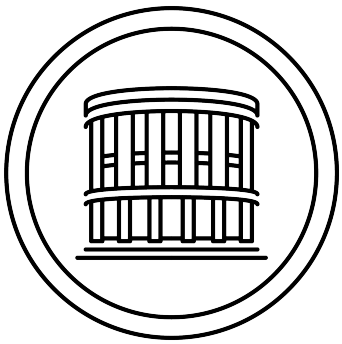
Comenius University in Bratislava

Faculty of Mathematics, Physics and Informatics

Department of Applied Mathematics and Statistics



Multivariate Markovian models of biological processes

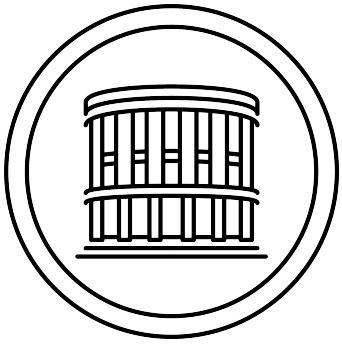
Dissertation Thesis

2022 Candan Çelik

Univerzita Komenského v Bratislave

Fakulta Matematiky, Fyziky a Informatiky

Katedra Aplikovanej Matematiky a Štatistiky



Viacrozmerné markovovské modely biologických procesov

Dizertačná Práca

Študijný program:	Aplikovaná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. Mgr. Pavol Bokes, PhD.

2022 Candan Çelik

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Candan Çelik
Study programme:	Applied Mathematics (Single degree study, Ph.D. III. deg., full time form)
Field of Study:	Mathematics
Type of Thesis:	Dissertation thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Multivariate Markovian models of biological processes

Annotation: The thesis will use analytic and simulation methods to study the Chapman-Kolmogorov equation that governs the dynamics of molecular distributions in complex biochemical systems.

Tutor:	doc. Mgr. Pavol Bokes, PhD.
Department:	FMFI.KAMŠ - Department of Applied Mathematics and Statistics
Head of department:	prof. RNDr. Marek Fila, DrSc.

Electronic version available:
archív

Assigned: 22.01.2018

Approved: 22.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
Guarantor of Study Programme

.....
Student

.....
Tutor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Candan Çelik
Študijný program: aplikovaná matematika (Jednoodborové štúdium, doktorandské III. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: dizertačná
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Multivariate Markovian models of biological processes
Viacrozmerné Markovovské modely biologických procesov

Anotácia:

Školiteľ: doc. Mgr. Pavol Bokes, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Spôsob sprístupnenia elektronickej verzie práce:
archív

Dátum zadania: 22.01.2018

Dátum schválenia: 22.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
školiteľ

Acknowledgements

First and foremost, I would like to thank my supervisor Pavol Bokes, who allowed me to produce this thesis work and many other insights he has provided during the exciting four years of developing my academic skills. I appreciate their extraordinary support.

I would like to thank Comenius University and the Faculty of Mathematics, Physics and Informatics for funding this doctoral project and allowing me to attend international conferences and summer schools.

I owe particular thanks to Rabeno Kuryel, who has been a lifelong learning partner, not only for the lectures I had the opportunity to take during my master's but also for broadening my horizons through his beneficial discussions.

I would like to express my special thanks to my outstanding parents, Hidayet and Sukriye, who have always believed in me and maintained their everlasting support. My sisters Gul, Gulden, Oya and Zekiye, I am grateful to you all for being always there for me. Your stunning daughters, Nehir and Oyku, made it much more enjoyable with their lovely chats. My centenarian grandma, who has always been supportive with her splitting jokes, deserves my special thanks.

I am particularly thankful for having the most thoughtful person ever, Radka, who never withheld her invaluable support in many ways during my studies.

Last but not least, having friends Adrian, Jakub, Martin, Patrik, and Petra has been a privilege for me with their helpful discussions.

Bratislava, April 2022

.....

Candan Çelik

Declaration

I hereby declare that I have produced this Dissertation Thesis without the prohibited assistance of third parties and without using aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such.

This thesis was conducted from 2018 to 2022 under the supervision of doc. Mgr. Pavol Bokes, PhD. at the Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovak Republic, and is comprised of the following publications:

- **Çelik, C.**, Bokes, P., Singh, A. (2022). "Translation regulation by RNA stem-loops can reduce gene expression noise". *Submitted for publication to BMC Bioinformatics*.
- **Çelik, C.**, Bokes, P., Singh, A. (2021). "Protein Noise and Distribution in a Two-Stage Gene-Expression Model Extended by an mRNA Inactivation Loop". In: Cinquemani, E., Paulevé, L. (eds) Computational Methods in Systems Biology. CMSB 2021. Lecture Notes in Computer Science(), vol 12881. Springer, Cham.
- **Çelik, C.**, Bokes, P., Singh, A. (2020). "Stationary Distributions and Metastable Behaviour for Self-regulating Proteins with General Lifetime Distributions". In: Abate, A., Petrov, T., Wolf, V. (eds) Computational Methods in Systems Biology. CMSB 2020. Lecture Notes in Computer Science(), vol 12314. Springer, Cham.

This document has not previously been presented in identical or similar form to any other Slovak or foreign examination board.

Bratislava, April 2022

.....

Candan Çelik

Abstract

Mathematical models of biochemical processes are essential tools to understanding the dynamics of intercellular events in living organisms. The copy number of species in such a system fluctuates in time due to the random occurrence of chemical reactions, leading to variability in the population of living cells. Therefore, characterising the number of species at a certain time requires stochastic models of particular interest. To this end, numerous modelling approaches have been proposed in both deterministic and stochastic settings to elucidate the dynamic behaviour of biochemical reaction networks.

Gene expression, the production process of gene products such as proteins, is one of such biological mechanisms intensively studied over the last decades. The (basic) two-stage model gives the essential description of gene expression, which entails the dynamics of transcription and translation processes involving mRNA and protein species. In the stochastic context, the dynamics of species in gene expression is given by the chemical master equation (CME). Despite its simple structure, finding a solution to the CME is often challenging, and analytical solutions are inaccessible for most systems. Consequently, numerical methods, including stochastic simulations, emerge as a remedy to the underlying problem.

On the other hand, with the recent advances in technology, experimental studies open a new window into the more advanced models capable of capturing the detailed dynamics of molecular systems. Accordingly, the present biochemical reaction models need to be revisited. In particular, gene expression is far more complex than the basic two-stage model. The mRNA molecules, for example, can switch between their active and inactive states.

In this thesis, we consider and study various structured chemical reaction systems tailored for gene expression, generalising the results of the basic two-stage model. Specifically, we first begin with a stochastic gene-expression model that accounts for a self-regulating protein molecule with exponential and phase-type lifetimes. We show that the one-dimensional and multiclass-multistage models exhibit the same stationary distribution in the case of non-bursty production of protein regime.

Second, we focus on extending the basic two-stage model to that involving an mRNA inactivation loop, where mRNA molecules are allowed to transition between

active and inactive states. For this model, we present a systematic mathematical derivation of the generating function of the stationary distribution of the species, providing the marginal protein distribution given in terms of special functions.

Next, we characterise the protein noise in terms of the two metrics, Fano factor and the squared coefficient of variation, concluding that the extended model exhibits less protein noise than the basic two-stage model. Importantly, we demonstrate how the models studied here play an important role in modelling the formation of stem–loops, thus controlling noise.

Finally, we present a generalisation of the two-stage and the extended model, which involves multiple mRNA states. We give a detailed mathematical analysis of the model, obtain marginal molecular distributions, and provide an additional gene expression example, which can be obtained from the generalised model.

Overall, in this thesis, we develop and study various gene expression models that may contribute to understanding the stochastic dynamics of biochemical reaction networks arising in relevant research fields.

Keywords: biochemical processes • Master equation • stochastic simulation • gene expression

Abstrakt

Matematické modely biochemických procesov sú základnými nástrojmi na chápanie dynamiky medzibunkových udalostí v živých organizmoch. Fluktuácia počtu kópií druhov v takomto systéme v čase je dôsledkom náhodného výskytu chemických reakcií, čo vedie ku variabilite populácie živých buniek. Preto si charakterizácia počtu druhov v určitom čase vyžaduje stochastické modely, ktoré majú osobitný záujem. Na tento účel bolo navrhnuté mnozstvo prístupov k modelovaniu aj v deterministických, aj v stochastických nastaveniach, aby objasnili dynamické správanie sietí biochemických reakcií.

Génová expresia, proces tvorby génových produktov, ako sú proteíny, je jedným z takýchto biologických mechanizmov intenzívne študovaných v posledných desaťročiach. (Základný) dvojstupňový model poskytuje podstatný popis génovej expresie, ktorý zobrazuje dynamiku transkripčných a translačných procesov zahŕňajúcich mRNA a proteínové druhy. V stochastickom kontexte, dynamika druhov v génovej expresii je daná chemickou Master rovnicou (CME). Napriek jednoduchej štruktúre je hľadanie riešenia pre CME často náročné a analytické riešenia sú pre väčšinu systémov nedostupné. V dôsledku toho sa ako náprava objavujú numerické metódy, vrátane stochastických simulácií.

Na druhú stranu, vďaka nedávnym pokrokom v technológii sa otvárajú experimentálne štúdie nové okno do pokročilejších modelov, ktoré sú schopné zachytiť detailnú dynamiku molekulárnych systémov. V súlade s tým je potrebné prehodnotiť súčasné modely biochemických reakcií. Najmä génová expresia je oveľa zložitejšia ako základný dvojstupňový model. Napríklad, molekuly mRNA sa môžu prepínať medzi aktívnym a neaktívnym stavmi.

V tejto práci uvažujeme a študujeme rôzne štruktúrované chemické reakčné systémy prispôsobené pre génovú expresiu, následne zovšeobecňujeme výsledky základného dvojstupňového modelu. Najprv začíname so stochastickým modelom génovej expresie, ktorý zodpovedá za samoregulačnú proteínovú molekulu s exponenciálnym a fázovým typom života. Ukazujeme, ze keď režim produkcie proteínu je neprerušovaným (bez náhodnych pulzov), jednorozmerné a viactriedne-viacstupňové modely vykazujú rovnaké stacionárne rozdelenie.

Následne sa zameriavame na rozšírenie základného dvojstupňového modelu na

model zahŕňajúci inaktivačnú slučku mRNA, kde je pre molekuly mRNA umožnený prechod medzi aktívnymi a neaktívnymi stavmi. Pre tento model uvádzame systematické matematické odvodenie generujúcej funkcie stacionárneho rozdelenia, poskytujúce marginálne rozdelenie proteínu vyjadrené pomocou špeciálnych funkcií.

Ďalej charakterizujeme proteínový šum z hľadiska dvoch metrík, Fano faktora a štvorcového variačného koeficientu, pričom sme dospeli k záveru, že rozšírený model vykazuje menší proteínový šum ako základný dvojstupňový model. Toto je dôležitá demonštrácia toho, ako študované modely hrajú dôležitú úlohu pri modelovaní utvárania kmeňových slučiek, tým kontrolujúc šum.

Nakoniec uvádzame zovšeobecnenie dvojstupňového a rozšíreného modelu, ktorý zahŕňa viacero stavov mRNA. Poskytujeme podrobnú matematickú analýzu modelu, získavame marginálne molekulárne rozdelenia a poskytujeme ďalší príklad génovej expresie, ktorý možno získať zo zovšeobecneného modelu.

Celkovo v tejto práci vyvíjame a študujeme rôzne modely génovej expresie, ktoré môžu prispieť k pochopeniu stochastickej dynamiky sietí biochemických reakcií vznikajúcich v relevantných oblastiach výskumu.

Kľúčové slová: biochemické procesy • Master rovnica • stochastické simulácie • génová expresia

List of Abbreviations

BDP	birth–death process
CME	chemical master equation
ODE	ordinary differential equation
PDE	partial differential equation
SSA	stochastic simulation algorithm

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent decades, the study of complex systems has gained significant attention in many research fields, including physics, chemistry and biological sciences. In particular, the models involving chemical reaction kinetics have become an in-demand research topic of interest. In simplest assumptions, the dynamics of such chemically reacting systems are modelled deterministically [1]. However, the number of species in a reaction system is often subject to inherent stochasticity, which requires a rigorous understanding of mathematical modelling. To this end, many different deterministic and stochastic approaches have been proposed to find a solution to the underlying problem, analytical or numerical.

The deterministic description of the dynamics of biochemical systems is given by the reaction rate equations, which consist of a set of ordinary differential equations. In particular, when the system of interest involves a large number of molecules, deterministic approaches are adequate to govern the dynamics of the mean number of species in the system. Unsurprisingly, the stochastic effects come into play for the systems with a low copy number of molecules and many other biological processes, which are inherently subject to random fluctuations due to the production and degradation reactions, e.g., gene expression [2]. These stochastic effects are characterised by the associated probability distributions, whose description is given by the chemical master equation (CME) capturing the full random dynamics.

The CME consists of a finite set of differential equations describing the time evolution of the probability of being in a state at a given time. Except for simple cases, no analytical solutions to the CME are obtained. Consequently, stochastic simulation algorithms that enable us to generate sample time trajectories of species

in a system have been developed as an alternative approach rather than seeking a solution to the underlying CME [3, 4]. However, using a stochastic simulation algorithm can come with a high computational cost for systems including many species.

One of the widely-studied chemical reaction networks where random fluctuations in the number of species play an essential role is the stochastic model of gene expression. The model describes the synthesis of the gene products such as protein and messenger ribonucleic acid (mRNA). From a modelling viewpoint, the simplest gene expression mechanism consists of two main steps: transcription and translation; it is therefore referred to as the two-stage gene-expression model in the literature. While an mRNA molecule is produced during the transcription, translation of the "transcribed information" takes place to produce protein in the second step. Due to the random occurrence of the production and decay reactions, the copy number of species in the two-stage model of gene expression fluctuates in time. These random fluctuations are referred to as noise.

Noise in gene expression can stem from many different sources. For example, transcription factors that are gene-specific proteins that bind to DNA to regulate the transcription rate play an essential role in gene expression noise [5, 6]. Recent studies have shown that the lifetime of such regulatory molecules can assume a far more complex model than simple exponential, where the (basic) two-stage model remains inadequate for tracking the system's full dynamics [7, 8]. Another biologically possible scenario by which stochasticity can be regulated is the inclusion of an mRNA activation loop into the (basic) two-stage model, turning active mRNA molecules into inactive and vice versa.

As outlined above, motivated by relevant studies in the literature, we focus on multivariate gene expression models in this thesis.

1.2 Organisation

This thesis is organised as follows. Chapter 2 presents fundamental mathematical concepts and methods arising in biochemical systems, which will be used throughout the following chapters. In Section 2.1, we review the two-state chain

and introduce the associated Master equation together with its solution. We then give a simple example of a chemical reaction system that can be modelled using the two-state chain in Section 2.2. We next, in Sections 2.3–2.4, present the relationship between chemical reaction networks and queueing theory, providing the necessary elementary information about some simple queuing models we will use in the subsequent chapters. In Section 2.5, we present the generating function technique for solving a particular class of systems of differential equations. We demonstrate in detail the stochastic simulation algorithm, of which we give two specific examples in Section 2.6. Next, we present how reaction kinetics works in general settings in Section 2.7. After briefly reviewing gene expression in Section 2.8, we finalise this chapter by introducing the basic two-stage gene-expression model, which will be frequently used throughout this thesis.

In Chapter 3 we study a stochastic gene-expression model for a self-regulating transcription factor whose lifespan (or time till degradation) follows a general distribution modelled as per a multi-dimensional phase-type process. In Section 3.2, we formulate, both in the deterministic and stochastic settings, a one-dimensional model for the abundance of a transcription factor with a memoryless lifetime, allowing the production rate to vary with the copy number. We show that steady states of the deterministic model are given by the fixed points of the feedback response function. We next characterise the steady-state behaviour of a structured multiclass–multistage model that accounts for complex lifetime pathways in Section 3.3. We demonstrate that the deterministic fixed points and the stochastic stationary distribution for the one-dimensional framework remain valid for the total protein amount in the multi-dimensional setting. We provide explicit counter-examples in Section 3.4 to show that the distribution invariance result rests on the assumption of non-bursty production of protein. In Section 3.5, we approximate the stochastic protein distribution by a mixture of Gaussians and study the rates of metastable transitions between the Gaussian modes in the one- and multi-dimensional settings.

Chapter 4 focuses on a two-stage stochastic gene expression model that extends the standard model by an mRNA inactivation loop. The model is introduced in Section 4.2; the stationary means are obtained from the system of deterministic rate equations; the CME is formulated. We give a detailed mathematical breakdown of

the associated CME and rederive the stationary means using factorial cumulants in Section 4.3. Next, we express the Fano factor (variance-to-mean ratio) in terms of these cumulants in Section 4.4. Additionally, we provide a special-function representation of the sought-after joint generating function in Section 4.5. We then obtain the marginal mRNA and protein distributions employing the generating function of the stationary distribution in Section 4.6. Furthermore, we perform stochastic simulations to compare the theoretical results with those obtained by the simulations.

Chapter 5 presents the application of gene-expression model, which is studied in the previous chapter. More specifically, we briefly summarise our motivation for studying this model by providing a biological example of the formation of stem–loops in Section 5.1. Next, we characterise the noise in the basic two-stage model and the extended model in terms of the Fano factor and the squared coefficient of variation in Section 5.2. In Section 5.3, we study noise control by stem–loops in detail, obtaining the Fano factor and the CV^2, which leads us to conclude that incorporating an mRNA inactivation loop into the basic two-stage model decreases the protein noise.

In Chapter 6, we generalise the extended model of Chapter 4 to a structured multivariate model which considers multiple mRNA states. The model and its corresponding CME are formulated in Section 6.2. Next, we focus on a comprehensive mathematical analysis of the model to obtain a solution to the corresponding PDE in Section 6.3. In the next Section 6.4, we obtain the marginal mRNA distributions and determine the moments of the protein distribution. We then discuss two distinct gene-expression models, the mRNA inactivation loop and the multiphasic model, which can be obtained as special cases from the structured model in Sections 6.5–6.6.

The thesis is summarised and concluded in Chapter 7.

Chapter 2

Preliminaries

This chapter introduces fundamental concepts and methods that often arise in the mathematical modelling of chemical reaction systems. These introductory topics will be used in the following chapters. First, we start with the simplest Markovian model, i.e. the two-state model, that is often used to describe random switching in chemical reaction systems. We then present step-by-step how to model a degradation reaction in deterministic and stochastic contexts, providing its corresponding ODE and chemical master equation. Next, we show the relationship between chemical reaction networks and queueing systems by introducing some basic Markovian models of queues. Then, we present *the generating function method* for solving the CME. Additionally, we provide the SSA for obtaining sample time trajectories of species in a system and discuss general reaction kinetics. Finally, we briefly overview the stochastic gene expression process on which the following chapters will be based.

2.1 The two-state model

A two-state chain is one of the simplest models widely used in stochastic modelling. As its name indicates, the chain involves two states, say state 1 and state 2, where random transitions occur from each of which at constant rates p and q, respectively. A graphical representation of the model is shown in Figure 2.1. The rate constant p is defined so that $p\mathrm{d}t$ gives the probability that a single transition from state 1 to state 2 occurs in the time interval $[t, t+\mathrm{d}t)$ where t stands for time and $\mathrm{d}t$ is an infinitesimally small time step.

A question of interest is that, starting at time $t = 0$, how the probabilities $p_1(t)$ and $p_2(t)$ of being found in state 1 and 2 at time t, respectively, evolve as time goes on. To answer this question, let us determine all possible transitions as follows: i)

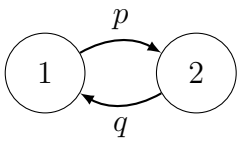


Figure 2.1: Schematic of the two-state chain. Random transitions occur between states 1 and 2 at constant rates p and q, respectively.

the system can be in state 1 at time t and have not moved to state 2 during the time interval $[t, t+\mathrm{d}t)$ or ii) it can be in state 2 and can move to state 1 in the time interval $[t, t+\mathrm{d}t)$. Note that we here neglect the terms of $\mathcal{O}(\mathrm{d}t^2)$ due to the fact that these terms will be vanishing in the limit $\mathrm{d}t \to 0$. Hence, we obtain

$$p_1(t+\mathrm{d}t) = p_1(t)[1 - p\mathrm{d}t] + p_2(t)q\mathrm{d}t. \tag{2.1}$$

Rearranging (2.1) and taking the limit $\mathrm{d}t \to 0$ yields the following ordinary differential equation (ODE)

$$\frac{\mathrm{d}p_1(t)}{\mathrm{d}t} = -pp_1(t) + qp_2(t) \tag{2.2}$$

for $p_1(t)$. Likewise, we can obtain

$$\frac{\mathrm{d}p_2(t)}{\mathrm{d}t} = -qp_2(t) + pp_1(t) \tag{2.3}$$

for $p_2(t)$. It is worth noting that the set of equations (2.2)–(2.3) is nothing else than a linear ODE system, but it is referred to as *the differential Chapman-Kolmogorov equation* in mathematics, *the master equation* in physics, and *the chemical master equation* in chemistry.

The ODE system (2.2)–(2.3) is subject to initial conditions

$$p_1(0) = p_1^{(0)}, \quad p_2(0) = p_2^{(0)}, \tag{2.4}$$

where $p_1^{(0)}$ and $p_2^{(0)}$ are prescribed distributions that characterise the state of the system at the initial time, satisfying the normalisation condition $p_1^{(0)} + p_2^{(0)} = 1$. In particular, when the process of interest is initialised in state 1, equation (2.4) takes the form of

$$p_1^{(0)} = 1, \quad p_2^{(0)} = 0. \tag{2.5}$$

For the sake of simplicity, the system (2.2)–(2.3) can be expressed in matrix form as

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathbf{Q}^T\mathbf{p}, \tag{2.6}$$

where the transition matrix is given by

$$\mathbf{Q} = \begin{pmatrix} -p & p \\ q & -q \end{pmatrix},$$

and

$$\mathbf{p}(t) = \begin{pmatrix} p_1(t) & p_2(t) \end{pmatrix}^T$$

is the state vector.

The explicit solution to the system (2.2)–(2.3) subject to the initial conditions (2.4) satisfying (2.5) can be obtained as

$$p_1(t) = \frac{q + p\mathrm{e}^{-(p+q)t}}{p+q}, \tag{2.7}$$

$$p_2(t) = \frac{p\left(1 - \mathrm{e}^{-(p+q)t}\right)}{p+q}. \tag{2.8}$$

At steady state, the system (2.2)–(2.3) reduces to the algebraic expressions

$$0 = -pp_1^{ss} + qp_2^{ss},$$

$$0 = -qp_2^{ss} + pp_1^{ss},$$

where $p_1(t) \equiv p_1^{ss}$ and $p_2(t) \equiv p_2^{ss}$ are the stationary probabilities that are satisfied by

$$p_1^{ss} = \frac{q}{p+q}, \quad p_2^{ss} = \frac{p}{p+q}. \tag{2.9}$$

We note that stationary probabilities (2.9) coincide with the limit values of the time-dependent results (2.7)–(2.8) as $t \to \infty$.

2.2 A simple model of degradation

Let us present an elementary chemical reaction system, which can be thought of as a collection of independent two-state chains. Consider the following degradation reaction given by

$$\mathrm{X} \xrightarrow{k} \emptyset, \tag{2.10}$$

where X is a chemical species of interest decaying at constant reaction rate k. The symbol $\emptyset$ stands for chemical species that are of no interest. The reaction (2.10) can deterministically be modelled by

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -kX, \tag{2.11}$$

where $X = X(t)$ denotes the number of X molecules at time t. The left hand side of (2.11) denotes the change in the number of X with respect to time, whereas the right hand side implies that the change is proportional to the decay rate k . The solution is, clearly, $X(t) = X(0)\mathrm{e}^{-kt}$.

In a stochastic setting, the reaction rate k is reinterpreted so that $k\mathrm{d}t$ gives the probability of degradation of a species X in the time interval $[t, t+\mathrm{d}t)$ [9]. Particularly, the probability that exactly one degradation reaction occurs during the time interval $[t, t + \mathrm{d}t)$ is given by $X(t)k\mathrm{d}t$. Consequently, we have the following possible reactions occuring in $[t, t + \mathrm{d}t)$:

$$\mathbb{P}(\text{exactly one reaction}) = X(t)k\mathrm{d}t + \mathcal{O}(\mathrm{d}t^2), \tag{2.12}$$

$$\mathbb{P}(\text{no reactions}) = 1 - X(t)k\mathrm{d}t + \mathcal{O}(\mathrm{d}t^2), \tag{2.13}$$

$$\mathbb{P}(\text{two or more reactions}) = \mathcal{O}(\mathrm{d}t^2), \tag{2.14}$$

where $\mathbb{P}$ stands for the probability that the event of interest occurs.

The Chemical Master Equation Let $p_n(t)$ denote the probability of having n molecules of X in the system at time t. The probability that one of the reactions (2.12)–(2.14) occurs in $[t, t + \mathrm{d}t)$ is then given by

$$p_n(t + \mathrm{d}t) = p_n(t)(1 - kn\mathrm{d}t) + p_{n+1}(t)k(n + 1)\mathrm{d}t. \tag{2.15}$$

Rearranging the terms in (2.15) and dividing both sides by $\mathrm{d}t$ yields

$$\frac{p_n(t + \mathrm{d}t) - p_n(t)}{\mathrm{d}t} = k(n + 1)p_{n+1}(t) - knp_n(t). \tag{2.16}$$

Taking the limit $\mathrm{d}t \to 0$ of (2.16), we arrive at

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = k(n + 1)p_{n+1} - knp_n, \quad n \geq 0. \tag{2.17}$$

Equation (2.17) is known as "the chemical master equation (CME)" [10], and denotes an infinite system of ODEs for the probabilities p_n, where $n = 0, 1, 2, \ldots$. Given an initial condition such as

$$p_n(0) = \begin{cases} 1 & \text{if } n = n_0, \\ 0 & \text{otherwise}, \end{cases} \tag{2.18}$$

where $n \leq n_0$, reduces (2.17) to a finite set of ODEs. Setting $n = n_0$ gives

$$\frac{\mathrm{d}p_{n_0}}{\mathrm{d}t} = -kn_0p_{n_0}, \quad \text{where} \quad p_{n_0}(0) = 1,$$

which admits the solution

$$p_{n_0}(t) = \mathrm{e}^{-kn_0t}. \tag{2.19}$$

Inserting (2.19) into (2.17) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{n_0-1}(t) = -k(n_0-1)p_{n_0-1}(t) + kn_0\mathrm{e}^{-kn_0t}. \tag{2.20}$$

Using the method of variation of constants, we seek for a solution to (2.20). The solution $p_h(t)$ to the homogeneous equation

$$p_{n_0-1}(t) + k(n_0-1)p_{n_0-1}(t) = 0 \tag{2.21}$$

is obtained as

$$p_h(t) = \mathrm{e}^{-k(n_0-1)t}. \tag{2.22}$$

We then assume that the general solution to (2.20) is in the form of

$$p_{n_0-1}(t) = v(t)p_h(t). \tag{2.23}$$

To obtain $v(t)$, we substitute (2.23) into (2.20), and solve for $v(t)$. Elementary calculations yields

$$\begin{aligned} v(t) &= \int \frac{kn_0\mathrm{e}^{-kn_0t}}{\mathrm{e}^{-k(n_0-1)t}}\mathrm{d}t \\ &= -kn_0\int \mathrm{e}^{-kt}\mathrm{d}t \\ &= -n_0\mathrm{e}^{-kt} + C, \quad C \text{ is constant.} \end{aligned}$$

Imposing initial condition $p_{n_0-1}(0) = 0$ gives

$$v(t) = n_0\left(1-\mathrm{e}^{-kt}\right). \tag{2.24}$$

Substituting (2.22) and (2.24) into (2.23) reads

$$p_{n_0-1}(t) = n_0\left(1-\mathrm{e}^{-kt}\right)\mathrm{e}^{-k(n_0-1)t},$$

which is the sought-after solution to (2.20). Furthermore, we can inductively show that

$$p_n(t) = \binom{n_0}{n}\left(1-\mathrm{e}^{-kt}\right)^{n_0-n}\mathrm{e}^{-knt}, \tag{2.25}$$

which corresponds to the *Binomial distribution* [11]. Equation (2.25) can be used to quantify the characteristics (i.e. the mean and variance) of the degradation reaction (2.10).

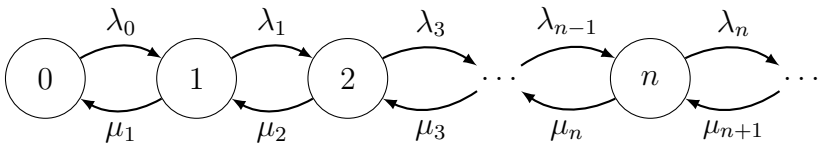


Figure 2.2: A schematic of a birth-death process. Birth and death transitions occur at rates λ_n and μ_n, respectively.

The chemical master equation (2.17) can be written in matrix form as

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathbf{Q}^\top \mathbf{p},$$

where $\mathbf{p}(t) = \begin{bmatrix} p_0(t) & p_1(t) & \dots & p_n(t) & \dots \end{bmatrix}^\top$ is the vector of probabilities, and $\mathbf{Q}$ is given by

$$\mathbf{Q} = k \begin{bmatrix} 0 & 0 & 0 & \dots \\ 1 & -1 & 0 & \dots \\ & 2 & -2 & \dots \\ & & & \ddots \end{bmatrix}.$$

We here note that the matrix $\mathbf{Q}$ and the vector $\mathbf{p}$ are different from those in (2.6).

2.3 Birth–Death processes

Birth–death processes (BDPs) have been extensively studied to model population dynamics in biology, chemistry, and other related fields such as mathematics. Formally, BDPs are a specific class of continuous-time Markov chains in which the number of species of interest in a system is modelled. The states (n) of the Markov chain are the nonnegative integers from which only transitions to adjacent states are permitted. A schematic representation of the transition rates in a BDP is given in Figure 2.2. In particular, BDPs have been widely used in queueing theory to model arrivals and departures in a waiting system. Following queueing theory terminology, the states represent the number of "customers" in the system. A customer's arrival to the system is referred to as "birth", whereas a "death" corresponds to a customer's departure. The time until the next arrival and departure are exponential random variables with rates λ_n and μ_n. Upon an arrival and departure, the system moves from state n to $n+1$ and n to $n-1$, respectively. Below, we will derive a system of difference-differential equations that describes the dynamics of birth-death process.

Let $p_n(t)$ denote the probability that the number of customers (or "population") is equal to n at time t. Assuming that births and deaths are independent, we can make the following statements for a birth transition in $(t, \Delta t]$:

$$\mathbb{P}(\text{one birth}) = \lambda_n \Delta t + o(\Delta t) \tag{2.26}$$

$$\mathbb{P}(\text{no birth}) = 1 - \lambda_n \Delta t + o(\Delta t), \tag{2.27}$$

$$\mathbb{P}(\text{two or more birth}) = o(\Delta t), \tag{2.28}$$

where $n \geq 0$, and Δt is an infinitesimally small time step. By $o(\Delta t)$, we mean that the probability that any other birth or death other than those stated above occurs is negligibly small, i.e., of order $o(\Delta t)$. Likewise, for $n > 0$, we have

$$\mathbb{P}(\text{one death}) = \mu_n \Delta t + o(\Delta t) \tag{2.29}$$

$$\mathbb{P}(\text{no death}) = 1 - \mu_n \Delta t + o(\Delta t), \tag{2.30}$$

$$\mathbb{P}(\text{two or more death}) = o(\Delta t) \tag{2.31}$$

for a death transition in the time interval $(t, \Delta t]$. The system of difference-differential equations for this BDP can be obtained via the expression for $p_n(t+\Delta t)$ and taking the limit of $[p_n(t + \Delta t) - p_n(t)]/\Delta t$ as $\Delta t \to 0$ (c.f. Section (2.2)). It can also be obtained by writing down the *balance equations,* which state that the *outflux* rate of transitions from a given state must be equal to the *influx* rate of transitions. In this case, the influx rate is given by

$$\lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t), \tag{2.32}$$

whereas the outflux rate is given by

$$(\lambda_n + \mu_n) p_n(t). \tag{2.33}$$

Combining (2.32) and (2.33), we obtain the *net* probability flow into state n as

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t), \quad n \geq 1, \tag{2.34}$$

$$\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = \mu_1 p_1(t) - \lambda_0 p_0(t). \tag{2.35}$$

Here we note that (2.35) gives the flow rate at the boundary state. At steady state, i.e. equating the time derivatives in (2.34) and (2.35) to zero, a solution to the system (2.34)–(2.35) can be obtained as follows. We rewrite the equations (2.34)–(2.35)

recursively as

$$p_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} p_n - \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1}, \quad n \geq 1, \tag{2.36}$$

$$p_1 = \frac{\lambda_0}{\mu_1} p_0, \tag{2.37}$$

Substituting (2.37) into (2.36) gives

$$p_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0. \tag{2.38}$$

Likewise, using recursion, we get

$$p_3 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0, \tag{2.39}$$

which, intuitively, leads to

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}, \quad \text{for } n \geq 1. \tag{2.40}$$

Formula (2.40) can be proven by using mathematical induction. The normalisation condition

$$\sum_{i=0}^{\infty} p_i = 1$$

implies that

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1}. \tag{2.41}$$

Note that equation (2.41) provides a necessary and sufficient condition for the existence of a steady-state solution when

$$1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} < \infty.$$

In a BDP, the transition rates (2.26)–(2.28) and (2.29)–(2.31) lead to the following generator matrix:

$$\mathbf{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\mu_1 + \lambda_1) & \lambda_1 & & \\ & \mu_2 & -(\mu_2 + \lambda_2) & \lambda_2 & \\ & & & \ddots & \end{bmatrix}, \tag{2.42}$$

which is a (tridiagonal) square band matrix whose elements only on the main diagonal, the subdiagonal, and the superdiagonal are nonzero. In the following sections, we shall briefly introduce a few simple queueing models that underline the processes we will be studying in the following chapters.

2.4 Basic Markovian queueing systems

A chemical reaction system can be interpreted as an infinite server queueing network [12]. For a chemical species of interest in the system, the reactions production, decay, and conversion correspond to an "*arrival*", "*departure*", and the "*movement*" of a "*customer*" from one state (or station) to another, respectively. The movement of the customers (species) in the system is independent of each other. Following these key points, it can be inferred that queueing theory thereby provides insights into such reaction networks in stochastic processes. Here we introduce some of simple queueing models which we will be using in the following chapters.

The M/M/1 queue Let us first consider the simplest single-server queueing model: the $M/M/1$ queue. It is used to describe the waiting times in a queueing system such as communication networks. The two letters M in its abbreviation stand for *memorylessness* or *Markovian* while the number 1 denotes the number of servers. The inter-arrival and the service times are assumed to be exponentially distributed. The arrivals occur in accordance with Poisson distribution with rate λ

$$\mathrm{P}[N(t) = k] = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots,$$

where $N(t)$ is the number of arrivals [11]. The service times have also an exponential distribution with rate parameter μ. The $M/M/1$ queue model can be described as a continuos time Markov chain with the generator matrix

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & & \ddots & \end{bmatrix},$$

whose state space $S = \{0, 1, 2, \ldots\}$ corresponds to the number of customers in the system [13].

The M/M/c queue Next, we consider the $M/M/c$ queue which consists of c identical servers providing independent and identically distributed exponential service at rate μ. The arrival of customers follows a Poisson process with rate λ.

Costumers are served according to principle "first come first served" as long as there is a free server; otherwise, costumers join the queue. Let assume that the number n of customers are greater then the number c of servers. The generator matrix for this process can then be written as

$$\mathbf{Q} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c \\ c+1 \\ \vdots \end{array} \begin{bmatrix} -\lambda & \lambda & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & \\ & & \ddots & & & \\ & c\mu & -(c\mu+\lambda) & \lambda & & \\ & & c\mu & -(c\mu+\lambda) & \lambda & \\ & & & & \ddots \end{bmatrix},$$

which is a tridiagonal matrix having the same rates when all servers are unavailable.

The $\mathrm{M}/\mathrm{M}/\infty$ queue We now focus on the $M/M/\infty$ queue which has infinitely many servers. A schematic of this queue is given in Figure 2.3. The model can be thought as the limit of c servers as $c \to \infty$. The arrivals follow a Poisson process, and the service times have an exponential distribution. As before, the arrival and service rates are denoted by λ and μ, respectively. The generator matrix is given by

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & \\ & & 3\mu & -(3\mu+\lambda) & \lambda \\ & & & & \ddots \end{bmatrix}. \tag{2.43}$$

The differential-difference equations for the $M/M/\infty$ queue can be obtained as

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = \lambda p_{n-1} + \mu(n+1)p_{n+1} - (\lambda + \mu n)p_n, \quad n \geq 1, \tag{2.44}$$

$$\frac{\mathrm{d}p_0}{\mathrm{d}t} = -\lambda p_0 + \mu p_1. \tag{2.45}$$

The time-dependent system of equations (2.44)–(2.45) can be solved using the generating function method [14], which will be reviewed in Section 2.5.

Note that in the context of chemical reactions the $M/M/\infty$ queue can be thought as a birth and death process (cf. Section 2.3) given by

$$\emptyset \xrightarrow{\lambda} \mathrm{X}, \quad \mathrm{X} \xrightarrow{\mu} \emptyset, \tag{2.46}$$

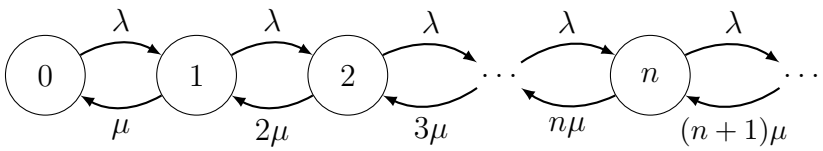


Figure 2.3: Transition rates diagram for the $M/M/\infty$ queue.

where the chemical species X is produced from an inexhaustible source of constituents denoted by $\emptyset$; afterwards, it is degraded.

Rather than finding the full time dependent probabilities, we focus on the stationary case, i.e. we solve in $\mathbf{p}$ the algebraic system

$$\mathbf{Q}^\top \mathbf{p} = 0, \tag{2.47}$$

where $\mathbf{Q}^\top$ is the transpose of the generator matrix (2.43) and $\mathbf{p}$ is the state vector. By (2.47), we have

$$\begin{aligned} \lambda p_0 &= \mu p_1, \\ (\lambda + \mu) p_1 &= \lambda p_0 + 2\mu p_2, \quad n = 1, \\ (\lambda + n\mu) p_n &= \lambda p_{n-1} + (n+1)\mu p_{n+1}, \quad n = 2, 3, \ldots \end{aligned} \tag{2.48}$$

which admit the solution

$$p_n = \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n p_0, \quad n = 0, 1, 2, \ldots. \tag{2.49}$$

The normalisation condition $\sum_0^\infty p_n = 1$ implies that

$$p_0 = \exp\left(-\frac{\lambda}{\mu}\right). \tag{2.50}$$

Inserting (2.50) into (2.49) gives

$$p_n = \exp\left(-\frac{\lambda}{\mu}\right) \frac{(\lambda/\mu)^n}{n!}, \quad n = 0, 1, 2, \ldots, \tag{2.51}$$

which implies that the stationary distribution of the number of customers in a $M/M/\infty$ queue is Poisson with mean λ/μ.

The step operator formalism Master equations (2.17) can be expressed in a more systematic and compact form using the step operator notation [10]. The linear operator $\mathbb{E}$ is defined by any function $f(n)$ where n takes integer values such that

$$\mathbb{E}^k[f(n)] = f(n+k), \quad \text{and} \quad \mathbb{E}^{-k}[f(n)] = f(n-k). \tag{2.52}$$

In particular,

$$\mathbb{E}[f(n)] = f(n+1) \quad \text{and} \quad \mathbb{E}^{-1}[f(n)] = f(n-1). \tag{2.53}$$

The master equation (2.17) for the decay process (2.10) can be rewritten as

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = k(\mathbb{E}-1)np_n, \quad n \geq 0. \tag{2.54}$$

Similarly, the master equation (2.44) for the $M/M/\infty$ can be expressed as

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = \lambda\left(\mathbb{E}^{-1}-1\right)p_n + \mu\left(\mathbb{E}-1\right)np_n, \quad n \geq 1. \tag{2.55}$$

Equations (2.54) and (2.55) are subject to the initial condition (2.18).

In the next section, we will describe the generating function method for solving the chemical master equation.

2.5 Generating functions

In this section we present the generating function technique for solving the system of equations given by (2.44). Generating functions are useful, in particular, for solving a certain class of difference equations.

Let $G(z,t)$ denote the generating function for a probability distribution $p_n(t)$ defined as

$$G(z,t) = \sum_{n=0}^{\infty} z^n p_n(t). \tag{2.56}$$

Differentiating (2.56) with respect to z gives

$$\frac{\partial G}{\partial z} = \sum_{n=0}^{\infty} nz^{n-1}p_n, \quad \frac{\partial^2 G}{\partial z^2} = \sum_{n=0}^{\infty} n(n-1)z^{n-2}p_n. \tag{2.57}$$

Evaluating (2.57) at $z=1$ gives

$$\left.\frac{\partial G(z,t)}{\partial z}\right|_{z=1} = \sum_{n=0}^{\infty} np_n(t) = \langle n(t)\rangle, \tag{2.58}$$

$$\left.\frac{\partial^2 G(z,t)}{\partial z^2}\right|_{z=1} = \sum_{n=0}^{\infty} n(n-1)p_n(t) = \langle n^2(t)\rangle - \langle n(t)\rangle. \tag{2.59}$$

Equations (2.58) and (2.59) correspond to the (first and second) factorial moments of $p_n(t)$; the function $F(u,t) = G(1+u,t)$ is thereby referred as to "*factorial moment generating function*" [11, 15]. Note that the normalisation condition implies that

$$G(z,t)|_{z=1} = \sum_{n=0}^{\infty} p_n(t) = 1. \tag{2.60}$$

Let us now focus on solving (2.44) using the generating function (2.56). To do so, we multiply (2.44) by the factor z^n and sum over all n to obtain

$$\frac{\partial G(z,t)}{\partial t} = \lambda \sum_{n=1}^{\infty} z^n p_{n-1} + \mu \sum_{n=0}^{\infty} (n+1) z^n p_{n+1} - \lambda \sum_{n=0}^{\infty} z^n p_n - \mu \sum_{n=0}^{\infty} n z^n p_n. \quad (2.61)$$

Factoring z out of the first and last sums on the right hand side of (2.61) so that we can utilise the formulae (2.56)–(2.57) gives

$$\frac{\partial G(z,t)}{\partial t} = \lambda z \sum_{n=1}^{\infty} z^{n-1} p_{n-1} + \mu \sum_{n=0}^{\infty} (n+1) z^n p_{n+1} - \lambda \sum_{n=0}^{\infty} z^n p_n - \mu z \sum_{n=1}^{\infty} n z^{n-1} p_n, \quad (2.62)$$

which can be rewritten as

$$\frac{\partial G}{\partial t} = \lambda z G + \mu \frac{\partial G}{\partial z} - \lambda G - \mu z \frac{\partial G}{\partial z}. \quad (2.63)$$

Rearranging (2.63), we arrive at

$$\frac{\partial G}{\partial t} = (z-1)\left(\lambda G - \mu \frac{\partial G}{\partial z}\right), \quad (2.64)$$

which is a first order partial differential equation (PDE) and can be solved using the method of characteristics. However, this is out of our interest; we focus on solving it at steady state, i.e., $\partial G/\partial t = 0$.

Importantly, equation (2.63) can formally be obtained from (2.55) by way of the following relations:

$$\mathbb{E}^{\pm 1} \equiv z^{\mp 1}, \quad n \equiv z \frac{\partial}{\partial z}, \quad (2.65)$$

where $\mathbb{E}^{\pm 1}$ is the step operator defined by (2.52). By (2.65), we present the implicit correspondence between the step operator $\mathbb{E}$ and variable z in finding the PDE. More clearly, the step operator $\mathbb{E}$ coincides with the reciprocal of z, whereas $\mathbb{E}^{-1}$ increases the power of z by one. Notably, relations (2.65) are useful when we deal with multivariate functions.

At steady state, equation (2.64) turns into an ordinary differential equation (ODE)

$$\frac{\partial G}{\partial z} = \frac{\lambda}{\mu} G, \quad (2.66)$$

which has a solution in the form of

$$G(z) = C \exp\left(\frac{\lambda}{\mu} z\right), \quad (2.67)$$

where C is a constant to be determined using the normalisation condition (2.60). By doing so, we obtain

$$G(z) = \exp\left(\frac{\lambda}{\mu}(z-1)\right). \tag{2.68}$$

Expanding (2.68) in a Taylor series gives

$$G(z) = \exp\left(-\frac{\lambda}{\mu}\right)\sum_{n=0}^{\infty}\frac{(\lambda/\mu)^n}{n!}z^n. \tag{2.69}$$

Comparing (2.69) with (2.56), we recover the steady-state probability distribution as

$$p_n = \exp\left(-\frac{\lambda}{\mu}\right)\frac{(\lambda/\mu)^n}{n!}, \tag{2.70}$$

which is a Poisson distribution. We here note that (2.70) is consistent with (2.51), which obtained from the set of recursive equations (2.48).

2.6 Stochastic simulation

Although the CME is useful to provide information about the process of interest, its explicit solution can be stringent to obtain except for some simple cases. Therefore, one needs efficient numerical methods for simulating the underlying stochastic process whose description is given by the CME. The stochastic simulation algorithm (SSA) is one of those methods that is used to generate time trajectories of the species in a chemical reaction system [4, 3]. Numerous variants of the SSA have been proposed in the literature [1, 16]. Some of those include the direct method, the first reaction method, and the next reaction method [17]. However, for systems with a large number of species, Gillespie's SSA becomes computationally expensive. Below, we shall introduce the SSA for the degradation reaction (2.10) and the two-stage model.

The SSA for the degradation reaction given by (2.10) can be written as follows [9]. Given that an initial number of species X at time $t = 0$, we generate a random number r uniformly distributed in $(0, 1)$. The first task is to determine the time of next reaction using equation (2.71). Note that $\ln\left(\frac{1}{r}\right)$ is exponentially distributed with unit mean; multiplying this with the inverse of the total reaction rate (here $kX(t)$) sets the timescale of reaction occurrence. Then the number of X species is decreased by 1 at time $t + \tau$. This procedure repeats itself until there is no species X left. Two

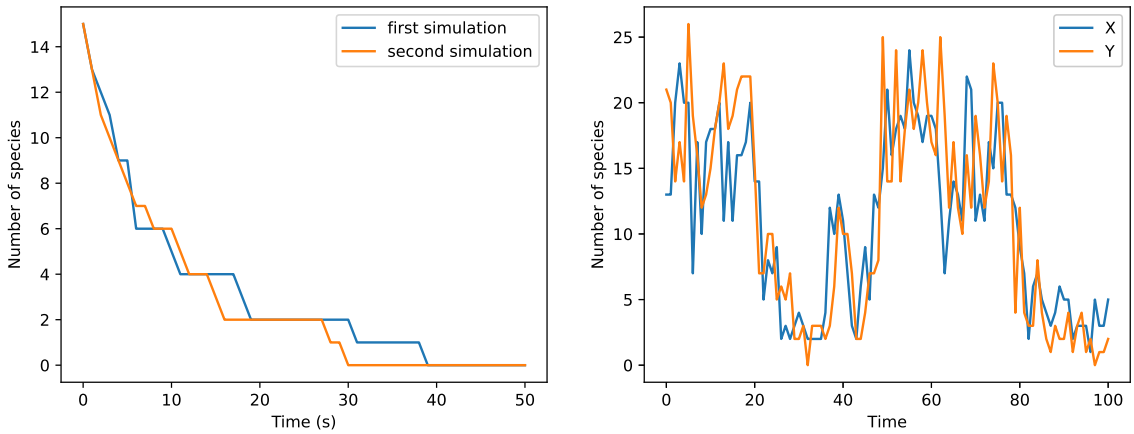


Figure 2.4: Sample time trajectories for the degradation and the two-stage model. *Left:* Stochastic simulation of the degradation reaction (2.10) plotted for the decay rate $k = 0.1$, and the initial number of species is set to 15, i.e., $X(0) = 15$. *Right:* A realisation of two-stage process defined by (2.72) with the rate parameters $k_2 = k_3 = 2$. The parameter values for (2.73) are $H = 4$, $a_0 = 0.3$, $a_1 = 1.6$ and the system size parameter is $\Omega = 20$.

realisations of SSA for the degradation reaction (2.10) is given in the left panel of Figure 2.4.

Algorithm 1: SSA for the decay process

1 Start with $X(0) = n_0$ and $t = 0$.

2 Generate a random number r uniformly distributed in (0, 1).

3 Compute the next reaction time, $t + \tau$, where

$$\tau = \frac{1}{X(t)k} \ln \left[\frac{1}{r} \right]. \tag{2.71}$$

4 Compute $X(t + \tau) = X(t) - 1$.

Then proceed to Step 2 for time $t + \tau$.

Now, let us present a more complicated model of a chemical reaction system, including production and degradation reactions. The model is known as the two-stage process and is given by the following set of chemical reactions

$$\begin{aligned} \emptyset &\xrightarrow{k_1} \mathrm{X}, \\ \mathrm{X} &\xrightarrow{k_2} \mathrm{Y}, \\ \mathrm{Y} &\xrightarrow{k_3} \emptyset, \end{aligned} \tag{2.72}$$

occurring at reaction rates k_1, k_2, and k_3, respectively. The epithet two-stage refers the two stages of a molecule's lifetime, first denoted by X and then by Y, and should not be confused with the stages of transcription and translation of a gene expression model (Section 2.8). The current two-stage model will be generalised and explored in Chapter 3. In (2.72), the first reaction states that the probability of production of species X at a rate k_1 in the time interval $[t, t+\mathrm{d}t)$ is equal to $k_1\mathrm{d}t$. The second reaction states that the species X becomes Y with probability $k_2 X\mathrm{d}t$ during the time interval $(t, t+\mathrm{d}t)$. The last reaction stands for the degradation of Y with a rate constant k_3 (cf. (2.10)). The reaction rates $k_2 = 2$ and $k_3 = 2$ are chosen so that production and degradation rates vary with the number of species as $2X$ and $2Y$, respectively. The reaction rate k_1 is defined by a sigmoid function

$$k_1(X, Y) = \Omega\left(a_0 + \frac{a_1(X+Y)^H}{\Omega^H + (X+Y)^H}\right), \tag{2.73}$$

where H, a_0, and a_1 are the parameter values, and Ω is the system-size parameter.

The deterministic description of the chemical reactions (2.72) is given in the form of system of ODEs

$$\frac{\mathrm{d}X}{\mathrm{d}t} = k_1(X, Y) - k_2 X, \tag{2.74}$$

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = k_2 X - k_3 Y, \tag{2.75}$$

subject to initial conditions

$$X(0) = X_0,$$
$$Y(0) = Y_0,$$

where $X = X(t)$ and $Y = Y(t)$ denote the number of X and Y species at time t, respectively. The SSA procedure corresponding to the system of chemical reactions (2.72) can be given as follows: for a given initial number of species and time, two random numbers are drawn from a uniform distribution, and the propensity function of each reaction is calculated. Then the time when the next reaction occurs is computed via (2.76). Subsequently, a distinct random number r_2 is generated to determine which reaction in the system will occur next (see (2.77)–(2.78)). The fraction α_i/α gives the probability that the i-th reaction occurs. When i-th reaction fires, the number of species is updated accordingly. Here we

Algorithm 2: SSA for the two-stage model

1 Start with $X(0) = n_0$ and $Y(0) = m_0$ at time $t = 0$.

2 Generate two random numbers r_1 and r_2 uniformly distributed in (0, 1).

3 Compute the propensity function of each reaction by

$$\alpha_1 = k_1,$$

$$\alpha_2 = X(t)k_2,$$

$$\alpha_3 = Y(t)k_3.$$

Then, compute $\alpha = \alpha_1 + \alpha_2 + \alpha_3$.

4 Compute the next reaction time, $t + \tau$, where

$$\tau = \frac{1}{\alpha}\ln\left[\frac{1}{r_1}\right]. \tag{2.76}$$

5 Compute the number of molecules at time $t + \tau$ as

$$X(t+\tau) = \begin{cases} X(t) - 1 & \text{if } 0 \le r_2 < \alpha_2/\alpha; \\ X(t) + 1 & \text{if } \alpha_2/\alpha \le r_2 < (\alpha_1 + \alpha_2)/\alpha; \\ X(t) & \text{if } (\alpha_1 + \alpha_2)/\alpha \le r_2 < 1; \end{cases} \tag{2.77}$$

and

$$Y(t+\tau) = \begin{cases} Y(t) + 1 & \text{if } 0 \le r_2 < \alpha_2/\alpha; \\ Y(t) & \text{if } \alpha_2/\alpha \le r_2 < (\alpha_1 + \alpha_2)/\alpha; \\ Y(t) - 1 & \text{if } (\alpha_1 + \alpha_2)/\alpha \le r_2 < 1; \end{cases} \tag{2.78}$$

Then proceed to Step 2 for time $t + \tau$.

simulated the system of chemical reactions (2.72) by taking the initial number of species as $X(0) = 15$ and $Y(0) = 15$. A plot of this simulation is shown in the right panel of Figure 2.4. The Python package `Gillespy2` was used to generate time trajectories [18].

2.7 General reaction kinetics

As illustrated in the previous sections, biochemical systems, including gene expression, can be modelled as chemical reaction networks that contain a variety of *chemical species*. In such a network, the *copy number* of a species denotes the amount of its population. The chemical reactions represent how the species in a system interact. Each species has a *stoichiometric coefficient* denoting the amount of each constituent that will be degraded or produced in a reaction system. It is typically a negative number for reactants and is positive for products. Below, we introduce the fundamental concepts of general reaction kinetics.

Let us consider a set of M chemical reactions involving N species whose copy numbers at time t are stored in the state vector $\mathbf{X}(t) = \begin{bmatrix} X_1(t) & X_2(t) & \dots & X_N(t) \end{bmatrix}^\top$. When a reaction fires, the copy numbers are updated as per their stoichiometric coefficients, leading to the net state change in its stoichiometric vector. Once reaction j occurs, the new state vector is updated as per

$$\mathbf{X}(t) = \mathbf{X}(t^{\mathrm{p}}) + \mathbf{s}_j,$$

where t^{p} denotes the preceding time for the reaction event. The vectors $\mathbf{s}_j$, for $j = 1, \dots, M$, are obtained via $\mathbf{s}_j = \mathbf{s}_j^{\mathrm{p}} - \mathbf{s}_j^{\mathrm{r}}$, where $\mathbf{s}_j^{\mathrm{r}}$ and $\mathbf{s}_j^{\mathrm{p}}$ are the vectors of the reactant and product stoichiometric coefficients for reaction j, respectively. Consequently, the stoichiometric matrix $\mathbf{S}$ that consists of integer stoichiometric coefficients can be formed by stacking up the stoichiometric vectors column-wise. Thus, the resulting matrix is of the size $N \times M$, whose columns correspond to the reactions, and rows correspond to the compounds. The reactions in a well-mixed system follow the *law of mass action*, which states that the probability that i-th reaction occurs in the time interval $[t, t + \mathrm{d}t)$ is proportional to the product of infinitesimally small time step $\mathrm{d}t$ and a function of reactant copy numbers. This is

typically given by the *propensity function* $f_i(\mathbf{X})$ defined as [19]

$$f_i(\mathbf{X}(t)) = k \times \text{total combinations in } \mathbf{X}(t) \text{ for each } i,$$

where k is a non-negative constant called as the *kinetic rate parameter*.

Let us consider the two-stage model (2.72), where we have species X and Y, evolving as per the reaction scheme

$$\emptyset \xrightarrow{k_1} \mathrm{X}, \quad \mathrm{X} \xrightarrow{k_2} \mathrm{Y}, \quad \mathrm{Y} \xrightarrow{k_3} \emptyset, \tag{2.79}$$

where k_1, k_2 and k_3 are the kinetic rate parameters (or reaction rate constants). In (2.79), species X is produced from a *pool*, X becomes Y, and then Y is degraded, respectively. Let $\tilde{\mathbf{X}}(t) = \begin{bmatrix} X(t), & Y(t) \end{bmatrix}^\top$ be the state vector. The propensity functions read

$$f_1(\tilde{\mathbf{X}}(t)) = k_1, \quad f_2(\tilde{\mathbf{X}}(t)) = k_2 X(t) \quad \text{and} \quad f_3(\tilde{\mathbf{X}}(t)) = k_3 Y(t). \tag{2.80}$$

Here we note that the propensity function $f_1(\tilde{\mathbf{X}}(t)) = k_1$ is specifically defined by (2.73) for the two-stage model. The stoichiometric vectors are given by

$$\mathbf{s}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{s}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \mathbf{s}_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \tag{2.81}$$

which form the stoichiometric matrix

$$\mathbf{S} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}.$$

The probability $P(\mathbf{x}, t)$ of having $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}$ molecules at time t satisfies the (general form of the) CME

$$\frac{\mathrm{d}P(\mathbf{x}, t)}{\mathrm{d}t} = \sum_{j=1}^{M} f_j(\mathbf{x} - \mathbf{s}_j) P(\mathbf{x} - \mathbf{s}_j, t) - P(\mathbf{x}, t) \sum_{j=1}^{M} f_j(\mathbf{x}), \tag{2.82}$$

where $\mathbf{s}_j$ is the j-th column of the stoichiometric matrix $\mathbf{S}$ [20, 21]. Note that $P(\mathbf{x}, t)$ is a function of N integers, i.e. $x_1, x_2, \dots, x_N$, and of a continuos variable t. Therefore, we write ordinary derivatives instead of partial derivatives. Equation (2.82) can be rewritten as

$$\frac{\mathrm{d}P(\mathbf{x}, t)}{\mathrm{d}t} = \sum_{j=1}^{M} \left(\prod_{i=1}^{N} \mathbb{E}_i^{-S_{ij}} - 1 \right) f_j(\mathbf{x}) P(\mathbf{x}, t), \tag{2.83}$$

where the step operators $\mathbb{E}_i^{-S_{ij}}$ [10, 14] are defined on a general function g by

$$\mathbb{E}_i^{-S_{ij}} g(x_1, \ldots, x_i, \ldots, x_N) = g(x_1, \ldots, x_i - S_{ij}, \ldots, x_N).$$

For the specific model (2.79), inserting (2.80) and (2.81) into (2.82) gives the CME

$$\begin{aligned} \frac{\mathrm{d}P(x,y,t)}{\mathrm{d}t} =& k_1 P(x-1,y,t) + k_2(x+1)P(x+1,y-1,t) \\ &+ k_3(y+1)P(x,y+1,t) - (k_1 + k_2 x + k_3 y)P(x,y,t). \end{aligned} \tag{2.84}$$

Denoting by $\mathbb{E}_1$ and $\mathbb{E}_2$ the step operators in the variables x and y, the CME (2.84) can then be recast as

$$\frac{\mathrm{d}P(x,y,t)}{\mathrm{d}t} = k_1(\mathbb{E}_1^{-1} - 1)P + k_2(\mathbb{E}_1\mathbb{E}_2^{-1} - 1)xP + k_3(\mathbb{E}_2 - 1)yP. \tag{2.85}$$

In the next section, we will briefly introduce the gene expression process and its corresponding mathematical model.

2.8 Gene expression

Gene expression is a process by which the information in DNA is transformed via cellular machinery into functional gene products such as mRNA and protein [2]. The process comprises two key steps, transcription and translation, whose control plays an essential role in producing a protein molecule. Transcription is conducted by an RNA polymerase enzyme, leading to the creation of an RNA transcript called messenger RNA (mRNA) from the DNA in a gene. After the mRNA has copied the transcribed *information* from the DNA, translation, the second major step, in which the mRNA is *read* to form a sequence of amino acids during the protein-making process, takes place. In addition, the levels of gene products, i.e. mRNA and protein, determine the fate of cells and give rise to variation in the cell population. Therefore, quantifying the amount of these molecules is a question of intense research. Although gene expression is a far more complex process, we here make a simplified description of the model as elementary reactions. We neglect processes such as the binding of ribosomes and RNA polymerase; we focus on the simplest case in which the model incorporates transcription, translation, and decay reactions.

The two-stage model The dynamics of the two-stage gene-expression model is given by the reaction scheme [22]

$$\begin{aligned} &\emptyset \xrightarrow{\lambda_1} \text{mRNA}, \quad \text{mRNA} \xrightarrow{\gamma_1} \emptyset, \\ &\text{mRNA} \xrightarrow{\lambda_2} \text{mRNA} + \text{protein}, \quad \text{protein} \xrightarrow{\gamma_2} \emptyset, \end{aligned} \tag{2.86}$$

where λ_1 is the mRNA production rate, λ_2 is the protein translation rate, and γ_1 and γ_2 are the decay rate constants of mRNA and protein species, respectively. Here and below, 1 and 2 in the subscript indicate the mRNA and protein species, respectively.

The deterministic description of the reaction system (2.86) is given by

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \lambda_1 - \gamma_1 M, \quad \frac{\mathrm{d}N}{\mathrm{d}t} = \lambda_2 M - \gamma_2 N, \tag{2.87}$$

where M and N denote the levels of mRNA and protein, respectively. Given that an initial condition, an explicit solution to the system (2.87) can be obtained [22].

In the stochastic context, the probability $P(m, n, t)$ of observing m mRNA and n protein molecules at time t satisfies the CME

$$\begin{aligned} \frac{\mathrm{d}P(m,n,t)}{\mathrm{d}t} =& \lambda_1(P(m-1,n,t) - P(m,n,t)) + \gamma_1((m+1)P(m+1,n,t) \\ &- mP(m,n,t)) + \lambda_2 m(P(m,n-1,t) - P(m,n,t)) \\ &+ \gamma_2((n+1)P(m,n+1,t) - nP(m,n,t)), \end{aligned} \tag{2.88}$$

subject to initial condition

$$p(m,n,0) = \delta_{m,m_0}\delta_{n,n_0}, \tag{2.89}$$

where $\delta_{i,j}$ represents the Kronecker delta symbol, which is one if $i = j$ and zero otherwise; m_0 and n_0 are the initial mRNA and protein amounts, respectively. Note that equation (2.88) can be rewritten in a more compact form as

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \lambda_1(\mathbb{E}_1^{-1} - 1)P + \gamma_1(\mathbb{E}_1 - 1)mP + \lambda_2 m(\mathbb{E}_2^{-1} - 1)P + \gamma_2(\mathbb{E}_2 - 1)nP, \tag{2.90}$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the step operators in the variables m and n, respectively.

In what follows, we are interested in finding a solution to the CME (2.88). To that end, let us define by

$$G(x,y,t) = \sum_m \sum_n x^m y^n P(m,n,t)$$

the generating function. Then the CME (2.88) can be transformed into the PDE for the generating function of the stationary distribution, which is given by (see [22] for

details)

$$\frac{\partial G}{\partial t} = \lambda_1(1-x)G + (\gamma_1(x-1) + \lambda_2 x(1-y))\frac{\partial G}{\partial x} + \gamma_2(y-1)\frac{\partial G}{\partial y}. \tag{2.91}$$

Note that (2.91) can formally be obtained from (2.90) by the following transformation rules: $P \equiv G$; $\mathbb{E}_1^{\pm 1} \equiv x^{\mp 1}$; $\mathbb{E}_2^{\pm 1} \equiv y^{\mp 1}$; $m \equiv x\frac{\partial}{\partial x}$; $n \equiv y\frac{\partial}{\partial y}$ (cf. Eq. (2.65)).

At steady state, equation (2.91) reads

$$(\gamma_1(x-1) + \lambda_2 x(1-y))\frac{\partial G}{\partial x} + \gamma_2(y-1)\frac{\partial G}{\partial y} = \lambda_1(x-1)G. \tag{2.92}$$

An analytical solution to (2.92) has been obtained in [22] as

$$G(x,y) = \exp\left(\alpha\beta\int_1^y M\left(1, 1+\lambda, \beta(s-1)\right)\mathrm{d}s + \alpha(x-1)M\left(1, 1+\lambda, \beta(y-1)\right)\right), \tag{2.93}$$

where

$$M(a,b,z) = \sum_{n=0}^{\infty}\frac{(a)_n z^n}{(b)_n n!}$$

is Kummer's function [23], $(x)_n = x(x+1)(x+2)\ldots(x+n-1)$ which for $(x)_0 = 1$ is the rising factorial (or Pochhammer symbol), and $\alpha = \lambda_1/\gamma_1$, $\beta = \lambda_2/\gamma_2$, and $\lambda = \gamma_1/\gamma_2$.

The results and fundamental concepts given in this chapter will be used in the following chapters. In particular, we shall focus on a gene expression model consisting of complex lifetime pathways in the subsequent chapter.

Chapter 3

Multiclass–multistage model

In this chapter, we consider a stochastic gene-expression model for a self-regulating transcription factor whose lifespan (or time till degradation) follows a general distribution modelled as per a multi-dimensional phase-type process. We show that at steady state the protein copy-number distribution is the same as in a one-dimensional model with exponentially distributed lifetimes. This invariance result holds only if molecules are produced one at a time: we provide explicit counterexamples in the bursty production regime. Additionally, we consider the case of a bistable genetic switch constituted by a positively autoregulating transcription factor. The switch alternately resides in states of up- and downregulation and generates bimodal protein distributions. In the context of our invariance result, we investigate how the choice of lifetime distribution affects the rates of metastable transitions between the two modes of the distribution. The phase-type model, being non-linear and multi-dimensional whilst possessing an explicit stationary distribution, provides a valuable test example for exploring dynamics in complex biological systems.

The content of this chapter has been published in *Lecture Notes in Computer Science*, vol 12314, pp. 27–43. Springer, 2020 [24].

3.1 Introduction

Biochemical processes at the single-cell level involve molecules such as transcription factors that are present at low copy numbers [25, 26]. The dynamics of these processes is well described by stochastic Markov processes in continuous time with discrete state space [27, 28, 29]. While few-component or linear-kinetics systems [30] allow for exact analysis, in more complex system one often uses

approximative methods [31], such as moment closure [32], linear-noise approximation [33, 34], hybrid formulations [35, 36, 37], and multi-scale techniques [38, 39].

In simplest Markovian formulations, the lifetime of a regulatory molecule is memoryless, i.e. exponentially distributed [22, 40]. However, non-exponential decay patterns have been observed experimentally for both mRNA transcripts and proteins [7, 8]. Therefore, in this chapter we shall consider lifetime distributions that can assume far more complex forms than the simple exponential. Previous studies of gene-expression models with delayed degradation also provide examples of non-exponential lifetime distributions [41, 42].

In Section 3.2, we formulate, both in the deterministic and stochastic settings, a one-dimensional model for the abundance of a transcription factor with a memoryless lifetime. Since many transcription factors regulate their own gene expression [43], we allow the production rate to vary with the copy number. We show that the deterministic solutions tend to the fixed points of the feedback response function; in the stochastic framework, we provide the stationary distribution of the protein copy number.

In Section 3.3, we characterise the steady-state behaviour of a structured model that accounts for complex lifetime pathways. The model is multidimensional, each dimension corresponding to a different class and stage of a molecule's lifetime; the chosen structure accounts for a wide class of phase-type lifetime distributions [44, 45]. We demonstrate that the deterministic fixed points and the stochastic stationary distribution that were found for the one-dimensional framework remain valid for the total protein amount in the multi-dimensional setting.

We emphasise that the distribution invariance result rests on the assumption of non-bursty production of protein. The case of bursty production is briefly discussed in Section 3.4, where explicit counter-examples are constructed by means of referring to explicit mean and variance formulae available from literature for systems without feedback [46, 47].

In the final Section 3.5, we approximate the stochastic protein distribution by a mixture of Gaussians with means at deterministic fixed points and variances given by the linear-noise approximation [10, 48]. Additionally, we study the rates of

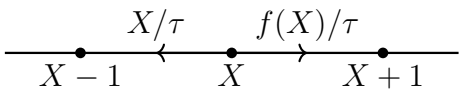


Figure 3.1: A diagram of the one-dimensional model. The number of molecules X can decrease by one or increase by one. The stochastic rates (or propensities) of these transitions are indicated above the transition edges.

metastable transitions [49, 50] between the Gaussian modes in the one-dimensional and structured settings.

3.2 One-dimensional model

Deterministic framework. The dynamics of the abundance of protein X at time t can be modelled deterministically by an ordinary differential equation

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \tau^{-1}\left(f(X) - X\right), \tag{3.1}$$

which states that the rate of change in X is equal to the difference of production and decay rates. The decay rate is proportional to X; the factor of proportionality is the reciprocal of the expected lifetime τ. The rate of production per unit protein lifetime is denoted by $f(X)$ in (3.1); the dependence of the production rate on the protein amount X implements the feedback in the model. Equating the right-hand side of (3.1) to zero yields

$$f(X) = X, \tag{3.2}$$

meaning that steady states of (3.1) are given by the fixed point of the production response function $f(X)$.

Stochastic framework. The stochastic counterpart of (3.1) is the Markov process with discrete states $X \in \mathbb{N}_0$ in continuous time with transitions $X \to X-1$ or $X \to X+1$, occurring with rates X/τ and $f(X)/\tau$ respectively (see the schematic in Figure 3.1). Note that in case of a constant production rate, i.e. $f(X) \equiv \lambda$, the model turns into the immigration-and-death process [51]; in queueing theory this is also known as $M/M/\infty$ queue [12], (see also Section 2.4). The stationary distribution of the immigration–death process is known to be Poissonian with mean equal to λ [51].

For a system with feedback, the probability $P(X,t)$ of having X molecules at time t satisfies the master equation

$$\frac{\mathrm{d}P(X,t)}{\mathrm{d}t} = \tau^{-1}\left(\mathbb{E}^{-1}-1\right) f(X)P(X,t) + \tau^{-1}\left(\mathbb{E}-1\right) XP(X,t), \tag{3.3}$$

in which $\mathbb{E}$ is the van-Kampen step operator [10]. Inserting $P(X,t)=\pi(X)$ into (3.3) and solving the resulting difference equation, one finds a steady-state distribution in the explicit form

$$\pi(X) = \pi(0)\frac{\prod_{k=0}^{X-1} f(k)}{X!}. \tag{3.4}$$

The probability $\pi(0)$ of having zero molecules plays the role of the normalisation constant in (3.4), which can be uniquely determined by imposing the normalisation condition $\pi(0)+\pi(1)+\ldots=1$. Note that inserting $f(X)\equiv\lambda$ into (3.4) results in the aforementioned Poissonian distribution with $\pi(0)=\mathrm{e}^{-\lambda}$.

3.3 Multiclass–multistage model

In this section, we introduce a structured multiclass–multistage model which is an extension of one-dimensional model introduced in the previous section. The fundamentals of the multidimensional model are as shown in Figure 3.2. A newly produced molecule is assigned into one of K distinct classes. Which class is selected is chosen randomly according to a discrete distribution $p_1,\ldots,p_K$. The lifetime of a molecule in the i-th class consists of S_i stages. The holding time in any of these stages is memoryless (exponential), and parametrised by its mean τ_{ij}, where i indicates which class and j indicates which stage. Note that

$$\tau = \sum_{i=1}^{K}\sum_{j=1}^{S_i} p_i\tau_{ij} \tag{3.5}$$

gives the expected lifetime of a newly produced molecule. After the last (S_i-th) stage, the molecule is degraded. The total distribution of a molecule lifetime is a mixture, with weights p_i, of the lifetime distributions of the individual classes, each of which is a convolution of exponential distributions of the durations of the individual stages; such distributions are referred to as phase-type distribution and provide a wide family of distribution to approximate practically any distribution of a positive random variable [44].

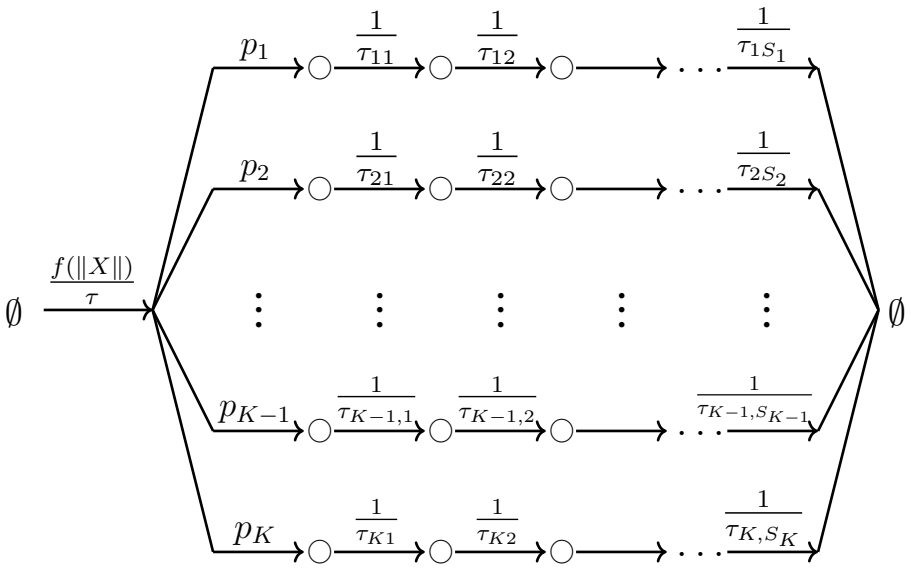


Figure 3.2: A schematic representation of multiclass–multistage model. A newly produced molecule is randomly assigned, according to a prescribed distribution $p_1, \ldots, p_K$, into one of K distinct classes. The lifetime of a molecule in the i-th class consists of S_i consecutive memoryless stages, and ends in the degradation of the molecule. The expected holding time in the j-th stage of the i-th class is τ_{ij}. The production rate is a function of the total number $\|X\|$ of molecules across all stages and classes.

We denote by X_{ij} the number of molecules in the i-th class and the j-th stage of their lifetime, by

$$X = (X_{11}, \ldots, X_{1S_1}, X_{21}, \ldots, X_{2S_2}, \ldots, X_{K1}, \ldots, X_{KS_K})$$

the $\sum_{i=1}^{K} S_i$-dimensional copy-number vector, and by

$$\|X\| = \sum_{i=1}^{K} \sum_{j=1}^{S_i} X_{ij} \tag{3.6}$$

the total number of molecules across all classes and stages.

Deterministic framework. The deterministic description of the structured model is given by a system of coupled ordinary differential equations

$$\frac{\mathrm{d}X_{i1}}{\mathrm{d}t} = \frac{p_i f\left(\|X\|\right)}{\tau} - \frac{X_{i1}}{\tau_{i1}}, \quad i = 1, \ldots, K, \tag{3.7}$$

$$\frac{\mathrm{d}X_{ij}}{\mathrm{d}t} = \frac{X_{ij-1}}{\tau_{ij-1}} - \frac{X_{ij}}{\tau_{ij}}, \quad i = 1, \ldots, K \text{ and } j = 2, \ldots, S_i. \tag{3.8}$$

The right-hand sides of (3.7)–(3.8) are each equal to the difference of appropriate arrival and departure rates at/from a particular compartment of the structured model. The departure rates are proportional to the number of molecules in the compartment, with the reciprocal of the holding time giving the factor of proportionality. The arrival rate takes a different form for the first stages (3.7) and for the other stages (3.8). For the first stage, the arrival is obtained by the product of the production rate $f(\|X\|)/\tau$ and the probability p_i of selecting the i-th class. For the latter stages, the arrival rate is equal to the departure rate of the previous stage.

Equating (3.7)–(3.8) to zero, we find that

$$\frac{p_i f\left(\|X\|\right)}{\tau} = \frac{X_{i1}}{\tau_{i1}} = \frac{X_{i2}}{\tau_{i2}} = \ldots = \frac{X_{ij}}{\tau_{ij}} \tag{3.9}$$

hold at steady state, from which it follows that

$$X_{ij} = \frac{p_i \tau_{ij} f\left(\|X\|\right)}{\tau}. \tag{3.10}$$

Summing (3.10) over $i = 1, \ldots, K$ and $j = 2, \ldots, S_i$, and using (3.5) and (3.6), yield

$$\|X\| = f\left(\|X\|\right) \tag{3.11}$$

for the total protein amount (3.6). Thus, the protein amount at steady state is obtained, like in the one-dimensional model, by calculating the fixed points of the feedback response function.

Combining (3.11) and (3.9) we find

$$X_{ij} = \frac{p_i \tau_{ij} \|X\|}{\tau}, \tag{3.12}$$

which means that at steady state the total protein amount is distributed among the compartments proportionally to the product of class assignment probability and the mean holding time of the particular compartment.

Stochastic framework. Having demonstrated that the stationary behaviour of the one-dimensional and the structured multi-dimensional models is the same in the deterministic framework, we next aim to demonstrate that the same is also true in the stochastic context. Prior to turning our attention to the feedback system, it is again instructive to discuss the case without regulation, i.e. $f(\|X\|) \equiv \lambda$; the new molecule arrivals are then exponentially distributed. In the language of queueing theory, the process can be reinterpreted as the $M/G/\infty$ queue with exponential arrivals of customers, a general phase-type distribution of service times, and an infinite number of servers. It is well known that the steady-state distribution of an $M/G/\infty$ queue is Poisson with mean equal to λ [52]. Thus, without feedback, we obtain the very same Poisson(λ) distribution that applies in the one-dimensional case.

In the feedback case, the probability $P(X,t)$ of having $X = (X_{11}, \ldots, X_{K,S_K})$ copy numbers in the individual compartments at any time t satisfies the master equation

$$\begin{aligned}
\frac{\mathrm{d}P(X,t)}{\mathrm{d}t} =& \tau^{-1} \sum_{i=1}^{K} p_i \left(\mathbb{E}_{i1}^{-1} - 1\right) f(\|X\|) P(X,t) && (3.13)\\
&+ \sum_{i=1}^{K} \sum_{j=1}^{S_i - 1} \tau_{ij}^{-1} \left(\mathbb{E}_{ij}\mathbb{E}_{ij+1}^{-1} - 1\right) X_{ij} P(X,t) && (3.14)\\
&+ \sum_{i=1}^{K} \tau_{iS_i}^{-1} \left(\mathbb{E}_{iS_i} - 1\right) X_{iS_i} P(X,t). && (3.15)
\end{aligned}$$

The right-hand-side terms (3.13), (3.14), and (3.15) stand for the change in probability mass function due to the production, moving to next stage, and decay reactions, respectively. Note that $\mathbb{E}_{ij}$ is a step operator which increases the copy

number of molecules in the i-th class at the j-th stage by one [10]. Likewise, $\mathbb{E}_{ij}^{-1}$ decreases the same copy number by one. Rearrangement of terms in the master equation yields

$$\begin{aligned}\frac{\mathrm{d}P(X,t)}{\mathrm{d}t} =& \sum_{i=1}^{K}\left(\tau^{-1}p_i\mathbb{E}_{i1}^{-1}f(\|X\|)P(X,t) - \tau_{i1}^{-1}X_{i1}P(X,t)\right)\\ &+\sum_{i=1}^{K}\sum_{j=1}^{S_i-1}\left(\tau_{ij}^{-1}\mathbb{E}_{ij}\mathbb{E}_{ij+1}^{-1}X_{ij}P(X,t) - \tau_{ij+1}^{-1}X_{ij+1}P(X,t)\right)\\ &+\sum_{i=1}^{K}\tau_{iS_i}^{-1}\mathbb{E}_{iS_i}X_{iS_i}P(X,t) \;-\; \tau^{-1}f(\|X\|)P(X,t).\end{aligned}$$

Equating the derivative to zero, we derive for the stationary distribution $\pi(X)$ an algebraic system

$$\begin{aligned}0 =& \sum_{i=1}^{K}\left(\tau^{-1}p_i\mathbb{E}_{i1}^{-1}f(\|X\|)\pi(X) - \tau_{i1}^{-1}X_{i1}\pi(X)\right)\\ &+\sum_{i=1}^{K}\sum_{j=1}^{S_i-1}\left(\tau_{ij}^{-1}\mathbb{E}_{ij}\mathbb{E}_{ij+1}^{-1}X_{ij}\pi(X) - \tau_{ij+1}^{-1}X_{ij+1}\pi(X)\right)\\ &+\sum_{i=1}^{K}\tau_{iS_i}^{-1}\mathbb{E}_{iS_i}X_{iS_i}\pi(X) \;-\; \tau^{-1}f(\|X\|)\pi(X).\end{aligned} \tag{3.16}$$

Clearly, it is sufficient that

$$\begin{aligned}\tau^{-1}p_i\mathbb{E}_{i1}^{-1}f(\|X\|)\pi(X) &= \tau_{i1}^{-1}X_{i1}\pi(X),\\ \tau_{ij}^{-1}\mathbb{E}_{ij}\mathbb{E}_{ij+1}^{-1}X_{ij}\pi(X) &= \tau_{ij+1}^{-1}X_{ij+1}\pi(X),\\ \sum_{i=1}^{K}\tau_{iS_i}^{-1}\mathbb{E}_{iS_i}X_{iS_i}\pi(X) &= \tau^{-1}f(\|X\|)\pi(X)\end{aligned} \tag{3.17}$$

hold for $\pi(X)$ in order that (3.16) be satisfied. One checks by direct substitution that

$$\pi(X) \propto \prod_{k=0}^{\|X\|-1} f(k) \times \prod_{i=1}^{K}\prod_{j=1}^{S_i}\frac{(p_i\tau_{ij}/\tau)^{X_{ij}}}{X_{ij}!} \tag{3.18}$$

satisfies equations in (3.17); therefore, (3.18) represents the stationary distribution of the structured model. In order to interpret (3.18), we condition the joint distribution on the total protein copy number, writing

$$\pi(X) = \pi_{\mathrm{cond}}(X \mid \|X\|)\pi_{\mathrm{tot}}(\|X\|), \tag{3.19}$$

in which the conditional distribution is recognised as the multinomial [15]

$$\pi_{\mathrm{cond}}(X \mid \|X\|) = \binom{\|X\|}{X}\prod_{i=1}^{K}\prod_{j=1}^{S_i}(p_i\tau_{ij}/\tau)^{X_{ij}}, \tag{3.20}$$

and the total copy number distribution is given by

$$\pi_{\text{tot}}(\|X\|) = \pi_{\text{tot}}(0)\frac{\prod_{k=0}^{\|X\|-1} f(k)}{\|X\|!}. \tag{3.21}$$

By (3.20), the conditional means of X_{ij} coincide with the deterministic partitioning of the total copy number (3.12). Importantly, comparing (3.21) to (3.4), we conclude that the one-dimensional and multi-dimensional models generate the same (total) copy number distributions.

3.4 Bursting

The independence of stationary distribution on the lifetime distribution relies on the assumption of non-bursty production of protein that has implicitly been made in our model. In this section, we allow for the synthesis of protein in bursts of multiple molecules at a single time [53, 54]. Referring to previously published results [46, 47], we provide a counterexample that demonstrates that in the bursty case different protein lifetime distributions can lead to different stationary copy-number distributions. The counterexample can be found even in the absence of feedback.

Bursty production means that the number of molecules can increase within an infinitesimally small time interval of length $\mathrm{d}t$ from X to $X + j$, where $j \geq 1$, with probability $\lambda\tau^{-1}b_j\mathrm{d}t$, in which λ is the burst frequency (a constant in the absence of feedback), τ is the mean protein lifetime, and $b_j = \text{Prob}[B = j]$ is the probability mass function of the burst size B. Protein molecules degrade independently of one another. The distribution of their lifetime T can in general be described by the survival function $G(t) = \text{Prob}[T > t]$; the mean lifetime thereby satisfies

$$\tau = -\int_0^\infty tG'(t)\mathrm{d}t = \int_0^\infty G(t)\mathrm{d}t. \tag{3.22}$$

The copy protein number X at a given time is given by the number of products that have been produced in a past burst and survived until the given time; this defines a random process, cf. [46], whose steady-state moments are provided below. In queueing theory, bursty increases in the state variable are referred to as batch customer arrivals. Specifically, a bursty gene-expression model without feedback and with general lifetime distribution corresponds to the $M^X/G/\infty$ queue with

memoryless (exponential) batch arrivals, general service distribution, and an infinite number of servers.

Previous analyses [46, 47] show that the steady-state protein mean $\langle X\rangle$ and the Fano factor $F = \mathrm{Var}(X)/\langle X\rangle$ are given by

$$\langle X\rangle = \lambda\langle B\rangle, \quad F = 1 + K_{\mathrm{s}}\left(\frac{\langle B^2\rangle}{\langle B\rangle} - 1\right), \tag{3.23}$$

where

$$K_s = \frac{\int_0^\infty G^2(t)\mathrm{d}t}{\tau} \tag{3.24}$$

is referred to as the senescence factor. Elementary calculation shows that $K_s = 1/2$ if the lifetime distribution is exponential with survival function $G(t) = \mathrm{e}^{-t/\tau}$ and that $K_s = 1$ if the lifetime distribution is deterministic with survival function $G(t) = 1$ for $t < \tau$ and $G(t) = 0$ for $t \geq \tau$. Thus, although two lifetime distributions result in the same value of the stationary mean protein copy number, they give a different value of the noise (the Fano factor); therefore the copy-number distributions are different.

3.5 Metastable transitioning

Transcription factors that self-sustain their gene expression by means of a positive feedback loop can act as a simple genetic switch [55, 56]. A positive-feedback switch can be in two states, one in which the gene is fully activated through its feedback loop, while in the other the gene is expressed at a basal level. The switch serves as a basic memory unit, retaining the information on its initial state on long timescales, and very slowly relaxing towards an equilibrium distribution. It is therefore important to investigate not only the stationary, but also transient distributions, which are generated by a positively autoregulating transcription factor.

Following previous studies [57, 58, 59], we model positive feedback by the Hill function response curve

$$f(X) = \Omega\left(a_0 + \frac{a_1 X^H}{\Omega^H + X^H}\right), \tag{3.25}$$

in which a_0 and a_1 represent the basal and regulable production rates, H is the cooperativity coefficient, and Ω gives the critical amount of protein required for half-stimulation of feedback. Provided that $H > 1$, one can find a_0 and a_1 such that

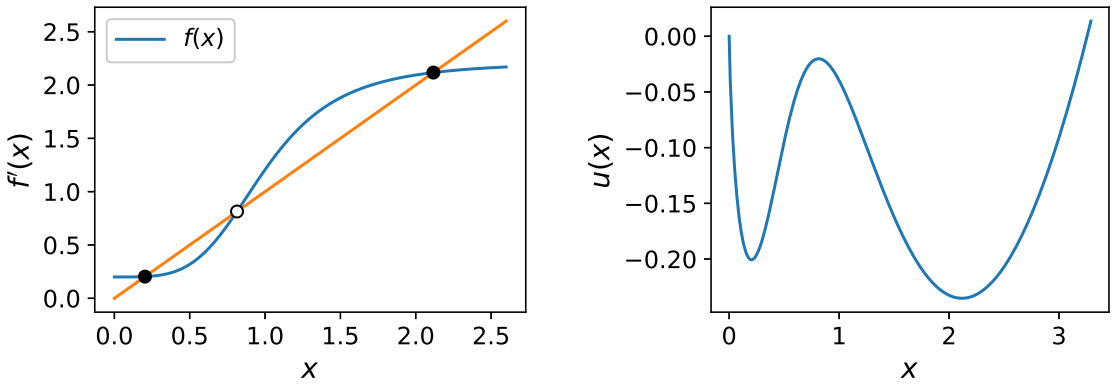


Figure 3.3: *Left:* A sigmoid feedback response function (blue curve) intersects the diagonal (orange line) in multiple fixed points. Ones that are stable to the rate equation (3.1) (full circles) are interspersed by unstable ones (empty circle). *Right:* The potential $u(x)$, defined by (3.33), is a Lyapunov function of the rate equation (3.1). The local minima, or the troughs/wells, of the potential are situated at its stable fixed points; the local maximum, or the barrier, of the potential coincides with the unstable fixed point. *Parameter values for both panels:* We use the Hill-type response (3.25) with $a_0 = 0.3$, $a_1 = 1.6$, $H = 4$, $\Omega = 50$.

(3.25) possesses three distinct fixed points $X_- < X_0 < X_+$, of which the central is unstable and the other two are stable (Figure 3.3, left). The two stable fixed points provide alternative large-time outcomes of the deterministic models (3.1) and (3.7)–(3.8).

Bistability of deterministic models translates into bimodal distributions in the stochastic framework. For large values of Ω, the bimodal protein distribution can be approximated by a mixture of Gaussian modes which are located at the stable fixed points $X_\pm$ (see Figure 3.4), cf. [10, 48],

$$P(X,t) \sim p_-(t)\frac{\mathrm{e}^{-\frac{(X-X_-)^2}{2\sigma_-^2}}}{\sqrt{2\pi}\sigma_-} + p_+(t)\frac{\mathrm{e}^{-\frac{(X-X_+)^2}{2\sigma_+^2}}}{\sqrt{2\pi}\sigma_+}. \tag{3.26}$$

The mixture approximation (3.26) is determined not only by the locations $X_\pm$, but also on the variances $\sigma_\pm^2$ and the weights $p_\pm(t)$ of the two modes (which are given below). The weights in (3.26) are allowed to vary with time in order to account for the slow, metastable transitions that occur between the distribution modes.

The invariance result for stationary distributions derived in the preceding

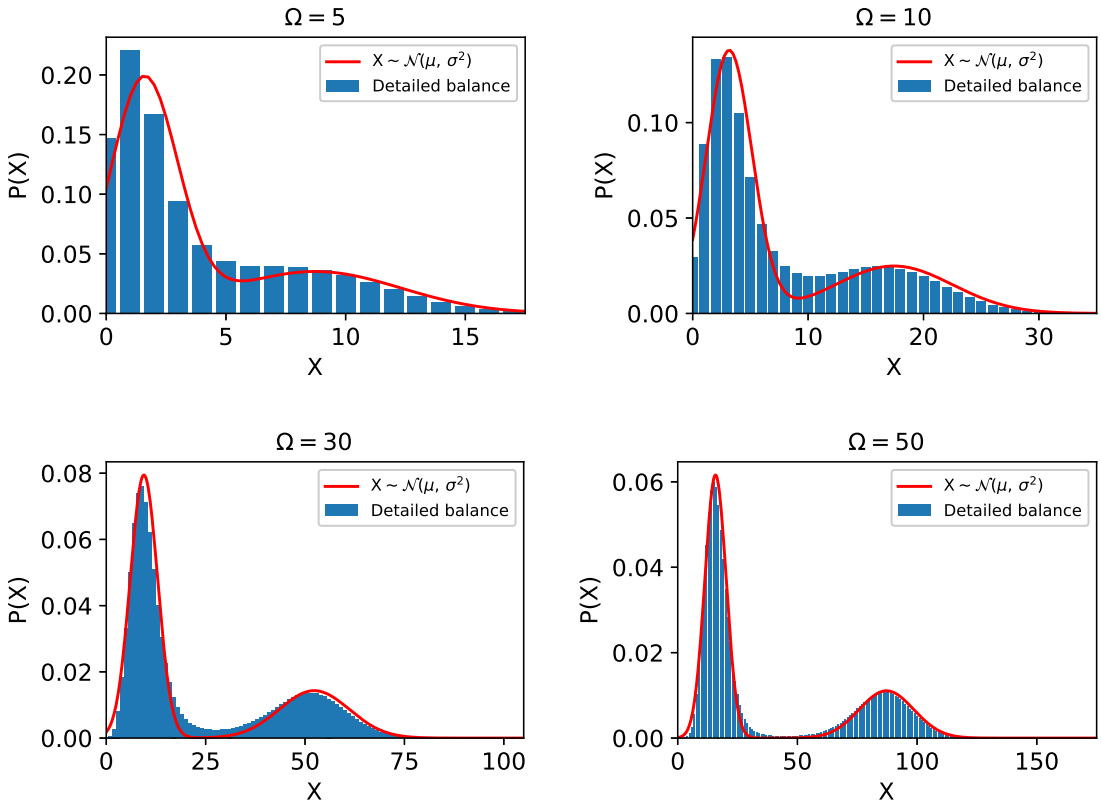


Figure 3.4: Exact stationary protein distribution (3.4) and the Gaussian-mixture approximation (3.26) in varying system-size conditions. The means of the Gaussians are given by the stable fixed points of $f(X)$; the variances are given by linear-noise approximation (3.27). The mixture weights are given by $p_+(\infty) = T_+/(T_+ + T_-)$, $p_-(\infty) = T_-/(T_+ + T_-)$, where the residence times are given by the Arrhenius-type formula (3.32). We use a Hill-type response (3.25) with $a_0 = 0.3$, $a_1 = 1.6$, $H = 4$, and Ω shown in panel captions.

sections implies that, in the limit of $t \to \infty$, the protein distribution (3.26) becomes independent of the choice of the protein lifetime distribution. In particular, the same variances $\sigma_\pm^2$ and the same limit values $p_\pm(\infty)$ of the weights will apply for exponentially distributed and phase-type decay processes. In what follows, we first consult literature to provide results $\sigma_\pm^2$ and $p_\pm(t)$ that apply for the one-dimensional model with exponential decay. Next, we use stochastic simulation to investigate the effect of phase-type lifespan distributions on the relaxation rate of $p_\pm(t)$ to the stationary values.

The variances of the modes are obtained by the linear-noise approximation [60,

61] of the master equation (3.3), which yields

$$\sigma_{\pm}^2 = \frac{X_{\pm}}{1 - f'(X_{\pm})}; \tag{3.27}$$

the right-hand side of (3.27) is equal to the ratio of a fluctuation term (equal to the number of molecules) to a dissipation term (obtained by linearising the rate equation (3.1) around a stable fixed point).

The metastable transitions between the distribution modes can be described by a random telegraph process (cf. Figure 3.5, left), cf. [11],

$$\ominus \underset{1/T_+}{\overset{1/T_-}{\rightleftharpoons}} \oplus, \tag{3.28}$$

in which the lumped states $\ominus$ and $\oplus$ correspond to the basins of attractions of the two stable fixed points; T_- and T_+ are the respective residence times. The mixture weights $p_-(t)$ and $p_+(t)$ in (3.26) are identified with the probabilities of the lumped states in (3.28); these satisfy the Chapman–Kolmogorov equations [62]

$$\frac{\mathrm{d}p_-}{\mathrm{d}t} = -\frac{p_-}{T_-} + \frac{p_+}{T_+}, \quad \frac{\mathrm{d}p_+}{\mathrm{d}t} = \frac{p_-}{T_-} - \frac{p_+}{T_+}, \tag{3.29}$$

which admit an explicit solution

$$p_+(t) = \frac{T_+}{T_+ + T_-} + \left(p_+(0) - \frac{T_+}{T_+ + T_-}\right) \exp\left(-\left(\frac{1}{T_+} + \frac{1}{T_-}\right) t\right), \tag{3.30}$$

$$p_-(t) = \frac{T_-}{T_+ + T_-} + \left(p_-(0) - \frac{T_-}{T_+ + T_-}\right) \exp\left(-\left(\frac{1}{T_+} + \frac{1}{T_-}\right) t\right). \tag{3.31}$$

The initial probability $p_+(0) = 1 - p_-(0)$ is set to one or zero in (3.30)–(3.31) depending on whether the model is initialised in the neighbourhood of the upper or the lower stable fixed point.

With (3.30)–(3.31) at hand, the problem of determining the mixture weights in (3.26) is reduced to that of determining the residence times $T_{\pm}$. Previous large-deviation and WKB analyses of the one-dimensional model [63, 64, 65] provide an Arrhenius-type formula

$$T_{\pm} = 2\pi\tau X_{\pm}^{-1}\sigma_{\pm}\sqrt{-\sigma_0^2}\exp(u(X_0) - u(X_{\pm})). \tag{3.32}$$

Formula (3.32) features, on top of the familiar symbols (the mean lifetime τ, fixed points $X_{\pm}$ and X_0, linearised variances $\sigma_{\pm}$, and the Ludolph-van-Ceulen constant π), two new symbols: a value σ_0^2 and a function $u(X)$. The value σ_0^2 is readily calculated

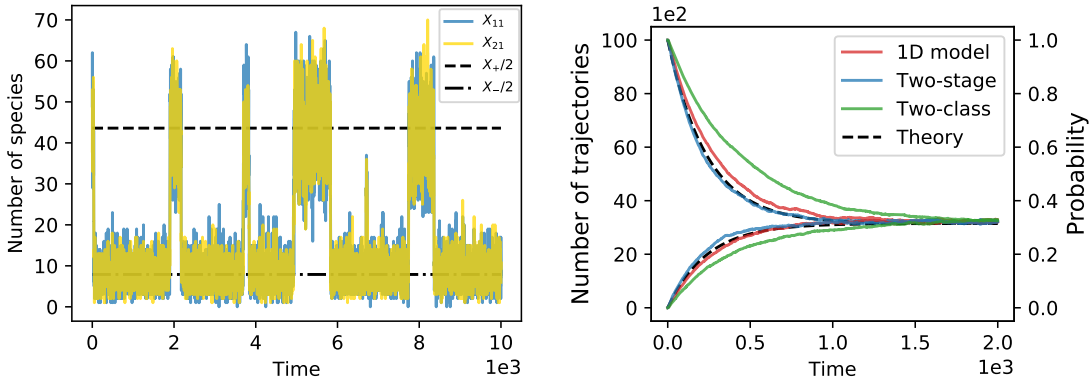


Figure 3.5: *Left:* Large-time stochastic trajectories of a structured two-class model with parameters as given below. The horizontal lines represent deterministic fixed points as given by (3.11)–(3.12). *Right:* The number of trajectories, out of 10^4 simulation repeats, that reside in the basin of attraction of the upper stable fixed point as function of time. Simulation is initiated at the upper stable fixed point (the decreasing function) or at the lower stable fixed point (the increasing function). The dashed black curve gives the theoretical probability (3.30) with initial condition $p_+(0) = 1$ (the decreasing solution) or $p_+(0) = 0$ (the increasing solution). *Parameter values:* The Hill-function parameters are: $\Omega = 50$, $H = 4$, $a_0 = 0.3$, $a_1 = 1.6$. The mean lifetime is $\tau = 1$. The two-stage model parameters are: $K = 1$, $p_1 = 1$, $S_1 = 2$, $\tau_{11} = \tau_{12} = 0.5$. The two-class model parameters are: $K = 2$, $p_1 = 1/6$, $p_2 = 5/6$, $S_1 = S_2 = 1$, $\tau_{11} = 3$, $\tau_{21} = 3/5$.

by inserting 0 instead of $\pm$ into the fluctuation–dissipation relation (3.27); note that for the unstable fixed point X_0, the denominator in (3.27) is negative (cf. Figure 3.3, left), which renders the whole fraction also negative.

In analogy with the Arrhenius law, the function $u(X)$ represents an "energy" of state X, and is given here explicitly by an indefinite integral [63, 64, 65]

$$u(X) = \int \ln\left(\frac{X}{f(X)}\right) \mathrm{d}X. \tag{3.33}$$

Note that the derivative of (3.33),

$$u'(X) = \ln\left(\frac{X}{f(X)}\right), \tag{3.34}$$

is zero if $f(X) = X$, i.e. at the fixed points of the feedback response function, is negative if $f(X) > X$ and positive if $f(X) < X$. Substituting into (3.33) the solution

$X = X(t)$ to the deterministic rate equation (3.1) and evaluating the time derivative, we find

$$\frac{\mathrm{d}u(X(t))}{\mathrm{d}t} = u'(X(t))\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \tau^{-1}(f(X) - X)\ln\left(\frac{X}{f(X)}\right)\Bigg|_{X=X(t)} \leq 0, \qquad (3.35)$$

with equality in (3.35) holding if and only if X is a fixed point of the feedback response function $f(X)$. Therefore, the energy function $u(X)$ is a Lyapunov function of the ordinary differential equation (3.1) (Figure 3.3, right). The exponentiation in (3.32) dramatically amplifies the potential difference between the stable and the unstable fixed points. For example, a moderately large potential barrier, say 5 (which is about the height of the potential barrier in Figure 3.3, right), introduces a large factor $\mathrm{e}^5 \approx 150$ in (3.32). This confirms an intuition that metastable transitions between the distribution modes are very (exponentially) slow.

The random telegraph solution (3.30) is compared in Figure 3.5 to the residence of stochastically generated trajectories in the basin of attraction of the upper fixed point. The agreement is close for simulations of the one-dimensional model (with an exponential lifetime) and for a structured model with one class and two stages (with an Erlangian lifetime). For a two-class model (with an exponential mixture lifetime), the transitioning also occurs on the exponentially slow timescale, but is perceptibly slower. Sample trajectories were generated in Python's package for stochastic simulation of biochemical systems GillesPy2 [18]. The one-dimensional model was initiated with $\lfloor X_+ \rfloor$ molecules. The two species in the two-stage and two-class models were initiated to S and $\lfloor X_+ \rfloor - S$, where S was drawn from the binomial distribution Binom($\lfloor X_+ \rfloor$, 0.5).

3.6 Discussion

In this chapter we studied a stochastic chemical reaction system for a self-regulating protein molecule with exponential and phase-type lifetimes. We demonstrated that the exponential and phase-type models support the same stationary distribution of the protein copy number. While stationary distributions of similar forms have previously been formulated in the context of queueing theory [12, 66, 67], this chapter provides a self-contained and concise treatment of the one-dimensional

model and the multi-dimensional structured model that is specifically tailored for applications in systems biology.

We showed that the invariance result rests on the assumption of non-bursty production of protein. We have demonstrated that, in the presence of bursts, exponential and deterministic lifetimes generate stationary protein-level distributions with different variances.

Deterministic modelling approaches are used in systems biology as widely as stochastic ones. Therefore, we complemented the stationary analysis of the stochastic Markov-chain models by a fixed-point analysis of deterministic models based on differential equations. The result is that, irrespective of lifetime distribution, the deterministic protein level is attracted, for large times, to the stable fixed points of the feedback response function. Connecting the stochastic and deterministic frameworks, we demonstrated that the stationary distribution of the Markovian model is sharply peaked around the fixed points of the deterministic equation. We showed that the distribution can be approximated by a mixture of Gaussian modes with means given by the deterministic fixed points and variances that are consistent with the traditional linear-noise analysis results.

Next, we focused on the transitions between the distribution modes. These occur rarely with rates that are exponentially small. We compared an asymptotic result, derived in previous literature for the one-dimensional model, to stochastic simulation results of the one-dimensional model and two specific structured models: we chose a model with one class and two stages and a model with two classes each with one stage. The simulation results of the one-dimensional and two-stage models agreed closely to the theoretical prediction; intriguingly, the agreement with theory was closer for the two-stage model. On the other hand, a two-class model showed slower transitioning rates. The theoretical asymptotic results have been derived in [63, 64, 65] only for the one-dimensional model. Large deviations in multi-dimensional models are much harder to quantify than one-variable ones. We believe that the current model, being multi-dimensional while possessing a tractable steady-state distribution, provides a convenient framework on which such methodologies can be developed.

In summary, our study provides an invariance-on-lifetime-distribution result in the deterministic and stochastic contexts for a non-bursty regulatory protein. While the

main results concern the stationary behaviour, our study also performs simulation, and opens avenue for future enquiries, into the transient transitioning dynamics.

Chapter 4

The extended two-stage model

This chapter presents a two-stage stochastic gene expression model that extends the standard model by an mRNA inactivation loop. The extended model considers that mRNA molecules can transition between their active/inactive states. We provide an extensive mathematical analysis of the joint steady-state distribution of active and inactive mRNA and protein species. We determine its generating function and derive a recursive formula for the protein distribution. In addition to obtaining the stationary means from the deterministic model, we calculate the steady-state Fano factor and express it as a function of the model parameters. Additionally, we use the analytical formula for the generating function to determine the marginal distribution for each species. The results of the analytical formula are then cross-validated by kinetic Monte-Carlo simulation.

The contents of this chapter have been published in *Lecture Notes in Computer Science*, vol 12881, pp. 215–229. Springer, 2021 [68].

4.1 Introduction

As many other biochemical mechanisms, gene expression in which protein synthesis occurs is inherently stochastic due to random fluctuations in the copy number of gene products, e.g. proteins [2]. From the viewpoint of biochemical reactions, in simplest formulations, gene expression consists of two main steps: transcription and translation. While RNA polymerase enzymes produce mRNA molecules in the former, protein synthesis takes place by ribosomes in the latter, each reaction corresponding to the production and decay of relevant species. Additionally, the two-stage model can be extended by the regulation of transcription factors, which affect gene expression by modulating the binding rate of RNA polymerase [69].

Over the last decades, the two-stage model of gene expression has been extensively studied to understand how the stochastic phenomenon in cellular processes takes place [70, 71, 21, 72]. Specifically, quantifying the number of species in terms of probability distributions has become an interesting and challenging endeavour due to the subtleties involved in finding a solution to the underlying problem. On the other hand, the fluctuations in mRNA and protein levels are considered as a major source of noise, leading to cell-to-cell variability in gene regulatory networks [73, 74, 75, 76, 77, 6, 78]. The noise emerges from different sources, namely *intrinsic* and *extrinsic* noise [79, 80]; yet, structural elements such as stem-loops can also contribute to noise by forming in an untranslated region of mRNA [81]. The untranslated regions of mRNAs often contain these stem-loops that can reversibly change configurations making individual mRNAs translationally active/inactive.

Numerous modelling approaches have been proposed based on deterministic and stochastic frameworks, including the hybrid ones as a combination of the preceding two [82, 83, 84]. Only a few of those provide an explicit solution to the two-stage gene-expression model [22, 72]; most of the studies are based on Monte Carlo simulations, which are usually computationally expensive.

As a generalisation of the two-stage model, some studies in the literature consider a set of multiple gene states and investigate the dynamics of stochastic transitions among these states [85, 86]. Nevertheless, to the best of our knowledge, none of these studies takes an mRNA inactivation into account. Here we extend the two-stage model by an mRNA inactivation loop, by which we mean that after transcription species can switch between active and inactive states. In other words, there exists a pair of reversible chemical reactions occurring at constant rates by turning active mRNA species into inactive ones, and vice versa. Subsequently, the active mRNA is translated, while the inactive mRNA stays dormant. The schematic of reactions describing the model is given in (4.1). Here we thereafter refer to the aforementioned model as *the extended model*. A possible biological scenario that can implement this extended model is by a regulatory RNA that temporarily blocks mRNA function [87].

This chapter is organised as follows. In Section 4.2, the stationary means of active mRNA, inactive mRNA, and protein are obtained from a deterministic formulation the model; the master equation of the stochastic model is formulated,

and transformed into a partial differential equation for the generating function. In Section 4.3, the partial differential equation is transformed into one for the factorial cumulant generation function and a power series solution is found; recursive expressions for the coefficients — the factorial cumulants of the three molecular species — are thereby provided. In Section 4.4, the protein Fano factor is expressed in terms of the first two factorial cumulants, which will be used in the subsequent chapter to analyse the noise-reduction effect of the mRNA inactivation loop. The generating function of the stationary distribution of active mRNA, inactive mRNA and protein amounts is represented in the special-function form in Section 4.5. The marginal protein and active and inactive mRNA distributions are derived in Section 4.6. The chapter is concluded in Section 4.7.

4.2 Model formulation

The extended model involves three species, mRNA, inactive mRNA (imRNA for short), and protein, and consists of the reactions

$$\begin{aligned} &\emptyset \underset{\gamma_1}{\overset{\lambda_1}{\rightleftharpoons}} \mathrm{mRNA}, \quad \mathrm{mRNA} \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} \mathrm{imRNA}, \quad \mathrm{imRNA} \xrightarrow{\tilde{\gamma}_1} \emptyset, \\ &\mathrm{mRNA} \xrightarrow{\lambda_2} \mathrm{mRNA} + \mathrm{protein}, \quad \mathrm{protein} \xrightarrow{\gamma_2} \emptyset. \end{aligned} \tag{4.1}$$

The reactions in (4.1) correspond to mRNA transcription and decay, mRNA activation and inactivation, inactive mRNA decay, protein translation, and protein decay, respectively.

Due to the linearity of kinetics in (4.1), the mean levels of the mRNA (m), inactive mRNA ($\tilde{m}$) and protein (n) exactly satisfy the system of deterministic rate equations

$$\begin{aligned} \frac{\mathrm{d}\langle m\rangle}{\mathrm{d}t} &= \lambda_1 - (\gamma_1+\alpha)\langle m\rangle + \beta\langle\tilde{m}\rangle, \\ \frac{\mathrm{d}\langle\tilde{m}\rangle}{\mathrm{d}t} &= \alpha\langle m\rangle - (\tilde{\gamma}_1+\beta)\langle\tilde{m}\rangle, \\ \frac{\mathrm{d}\langle n\rangle}{\mathrm{d}t} &= \lambda_2\langle m\rangle - \gamma_2\langle n\rangle. \end{aligned} \tag{4.2}$$

Setting time derivatives in (4.2) to zero, and solving the resulting algebraic system, the stationary means are obtained as

$$\langle m\rangle = \frac{\lambda_1}{\gamma_1^{\mathrm{eff}}}, \quad \langle\tilde{m}\rangle = \frac{\alpha}{\tilde{\gamma}_1+\beta}\langle m\rangle, \quad \langle n\rangle = \frac{\lambda_2}{\gamma_2}\langle m\rangle, \tag{4.3}$$

for the mRNA, inactive mRNA, and protein respectively, where

$$\gamma_1^{\text{eff}} = \gamma_1 + \frac{\alpha\tilde{\gamma}_1}{\tilde{\gamma}_1 + \beta} \tag{4.4}$$

denotes the effective rate of mRNA decay. Owing to the linearity of reaction rates, one can find a closed system of differential equations not only for means, but also for higher-order moments [88, 89]; however these equations are typically less revealing than the mean dynamics. Here we take a different approach and quantify the protein noise as a by-product of a generating-function analysis in Section 4.4.

The probability $p_{m,\tilde{m},n}(t)$ of having m mRNA, $\tilde{m}$ inactive mRNA, and n protein molecules at time t satisfies the chemical master equation

$$\begin{aligned}\frac{\mathrm{d}p_{m,\tilde{m},n}}{\mathrm{d}t} =& \lambda_1(p_{m-1,\tilde{m},n} - p_{m,\tilde{m},n}) + \alpha((m+1)p_{m+1,\tilde{m}-1,n} - mp_{m,\tilde{m},n}) \\ &+ \tilde{\gamma}_1((\tilde{m}+1)p_{m,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}) + \lambda_2 m(p_{m,\tilde{m},n-1} - p_{m,\tilde{m},n}) \\ &+ \gamma_2((n+1)p_{m,\tilde{m},n+1} - np_{m,\tilde{m},n}) + \gamma_1((m+1)p_{m+1,\tilde{m},n} - mp_{m,\tilde{m},n}) \\ &+ \beta((\tilde{m}+1)p_{m-1,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}).\end{aligned} \tag{4.5}$$

Equating the left-hand side of (4.5) to zero yields the steady-state master equation

$$\begin{aligned}0 =& \lambda_1(p_{m-1,\tilde{m},n} - p_{m,\tilde{m},n}) + \alpha((m+1)p_{m+1,\tilde{m}-1,n} - mp_{m,\tilde{m},n}) \\ &+ \tilde{\gamma}_1((\tilde{m}+1)p_{m,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}) + \lambda_2 m(p_{m,\tilde{m},n-1} - p_{m,\tilde{m},n}) \\ &+ \gamma_2((n+1)p_{m,\tilde{m},n+1} - np_{m,\tilde{m},n}) + \gamma_1((m+1)p_{m+1,\tilde{m},n} - mp_{m,\tilde{m},n}) \\ &+ \beta((\tilde{m}+1)p_{m-1,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}).\end{aligned} \tag{4.6}$$

We additionally require that the normalising condition

$$\sum_{m,\tilde{m},n} p_{m,\tilde{m},n} = 1 \tag{4.7}$$

hold.

We aim to find the moments of the probability distribution $p_{m,\tilde{m},n}$ by using the generating function approach [14]. In order to solve (4.6)–(4.7), we employ the probability generating function

$$G(x,y,z) = \sum_{m,\tilde{m},n} x^m y^{\tilde{m}} z^n p_{m,\tilde{m},n} \tag{4.8}$$

for the probability distribution $p_{m,\tilde{m},n}$. Multiplying (4.6) by the factor $x^m y^{\tilde{m}} z^n$ and

summing over $m, \tilde{m}$ and n yields

$$\begin{aligned}\lambda_1(1-x)G = &\left(\lambda_2 x(z-1) + \gamma_1(1-x) + \alpha(y-x)\right)\frac{\partial G}{\partial x} \\ &+ \left(\tilde{\gamma}_1(1-y) + \beta(x-y)\right)\frac{\partial G}{\partial y} + \gamma_2(1-z)\frac{\partial G}{\partial z}.\end{aligned} \tag{4.9}$$

Equation (4.9) is subject to

$$G(1,1,1) = 1, \tag{4.10}$$

which is implied by the normalisation condition (4.7).

4.3 Factorial cumulant generating function

In order to find a particular solution to (4.9)–(4.10), we change the variables according to

$$x = 1+u, \quad y = 1+v, \quad z = 1+w, \quad G = \exp(\varphi), \tag{4.11}$$

and obtain that the factorial cumulant generating function [15] $\varphi = \varphi(u,v,w)$ is a solution of the inhomogeneous linear partial differential equation (PDE),

$$\lambda_1 u = \left(-\lambda_2(1+u)w + \gamma_1 u + \alpha(u-v)\right)\frac{\partial \varphi}{\partial u} + \left(\tilde{\gamma}_1 v + \beta(v-u)\right)\frac{\partial \varphi}{\partial v} + \gamma_2 w\frac{\partial \varphi}{\partial w} \tag{4.12}$$

subject to

$$\varphi(0,0,0) = 0. \tag{4.13}$$

In order to solve (4.12)–(4.13) we shall employ the ansatz

$$\varphi(u,v,w) = \varphi_{00}(w) + u\varphi_{10}(w) + v\varphi_{01}(w). \tag{4.14}$$

We immediately obtain the partial derivatives

$$\frac{\partial \varphi}{\partial u} = \varphi_{10}(w), \quad \frac{\partial \varphi}{\partial v} = \varphi_{01}(w), \quad \frac{\partial \varphi}{\partial w} = \varphi'_{00}(w) + u\varphi'_{10}(w) + v\varphi'_{01}(w). \tag{4.15}$$

Inserting (4.15) into (4.12) and rearranging the terms yields an inhomogeneous system of ODEs

$$\begin{aligned}&\gamma_2 w\varphi'_{00} - \lambda_2 w\varphi_{10} = 0, \\ &\gamma_2 w\varphi'_{10} + (\gamma_1 + \alpha - \lambda_2 w)\varphi_{10} - \beta\varphi_{01} = \lambda_1, \\ &\gamma_2 w\varphi'_{01} + (\tilde{\gamma}_1 + \beta)\varphi_{01} - \alpha\varphi_{10} = 0.\end{aligned} \tag{4.16}$$

Let us assume that the functions φ_{00}, φ_{10}, and φ_{01} are of the power series form, i.e.,

$$\varphi_{00}(w) = \sum_{k=0}^{\infty} a_k w^k, \quad \varphi_{10}(w) = \sum_{k=0}^{\infty} b_k w^k, \quad \varphi_{01}(w) = \sum_{k=0}^{\infty} c_k w^k. \tag{4.17}$$

The coefficients a_k, b_k, and c_k give the factorial cumulants of the joint molecular distribution [15]. Note that $a_0 = 0$ follows immediately from the normalisation condition (4.13). Evaluating the derivatives in (4.17) and substituting into (4.16), we obtain the following recurrence equations:

$$a_k = \frac{\lambda_2}{k\gamma_2} b_{k-1}, \quad k \geq 1, \tag{4.18}$$

$$(\gamma_1 + \alpha)b_0 - \beta c_0 - \lambda_1 + \sum_{k=1}^{\infty} (\gamma_2 k b_k + (\gamma_1 + \alpha)b_k - \lambda_2 b_{k-1} - \beta c_k) w^k = 0, \tag{4.19}$$

$$(\tilde{\gamma}_1 + \beta)c_0 - \alpha b_0 + \sum_{k=1}^{\infty} (\gamma_2 k c_k + (\tilde{\gamma}_1 + \beta)c_k - \alpha b_k) w^k = 0. \tag{4.20}$$

Since we consider (4.17) as a solution to (4.12), all the coefficients in (4.19)–(4.20) must be zero. Thus, we get

$$(\gamma_1 + \alpha + \gamma_2 k)b_k - \lambda_2 b_{k-1} - \beta c_k = 0, \tag{4.21}$$

$$(\tilde{\gamma}_1 + \beta + \gamma_2 k)c_k - \alpha b_k = 0, \tag{4.22}$$

for b_k and c_k. Solving the algebraic system (4.21)–(4.22) in b_k, $k \geq 1$, yields

$$(\gamma_2^2 k^2 + \gamma_2(\tilde{\gamma}_1 + \gamma_1 + \beta + \alpha)k + \tilde{\gamma}_1\gamma_1 + \gamma_1\beta + \tilde{\gamma}_1\alpha)b_k = \lambda_2(\tilde{\gamma}_1 + \beta + k\gamma_2)b_{k-1},$$

i.e.

$$b_k = \frac{\lambda_2(\tilde{\gamma}_1 + \beta + k\gamma_2)}{\gamma_2^2 k^2 + \gamma_2(\tilde{\gamma}_1 + \gamma_1 + \beta + \alpha)k + \tilde{\gamma}_1\gamma_1 + \gamma_1\beta + \tilde{\gamma}_1\alpha} b_{k-1}, \tag{4.23}$$

where the zeroth term of the sequence b_k is obtained, by equating the terms out of the sums in (4.19) and (4.20) to zero, as

$$b_0 = \frac{\lambda_1(\tilde{\gamma}_1 + \beta)}{(\gamma_1 + \alpha)(\tilde{\gamma}_1 + \beta) - \beta\alpha} = \frac{\lambda_1}{\gamma_1^{\text{eff}}}. \tag{4.24}$$

Equation (4.24) thus rederives the stationary mRNA mean (4.3) by means of factorial cumulant analysis; similarly, c_0 and a_1 can be identified as the stationary imRNA and protein means. Thus, the sequence b_k can be calculated iteratively from (4.23) starting from the initial condition (4.24). Having calculated b_k, the sequence a_k and c_k can be evaluated via (4.18) and (4.22). In Section 4.5, we will utilise these formulas to obtain a special-function representation of the generating function. Before doing that, we show that the first two terms of these sequences determine protein variability.

4.4 Protein variability

As outlined in the previous section, the first-order cumulants b_0, c_0, and a_1 ($a_0 = 0$ by normalisation condition), coincide with the stationary mRNA, imRNA, and protein mean values. In this section, we use the second-order cumulants to describe the stationary noise in our model. The noise in mRNA and imRNA is Poissonian (see Section 4.6 for details) and therefore uninteresting: we focus on the protein noise.

Below, we express the Fano factor in terms of the first and second order cumulants, which is independent of the specifics of the current model. We will use this formula to analyse the noise reduction effect of the inactivation loop in Chapter 5 (cf. Section 5.3).

Expressing the Fano factor in terms of the cumulants. The generating function is expanded by the Taylor formula as

$$G(1,1,z) = G(1,1,1) + \frac{\partial G}{\partial z}(1,1,1)(z-1) + \frac{1}{2}\frac{\partial^2 G}{\partial z^2}(1,1,1)(z-1)^2 + \mathcal{O}(z-1)^3. \quad (4.25)$$

Differentiating (4.8) with respect to z and setting $(x,y,z) = (1,1,1)$ links the derivatives of the generating function to the factorial moments:

$$\frac{\partial G}{\partial z}(1,1,1) = \langle n \rangle, \quad \frac{\partial^2 G}{\partial z^2}(1,1,1) = \langle n(n-1) \rangle. \quad (4.26)$$

Inserting (4.10) and (4.26) into (4.25), we have

$$G(1,1,z) = 1 + \langle n \rangle (z-1) + \frac{\langle n(n-1) \rangle}{2}(z-1)^2 + \mathcal{O}(z-1)^3. \quad (4.27)$$

On the other hand, (4.11), (4.14), and (4.17) imply

$$\begin{aligned} G(1,1,z) &= \exp\left(a_1(z-1) + a_2(z-1)^2 + \mathcal{O}(z-1)^3\right) \\ &= \left(1 + a_1(z-1) + \frac{a_1^2}{2}(z-1)^2\right)\left(1 + a_2(z-1)^2\right) + \mathcal{O}(z-1)^3 \\ &= 1 + a_1(z-1) + \left(a_2 + \frac{a_1^2}{2}\right)(z-1)^2 + \mathcal{O}(z-1)^3. \end{aligned} \quad (4.28)$$

Comparing (4.27) and (4.28) gives

$$\langle n \rangle = a_1, \quad \langle n(n-1) \rangle = 2a_2 + a_1^2.$$

The Fano factor,

$$\mathrm{F} = \frac{\langle n^2 \rangle}{\langle n \rangle} - \langle n \rangle = \frac{\langle n(n-1) \rangle}{\langle n \rangle} + 1 - \langle n \rangle = \frac{2a_2}{a_1} + 1, \quad (4.29)$$

is thus expressed in terms of the first two factorial cumulants a_1 and a_2.

Substituting (4.18) and (4.23) into (4.29) and simplifying gives

$$\mathrm{F} = 1 + \frac{b_1}{b_0} = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1 + \frac{\alpha(\gamma_2+\tilde{\gamma}_1)}{\gamma_2+\tilde{\gamma}_1+\beta}}. \tag{4.30}$$

Formula (4.30) provides the steady-state protein Fano factor as function of the model parameters (degradation rate constants $\gamma_1, \tilde{\gamma}_1, \gamma_2$ of active/inactive mRNA and protein; inactivation/activation rate constants α, β; translation rate constant λ_2).

In the next section, we go beyond the mean and noise statistics (the first and second order factorial cumulants), using the higher order cumulants to find a special-function representation of the generating function of the joint distribution of mRNA, imRNA, and protein copy numbers.

4.5 Special-function representation

Factorising the second-order polynomial in k in the denominator of (4.23) gives

$$b_k = \lambda_2 \frac{\tilde{\gamma}_1 + \beta + k\gamma_2}{\gamma_2^2 (k + r_1)(k + r_2)} b_{k-1} \quad \text{for } k \geq 1, \tag{4.31}$$

where

$$r_{1,2} = \frac{\gamma_1 + \alpha + \tilde{\gamma}_1 + \beta \pm \sqrt{(\tilde{\gamma}_1 + \beta - \gamma_1 - \alpha)^2 + 4\beta\alpha}}{2\gamma_2}.$$

Note that the sequence b_k in (4.31) can be rewritten as

$$b_k = b_0 \frac{(1+\tau)_k}{(1+r_1)_k (1+r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1, \tag{4.32}$$

where we set $\tau = (\tilde{\gamma}_1 + \beta)/\gamma_2$ for the sake of simplicity and the polynomial

$$(x)_k = x(x+1)(x+2)\dots(x+k-1), \quad (x)_0 = 1$$

represents the rising factorial, also called the Pochammer symbol.

We next find the remaining sequences a_k and c_k. Inserting (4.32) into (4.18) gives

$$a_k = \frac{b_0 r_1 r_2}{\tau} \frac{(\tau)_k}{k (r_1)_k (r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1. \tag{4.33}$$

Similarly, substituting (4.32) into (4.22) yields

$$c_k = \frac{\alpha b_0}{\tilde{\gamma}_1 + \beta} \frac{(\tau)_k}{(1+r_1)_k (1+r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1, \tag{4.34}$$

where $c_0 = \frac{\alpha b_0}{\tilde{\gamma}_1+\beta}$, which can be obtained by combining (4.20) and (4.24).

Having found the sequences in (4.17), we next return to the original variables in (4.11) to obtain the generating function of the stationary distribution of active mRNA, inactive mRNA, and protein amounts, which is given by

$$\begin{aligned} &G(x,y,z) \\ &= \exp\left(\sum_{k\geq 1} a_k(z-1)^k + (x-1)\sum_{k\geq 0} b_k(z-1)^k + (y-1)\sum_{k\geq 0} c_k(z-1)^k\right). \end{aligned} \tag{4.35}$$

Equation (4.35) can be rewritten as

$$\begin{aligned} G(x,y,z) = &\exp\left(\frac{b_0\lambda_2}{\gamma_2}\int_1^z {}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right)ds\right. \\ &+b_0(x-1){}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right) \\ &\left.+\frac{\alpha b_0}{\tilde{\gamma}_1+\beta}(y-1){}_2F_2\left(\begin{matrix}1, \tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right)\right) \end{aligned} \tag{4.36}$$

in terms of the generalised hypergeometric functions defined by [23]

$${}_pF_q\left(\begin{matrix}a_1, \ldots, a_p\\ b_1, \ldots, b_q\end{matrix}; \tilde{z}\right) = \sum_{n=0}^{\infty} \frac{(a_1)_n \ldots (a_p)_n}{(b_1)_n \ldots (b_q)_n}\frac{\tilde{z}^n}{n!}. \tag{4.37}$$

Equation (4.36) provides the sought-after special function representation of the joint generating function. In the following section, we focus on specific one-dimensional sections of the joint generating function that give the generating functions of the three marginal distributions.

4.6 Marginal distributions

In this section, we use the analytic formula (4.36) for the generating function to determine the marginal active and inactive mRNA, and protein distributions. To do so, we first set $y = z = 1$ in (4.36) and obtain

$$G(x) = G(x,1,1) = \exp(b_0(x-1))$$

for the marginal active mRNA distribution. Similarly, setting $x = z = 1$ in (4.36) yields the marginal inactive mRNA distribution

$$G(y) = G(1,y,1) = \exp\left(\frac{\alpha b_0}{\tilde{\gamma}_1+\beta}(y-1)\right).$$

Finally, we set $x = y = 1$ in (4.36) and get the marginal protein generating function $G(z)$ as

$$G(z) = G(1, 1, z) = \exp(\psi(z)),$$

where ψ is given by

$$\psi(z) = \frac{b_0 \lambda_2}{\gamma_2} \int_1^z {}_2F_2\left(\begin{matrix} 1, 1+\tau \\ 1+r_1, 1+r_2 \end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right) ds. \tag{4.38}$$

In order to obtain the marginal protein distribution, we exploit its generating function

$$p_{\cdot,\cdot,n} = \left.\frac{\mathrm{D}^n(G(z))}{n!}\right|_{z=0}, \tag{4.39}$$

where D stands for the differential operator d/dz and $p_{\cdot,\cdot,z}$ gives the probability of having z protein molecules and any number of active and inactive amount of mRNA. The first derivative of the composite function $G(z)$ in (4.39) is obtained by chain rule as

$$\frac{\mathrm{d}G(z)}{\mathrm{d}z} = G(z)\frac{\mathrm{d}\psi(z)}{\mathrm{d}z}. \tag{4.40}$$

For the n-th derivative, we evaluate the $(n-1)$th derivative of (4.40) according to the Leibniz rule, thus we have

$$\mathrm{D}^n(G(z)) = \sum_{i=0}^{n-1} \binom{n-1}{i} \mathrm{D}^i(G(z))\mathrm{D}^{n-i}(\psi(z)). \tag{4.41}$$

Next, we determine the rth–r is an arbitrary positive integer–derivative of the function $\psi(z)$, which is given by

$$\mathrm{D}^r(\psi(z)) = b_0 \left(\frac{\lambda_2}{\gamma_2}\right)^r \frac{(r-1)!(1+\tau)_{r-1}}{(1+r_1)_{r-1}(1+r_2)_{r-1}} {}_2F_2\left(\begin{matrix} r, \tau+r \\ r_1+r, r_2+r \end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right), \tag{4.42}$$

in which we used the formula

$$\frac{\mathrm{d}^s}{\mathrm{d}\tilde{z}^s} {}_pF_q\left(\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}; \tilde{z}\right) = \frac{\prod_{i=1}^p (a_i)_s}{\prod_{j=1}^q (b_j)_s} {}_pF_q\left(\begin{matrix} a_1+s, \ldots, a_p+s \\ b_1+s, \ldots, b_q+s \end{matrix}; \tilde{z}\right)$$

for the s-th derivative of the generalised hypergeometric function ${}_pF_q$. Inserting the derivatives in (4.42) into (4.41), taking $z = 0$, and rearranging the resulting equation according to (4.39) gives the formula for the marginal protein probabilities

$$\begin{aligned} p_{\cdot,\cdot,n} = & \frac{b_0 \lambda_2}{n\gamma_2} \sum_{i=0}^{n-1} \left(\frac{\lambda_2}{\gamma_2}\right)^{n-i-1} \frac{(1+\tau)_{n-i-1}}{(1+r_1)_{n-i-1}(1+r_2)_{n-i-1}} \\ & \times {}_2F_2\left(\begin{matrix} n-i, \tau+n-i \\ n-i+r_1, n-i+r_2 \end{matrix}; -\frac{\lambda_2}{\gamma_2}\right) p_{\cdot,\cdot,i}, \end{aligned} \tag{4.43}$$

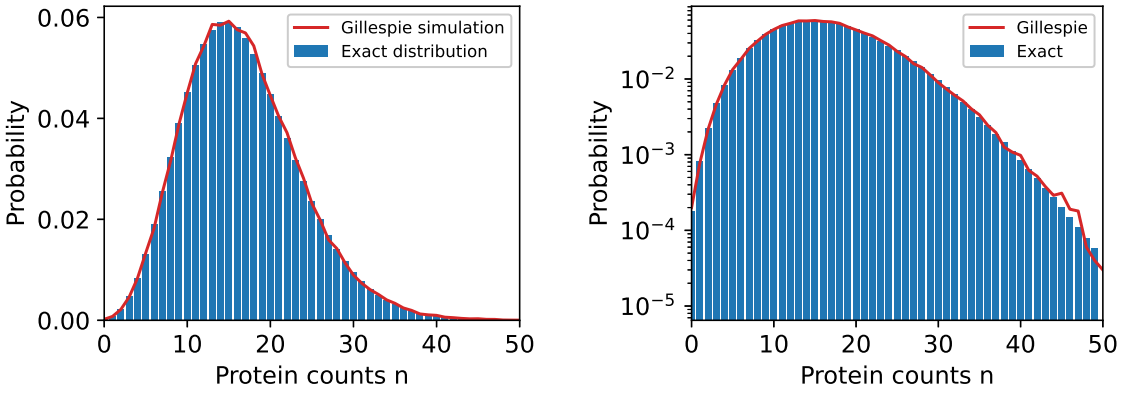


Figure 4.1: *Left:* Comparison of the probability mass function (4.43) of the marginal protein distribution and the probability calculated by Gillespie's stochastic simulation algorithm (the solid line). *Right:* A logarithmic scale plot of the probability, out of 10^5 repeats, obtained by the two approaches. *Parameter values:* The kinetic parameters are: $\lambda_1 = 5$, $\alpha = \gamma_1 = \beta = \tilde{\gamma_1} = \gamma_2 = 1$, $\lambda_2 = 5$.

where the first term of the series is given by

$$p_{\cdot,\cdot,0} = G(0) = \exp\left(-\frac{b_0\lambda_2}{\gamma_2}\int_0^1 {}_2F_2\left(\begin{matrix}1, 1+\tau \\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right)ds\right). \tag{4.44}$$

In order to calculate and compare the marginal protein probabilities (4.43) with those obtained by stochastic simulations based on Gillespie's algorithm, we implement the recursive formula (4.43) in a high-level programming language, Python, together with using its numerical computing library NumPy and plotting library Matplotlib. The probabilities in (4.43) are calculated iteratively starting from its first term given by (4.44) up to $n = 50$. In Figure 4.1, the right panel compares the theoretical probability distribution (4.43) (blue bars) with the one obtained using stochastic simulations (solid line) at the timepoint $t = 100$, while the left panel shows the same comparison but on a logarithmic scale. The number of Gillespie iterations was set to 10^5 in the Python package GillesPy2 [18]. The initial number of active and inactive mRNA and protein was set to 5. A Python routine `mpmath.hyp2f2` used to calculate the generalised hypergeometric function ${}_2F_2$ in (4.43)–(4.44).

4.7 Conclusion

In this chapter, we analysed a formulation of the two-stage model for gene expression that extends the classical version [22, 40] by an mRNA inactivation loop. The principal results of our analysis are the characterisation of the mean and noise behaviour, as well as the underlying probability distribution. The principal tool is the factorial cumulant generating function and the factorial cumulant expansion.

We have provided a comprehensive classification of the underlying probability distributions. Unsurprisingly, the distributions of the active and inactive mRNA are Poissonian. On the other hand, the protein distribution is highly non-trivial, and is characterised in terms of the generalised hypergeometric series. The characterisation is used to derive a recursive expression for the protein probability mass function. The recursive formula is found to be consistent with kinetic Monte-Carlo simulation (by means of the Gillespie direct method).

In summary, this chapter provides a systematic mathematical analysis of an mRNA–protein model for gene expression extended by an inactive mRNA species, and hints at possible functional roles of mRNA inactivation loop in the control of low copy number gene-expression noise.

In the next chapter, we shall employ the extended model and analyse the protein noise in the model. To do so, we utilise the two popular metrics: the Fano factor and the squared coefficient of variation. We cross-validate our theoretical results with the data from a recent experimental study.

Chapter 5

Effects of stem–loops on protein noise

This chapter concerns the application of gene-expression models. Specifically, we characterise noise in the basic two-stage model in terms of two noise metrics, the Fano factor and the squared coefficient of variation. Next, we compare protein noise in the basic two-stage model and the extended model. The main example pertains to the formation of stem loops; here, we reinterpret previous data and provide additional insights, summarising some of the key results of our mathematical analysis.

The content of this chapter will be published in *BMC Bioinformatics*.

5.1 Introduction

The motivation for our mathematical analysis stems from a recent experimental study [81] on the influence of RNA stem loops on gene expression noise. Stem loops appear when two palindromic sequences on the chain of nucleic acids align and form hydrogen bonds. The aligned palindromic sequences then form the "stem" and the nucleic acids in between form the "loop" of a stem loop. Another term is "hairpin loop" because of resemblance.

The authors of [81] have constructed several variants of a gene encoding for a fluorescent reporter protein. Although the constructs encode for the same reporter protein, they differ in palindromic sequences in the untranslated region at the 5' end of the gene (5'UTR). The formation of a stem–loop interferes with translation; the higher the stability of a stem–loop, the greater the interference; the lower the mean. The authors also show that this is associated with an increase in the coefficient of variation.

5.2 Application of two-stage model

Previous theoretical studies indicate that different noise metrics can lead to different interpretations of the effects of a particular mechanism on gene expression noise. The most common are the squared coefficient of variation and the Fano factor defined by

$$\mathrm{CV}^2 = \frac{\langle P^2\rangle - \langle P\rangle^2}{\langle P\rangle^2}, \quad \mathrm{F} = \frac{\langle P^2\rangle - \langle P\rangle^2}{\langle P\rangle},$$

where P stands for the reporter protein and $\langle . \rangle$ are the averaging brackets. In Figure 5.1, in addition to showing the dependence of the CV^2 on mean (thus reproducing Figure 6 of [81]), we also show the dependence of $\mathrm{F} = \langle P\rangle \mathrm{CV}^2$ on the mean. Notably, decreasing the mean (which is associated with greater stem loop stability) decreases the Fano factor.

In order to explain the apparently contradictory interpretations, we fit the basic two-stage (transcription-translation) model of gene expression [40, 82]. The model is described in full mathematical detail in Section 2.8. For the purposes of the current section, we mention that it predicts that the stationary protein mean and Fano factor of the form

$$\langle P\rangle = \frac{\lambda_1\lambda_2}{\gamma_1\gamma_2}, \quad \mathrm{F} = 1 + \frac{\lambda_2}{\gamma_1+\gamma_2},$$

where λ_1 is the mRNA production rate, λ_2 is the protein translation rate, γ_1 and γ_2 are the decay rate constants of mRNA and protein species, respectively. Provided that the protein is more stable than the mRNA ($\gamma_2 \ll \gamma_1$), we can simplify to

$$\mathrm{F} = 1 + \frac{\lambda_2}{\gamma_1} = 1 + \frac{\gamma_2\langle P\rangle}{\lambda_1}, \quad \mathrm{CV}^2 = \frac{\mathrm{F}}{\langle P\rangle} = \frac{1}{\langle P\rangle} + \frac{\gamma_2}{\lambda_1}. \tag{5.1}$$

Stem loops do not affect the transcription rate λ_1 or the protein stability γ_2, but they can affect the protein mean through translation rate λ_2 and mRNA decay rate γ_1. Thus, the two-stage model predicts an increasing linear dependence of the Fano factor, and a decreasing hyperbolic dependence of the CV^2, on the mean. In Figure 5.1, the Fano factor data are fit by a straight line using simple linear regression. The regression coefficients are reused for the hyperbolic dependence of the CV^2. The fits seem to be satisfactory, leading us to attribute the changes in the noise to the decrease of mean rather than an active control of noise by the stem–loop mechanism.

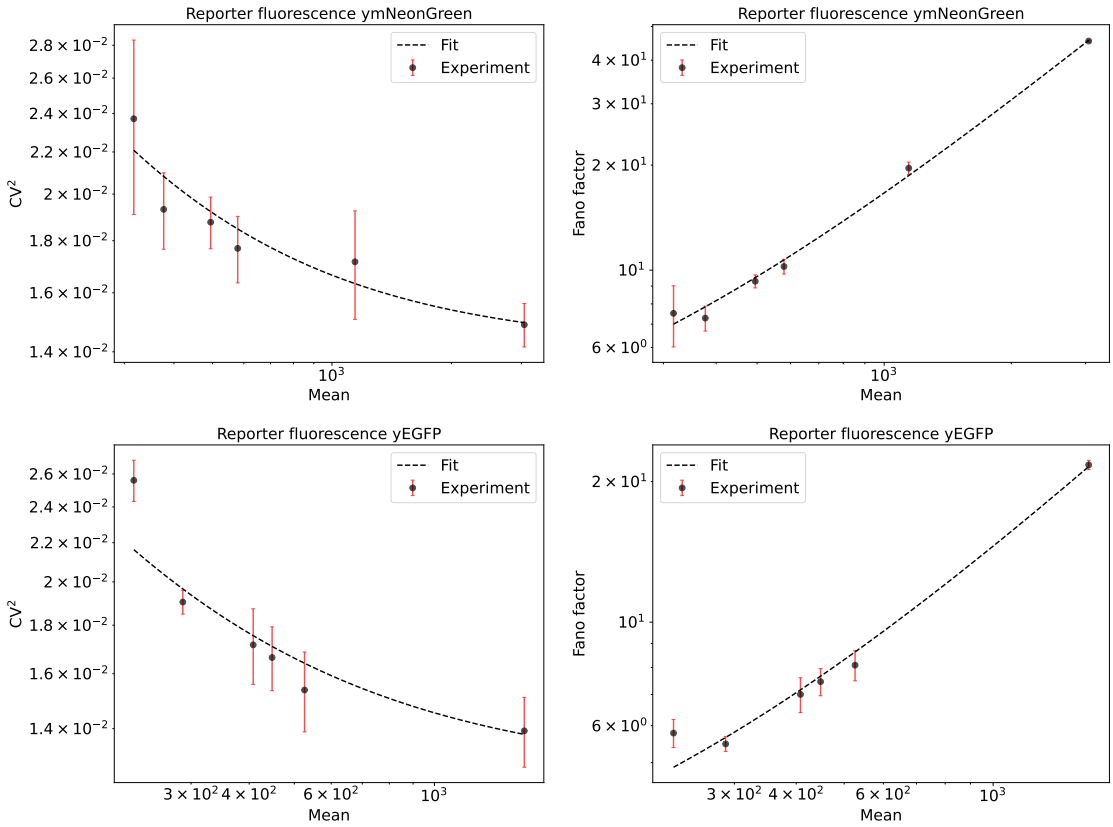


Figure 5.1: Dependence of protein noise on protein mean for different 5'UTR constructs. The yEGFP reporter (bottom) and the ymNeonGreen reporter (top) constructs are treated separately. The use of a log-log scale is adopted from [81]. The dots give the experimental values taken from [81] (see Table 5.1). Each dot is a result of multiple experiments, and the error bars indicate the standard deviation. These were obtained from the standard deviation of the (nonsquared) coefficient of variation by Taylor formula: $SD_{CV^2} = 2CVSD_{CV}$, $SD_F = \langle P \rangle SD_{CV^2}$. The dashed lines give the linear and hyperbolic dependence of the F and CV^2, respectively, which are predicted by the two-stage gene expression model (cf. (5.1)). The protein translation rate λ_2 and the mRNA decay rate γ_1 are being varied to change the mean levels. Note that the use of the log-log scale results in a slight curvature of the line (with a nonzero intercept).

5.3 Noise control by stem–loop

Let us address the question of noise control by stem–loop formation theoretically. For reasons of mathematical elegance, we will introduce in Section 6.2 and analyse in Sections 6.3–6.4 a general model that extends the basic two-stage by multiple transcript states. Here we discuss the special case with two states, one of them translationally active (without a stem–loop), the other translationally inactive (with a stem–loop). This special case was analysed in Chapter 4. Using standard methods, we derived that the mean is given by

$$\langle P \rangle = \frac{\lambda_2 \lambda_1}{\gamma_2 \gamma_1^{\text{eff}}},$$

where

$$\gamma_1^{\text{eff}} = \gamma_1 + \frac{\alpha \tilde{\gamma_1}}{\tilde{\gamma_1} + \beta} \tag{5.2}$$

gives an effective mRNA decay rate constant. The Fano factor satisfies

$$\mathrm{F} = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1 + \frac{\alpha(\gamma_2 + \tilde{\gamma_1})}{\gamma_2 + \tilde{\gamma_1} + \beta}}.$$

The above equations give the steady-state protein mean and Fano factor as function of the model parameters (degradation rate constants $\gamma_1, \tilde{\gamma_1}, \gamma_2$ of active/inactive mRNA and protein; inactivation/activation rate constants α, β; translation rate constant λ_2). The formula for the mean implies, in particular, that making the stem–loop more stable (i.e. decreasing β) decreases the mean. The noise requires a more subtle analysis, which is given below.

In order to compare the protein noise in the current model to that exhibited by the classical two-stage model (without the inactivation–activation loop) we define the baseline Fano factor as

$$\mathrm{F}_0 = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1^{\text{eff}}} = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1 + \frac{\alpha \tilde{\gamma_1}}{\tilde{\gamma_1} + \beta}}, \tag{5.3}$$

which can be obtained from (4.30) by first setting $\alpha = 0$ (no inactivation) and then replacing the mRNA decay rate γ_1 by its effective value (5.2). Adjusting the mRNA decay rate maintains the same species means in the baseline model like in the full model extended by the inactivation loop. Note that a comparison in protein variance

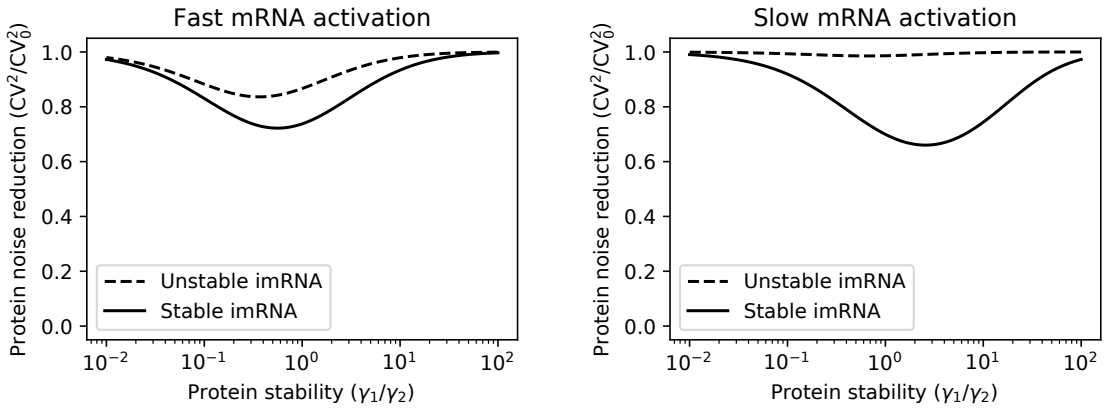


Figure 5.2: Fractional protein noise reduction by the mRNA inactivation loop as function of protein stability. The ordinate gives the protein noise (the squared coefficient of variation) in the two-stage model extended by the mRNA inactivation loop relative to the protein noise in a baseline two-stage model without the mRNA inactivation loop (adjusting the mRNA decay rate to obtain the same species means). The protein mean is set to $\langle n \rangle = 500$; the mRNA mean is $\langle m \rangle = 10$; the imRNA decay rate is either the same as that of active mRNA ($\tilde{\gamma}_1 = \gamma_1$; dashed line) or set to zero ($\tilde{\gamma}_1 = 0$; solid line). The inactivation and activation rates are $\alpha = 3$, $\beta = 3$ (left panel) or $\alpha = 1$, $\beta = 0.1$ (right panel); we thereby set $\gamma_1 = 1$ without loss of generality.

	yEGFP		ymNeonGreen	
Construct	μ	CV (%)	μ	CV (%)
L0 (P_{TEF1})	1560	11.8 ± 0.5	3050	12.2 ± 0.3
U	526	12.4 ± 0.6	-	-
M1Ug	408	13.1 ± 0.6	-	-
M3g	226	16.0 ± 0.4	-	-
G_{10}	448	12.9 ± 0.5	-	-
G_{14}	-	-	317	15.4 ± 1.5
M3Wn	-	-	1143	13.1 ± 0.8
M3n	-	-	579	13.3 ± 0.5
M3Un	-	-	377	13.9 ± 0.6
L0 (P_{PAB1})	288	13.8 ± 0.2	495	13.7 ± 0.4

Table 5.1: Protein mean and noise (CV) values for the yEGFP and the ymNeonGreen reporters obtained from [81]. The hyphen symbol denotes undetermined values.

between the extended and canonical two-stage model can also be done by the mRNA autocovariance function [90].

The protein variability formulae (4.30) and (5.3) can equivalently be expressed in terms of the squared coefficient of variation [91, 92] $\mathrm{CV}^2 = \mathrm{F}/\langle n\rangle$ and $\mathrm{CV}_0^2 = \mathrm{F}_0/\langle n\rangle$. Combining (4.3) and (4.30)–(5.3), we find

$$\mathrm{CV}^2 = \frac{1}{\langle n\rangle} + \frac{1}{\langle m\rangle}\frac{\gamma_2}{\gamma_2+\gamma_1+\frac{\alpha(\gamma_2+\tilde{\gamma}_1)}{\gamma_2+\tilde{\gamma}_1+\beta}}, \tag{5.4}$$

$$\mathrm{CV}_0^2 = \frac{1}{\langle n\rangle} + \frac{1}{\langle m\rangle}\frac{\gamma_2}{\gamma_2+\gamma_1+\frac{\alpha\tilde{\gamma}_1}{\tilde{\gamma}_1+\beta}} \tag{5.5}$$

for the protein coefficient of variation and its baseline value (no activation loop).

Comparing (5.4) to (5.5), we see that $\mathrm{CV}^2 < \mathrm{CV}_0^2$, allowing us to conclude that the inclusion of the mRNA inactivation loop decreases protein noise. However, the two coefficients will be very close in many parameter regimes; the necessary conditions for observing a significant difference are given by

$$\tilde{\gamma}_1 \lesssim \min\{\beta, \gamma_2\}, \quad \max\{\gamma_1, \gamma_2\} \lesssim \alpha, \tag{5.6}$$

where by "$\lesssim$" we mean smaller than or of similar magnitude. Thus, in order to obtain

significant reduction of noise, we require that an individual active mRNA molecule be more likely to be inactivated than degraded, and that an individual inactive mRNA molecule be more likely to be activated than degraded. Additionally, we require that inactive mRNA be more stable than protein (which is possible if inactivation protects the mRNA from decay).

One particular consequence of the necessary conditions (5.6) is that the fraction protein noise reduction, $\mathrm{CV}^2/\mathrm{CV}_0^2$, is a non-monotonous function of protein stability: it tends to one for highly unstable or highly stable proteins, and is less than one for proteins of optimal stability (cf. Figure 5.2). The optimal value of protein stability critically depends on the rate constant β of mRNA activation. In case of fast mRNA activation, the optimum noise reduction is achieved by unstable proteins (less stable than mRNA; Figure 5.2, left panel. In case of slow mRNA activation, the optimum can be achieved by stable proteins (Figure 5.2, right panel). However, slow activation ($\beta \ll 1$) imposes, via (5.6), a stringent condition on the stability of inactivated mRNA. Indeed, the right panel of Figure 5.2 demonstrates that there is hardly any reduction of noise if the inactive mRNA is unstable.

5.4 Conclusion

In this chapter, we focused on the applications of gene expression models, which we introduced in the previous chapters. Our motivation for studying these models was the influence of stem loops on gene expression noise. In particular, we comprehensively studied the characterisation of protein noise in the basic two-stage and the extended model.

We first presented the two noise metrics for the basic two-stage model: the Fano factor and the squared coefficient of variation. Our theoretical results were fitted to data from an experimental study, providing an illustrative comparison that allows us to conclude that the changes in noise are due to the decrease in mean rather than an active control of noise by the stem–loop mechanism.

Next, we considered the inactivation loop model, showing that the incorporation of the mRNA inactivation loop into the classical two-stage model for gene expression reduces the protein noise. However, in order for the reduction be substantial, several

restrictions on the parameter rates have to be in place. In particular, the protein cannot be too stable or unstable, but its stability has to be optimally chosen. The resulting optimal value of protein stability is typically unrealistically low (lower than mRNA stability, in particular). In order to obtain an optimal stability that is greater than mRNA stability, one has to assume that inactivation protects the mRNA from degradation and activation is slow. Thus, our noise analysis points towards a potential role of the mRNA inactivation loop in gene expression noise control; at the same time, it delineates the limits of its application.

Overall, this chapter provides a comprehensive noise analysis of two different gene-expression mechanisms by which the noise in relevant biological scenarios can be controlled.

Chapter 6

The generalised model

In this chapter, we study a structuration/generalisation of a stochastic gene-expression model in which mRNA molecules can be found in one of its finite number of different states and can transition among these states. We give the complete mathematical description of the model in Section 6.2. Next, we characterise and derive non-trivial analytical expressions for the steady-state protein distribution in Section 6.3. Furthermore, we obtain the marginal mRNA and protein distributions and provide the protein Fano factor. In particular, we show that two different gene-expression models, the extended model and the model with multiphasic mRNA lifetime, can be obtained from the structured/generalised model.

The content of this chapter will be published in *BMC Bioinformatics*.

6.1 Introduction

As a generalisation of the (classical) two-stage model, some studies in the literature consider a set of multiple gene states and investigate the dynamics of stochastic transitions among these states [85, 86]. In Chapter 4 we studied an extension of the classical model with two distinct mRNA states. Here, we study a structuration/ generalisation of the classical two-stage gene-expression model, which takes into account multiple mRNA states. More specifically, after being transcribed, mRNA molecules are considered to be transitioning among their different states at constant reaction rates. Subsequently, the nascent mRNA molecule is translated, and protein is degraded. The schematic of the reactions describing this system is given (6.1). In what follows, we formulate the structured model in terms of the Chemical Master Equation (CME), and by solving it, we obtain quantitative insights into the characteristics of the copy number distributions. Additionally, we present two

specific examples which can be obtained from the structured/generalised model: the mRNA inactivation loop model and the multiphasic mRNA model. We complement our analysis by performing stochastic simulations that validate our theoretical results.

This chapter is structured as follows. In Section 6.2, a brief review of the basic two-stage gene-expression model is given in deterministic and stochastic settings; the associated CME to this model, which is then transformed into a partial differential equation (PDE) for the generating function, is formulated; the core of this chapter, in which a generalisation of the two-stage model and its corresponding CME and PDE are given, is introduced. In Section 6.3, a power series solution to the PDE is found. In Section 6.4, the marginal mRNA and protein distributions are obtained using the analytical formula for the generation function; moments of the protein distributions are determined utilising the factorial cumulants, thereby the protein distribution is recovered. In Section 6.5, an example of the generalised model, explaining how the model is reduced to the one in which only two mRNA states (namely active and inactive) are considered, is given. In Section 6.6, a particular case of the generalised model which takes multiphasic mRNA lifetime into account is provided. The chapter is concluded in Section 6.7.

6.2 Model formulation

A generalisation of the basic two-stage model (2.86) is given by the following set of reactions:

$$\begin{aligned}
&\emptyset \xrightarrow{\lambda_i^m} \mathrm{mRNA_i} \xrightarrow{\gamma_i^m} \emptyset, \quad i = 1, \ldots, K, \\
&\mathrm{mRNA_i} \xrightarrow{q_{ij}} \mathrm{mRNA_j}, \quad i, j = 1, \ldots, K, \quad i \neq j, \\
&\mathrm{mRNA_i} \xrightarrow{\lambda_i^p} \mathrm{mRNA_i} + \mathrm{protein}, \quad i = 1, \ldots, K, \\
&\mathrm{protein} \xrightarrow{\gamma^p} \emptyset,
\end{aligned} \tag{6.1}$$

where λ_i^m and γ_i^m are the production and decay rates for an mRNA molecule in i-th state, respectively. The term q_{ij}, $i \neq j$, denotes the mRNA transition rate from state i to state j, λ_i^p is the protein translation rate, and γ^p is the protein decay rate. In (6.1), the reactions correspond to mRNA transcription and decay, transitions among these mRNA molecules, protein translation, and protein decay, respectively. Throughout

this chapter, we refer the model described by (6.1) to as the *generalised two-stage model*, by which we mean that the model is considered as an extension of the classical two-stage model by structuration of mRNA.

For the generalised model (6.1), the probability $P(\mathbf{m}, n, t)$ of observing m_1 mRNA copies in state 1, m_2 mRNA copies in state 2, and so on, at given time t satisfies the following CME,

$$\begin{aligned}\frac{\mathrm{d}P(\mathbf{m}, n, t)}{\mathrm{dt}} = &\sum_{i=1}^{K} \Bigg(\lambda_i^m(\mathbb{E}_i^{-1} - 1)P + \gamma_i^m(\mathbb{E}_i - 1)m_i P + \sum_{j=1}^{K} q_{ij}(\mathbb{E}_i\mathbb{E}_j^{-1} - 1) \\ &\times m_i P + \lambda_i^p(\mathbb{E}_{K+1}^{-1} - 1)m_i P \Bigg) + \gamma^p(\mathbb{E}_{K+1} - 1)nP,\end{aligned} \tag{6.2}$$

where $\mathbf{m} = \begin{bmatrix} m_1 & m_2 & m_3 & \dots & m_K \end{bmatrix}$. Note that the step operator [10] $\mathbb{E}_i$ in (6.2) is in the variable m_i, whereas $\mathbb{E}_{K+1}$ in the variable n; $\mathbb{E}_i\mathbb{E}_j^{-1} - 1 = 0$ for $i = j$.

The multivariate probability generating function is given by

$$G(\mathbf{x}, y, t) = \sum_{m_1} \cdots \sum_{m_K} \sum_{n} P(\mathbf{m}, n, t) x_1^{m_1} x_2^{m_2} \cdots x_K^{m_K} y^n, \tag{6.3}$$

where $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_K \end{bmatrix}$. Multiplying (6.2) by $x_1^{m_1} x_2^{m_2} \dots x_K^{m_K} y^n$ and summing over all $m_1, m_2, \dots, m_K, n$, we arrive at the PDE

$$\begin{aligned}\frac{\partial G(\mathbf{x}, y, t)}{\partial t} = &\sum_{i=1}^{K} \Bigg(\lambda_i^m(x_i - 1)G + \gamma_i^m(1 - x_i)\frac{\partial G}{\partial x_i} + \sum_{j=1}^{K} q_{ij}(x_j - x_i)\frac{\partial G}{\partial x_i} \\ &+ \lambda_i^p(y - 1)x_i\frac{\partial G}{\partial x_i} \Bigg) + \gamma^p(1 - y)\frac{\partial G}{\partial y}.\end{aligned} \tag{6.4}$$

Note that the step operators $\mathbb{E}_i^{\pm 1}$ in (6.2) coincide with the variables $x_i^{\mp 1}$ while the number of species m_i correspond to the terms $x_i\partial_{x_i}$ in (6.4) for the generating function. In the next section, we will seek a solution to the PDE (6.4).

6.3 Solution

In this section, we shall provide a step-by-step breakdown of our solution method for solving the PDE (6.4). We are interested in the steady state; therefore, we set the time

derivative in (6.4) to zero and rearrange the resulting equation to obtain

$$\sum_{i=1}^{K}\left(\lambda_i^m(x_i-1)G+\left(\gamma_i^m(1-x_i)+\sum_{j=1}^{K}q_{ij}(x_j-x_i)+\lambda_i^p(y-1)x_i\right)\frac{\partial G}{\partial x_i}\right) + \gamma^p(1-y)\frac{\partial G}{\partial y}=0 \tag{6.5}$$

for the time-independent generating function $G(\mathbf{x},y)$ of the stationary distribution. The probability normalisation condition translates to $G(1,\ldots,1)=1$. Changing the variables according to

$$x_i=1+u_i,\quad y=1+v,\quad G=\exp(\varphi) \tag{6.6}$$

allows us to transform (6.5) into

$$\sum_{i=1}^{K}\left(\lambda_i^m u_i+\left(\lambda_i^p v(1+u_i)-\gamma_i^m u_i+\sum_{j=1}^{K}q_{ij}(u_j-u_i)\right)\frac{\partial\varphi}{\partial u_i}\right)=\gamma^p v\frac{\partial\varphi}{\partial v}, \tag{6.7}$$

which is subject to the normalisation condition

$$\varphi(\mathbf{0})=0. \tag{6.8}$$

Below, we focus on seeking a solution to (6.7)–(6.8) using a suitable ansatz.

Let us first consider that the solution is of the form

$$\varphi(u_1,u_2,u_3,\ldots,u_K,v)=\varphi_0(v)+u_1\varphi_1(v)+\ldots+u_K\varphi_K(v). \tag{6.9}$$

With this in mind, we obtain from (6.9) that

$$\frac{\partial\varphi}{\partial u_i}=\varphi_i(v),\quad \frac{\partial\varphi}{\partial v}=\varphi_0'(v)+u_1\varphi_1'(v)+\ldots+u_K\varphi_K'(v). \tag{6.10}$$

Inserting the partial derivatives (6.10) into (6.7), we get

$$\sum_{i=1}^{K}\left(\lambda_i^m u_i+\left(\lambda_i^p v(1+u_i)-\gamma_i^m u_i+\sum_{j=1}^{K}q_{ij}(u_j-u_i)\right)\varphi_i-\gamma^p v u_i\varphi_i'\right) = \gamma^p v\varphi_0'. \tag{6.11}$$

Equation (6.11) can be rewritten as

$$\left(\gamma^p\varphi_0'-\sum_{i=1}^{K}\lambda_i^p\varphi_i\right)v + \sum_{i=1}^{K}\left(\gamma^p v\varphi_i'+\left(\gamma_i^m-\lambda_i^p v+\sum_{j=1}^{K}q_{ij}\right)\varphi_i-\sum_{j=1}^{K}q_{ji}\varphi_j-\lambda_i^m\right)u_i=0. \tag{6.12}$$

In order that (6.12) hold, we must necessarily have

$$\sum_{i=1}^{K} \lambda_i^p \varphi_i - \gamma^p \varphi_0' = 0, \tag{6.13}$$

$$\gamma^p v \varphi_i' + \left(\gamma_i^m - \lambda_i^p v + \sum_{j=1}^{K} q_{ij} \right) \varphi_i - \sum_{j=1}^{K} q_{ji} \varphi_j = \lambda_i^m. \tag{6.14}$$

Next, we solve the system of ODEs (6.13)–(6.14) using the power series method.

Let us assume that the functions φ_0 and φ_i are of the power series form, i.e.,

$$\varphi_0(v) = \sum_{n=0}^{\infty} a_n v^n, \quad \varphi_i(v) = \sum_{n=0}^{\infty} b_n^{(i)} v^n \tag{6.15}$$

for $i \in \{1, \ldots, K\}$. Differentiating (6.15) term by term we get

$$\varphi_0'(v) = \sum_{n=1}^{\infty} n a_n v^{n-1}, \quad \varphi_i'(v) = \sum_{n=1}^{\infty} n b_n^{(i)} v^{n-1}. \tag{6.16}$$

Inserting (6.15) and (6.16) into (6.14), and collecting same powers of v, we obtain the following system of recurrence relations

$$\left(\gamma_i^m + \sum_{j=1}^{K} q_{ij} + n\gamma^p \right) b_n^{(i)} - \sum_{j=1}^{K} q_{ji} b_n^{(j)} = \lambda_i^p b_{n-1}^{(i)} \tag{6.17}$$

for $b_n^{(i)}$, $i = 1, \ldots, K$. Equations (6.17) can be rewritten in matrix form as

$$(\mathbf{A} - \mathbf{Q}^\top + n\gamma^p \mathbf{I}) X_n = \mathbf{B} X_{n-1}, \quad n \geq 1, \tag{6.18}$$

where $\mathbf{I}$ is the identity matrix and the vector X_n is defined as

$$X_n = \left[b_n^{(1)}, \quad b_n^{(2)}, \quad b_n^{(3)}, \quad \ldots, \quad b_n^{(K)} \right]^\top. \tag{6.19}$$

In (6.18), $\mathbf{A}$ is a $K \times K$ matrix defined by

$$\mathbf{A}_{ij} := \begin{cases} \gamma_i^m & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \tag{6.20}$$

$\mathbf{Q}$ is a $K \times K$ matrix defined by

$$\mathbf{Q}_{ij} := \begin{cases} -\sum_{k \neq i} q_{ik} & \text{for } i = j, \\ q_{ij} & \text{for } i \neq j, \end{cases} \tag{6.21}$$

and $\mathbf{B}$ is a $K \times K$ matrix defined by

$$\mathbf{B}_{ij} := \begin{cases} \lambda_i^p & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \tag{6.22}$$

In order to solve the recurrence relations (6.18) initial conditions are needed. These can be obtained from (6.14) by setting $v = 0$ for each $i \in \{1, 2, \ldots, K\}$. The resulting system of linear equations is given in matrix form as

$$(\mathbf{A} - \mathbf{Q}^\top) X_0 = C, \tag{6.23}$$

where C is a column vector defined as $C = \begin{bmatrix} \lambda_1^m & \lambda_2^m & \ldots & \lambda_K^m \end{bmatrix}^\top$.

Solving the system of algebraic equations (6.18) under the initial conditions (6.23) yields the terms of $b_n^{(i)}$; the sequence a_n can be obtained by substituting (6.15) and (6.16) into (6.13) and collecting same powers of v. By doing so, we get

$$a_n = \frac{1}{n\gamma^p} \sum_{i=1}^{K} \lambda_i^p b_{n-1}^{(i)}, \quad n \geq 1. \tag{6.24}$$

Here the normalisation condition (6.8) implies that $a_0 = \varphi_0(0) = \varphi(\mathbf{0}) = 0$. Having found the sequences a_n and $b_n^{(i)}$, we combine (6.9) and (6.15) to obtain

$$\varphi(u, v) = \sum_{n=1}^{\infty} a_n u^n + \sum_{i=1}^{K} v_i \sum_{n=0}^{\infty} b_n^{(i)} u^n. \tag{6.25}$$

We return to the original variables in (6.25) via (6.6) to obtain the generating function of the stationary distribution of mRNA and protein amounts, which is given by

$$G(\mathbf{x}, y) = \exp\left(\sum_{n=1}^{\infty} a_n (y-1)^n + \sum_{i=1}^{K} (x_i - 1) \sum_{n=0}^{\infty} b_n^{(i)} (y-1)^n \right). \tag{6.26}$$

Equation (6.26) provides the sought-after steady-state solution to the PDE (6.4) and will be used in the following section.

6.4 Marginal distributions and moments

In this section, we use the analytical formula for the generating function (6.26) to obtain marginal mRNA distributions. Subsequently, we determine the moments of the protein distribution by way of the factorial cumulants, allowing us to recover the protein distribution. Additionally, we derive the protein Fano factor (variance-to-mean ratio) and express it in terms of the first two factorial moments.

Marginal mRNA distributions. In the generating function (6.26), if we take $y = 1$, then we get the marginal mRNA distributions as

$$G^m(\mathbf{x}) = G(\mathbf{x}, 1) = \exp\left(\sum_{i=1}^{K} b_0^{(i)}(x_i - 1)\right) = \prod_{i=1}^{K} \exp\left(b_0^{(i)}(x_i - 1)\right), \tag{6.27}$$

from which we conclude that the steady state mRNA distributions are independent Poissons with means

$$\langle m_i \rangle = b_0^{(i)}. \tag{6.28}$$

Marginal protein distribution. We can recover the generating function of the marginal protein distribution, by inserting $x_i = 1$, $i = 1, \ldots, K$, into (6.26), as

$$G(y) = G(\mathbf{1}, y) = \exp\left(\sum_{n=1}^{\infty} a_n (y-1)^n\right), \tag{6.29}$$

where $\mathbf{1}$ is a K-dimensional row vector of ones.

Next, we determine the moments of the protein distributions. The factorial (combinatorial) moments h_n are obtained by expanding the generating function into a power series around $y = 1$:

$$G(y) = \sum_{n=0}^{\infty} h_n (y-1)^n. \tag{6.30}$$

We aim to calculate the factorial moments h_n by way of the factorial cumulants a_n. To that end, we first differentiate (6.29) to obtain

$$\mathrm{D}G(y) = G(y)\mathrm{D}\ln G(y), \tag{6.31}$$

where D denotes the differential operator $\mathrm{d}/\mathrm{d}y$. Then, taking the $(n-1)$th derivative of (6.31), we get

$$\mathrm{D}^n G(y) = \sum_{i=0}^{n-1} \binom{n-1}{i} \mathrm{D}^i G(y) \mathrm{D}^{n-i} \ln G(y), \tag{6.32}$$

which can be recast as

$$\frac{\mathrm{D}^n G(y)}{n!} = \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \frac{\mathrm{D}^i G(y)}{i!} \frac{\mathrm{D}^{n-i}(\ln G(y))}{(n-i)!}. \tag{6.33}$$

Evaluating (6.33) at $y = 1$ gives the factorial moments of the protein distribution

$$h_n = \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) a_{n-i} h_i, \quad \text{for } n \geq 1, \tag{6.34}$$

where $h_0 = 1$. The terms of h_n can be recursively obtained by inserting (6.24) into (6.34). Subsequently, we recover the protein distribution exploiting the recurrence method [93] given by

$$p(n) = \sum_{j=1}^{\infty} \frac{(j+1)_n}{n!} h_{n+j} (-1)^j, \tag{6.35}$$

where $(x)_n$, n being a nonnegative integer, denotes the rising factorial or namely Pochhammer symbol.

Moments. The mRNA distributions in (6.27) are Poissonian. Therefore, mRNA Fano factor is equal to 1. The protein mean and Fano factor can be derived from the factorial moments (6.34). The first two factorial moments are given by

$$\langle n \rangle = h_1 = a_1 \quad \text{and} \quad \langle n(n-1) \rangle = 2h_2 = 2a_2 + a_1^2, \tag{6.36}$$

respectively. The Fano factor,

$$\mathrm{F} = \frac{\langle n^2 \rangle}{\langle n \rangle} - \langle n \rangle = \frac{\langle n(n-1) \rangle}{\langle n \rangle} + 1 - \langle n \rangle = \frac{2a_2}{a_1} + 1, \tag{6.37}$$

is thus expressed in terms of the first two factorial cumulants a_1 and a_2.

6.5 The mRNA inactivation loop model

In this section, we reconsider the mRNA inactivation loop model, which we previously studied in detail in Chapter 4. Here, we present how the structured/generalised model can be used to encapsulate this model. In particular, we rederive the (full) high-order cumulants for the extended model using the generalised model.

As mentioned, the inactivation loop model can be readily obtained from the generalised model (6.1) by taking $K = 2$, which accounts for only two mRNA states denoting the active mRNA state m_1 and the inactive mRNA state m_2. In what follows, we assume that a newly produced mRNA is active, i.e. that the transcription rate satisfies

$$\lambda_i^m = \lambda^m \delta_{i,1}, \quad \text{for } i = 1, 2. \tag{6.38}$$

Additionally, we assume that proteins are translated only from an active mRNA, so that we have

$$\lambda_i^p = \lambda^p \delta_{i,1}, \quad \text{for } i = 1, 2, \tag{6.39}$$

for the translation rate. Here, $\delta_{i,j}$ denotes the Kronecker delta symbol.

Cumulants. We aim to recover expressions for the inactivation loop model from the generalised model. The system of algebraic equations for this model follow from (6.23), taking the form of

$$(\gamma_1^m + q_{12})b_0^{(1)} - q_{21}b_0^{(2)} = \lambda^m, \tag{6.40}$$

$$(\gamma_2^m + q_{21})b_0^{(2)} - q_{12}b_0^{(1)} = 0, \tag{6.41}$$

from which we recover

$$b_0^{(1)} = \frac{\lambda^m(\gamma_2^m + q_{21})}{(\gamma_1^m + q_{12})(\gamma_2^m + q_{21}) - q_{12}q_{21}}. \tag{6.42}$$

Combining (6.42) with (6.29) we find

$$\langle m_1 \rangle = \frac{\lambda^m}{\gamma_{\text{eff}}^m}, \tag{6.43}$$

where

$$\gamma_{\text{eff}}^m = \gamma_1^m + \frac{q_{12}\gamma_2^m}{\gamma_2^m + q_{21}} \tag{6.44}$$

is the effective rate of mRNA decay. The recurrence relations (6.18) read

$$(\gamma_1^m + q_{12} + n\gamma^p)b_n^{(1)} - \lambda^p b_{n-1}^{(1)} - q_{21}b_n^{(2)} = 0, \tag{6.45}$$

$$(\gamma_2^m + q_{21} + n\gamma^p)b_n^{(2)} - q_{12}b_n^{(1)} = 0, \tag{6.46}$$

for $n \geq 1$. Solving the algebraic system (6.45)–(6.46) in $b_n^{(1)}$ yields

$$b_n^{(1)} = \frac{\lambda^p(\gamma_2^m + q_{21} + n\gamma^p)}{\gamma^{p2}n^2 + \gamma^p(\gamma_2^m + \gamma_1^m + q_{21} + q_{12})n + \gamma_2^m\gamma_1^m + \gamma_1^m q_{21} + \gamma_2^m q_{12}} b_{n-1}^{(1)}, \tag{6.47}$$

which is a recursive expression whose first term (i.e. zeroth) is given by (6.42). The sequence a_n can be obtained from (6.24) as

$$a_n = \frac{\lambda^p}{n\gamma^p} b_{n-1}^{(1)}, \quad n \geq 1. \tag{6.48}$$

Note that equations (6.47) and (6.48) rederive their one-dimensional counterparts (4.18) and (4.23) given in Chapter 4 (cf. Section 4.3).

6.6 Multiphasic mRNA lifetime

In this section, we consider that mRNA molecules posses $K > 2$ stages of their lifetime, where the transition rates correspond to the ageing of an mRNA molecule. We can

represent this process, which we refer to hereafter as *the multiphasic model*, by the following reaction scheme

$$\begin{aligned}
&\emptyset \xrightarrow{\lambda^m} \mathrm{mRNA_1} \xrightarrow{K\gamma^m_{\mathrm{eff}}} \mathrm{mRNA_2} \xrightarrow{K\gamma^m_{\mathrm{eff}}} \cdots \xrightarrow{K\gamma^m_{\mathrm{eff}}} \mathrm{mRNA_K} \xrightarrow{K\gamma^m_{\mathrm{eff}}} \emptyset, \\
&\mathrm{mRNA_i} \xrightarrow{\lambda^p} \mathrm{mRNA_i} + \mathrm{protein}, \quad i = 1, \ldots, K, \\
&\mathrm{protein} \xrightarrow{\gamma^p} \emptyset.
\end{aligned} \tag{6.49}$$

By (6.49), there are K stages of an mRNA's molecule lifetime, each of which lasts $1/K\gamma^m_{\mathrm{eff}}$ on average. The total mRNA lifetime is then $1/\gamma^m_{\mathrm{eff}}$; γ^m_{eff} is thereby interpreted as the effective mRNA decay rate.

The multiphasic model (6.49) is obtained by making the following choices in the general model statement (6.1):

$$\lambda^m_i = \begin{cases} \lambda^m & \text{for } i = 1, \\ 0 & \text{for } i \neq 1, \end{cases} \tag{6.50}$$

and

$$\gamma^m_i = \begin{cases} K\gamma^m_{\mathrm{eff}} & \text{if } i = K, \\ 0 & \text{otherwise.} \end{cases} \tag{6.51}$$

The transition matrix $\mathbf{Q}$ (6.21) for the multiphasic model takes the form of

$$\mathbf{Q} = K\gamma^m_{\mathrm{eff}} \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & 0 \end{pmatrix}, \tag{6.52}$$

and the matrix $\mathbf{A}$ (6.20) is given by

$$\mathbf{A} = K\gamma^m_{\mathrm{eff}} \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 1 \end{pmatrix}. \tag{6.53}$$

Inserting (6.53) and (6.52) into (6.23), we obtain the system of recurrence equations

$$K\gamma_{\text{eff}}^m \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & \ddots & & \\ & & \ddots & 1 & \\ & & & -1 & 1 \end{pmatrix} \begin{pmatrix} b_0^{(1)} \\ b_0^{(2)} \\ \vdots \\ b_0^{(i)} \\ \vdots \\ b_0^{(K)} \end{pmatrix} = \begin{pmatrix} \lambda^m \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{6.54}$$

from which, upon taking the i-th row of (6.54) and solving the recursive equations

$$-K\gamma_{\text{eff}}^m b_0^{(i-1)} + K\gamma_{\text{eff}}^m b_0^{(i)} = 0, \quad \text{for } 2 \leq i \leq K, \tag{6.55}$$

where $b_0^{(1)} = \lambda^m / K\gamma_{\text{eff}}^m$, we recover

$$b_0^{(i)} = \frac{\lambda^m}{K\gamma_{\text{eff}}^m}. \tag{6.56}$$

Formula (6.56) gives the mean of mRNA molecule in the i-th state of its lifetime. Note that the matrix $\mathbf{B}$ (6.22) takes the form of $\mathbf{B} = \lambda^p \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Having found the first moments (i.e. means) (6.56), we then determine the second moments. Taking $n = 1$ in (6.18), we have

$$\begin{pmatrix} K\gamma_{\text{eff}}^m + \gamma^p & & & & \\ -K\gamma_{\text{eff}}^m & K\gamma_{\text{eff}}^m + \gamma^p & & & \\ & -K\gamma_{\text{eff}}^m & \ddots & & \\ & & \ddots & K\gamma_{\text{eff}}^m + \gamma^p & \\ & & & -K\gamma_{\text{eff}}^m & K\gamma_{\text{eff}}^m + \gamma^p \end{pmatrix} \begin{pmatrix} b_1^{(1)} \\ b_1^{(2)} \\ \vdots \\ b_1^{(i)} \\ \vdots \\ b_1^{(K)} \end{pmatrix} = \frac{\lambda^p \lambda^m}{K\gamma_{\text{eff}}^m} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{6.57}$$

from which we obtain the first term of the sequence $b_1^{(i)}$ as

$$b_1^{(1)} := u = \frac{\lambda^p \lambda^m}{K\gamma_{\text{eff}}^m (K\gamma_{\text{eff}}^m + \gamma^p)}. \tag{6.58}$$

Equation (6.57) implies that

$$-K\gamma_{\text{eff}}^m b_1^{(i-1)} + (K\gamma_{\text{eff}}^m + \gamma^p) b_1^{(i)} = \frac{\lambda^p \lambda^m}{K\gamma_{\text{eff}}^m}, \quad \text{for } 2 \leq i \leq K, \tag{6.59}$$

which can equivalently be rewritten as

$$b_1^{(i)} = u + v b_1^{(i-1)}, \quad 2 \leq i \leq K, \tag{6.60}$$

where we set

$$v = \frac{K\gamma_{\text{eff}}^m}{K\gamma_{\text{eff}}^m + \gamma^p} \tag{6.61}$$

for simplicity. Combining (6.60) and (6.58), we obtain

$$b_1^{(i)} = \frac{u}{1-v} + v^{i-1}\left(u - \frac{u}{1-v}\right), \quad 1 \leq i \leq K, \tag{6.62}$$

from which all the elements of $b_1^{(i)}$ (thereby the second moments) can be iteratively obtained. It is worth noting that one can derive higher moments using formula (6.18), but we limit our study to the first two moments.

Next, we focus on calculating the first two terms of the sequence a_n (6.24). Setting $n = 1, 2$ in (6.24) and inserting (6.56) and (6.62) into the resulting equations, respectively, we get

$$a_1 = \frac{\lambda^p \lambda^m}{\gamma^p \gamma_{\text{eff}}^m} \quad \text{and} \quad a_2 = \frac{\lambda^p u\left(K + v\left(-1 - K + v^K\right)\right)}{2\gamma^p(1-v)^2}. \tag{6.63}$$

Having found the first two terms of a_n, we are now ready to calculate the Fano factor. Inserting (6.63) into (6.37), and substituting (6.58) and (6.61) into the resulting expression yields

$$\mathrm{F_m} = 1 + \frac{\lambda^p}{\gamma^p}\left(1 + \frac{\gamma_{\text{eff}}^m}{\gamma^p}\left(-1 + \left(\frac{K\gamma_{\text{eff}}^m}{K\gamma_{\text{eff}}^m + \gamma^p}\right)^K\right)\right), \tag{6.64}$$

where $\mathrm{F_m}$ stands for the multiphasic Fano factor.

6.7 Conclusion

In this chapter, we formulated and analysed a structuration of the two-stage gene expression model that considers multiple mRNA states. We demonstrated that this generalised model (in the sense of having multiple mRNA states) can be used to capture the dynamics of simpler models such as the inactivation loop model and the multiphasic mRNA model, which were analysed in detail as a particular interest of this chapter.

We introduced the generalised model and formulated its mathematical description via the CME. Next, we focused on seeking a solution to the corresponding partial differential equation obtained by transforming the CME using the generating function approach. A suitable ansatz was employed for converting

the PDE to a system of ODEs. Subsequently, using the power series method, we sought a solution to the ODE system, which is then expressed in matrix form as a system of recurrence equations. We recovered the generating function of the stationary distribution of mRNA and protein amounts by means of the coefficients of power series, which were obtained by solving the recurrence relations under the initial conditions.

Furthermore, the sought-after solution was then used to characterise the marginal protein and mRNA distributions. To determine the protein distribution, we used the factorial moments calculated from the factorial cumulants. Additionally, we demonstrated that the mRNA distributions are Poissonian; therefore, we derived the protein mean and Fano factor and thus expressed it in terms of the first two factorial moments. We then provided two different examples to which the generalised model and its results can be applied.

The first example was the two-stage gene-expression model extended by an mRNA inactivation loop. We demonstrated that under a suitable choice of parameters, the generalised model and its results can be used to recover the expressions for the extended model. In particular, we rederived the formulae for the cumulants and showed that they agree with those obtained for the extended model.

As a second example, by making suitable parameter choices in the generalised model, we presented the multiphasic model in which an mRNA molecule is assumed to be transitioning through its lifetime stages. The solution formula derived previously for the generalised model and the associated matrices (e.g. the transition matrix) for this model were used to determine, upon solving the recursive equations, the first two moments of mRNA distributions, allowing us to calculate the Fano factor for the multiphasic model.

In summary, this chapter provides a systematic mathematical breakdown for protein and mRNA distributions in a structured gene expression model, which takes into account multiple mRNA states. We believe that the model and its results can be used in understanding the dynamics of underlying biochemical processes.

Chapter 7

Conclusion

Chemical reaction networks describing many biological processes require a rigorous mathematical formulation of the underlying model. One of such mechanisms is the stochastic gene-expression process, which describes the production of gene products such as proteins. Due to the inherent occurrence of chemical reactions, the number of species involved varies over time, leading to an interest in quantifying it. However, this task becomes more and more challenging as the number of species in the system of interest increases. From a mathematical point of view, simple models remain inadequate to elucidate the problem, giving rise to the study of more complex models that account for multivariate dynamics. In this thesis, we have addressed and studied such models of stochastic gene expression.

First, we reviewed fundamental mathematical concepts and methods common in biochemical reaction systems. These include general reaction kinetics, whose deterministic and stochastic description is given by a system of ordinary differential equations and the chemical master equation, respectively. As a standard technique for solving the CME, we presented the generating function method. In particular, we introduced the basic two-stage model of gene expression, providing its corresponding ODE system and chemical master equation. Additionally, we provided the stochastic simulation algorithm for obtaining sample time trajectories of species in a system.

We next focused on a specific multiclass–multistage model that considers complex lifetime pathways for a self-regulating transcription factor. We have shown that the one-dimensional and the structured models exhibit the same stationary protein distribution in the non-bursty regime. Our theoretical results were cross-validated by performing stochastic simulations.

Motivated by relevant studies in the literature, we have extended the basic

two-stage gene expression model by including an mRNA inactivation loop. We provided a comprehensive mathematical analysis in both deterministic and stochastic frameworks; obtained the statistical measures, which we then used to quantify the noise. Consequently, our noise analysis demonstrated that the inclusion of an mRNA inactivation loop into the classical model reduces the protein noise. In what follows, we have generalised the extended model to the one that considers multiple mRNA states. We have shown that the generalised model can govern the dynamics of simpler models, such as the extended and the multiphasic model. Notably, our results provide non-trivial expressions for the steady-state protein distribution.

In summary, we think that the models and their extensive mathematical analysis studied in this thesis will contribute to understanding the dynamics of biochemical mechanisms in relevant research fields.

Bibliography

[1] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[2] 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.

[3] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[4] 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.

[5] 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.

[6] Saurabh Modi, Supravat Dey, and Abhyudai Singh. Noise suppression in stochastic genetic circuits using pid controllers. *PLOS Computational Biology*, 17(7):1–25, 07 2021.

[7] Carlus Deneke, Reinhard Lipowsky, and Angelo Valleriani. Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger rna. *PLOS ONE*, 8(2):1–12, 02 2013.

[8] Erik McShane, Celine Sin, Henrik Zauber, Jonathan N. Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, Wei Chen, Zuzana Storchova, Joseph A. Marsh, Angelo Valleriani, and Matthias Selbach. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell*, 167(3):803–815.e21, October 2016.

[9] Radek Erban, Jonathan Chapman, and Philip Maini. A practical guide to stochastic simulations of reaction-diffusion processes. *arXiv preprint arXiv:0704.1908*, 2007.

[10] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, August 2011.

[11] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[12] Donald Gross, John F. Shortle, James M. Thompson, and Carl M. Harris. *Fundamentals of Queueing Theory*. Wiley-Interscience, USA, 4th edition, 2008.

[13] U. Narayan Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhäuser, 2015.

[14] Crispin Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer-Verlag, Berlin Heidelberg, 4 edition, 2009.

[15] Norman L Johnson, Samuel Kotz, and Adrienne W Kemp. *Univariate discrete distributions*. John Wiley & Sons, 2005.

[16] Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, 121(9):4059–4067, 2004.

[17] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889, 2000.

[18] John H. Abel, Brian Drawert, Andreas Hellander, and Linda R. Petzold. Gillespy: A python package for stochastic model building and simulation. *IEEE Life Sciences Letters*, 2(3):35–38, 2016.

[19] David J. Warne, Ruth E. Baker, and Matthew J. Simpson. Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of The Royal Society Interface*, 16(151):20180943, February 2019.

[20] Ramon Grima. Construction and accuracy of partial differential equation approximations to the chemical master equation. *Physical Review E*, 84(5):056109, November 2011.

[21] 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.

[22] Pavol Bokes, John R. King, Andrew T. A. Wood, and Matthew Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of Mathematical Biology*, 64(5):829–854, April 2012.

[23] Milton Abramowitz, Irene A. Stegun, and Robert H. Romer. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. *American Journal of Physics*, 56(10):958–958, October 1988.

[24] Candan Çelik, Pavol Bokes, and Abhyudai Singh. Stationary distributions and metastable behaviour for self-regulating proteins with general lifetime distributions. In Alessandro Abate, Tatjana Petrov, and Verena Wolf, editors, *Computational Methods in Systems Biology*, pages 27–43, Cham, 2020. Springer International Publishing.

[25] William J. Blake, Mads KÆrn, Charles R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, April 2003.

[26] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science (New York, N.Y.)*, 329:533–538, July 2010.

[27] Eugenio Cinquemani. Identifiability and reconstruction of biochemical reaction networks from population snapshot data. *Processes*, 6(9), 2018.

[28] Mahendra Kumar Prajapat and Andre S. Ribeiro. Added value of autoregulation and multi-step kinetics of transcription initiation. *Royal Society Open Science*, 5(11):181170, November 2018.

[29] Calin Guet, Thomas A. Henzinger, Claudia Igler, Tatjana Petrov, and Ali Sezgin. Transient memory in gene regulation. In Luca Bortolussi and Guido Sanguinetti, editors, *Computational Methods in Systems Biology*, pages 155–187, Cham, 2019. Springer International Publishing.

[30] Eugenio Cinquemani. Stochastic reaction networks with input processes: Analysis and application to gene expression inference. *Automatica*, 101:150–156, 2019.

[31] Luca Bortolussi, Roberta Lanciani, and Laura Nenzi. Model checking markov population models by stochastic approximations. *Information and Computation*, 262:189–220, 2018. GandALF 2016.

[32] Michael Backenköhler, Luca Bortolussi, and Verena Wolf. Control variates for stochastic simulation of chemical reaction networks. In Luca Bortolussi and Guido Sanguinetti, editors, *Computational Methods in Systems Biology*, pages 42–59, Cham, 2019. Springer International Publishing.

[33] Alexander Andreychenko, Luca Bortolussi, Ramon Grima, Philipp Thomas, and Verena Wolf. *Distribution Approximations for the Chemical Master Equation: Comparison of the Method of Moments and the System Size Expansion*, pages 39–66. Springer International Publishing, Cham, 2017.

[34] Pavol Bokes, Michal Hojcka, and Abhyudai Singh. Buffering gene expression noise by microrna based feedforward regulation. In Milan Češka and David Šafránek, editors, *Computational Methods in Systems Biology*, pages 129–145, Cham, 2018. Springer International Publishing.

[35] Guilherme C. P. Innocentini, Fernando Antoneli, Arran Hodgkinson, and Ovidiu Radulescu. Effective computational methods for hybrid stochastic gene networks. In Luca Bortolussi and Guido Sanguinetti, editors, *Computational Methods in Systems Biology*, pages 60–77, Cham, 2019. Springer International Publishing.

[36] Guilherme C. P. Innocentini, Arran Hodgkinson, and Ovidiu Radulescu. Time dependent stochastic mrna and protein synthesis in piecewise-deterministic models of gene networks. *Frontiers in Physics*, 6, 2018.

[37] Pavel Kurasov, Alexander Lück, Delio Mugnolo, and Verena Wolf. Stochastic hybrid models of gene regulatory networks – a pde approach. *Mathematical Biosciences*, 305:170–177, 2018.

[38] Michalis Michaelides, Jane Hillston, and Guido Sanguinetti. Statistical abstraction for multi-scale spatio-temporal systems. In Nathalie Bertrand and Luca Bortolussi, editors, *Quantitative Evaluation of Systems*, pages 243–258, Cham, 2017. Springer International Publishing.

[39] Michalis Michaelides, Jane Hillston, and Guido Sanguinetti. Geometric fluid approximation for general continuous-time Markov chains. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2229):20190100, September 2019.

[40] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[41] Dmitri Bratsun, Dmitri Volfson, Lev S. Tsimring, and Jeff Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences*, 102(41):14593–14598, 2005.

[42] Jacek Miękisz, Jan Poleszczuk, Marek Bodnar, and Urszula Foryś. Stochastic models of gene expression with delayed degradation. *Bulletin of Mathematical Biology*, 73(9):2231–2247, September 2011.

[43] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, Boca Raton, 2 edition, August 2019.

[44] Mohammad Soltani, Cesar A. Vargas-Garcia, Duarte Antunes, and Abhyudai Singh. Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLOS Computational Biology*, 12(8):1–23, 08 2016.

[45] Svenja Lagershausen. *Performance analysis of closed queueing networks*. Lecture Notes in Economics and Mathematical Systems. Springer, 2013.

[46] Tao Jia and Rahul V. Kulkarni. Intrinsic Noise in Stochastic Models of Gene Expression with Molecular Memory and Bursting. *Physical Review Letters*, 106(5):058102, February 2011.

[47] L. Liu, B. R. K. Kashyap, and J. G. C. Templeton. On the GIX/G/∞ system. *Journal of Applied Probability*, 27(3):671–683, September 1990.

[48] Pavol Bokes, Alessandro Borri, Pasquale Palumbo, and Abhyudai Singh. Mixture distributions in a stochastic gene expression model with delayed feedback: a WKB approximation approach. *Journal of Mathematical Biology*, 81(1):343–367, July 2020.

[49] Satya Swarup Samal, Jeyashree Krishnan, Ali Hadizadeh Esfahani, Christoph Lüders, Andreas Weber, and Ovidiu Radulescu. *Metastable Regimes and Tipping Points of Biochemical Networks with Potential Applications in Precision Medicine*, pages 269–295. Springer International Publishing, Cham, 2019.

[50] Jay M. Newby and Jonathan Chapman. Metastable behavior in markov processes with internal states. *Journal of Mathematical Biology*, 69:941–976, 2014.

[51] David G. Kendall. Stochastic Processes and Population Growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282, 1949.

[52] James R Norris. *Markov chains*. Cambridge university press, 1998.

[53] Long Cai, Nir Friedman, and X Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[54] 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.

[55] Attila Becskei, Bertrand Séraphin, and Luis Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO Journal*, 20(10):2528–2535, May 2001.

[56] J.S. Griffith. Mathematics of cellular control processes ii. positive feedback to one gene. *Journal of Theoretical Biology*, 20(2):209–216, 1968.

[57] Nir Friedman, Long Cai, and X. Sunney Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters*, 97(16):168302, October 2006.

[58] Pavol Bokes, Yen Ting Lin, and Abhyudai Singh. High Cooperativity in Negative Feedback can Amplify Noisy Gene Expression. *Bulletin of Mathematical Biology*, 80(7):1871–1899, July 2018.

[59] Pavol Bokes and Abhyudai Singh. Controlling noisy expression through auto regulation of burst frequency and protein stability. In Milan Češka and Nicola Paoletti, editors, *Hybrid Systems Biology*, pages 80–97, Cham, 2019. Springer International Publishing.

[60] 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.

[61] Ioannis Lestas, Johan Paulsson, Nicholas E Ross, and Glenn Vinnicombe. Noise in gene regulatory networks. *IEEE Transactions on Automatic Control*, 53(Special Issue):189–200, 2008.

[62] Pavol Bokes. Postponing production exponentially enhances the molecular memory of a stochastic switch. *European Journal of Applied Mathematics*, 33(1):182–199, February 2022.

[63] Robert Hinch and S. Jon Chapman. Exponentially slow transitions on a Markov chain: the frequency of Calcium Sparks. *European Journal of Applied Mathematics*, 16(4):427–446, August 2005.

[64] Carlos Escudero and Alex Kamenev. Switching rates of multistep reactions. *Physical Review E*, 79(4):041149, April 2009.

[65] Peter Hanggi, Hermann Grabert, Peter Talkner, and Harry Thomas. Bistable systems: Master equation versus Fokker-Planck modeling. *Physical Review A*, 29(1):371–378, January 1984.

[66] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[67] James R. Jackson. Jobshop-like Queueing Systems. *Management Science*, 10(1):131–142, 1963.

[68] Candan Çelik, Pavol Bokes, and Abhyudai Singh. Protein noise and distribution in a two-stage gene-expression model extended by an mrna inactivation loop. In Eugenio Cinquemani and Loïc Paulevé, editors, *Computational Methods in Systems Biology*, pages 215–229, Cham, 2021. Springer International Publishing.

[69] 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.

[70] J. Peccoud and B. Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[71] Hodjat Pendar, Thierry Platini, and Rahul V. Kulkarni. Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes. *Physical Review E*, 87(4):042720, April 2013.

[72] Vahid Shahrezaei and Peter S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, December 2008.

[73] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, April 2012.

[74] Jonathan M. Raser and Erin K. O'Shea. Noise in Gene Expression: Origins, Consequences, and Control. *Science*, 309(5743):2010–2013, September 2005.

[75] Alvaro Sanchez, Sandeep Choubey, and Jane Kondev. Regulation of noise in gene expression. *Annual Review of Biophysics*, 42:469–491, 2013.

[76] Roy D. Dar, Sydney M. Shaffer, Abhyudai Singh, Brandon S. Razooky, Michael L. Simpson, Arjun Raj, and Leor S. Weinberger. Transcriptional bursting explains the noise–versus–mean relationship in mrna and protein levels. *PLOS ONE*, 11(7):1–5, 07 2016.

[77] LaTasha C. R. Fraser, Ryan J. Dikdan, Supravat Dey, Abhyudai Singh, and Sanjay Tyagi. Reduction in gene expression noise by targeted increase in accessibility at gene loci. *Proceedings of the National Academy of Sciences*, 118(42):e2018640118, 2021.

[78] 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, 2021.

[79] 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, October 2002.

[80] Philipp Thomas. Intrinsic and extrinsic noise of gene expression in lineage trees. *Scientific Reports*, 9(1):474, January 2019.

[81] Estelle Dacheux, Naglis Malys, Xiang Meng, Vinoy Ramachandran, Pedro Mendes, and John E. G. McCarthy. Translation initiation events on structured eukaryotic mRNAs generate gene expression noise. *Nucleic Acids Research*, 45(11):6981–6992, June 2017.

[82] Pavol Bokes, John R. King, Andrew T. A. Wood, and Matthew Loose. Transcriptional bursting diversifies the behaviour of a toggle switch: hybrid simulation of stochastic gene expression. *Bulletin of mathematical biology*, 75(2):351–371, 2013.

[83] Pavel Kurasov, Delio Mugnolo, and Verena Wolf. Analytic solutions for stochastic hybrid models of gene regulatory networks. *Journal of Mathematical Biology*, 82(1):1–29, 2021.

[84] Abhyudai Singh and Joao P. Hespanha. Stochastic hybrid systems for studying biochemical processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1930):4995–5011, 2010.

[85] Jingwei Li, Hao Ge, and Yunxin Zhang. Fluctuating-rate model with multiple gene states. *Journal of Mathematical Biology*, 81(4):1099–1141, 2020.

[86] Tianshou Zhou and Tuoqi Liu. Quantitative analysis of gene expression systems. *Quantitative Biology*, 3(4):168–181, 2015.

[87] María Rodríguez Martínez, Jordi Soriano, Tsvi Tlusty, Yitzhak Pilpel, and Itay Furman. Messenger rna fluctuations and regulatory rnas shape the dynamics of a negative feedback loop. *Phys. Rev. E*, 81:031924, Mar 2010.

[88] Abhyudai Singh and João P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control*, 56(2):414–418, 2011.

[89] Mohammad Soltani, Cesar Augusto Vargas-Garcia, and Abhyudai Singh. Conditional moment closure schemes for studying stochastic dynamics of genetic circuits. *IEEE Transactions on Biomedical Circuits and Systems*, 9(4):518–526, 2015.

[90] 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, October 2006.

[91] Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, January 2004.

[92] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[93] Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics*, 152(14):144106, April 2020.