aligned}
\frac{d\langle n_1\rangle}{dt} &= \sum_{m=0}^{\infty} mP(m)s_b(t) - k_1\langle n_1\rangle = bs_b(t) - k_1\langle n_1\rangle, \\
\frac{d\langle n_j\rangle}{dt} &= k_{j-1}\langle n_{j-1}\rangle - k_j\langle n_j\rangle \quad \text{for } \; j = 2, \dots, L,
\end{aligned} \tag{B.3.1}$$

where we have used the definition $r(t) = s_b(t)$. We apply the Laplace transform to the above equations and obtain the system

$$\begin{aligned}
N_1 &= b\frac{1}{s+k_1}S_b(s), \\
N_j &= k_{j-1}\frac{1}{s+k_j}N_{j-1} \; \text{ for } \; j = 2, \dots, L,
\end{aligned} \tag{B.3.2}$$

where $\mathcal{L}(f(t)) = \int_0^\infty f(t)e^{-st}dt = F(s)$ is the Laplace transform and $\mathcal{L}(\langle n_j\rangle) = N_j$, as well as $\mathcal{L}(s_b(t)) = S_b(s)$. Here, we have used the initial conditions $\langle n_j\rangle|_{t=0} = 0$ for $j = 1, \dots, L$, which indicates zero mRNA molecules in the system initially. Now, we apply the inverse Laplace transform, $\mathcal{L}^{-1}(F(s)) = f(t)$, to Eq. (B.3.2) to obtain the following system:

$$\begin{aligned}
\langle n_1\rangle &= b\mathcal{L}^{-1}\Big(\frac{1}{s+k_1}\Big) * \mathcal{L}^{-1}(S_b(s)) = be^{-k_1 t} * s_b(t), \\
\langle n_j\rangle &= k_{j-1}\mathcal{L}^{-1}\Big(\frac{1}{s+k_j}\Big) * \mathcal{L}^{-1}(N_{j-1}(s)) = k_{j-1}e^{-k_j t} * \langle n_{j-1}\rangle \quad \text{for } \; j = 2, \dots, L,
\end{aligned} \tag{B.3.3}$$

where $*$ denotes the convolution operator. Then, the system in Eq. (B.3.3) can be written as

$$\begin{aligned}\langle n_1\rangle &= b\int_0^t e^{-k_1 x} s_b(t-x)dx,\\ \langle n_j\rangle &= k_{j-1}\int_0^t e^{-k_j x}\langle n_{j-1}(t-x)\rangle dx \quad \text{for } j=2,\ldots,L.\end{aligned} \tag{B.3.4}$$

Evaluating the first integral in the cyclo-stationary limit of $t\to\infty$, we obtain

$$\langle n_1\rangle = bs_b k_1^{-1}\Re[1+\varepsilon z_1 e^{i\omega t}e^{i\varphi}], \quad \text{with } z_1=\frac{k_1}{k_1+i\omega}. \tag{B.3.5}$$

Note that we have expressed the time-dependent signal in the form $s_b(t)=\sigma\Re[1+\varepsilon e^{i\omega t}e^{i\varphi}]$ here, where $\Re[z]$ again denotes the real part of the complex number, z. Similarly, we can find the solution for each $\langle n_j\rangle$ from Eq. (B.3.4) in the limit of large times, where we also use the solution for $\langle n_1\rangle$ from Eq. (B.3.5). Then, one can easily show that for $j=2,\ldots,L$,

$$\langle n_j\rangle = bs_b k_j^{-1}\Re\left[1+\varepsilon\prod_{q=1}^{j} z_q e^{i\omega t}e^{i\varphi}\right] \quad \text{with} \quad z_q=\frac{k_q}{k_q+i\omega}. \tag{B.3.6}$$

We can simplify the expressions in Eq. (B.3.5) and Eq. (B.3.6) by expressing the complex numbers z_q in polar form as

$$z_q=\frac{k_q}{k_q+i\omega}=\frac{k_q}{\sqrt{k_q^2+\omega^2}}e^{i\theta_q}=|z_q|e^{i\theta_q}, \quad \text{with} \quad \theta_q=-\tan^{-1}\left(\frac{\omega}{k_q}\right). \tag{B.3.7}$$

Then, we use the identity

$$\prod_{q=1}^{j} z_q=\prod_{q=1}^{j}|z_q|\exp\left[i\sum_{q=1}^{j}\theta_q\right]=K_j e^{i\Theta_j}. \tag{B.3.8}$$

Hence, Eq. (B.3.5) and Eq. (B.3.6) simplify to

$$\langle n_j\rangle = bs_b k_j^{-1}(1+\varepsilon K_j\cos(\omega t+\varphi+\Theta_j)) \quad \text{for } j=1,\ldots,L. \tag{B.3.9}$$

Since Eq. (B.3.9) represents a wave for each stage j, we can also rewrite the above as

$$\langle n_j\rangle = \langle \bar{n_j}\rangle + A_j^m\cos(\omega t+\varphi+\Theta_j) \quad \text{for } j=1,\ldots,L, \tag{B.3.10}$$

where $\langle \bar{n_j}\rangle = \int_0^\tau \langle n_j\rangle dt/\tau = bs_b k_j^{-1}$ is the time-averaged mean over one period of time, $A_j^m = bs_b k_j^{-1}\varepsilon K_j$ is the amplitude, and $\varphi+\Theta_j$ is the phase of the time-dependent oscillatory part.

Here, we note that one can easily show that the amplitude A_j^m is a decreasing function of ω. Also, for $\omega \gg k_q$ $(q=1,\ldots,j)$, it follows that $A_j^m=0$, i.e. that the mean is constant and equal to the time-averaged mean. Additionally, we have that

$$\frac{A_j^m}{A_{j+1}^m}=\sqrt{\left(\frac{k_{j+1}}{k_j}\right)^2+\left(\frac{\omega}{k_j}\right)^2}, \tag{B.3.11}$$

which indicates that the amplitude can increase or decrease with the life-cycle stage j, depending on the values of the parameters k_{j+1}, k_j, and ω.

B.4 Exact solution of the Lyapunov equation for the RM

The Lyapunov equation mentioned in Eq. (4.12) can be solved explicitly for the covariance matrix **C** in the cyclo-stationary limit. The non-zero elements of **J** are given by

$$J_{ii} = -k_i \quad \text{for } i = 1, \ldots, L \quad \text{and} \quad J_{i,i-1} = k_{i-1} \quad \text{for } i = 2, \ldots, L, \tag{B.4.1}$$

while the non-zero elements D_{ij} of **D** read

$$\begin{aligned} D_{11} &= \sum_{m=0}^{\infty} m^2 f_1 + f_2, \\ D_{i,i+1} &= -f_{i+1} && \text{for } i = 1, \ldots, L-1, \\ D_{ii} &= f_i + f_{i+1} && \text{for } i = 2, \ldots, L-1, \quad \text{and} \\ D_{i,i-1} &= -f_i && \text{for } i = 2, \ldots, L, \end{aligned} \tag{B.4.2}$$

where f_i is defined in Eq. (4.5) and given in Eq. (B.3.9) when evaluated at $\langle n_i \rangle$. Also, we have that

$$\sum_{m=0}^{\infty} m^2 f_1 = s_b(t) \sum_{m=0}^{\infty} m^2 P(m) = s_b(t)(Var(m) + \langle m \rangle^2) = s_b(t)(b + 2b^2), \tag{B.4.3}$$

because $P(m)$ is a geometric distribution with mean burst size b. With these definitions, we can express the Lyapunov equation as a system of L^2 differential equations:

$$\frac{dC_{ij}}{dt} = \sum_{q=1}^{L} (J_{iq} C_{qj} + J_{jq} C_{iq}) + D_{ij} \quad \text{for } i, j = 1, \ldots, L. \tag{B.4.4}$$

Considering only the non-zero elements of **J**, Eq. (B.4.4) simplifies to

$$\frac{dC_{ij}}{dt} = C_{ij}(J_{ii} + J_{jj}) + C_{i-1,j} J_{i,i-1} + C_{i,j-1} J_{j,j-1} + D_{ij}. \tag{B.4.5}$$

A further simplification is achieved by considering only the non-zero elements of **D**:

$$\begin{aligned} \dot{C}_{11} &= C_{11} 2J_{11} + D_{11}, \\ \dot{C}_{12} &= C_{12}(J_{11} + J_{22}) + C_{11} J_{21} + D_{12}, \\ \dot{C}_{1j} &= C_{1j}(J_{11} + J_{jj}) + C_{1,j-1} J_{j,j-1} && \text{for } j = 3, \ldots, L, \\ \dot{C}_{ii} &= C_{ii} 2J_{ii} + C_{i-1,i} 2J_{i,i-1} + D_{ii} && \text{for } i = 2, \ldots, L, \\ \dot{C}_{i,i+1} &= C_{i,i+1}(J_{ii} + J_{i+1,i+1}) + C_{i-1,i+1} J_{i,i-1} + C_{ii} J_{i+1,i} + D_{i,i+1} && \text{for } i = 2, \ldots, L-1, \quad \text{and} \\ \dot{C}_{ij} &= C_{ij}(J_{ii} + Jjj) + C_{i-1,j} J_{i,i-1} + C_{i,j-1} J_{j,j-1} && \text{for } i, j = 2, \ldots, L \text{ and } j \geq i+2, \end{aligned} \tag{B.4.6}$$

where the overdot denotes differentiation with respect to time t. Substituting the definitions of J_{ij} and D_{ij} into Eq. (B.4.6), we obtain

$$\begin{aligned} \dot{C}_{11} &= -C_{11} 2k_1 + s_b(t)(b + 2b^2) + k_1 \langle n_1 \rangle, \\ \dot{C}_{12} &= -C_{12}(k_1 + k_2) + C_{11} k_1 - k_1 \langle n_1 \rangle, \\ \dot{C}_{1j} &= -C_{1j}(k_1 + k_j) + C_{1,j-1} k_{j-1} && \text{for } j = 3, \ldots, L, \\ \dot{C}_{ii} &= -C_{ii} 2k_i + C_{i-1,i} 2k_{i-1} + k_{i-1} \langle n_{i-1} \rangle + k_i \langle n_i \rangle && \text{for } i = 2, \ldots, L, \\ \dot{C}_{i,i+1} &= -C_{i,i+1}(k_i + k_{i+1}) + C_{i-1,i+1} k_{i-1} + C_{ii} k_i - k_i \langle n_i \rangle && \text{for } i = 2, \ldots, L-1, \quad \text{and} \\ \dot{C}_{ij} &= -C_{ij}(k_i + k_j) + C_{i-1,j} k_{i-1} + C_{i,j-1} k_{j-1} && \text{for } i, j = 2, \ldots, L \text{ and } j \geq i+2. \end{aligned} \tag{B.4.7}$$

Now, we apply the Laplace transform to Eq. (B.4.7) with $\mathcal{L}(\langle n_j \rangle) = N_j$, $\mathcal{L}(s_b(t)) = S_b(s)$, and $\mathcal{L}(C_{ij}) = c_{ij}$, which gives us the following system of equations:

$$\begin{aligned}
(s+2k_1)c_{11} &= k_1 N_1 + S_b(s)(b+2b^2), \\
(s+k_1+k_2)c_{12} &= c_{11}k_1 - k_1 N_1, \\
(s+k_1+k_j)c_{1j} &= c_{1,j-1}k_{j-1} && \text{for } j = 3,\dots,L, \\
(s+2k_i)c_{ii} &= c_{i-1,i}2k_{i-1} + k_{i-1}N_{i-1} + k_i N_i && \text{for } i = 2,\dots,L, \\
(s+k_i+k_{i+1})c_{i,i+1} &= c_{i-1,i+1}k_{i-1} + c_{ii}k_i - k_i N_i && \text{for } i = 2,\dots,L-1, \quad \text{and} \\
(s+k_i+k_j)c_{ij} &= c_{i-1,j}k_{i-1} + c_{i,j-1}k_{j-1} && \text{for } i,j = 2,\dots,L \text{ and } j \geq i+2.
\end{aligned} \tag{B.4.8}$$

Here, we have used the initial conditions $\langle n_j \rangle|_{t=0} = 0$ and $C_{ij}|_{t=0} = 0$ for $i,j = 1,\dots,L$. Now, in Laplace space, we can use the expressions from Eq. (B.3.2), which gives us

$$\begin{aligned}
bS_b(s) &= (s+k_1)N_1 \quad \text{and} \\
k_{j-1}N_{j-1} &= (s+k_j)N_j \quad \text{for } j = 2,\dots,L.
\end{aligned} \tag{B.4.9}$$

We substitute the above expressions into Eq. (B.4.8) and hence obtain the following simplified system:

$$\begin{aligned}
c_{11} &= N_1 + \frac{1}{s+2k_1}S_b(s)2b^2, \\
c_{12} &= (c_{11} - N_1)\frac{k_1}{s+k_1+k_2}, \\
c_{1j} &= c_{1,j-1}\frac{k_{j-1}}{s+k_1+k_j} && \text{for } j = 3,\dots,L, \\
c_{ii} &= c_{i-1,i}\frac{2k_{i-1}}{s+2k_i} + N_i && \text{for } i = 2,\dots,L, \\
c_{i,i+1} &= c_{i-1,i+1}\frac{k_{i-1}}{s+k_i+k_{i+1}} + (c_{ii} - N_i)\frac{k_i}{s+k_i+k_{i+1}} && \text{for } i = 2,\dots,L-1, \quad \text{and} \\
c_{ij} &= c_{i-1,j}\frac{k_{i-1}}{s+k_i+k_j} + c_{i,j-1}\frac{k_{j-1}}{s+k_i+k_j} && \text{for } i,j = 2,\dots,L \text{ and } j \geq i+2.
\end{aligned} \tag{B.4.10}$$

Now, we define new functions f_{ij} such that

$$c_{ij} = \delta_{ij}N_i + f_{ij} \quad \text{for } i,j = 1,\dots,L; \tag{B.4.11}$$

then, the system in Eq. (B.4.10) transforms to

$$\begin{aligned}
c_{11} &= N_1 + \frac{1}{s+2k_1}S_b(s)2b^2 = N_1 + f_{11}, \\
c_{12} &= f_{11}\frac{k_1}{s+k_1+k_2} = f_{12}, \\
c_{1j} &= f_{1,j-1}\frac{k_{j-1}}{s+k_1+k_j} = f_{1j} && \text{for } j = 3,\dots,L, \\
c_{ii} &= f_{i-1,i}\frac{2k_{i-1}}{s+2k_i} + N_i = N_i + f_{ii} && \text{for } i = 2,\dots,L, \\
c_{i,i+1} &= f_{i-1,i+1}\frac{k_{i-1}}{s+k_i+k_{i+1}} + f_{ii}\frac{k_i}{s+k_i+k_{i+1}} = f_{i,i+1} && \text{for } i = 2,\dots,L-1, \quad \text{and} \\
c_{ij} &= f_{i-1,j}\frac{k_{i-1}}{s+k_i+k_j} + f_{i,j-1}\frac{k_{j-1}}{s+k_i+k_j} = f_{ij} && \text{for } i,j = 2,\dots,L \text{ and } j \geq i+2.
\end{aligned} \tag{B.4.12}$$

From the above transformation, it is clear that the function f_{ij} is naturally defined as

$$\begin{aligned} f_{11} &= \frac{1}{s+2k_1} S_b(s) 2b^2, \\ f_{1j} &= f_{1,j-1} \frac{k_{j-1}}{s+k_1+k_j} \qquad \text{for } \; i,j = 2,\ldots,L, \quad \text{and} \\ f_{ij} &= f_{i-1,j} \frac{k_{i-1}}{s+k_i+k_j} + f_{i,j-1} \frac{k_{j-1}}{s+k_i+k_j} \qquad \text{for } \; i,j = 2,\ldots,L. \end{aligned} \tag{B.4.13}$$

Next, we apply the inverse Laplace transform to Eq. (B.4.11) and obtain the expression

$$C_{ij} = \delta_{ij}\langle n_i \rangle + \mathcal{L}^{-1}(f_{ij}) \quad \text{for } \; i,j = 1,\ldots,L. \tag{B.4.14}$$

For the full solution of the covariance matrix $\mathbf{C}$, we need to find the expressions for $F_{ij} = \mathcal{L}^{-1}(f_{ij})$. By going through the same steps as in Section B.3 – use of the convolution property and introduction of a time-dependent signal in the form $s_b(t) = \sigma\Re[1 + \varepsilon e^{i\omega t} e^{i\varphi}]$, followed by evaluation of the integral over $[0,\infty)$ to enforce the cyclo-stationary limit – we find that

$$F_{1j} = 2b^2\sigma\Re[g_{1j}^0 + \varepsilon g_{1j} e^{i\omega t} e^{i\varphi}] \quad \text{for } \; j = 1,\ldots,L. \tag{B.4.15}$$

Here, g_{1j} is a complex function that is defined as

$$g_{1j} = \prod_{q=1}^{j} \frac{k_{q-1}}{k_1 + k_q + i\omega} \qquad \text{with } \; k_0 = 1, \tag{B.4.16}$$

and $g_{1j}^0 = g_{1j}|_{\{\omega=0\}}$. (The superscript 0 indicates that the expression is independent of the parameter ω and that it is a real function.) Now, we assume that

$$F_{ij} = 2b^2\sigma\Re[g_{ij}^0 + \varepsilon g_{ij} e^{i\omega t} e^{i\varphi}] \quad \text{for } \; i,j = 2,\ldots,L. \tag{B.4.17}$$

By combining the third equation in Eq. (B.4.13) and Eq. (B.4.17), one can easily find the recurrence relation

$$g_{ij} = g_{i-1,j} \frac{k_{i-1}}{k_i + k_j + i\omega} + g_{i,j-1} \frac{k_{j-1}}{k_i + k_j + i\omega} \quad \text{for } \; i,j = 2,\ldots,L, \tag{B.4.18}$$

with initial conditions g_{1j} as defined in Eq. (B.4.16). Summarising our results, we have that the elements of the covariance matrix $\mathbf{C}$ for $i,j = 1,\ldots,L$ have the following closed-form expressions:

$$\begin{aligned} C_{ij} &= \delta_{ij}\langle n_i \rangle + 2b^2\sigma\Re[g_{ij}^0 + \varepsilon g_{ij} e^{i\omega t} e^{i\varphi}] \\ &= \delta_{ij}\langle n_i \rangle + b^2\sigma k_j^{-1}\Re[2k_j g_{ij}^0 + \varepsilon 2k_j g_{ij} e^{i\omega t} e^{i\varphi}], \end{aligned} \tag{B.4.19}$$

where the complex function g_{ij} is defined by the recurrence relation in Eq. (B.4.18) and $g_{ij}^0 = g_{ij}|_{\{\omega=0\}}$ is a real function. The solution of this recurrence relation is given by

$$\begin{aligned} g_{ij} &= \sum_{q_i=i}^{q_{i+1}=j} \sum_{q_{i-1}=i-1}^{q_i} \cdots \sum_{q_3=3}^{q_4} \sum_{q_2=2}^{q_3} a(i,q_{i+1},q_i)a(i-1,q_i,q_{i-1})\ldots a(3,q_4,q_3)a(2,q_3,q_2)g_{1,q_2} \\ &= \sum_{q_i=i}^{q_{i+1}=j} \sum_{q_{i-1}=i-1}^{q_i} \cdots \sum_{q_3=3}^{q_4} \sum_{q_2=2}^{q_3} \Big[\prod_{m=2}^{i} a(m,q_{m+1},q_m) \Big] g_{1,q_2}, \end{aligned} \tag{B.4.20}$$

where $m = 2, 3, \ldots, i$, $q_{m+1} \geq m$, $m \leq q_m \leq q_{m+1}$, and

$$g_{1p} = \prod_{q=1}^{p} \frac{k_{q-1}}{k_1 + k_q + i\omega} \qquad \text{with } k_0 = 1 \text{ for } p = 1, \ldots, L,$$

$$a(m, m, m) = 2\frac{k_{m-1}}{2k_m + i\omega} \qquad \text{if } q_{m+1} = q_m = m \text{ and } m = 2, 3, \ldots, L,$$

$$a(m, q_m, q_m) = \frac{k_{m-1}}{k_m + k_{q_m} + i\omega} \qquad \text{if } q_{m+1} = q_m \text{ and } q_m \geq m + 1, \quad \text{and}$$

$$a(m, q_{m+1}, q_m) = a(m, q_m, q_m) \prod_{p=q_m}^{q_{m+1}-1} \frac{k_p}{k_m + k_{p+1} + i\omega} \qquad \text{if } q_{m+1} \geq q_m + 1 \text{ and } q_m \geq m.$$

For $i = j$, we can obtain expressions for the variance from Eq. (B.4.19), which are given by

$$C_{jj} = Var(n_j) = \langle n_j \rangle + b^2 \sigma k_j^{-1} (G_j^0 + \varepsilon G_j \cos(\omega t + \varphi + \Phi_j)) \quad \text{for } j = 1, \ldots, L, \tag{B.4.21}$$

where $G_j e^{i\Phi_j} = 2k_j g_{jj}$ and $G_j^0 = 2k_j g_{jj}|_{\{\omega=0\}}$.

B.5 The modified stochastic simulation algorithm

For the stochastic simulations of our model as presented in the paper, we used the modified Gillespie algorithm for time-dependent propensity functions [216]. The main steps in the algorithm for generating one trajectory are as follows:

1. Initialize the time $t = 0$ and the number of molecules of each species, $\vec{x}(t = 0) = (n_1, \ldots, n_{N_S})$, where N_S is the total number of species, and perform the following steps as long as $t \leq t_{\max}$.

2. Generate two independent uniform random numbers on the interval $(0, 1)$: r_1, which defines the time that passes until the next reaction occurs; and r_2, which defines the next reaction that occurs.

3. The time that passes until the next reaction occurs, Δ, is exponentially distributed with parameter

$$f_0 = \sum_{i=1}^{N_R} \int_t^{t+\Delta} \tilde{f}_i(\vec{x}(t), s) ds, \tag{B.5.1}$$

where the propensities $\tilde{f}_i(\vec{x}(t), t)$ for $i = 1, 2, \ldots, N_R$ depend explicitly on time and N_R is the total number of reactions. Note that $\vec{x}(t)$ is constant in the above integral, as no reactions take place within the time interval $[t, t + \Delta)$. In our case, the above integral can be solved analytically; we write the propensity functions of the RM as

$$\tilde{f}_1 = \tilde{s}_b(t) = \sigma\Big(1 + \sum_{n=1}^{N_S} A_n \cos(\omega n t + \varphi_n)\Big) \quad \text{and} \quad \tilde{f}_i = k_{i-1} n_{i-1} \quad \text{for } i = 2, \ldots, N_R, \tag{B.5.2}$$

where we have $N_R = L + 1$ for our model, with n_{i-1} the entries of the array $\vec{x}(t)$. Here, we have considered the general case for our time-dependent signal, which can be written in truncated Fourier form; see Section 4.4, with $\tilde{s}_b(t)$ as in Eq. (4.30). Note that in our case, only the propensity function $\tilde{f}_1$ is explicitly dependent on time. The solution of the integral

in Eq. (B.5.1) is then given by

$$\begin{aligned} f_0 &= \sum_{i=2}^{N_R} \tilde{f}_i \int_t^{t+\Delta} ds + \int_t^{t+\Delta} \tilde{f}_1 ds \\ &= \sum_{i=2}^{N_R} \tilde{f}_i \Delta + \sigma \Big[\Delta + \sum_{n=1}^{N_S} A_n \frac{1}{n\omega} \Big(\sin(\omega n(t+\Delta) + \varphi_n) - \sin(\omega n t + \varphi_n) \Big) \Big]. \end{aligned} \tag{B.5.3}$$

In order to obtain the time interval Δ, we solve the algebraic equation

$$f_0 = \ln(1/r_1), \tag{B.5.4}$$

using the bisection method [217].

4. Identify which reaction is going to occur next by picking the reaction index j to satisfy the inequality

$$\sum_{i=1}^{j-1} \frac{\tilde{f}_i(\vec{x}(t), t+\Delta)}{f_\Delta} < r_2 \leq \sum_{i=1}^{j} \frac{\tilde{f}_i(\vec{x}(t), t+\Delta)}{f_\Delta}, \quad \text{where} \quad f_\Delta = \sum_{i=1}^{N_R} \tilde{f}_i(\vec{x}(t), t+\Delta). \tag{B.5.5}$$

5. According to which reaction has occurred, update the species vector, $\vec{x}(t)$.

6. Update the time by replacing t with $t + \Delta$.

7. If $t \leq t_{\max}$, then go back to step 2; otherwise, end.

B.6 Derivation of the exact mRNA distribution for the EM

In this section, we consider the simple reaction scheme given in Eq. (4.7) for the EM, and we derive its exact time-dependent mRNA distribution. Throughout the following derivations, we consider j as some user-input (fixed) parameter which depends on the life-cycle stage of interest. We begin with the PDE for the probability generating function given in Eq. (4.20), which we convert to the following system of ODEs by using the method of characteristics:

$$\begin{aligned} &\partial_s t = 1, \quad \text{which implies } \; t = s, \\ &\partial_s u = k_j u, \quad \text{which implies } \; u = u^0 e^{k_j s}, \; \text{ with } \; u^0 = u(0), \\ &\partial_s F = \bar{\sigma}_j (1 + \bar{\varepsilon}_j \cos(\omega s + \bar{\varphi}_j)) \frac{\bar{b}_j u}{1 - \bar{b}_j u} F. \end{aligned} \tag{B.6.1}$$

The last ODE in the above system can be rewritten as

$$\frac{\partial_s F}{F} = \bar{\sigma}_j \frac{\bar{b}_j u^0 e^{k_j s}}{1 - \bar{b}_j u^0 e^{k_j s}} + \bar{\sigma}_j \bar{\varepsilon}_j \cos(\omega s + \bar{\varphi}_j) \Big(\frac{1}{1 - \bar{b}_j u^0 e^{k_j s}} - 1 \Big), \tag{B.6.2}$$

which admits the solution

$$\ln(F) = \bar{\sigma}_j (h_0 - h_1 + h_2) + C_0. \tag{B.6.3}$$

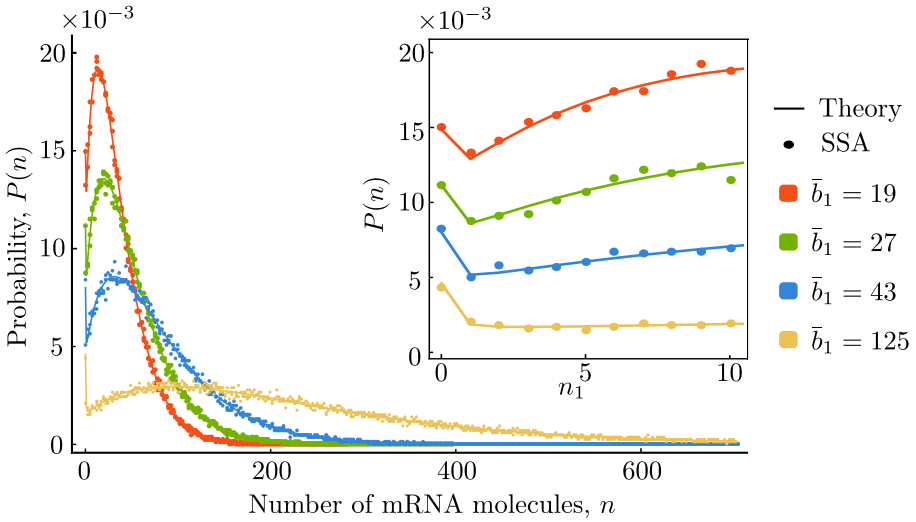


Figure B.6.1: *The mRNA distribution from the EM and the RM can display bimodality.* We show the mRNA distribution for species M_1 from the EM, as found from stochastic simulation (SSA; points) and predicted exactly by our theory (Eq. (4.23) together with Eq. (4.21); solid lines). Note that the above is identical to what is found in the RM, since the two models agree exactly in their prediction of the distribution for M_1. We illustrate results for four different values of the mean burst size, $\bar{b}_1 \in \{19,\ 27,\ 43,\ 130\}$. The parameter values that have been used are $\xi = 0$ (cyclo-stationary limit), $\bar{\varepsilon}_1 = 0.99$, $\bar{\sigma}_1 = 5\ \text{min}^{-1}$, $\omega = 1.82\ \text{min}^{-1}$, $\kappa_1 = 4.2\ \text{min}^{-1}$, $\bar{\varphi}_1 = \pi/2$, and $t = 9.2$ min. The inset shows a zoomed-in version of the distribution for small numbers of mRNA molecules.

Here, C_0 is some constant of integration to be determined later, with

$$h_0 = \int \frac{\bar{b}_j u^0 e^{k_j s}}{1 - \bar{b}_j u^0 e^{k_j s}} ds = \ln(1 - \bar{b}_j u^0 e^{k_j s})^{-\frac{1}{k_j}},$$

$$h_1 = \int \bar{\varepsilon}_j \cos(\omega s + \bar{\varphi}_j) ds = \bar{\varepsilon}_j \omega^{-1} \sin(\omega s + \bar{\varphi}_j), \quad \text{and}$$

$$h_2 = \int \bar{\varepsilon}_j \cos(\omega s + \bar{\varphi}_j) \frac{1}{1 - \bar{b}_j u^0 e^{k_j s}} ds = \bar{\varepsilon} \omega^{-1} \Im[e^{i(\omega s + \bar{\varphi}_j)} {}_2F_1(1, i\omega k_j^{-1}, 1 + i\omega k_j^{-1}, \bar{b} u^0 e^{k_j s})], \tag{B.6.4}$$

which are well defined when $\omega \neq 0$ and $k_j \neq 0$. Here, ${}_2F_1$ is the hypergeometric function of the second kind [44,63] and $\Im[z]$ denotes the imaginary part of a complex number, z. Note that in the expression for h_2, we have used the identity $\cos(\omega s + \bar{\varphi}_j) = (e^{i(\omega s + \bar{\varphi}_j)} + e^{-i(\omega s + \bar{\varphi}_j)})/2$, as well as the following identity for hypergeometric functions:

$$\frac{1}{1 - z} = \frac{z}{a} \frac{d}{dz} {}_2F_1(1, a; a + 1; z) + {}_2F_1(1, a; a + 1; z), \tag{B.6.5}$$

which one can easily derive from Equation 15.5.20 in [44]. Next, we need to determine the constant of integration, C_0. The generating function F has to satisfy the initial condition $F(s = 0) = 1$ or, equivalently, $F(t = 0) = 1$, which stems from the initial condition, $P(n = 0; t = 0)$, as there are zero mRNA molecules in the system at time zero. Applying this initial condition to Eq. (B.6.3), we have that

$$C_0 = -\bar{\sigma}_j [h_0|_{\{s=0\}} - h_1|_{\{s=0\}} + h_2|_{\{s=0\}}]. \tag{B.6.6}$$

Substituting Eq. (B.6.6) into Eq. (B.6.3) and simplifying, we obtain the following solution:

$$F(u^0; s) = \exp[\bar{\sigma}_j (h_0 - h_0|_{\{s=0\}} - (h_1 - h_1|_{\{s=0\}}) + h_2 - h_2|_{\{s=0\}})]. \tag{B.6.7}$$

Next, we apply the inverse transformation $s = t$ and $u^0 = ue^{-k_j t} = u/\xi_j$ to obtain our final expression for the generating function,

$$F(u;t) = \Big(\frac{1-\bar{b}_j u\xi_j}{1-\bar{b}_j u}\Big)^{\frac{\bar{\sigma}_j}{k_j}} \exp[\bar{\sigma}_j\bar{\varepsilon}_j\omega^{-1}(f_1(t)+f_2(u,t))], \tag{B.6.8}$$

where

$$f_1(t) = \sin(\bar{\varphi}_j) - \sin(\omega t + \bar{\varphi}_j) \quad \text{and}$$
$$f_2(u,t) = \Im[e^{i(\omega t+\bar{\varphi}_j)}{}_2F_1(1, i\omega k_j^{-1}, 1+i\omega k_j^{-1}, \bar{b}_j u)] - \Im[e^{i\bar{\varphi}_j}{}_2F_1(1, i\omega k_j^{-1}, 1+i\omega k_j^{-1}, \bar{b}_j u\xi_j)].$$

The mRNA distribution is then found from the formula $P(n;t) = \frac{1}{n!}\frac{d^n}{du^n}F(u;t)\big|_{\{u=-1\}}$.

Bibliography

[1] Gordon L Hager, James G McNally, and Tom Misteli. Transcription dynamics. Molecular Cell, 35(6):741–753, 2009.

[2] Robert G Roeder. The role of general initiation factors in transcription by RNA polymerase ii. Trends in Biochemical Sciences, 21(9):327–335, 1996.

[3] DB Nikolov and SK Burley. RNA polymerase ii transcription initiation: a structural view. Proceedings of the National Academy of Sciences, 94(1):15–22, 1997.

[4] Edward A Felinski, Jeonga Kim, Jingfang Lu, and Patrick G Quinn. Recruitment of an RNA polymerase ii complex is mediated by the constitutive activation domain in creb, independently of creb phosphorylation. Molecular and Cellular Biology, 21(4):1001–1010, 2001.

[5] Daniel R Larson, Daniel Zenklusen, Bin Wu, Jeffrey A Chao, and Robert H Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science, 332(6028):475–478, 2011.

[6] Thomas D Pollard, William C Earnshaw, Jennifer Lippincott-Schwartz, and Graham Johnson. Cell biology E-book. Elsevier Health Sciences, 2016.

[7] Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of pol ii elongation and its interplay with promoter proximal pausing, chromatin, and exons. elife, 3:e02407, 2014.

[8] Telmo Henriques, Daniel A Gilchrist, Sergei Nechaev, Michael Bern, Ginger W Muse, Adam Burkholder, David C Fargo, and Karen Adelman. Stable pausing by rna polymerase ii provides an opportunity to target and integrate regulatory signals. Molecular cell, 52(4):517–528, 2013.

[9] Qiang Zhou, Tiandao Li, and David H Price. RNA polymerase ii elongation control. Annual Review of Biochemistry, 81:119–143, 2012.

[10] Eun-Jung Cho, Toshimitsu Takagi, Christine R Moore, and Stephen Buratowski. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase ii carboxy-terminal domain. Genes & Development, 11(24):3319–3326, 1997.

[11] Diana F Colgan and James L Manley. Mechanism and regulation of mRNA polyadenylation. Genes & Development, 11(21):2755–2766, 1997.

[12] Nuno André Faustino and Thomas A Cooper. Pre-mRNA splicing and human disease. Genes & Development, 17(4):419–437, 2003.

[13] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. Cell, 163(7):1596–1610, 2015.

[14] Thomas Stoeger, Nico Battich, and Lucas Pelkmans. Passive noise filtering by cellular compartmentalization. Cell, 164(6):1151–1161, 2016.

[15] Carol J Wilusz, Michael Wormington, and Stuart W Peltz. The cap-to-tail guide to mRNA turnover. Nature Reviews Molecular Cell Biology, 2(4):237, 2001.

[16] Carolyn J Decker and Roy Parker. A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. Genes & Development, 7(8):1632–1643, 1993.

[17] Allan Jacobson and Stuart W Peltz. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Annual Review of Biochemistry, 65(1):693–739, 1996.

[18] Juan M Pedraza and Johan Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. Science, 319(5861):339–343, 2008.

[19] DAN CAO and ROY PARKER. Computational modeling of eukaryotic mRNA turnover. RNA, 7(9):1192–1212, 2001.

[20] Clare A Beelman and Roy Parker. Degradation of mRNA in eukaryotes. Cell, 81(2):179–183, 1995.

[21] Cécile Bousquet-Antonelli, Carlo Presutti, and David Tollervey. Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell, 102(6):765–775, 2000.

[22] Jonathan A Bernstein, Arkady B Khodursky, Pei-Hsun Lin, Sue Lin-Chao, and Stanley N Cohen. Global analysis of mRNA decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays. Proceedings of the National Academy of Sciences, 99(15):9697–9702, 2002.

[23] Yulei Wang, Chih Long Liu, John D Storey, Robert J Tibshirani, Daniel Herschlag, and Patrick O Brown. Precision and functional specificity in mRNA decay. Proceedings of the National Academy of Sciences, 99(9):5860–5865, 2002.

[24] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X Sunney Xie. Probing gene expression in live cells, one protein molecule at a time. Science, 311(5767):1600–1603, 2006.

[25] T Martin Schmeing and Venki Ramakrishnan. What recent ribosome structures have revealed about the mechanism of translation. Nature, 461(7268):1234–1242, 2009.

[26] Daniel R Larson, Robert H Singer, and Daniel Zenklusen. A single molecule view of gene expression. Trends in cell biology, 19(11):630–637, 2009.

[27] Hans-Hermann Gerdes and Christoph Kaether. Green fluorescent protein: applications in cell biology. FEBS letters, 389(1):44–47, 1996.

[28] Mark R Soboleski, Jason Oaks, and William P Halford. Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells. The FASEB journal, 19(3):1–20, 2005.

[29] Alexander P Young, Daniel J Jackson, and Russell C Wyeth. A technical review and guide to rna fluorescence in situ hybridization. PeerJ, 8:e8806, 2020.

[30] Andrea M Femino, Fredric S Fay, Kevin Fogarty, and Robert H Singer. Visualization of single RNA transcripts in situ. Science, 280(5363):585–590, 1998.

[31] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. PLoS Biology, 4(10):e309, 2006.

[32] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. eLife, 5:e12175, 2016.

[33] Manuele Castelnuovo, Samir Rahman, Elisa Guffanti, Valentina Infantino, Françoise Stutz, and Daniel Zenklusen. Bimodal expression of pho84 is modulated by early termination of antisense transcription. Nature Structural & Molecular Biology, 20(7):851, 2013.

[34] Saumil J Gandhi, Daniel Zenklusen, Timothée Lionnet, and Robert H Singer. Transcription of functionally related constitutive genes is not coordinated. Nature Structural & Molecular Biology, 18(1):27, 2011.

[35] Ciaran Condon, Sarah French, Craig Squires, and Catherine L Squires. Depletion of functional ribosomal RNA operons in escherichia coli causes increased expression of the remaining intact copies. The EMBO Journal, 12(11):4305–4315, 1993.

[36] Justina Voulgaris, Sarah French, Richard L Gourse, Craig Squires, and Catherine L Squires. Increased rrn gene dosage causes intermittent transcription of rRNA in escherichia coli. Journal of Bacteriology, 181(14):4170–4175, 1999.

[37] Aziz El Hage, Sarah L French, Ann L Beyer, and David Tollervey. Loss of topoisomerase i leads to r-loop-mediated transcriptional blocks during ribosomal RNA synthesis. Genes & Development, 24(14):1546–1558, 2010.

[38] SUSAN L Gotta, OL Miller, and SARAH L French. rRNA transcription rate in escherichia coli. Journal of Bacteriology, 173(20):6647–6649, 1991.

[39] SARAH L French and OL Miller. Transcription mapping of the escherichia coli chromosome by electron microscopy. Journal of Bacteriology, 171(8):4207–4216, 1989.

[40] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. Science, 297(5584):1183–1186, 2002.

[41] Berend Snijder and Lucas Pelkmans. Origins of regulated cell-to-cell variability. Nature Reviews Molecular Cell Biology, 12(2):119, 2011.

[42] Narendra Maheshri and Erin K O'Shea. Living with noisy genes: how cells function reliably with inherent variability in gene expression. Annu. Rev. Biophys. Biomol. Struct., 36:413–434, 2007.

[43] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences, 105(45):17256–17261, 2008.

[44] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.4 of 2022-01-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[45] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. Cell, 123(6):1025–1036, 2005.

[46] Roy D Dar, Brandon S Razooky, Abhyudai Singh, Thomas V Trimeloni, James M McCollum, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proceedings of the National Academy of Sciences, 109(43):17454–17459, 2012.

[47] Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. Nature, 565(7738):251, 2019.

[48] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-rna counting reveals alternative modes of gene expression in yeast. Nature structural & molecular biology, 15(12):1263–1271, 2008.

[49] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. RNA velocity of single cells. Nature, 560(7719):494, 2018.

[50] Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. Science, 342(6163):1188–1193, 2013.

[51] Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, et al. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. Nature Communications, 7:12248, 2016.

[52] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. Molecular Cell, 58(1):147–156, 2015.

[53] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. Science, 332(6028):472–474, 2011.

[54] Alvaro Sanchez, Sandeep Choubey, and Jane Kondev. Regulation of noise in gene expression. Annual Review of Biophysics, 42:469–491, 2013.

[55] Nacho Molina, David M Suter, Rosamaria Cannavo, Benjamin Zoller, Ivana Gotic, and Félix Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. Proceedings of the National Academy of Sciences, 110(51):20563–20568, 2013.

[56] Lok-hang So, Anandamohan Ghosh, Chenghang Zong, Leonardo A Sepúlveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in escherichia coli. Nature Genetics, 43(6):554, 2011.

[57] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics, 6(6):451, 2005.

[58] Minoru SH Ko. A stochastic model for gene induction. Journal of Theoretical Biology, 153(2):181–194, 1991.

[59] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. Science, 304(5678):1811–1814, 2004.

[60] William J Blake, Mads Kærn, Charles R Cantor, and James J Collins. Noise in eukaryotic gene expression. Nature, 422(6932):633, 2003.

[61] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. Theoretical Population Biology, 48(2):222–234, 1995.

[62] Srividya Iyer-Biswas, Fernand Hayot, and Ciriyam Jayaprakash. Stochasticity of gene products from transcriptional pulsing. Physical Review E, 79(3):031911, 2009.

[63] Milton Abramowitz, Irene A Stegun, Michael Danos, and Johann Rafelski. Pocketbook of mathematical functions: Abridged edition of handbook of mathematical functions. Verlag Harri Deutsch, 1984.

[64] Edward Tunnacliffe and Jonathan R Chubb. What is a transcriptional burst? Trends in Genetics, 36(4):288–297, 2020.

[65] Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun Raj. Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. Molecular Cell, 73(3):519–532, 2019.

[66] Jacques P Bothma, Hernan G Garcia, Emilia Esposito, Gavin Schlissel, Thomas Gregor, and Michael Levine. Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living drosophila embryos. Proceedings of the National Academy of Sciences, 111(29):10598–10603, 2014.

[67] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander Van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. Science, 339(6119):584–587, 2013.

[68] Iris Jonkers and John T Lis. Getting up to speed with transcription elongation by RNA polymerase ii. Nature Reviews Molecular Cell Biology, 16(3):167–177, 2015.

[69] Sharon Yunger, Liat Rosenfeld, Yuval Garini, and Yaron Shav-Tal. Single-allele analysis of transcription kinetics in living mammalian cells. Nature Methods, 7(8):631, 2010.

[70] Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. Deciphering transcriptional dynamics in vivo by counting nascent rna molecules. PLoS computational biology, 11(11), 2015.

[71] Sandeep Choubey. Nascent RNA kinetics: Transient and steady state behavior of models of transcription. Physical Review E, 97(2):022402, 2018.

[72] Heng Xu, Samuel O Skinner, Anna Marie Sokac, and Ido Golding. Stochastic kinetics of nascent RNA. Physical Review Letters, 117(12):128101, 2016.

[73] Abhyudai Singh and Pavol Bokes. Consequences of mRNA transport on stochastic variability in protein levels. Biophysical Journal, 103(5):1087–1096, 2012.

[74] Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear retention of mrna in mammalian tissues. Cell reports, 13(12):2653–2662, 2015.

[75] Li-ping Xiong, Yu-qiang Ma, and Lei-han Tang. Attenuation of transcriptional bursting in mRNA transport. Physical Biology, 7(1):016005, 2009.

[76] Marc Sturrock, Shiyu Li, and Vahid Shahrezaei. The influence of nuclear compartmentalisation on stochastic dynamics of self-repressing gene expression. Journal of Theoretical Biology, 424:55–72, 2017.

[77] Amir Mor, Shimrit Suliman, Rakefet Ben-Yishay, Sharon Yunger, Yehuda Brody, and Yaron Shav-Tal. Dynamics of single mrnp nucleocytoplasmic transport and export through the nuclear pore in living cells. Nature Cell Biology, 12(6):543, 2010.

[78] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. Nature, 473(7347):337, 2011.

[79] Uri Alon. Network motifs: theory and experimental approaches. Nature Reviews Genetics, 8(6):450, 2007.

[80] Attila Becskei and Luis Serrano. Engineering stability in gene networks by autoregulation. Nature, 405(6786):590, 2000.

[81] Hana El-Samad and Mustafa Khammash. Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks. Biophysical Journal, 90(10):3749–3761, 2006.

[82] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-fos transcriptional bursts. Cell reports, 8(1):75–83, 2014.

[83] Nicolaas Godfried Van Kampen. Stochastic processes in physics and chemistry, volume 1. Elsevier, 1992.

[84] Donald A McQuarrie. Stochastic approach to chemical kinetics. Journal of Applied Probability, 4(3):413–478, 1967.

[85] Daniel T Gillespie. A rigorous derivation of the chemical master equation. Physica A: Statistical Mechanics and its Applications, 188(1-3):404–425, 1992.

[86] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. Journal of Physics A: Mathematical and Theoretical, 50(9):093001, 2017.

[87] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 22(4):403–434, 1976.

[88] Daniel T Gillespie. Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem., 58:35–55, 2007.

[89] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25):2340–2361, 1977.

[90] Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. Physical Review E, 92(4):042124, 2015.

[91] Parimal Mukhopadhyay. An introduction to the theory of probability. World Scientific, 2012.

[92] George F Carrier and Carl E Pearson. Partial differential equations: theory and technique. Academic Press, 2014.

[93] Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. Journal of Differential Equations, 31(1):53–98, 1979.

[94] Geertje Hek. Geometric singular perturbation theory in biological practice. Journal of Mathematical Biology, 60(3):347–386, 2010.

[95] Nikola Popović, Carsten Marr, and Peter S Swain. A geometric analysis of fast-slow models for stochastic gene expression. Journal of Mathematical Biology, 72(1-2):87–122, 2016.

[96] Pablo A Iglesias and Brian P Ingalls. Control theory and systems biology. MIT press, 2010.

[97] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. The Journal of chemical physics, 124(4):044104, 2006.

[98] Johan Elf and Måns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. Genome Research, 13(11):2475–2484, 2003.

[99] NG van Kampen. A power series expansion of the master equation. Canadian Journal of Physics, 39(4):551–567, 1961.

[100] Patrick B Warren, Sorin Tănase-Nicola, and Pieter Rein ten Wolde. Exact results for noise power spectra in linear biochemical reaction networks. The Journal of Chemical Physics, 125(14):144904, 2006.

[101] Ramon Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. The Journal of chemical physics, 133(3):07B604, 2010.

[102] Rajesh Ramaswamy, Nélido González-Segredo, Ivo F Sbalzarini, and Ramon Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. Nature communications, 3(1):1–8, 2012.

[103] C. W. Gardiner. Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. Springer, New York, 2004.

[104] Daniel T Gillespie. The chemical langevin equation. The Journal of Chemical Physics, 113(1):297–306, 2000.

[105] Ramon Grima, Philipp Thomas, and Arthur V Straube. How accurate are the nonlinear chemical fokker-planck and chemical langevin equations? The Journal of chemical physics, 135(8):084103, 2011.

[106] Neil Fenichel and JK Moser. Persistence and smoothness of invariant manifolds for flows. Indiana University Mathematics Journal, 21(3):193–226, 1971.

[107] Christopher KRT Jones. Geometric singular perturbation theory. In Dynamical Systems, pages 44–118. Springer, 1995.

[108] Pedro Toniol Cardin and Marco Antonio Teixeira. Fenichel theory for multiple time scale singular perturbation problems. SIAM Journal on Applied Dynamical Systems, 16(3):1425–1452, 2017.

[109] Frits Veerman, Carsten Marr, and Nikola Popović. Time-dependent propagators for stochastic models of gene expression: an analytical method. Journal of mathematical biology, 77(2):261–312, 2018.

[110] Pavol Bokes, John R King, Andrew TA Wood, and Matthew Loose. Multiscale stochastic modelling of gene expression. Journal of Mathematical Biology, 65(3):493–520, 2012.

[111] Martin Krupa and Peter Szmolyan. Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions. SIAM journal on mathematical analysis, 33(2):286–314, 2001.

[112] Niraj Kumar and Rahul V Kulkarni. Constraining the complexity of promoter dynamics using fluctuations in gene expression. Physical Biology, 17(1):015001, 2019.

[113] Tianshou Zhou and Jiajun Zhang. Analytical results for a multistate gene model. SIAM Journal on Applied Mathematics, 72(3):789–818, 2012.

[114] Joseph Rodriguez, Gang Ren, Christopher R Day, Keji Zhao, Carson C Chow, and Daniel R Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. Cell, 176(1-2):213–226, 2019.

[115] Jiajun Zhang and Tianshou Zhou. Stationary moments, distribution conjugation and phenotypic regions in stochastic gene transcription. Mathematical biosciences and engineering: MBE, 16(5):6134–6166, 2019.

[116] Adam M Corrigan, Edward Tunnacliffe, Danielle Cannon, and Jonathan R Chubb. A continuum model of transcriptional bursting. Elife, 5:e13051, 2016.

[117] Wanqing Shao and Julia Zeitlinger. Paused rna polymerase ii inhibits new transcriptional initiation. Nature genetics, 49(7):1045, 2017.

[118] Saskia Gressel, Björn Schwalb, Tim Michael Decker, Weihua Qin, Heinrich Leonhardt, Dirk Eick, and Patrick Cramer. Cdk9-dependent rna polymerase ii pausing controls transcription initiation. Elife, 6:e29736, 2017.

[119] Heng Xu, Samuel O Skinner, Anna Marie Sokac, and Ido Golding. Stochastic kinetics of nascent rna. Phys. Rev. Lett., 117(12):128101, 2016.

[120] Rob Phillips, Nathan M Belliveau, Griffin Chure, Hernan G Garcia, Manuel Razo-Mejia, and Clarissa Scholes. Figure 1 theory meets figure 2 experiments in the study of gene expression. Annual review of biophysics, 48:121–163, 2019.

[121] Jonathan R Chubb, Tatjana Trcek, Shailesh M Shenoy, and Robert H Singer. Transcriptional pulsing of a developmental gene. Current biology, 16(10):1018–1025, 2006.

[122] Zhixing Cao, Tatiana Filatova, Diego A Oyarzún, and Ramon Grima. A stochastic model of gene expression with polymerase recruitment and pause release. Biophysical Journal, 119(5):1002–1014, 2020.

[123] Tatiana Filatova, Nikola Popovic, and Ramon Grima. Statistics of nascent and mature rna fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination. Bulletin of Mathematical Biology, 83(1):1–62, 2021.

[124] Ibrahim I Cisse, Ignacio Izeddin, Sebastien Z Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of rna polymerase ii clustering in live human cells. Science, 341(6146):664–667, 2013.

[125] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods, 5(10):877–879, 2008.

[126] Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. Proceedings of the National Academy of Sciences, 117(9):4682–4692, 2020.

[127] Tineke L Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R Larson. Transcription dynamics in living cells. Annual Review of Biophysics, 45:25–47, 2016.

[128] Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. eLife, 3:e03939, 2014.

[129] Ineke Brouwer and Tineke L Lenstra. Visualizing transcription: key to understanding gene expression dynamics. Current Opinion in Chemical Biology, 51:122–129, 2019.

[130] Stefan Klumpp and Terence Hwa. Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. Proceedings of the National Academy of Sciences, 105(47):18159–18164, 2008.

[131] Tiina Rajala, Antti Häkkinen, Shannon Healy, Olli Yli-Harja, and Andre S Ribeiro. Effects of transcriptional pausing on gene expression dynamics. PLoS Computational Biology, 6(3), 2010.

[132] Md Zulfikar Ali, Sandeep Choubey, Dipjyoti Das, and Robert C Brewster. Probing mechanisms of transcription elongation through cell-to-cell variability of RNA polymerase. Biophysical Journal, 2020.

[133] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. Journal of Mathematical Biology, 54(1):1–26, 2007.

[134] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. Proceedings of the National Academy of Sciences, 109(21):8340–8345, 2012.

[135] Nancy R Forde, David Izhaky, Glenna R Woodcock, Gijs JL Wuite, and Carlos Bustamante. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by escherichia coli RNA polymerase. Proceedings of the National Academy of Sciences, 99(18):11682–11687, 2002.

[136] Karen Adelman, Arthur La Porta, Thomas J Santangelo, John T Lis, Jeffrey W Roberts, and Michelle D Wang. Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. Proceedings of the National Academy of Sciences, 99(21):13538–13543, 2002.

[137] Claudia Cianci, Stephen Smith, and Ramon Grima. Molecular finite-size effects in stochastic models of equilibrium chemical systems. The Journal of Chemical Physics, 144(8):084101, 2016.

[138] Stephen Smith, Claudia Cianci, and Ramon Grima. Macromolecular crowding directs the motion of small molecules inside cells. Journal of The Royal Society Interface, 14(131):20170047, 2017.

[139] Gennady Gorin, Mengyu Wang, Ido Golding, and Heng Xu. Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics. PLoS One, 15(3):e0230736, 2020.

[140] Kaan Öcal, Ramon Grima, and Guido Sanguinetti. Parameter estimation for biochemical reaction networks using Wasserstein distances. Journal of Physics A: Mathematical and Theoretical, 2019.

[141] Sarah Sainsbury, Carrie Bernecky, and Patrick Cramer. Structural basis of transcription initiation by rna polymerase ii. Nature reviews Molecular cell biology, 16(3):129–143, 2015.

[142] Juraj Szavits-Nossan and Ramon Grima. Predicting variability in nascent rna from transcription initiation kinetics. bioRxiv, 2022.

[143] Robert Landick. Transcriptional pausing without backtracking. Proceedings of the National Academy of Sciences, 106(22):8797–8798, 2009.

[144] Stefan Klumpp. Pausing and backtracking in transcription under dense traffic conditions. Journal of Statistical Physics, 142(6):1252–1267, 2011.

[145] Carolyn T MacDonald, Julian H Gibbs, and Allen C Pipkin. Kinetics of biopolymerization on nucleic acid templates. Biopolymers: Original Research on Biomolecules, 6(1):1–25, 1968.

[146] Carolyn T MacDonald and Julian H Gibbs. Concerning the kinetics of polypeptide synthesis on polyribosomes. Biopolymers: Original Research on Biomolecules, 7(5):707–725, 1969.

[147] RKP Zia, JJ Dong, and B Schmittmann. Modeling translation in protein synthesis with tasep: A tutorial and recent developments. Journal of Statistical Physics, 144(2):405–428, 2011.

[148] Simon Scott and Juraj Szavits-Nossan. Power series method for solving tasep-based models of mrna translation. Physical biology, 17(1):015004, 2019.

[149] Jingkui Wang, Benjamin Pfeuty, Quentin Thommen, M Carmen Romano, and Marc Lefranc. Minimal model of transcriptional elongation processes with pauses. Physical Review E, 90(5):050701, 2014.

[150] Robert C Mines, Tomasz Lipniacki, and Xiling Shen. Slow nucleosome dynamics set the transcriptional speed limit and induce rna polymerase ii traffic jams and bursts. PLoS computational biology, 18(2):e1009811, 2022.

[151] Aafke A van den Berg and Martin Depken. Crowding-induced transcriptional bursts dictate polymerase and nucleosome density profiles along genes. Nucleic acids research, 45(13):7623–7632, 2017.

[152] Tatiana Filatova, Nikola Popović, and Ramon Grima. Modulation of nuclear and cytoplasmic mrna fluctuations by time-dependent stimuli: analytical distributions. Mathematical biosciences, 347:108828, 2022.

[153] Benjamin T Donovan, Anh Huynh, David A Ball, Heta P Patel, Michael G Poirier, Daniel R Larson, Matthew L Ferguson, and Tineke L Lenstra. Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. The EMBO journal, 38(12):e100809, 2019.

[154] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences, 99(20):12795–12800, 2002.

[155] Damien Nicolas, Nick E Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? Molecular BioSystems, 13(7):1280–1290, 2017.

[156] Joseph Rodriguez and Daniel R Larson. Transcription in living cells: molecular mechanisms of bursting. Annual review of biochemistry, 89:189–212, 2020.

[157] Chen Jia, Abhyudai Singh, and Ramon Grima. Concentration fluctuations due to size-dependent gene expression and cell-size control mechanisms. bioRxiv, 2021.

[158] Lucy Ham, Rowan D Brackston, and Michael PH Stumpf. Extrinsic noise and heavy-tailed laws in gene expression. Physical review letters, 124(10):108101, 2020.

[159] Svitlana Braichenko, James Holehouse, and Ramon Grima. Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models. Journal of the Royal Society Interface, 18(183):20210510, 2021.

[160] Jaroslav Albert. A detailed model of gene promoter dynamics reveals the entry into productive elongation to be a highly punctual process. arXiv preprint arXiv:2201.13092, 2022.

[161] Qiwen Sun, Feng Jiao, Genghong Lin, Jianshe Yu, and Moxun Tang. The nonlinear dynamics and fluctuations of mrna levels in cell cycle coupled transcription. PLoS computational biology, 15(4):e1007017, 2019.

[162] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. Effects of cell cycle variability on lineage and population measurements of messenger rna abundance. Journal of the Royal Society Interface, 17(168):20200360, 2020.

[163] Juraj Szavits-Nossan and Ramon Grima. Mean-field theory accurately captures the variation of copy number distributions across the mrna life cycle. Physical Review E, 105(1):014410, 2022.

[164] Madeline Smith, Mohammad Soltani, Rahul Kulkarni, and Abhyudai Singh. Modulation of stochastic gene expression by nuclear export processes. In 2021 60th IEEE Conference on Decision and Control (CDC), pages 655–660. IEEE, 2021.

[165] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. Physical Review E, 102(2):022409, 2020.

[166] Alexander Hoffmann, Andre Levchenko, Martin L Scott, and David Baltimore. The iκb-nf-κb signaling module: temporal control and selective gene activation. science, 298(5596):1241–1245, 2002.

[167] Long Cai, Chiraj K Dalal, and Michael B Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. Nature, 455(7212):485–490, 2008.

[168] Nan Hao and Erin K O'shea. Signal-dependent dynamics of transcription factor translocation controls gene expression. Nature structural & molecular biology, 19(1):31–39, 2012.

[169] Pawel Paszek, Dean A Jackson, and Michael RH White. Oscillatory control of signalling molecules. Current opinion in genetics & development, 20(6):670–676, 2010.

[170] Arvind Murugan, Kabir Husain, Michael J Rust, Chelsea Hepler, Joseph Bass, Julian MJ Pietsch, Peter S Swain, Siddhartha G Jena, Jared E Toettcher, Arup K Chakraborty, et al. Roadmap on biology in time varying environments. Physical biology, 18(4):041502, 2021.

[171] Alexander M Berezhkovskii, Christine Sample, and Stanislav Y Shvartsman. How long does it take to establish a morphogen gradient? Biophysical journal, 99(8):L59–L61, 2010.

[172] Oliver Grimm, Mathieu Coppey, and Eric Wieschaus. Modelling the bicoid gradient. Development, 137(14):2253–2264, 2010.

[173] Michail E Kavousanakis, Jitendra S Kanodia, Yoosik Kim, Ioannis G Kevrekidis, and Stanislav Y Shvartsman. A compartmental model for the bicoid gradient. Developmental biology, 345(1):12–17, 2010.

[174] Stanislav Y Shvartsman and Ruth E Baker. Mathematical models of morphogen gradients and their effects on gene expression. Wiley Interdisciplinary Reviews: Developmental Biology, 1(5):715–730, 2012.

[175] Heng Xu, Leonardo A Sepúlveda, Lauren Figard, Anna Marie Sokac, and Ido Golding. Combining protein and mrna quantification to decipher transcriptional regulation. Nature methods, 12(8):739–742, 2015.

[176] Jitendra S Kanodia, Richa Rikhy, Yoosik Kim, Viktor K Lund, Robert DeLotto, Jennifer Lippincott-Schwartz, and Stanislav Y Shvartsman. Dynamics of the dorsal morphogen gradient. Proceedings of the National Academy of Sciences, 106(51):21707–21712, 2009.

[177] Congxin Li, François Cesbron, Michael Oehler, Michael Brunner, and Thomas Höfer. Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. Cell systems, 6(4):409–423, 2018.

[178] Justine Dattani and Mauricio Barahona. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. Journal of The Royal Society Interface, 14(126):20160833, 2017.

[179] Zhixing Cao and Ramon Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. Nature communications, 9(1):1–15, 2018.

[180] Feng Jiao, Genghong Lin, and Jianshe Yu. Approximating gene transcription dynamics using steady-state formulas. Physical Review E, 104(1):014401, 2021.

[181] Jakub Jędrak and Anna Ochab-Marcinek. Time-dependent solutions for a stochastic model of gene expression with molecule production in the form of a compound poisson process. Physical Review E, 94(3):032401, 2016.

[182] Maike MK Hansen, Ravi V Desai, Michael L Simpson, and Leor S Weinberger. Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. Cell systems, 7(4):384–397, 2018.

[183] Yu Rim Lim, Ji-Hyun Kim, Seong Jun Park, Gil-Suk Yang, Sanggeun Song, Suk-Kyu Chang, Nam Ki Lee, and Jaeyoung Sung. Quantitative understanding of probabilistic behavior of living cells operated by vibrant intracellular networks. Physical Review X, 5(3):031014, 2015.

[184] Seong Jun Park, Sanggeun Song, Gil-Suk Yang, Philip M Kim, Sangwoon Yoon, Ji-Hyun Kim, and Jaeyoung Sung. The chemical fluctuation theorem governing gene expression. Nature communications, 9(1):1–12, 2018.

[185] Sanggeun Song, Gil-Suk Yang, Seong Jun Park, Sungguan Hong, Ji-Hyun Kim, and Jaeyoung Sung. Frequency spectrum of chemical fluctuation: A probe of reaction mechanism and dynamics. PLoS computational biology, 15(9):e1007356, 2019.

[186] Filipe Tostevin, Wiet de Ronde, and Pieter Rein Ten Wolde. Reliability of frequency and amplitude decoding in gene regulation. Physical review letters, 108(10):108104, 2012.

[187] Otto G Berg. A model for the statistical fluctuations of protein numbers in a microbial population. Journal of theoretical biology, 71(4):587–603, 1978.

[188] Johan Paulsson, Otto G Berg, and Måns Ehrenberg. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proceedings of the National Academy of Sciences, 97(13):7148–7153, 2000.

[189] Qiuying Li, Lifang Huang, and Jianshe Yu. Modulation of first-passage time for bursty gene expression via random signals. Mathematical Biosciences & Engineering, 14(5&6):1261, 2017.

[190] William A Gardner, Antonio Napolitano, and Luigi Paura. Cyclostationarity: Half a century of research. Signal processing, 86(4):639–697, 2006.

[191] Alexandre Ferreira Ramos, Guilherme CP Innocentini, and José Eduardo Martinho Hornos. Exact time-dependent solutions for a self-regulating gene. Physical Review E, 83(6):062902, 2011.

[192] Chen Jia and Ramon Grima. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. The Journal of chemical physics, 152(17):174110, 2020.

[193] Chen Jia and Youming Li. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. bioRxiv, 2022.

[194] Darren J Wilkinson. Stochastic modelling for systems biology. Chapman and Hall/CRC, 2018.

[195] Lewis Wolpert, Cheryll Tickle, and Alfonso Martinez Arias. Principles of development. Oxford University Press, USA, 2015.

[196] Thomas Gregor, Eric F Wieschaus, Alistair P McGregor, William Bialek, and David W Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. Cell, 130(1):141–152, 2007.

[197] Sven Bergmann, Oded Sandler, Hila Sberro, Sara Shnider, Eyal Schejter, Ben-Zion Shilo, and Naama Barkai. Pre-steady-state decoding of the bicoid morphogen gradient. PLoS biology, 5(2):e46, 2007.

[198] Jonathan Liu, Donald Hansen, Elizabeth Eck, Yang Joon Kim, Meghan Turner, Simon Alamos, and Hernan G Garcia. Real-time single-cell characterization of the eukaryotic transcription cycle reveals correlations between rna initiation, elongation, and cleavage. PLoS computational biology, 17(5):e1008999, 2021.

[199] Hernan G Garcia, Mikhail Tikhonov, Albert Lin, and Thomas Gregor. Quantitative imaging of transcription in living drosophila embryos links polymerase activity to patterning. Current biology, 23(21):2140–2145, 2013.

[200] James Holehouse, Zhixing Cao, and Ramon Grima. Stochastic modeling of autoregulatory genetic feedback loops: A review and comparative study. Biophysical Journal, 118(7):1517–1525, 2020.

[201] Olivia Padovan-Merhar, Gautham P Nair, Andrew G Biaesch, Andreas Mayer, Steven Scarfone, Shawn W Foley, Angela R Wu, L Stirling Churchman, Abhyudai Singh, and Arjun Raj. Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. Molecular cell, 58(2):339–352, 2015.

[202] Samuel Marguerat and Jürg Bähler. Coordinating genome expression with cell size. Trends in Genetics, 28(11):560–565, 2012.

[203] Harry A Crissman and John A Steinkamp. Rapid, simultaneous measurement of dna, protein, and cell volume in single cells from large mammalian cell populations. The Journal of cell biology, 59(3):766, 1973.

[204] Hermannus Kempe, Anne Schwabe, Frédéric Crémazy, Pernette J Verschure, and Frank J Bruggeman. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. Molecular biology of the cell, 26(4):797–804, 2015.

[205] Jie Lin and Ariel Amir. Homeostasis of protein and mrna concentrations in growing cells. Nature communications, 9(1):1–11, 2018.

[206] Naama Brenner, Erez Braun, Anna Yoney, Lee Susman, James Rotella, and Hanna Salman. Single-cell protein dynamics reproduce universal fluctuations in cell populations. The European Physical Journal E, 38(9):1–9, 2015.

[207] Casper HL Beentjes, Ruben Perez-Carrasco, and Ramon Grima. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. Physical Review E, 101(3):032403, 2020.

[208] Chen Jia and Ramon Grima. Frequency domain analysis of fluctuations of mrna and protein copy numbers within a cell lineage: theory and experimental validation. Physical Review X, 11(2):021032, 2021.

[209] Damien Hall and Allen P Minton. Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1649(2):127–139, 2003.

[210] S Schnell and TE Turner. Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Progress in biophysics and molecular biology, 85(2-3):235–260, 2004.

[211] R Grima. Intrinsic biochemical noise in crowded intracellular conditions. The Journal of Chemical Physics, 132(18):05B604, 2010.

[212] Cheemeng Tan, Saumya Saurabh, Marcel P Bruchez, Russell Schwartz, and Philip LeDuc. Molecular crowding shapes gene expression in synthetic cellular nanosystems. Nature nanotechnology, 8(8):602–608, 2013.

[213] Ulysse Herbach, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. Inferring gene regulatory networks from single-cell data: a mechanistic approach. BMC systems biology, 11(1):1–15, 2017.

[214] Aleksandra M Walczak, Masaki Sasai, and Peter G Wolynes. Self-consistent proteomic field theory of stochastic gene switches. Biophysical journal, 88(2):828–850, 2005.

[215] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. Biosystems, 96(1):86–103, 2009.

[216] Vahid Shahrezaei, Julien F Ollivier, and Peter S Swain. Colored extrinsic fluctuations and stochastic gene expression. Molecular systems biology, 4(1):196, 2008.

[217] James F Epperson. An introduction to numerical methods and analysis. John Wiley & Sons, 2021.