Stochastic compartmental models and $CD8^+$ T cell exhaustion

Flavia Feliciangeli

Department of Applied Mathematics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

July, 2023

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Joint Publications

Almost all of the work in Chapter 3 has been refereed and published, as follows:

- Feliciangeli F., Dreiwi H., López-García M., Castro Ponce M., Molina-París C. and Lythe G., (2022), Why are cell populations replenished via intermediate compartments?, *J. R. Soc. Interface*, https://doi.org/10.1098/rsif.2022.0629

Almost all of the work in Chapter 5 has been refereed and submitted for publication, as follows:

- Dreiwi H., Feliciangeli F., Castro Ponce M., Lythe G., Molina-París C. and López-García M., (2023), Stochastic journeys of cell progenies through compartments: self-renewal, symmetric and asymmetric division, *submitted to Scientific Reports*, a preliminary version is also available at https://doi.org/10.48550/arXiv.2111.11753

Preface

My PhD at the University of Leeds comes as part of the QuanTII (Quantitative T cell Immunology and Immunotherapy) ITN European network funded by the Horizon 2020 programme of the European Union (No.764698, https://quantii.leeds.ac.uk/). This training and research H2020 network has both academic and non-academic beneficiaries and I have been hosted, since the beginning of my PhD, by Bayer AG with headquarters in Leverkusen (Germany). As an early-stage researcher of QuanTII, my research project was focused on T cell immunotherapies and, in particular, on developing new mathematical models of T cell exhaustion during chronic viral infection or cancer; indeed, a deeper understanding of this biological process might bring to the development of novel treatments and relevant for immunotherapies.

After the first virtual onboarding period, I have been visiting the University of Leeds for three months during which I met my professors in person and attended required modules. During this period, Professor Carmen Molina-París and I started a collaboration with Professor Susan Kaech and a post-doctoral scientist in her group, Dr. Thomas Mann, both at the Salk Institute for Biological Studies in La Jolla, California (US). Kaech Lab aims to shed light into the mechanisms of development of immunological memory and T cell exhaustion as well as to enhance the efficacy of immunotherapies, including checkpoint blockade. After months of virtual collaboration, I obtained a scholarship to have the opportunity to visit Kaech Lab and Dr. Mann. The *in vitro* experiments Dr. Mann carried out were based on a novel optogenetic technique and were intended to stimulate CAR-T cells and analyse different emerging populations from a "functional" to an "exhausted" phenotype. The aim of the visit, planned for summer 2020, was to

enhance the mathematical and computational models of $CD8^+$ T cells differentiation during exhaustion that were set up during the virtual collaboration. Model inference and parameter estimation should have been based on Dr. Mann experimental measurements. This was considered, as part of QuanTII secondment, one of the requirements of the network to ensure PhD student to receive interdisciplinary training in the fields relevant for their research. This work would have constituted the main content of my PhD thesis. However, the originally planned experiments revealed the complexity of the exhaustion landscape. Specificity of population properties and experimental challenges, that were unforeseeable before, came up and the focus of the experiment has been then refined to investigate a signalling cascade upon cell activation first. Moreover, the SARS-CoV-2 pandemic hit only few months before my arranged visit; the laboratory was closed for several weeks and the experiment delayed. Nevertheless, the collaboration has been fostered by regular weekly virtual meetings to push an exchange of both biological and mathematical knowledge. This work, carried out during the first year of my PhD, constitutes the content of Chapter 6. Importantly, I want to underline that the mathematical techniques that I learnt with the idea of applying them to model $CD8^+$ T cell exhaustion process, in collaboration with the Salk Institute, could be also applied to parameterise different deterministic mathematical models.

In addition, I was offered a research visiting status by CNLS (Center for Non Linear Studies in the Theoretical Division) at Los Alamos National Laboratory in New Mexico (USA), where my Professor C. Molina-París currently is. Unfortunately, also this visit was postponed given the Presidential Proclamation and travel ban during the COVID-19 pandemic. However, the collaboration has been fostered by regular weekly virtual meetings; this was precious for the conceptualisation and supervision of the work published in Feliciangeli *et al.* (2022). This work of a more theoretical nature, and reported in Chapter 3, has been particularly pushed by my supervisors and I during the SARS-CoV-2 pandemic and travel ban. From this initial work, I started further analysis to consider time dependency or different biological transitions. These further studies are reported in Chapters 4 and 5 which is published in Dreiwi *et al.* (2021). This explains why

my PhD thesis contains mathematical models of such different mathematical nature: from a stochastic, more theoretical model to a deterministic one applied to experimental data.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764698.

First and foremost, I thank my supervisors Grant Lythe, Martín López-García, Carmen Molina-París from the University of Leeds for their constant support and source of inspiration during my PhD studies. I will be always impressed with your enthusiasm for mathematical biology you shared with me and that I will try to make it mine. I also want to thank them for their patience and guidance and for closely following me even remotely throughout these four years. Also, I would like to thank my industrial supervisors Christoph Niederalt and Jan Schlender from Bayer AG for constantly supporting me and sharing their experience

I also thank Thomas Mann from Kaech Lab at the Salk Institute for Biological Studies. I am extremely grateful to him for huge amount of time he spent explaining me biology of cellular immunology and basis of flow cytometry. I will never forget the hours we spent discussing the state of the art of T cell immunology and how to include those concepts in a mathematical model. Our virtual meetings were always stimulating and I will not be where I am now without them.

Thank you also to all the other ESRs from QuanTII. This network provided me many insights on different and interesting research topics; also, it gave me the opportunity to be in contact with inspiring scientist and to understand the importance of public outreach sessions. Finally, I would like to thank Jessica Brennan, QuanTII administrator, for her incredible support and for organising the QuanTII meetings.

Finally, I give thanks to my family and Matteo who have seen me in very different moods (from crazy, to happy, to depressed, to excited) and really always supported me throughout my studies, and without whom this work would not have been possible. Lastly, I want to thank my friends spread somewhere around Italy and Germany. With a special thank to Iva, Francesca, le Pulzelle, Dario and Marta.

Abstract

In this PhD thesis, mathematical models for cell differentiation are presented. Cell differentiation is a widely observed process in cellular biology allowing a small pool of not specialised cells to develop and maintain a bigger population of cells with a specific function. Different mathematical techniques are employed in this thesis, to study cell differentiation process.

We propose a time-independent stochastic mathematical model to represent a general differentiation process via a sequence of compartments. Since we are interested in the ultimate fate of the system, we define a discrete-time branching processes and consider the impact, on the final population, of cells passing through only one or multiple compartments. Further, we include time dependency and define a continuous-time Markov chain to analyse cells dynamics along the sequence of compartments over time. Also, we focus on the journey of a single cell over time and compute a number of summary statistics of interest. Moreover, the impact of different types of differentiation events is considered and numerical results inspired by biological applications, mainly related to immunology, are summarised to illustrate our theoretical approach and methods.

In the last Chapter, we focus on a specific cell differentiation process: cells of the immune system have been observed to differentiate towards a dysfunctional state, called *exhaustion*, during a chronic infection or cancer. One of the aims of this PhD thesis is to shed light into the exhaustion-differentiation process of $CD8^+$ T cells and its reversibility which is a topic of interest for the current and future development of immunotherapies. In particular, based on data collected by the Kaech Lab, several deterministic mathematical models are defined to investigate cells' trajectory towards the exhausted state as well as the duration of the antigen signal at early time point of stimulation.

Contents

CONTENTS

List of Figures

LIST OF FIGURES

LIST OF FIGURES

Chapter 1

Biological introduction

Humans, animals, plants, and even fungi are multicellular organisms made up of several types of cells specialised for various functions and working together to perform tasks and maintain the body in a steady state of internal conditions, called homeostasis. All cells have the same DNA (genetic information), and are only the variations in their gene expression (which genes are turned on or off) that cause cells to acquire different functions. *Cell differentiation* is the process, controlled through changes in gene expression, of cells becoming more specialised.

At the beginning of the differentiation process, cells retain *pluripotency* as the potential to become any cell type; this is slowly lost as cells differentiate gaining specialised functions. *Stem cells* are special human cells retaining the potential to develop into many different types of cells. This term, however, has a specific meaning in biology and stem cells, categorised into embryonic and adult stem cells, are a clearly defined type of cells. A terminology inspired by this specific cell type is nevertheless commonly applied even to other differentiation processes. So, generally speaking we can say that *stem-like* cells are able to differentiate into multiple cells of a restricted lineage and, thus, give rise to more matured cells performing specific functions. Within the human body there are a large number of differentiation processes occurring, wherever specialised cells are generated. Mathematical models of cell differentiation can be defined appropriately for different types of differentiation process and, then, used to explain biological data. Moreover, by fitting the model to experimental data, one can shed light into cell behaviour relating cellular rates of more progenitor or mature cells, as well as

into differentiation timelines or to related pathological conditions. Examples of different types of differentiation, analysed from a mathematical perspective, are: keratinocyte differentiation and psoriasis pathogenesis (Zhang *et al.*, 2015), differentiation of cells of the human gastrointestinal tract (Murray *et al.*, 2011), blood cell differentiation and hematopoiesis (Barile *et al.*, 2020; Busch *et al.*, 2015), cancer differentiation as in chronic myeloid leukaemia (Michor *et al.*, 2005), T cell development and thymus selection (Sawicka *et al.*, 2014; Yates, 2014) or the T cell dysfunctional exhaustion process, detailed in this PhD thesis.

A differentiation process is not restricted to different cell types, but it might involve different spatial locations as well, such as in the keratinocyte differentiation process. The epidermis, the outermost layer of skin, is organised into a stratified structure of cells, called keratinocytes, characterised by several differentiated stages. During its lifespan, a keratinocyte spatially transits from a more inner to an outer stratum, experiencing several biochemical and morphological changes (Montagna *et al.*, 1992). Similarly, for the colonic cell differentiation process. The endothelial tissue of the human gastrointestinal tract consists of several cell types with different functions, from an outermost layer that acts as a protective barrier and absorbs nutrients to an inner one of proliferative stem-like cells. In their model, Murray *et al.* (2011) aim to investigate cell proliferation and movement of cells in the intestinal crypt (invaginations typical of the colonic tract).

Moreover, a same differentiation hierarchy can apply to both a healthy and cancerous system, showing that a cancer population of cells might arise from cell mutations generating a subpopulation with different cellular events rates. This is the case considered in Michor *et al.* (2005), where a mathematical model representing stem-like cells, progenitors, differentiated and terminally differentiated cells is defined. If this hierarchy applies to both normal and leukaemic cells, differences arise from leukaemic stem-like cells highly dividing by self-renewal.

In many of these examples, the immune system is somehow involved and considered. This is a complex system composed by a variety of different cell types and organs harmonised and working together to ensure our health and fights external dangerous agents. A key role is played by the white blood cells (also known as leukocytes) that, as part of the immune system, circulate in the

blood, locate the site of an infection and defend the body against infections. T cell, also called T lymphocyte, are a specific type of leukocyte that originate in the bone marrow and mature in the thymus. In particular, $CD8^+$ T cells, also known as *killer T cells*, are immune cells that are specialised in direct killing external agents. During cancer or chronic infections, $CD8^+$ T cells progressively differentiate into a dysfunctional state, called *exhausted state*, leading to cancer or pathogen immune evasion (Janeway *et al.*, 1999).

A differentiation process, arising from diverse stimuli both in space and time, is not always an irreversible process: in some cases, specific set of genes can be reprogrammed so that the cell is back to a more primitive state. This is the hope for the dysfunctional process of exhaustion. Indeed, reversing T cell exhaustion would have the potential of restoring an effector immune response; thus, several cancer immunotherapies are directed to reinvigorate exhausted $CD8^+$ T cells. However, on the one side, the molecular and cellular mechanisms which drive T cell exhaustion are not clearly identified; on the other side, it has not clearly determined if, and when, a *point of no return* for T cell exhaustion-differentiation pathway exists as well as if there is only, or more, subsets of exhausted T cells that positively respond to immunotherapy. The partial understanding of T cell exhaustion process leads to drawbacks within cancer immunotherapy. Despite important improvements in cancer treatment, disease relapse eventually occurs in many patients, with higher occurrence in specific tumour type and stage, and according to patient conditions. Thus, a better understanding of the process of exhaustion development as well as of the impact of immunotherapies on this dysfunctional state, would be beneficial for enhancing patient survival (Barber *et al.*, 2006; Chow *et al.*, 2022; Hashimoto *et al.*, 2018). Parallel to this, a new treatment arose: transfer into the patient of autologous (from patient itself) T cells engineered with a chimeric antigen receptor (CAR T cells) directed against specific tumour markers. Few studies have been performed on exhausted CARs, but this could be of crucial importance to improve the efficacy of novel CAR T cells treatments. Fundamental questions remain still open about how and why T cells become dysfunctional under a persistent stimulation, such as in cancer and chronic infections. Moreover, the antigen load, the degree of exhaustion as

well as the type of infection seem to impact the reversing of exhausted T cells to a functional effector state.

Within this PhD thesis we particularly focus on immune processes in the context of T cell development and $CD8^+$ T cell exhaustion-differentiation process (towards a dysfunctional state). In particular, this Chapter comes as a review of the biological literature relevant for this PhD project. In Section 1.1 we outline the main concepts about immune system activation and, in particular, about the role of $CD8^+$ T cells; here, biological details about the thymocytes selection process are also given. In Section 1.2 we focus on the dysfunctional state of exhaustion, its reversibility and its drawbacks in cancer immunology. Some of the aforementioned types of differentiation process are considered as an application of theoretical techniques in this PhD thesis; for example, the developed mathematical techniques are illustrated with a case study about hematopoiesis (see Section 5.2.2). Deterministic mathematical models for $CD8^+$ T cell exhaustion process are defined in Chapter 6 based on the collaboration with Prof. Kaech of the Salk Institute for Biological Studies (California, USA). Moreover, the reader can find a mathematical application of the thymocytes selection process in Sections 5.2.3 and 3.5. In Chapters 3, 4 and 5 we propose a general stochastic model of cell division, death and differentiation (or migration) across an ordered sequence of compartments that can represent different types of biological differentiation processes.

As mentioned, cell differentiation is driven by changes in gene transcription (gene expression), triggered by cells receiving multiple stimuli from surrounding environment and cells. This results in a complex mechanism and is the subject of ongoing research. Recent studies highlight the role of asymmetric division (Berika *et al.*, 2014) as taking part in cellular differentiation in niche-adjacent cells, where it is more likely that two halves of a cell receive different signals. When cells divides by asymmetric division, the two daughter cells will have different developmental fates. Also, embryonic stem cells are known to divide by asymmetric cell division, so that one of the two daughters will differentiate into a specific lineage and the other cell may continue to remain as a stem cell still dividing asymmetrically. This particular type of cellular division is considered in the mathematical models defined in Chapters 3 and 5.

1.1 The body's security force

1.1.1 The immune system and chronic infection

Viruses and bacteria on Earth are estimated to be more than 10^{31} and they are essentially in every environment (Mushegian, 2020); when eating, breathing and just living, we are constantly exposed to other microorganisms. Fortunately, some of them can even be beneficial and only a few have a negative effect on us, ending up infecting the body and causing diseases. Those are the *pathogens*, defined as microbial organisms that cause diseases to the host (human, animal or plant). Viruses, bacteria, fungi, and parasites are the most common types of pathogens. They can cause a variety of different diseases: from a more to a less severe infection, as well as a short, longer or recurrent infection persisting in our body.

The immune system recognises a pathogen by the expressed antigens, that are molecules on the surface of the pathogen that uniquely identify its type. As soon as the pathogen is recognised, the immune system is triggered and starts producing cells that are able to attack and eliminate the pathogen. At the same time, the immune system starts producing small molecules, called *antibodies*, that bind to the pathogen and allow more types of cells of the immune system to recognise and attack it. This antigen-antibody trigger is the foundation of the immune system specificity; indeed, antigens are expressed either on the surface of pathogens but also on allergens, proteins, tumour cells or normal cells. Generally speaking an antigen is, indeed, a marker (usually a protein) that the immune system recognises as "foreign" or "dangerous" and, thus, it triggers activation of the immune response against it.

Technically, the immune system is divided into two subsystems, the *innate* and the *adaptive* immune system. The innate system is the first line of defence against pathogens. Cells of the innate system are able to recognise antigens and mount a quick response. The adaptive immune system takes longer to mount an effective response upon pathogen recognition but it has evolved to provide a versatile response, thanks to the ability to generate pathogen-specific receptors. In addition, the adaptive immune system provides protection against subsequent re-

infections of the same pathogen. It is, indeed, characterised by an immunological memory and the latent presence of an antigen-specific population of memory immune cells that can rapidly respond upon secondary or subsequent infection. In a subsequent response to the same antigen, previously generated memory cells are rapidly activated leading to a response that is quicker and more effective than the primary one. Dendritic cells, monocytes, macrophages, granulocytes, and natural killer T cells are the cellular components of the innate immune system; whereas, T and B lymphocytes are the cellular elements of the adaptive immune system. The synergy between the adaptive and innate immune systems is critical for the elimination of infectious agents: the flexibility and the immunologic memory of the adaptive immune system relies upon the innate immune system for a proper initiation and direction of antigen responses.

T lymphocytes are antigen-specific cells of the immune system developing in the thymus and having a key role in determining either pathogen clearance or its persistence. During a typical acute infection, a rapid and large pathogen production is observed. The immune system gets activated and is able to efficiently clear the infection in a short time (within around 2 weeks) (Janeway *et al.*, 1999). However, if the primary infection is not cleared and the pathogen continues replicating either latently or in specific niches, the infection persists leading to a chronic infection. In this case, the immune system is constantly activated without reaching pathogen clearance; examples of chronic infections are Herpesvirus, HIV or Hepatitis C (Janeway *et al.*, 1999).

In order to accomplish both short and long-term goals in fighting the pathogen, T cells have developed a specific ability, called *T cell plasticity*: according to the environmental conditions (such as nutrients, oxygen and glucose levels and signals released from surrounding cells), T cells are able to differentiate into several types (Buchholz *et al.*, 2013; Gerlach *et al.*, 2013). Cells' plasticity is higher for less differentiated cells and decreases as cells differentiate further to a more stable terminal differentiation state. This ability allows T cells to give rise, upon infection, to heterogeneous subsets of cells, each characterised by different functions. With the advancement of flow cytometry technology, it was possible to analyse, on one single cell, a greater number of expressed proteins simultaneously; in this way, specific differentiation subsets of an heterogeneous pool of cells can

be identified (Kaech *et al.*, 2002a; Sallusto *et al.*, 1999; Wherry *et al.*, 2003). With the advent of single-cell technologies, it was possible to look at DNA and RNA sequences and thus, analyse the transcriptomes, epigenomes, as well as clonotypes (TCR repertoire) of T cells (Beltra *et al.*, 2020; Chen *et al.*, 2019; Pauken *et al.*, 2016). These technologies open the door to a deconvolution of T cell heterogeneity but, given the broadness of the topic, a full understanding of the undergoing molecular and signalling processes might be still lacking.

In both an acute and chronic infection, subsets of T cells for either a short and long-term control of the pathogen have been observed. However, if in an acute infection the antigen is cleared, the inflammation is resolved and long-term memory cells are generated upon pathogen clearance. In a chronic infection the long-term fate of T cells is a dysfunctional state, usually called *exhausted state* (Chen *et al.*, 2019; Khan *et al.*, 2019; Seo *et al.*, 2019; Thommen *et al.*, 2018). As pathogens, tumour cells are identified as "not-in-host" cells and are targeted by the immune system. Similarly than in a chronic infection, during cancer progression, the antigen stimulation persists and the immune system might fail to efficiently control tumour development (McLane *et al.*, 2019).

As mentioned, T cells can be divided, according to their functionality, into a number of different subsets. For the purpose of this PhD thesis, we focus on reviewing $CD8^+$ T cell with an in-depth look at their exhaustion process.

1.1.2 $CD8^+$ T cells killing force

Cells of the immune system originate in the bone marrow and migrate to different parts of the body via the blood or the lymphatic system. Stem cells that reach the thymus, a lymphoid organ of the immune system, differentiate into thymus-derived cells (T cells). The thymus generates cells of the immune system and has a great variety of spanning cells of different types. In particular, the thymus is responsible for the maturation of stem-cell to mature CD4 or CD8 T cells, each with a unique *T cell receptor* (TCR), that is a set of proteins on the cell's surface responsible for receiving signals. (Pham *et al.*, 2015; Singer *et al.*, 2008). A critical step in T cell maturation is making a functional TCR. Figure 1.1 summarises the process of thymocytes development in the thymus. This starts with lymphoid

precursor cells; these cells lack expression of both CD4 and CD8 co-receptors and are, thus, known as double-negative (DN) thymocytes. DN thymocytes undergo several differentiation steps, each of them defined by the cell-surface expression of developmentally regulated markers, before reaching the so called pre-selection DP thymocytes (pre-DP). In particular, according to thymocytes expression of two receptors (CD25 and CD44), they can be further subdivided into DN1, DN2, DN3, DN4 stages (Bailis & Pear, 2014; Ciofani & Zúñiga-Pflücker, 2007). At first, cells enter the thymus as DN1 and, after a gene rearrangement at the TCRβ gene locus is initiated, they progress to DN2. At the DN3 stage, cells start expressing the pre-T cell receptor (pre-TCR), composed of pre-TCRα and TCRβ. Here, there is a checkpoint as cells are now allowed to undergo *β-selection* process. Cells entering this transition are subjected to dramatic changes in their transcriptional profile, thus some studies differentiate the stages before and after the β-selection into DN3a and DN3b stages (Pham *et al.*, 2015). Cells arising from this checkpoint proceed to DN4 stage. Further development involves the up-regulation of both CD4 and CD8 co-receptors to generate double positive (DP) cells. DP cells go through gene rearrangement at the TCRα gene locus; when a mature TCR complex is expressed, thymocytes undergo the so called *positive selection*. During this process, DP thymocytes undergo major histocompatibility complex (MHC)-mediated selection. Post-DP cells that are positively selected transition to the single positive (SP) stage, where they can express either the CD4 or CD8 co-receptor. Thymocyte development is completed after the *negative selection* process: here self-reactive SP thymocytes are eliminated (Ciofani & Zúñiga-Pflücker, 2007).

At the end of the process, only cells with a functional receptor, capable of recognising foreign antigen, survive and leave the thymus. Cells exiting the thymus, called *näive* T cells, express either CD4 or CD8 molecules on their surface and they are functionally monospecific for antigen recognition. Näive T cells enter the periphery, meaning that they circulate in the blood, through the body, rather than being localised to specific organs (such as the spleen or lymph nodes). Upon encounter with a foreign antigen, they start differentiating and, thus, specialising into different cell types. They can now be called mature T cells. According to their function, T cells can be divided into helper T cells (called $CD4^+$ T cells),

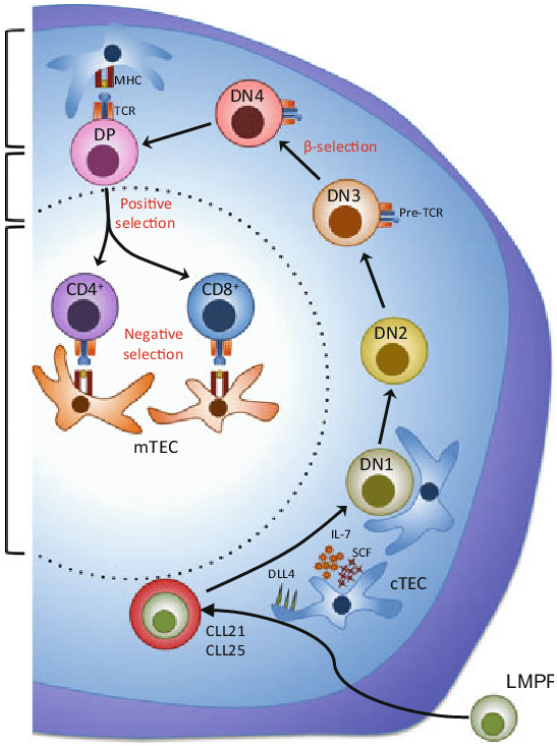


Figure 1.1: Stages of thymocyte development. A lymphoid primed multipotent progenitor (LMPP) that has migrated from the bone marrow, enters the thymus and starts the differentiation process towards thymocyte maturation. Thymocytes with a mature TCR complex undergo positive and negative selection to eliminate both cells with TCRs incapable of interacting with host-MHC (major histocompatibility complex from endogenous cells) as well as self-reactive cells. Figure reprinted with permission (RightsLink service) from Bailis & Pear (2014).

regulatory $CD4^+$ T cells, cytotoxic T cells (namely $CD8^+$ T cells or killer T cells) and memory T cells. For this PhD thesis, we focus on $CD8^+$ T cells, specialised in killing foreign agents as well as infected or tumour cells. We will review in the remaining of this Section, the main concepts of $CD8^+$ T cells activation and signalling program; Section 1.2 is dedicated to the $CD8^+$ T cells exhaustion dysfunctional state.

T cell activation

The activation of näive T cells by antigen is one of the central events in an adaptive immune response and quite a complex process. $CD8^+$ T cells leaving the thymus are characterised by one functional receptor specificity; they circulate through the body until they recognise the specific cognate antigen on the surface of an *antigen presenting cell* (APC), a second type of cells of the immune system. When APCs come across an antigen, they break it apart and show on their surface that specific part that can be recognised by specialised T cells. Indeed, T cell receptors (TCRs) cannot recognise a "free" antigen, but they can bind to it only using a specific structure, the major histocompatibility complex (MHC), placed on the surface of the APC (see Figure 1.2). The encounter and binding between a T cell and an APC initiates an adaptive immune response. A correct activation of the T cell takes place only in the presence of a second binding with a *co-stimulatory receptor*, such as CD28 (see Figure 1.2). Once the cell is activated, it starts proliferating; a single näive $CD8^+$ T cell can produce even 10^4 daughter cells (Kaech *et al.*, 2002b). These cells have extraordinary effector functions in killing infected cells showing that specific antigen on their surface. However, they do not live long: following pathogen's clearance, more than 95% of the effector cells die while the remaining small pool of $CD8^+$ T cells ultimately develops into long-lived memory T cells. Further details are given in Janeway *et al.* (1999).

Activated T cells express specific molecules to prevent their continuous and persistent activation; these are called *co-inhibitory receptors* (IRs) and includes, for example, CTLA-4 and PD-1 (Hashimoto *et al.*, 2018; Janeway *et al.*, 1999). The delicate balance between co-stimulatory and co-inhibitory receptors signals determines the T cell activity. The TCR signalling is triggered by the recognition of the cognate antigen peptide (small protein), but the T cell - APC binding alone is not enough: T cell differentiation and survival and, thus, cell fade, are driven by co-stimulatory and co-inhibitory receptors signalling (Chen & Flies, 2013). Mueller *et al.* (1989) show for the first time how co-stimulatory signals are necessary for a functional T cell activation. Moreover, the expression of many co-stimulatory and co-inhibitory molecules on the surface of T cells is also induced after activation: co-signals are continuously varied in response to dynamic

changes in environmental conditions (Chen & Flies, 2013; Chung *et al.*, 2021).

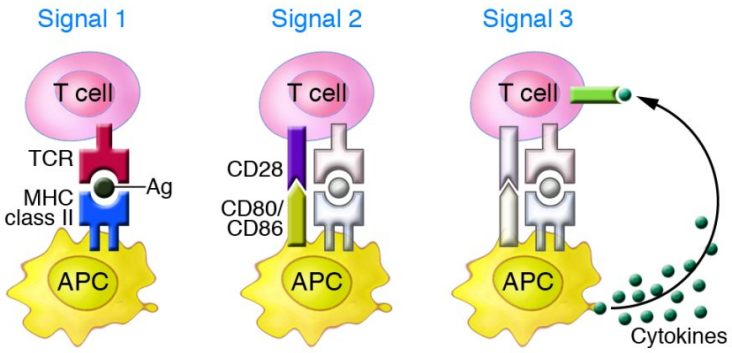


Figure 1.2: Three signals are required for T cell activation. Presentation, to a T cell, of the antigen peptide (Ag) via an MHC molecule on an APC cell; the antigen is recognised by the antigen-specific TCR. This binding needs to be stabilised by a co-stimulatory signal resulting from a second binding between specific molecules on the surface of APCs and T cells. For example, the binding between CD28 and CD80, on a T cell and APC respectively, triggers an activatory signal, whereas other interactions, such as with CTLA4, generate an inhibitory signal (see Section 1.2). Finally, a signal from cytokines produced by the antigen presenting cell interacting with specific receptors on the T cell is required to drive the T cell's phenotype. Figure reprinted with permission (RightsLink service) from Gutcher & Becher (2007).

Cell signalling and programming instructions

During an immune response, cells of the immune system communicate via different signals. As soon as the pathogen is recognised and the immune system activated, immune cells start producing *cytokines*. Cytokines are small molecules, made by proteins, that cells release in order to communicate with other cells (such as survival, proliferation, differentiation, functional activity and death signal); moreover, cytokines can be used to influence and stimulate cells' behaviour, e.g. movement towards the site of infection. The most common cytokines are *interleukins* and *interferons*. For example, Interleukin-2 (IL-2) is known to facilitate the growth of immune cells and enhance their division rate. Whereas,

interferon-gamma (IFN-γ) is a cytokine secreted mainly by activated $CD4^+$ and $CD8^+$ T cells and whose expression is induced by other cytokines such as IL-12 and IL-15. Secreted IFN-γ acts in upregulating MHC molecules and, thus, it plays a role in regulating the antigen presentation during an infection or in tumour micro-environment. However, what was just briefly described is an over-simplification: cytokines biology is usually complex; they interact and influence each other and might have different effects under different environmental contour conditions (Dinarello, 2007; Janeway *et al.*, 1999).

As ultimate effect, extracellular signals as cytokines are converted into intracellular signals as *transcription factors* (TFs). TFs are proteins that regulate gene transcription, the process by which genes of DNA are copied to make an RNA molecule. The transcription of a gene is followed by its translation into a protein that will be expressed by the cell and, eventually, influence its function. In particular, TFs bind to the DNA sequence around the gene of interest and they can either promote or repress the beginning of the transcription. In this way, TFs can alter proteins synthesis (production) and, thus, cellular functions. Overall, cell behaviour is determined by a delicate and complex balance: on the one hand, signals and cytokines are transformed into specific nuclear information and consequently determine TFs production; on the other hand, cytokines production is regulated by transcription factors as a response to environmental stimuli (Janeway *et al.*, 1999; Muegge & K.Durum, 1990).

During T cells activation and differentiation, triggered by pathogen encounter, cells with an identical genome (DNA sequence) transcribe, express and release different proteins according to their functions and location; that is, T cells of different types secrete different cytokines and produce different TFs to modulate their function. These differences can be used for cells' classification. T cell's *phenotype* results from cell cytokine production and cell surface markers, as well as the inter-play between the multiple transcription factors that are co-expressed within the cell (Evans & Jenner, 2013). Theoretically speaking, the cell's phenotype can uniquely indicate the state of a cell, that is its type and its functional activity.

Unfortunately, determination of cells' phenotype relies onto advances in experimental techniques. Fluorescence microscopy is a broadly adopted method

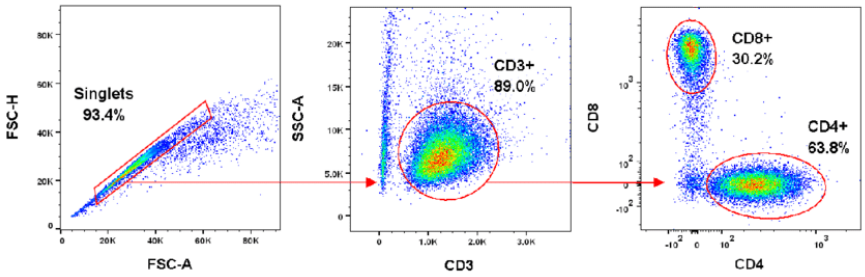


Figure 1.3: An example of a T cell density plot (Manescu *et al.*, 2020). The gating strategy allows to identify, from left to right, single cells, $CD3^+$ cells as well as $CD4^+$ (bottom right, expressing low CD8) and the cytotoxic T cells on the top left, expressing high CD8 and low CD4. Figure reprinted under the Creative Commons License.

used for analysing cells and investigating their heterogeneity. In particular, flow cytometry measures cell light scatters and the fluorescence intensities produced by fluorescent-labelled molecules that have been previously loaded in the cells sample. Figure 1.3 reports an example of how, using two labels, it is possible to identify two different populations of T cells expressing CD4 and CD8 transmembrane proteins, respectively. If on the one hand flow cytometry provides single-cell resolution, from the other the analysis can be carried out only for some pre-selected markers that are known to characterise cell types. On the contrary, single-cell sequencing technologies allow DNA and RNA quantification: an analysis of the genome can be done without need of prior knowledge about cells' type. This flexibility, however, comes at a high cost per cell analysis; moreover, a pre-enrichment of the cells (cells co-culture with nutrients) is often require to increase sample size. Consequently, some subpopulations might be lost due to their evolution during the enrichment period. Unfortunately, both these methods provide only a single screenshot of the cells' states, without giving insights into the spatio-temporal dynamics. An analysis of cells' states at different time points is only possible if multiple measurements of the cells sample are acquired; this comes with a higher cost in performing the experiment as well as an increase in its complexity.

1.2 CD8$^+$ T cell exhaustion

In contrast to acute infections, in persistent chronic infections and cancer, most effector CD8$^+$ T cells become dysfunctional in the long-term. These cells have been first observed at the end of the last century in Rafi Ahmed's lab (Zajac *et al.*, 1998) and have brought immunologists to start wondering about how T cell response differs between an acute and chronic infection. In chronic infections and cancer, it has been observed that virus-specific CD8$^+$ T cells persist but they lose their effector functions and turn into a less functional state, commonly referred as *exhausted state* (see Figure 1.4). Only more recent biological studies highlighted the complexity of the exhaustion paradigm. In particular, studies carried out by John Wherry's lab on lymphocytic choriomeningitis virus (LCMV) chronic infection were the first to characterise T cell exhaustion by a loss of inflammatory cytokines, a high expression of inhibitory surface receptors, as well as a modified transcriptional and epigenetic program (Wherry, 2011; Wherry *et al.*, 2007).

Moreover, Shin & Wherry (2007) were the first to identify different populations of exhausted cells. Only later, the process leading to T cell exhaustion has been observed to be heterogeneous by many following studies (Im *et al.*, 2016; McLane *et al.*, 2019; Miller *et al.*, 2019; Paley *et al.*, 2012; Wherry, 2011). In these terms, CD8$^+$ T cell exhaustion is a differentiation process, observed during chronic infections, where cell lineage is comprised of stem-like progenitor exhausted cells that, eventually, differentiate into a terminal and dysfunctional exhausted state. However, the differentiation to this dysfunctional state might also be beneficial for the body. If at the beginning of a chronic stimulation (such as HIV, hepatitis or malaria) the immune system tries to destroy the pathogen and eliminate the infection, with the differentiation of immune cells toward an exhausted state the immune system tries to learn how to manage the infection, keeping the virus in check without causing too much damage (Gao *et al.*, 2022). T cell exhaustion is also a major problem in human cancer; exhausted cells have been later observed also in the tumour micro-environment characterised by a persistent antigen signal, similarly to a chronic infection with a constant pathogen load (Schietinger *et al.*, 2016; Weinberg, 2013). Tumour cells have indeed develop a mechanism to trigger T cell exhaustion and slow down the reaction of the immune system against

them. We also see T cell exhaustion in autoimmunity, when immune cells are also inappropriately chronically stimulated by parts of our own body. Here, when T cell exhaustion is more severe, the immune response against our own cells is lower and thus the symptoms of autoimmunity tend to decrease. Overall, in all these three scenarios, the immune system needs to control and balance immune cells count and development of immunopathologies.

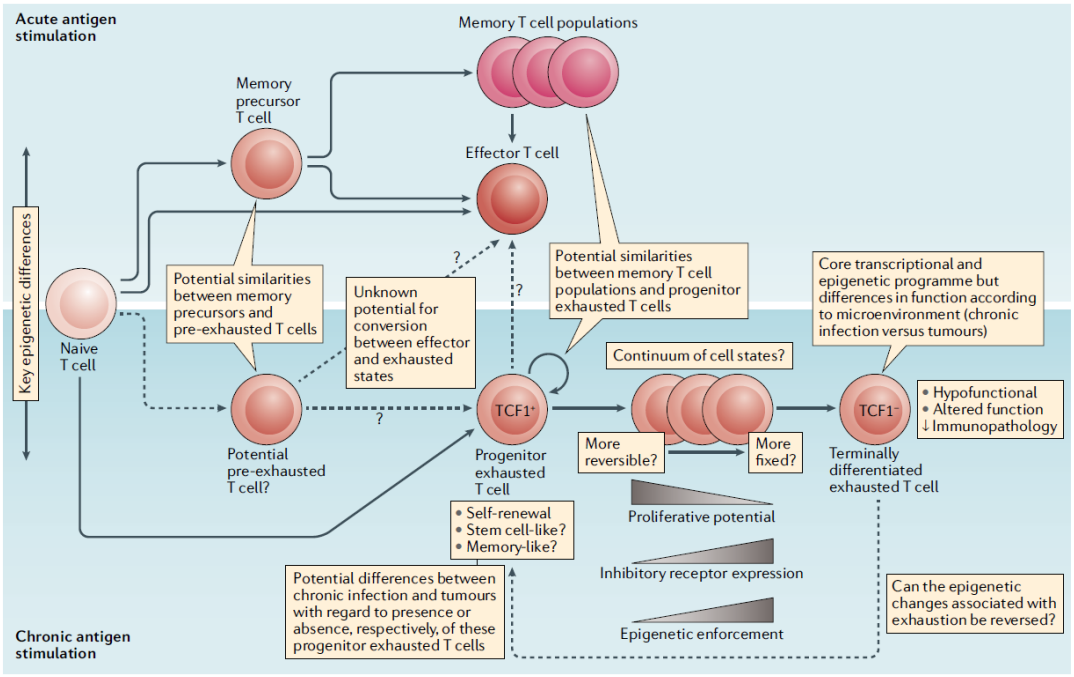


Figure 1.4: Overview of the T cell differentiation process during an acute (on top, light blue) and chronic (at the bottom, darker blue) infection. This Figure highlights the divergence of the differentiation pathway in an acute and chronic infections, as well as some similarities (such as between memory cells and progenitor exhausted cells). Moreover, the loss of effector functions (proliferative potential, IRs expression, epigenetic changes) during the differentiation pathway from a progenitor to a more terminally exhausted state, is represented in a qualitative way by the grey bars. Finally, the possibility, crucial in immunotherapies, of restoring a functional response of terminally differentiated exhausted cells is outlined: a dashed arrow depicts an hypothesesd reverse transition. Figure reprinted with permission (RightsLink service) from Blank *et al.* (2019).

As depicted in Figure 1.2, a signal from antigen, a co-stimulatory signal, i.e. IRs, and a signal from inflammation and soluble mediators, i.e. cytokines, constitute the basic framework for T cell activation. Alterations of antigen and co-stimulatory signals, typical of a chronic infection, will lead T cells to an exhausted state, functionally different from the effector state observed in an acute infection (a schema is reported in Figure 1.4). Moreover, during a chronic infection, the landscape of cytokines (that is molecules floating in the extracellular medium and influencing T cell activation and differentiation) released by cells is complex and still not completely explored. For example, cytokines that are known to play a role in the exhaustion fate are IL-10 and IL-2, promoting and antagonising exhaustion, respectively (McLane *et al.*, 2019). The possibility of tuning co-stimulatory signals or cytokine levels and, thus, boosting the immune response, is being exploited by many therapeutic strategies for cancer's treatment.

In simple words, exhausted T cells are functionally impaired, unable to effectively produce cytokines and clear the infected or cancer cells. More technically, they have been defined by the following set of features (McLane *et al.*, 2019; Wherry, 2011; Wherry *et al.*, 2007):

1. Loss of effector functions in a hierarchical manner. At early stages of exhaustion, T cells lose the ability to produce cytokines IL-2 and TNF (tumour necrosis factor); a decrease in their cytotoxicity, that is the ability of being toxic for infected or cancer cells, is also observed. This involves a reduction of perforin and granzyme B release and, at later stages, a decrease in the production of IFN-γ. Final stages might bring cellular death via apoptosis.

2. Loss of the ability to undergo homeostatic proliferation driven by IL-7 and IL-15 cytokines. Exhausted T cells are, instead, maintained by a persistent antigen signal that stimulates the proliferation of early exhausted cells.

3. Expression of multiple inhibitory receptors (IRs) in a sustained manner. Inhibitory receptors can inhibit the immune activation signal and, thus, they constitute a way cells can dampen their response and keep the immune system in balance preventing immunopathology. Within the most studied

IRs, there are PD-1 (programmed cell death 1) and CTLA-4 (cytotoxic T lymphocyte-associated Antigen 4). If in an acute infection they are transiently expressed by T cells, in a chronic infection PD-1 and CTLA-4 are constantly produced by cells after activation. Similarly, both Lag-3 and TIM-3 are IRs known to be expressed transiently during T cell activation and differentiation, but they are kept to a sustained level in case of a chronic infection. In general, the co-expression of multiple IRs is a common feature of exhausted cells and, in most cases, other IRs are co-expressed with PD-1. Therapeutic strategies have been developed exploiting the fact that, by blocking the interaction of PD-1 with its conjugate ligand PD-L1, functions of exhausted T cells can be partially restored (see Section 1.2.1).

4. Maintenance of a quiescence oxidative phosphorylation metabolism. Exhausted T cells are not able to switch their metabolism to aerobic glycolysis; the metabolic switch from a fatty acid oxidation pathway to glycolysis allows, instead, functional effector cells to sustain the required enzymatic activity and the transport of nutrients.

5. Distinct transcriptional and epigenetic profile compared to effector and memory T cells. Transcriptional changes include genes encoding IRs, transcription factors (as Tbet or TCF1), or molecules involved into cytokines production, co-stimulatory signals as well as metabolic behaviour. In particular, molecules, such as Tbet or TCF1, can be expressed in both functional effector or memory cells as well as in exhausted cells, but with a different biology. Differences in the epigenetic profile involve changes in the accessibility of chromatin regions, meaning there is a difference in cells' phenotype but not in their genotype. Chromatin changes affect proteins formation and, thus, cells' fate commitment.

Each one of these features, taken by itself, is not sufficient to clearly identify an exhausted cell; but, taken all together, they characterise exhausted cells. However, the mechanisms behind this change of program as well as the reasons of its origin still remain elusive. In recent years, several studies have been carried out

defining a multiplicity of exhausted subsets each characterised by specific markers (proteins expressed on cellular surface). These results do not always align and the inconsistencies between the subpopulations arising from different studies might suggest the exhaustion-differentiation process to be more heterogeneous and complex than what was originally thought (see Figure 1.4). One hypothesis is that multiple distinct exhausted subsets are formed as a way to better fight a chronic infection and adapt the immune response to minimise the damage and keep low the antigen load (Utzschneider *et al.*, 2013).

Recent techniques allow to analyse the transcriptional and epigenetic landscape of exhausted cells (Miller *et al.*, 2019; Pauken *et al.*, 2016) as well as the importance of some factors, such as TOX (Khan *et al.*, 2019; Scott *et al.*, 2019), in driving the exhausted-specific epigenetic program being established. Here, we would like to give the reader an overview of the most common hallmarks that have been used to analyse and characterise a functional-dysfunctional T cell population.

At the peak of the CD8$^+$ T cell response in an acute infection, effector cells can be divided into different subsets according to their expression of the IL-7 receptor and the inhibitory receptor KLRG1; in particular, at early stages, effector cells lack both IL-7 receptor and KLRG1 and they differentiate into short-lived effector cells (expressing KLRG1) and memory precursor effector cells (expressing IL-7 receptor) (Plumlee *et al.*, 2013). Moreover, upon T cell activation, the expression of the inhibitory receptor PD-1 (programmed death 1) is rapidly upregulated and decreased upon clearance (Sharpe & Pauken, 2018); this is not the case in a chronic infection where the expression of PD-1 is maintained substantially higher (Wherry & Kurachi, 2015). The role of PD-1 has been first investigated by Day *et al.* (2006); resulting in PD-1 being the first, and still one of the most important, markers for the dysfunctional exhausted state. Evidence supports the proposal of a progressive differentiation process where PD-1 has an intermediate expression in progenitor exhausted cells and a higher expression in a terminal stage of exhaustion (Abdel-Hakeem *et al.*, 2021; Blackburn *et al.*, 2008; Mann & Kaech, 2019). In particular, precursor exhausted cells have been described as a PD-1^{int} population, whereas a more terminal subset of cells as characterised by PD-1$^+$ (Blackburn *et al.*, 2008; He *et al.*, 2016; Paley *et al.*, 2012). Studies

of Utzschneider *et al.* (2016), Im *et al.* (2016) and Wu *et al.* (2016) focus, instead, on TCF1, that is a transcription factor, namely T cell factor 1, essential for normal T cell development. They identify a $TCF1^+$ progenitor exhausted population and a $TCF1^-$ terminally exhausted population of cells. However, a recent study of Chen *et al.* (2019) shows a non-homogeneous behaviour, over differentiation, of the expression of PD-1 and TCF1 as both being strongly expressed already at day 8 of a chronic infection. They, thus, suggest the influence of PD-1 in preserving $TCF1^+$ cells and their mutual importance in determining cells fate. Overall, the PD-1^+ $TCF1^-$ subset is associated with a terminally exhausted T cell population, whereas PD-1^{int} $TCF1^+$ cells are shown to be able to undergo antigen-driven proliferation and, thus, retain stem-like behaviours (Beltra *et al.*, 2020). Indeed, this latter subpopulation shares several features with multipotent PD-1^- $TCF1^+$ cells that are typical of an acute infection and can further differentiate into either the above mentioned $KLRG1^+$ cells or the memory precursor effector cells. Moreover, lineage-tracing studies performed by Angelosanto *et al.* (2012), demonstrated that terminal differentiated cells marked as $KLRG1^+$ survive poorly during chronic infection and cannot give rise to exhausted cells, whereas $KLRG1^-$ cells can give rise to either memory-type cells or to exhausted cells. In general, gene studies looking at gene chromatin regions highlight the key role of the interplay between PD-1 and the transcription factor TCF1 in coordinating and maintaining the formation of progenitor exhausted cells (Abdel-Hakeem *et al.*, 2021; Chen *et al.*, 2019; Im *et al.*, 2016; Kasmani *et al.*, 2022). In addition, they clearly define a distinct and stable (at day > 180 post clearance) epigenetic state of exhausted cells compared to effector and memory cells.

After PD-1, one of the most studied surface molecules to identify exhausted cells as well as enhance immunotherapy treatments is the co-inhibitory receptor TIM-3. This is involved in the production of the IFN-γ cytokine (decreased in exhausted cells, see point 1. of the exhaustion feature list) and, thus, it has a key role in defining the immune responses in different T cell subsets. Several works characterise terminally exhausted T cells as highly expressing PD-1 and TIM-3. In particular, Wherry *et al.* (2007) and Wu *et al.* (2016) found that $CD8^+$ T cells differentiate into $TCF1^+$ TIM-3^- and $TCF1^-$ TIM-3^+ even at early time points (before 30 days after infection) of a chronic infection. These two subsets define

a progenitor reversible exhausted state and a terminally exhausted state, respectively. In addition, Im *et al.* (2016) and Seo *et al.* (2019) characterise terminally exhausted T cells as highly expressing PD-1 and TIM-3; He *et al.* (2016) define a more exhausted subset of T cells as less expressing the chemokine receptor CXCR5 and highly expressing PD-1 and TIM-3. These observations, together with the ones mentioned in the previous paragraph, could lead to the characterisation of a terminally exhausted subset as PD-1$^+$ TCF1$^-$ TIM-3$^+$ and a progenitor subset of cells as PD-1^{int} TCF1$^+$ TIM-3$^-$. However, the key steps in T cell differentiation towards exhaustion are still not clearly defined, as several studies employ different markers and highlight different subpopulations; for example the work of Beltra *et al.* (2020) reveals four different developmental states in the exhaustion-differentiation process (from progenitor to intermediate to terminal) according to the expression of PD-1 and TCF1 inhibitory receptors, chemokine CXCR5 as well as Tbet, Eomes transcription factors. Similarly, Hudson *et al.* (2019) suggest the existence of an intermediate exhausted state subdividing the cell population according to TIM-3, TCF1 and CX3CR1 factors: the intermediate subpopulation, characterised by CX3CR1$^+$, is still able to proliferate, whereas the terminal population (CX3CR1$^-$) can only undergo little to no division and poorly contribute to viral control.

Moreover, details about how effector T cells become programmed, at early time of stimulation, to terminal exhaustion have remained poorly understood until now. Indeed, our understanding of the molecular and intracellular mechanisms by which, even well-known, inhibitory receptors (such as PD-1) control T cell exhaustion is incomplete. Recent studies broaden the spectrum, investigating other markers as well as the influence of cytokines. For example, Mo *et al.* (2021) found that the IL-2 cytokine, known to promote expansion of T cells, can, at the same time, drive terminal differentiation and induce expression of co-inhibitory receptors such as TIM-3 and PD-1 (both hallmarks of exhaustion, see point 3 of the exhaustion feature list). Moreover, it has been found that the population of PD-1^{int} TCF1$^+$ progenitor exhausted cells might be maintained by TOX activation and upregulation, highlighting the importance of this transcriptional factor (Alfei *et al.*, 2019; Khan *et al.*, 2019; Mann & Kaech, 2019).

In addition, Paley *et al.* (2012), Im *et al.* (2016), Utzschneider *et al.* (2016) and Beltra *et al.* (2020) assess the heterogeneity of exhausted $CD8^+$ T cells according to the expression of Tbet and Eomes transcription factors. During an acute infection, Tbet and Eomes are upregulated following T cell activation; they operate in contrasting ways driving the development of $KLRG1^+$ terminal effector cells or of the memory phenotype, respectively. Whereas, in a chronic infection, Tbet and Eomes are indispensable for the formation of exhausted cells. The behaviour of these two subpopulations, that is proliferating $TCF1^+$ and not proliferating exhausted $CD8^+$ T cells, is governed by these two transcription factors, Tbet and Eomes (Paley *et al.*, 2012): cells characterised by a high expression of Tbet, demonstrated extensive division, whereas the ones expressing high Eomes were associated with a more dysfunctional PD-1^+ phenotype. Also, findings by Li *et al.* (2018) support this hypothesis showing that high levels of Eomes are expressed by severely exhausted PD-1^+ cells. In addition, expression of the transcription factor NFAT has been associated with an anergy (cell inactivation) and exhausted state (Zhang *et al.*, 2020), and it is suggested to drive the production of IFN-γ (Teixeira *et al.*, 2005). On the other hand, the program of NFAT (mechanisms of NFAT accessing the genome) might be dictated by TOX, another transcription factor downstream of NFAT (Seo *et al.*, 2019), that we mentioned before as driving the formation of PD-1^{int} $TCF1^+$ cells.

Co-stimulatory signals, induced for example by CD28 and 4-1BB molecules, usually drive the activation, expansion, and differentiation of näive T cells; however, during chronic infections these same signals can lead to T cell exhaustion. For example, CD28 is associated with increased susceptibility to exhaustion as well as decreased T cell persistence compared to 4-1BB stimulation. Indeed, Kim *et al.* (2020) underline how co-stimulatory signals can antagonise a functional response of $CD8^+$ T cells and, in the context of a hepatocellular carcinoma tumour, they identify a population of exhausted T cells as expressing 4-1BB. In conclusion, if persistency of antigen is necessary for exhaustion to arise, both IRs, transcription factors and co-stimulatory receptors interact during T cell response, leading to a complex landscape and contributing to T cell exhaustion,

Another crucial point, deeply investigated but still not fully clarified, of T cell exhaustion research is the timeline of exhaustion to arise. For example, An-

gelosanto *et al.* (2012) show how 15 days after the initial infection, T cells are partially impaired but still retain some plasticity and effector functionality; whereas, at day 30 after infection, these same cells are fully committed to an irreversible exhausted state. This fits with the timeline for the transcriptional and epigenetic exhausted profile being established (Philip *et al.*, 2017). In addition, studies at early time points after infection have been performed. Surprisingly, Khan *et al.* (2019) found that the first signs of exhaustion arise within 4 days of infection: the exhaustion transcription factor TOX begins to be expressed at day 4 of LCMV chronic infection and cells expressing TIM-3 are visible within 6 days. However, the question if these cells are already committed or they still retain plasticity remains open. Utzschneider *et al.* (2020) suggest that exhaustion is imprinted early during T cell activation, and inherited by progenitor cells. They show that the divergence between a precursor and an effector type of cell is initiated early during T cell activation: precursors TCF1$^+$ TIM-3$^-$ (retaining high proliferative potential) and terminal TCF1$^-$ TIM-3$^+$ can be recognised already after three cell divisions; moreover, they hypotheses that KLRG1$^+$ cells are a subset of TCF1$^-$ TIM-3$^+$ cells.

In addition, a clear understanding of the stage at which CD8$^+$ T cells make the decision to change their intra-cellular state, responding to infection, and differentiate, and of the differences between an acute and chronic infection, is missing. The landscape is quite complex as, on one side, the persistency of antigen drives cells towards an exhausted state (lacking memory potential) and, on the other side, similarities in expressed molecules between exhausted and memory cells (e.g. TCF1) have been observed (Wherry & Kurachi, 2015). In particular, Angelosanto *et al.* (2012) investigate when cells differentiation pathway starts diverging towards exhaustion. Since an exhausted phenotype arises in case of a sustained antigen stimulation, the initial hypothesis focused on exhausted cells arising from terminal differentiated cells, that are effector cells subjected to multiple differentiation steps due to a sustained antigen stimulation. However, as mentioned above in this Section, both long-lived memory cells in an acute infection and progenitor exhausted cells might arise from a subset of cells characterised by a low expression of the KLRG1 inhibitory receptor. These two hypothesess have not been confirmed by clear studies yet, also due to the difficulties in the

analysis of memory cells (which are arising in long-term and thus require long experiments). Also, a related open question is whether memory cells can commit to an exhausted phenotype upon subsequent and sustained reinfection.

One of the problems in identifying, characterising and, thus, quantifying, subsets of cells is that some surface receptors that have been observed to discriminate a population are lacking of a corresponding antibody capable of fixed staining. This prevents the possibility of analysing via flow cytometry exactly that marker and requires the use of a workaround: use another protein with a known staining antibody assuming that is expressed (or not expressed) in the population of interest in the same way of the intended initial marker. For example, the chemokine receptor CXCR5 lacks a corresponding staining antibody. For this, Im *et al.* (2016) were pioneers in identifying two different subsets of $CD8^+$ T cells according to their expression of TCF1. In particular, the subset $CXCR5^+$ TIM-3^- was identified by $TCF1^+$ PD-1^+, whereas subset $CXCR5^-$ TIM-3^+ by $TCF1^-$ PD-1^+ cells. Note that, here, the stain on PD-1 has been used to distinguish näive $TCF1^+$ cells from the two subsets $CXCR5^+$ and $CXCR5^-$. Similarly, expression of PD-1 can be used as a surrogate marker to separate subpopulations according to Tbet and Eomes expression: cells PD-1^{int} are associated with $Tbet^+$ and cells PD-1^+ with $Eomes^+$, respectively (Paley *et al.*, 2012). Moreover, cellular subpopulations can be identified only via known markers: panels of specific antibodies (cellular markers) need to be decided before the flow cytometry analysis. This brings to limitations when analysing a population of cells whose expected subpopulations are still not well studied: one need to forecast relevant markers, according to the known functional behaviour of the subpopulations of interest. This is one of the reasons why several studies about the exhaustion-differentiation process are carried out considering different markers. For example, Hudson *et al.* (2019) analysed RNA data sets to systematically identify $CD8^+$ T cells exhaustion markers. Their results include surface proteins well-known to be associated with exhaustion, such as inhibitory receptors PD-1, TIM-3 and Lag3, as well as molecules with a less studied role in exhaustion. In particular, they focus on the gene encoding the protein CD101 and show how the TIM-3 subpopulation of terminal cells can be further subdivided into $TCF1^-$ TIM-3^+ $CD101^-$ and

TCF1$^-$ TIM-3$^+$ CD101$^+$ subpopulations. Since exhaustion is quite a novel research area, the involved molecular mechanisms are not completely understood yet and upcoming studies clarifying the underneath dynamics are expected.

1.2.1 Exhaustion reversibility and immunotherapies

In the previous Section, we have highlighted how the exhaustion-differentiation process leads to an heterogeneous pool of cells: functional cells progressively decline in their capabilities and diverge from their state towards a dysfunctional one. This transition comes with epigenetic modifications that sustain the different characterisation of exhausted T cells (Gray *et al.*, 2017; Philip *et al.*, 2017). Epigenetic changes are driven by environmental factors and do not imply changes in DNA sequence; they identify cell's status and which genes are "turned on" to make corresponding proteins and which, instead, "turned off". In this perspective, cell epigenetic status can tell the spectrum of functionalities the cell is taking over. Some epigenetic changes are stable, whereas some do not persist but can still change during development and they are not transmitted to cell progeny. Epigenetic studies show that exhausted T cell subpopulations differ from the effector or memory cells by approximately 6×10^3 open chromatin regions (Pauken *et al.*, 2016; Philip *et al.*, 2017); but it is not clear at which differentiation stage those changes are still plastic and when, instead, they become fixed. Consequently, current standing questions are when and at which stage exhaustion becomes irreversible as well as at what extent exhausted cells can be reinvigorated. Preventing epigenetic fixation or acting, i.e. via immune checkpoint blockade, when changes are still plastic might be a method to counteract T cell exhaustion.

Recent studies show how this modified epigenetic state can persist despite exhausted cells resting for a long time in absence of the antigen (Pauken *et al.*, 2016). However, the commitment to the exhausted phenotype seems to be reversible at early stages; that is, some T cells can recover and reacquire effector functions (killing and cytokine production, for instance). Barber *et al.* (2006) opened the door to the studies about exhausted CD8$^+$ T cells reinvigoration. They found that functionally exhausted CD8$^+$ T cells are characterised by a high expression

of the PD-1 inhibitory receptor; and they showed how, by blocking PD-1 interactions, exhausted cells could be rejuvenated (see Figure 1.5). More recently, *in vitro* studies employing advanced technologies (Pauken *et al.*, 2016; Sharpe & Pauken, 2018), highlight that the epigenetic profile of functionally reinvigorated (by PD-1 blockade) exhausted T cells differs only slightly from the exhausted one. Consequently, a burst of reinvigoration is observed at the beginning of the PD-1 blockade, but then, after the cessation of the treatment, $CD8^+$ T cells revert back to the previous exhausted state (proliferation and effector functions, IRs expression).

Ways of reversing the effects of T cell exhaustion, such as immune checkpoint blockade, i.e. PD-1 blockade (Tabana *et al.*, 2021), are explored by cancer immunotherapy techniques (Raskov *et al.*, 2021; Waldman *et al.*, 2020). Inhibitory receptors are fundamental to prevent autoimmunity, but many tumours exploit these mechanisms to evade the immune system. The tumour cell (expressing receptors as PD-L1 and PD-L2, for example) acquired the ability to bind to co-inhibitory receptors (in this case PD-1) and impair immune response: T cells perceive a negative, inhibitory signal and start losing their abilities showing exhausted features similar to those observed in a chronic infection (McLane *et al.*, 2019). Thus, immunotherapies have been developed to modulate immune activity by blocking the binding between tumour and T cells; in particular, research focused on the development of antibodies specifically formulated to target regulatory proteins, such as inhibitory receptors, expressed on the surface of exhausted cells.

PD-1 is known to contribute to T cell activation, determining cell fate as well as T cell return to homeostasis (Sharpe & Pauken, 2018). These observations have been translated to the clinic and PD-1 pathway blockade has been made the foundation of immunotherapy leading to the availability on the market of nivolumab and pembrolizumab. These two drugs, based on PD-1 blockade, are approved by the Food and Drug Administration (FDA) for treatment of some types of solid cancer (such as melanoma (Huang *et al.*, 2019) and non–small cell lung carcinoma). Despite the proven success in enhancing the survival of cancer patients, the response is not homogeneous among tumour types as well as

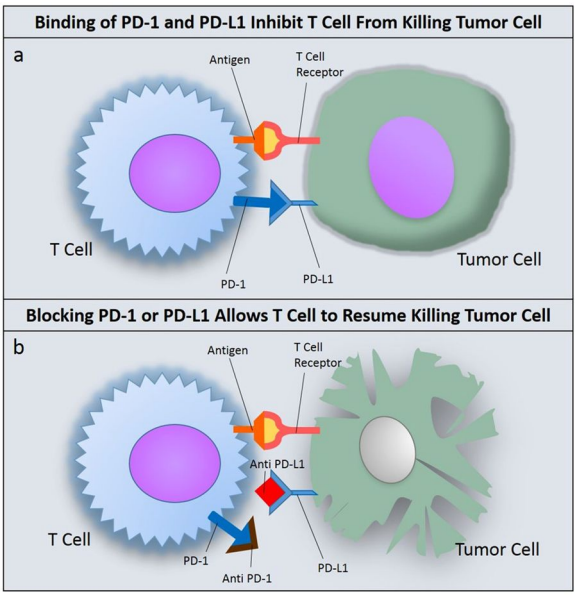


Figure 1.5: Besides specific antigen recognition (provided by MHC-TCR binding, reported in red), T cell activation requires a co-stimulation signal for a full response. Inhibitory receptors belong to the family of co-stimulatory signal delivering, instead, a negative signal to T cells: following their engagement, they negatively modulate T cell activation to maintain self-tolerance and prevent exacerbated activation and autoimmunity. Tumour cells acquire the ability to employ IRs signal to evade the immune system. In particular, PD-1 is one of the most well-known inhibitory receptor and normally expressed on T cells. Many tumour cells over-express PD-L1 (the corresponding binding receptor of PD-1) on their surface, as a mechanism to lead to T-cell dysfunction and prevent proper immune response (depicted in subfigure (a)). Immunotherapies targeting IRs signalling (in this case PD-1-PD-L1) have been developed: the administration of a specific antibody that is able to bind PD-1 or PD-L1 receptor results in blocking the inhibitory signal and, thus, in restoration of effector functions and significant reduction of tumour load (Barber *et al.*, 2006). Figure reprinted from Caldwell *et al.* (2017) under the Creative Commons License.

patients clinical conditions (Hudson *et al.*, 2019). Consequently, studies targeting others inhibitory receptors (such as CTLA-4 and PD-L1) have been carried out; in addition, combinations of different therapies have been considered, i.e.

multiple checkpoint blockade; checkpoint blockade and co-stimulatory signals; or checkpoint blockade and radiotherapy.

Moreover, these antibody treatments are shown to be effective only on restricted subpopulations. In particular, precursor TCF1$^+$ cells, still able to proliferate, can better respond to immune checkpoint blockade (Hudson *et al.*, 2019; Im *et al.*, 2016). A burst in proliferation after PD-1 blockade is observed for this subset of cells. In addition, it has been recently observed how mTOR inhibition improves the efficacy of PD-1 blockade therapy by suppressing more differentiated TIM-3$^+$ cells and promoting further formation of progenitor TCF1$^+$ cells (Ando *et al.*, 2023). Also, TOX knockdown (temporary stop or decrease of TOX expression) in tumour-specific CD8$^+$ T cell exhibits synergy with PD-1 blockade immunotherapy (Zhang *et al.*, 2020). Novel and different approaches are constantly studied and developed to face the complexities of tumour and immune system interactions. For example, Kumar *et al.* (2020) focus on combining PD-1 blockade therapy with a boost of mitochondrial activity. Indeed, mitochondria play a crucial role in managing the energy metabolism, and exhausted cells are impaired being unable to switch to an aerobic glycolysis metabolism (see point 4. of exhaustion feature list). Consequently, a boost in mitochondrial activation could enhance effects of PD-1 blockade therapy.

Overall, it is of crucial importance, for the development of any effective immunotherapies, to fully understand exhaustion mechanisms so that the opportunities to prevent or reverse T cell exhaustion therapeutically can be enhanced. In this sense, time is a key factor in cancer treatment: different patients survival responses have been reported, depending on if the treatment has been administered at early or late stage of exhaustion (Pauken & Wherry, 2015). Exhausted cells that, during the first one to three weeks of a chronic infection, have been adoptively transferred into uninfected mice, can recover and develop into long-lived memory-type cells, suggesting a still plastic epigenetic profile (Angelosanto *et al.*, 2012). A question that researchers are currently facing is whether progenitor TCF1$^+$ exhausted cells are already committed to terminal exhaustion or still able to revert their phenotype. Alternatively, a "precursor" reversible exhausted state should be identified: several hypotheses have been made about this

hypothetical state as a transition between a functional effector and a progenitor proliferating TCF1$^+$ exhausted state (Chen *et al.*, 2019).

1.2.2 Overview of the experiment performed by the Salk Institute of Biological Study

The exhaustion-differentiation process has been associated with progressive changes in cells epigenetic profile (Pauken *et al.*, 2016; Philip *et al.*, 2017; Wherry *et al.*, 2007). While the long-term profile has been more studied and characterised, the early events regulating exhaustion fate compared to an effector or memory one still remain not well identified. Moreover, the first studies about CD8$^+$ T cells exhaustion and antibody blockade treatments have been performed with endogenous CD8$^+$ T cells (cells arising inside the body). Only in recent years, with the advent of CAR T cell therapy as a potentially curative treatment, it has been observed that engineered CD8$^+$ T cells can become dysfunctional and acquire exhausted features in a similar way as endogenous CD8$^+$ T cells (Weber *et al.*, 2021).

For part of this doctoral project, a collaboration with Professor Susan Kaech and a post-doctoral scientist in her group, Dr. Thomas Mann (both at the Salk Institute for Biological Studies in La Jolla, California) has been carried out. The research of Prof. Kaech and Dr. Mann focuses on exploring the *in vitro* CD8$^+$ T cell differentiation towards an exhausted phenotype at early time points. In particular, they employ novel optogenetic approaches and use blue light to perform a precise and high-throughput investigation of how engineered T cell signalling dynamics influence cell fate towards an effector or exhausted state. As mentioned, the experimental conditions focus on early time scales: in the carried out experiments T cells are optogenetically stimulated in a range of duration intervals within 48 hours. Sample data are analysed by flow cytometry to identify the different populations of CAR T cells, from functional effector to a progenitor and terminally exhausted phenotypes.

In particular, Dr. Mann has made use of an optogenetically controllable version of a chimeric antigen receptor, named "OptoCAR", that is sensitive to blue light and, thus, allows to control signal dynamics in the T cell receptor

pathway. The engineered receptor can transduce antigen signal when the cell is in the dark, but signalling is inhibited when the cell is exposed to blue light. That is, when cells are in the dark, the antigen signalling is mediated by blue light which stimulates the binding of a light sensitive domain, attached to the engineered cellular receptor (light blue in Figure 1.6), and a peptide, attached to the signalling domains of the CAR. A schematic representation is reported in Figure 1.6. Going more into biochemical details, the so called *LOVTRAP optogenetic system* (depicted in red) is made of two different domains, the LOV and the Zdk fragments, that are inserted within the signalling chain from the cellular membrane to the CAR's signalling domain (where the signal starts the downstream pathway towards the nucleus). In light time, the two fragments are separated and the transduction is prevented. Re-association of the LOV-Zdk is required to reactivate the signal: in the dark, indeed, the two halves gradually re-bind (over a period of 5-10 minutes) so that, once bound, the signal can be transduced by the receptor.

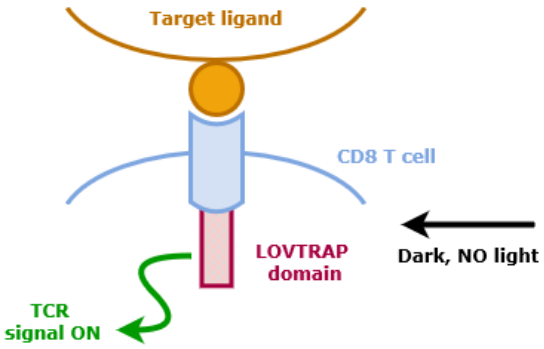


Figure 1.6: In orange the target cell; in light blue the $CD8^+$ T cell surface receptor. In the dark, the two halves of the LOVTRAP domain (in red) are bound and the TCR signalling cascade is induced. Whereas when cells are enlightened, the signal coming from the antigen-receptor binding is not spread (the LOV and the Zdk fragments are separated) and the signalling cascade not initiated.

For this experiment CD19 CAR T cells have been used, i.e. CARs have been engineered to recognise CD19 antigen (in orange in Figure 1.6). Cells have been cultivated under specific cytokine conditions (intermediate IL-2), CD19 peptide has been presented to the cells for a time window of 18 to 24 hours; then, cells

have been washed from the antigen before the start of the actual optogenetic experiment. Only cells showing CD19 receptor (a blasticidin selection has been applied) and cells that have positively received the OptoCAR vector (mScarlet fluorescence protein has been used as a transduction marker), and are thus OptoCAR+, are then placed in the wells ready for the blue-light stimulation period. The special optoPlate-96 used for the experiment is depicted in Figure 1.7. Cells that have been stimulated with different light duration, are then collected after 48 hours and analysed via flow cytometry. This method allows to gather single-cell resolution data on which proteins are expressed: by quantifying the variation in proteins' expression one can identify different subsets of cellular populations.

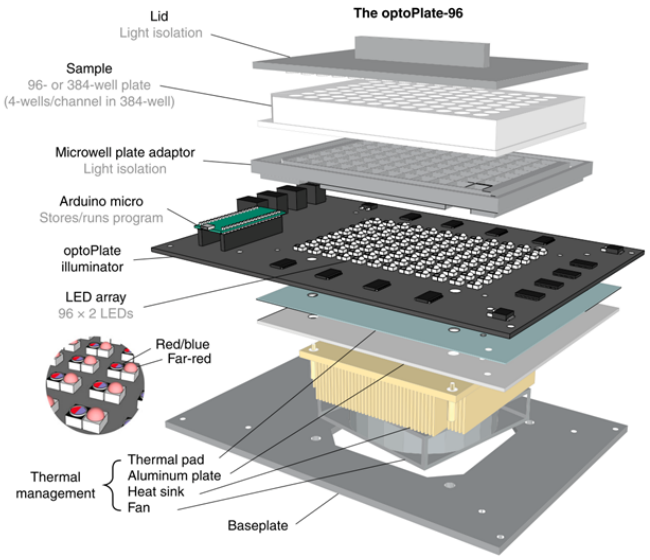


Figure 1.7: An expanded view of the optoPlate-96 showing its components. The 96-position LED array can independently illuminate each well. Figure reprinted with permission (RightsLink service) from Bugaj & Lim (2019)

A preliminary study to identify relevant exhaustion markers that could characterise effector and exhausted differentiation states has been performed by Dr. Mann. As outlined in previous Section, inhibitory receptors PD-1 and TIM-3 and the transcription factor TCF1 have a clear expression in the exhausted T cell population (Mann & Kaech, 2019; McLane *et al.*, 2019; Wherry *et al.*, 2007). The

data collected by the collaborators and used for the mathematical modelling reported in Chapter 6, focus on TIM-3$^-$TCF1$^+$ and TIM-3$^+$TCF1$^-$ populations, as characterising a progenitor reversible and a terminally exhausted state, respectively. Unfortunately, markers specific for functional effector differentiation states, could not be identified within reasonable time to be included in this PhD thesis.

After the 48 hours of optogenetic experiments, cells are withdrawn from the wells and analysed via flow cytometry. A preliminary step to the analysis is the cells' staining: cells are labelled with fluorescence antibodies for the target markers so that corresponding fluorescence intensities can be measured by the flow cytometer and thus cellular populations can be tracked. Unfortunately, a loss of cells has been observed during the staining procedure; in order to quantify it, some beads are added at the beginning of the staining. The collected number of beads out of the initial amount represents a correction factor for the total number of cells reported for each well.

The optogenetic experiment considered within this PhD project was repeated seven times between August and December 2020. In each experiment repeat, cells placed in different wells have been stimulated for different lengths of time. In particular, at the beginning of the optogenetic experiment, all the wells are placed in the dark (LOVTRAP domain bound); the time when light is switched on (antigen signalling interrupted) differs for each of the 96 wells of the plate. The provided data consider 36 different signal lengths administrated among the total duration of 48 hours. Regardless of the length of the light period, cells have been collected and analysed via flow cytometry only at the end of the 48 hours of the optogenetic experiment. Percentages of TIM-3$^-$TCF1$^+$ and TIM-3$^+$TCF1$^-$ exhausted population of cells have been measured for different antigen stimulation periods. The question of how the length of the antigen stimulus, and thus the duration of the dark delay, affects T cell differentiation towards an exhausted state, has been addressed by Dr. Mann.

Chapter 2

Mathematical background

This Chapter details the important concepts, definitions and methods about the theory of stochastic processes and Markov chains which are applied in this thesis to represent the intrinsic randomness involved in biological systems such as the T cell differentiation process and thymocytes development. In Sections 2.1 and 2.2 of this Chapter, we report some of the most relevant concepts in probability theory and the theory of stochastic processes, respectively. For large populations of cells, random fluctuations can typically be ignored and deterministic models can be used to study such biological systems. Experimental data sets and quantitative models can be brought together with Bayesian inference to quantify and identify molecular and cellular mechanisms or timescales. We give in Section 2.3 an overview of the statistical techniques which will be used in this thesis. Also, in Section 2.4 analytical methods to analyse the parameter space are reported. Finally, we present in Section 2.5 the key concepts and methods that are applied to study the differential equations arising in both the deterministic and stochastic processes of interest for this thesis.

2.1 Probability theory

In Chapters 3, 4 and 5 of this thesis, we use branching processes or continuous-time Markov chains to model cell differentiation. Thus, we present in this Section the probability theory required to understand such stochastic processes. In Section 2.2, stochastic processes will be introduced.

2.1.1 Random variables, independence and conditional probability

In probability theory, a *sample space* S is a set, a collection of objects. Each element of S is called a *sample point* and each subset of S is referred to as an *event*. For example, when tossing a coin we can get head (H) or tail (T). The sample space is $\{H, T\}$ and an event may be $\{H\}, \{T\}, \{H, T\}$ and $\emptyset$. In general, for any experiment, the sample space is the set of possible results from the experiment. Moreover, we can formally define the probability space underlying the random experiment.

Definition 2.1 (Properties of $\mathcal{A}$)**.** Let S be a sample space and $\mathcal{A}$ a collection of subsets in S. Then, $\mathcal{A}$ is called the σ-algebra of events and has the following properties:

- $S \in \mathcal{A}$,
- if $B \in \mathcal{A} \Rightarrow B^C \in \mathcal{A}$ where B^C denotes the complement of B,

$$B^C = \{s \,:\, s \in S,\, s \notin B\} \in \mathcal{A},$$

- if $B_n \in \mathcal{A}$ for $n = 1, 2, 3, \ldots \quad \Rightarrow \quad \bigcup_{n=1}^{+\infty} B_n \in \mathcal{A}$.

Definition 2.2 (Properties of $\mathbb{P}$)**.** The function $\mathbb{P} : \mathcal{A} \to [0, 1]$ defined on the σ-algebra $\mathcal{A}$ is called the *probability measure* or *probability function*. In particular, $\mathbb{P}$ and has the following properties:

- $\mathbb{P}(S) = 1$,
- if $\{B_1, B_2, \ldots\}$ are mutually exclusive events in $\mathcal{A}$, that is $B_i \cap B_j = \emptyset$ for $i \neq j$ and $i, j = 1, 2, \ldots$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{+\infty} B_i\right) = \sum_{i=1}^{+\infty} \mathbb{P}(B_i).$$

Definition 2.3. A *probability space* describing a random phenomenon is a triple $(S, \mathcal{A}, \mathbb{P})$, where S is the sample space, $\mathcal{A}$ is the collection of subsets or events in S, and $\mathbb{P}$ is a probability function defined on $\mathcal{A}$.

We now introduce the concepts of conditional probability and independence.

Definition 2.4 (Conditional probability)**.** Let $(S, \mathcal{A}, \mathbb{P})$ be a probability space and $B_1, B_2 \in \mathcal{A}$ two events. Given $\mathbb{P}(B_2) > 0$, the conditional probability of event B_1 given event B_2 is defined as

$$\mathbb{P}(B_1|B_2) = \frac{\mathbb{P}(B_1 \cap B_2)}{\mathbb{P}(B_2)}.$$

Definition 2.5. Let $(S, \mathcal{A}, \mathbb{P})$ be a probability space and $B_1, B_2 \in \mathcal{A}$ two events; B_1, B_2 are said to be independent if and only if

$$\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1)\mathbb{P}(B_2).$$

In other words, the occurrence of either one of the events does not affect the probability of occurrence of the other, $\mathbb{P}(B_1|B_2) = \mathbb{P}(B_1)$.

Theorem 1 (Law of total probability)**.** *Let $\{B_1, \ldots, B_n\}$ be a partition of a sample space S, then for any event A in the sample space S,*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

Theorem 1 has a key importance in the formulation of the first-step analysis carried out in Sections 5.1.2, 5.1.3. We can now formulate the Bayes theorem which is the most important concept used in Section 2.3 for Bayesian methods.

Theorem 2 (Bayes theorem)**.** *Let $\{B_1, \ldots, B_n\}$ be a partition of a sample space S, and let A be an event in the sample space S such that $\mathbb{P}(A) > 0$. Then, for $i = 1, \ldots, n$, the conditioned probability of B_i given event A is*

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}.$$

A central concept of probability theory is a random variable whose value depends on the possible outcomes in S.

Definition 2.6. Let $(S, \mathcal{A}, \mathbb{P})$ be a probability space. A *random variable* X is a real-valued function defined on the sample space S, $X : S \to \mathbb{R}$, such that $\forall a \in \mathbb{R}$

$$X^{-1}(-\infty, a] = \{s \in S : X(s) \leq a\} \in \mathcal{A}.$$

The *range* of X, also called *state space* or *support* of the random variable X is defined as the set of all the possible values that X can take

$$S_X = \{x \in \mathbb{R} : X(s) = x,\, s \in S\},$$

The random variable X is called a *discrete random variable* if the state space S_X is finite or countably infinite; if the state space is uncountably infinite, the random variable X is said to be a *continuous random variable*.

2.1.2 Cumulative distribution function and probability mass and density function

In this Section, we define the cumulative distribution which describes the probability that a random variable with a given probability distribution will be found at a value less than or equal to a certain value, as well as the probability mass and density functions which define the probability measure for a discrete and continuous random variable, respectively.

Definition 2.7. The *cumulative distribution function* (c.d.f.) of the random variable X is the function $F_X : \mathbb{R} \to [0, 1]$, defined by

$$F_X(x) = \mathbb{P}(X \in (-\infty, x]).$$

The function F_X is non-decreasing, right continuous, and satisfies

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Definition 2.8. If X is a discrete random variable, the *probability mass function* (p.m.f.) of X is defined, for each x in the state space of X, as the function $f_X : S_X \to [0, 1]$ such that

$$f_X(x) = \mathbb{P}(X = x).$$

The c.d.f. F_X of a discrete random variable X satisfies

$$F_X(x) = \sum_{a \in S_X,\, a \leq x} f_X(a)$$

where $F_X(x) = 0$ if $x < \inf\{a \in S_X\}$.
Similarly, if X is a continuous random variable, the *probability density function* (p.d.f.) of X is a non-negative, integrable function $f_X : \mathbb{R} \to [0, +\infty]$ such that, for $x \in \mathbb{R}$,

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy$$

where F_X is the cumulative distribution function of X.

When several random variables are associated to the same sample space, then $\mathbf{X} = (X_1, \ldots, X_n)$ is called *multivariate random variable* or *random vector*; where $X_i : S \to \mathbb{R}$ are n real-valued random variables. If each X_i is a discrete random variable, then $\mathbf{X}$ is called *discrete random vector*; similarly, for the continuous case. If the distribution of a single discrete random variable X can be characterised in terms of its probability mass function, when studying random vectors we can define the marginal probability mass function and the joint probability mass function, which characterise the distribution of the single discrete variable X_i and the joint distribution of all the entries of the random vector considered together, respectively. Similarly, the joint probability density function is defined for continuous random vectors.

Definition 2.9. If $\mathbf{X} = (X_1, \ldots, X_n)$ is a discrete random vector, the *joint cumulative distribution function* is defined as

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\left(X_i \leq x_i,\, i = 1, \ldots, n\right),$$

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $x_i \in S_{X_i}$. Moreover, the *joint probability mass function* is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\left(X_i = x_i,\, i = 1, \ldots, n\right), \quad x_i \in S_{X_i}\,.$$

For any $i = 1, \ldots, n$ and $x_i \in S_{X_i}$, the *marginal probability mass function* of X_i is defined as the sum of the joint p.m.f $f_{\mathbf{X}}(\mathbf{x})$ over the vector $\mathbf{x}$ such that $x_j \in S_{X_j}$ for all $j = 1, \ldots, n$ and their ith component is equal to x_i.

For a discrete random variable $\mathbf{X} = (X_1, X_2)$, that is

$$f_{X_1}(x_1) = \sum_{x_2 \in S_{X_2}} f_{\mathbf{X}}(x_1, x_2)\,.$$

Similarly, for continuous random variables.

Definition 2.10. A continuous random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to be a jointly continuous random vector if there exists a function $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$, called *joint probability density function*, such that

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \int_{y_n \le x_n} \cdots \int_{y_1 \le x_1} f_{\mathbf{X}}(y_1, y_2, \ldots, y_n)\, dy_1 \ldots dy_n\,.$$

For any $i = 1, \ldots, n$ and $x_i \in S_{X_i}$, the *marginal probability density function* of X_i is defined as the integral of the joint p.d.f. with respect to all the variables except x_i.

For a continuous random variable $\mathbf{X} = (X_1, X_2)$, that is

$$f_{X_1}(x_1) = \int_{S_{X_2}} f_{\mathbf{X}}(x_1, x_2) dx_2\,.$$

2.1.3 Expectation, standard deviation and covariance

The expectation, the standard deviation and the covariance are important concepts that help characterise the probability distribution of a random variable.

Definition 2.11. If X is a discrete random variable with p.m.f. f_X defined on state space S_X, then the expectation of X is defined as

$$\mathbb{E}(X) = \sum_{x \in S_X} x f_X(x).$$

Similarly, if X is a continuous random variable with p.d.f. f_X defined on state space S_X, then the expectation of X is defined as

$$\mathbb{E}(X) = \int_{S_X} x f_X(x)\, dx.$$

Given a random variable X, the expectation $\mathbb{E}(X^k)$ for $k \in \mathbb{N}$ is called the kth *moment* of X. The kth moment about the mean (or kth central moment) is defined as,

$$\mathbb{E}((X - \mathbb{E}(X))^k).$$

In particular, if $k = 2$ we define the variance.

Definition 2.12. The 2nd order central moment, or *variance*, of a random variable X is

$$\text{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

The *standard deviation* of a random variable X is defined as the square root of the variance, that is $\sigma = \sqrt{\text{var}(X)}$.

Theorem 3 (Law of total expectation)**.** *If X, Y are two random variables, then*

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).$$

In particular, if we condition on a partition $\{B_1, \ldots, B_n\}$ of the sample space S,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X|B_i)\mathbb{P}(B_i).$$

2.1.4 Generating functions

Generally speaking, a generating function is a way of writing a sequence of numbers. They arise in enumeration problems and discrete mathematics following the idea that a single function can encode an infinite sequence. The terms of the sequence are given by the power series of the function. More formally, if we let $\{g_k\}_{k\geq 0}$ be a sequence of numbers, then the generating function associated to this sequence is a function

$$G(z) = \sum_{k=0}^{+\infty} g_k z^k$$

of a symbolic argument $z \in \mathbb{R}$, $z \in [0, 1]$. For a discrete random variable X, the probability mass function $\mathbb{P}(X = k)$ is the element of the sequence g_k we would like to encode within a generating function. In Chapters 3 and 4 we make an extensive use of probability generating functions and their properties. We summarise below their main characteristics and refer to Wilf (2005) and Allen (2010) for more details.

Definition 2.13. If X is a discrete random variable taking non-negative integer values, we can define the *probability generating function* (p.g.f.) of X as

$$\phi_X(z) = \mathbb{E}(z^X) = \sum_{k=0}^{+\infty} \mathbb{P}(X = k)\, z^k .$$

where $z \in [0, 1]$ is symbolic argument.

If $\phi_X(z)$ is the probability generating function of a discrete random variable X, we can derive the p.m.f. by taking the kth derivative of ϕ_X,

$$\mathbb{P}(X = k) = \frac{1}{k!} \left(\frac{d^k}{dz^k} \phi_X(z) \right) \Bigg|_{z=0}, \quad k \geq 0, \quad z \in [0, 1] .$$

Moreover, we can express the normalisation of the p.m.f. of a random variable X in terms of its probability generating function as

$$\phi_X(1) = \sum_{k=0}^{+\infty} \mathbb{P}(X = k) = 1 .$$

Finally, the *kth factorial moment* of X is defined as

$$\mathbb{E}(X^k) = \mathbb{E}(X(X-1)\dots(X-k+1)) .$$

Consequently, given that $\phi''_X(1) = \mathbb{E}(X^2) - \mathbb{E}(X)$, the *variance* of a random variable X is given by (Grimmett & Welsh, 2014)

$$\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \phi''_X(1) + \phi'_X(1) - (\phi'_X(1))^2 .$$

If $\mathbf{X} = (X_1, \dots, X_n)$ is a multivariate discrete random variable, then the probability generating function of $\mathbf{X}$ is defined as

$$\phi_{\mathbf{X}}(z_1, \dots, z_n) = \sum_{k_1,\dots,k_n=0}^{+\infty} p_{\mathbf{X}}(k_1, \dots, k_n) z_1^{k_1} \dots z_n^{k_n} \tag{2.1}$$

where $p_{\mathbf{X}}(k_1, \dots, k_n)$ is the joint probability mass function of $\mathbf{X}$ and $z_i \in [0, 1]$ symbolic arguments, for $i = 1, \dots, n$.

2.1.5 Laplace-Stieltjes transform

In this Section, we report the definition of the Laplace-Stieltjes transform of a random variable and its different order moments (see Gross *et al.* (2008) for more details). An application can be found in Section 5.1.3.

Definition 2.14. Let X be a non-negative real-valued continuous random variable with probability density function f_X. The Laplace-Stieltjes transform (LST) of X is defined as

$$\mathcal{L}^\star[f_X](s) = \mathbb{E}(\mathrm{e}^{-sX}) = \int_0^{+\infty} \mathrm{e}^{-sx} f_X(x)\, dx\,,$$

for $s \geq 0$ symbolic argument. Note that, for $s = 0$, it holds $\mathcal{L}^\star[f_X](0) = \mathbb{P}(X < +\infty) = 1$, since f_X is a probability density function. Also, by taking the derivatives of the LST, we can find the kth order moment of X as

$$\mathbb{E}(X^k) = (-1)^k \left(\frac{d^k}{ds^k}\mathcal{L}^\star[f_X](s)\right)\Big|_{s=0}, \quad \text{for} \quad k \geq 1.$$

2.2 Stochastic processes

In this Section, stochastic processes are introduced; the theory is based mainly on the books Allen (2010), Bailey (1964), Kimmel & Axelrod (2002) and Kulkarni (2017).

Definition 2.15. A *stochastic process* is a collection of random variables

$$\mathcal{X} = \{X(t),\, t \in T\}\,,$$

on the common sample space S and indexed by the parameter $t \in T$. The random variables take values in a common set S_X, called the state space of the stochastic process. For a fixed $t \in T$, $X(t)$ denotes a single random variable defined on S. For a fixed element of S, a *sample path* or *stochastic realisation* of the process is $\{X(t), t \in T\}$ defined as a function on T.

Depending on whether the random variables and the index set are discrete or continuous, stochastic processes are usually classified into three different types:

both index set and the random variables are discrete-valued; the index set is continuous but the random variables are discrete-valued; both index set and random variables are continuous-valued. For the purpose of this thesis, we consider the first two types of stochastic processes, with discrete-valued random variables.

Throughout this thesis, the parameter t represents time; for example the random variable $X(t)$ can represent the number of births in a given population during the time period $[0, t]$. In this case, $T = \mathbb{R}^+ = [0, +\infty)$ and the state space is $\{0, 1, 2, \dots\}$ for a continuous-time stochastic process; or $T = \mathbb{N} \cup \{0\}$ in case of a discrete-time stochastic process.

Markov chains are a particular type of stochastic processes in which the current state of the process depends only on the immediately previous state and not on any other previous ones. A relevant type of stochastic processes, mainly used to model reproduction, are branching processes. In particular, a branching process is a Markov chain that models a population in which each individual in generation n produces a random number of offspring in generation $n+1$. Branching processes have been originally defined to model survival of family names but find numerous applications in biology, immunology, ecology, medicine and epidemiology (Kimmel & Axelrod, 2002).

In Section 2.2.1, we aim to provide the reader with the mathematical background about continuous-time Markov chains; in Section 2.2.3, a background about discrete-time branching processes is given. These mathematical techniques are then used in Chapters 3, 4 and 5 as a modelling tool for the cellular differentiation processes such as the ones described in Chapter 1.

2.2.1 Continuous-time Markov chains

Among stochastic processes, continuous-time Markov chains are widely used in biological applications as they can be theoretically and computationally analysed. The idea is to represent the size of a population by a continuous or discrete random variable where the time is continuous (the system can change state at any time). We stress here that, by convention in the theory of stochastic processes, we usually refer to *Markov chain* if the state space is discrete, and to *Markov process* when the state space is continuous. Following Norris (1997), Allen (2010), Bailey (1964)

and Kulkarni (2017), we recall below the main mathematical results as well as the key properties of continuous-time Markov chains.

Definitions and main results

Definition 2.16. Let $\mathcal{X} = \{X(t) : t \geq 0\}$ be a set of continuous-time random variables taking values in a discrete state space S_X. The stochastic process $\mathcal{X}$ is a *continuous-time Markov chain* (CTMC) if, for any sequence of real numbers satisfying $t_0 < t_1 < \cdots < t_{n+1}$,

$$\mathbb{P}\left(X(t_{n+1}) = i_{n+1} \,|\, X(t_n) = i_n, \ldots, X(t_0) = i_0\right) = \mathbb{P}\left(X(t_{n+1}) = i_{n+1} \,|\, X(t_n) = i_n\right),$$

where $i_k \in S_X$ for $k = 0, 1, 2, \ldots, n+1$. This is known as the *Markov property*.

Each random variable $X(t)$ has an associated probability distribution $\{p_i(t)\}_{t\geq 0}$ for $i \in S_X$, where $p_i(t) = \mathbb{P}\left(X(t) = i\right)$. Moreover, we can relate random variables at different times via transition probabilities $p_{ij}(s,t)$ indicating the probability that the Markov chain moves from state i at time s to state j at time t,

$$p_{ij}(s,t) = \mathbb{P}\left(X(t) = j \,|\, X(s) = i\right), \quad \text{for} \quad s < t \quad \text{and} \quad i, j \in S_X\,.$$

Definition 2.17. If the transition probabilities do not depend explicitly on the different times s, t but only on the length of the time interval $t - s$, they are called *stationary* or *homogeneous* transition probabilities. In this case, for $s < t$,

$$p_{ij}(t-s) = \mathbb{P}\left(X(t) = j \,|\, X(s) = i\right) = \mathbb{P}\left(X(t-s) = j \,|\, X(0) = i\right).$$

The matrix $\boldsymbol{P}(t) = [p_{ij}(t)]_{i,j\in S_X}$ is called *transition probability matrix*.

Theorem 4 (Characterisation of a CTMC)**.** *Let $\mathcal{X} = \{X(t),\ t > 0\}$ a CTMC and $\boldsymbol{a} = [a_i]_{i\in S_X}$ its initial distribution such that $a_i = \mathbb{P}\left(X(0) = i\right)$ for $i \in S_X$. Then $\mathcal{X}$ is completely described by the vector $\boldsymbol{a}$ and the matrices $\boldsymbol{P}(t)$ for all time $t \geq 0$.*

If a fixed initial distribution $\boldsymbol{a}$ of process $\mathcal{X}$ is given, that is $X(0) = i$ and $a_j = \delta_{ij}$, then the transition probability $p_{ij}(t)$ is the same as the state probability $p_j(t) = \mathbb{P}\left(X(t) = j \,|\, X(0) = i\right)$ for $i, j \in S_X$.

The process $\mathcal{X}$ is said *non-explosive* if the transition probabilities satisfy

$$\sum_{j \in S_X} p_{ij}(t) = 1, \quad t \geq 0 \quad \text{and} \quad i \in S_X \,.$$

In this case, the matrix $\boldsymbol{P}(t)$ is a *stochastic matrix*, that is all the elements of a row sum to 1. For more details about explosive processes we refer to Norris (1997); in this thesis, we will consider only homogeneous and non-explosive Markov chains.

Definition 2.18. A state $i \in S_X$ is said *absorbing state* for the Markov chain $\mathcal{X}$ if, for all $t \geq 0$, $p_{ii}(t) = 1$. Thus, once the state is entered, it is impossible to leave.

Transition matrix and Kolmogorov equation

A Markov chain is described by the set of its transition probabilities $p_{ij}(t)$ for each time $t \geq 0$. The transition matrix $\boldsymbol{P}(t)$ of a CTMC can be found by solving the so called Kolmogorov equation; this is a differential equation expressing the rate of change of the transition probabilities. Before defining the backward Kolmogorov equations, we need to introduce the generator matrix $\boldsymbol{Q}$ containing the infinitesimal transition rates q_{ij}. We assume the transition probabilities $p_{ij}(\Delta t)$ are continuous and differentiable for $t \geq 0$, consider a small time step Δt and derive the transition rates q_{ij} as follows; for $t > 0$ and $i, j \in S_X$,

$$q_{ij} = \lim_{\Delta t \to 0} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t} = \begin{cases} \lim\limits_{\Delta t \to 0} \dfrac{p_{ij}(\Delta t)}{\Delta t} & \text{if} \quad i \neq j \\ \lim\limits_{\Delta t \to 0} \dfrac{p_{ii}(\Delta t) - 1}{\Delta t} & \text{if} \quad i = j \end{cases}$$

given that $p_{ij}(0) = 1$ for $i = j$ and zero otherwise. Since $\boldsymbol{P}(t)$ is a stochastic matrix, by definition,

$$1 - p_{ii}(\Delta t) = \sum_{j \neq i} p_{ij}(\Delta t) = \sum_{j \neq i} (q_{ij} \Delta t + o(\Delta t)), \quad \text{for} \quad i, j \in S_X \,.$$

Thus, for $i = j$,

$$q_{ii} = \lim_{\Delta t \to 0} \frac{p_{ii}(\Delta t) - 1}{\Delta t} = \lim_{\Delta t \to 0} \frac{-\sum_{j \neq i} (q_{ij} \Delta t + o(\Delta t))}{\Delta t} = -\sum_{j \neq i} q_{ij} \,.$$

If the limit is finite, we get a relationship between $p_{ij}(\Delta t)$ and q_{ij}

$$p_{ij}(\Delta t) = q_{ij}\Delta t + o(\Delta t) + \delta_{ij}\,, \tag{2.2}$$

where δ_{ij} is the Kronecker's delta symbol. The generator matrix $\boldsymbol{Q}$ is defined as

$$\boldsymbol{Q} = \lim_{\Delta t \to 0^+} \frac{\boldsymbol{P}(\Delta t) - \boldsymbol{I}}{\Delta t}\,,$$

where $\boldsymbol{I}$ is the identity matrix of the infinitesimal transition matrix $\boldsymbol{P}(\Delta t)$.

Given Theorem 4, we can describe a CTMC by specifying its transition matrix $\boldsymbol{P}(t)$. Assuming $\boldsymbol{P}(t)$ is differentiable with respect to t, one can derive a set of differential equations whose solution gives the probability distribution of $X(t)$ for all the times.

Theorem 5 (Backward Kolmogorov equation)**.** *Let $\boldsymbol{P}(t)$ be the matrix of transition probabilities of a CTMC $\mathcal{X}$ with state space S_X and generator matrix $\boldsymbol{Q}$. Then, if $\boldsymbol{P}(t)$ is differentiable, the backward Kolmogorov differential equations are represented by the system*

$$\frac{d}{dt}p_{ij}(t) = \sum_{k \in S_X} q_{ik}\, p_{kj}(t), \quad i, j \in S_X\,.$$

Expressed in matrix form, we have

$$\frac{d}{dt}\boldsymbol{P}(t) = \boldsymbol{Q}\,\boldsymbol{P}(t)\,.$$

This system of equations is often referred to as the master equation (ME) or the chemical master equation (CME) of the stochastic process.

Proof. To derive this set of differential equations, we observe that the transition probabilities of the matrix $\boldsymbol{P}(t)$ satisfy

$$p_{ij}(t+s) = \sum_{k \in S_X} p_{ik}(s)\, p_{kj}(t), \quad \text{for } i, j \in S_X \text{ and } t, s \geq 0.$$

We differentiate probabilities $p_{ij}(t)$ with respect to time and apply Equation (2.2); for $i, j \in S_X$,

$$\begin{aligned}\frac{d}{dt}p_{ij}(t) &= \lim_{\Delta t \to 0} \frac{p_{ij}(t+\Delta t) - p_{ij}(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\sum_{k \in S_X} p_{ik}(\Delta t) p_{kj}(t) - p_{ij}(t)}{\Delta t} = \\ &= \lim_{\Delta t \to 0} \frac{\sum_{k \in S_X} \left(q_{ik}\Delta t + o(\Delta t) + \delta_{ik}\right) p_{kj}(t) - p_{ij}(t)}{\Delta t} = \sum_{k \in S_X} q_{ik}\, p_{kj}(t)\,.\end{aligned}$$

□

If a fixed initial distribution $\boldsymbol{a}$ of process $\mathcal{X}$ is given, that is $X(0) = i$ and $a_j = \delta_{ij}$, then we can express the Kolmogorov differential equations in terms of the state probabilities

$$\frac{d}{dt}\boldsymbol{p}(t) = \boldsymbol{Q}\,\boldsymbol{p}(t)$$

where $\boldsymbol{p}(t) = (p_0(t), p_1(t), \dots)^T$.

For finite state space continuous-time Markov chain, the Kolmogorov equations have a unique solution and thus it is possible to find an explicit solution $\boldsymbol{P}(t)$. In this case, the following theorem holds.

Theorem 6. *The transition probability matrix of a CTMC $\mathcal{X}$ on a finite state space S_X and with infinitesimal generator matrix $\boldsymbol{Q}$, is given by*

$$\boldsymbol{P}(t) = \mathrm{e}^{\boldsymbol{Q}t}, \quad t \geq 0\,.$$

However, in general, specifying the matrix $\boldsymbol{P}(t)$ can be quite hard (both analytically and numerically); another method (applied in Chapters 3 and 4) to obtain information about the probability distribution associated with a CTMC relies on partial differential equations and generating functions (see Section 2.1.4).

The Kolmogorov differential equations can be used to define a stationary probability distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)^T$ with $\pi_i \geq 0$ for $i \in S_X$ and such that

$$\boldsymbol{Q}\,\boldsymbol{\pi} = 0 \quad \text{and} \quad \sum_{i \in S_X} \pi_i = 1\,.$$

If the process is non-explosive, irreducible and positive recurrent, it can be proved that the stationary distribution is unique and positive (Allen, 2010). If the process has an absorbing state, a quasi-stationary distribution can be defined; prior to reaching the absorbing state, the probability distribution of $X(t)$ can be approximately stationary for a long period of time. In this case, we define $q_i(t)$ the quasi-stationary probability distribution conditioned on not-reaching the absorbing state $a \in S_X$ as, for $t \geq 0$,

$$q_i(t) = \frac{p_i(t)}{1 - p_a(t)}, \quad \text{for} \quad i \in S_X,\, i \neq a\,,$$

where we denote with $p_a(t)$ the state probability of reaching the absorbing state a given an initial distribution of the process $\mathcal{X}$.

Inter-event times

Knowing the distribution for the time between successive event, one can calculate the sample paths of a CTMC $\mathcal{X}$. In a discrete-time Markov chain, the state space S_X is a finite or infinite discrete set, such as $\{1, 2, \ldots, N\}$ or $\{1, 2, \ldots, +\infty\}$; whereas, in a continuous-time Markov chain the transition to a new state may occur at any time $t \geq 0$. In both cases, however, we can represent the progression of time as made by a "jump" from the current state to the following one.

Definition 2.19. Let W_n be the random variable for the time of the nth change of state of the process. Assuming $W_0 = 0$, the collection $\{W_n\}_{n=0}^{+\infty}$ is called *jump times* or *waiting times* of the process. Moreover, the random variable $T_n = W_n - W_{n-1}$ is called the nth *interevent time* or *sojourn time*.

One of the key properties of Markov processes is that their interevent times T_n are exponentially distributed with parameter $\sum_{j \in S_X, j \neq i} q_{ij}$, where q_{ij} represent the transition rates. We recall that the expectation of an exponentially distributed random variable with parameter λ is given by $\frac{1}{\lambda}$. Thus, assuming a CTMC $\mathcal{X}$ is in state i at the nth jump, the mean of random variable T_n is given by

$$\mathbb{E}(T_n) = \frac{1}{\sum_{j \in S_X, j \neq i} q_{ij}} .$$

The analytical explanation of interevent times of a Markov chain being exponentially distributed is based on the fact that the exponential distribution is the only continuous probability distribution for which the so called memoryless property holds (Allen, 2010):

$$\mathbb{P}\left(T_n \geq t + s \,|\, T_n \geq t\right) = \mathbb{P}\left(T_n \geq s\right), \quad \text{for any} \quad t, s \geq 0$$

First-step analysis

First-passage analysis aims to break down the possibilities resulting from the first-step (first transition) in the Markov chain $\mathcal{X}$. The law of total probability (see Section 2.1.1) and Markov property can be used to derive a set of relationships among some unknown variables of interest. For example, let $\mathcal{X} = \{X(t) : t \geq 0\}$

be a CTMC on the state space $S_X = \{0, 1, 2, \dots, \}$ and consider the random variable

$$T = \inf\{t \geq 0 : X(t) = 0\};$$

that is the first time at which the process $\mathcal{X}$ enters into state 0, for some initial state $X(0) \neq 0$. T is called the *first-passage time* to 0 and represents the first time at which the variable $X(t)$ is equal to 0. The same techniques can be used to study the first-passage time to any set $A \subset S_X$.

Given the random variable T one can study its cumulative distribution function and its expectation; moreover, the probability of reaching the state of interest (state 0 in this case) can be computed. Pioneer works focusing on computing first passage times in systems with discrete states are Castro *et al.* (2018); de la Higuera *et al.* (2019); Gómez-Corral *et al.* (2020); López-García *et al.* (2018). An application of these techniques can be found in Sections 5.1.3 for a stochastic compartment model of cell differentiation.

2.2.2 Birth-and-death process

Birth-and-death processes are one of the most well studied continuous-time Markov chains. Let $\mathcal{X} = \{X(t) : t \geq 0\}$ be a CTMC on state space S_X which might be either infinite $S_X = \{0, 1, 2, \dots, +\infty\}$ or finite $S_X = \{0, 1, 2, \dots, N\}$. Now, the random variable $X(t)$ represents the population size at time t. We assume there are only two types of events: a birth of an individual or a death. If the process $\mathcal{X}$ is in state n, a birth happens at rate λ_n and moves the process to state $n + 1$; whereas a death, happening at rate μ_n, moves the process to state $n-1$, as shown in Figure 2.1.

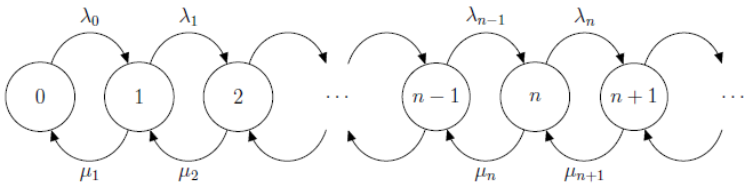


Figure 2.1: State diagram for a birth-and-death process.

Given the random variable $X(t)$ we denote the change of state in the process with

$$\Delta X(t) = X(t + \Delta t) - X(t) \quad \text{for} \quad t \geq 0\,;$$

thus, the infinitesimal transition probabilities, for $i, j \in S_X$, are

$$\begin{aligned} p_{ij}(\Delta t) &= \mathbb{P}\left(\Delta X(t) = j \middle| \, X(t) = i\right) \\ &= \begin{cases} \lambda_i \Delta t + o(\Delta t) & \text{for} \quad j = i + 1 \\ \mu_i \Delta t + o(\Delta t) & \text{for} \quad j = i - 1\,, \\ 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t) & \text{for} \quad j = i \\ o(\Delta t) & \text{otherwise} \end{cases} \end{aligned}$$

for Δt small enough and where $o(\cdot)$ is the Landau order symbol.
The infinitesimal generator matrix $\boldsymbol{Q}$ for the process $\mathcal{X}$ is a tridiagonal matrix and is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Moreover, one can consider immigration happening at rate ν. In this case, the infinitesimal transition probabilities are

$$\begin{aligned} p_{ij}(\Delta t) &= \mathbb{P}\left(\Delta X(t) = j \middle| \, X(t) = i\right) \\ &= \begin{cases} (\nu + \lambda_i)\Delta t + o(\Delta t) & \text{for} \quad j = i + 1 \\ \mu_i \Delta t + o(\Delta t) & \text{for} \quad j = i - 1\,. \\ 1 - (\nu + \lambda_i + \mu_i)\Delta t + o(\Delta t) & \text{for} \quad j = i \\ o(\Delta t) & \text{otherwise} \end{cases} \end{aligned}$$

2.2.3 Branching processes

Branching processes are a relevant type of stochastic processes, mainly used to model reproduction. In particular, a branching process is a Markov chain that models a population in which each individual in generation n produces a random number of offspring who belongs to generation $n + 1$. Branching processes were

originally defined to model survival of family names but find numerous applications in biology, immunology, ecology, medicine and epidemiology (Kimmel & Axelrod, 2002). The study of branching processes was introduced in 1873 when the mathematician Reverend Henry William Watson and the biometrician Francis Galton started a study about the survival of British family names in aristocracy. In particular, they were wondering about the proportion of surnames that will become extinct after r generations as well as about the number of instances of the same surname being held by m individuals (Allen, 2010).

When working with discrete-time branching processes and considering the progeny of an individual, we follow the common notation in the literature and use the index n instead of t, i.e. random variable X_n. Let $\{X_n : n \geq 0\}$ be a family of non-negative integer-valued random variables defined on a common state space S_X. We consider a single ancestor individual present at time $n = 0$ and we look at the number of its progeny produced at its death. Both the lifespan of the individual and the number of offspring can be considered as random variables. Then, the number of individuals present in the process at time $n > 0$ is equal to the sum of the progeny of the ancestor and their subsequent progeny. If each individual's lifespan is a fixed length of time, e.g. one unit of time for convenience, the branching process is known as a Galton-Watson process. Another important special case is when the lifespan is exponentially distributed or it is an arbitrary non-negative random variable, in the latter case the process is known as Bellman–Harris process (Harris, 1963; Kimmel & Axelrod, 2002). Regardless of the lifespan, multi-type branching processes can be defined if different types of individuals are considered within the same population.

A Galton-Watson branching process, as introduced in the XIX century, is a discrete-time Markov process defined as follows.

Definition 2.20. A *Galton-Watson branching process* follows the assumptions:

- Each individual in generation n gives birth to X_n offspring in the next generation, where X_n is a random variable that takes values in the countable state space $\{0, 1, 2, \dots\}$, with offspring distribution $\{p_k = \mathbb{P}(X_n = k)$, for $k = 0, 1, 2, \dots\}$.

- Each individual gives birth independently from all other individuals.

- The same offspring distribution applies to all n generations, $X_n = X$.

A discrete-time branching process can be used to describe the number of individual in one generation. Let Z_n be the random variable describing the number of individuals in the nth generation. We have

$$Z_{n+1} = \sum_{i=0}^{Z_n} X_i$$

where the random variable X_i is an independent realisation of X and represents the offspring of the ith member of the nth generation. Notice that the number Z_n of random variables that we sum is itself a random variable.

In a Galton-Watson process the lifespan of an individual is fixed to one unit, however, one can also consider the case of the lifespan being a random variable. If so, there is a probability that any individual in generation n dies and leaves X_n offspring, in a time interval of duration Δt. Throughout this thesis, we will focus on discrete-time and continuous-time branching processes where the lifespan is an exponentially-distributed random variable; an application to the cellular differentiation process can be found in Chapters 3 and 4, respectively.

In the single-type Galton-Watson process, each individual is of the same type and is replaced by its progeny. In order to model biological processes, we may need to define different types of individuals within the same population: a multi-type branching process allows for different kinds of individuals in each generation. For example, in the context of a Galton-Watson process, individuals might not be all alike: old men are likely to generate less offspring that younger ones.

Definition 2.21. For a branching process containing k types of individuals, we define a continuous-time *multi-type branching process* $\{\mathbf{Z}(t),\, t \geq 0\}$ where $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_k(t))$ and where $Z_i(t)$ is the number of individuals of type i, for $i = 1, \ldots, k$, at time $t \geq 0$.

2.2.4 Gillespie algorithm

Based on their property of an exponential waiting time distribution, a numerical method to generate stochastic realisations of CTMC was developed in Gillespie

(1976, 1977). Monte Carlo simulation methods are a class of computational algorithms relying on random numbers to generate a realisation of the process. In particular, the stochastic simulation algorithm (SSA), usually called Gillespie algorithm, uses two different random numbers to compute the time the system stays in the current state as well as which reaction occurs next (Erban *et al.*, 2007; Gillespie, 1976).

Let $\mathcal{X} = \{X(t) : t \geq 0\}$ be a CTMC defined on state space S_X and with infinitesimal generator matrix $\boldsymbol{Q} = [q_{ij}]_{i,j \in S_X}$. Algorithm 1 outlines the Gillespie algorithm for the process $\mathcal{X}$.

2.3 Bayesian inference

In this Section, we outline the main concepts of Bayesian inference and introduce the Approximate Bayesian Computation (ABC) algorithm. By a systematic comparison between observed and simulated data, the ABC algorithm is able to overcome the issue of intractable likelihood functions in data modelling. Further, we outline the principles of Bayesian model selection which can be used to decide which model, among the hypothesised ones, is the most viable. Chapter 6 shows an application to a model of $CD8^+$ T cell exhaustion-differentiation process.

Let X be a random variable defined on a sample space S; we can think of it as an outcome of a model which can arise from different hypotheses H_i, for $i = 1, \ldots, h$. We can make use of Bayes theorem (see Theorem 2) and compute the conditioned probability $\mathbb{P}(H_i | X = x^\star)$ for $i = 1, \ldots, h$, known as posterior belief about the hypothesis H_i based on the observed occurrence of $X = x^\star$. Usually, the Bayes theorem is reformulated in a statistical inference framework as follows. Let $\mathbf{X}$ be a discrete random vector that represents the discrete output of our mathematical model (the observed data), and let $\boldsymbol{\theta}$ be a continuum of hypotheses representing the possible values of the parameters of the model. Theorem 2 reads

$$\pi(\boldsymbol{\theta}|\mathbf{X}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{2.3}$$

where $\pi(\boldsymbol{\theta})$ is known as the *prior distribution* about the hypothesis $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbf{X})$ is the *posterior distribution*, and $\pi(\mathbf{X}|\boldsymbol{\theta})$ is the *likelihood*, the probability of obtaining the data $\mathbf{X}$ given the value of the parameters $\boldsymbol{\theta}$. Performing an inference

Algorithm 1: Gillespie algorithm (Gillespie, 1976)

1 Initialise the system. Set the initial conditions $t = 0$ and $i = X(0) = X_0$.

2 Generate two random numbers r_1, r_2 uniformly distributed in $[0, 1)$.

3 Assume that from state i a number of m reactions can occur, each moving the process to another state $(j_1, \ldots, j_m)$.

4 Compute the so called *propensity function* by summing all the transition rates to the states where the process can jump to from i in a single step. For $j \in (j_1, \ldots, j_m)$,

$$R = \sum_{j \neq i} q_{ij} \, .$$

5 Compute the time when the next reaction takes place as $t + \tau$ where τ is given by

$$\tau = \frac{1}{R} \log \frac{1}{r_1}.$$

6 Using the second generated random number, compute which reaction occurs at state i at time $t + \tau$. The lth reaction takes place, moving the process to state j_l, if

$$r_2 \geq \frac{1}{R} \sum_{k=1}^{l-1} q_{ij_k} \quad \text{and} \quad r_2 \leq \frac{1}{R} \sum_{k=1}^{l} q_{ij_k} \, .$$

7 Compute the new state according to the lth reaction that takes place $X(t + \tau) = j_l$. Update the current state of the process to $i = j_l$ and go back to step **1**. Repeat for all the time-steps until the final time $t = t_{max}$ is reached.

study we aim to evaluate the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{X})$, by using quantities on the right-hand-side of Equation (2.3). The prior distribution $\pi(\boldsymbol{\theta})$ encodes the users prior beliefs about the parameters; usually there is not a strong knowledge and thus, a uniform distribution is used. However, in special situations, a more informative distribution can be assumed, e.g. normal distribution. The integral in the denominator of Equation (2.3) corresponds to the probability of generating

the data,

$$\mathbb{P}(\mathbf{X}) = \int_{\boldsymbol{\theta}} \pi(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\,.$$

This integral does not depend on parameters and is very difficult to calculate, so $\mathbb{P}(\mathbf{X})$ is usually assumed as a normalising constant. Instead of Equation (2.3), we can then use the proportionality equation

$$\pi(\boldsymbol{\theta}|\mathbf{X}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta})\,.$$

Since the computation of the likelihood $\pi(\mathbf{X}|\boldsymbol{\theta})$ in a mathematical model often represents a challenge, alternative methods, such as Approximate Bayesian Computation (ABC for short) methods, have been developed to estimate the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{X})$ (Csilléry *et al.*, 2010; Toni *et al.*, 2009).

2.3.1 Approximate Bayesian Computation

Given a mathematical model $\mathcal{M}$ with parameter vector $\boldsymbol{\theta}$, and given the observed data $\mathbf{X}$, the Approximate Bayesian Computation method can infer the posterior distribution for the parameter values by making use of the experimental data and by updating the prior beliefs. The idea is to sample parameters $\boldsymbol{\theta}^\star$ from the prior distribution $\pi(\boldsymbol{\theta})$ and then simulate data from the model as $\mathbf{X}^\star \sim \pi(\mathbf{X}|\boldsymbol{\theta}^\star)$. A comparison between simulated data and the observed experimental data $\mathbf{X}$ can be done by making use of a distance function δ. The simplest form of this algorithm is based on the *rejection algorithm*: if the simulated data are sufficiently close to the observed data, that is the measure distance is smaller than a tolerance threshold ε, then the sample $(\boldsymbol{\theta}^\star, \mathbf{X}^\star)$ is accepted, otherwise it is rejected. Reiterating the method, $n_s \in \mathbb{N}$ samples of the posterior distribution are generated. For $\varepsilon = 0$ the algorithm corresponds to the exact rejection algorithm; for small enough choices of the tolerance rate ε, the ABC method will closely approximate the true posterior, but for too small choices of ε the algorithm leads to a high rejection rate. For more details we refer to Toni *et al.* (2009), Toni & Stumpf (2010). Moreover, given a model $\mathcal{M}$, different types of distance δ can be chosen according to the problem; the common choice is the Euclidean distance,

$$\delta(\mathbf{X}, \mathbf{X}^\star; \mathcal{M}) = \left[\sum_{j=1}^{n} (x_j^\star - x_j)^2\right]^{1/2}$$

where $\mathbf{X} = (x_1, \ldots, x_n)$ is the experimental data set and $\mathbf{X}^\star = (x_1^\star, \ldots, x_n^\star)$ the simulated data. Alternatively, instead of considering the distance between real and simulated data, one can define a distance between summary statistics, $\delta(S(\mathbf{X}^\star), S(\mathbf{X}); \mathcal{M})$.

We report the pseudocode of the ABC rejection algorithm in Algorithm 2.

Algorithm 2: ABC rejection algorithm (Toni *et al.*, 2009)

1 Initialise the system. Choose the sample size n_s, the acceptance threshold ε and the distance function δ. Set $i = 0$.
2 Sample $\boldsymbol{\theta}^\star$ from the prior distribution $\pi(\boldsymbol{\theta})$.
3 Simulate a data set $\mathbf{X}^\star$ with the model $\mathcal{M}$ from $\pi(\mathbf{X}|\boldsymbol{\theta}^\star)$.
4 Compute the distance $\delta(\mathbf{X}, \mathbf{X}^\star; \mathcal{M})$.
5 If $\delta(\mathbf{X}, \mathbf{X}^\star; \mathcal{M}) \leq \varepsilon$, accept and store $\boldsymbol{\theta}^\star$. Set $i = i + 1$.
6 If $i < n_s$ go to Step **2**, otherwise end the algorithm.

2.3.2 Bayesian model selection

Experimental data can be usually described by several mathematical models, with different complexity and dependency in the dynamics. The Bayesian model selection method allows to identify, given a number of potential models, which model structure and set of parameters can better explain the data. In particular, we focus on the ABC-rejection selection method which can assess the similarity between the observed and simulated data by making use of a distance function. Let assume that we have M different candidate models to reproduce a given experimental data set; let us explain the method for $M = 2$. The aim of Bayesian ABC model selection is to determine, via the ABC algorithm, which of the two models, $\mathcal{M}_1$ or $\mathcal{M}_2$, better represents the data. In particular, for $m = 1, 2$, we denote with $\pi(\mathcal{M}_m|\mathbf{X})$ and $\pi(\mathcal{M}_m)$ the marginal posterior distribution and the prior of model $\mathcal{M}_m$, respectively. The Bayes factor can be computed as,

$$B_{12} = \frac{\pi(\mathcal{M}_1|\mathbf{X})/\pi(\mathcal{M}_2|\mathbf{X})}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)} = \frac{\pi(\mathcal{M}_1|\mathbf{X})}{\pi(\mathcal{M}_2|\mathbf{X})},$$

where, in the last step, we assumed uniform priors for the two models $\mathcal{M}_1$ and $\mathcal{M}_2$.

For the general case of M models, parameters set $\boldsymbol{\theta}_m$ as well as a prior distribution $\pi(\mathcal{M}_m)$ specific for each model $\mathcal{M}_m$, for $m = 1, \ldots, M$, need to be defined. In case of no prior information, one can set a discrete uniform prior so that each model has prior probability $\frac{1}{M}$. Running the ABC rejection algorithm with acceptance threshold ε, we can approximate the marginal posterior distribution with the relative probability for model $\mathcal{M}_m$,

$$\pi(\mathcal{M}_m|\mathbf{X}) \approx \frac{f(\mathcal{M}_m)}{n_s}$$

where n_s is the posterior sample size and $f(\mathcal{M}_m)$ is the acceptance frequency for model $\mathcal{M}_m$ (Grelaud *et al.*, 2009). Algorithm 3 summarises the Bayesian model selection algorithm.

Algorithm 3: Bayesian model selection algorithm (Grelaud *et al.*, 2009)

1 Initialise the system. Choose the sample size n_s, the acceptance threshold ε and the distance function δ. Set $i = 0$.
2 Draw $\mathcal{M}_m^\star$ from the prior distribution $\pi(\mathcal{M}_m)$.
3 Sample $\boldsymbol{\theta}^\star$ from the prior distribution $\pi(\boldsymbol{\theta} \,|\, \mathcal{M}_m^\star)$.
4 Simulate a candidate data set $\mathbf{X}^\star$ from $\pi(\mathbf{X} \,|\, \boldsymbol{\theta}^\star, \mathcal{M}_m^\star)$.
5 Compute the distance $\delta(\mathbf{X}, \mathbf{X}^\star; \mathcal{M}_m^\star)$.
6 If $\delta(\mathbf{X}, \mathbf{X}^\star; \mathcal{M}_m^\star) \leq \varepsilon$, accept $(\mathcal{M}_m^\star, \mathbf{X}^\star)$, otherwise reject it. Set $i = i + 1$.
7 If $i < n_s$ go to Step **2**, otherwise end the algorithm.

2.4 Analysis of sensitivity and identifiability

Sensitivity analysis and identifiability analysis are two groups of methods used to determine whether a given model with set of parameters is too complex for the data. Assume to have a model $\mathcal{M}$ to describe some experimental data and, for simplicity of notation, we restrict to a scalar output. If the model has n parameters, we can represent it as a function $f : \boldsymbol{\theta} \to X^\star \in \mathbb{R}$ where $\boldsymbol{\theta} \in \mathbb{R}^n$.

Sensitivity analysis aims in answering the question of how the trajectory of a model output $X^\star$ is influenced by small perturbations of a single or multiple

parameter values. In particular, *local sensitivity* methods look at the individual contribution of a parameter, by computing the partial derivative $\frac{\partial X^\star}{\partial \theta_i}$ for $i = 1, \ldots, n$. Whereas, *global sensitivity* methods, by simultaneously varying parameters, can determine also the relative contribution of groups of parameters.

Moreover, via a parameter identifiability analysis we can investigate how accurately the model parameter values can be determined given the experimental data. In particular, we need to distinguish between a *practical identifiability* that depends on limitations set by the data, and a *structural identifiability* that depends on the structure of the model (Castro & de Boer, 2020).

2.4.1 Local sensitivity

Local sensitivity methods focus on changes of model output with respect to variations of a single parameter. We refer the reader to Section 5.2.3 for a local sensitivity analysis performed for a stochastic model of T cell thymic development process; for this we followed the approach suggested by Gómez-Corral & López-García (2018) and López-García *et al.* (2018).

Given a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^n$, the relative response to parameter θ_i, for $i = 1, \ldots, n$, is measured by the partial derivative

$$\frac{\partial X^\star}{\partial \theta_i},$$

where $X^\star$ is the simulated output. This ratio can be then normalised to a dimensionless values, so that the effect of different parameters on different descriptors can be easily compared.

Definition 2.22. Let $\mathcal{X}$ be a Markov process on state space S_X and assume it is characterised by the parameter rates $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$. Let m be a summary statistic of interest; the *elasticity* of m with respect to parameter θ_i is the ratio

$$\frac{\partial m / \partial \theta_i}{m / \theta_i}.$$

2.4.2 Global sensitivity

Global sensitivity analysis estimates the influence on model output of parameters changes over a large range of parameter space. In this Section, a brief description of the Sobol sensitivity analysis' method, one of the most used methodologies, is given (Saltelli *et al.*, 2008; Sobol, 2001).

Consider a model $\mathcal{M}$ represented by the function $f : \boldsymbol{\theta} \to X^\star \in \mathbb{R}$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n) \in \mathbb{R}^n$ and $X^\star$ the model scalar output. In particular, without loss of generality, we can assume each parameter distributed in $[0, 1]$ and write the mathematical model as

$$X^\star = f(\boldsymbol{\theta})$$

where $\boldsymbol{\theta} \in I^n$ and I^n is the n-dimensional unit hypercube. The idea of the Sobol method is to compute the model output for several set of parameters, sampled within their respective ranges. The variance associated to the simulated outputs is then decomposed into a sum of variances; each of those gives the contribution of a single input parameter to the total variance.

Mathematically, we consider the model output $X^\star$ as a random variable and the parameter vector $\boldsymbol{\theta}$ as a random vector of independent and uniform random variables over I^n. That is $\theta_i \sim U(0, 1)$ and $f : I^n \to \mathbb{R}$. If the function f is integrable over I^n, Sobol idea is to decompose the model into 2^n terms,

$$f(\theta_1, ..., \theta_n) = f_0 + \sum_{i=1}^{n} f_i(\theta_i) + \sum_{1 \le i < j \le n} f_{i,j}(\theta_i, \theta_j) + ... + f_{1,2,...,n}(\theta_1, \theta_2, ..., \theta_n)\,, \quad (2.4)$$

where each term is integrable and it is a function only of the factors in its index. Given a model f, there could be infinite choices for terms of the expansion; Sobol (2001) proves indeed that this decomposition is unique only under the assumption that all the terms have zero mean, i.e.

$$\int_0^1 f_k(\theta_k) d\theta_k = 0\,.$$

This is equivalent to asking the orthogonality between pairs i, j of terms,

$$\int_0^1 f_i(\theta_i) f_j(\theta_j) d\theta_i d\theta_j = 0\,.$$

Under this assumption the terms of the decomposition can be univocally calculated using the conditional expectations of the model output $X^\star$, such that $f_0 = \mathbb{E}(X^\star)$, $f_i(\theta_i) = \mathbb{E}(X^\star|\theta_i) - \mathbb{E}(X^\star)$, $f_{i,j}(\theta_i, \theta_j) = \mathbb{E}(X^\star|\theta_i, \theta_j) - f_i(\theta_i) - f_j(\theta_j) - \mathbb{E}(X^\star)$ and so on. Under orthogonality condition, $f(\boldsymbol{\theta})$ can be expressed in the so called ANOVA representation,

$$f(\boldsymbol{\theta}) = f_0 + \sum_{s=1}^{n} \sum_{i_1<\ldots<i_s}^{n} f_{i_1,\ldots,i_s}(\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_n}) \,.$$

Squaring and integrating over I^n we get

$$\int_0^1 f^2(\theta)d\theta - f_0^2 = \sum_{i=1}^{n} \int_0^1 f_i^2(\theta_i)d\theta_i + \sum_{1\le i<j\le n} \int_0^1 f_{i,j}^2(\theta_i, \theta_j)\, d\theta_i\, d\theta_j + \ldots + \int_0^1 f_{1,2,\ldots,n}^2(\theta_1, \theta_2, \ldots, \theta_n)\, d\theta_1, \ldots, d\theta_n \,.$$

Thus, for model output $X^\star$ we get,

$$V(X^\star = \sum_{i=1}^{n} V_i + \sum_{1\le i<j\le n} V_{i,j} + \cdots + V_{1,2,\ldots,n} \,.$$

Here, we denoted with $V(X^\star)$ the total output variance of $X^\star$; moreover, V_i represents the variance attributed only to the parameter θ_i, and $V_{i,j}$ the variance attributed to the interaction of the parameters θ_i and θ_j and similarly for higher order terms ($V_{1,2,\ldots,n}$ is the variance attributed to all the n parameters θ_i for $i = 1, \ldots, n$). Given the decomposed variances, we can define the *first-order Sobol index* S_i,

$$S_i = \frac{V_i}{V(X^\star)} \,,$$

which estimates the variance explained by a single parameter excluding interaction effects; it measures, indeed, the contribution of a single parameter to the output variance. Moreover, a *total-order Sobol index* ST_i can be defined as

$$ST_i = \frac{1}{V(X^\star} \left(V_i + \sum_{j=1}^{n} V_{i,j} + ... + V_{1,2,\ldots,n} \right) .$$

The total-order index takes into account the main, the second-order and higher order effects, estimating all the interactions between the parameters. For example, for a three-parameters model, we have $ST_1 = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3}$. The

total-Sobol index, thus, provides a broader picture of the model's behaviour. Similar definitions, for first and total Sobol index, can be given for a time-dependent model output $X^\star(t)$. In Chapter 6, we used the SALib package of Python to perform a sensitivity analysis and compute the Sobol indices of first and total order of a model for T cell exhaustion.

2.4.3 Structural identifiability

A parameter identifiability analysis can tell us how accurately parameter values of a given model can be determined given the experimental data. For example, in some situations, changing either of two parameters of the model can results in the same model output. In this case, the two parameters can not be independently identified. Consequently, structural identifiability is an important condition for model fitting. Here, we focus on the technique developed by Castro & de Boer (2020) to test the structural identifiability of a mathematical model.

Consider a mathematical model of ordinary differential equations, with m parameters and n variables,

$$\frac{dx_i}{dt} = f_i(x_1, ..., x_n; \theta_1, ..., \theta_m), \quad i = 1, \dots, n$$

with initial conditions $x_i(0) = x_{i,0}$, for $i = 1, \dots, n$. Assume to have $x_1, ..., x_r$ observable variables, with $1 \leq r \leq n$, and $x_{r+1}, ..., x_n$ latent variables, that is variables that cannot be measured experimentally. Each f_i of the system can be decomposed in a sum of $\bar{k}$ functionally independent functions f_{ik},

$$\frac{dx_i}{dt} = \sum_{k=1}^{\bar{k}} f_{ik}(\tilde{x}_k, \tilde{\theta}_k)\,, \tag{2.5}$$

where $\tilde{x}_k, \tilde{\theta}_k$ are the subset of variables and parameters of f_{ik} and functions f_{ik}, f_{il} are functionally independent for every $k \neq l$. For example, the terms $a\,x_1x_2$ and $b\,x_1x_2$ are not functionally independent (for variables x_1, x_2 and parameters a, b), whereas the terms $a\,x_1$ and $b\,x_1x_3$ are independent. We scale parameters and unobserved variables by the unknown scaling transformations

$$\theta_j \to u_{\theta_j}\theta_j \quad \text{and} \quad x_i \to u_{x_i}x_i\,,$$

for $j = 1, \ldots, m$ and $i = r+1, \ldots, n$, respectively. By substituting the scaled variables into Equations (2.5) and equating each functionally independent function, f_{ik}, to its scaled version, we get the identifiability equations of the model

$$f_{ik}(\tilde{x}_k, \tilde{\theta}_k) = f_{ik}(u_{\tilde{x}_k}\tilde{x}_k, u_{\tilde{\theta}_k}\tilde{\theta}_k), \quad \text{for} \quad i = 1, \ldots, r\,,$$
$$f_{ik}(\tilde{x}_k, \tilde{\theta}_k) = \frac{1}{u_{x_i}} f_{ik}(u_{\tilde{x}_k}\tilde{x}_k, u_{\tilde{\theta}_k}\tilde{\theta}_k), \quad \text{for} \quad i = r+1, \ldots, n\,.$$

We solve the system for $u_{\tilde{\theta}_k}$ and find parameters $\tilde{\theta}_k$ that are structurally identifiable: if $u_{\tilde{\theta}_k} = 1$, then parameter $\tilde{\theta}_k$ is identifiable.

We apply this method to study the identifiability of the deterministic mathematical model developed in Chapter 6 for the CD8$^+$ T cell exhaustion process. Full details of the method, including a proof can be found in Castro & de Boer (2020).

2.5 Differential equations

During the analysis of either stochastic and deterministic models we might need to address the study of differential equations. In particular, deterministic mathematical models for cell biology are naturally systems of ordinary differential equations representing dynamics of a population, e.g. population of cells, or receptors' copy numbers, or cytokines. Also, when working with stochastic processes and generating functions we might need to solve differential equations in terms of probability generating functions. We outline, in this Section, the main concepts we used for the analysis carried out in Chapters 4 and 6.

2.5.1 Steady states and stability

In Chapter 4 and Chapter 6 we define deterministic systems of differential equations to describe dynamics of cells populations. In particular, the model defined in Chapter 6 for the CD8$^+$ T cell exhaustion process, developed in collaboration with the Salk Institute of Biological Studies, consists of first-order differential equations,

$$\frac{d\mathbf{X}}{dt} = \boldsymbol{F}(\mathbf{X}(t), t)\,.$$

Here, the variable $\mathbf{X}(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$ represents n populations of cells over time and $\boldsymbol{F}(\mathbf{X}(t), t) = (f_1(\mathbf{X}(t), t), f_2(\mathbf{X}(t), t), \ldots, f_n(\mathbf{X}(t), t))^T$. In many cases, the solution of linear ODEs systems can be found analytically.

An important aspect we can analyse (also in the cases where an analytic solution is not straightforward to find) is the steady state solutions. In general, in a mathematical model there might be none, one or several steady states. When working with a mathematical model of a biological system, it is important to find the steady state solutions because they represent the homeostasis of the populations we are considering. As a simple example of homeostasis one can mention the temperature of our body which is maintained in a narrow range despite the large variation of the environment temperature. In our case of population of cells, when the solution reaches a steady state, the amount of cells is no longer changing over time. Mathematically, the steady state solution $\mathbf{X}^\star$ of an ODE model is computed as solution of algebraic equations and it satisfies

$$\boldsymbol{F}(\mathbf{X}^\star, t) = 0\,.$$

Steady states of a system of ODEs can be classified as either stable, asymptotically stable or unstable.

Definition 2.23. A steady state solution $\mathbf{X}^\star$ of a system of ODEs is *locally stable* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every solution $\mathbf{X}(t)$ with initial condition $\mathbf{X}(t_0)$,

$$\|\mathbf{X}(t_0) - \mathbf{X}^\star\| < \delta \quad \Rightarrow \quad \|\mathbf{X}(t) - \mathbf{X}^\star\| < \varepsilon$$

for all $t \geq t_0$, where $\|\cdot\|$ denotes the Euclidean distance in $\mathbb{R}^n$.

Definition 2.24. A steady state solution $\mathbf{X}^\star$ of a system of ODEs is *locally asymptotically stable* if it is locally stable and there exists $\delta > 0$ such that

$$\|\mathbf{X}(t_0) - \mathbf{X}^\star\| < \delta \Rightarrow \lim_{t \to +\infty} \|\mathbf{X}(t) - \mathbf{X}^\star\| = 0\,.$$

In order to study the asymptotic stability of the steady states we can apply the linearisation theory and determine the sign of the real part of the eigenvalues of the Jacobian matrix.

Theorem 7 (Stability Criterion). *A steady state* $\mathbf{X}^\star$ *of a system of ODEs is asymptotically stable if all the eigenvalues of* $J^\star$*, the Jacobian evaluated at* $\mathbf{X}^\star$*, have negative real parts. The steady state is unstable if at least one of the eigenvalues has a positive real part.*

2.5.2 Riccati equation

In Chapter 4, we define a time-dependent stochastic compartmental model for cells differentiation and consider the corresponding time-dependent probability generating functions and their differential equations which result to be of Riccati's type. In this Section, we recall the main concepts about Riccati differential equations.

Let $x(t)$ be a scalar variable and let us denote with $x^{(i)}(t)$ its ith derivative with respect to t, for $i = 1, \dots, n$. An ordinary differential equation of the nth order,

$$F(t, x(t), x'(t), \dots, x^{(n)}(t)) = 0\,,$$

is said to be *linear* if F is a linear function of the variables $x(t), x'(t), \dots, x^{(n)}(t)$. The generalised Riccati equation is a first-order non-linear ordinary differential equation and it is widely encountered in analytical mechanics and engineering (filtering theory) (Freiling, 2002). In particular, it is defined as

$$x'(t) + P(t)x(t) + Q(t)x^2(t) - R(t) = 0 \tag{2.6}$$

where $P(t), Q(t)$ and $R(t)$ are arbitrary functions of t. Depending on the functions $P(t), Q(t)$ and $R(t)$, several methods have been developed to solve the Riccati equations (Murphy, 1960; Pala & Ertas, 2017). In particular, if one of the following conditions (C1) - (C4) holds, Equation (2.6) can be easily solved by quadrature:

(C1) The coefficients $P(t), Q(t), R(t)$ are constant;

(C2) $R(t) = 0$, thus the equation is linear;

(C3) $P(t) = 0$, thus Equation (2.6) is a Bernoulli equation;

(C4) $Q(t) = 0$, $R(t)$ is constant and $P(t)$ is a power function, e.g. $P(t) = p\,t^n$.

If, as in our case in Chapter 4, none of the conditions (C1) - (C4) holds we fall in a more general case. If so, a change of variable is required so that the Riccati Equation (2.6) is transformed into a second-order linear ODE (Sugai, 1960). In particular, let $x(t) = -\dfrac{\bar{\mathrm{x}}'(t)}{\bar{\mathrm{x}}(t)R(t)}$ and solve

$$\bar{\mathrm{x}}''(t) - \left[Q(t) + \frac{R'(t)}{R(t)}\right]\bar{\mathrm{x}}'(t) + P(t)R(t)\,\bar{\mathrm{x}}(t) = 0\,.$$

Alternative transformation methods are reported by different authors in Allen & Stein (1964); Rao (1962) and Pala & Ertas (2017).

In Chapter 4 and the respective Appendix C, we follow Sugai (1960) and show how the Riccati equation arising from the probability generating functions of a time-dependent compartmental model can be transformed into a second-order linear, and thus more tractable, differential equation. In particular, the resulting equation is a second-order linear equation where the coefficients are functions of the independent variable. Here, we thus consider the homogeneous equation

$$\bar{P}(t)\,\bar{\mathrm{x}}''(t) + \bar{Q}(t)\,\bar{\mathrm{x}}'(t) + \bar{R}(t)\,\bar{\mathrm{x}}(t) = 0\,. \tag{2.7}$$

A point t_0 such that $\bar{P}(t_0) \neq 0$ is called an *ordinary point*, whereas, if $\bar{P}(t_0) = 0$, then t_0 is called a *singular point* of Equation (2.7). In this case, at least one of $\bar{Q}(t_0)$ and $\bar{R}(t_0)$ should not be zero for the structure of Equation (2.7) to hold.

Definition 2.25. A point t_0 is a *regular singular point* of Equation (2.7) if it is a singular point and if both

$$(t - t_0)\frac{\bar{Q}(t)}{\bar{P}(t)} \quad \text{and} \quad (t - t_0)^2\frac{\bar{R}(t)}{\bar{P}(t)} \tag{2.8}$$

have convergent Taylor series about t_0; that is both expressions (2.8) are analytic in $t = t_0$. If the coefficients are polynomials, the condition simplifies into Expressions (2.8) having a finite limit as $t \to t_0$.

Since Expressions (2.8) are analytic in the regular singular point t_0 and have a finite limit for $t \to t_0$, they have convergent power series expansions about an interval $\|t - t_0\| < \rho$, with $\rho > 0$. Thus, Equation (2.7) can be rewritten in terms of power series coefficients for $\bar{\mathrm{x}}'(t)$ and $\bar{\mathrm{x}}(t)$ (Boyce & di Prima, 2009). The

Frobenius method is thus based on the fact that it is natural to seek for solutions in the form of power series

$$\bar{\mathrm{x}}(t) = \sum_{n=0}^{+\infty} a_n \, \bar{\mathrm{x}}^{r+n}(t), \quad r \in \mathbb{R}, \quad a_0 \neq 0 \,.$$

Following (Boyce & di Prima, 2009, Chapter 5), we apply the method constructed by Frobenius for the analysis of the time-dependent probability generating functions of a compartmental mathematical model and report those results in Chapter 4 and Appendix C.

Chapter 3

Why are cell populations maintained via multiple intermediate compartments?

Cells are able to sense and respond to various internal and external stimuli and they grow, reproduce and specialise influenced by the surrounding environment. Also, populations of cells are maintained over host lifetimes and cell populations in organs and tissues are continuously replenished. Mechanisms of cell maintenance and cells differentiation are key to a healthy and functional body. A common observed pattern is that a small population of *progenitor cells* maintains, via a pathway of intermediate states, a larger population of *product* cells that perform a function in the host (MacLean *et al.*, 2018). As mentioned in Chapter 1, progenitor cells may be thought of as stem cells, defined by their ability to both self-renew and generate mature cells (Ogawa, 1993; Reya *et al.*, 2001). In different contexts (Johnston *et al.*, 2007; Robert *et al.*, 2021; Thomas-Vaslin *et al.*, 2008), the product cells may be termed "mature", "active", "fully differentiated" or "effector" cells. An example, detailed in Chapter 1, is the maturation and selection of T cells in the thymus: it takes place via a sequence of states from bone-marrow precursors to mature thymocytes leading, in the case of an adult mouse, to about one million T cells per day exiting the thymus (den Braber *et al.*, 2012). The numbers and phenotypes of descendants of individual näive T cells are highly variable, but the magnitude of total response is repro-

ducible when the output of many families is combined. The heterogeneity of a differentiation process can be quantified by the number of intermediate states in the process; those corresponds to different cell phenotypes, identified by cells surface markers. We refer to Chapter 1 for an overview of the known markers of $CD8^+$ T cell; for example, in a study of the long-term $CD8^+$ T cell response to persistent *T.gondii* infection, the surface markers CXCR3 and KLRG1 were used to identify an intermediate T cell subset between memory and effector cells (Chu *et al.*, 2016). Moreover, recent studies highlight the heterogeneity of the $CD8^+$ T cell exhaustion-differentiation process as characterised by at least a progenitor, an intermediate and a terminal exhausted state (Beltra *et al.*, 2020; Utzschneider *et al.*, 2020).

In many circumstances, the ability of product cells to perform their function is negatively affected by the number of rounds of cell divisions that separates them from their progenitor; every round of division brings, indeed, with it a risk of mutation. During each replication cycle, a cell replicates its DNA with a non-null probability of mistake and thus errors in gene expression can arise. We refer to *clonality* as the population of cells arising from one single cell. Variability of family sizes is unavoidable because the fates of individual cells are subject to chance (Duffy & Hodgkin, 2012; Gerlach *et al.*, 2013; Höfer *et al.*, 2016; Perié *et al.*, 2014). The variation in family size, from one progenitor to another, causes the population of product cells to be dominated by large families arising from few progenitor cells. It has been observed (Lyne *et al.*, 2021; Wainscoat & Fey, 1990) that excessive clonality may increase the risk of cancerous mutations becoming established. For example, as observed in Section 5.2.2, clonal hematopoiesis, observed in older mice and humans (Bowman *et al.*, 2018), is strictly linked to the accumulation of mutations in the haematopoietic population and an increasing risk of leukemia. Low rates of division of cells in early compartments of a lineage is conjectured to reduce the risk that potentially cancerous mutations accumulate. A quantification of cells development can be done considering the size of the population of cells as well as the number of cell cycle events that cells need to undergo. In Section 3.5 we illustrate a second example inspired by immunology, and consider the T cell development in the thymus (Pham *et al.*, 2015).

3. WHY ARE CELL POPULATIONS MAINTAINED VIA MULTIPLE INTERMEDIATE COMPARTMENTS?

Even if cells have nearly identical genome, they carry out their specific functions and exhibit very different phenotypes. New experimental techniques make it increasingly possible to track individual cells as well as their progeny (Marchingo *et al.*, 2016; Perié *et al.*, 2014) and, thus, open the opportunity to quantify cellular proliferation and differentiation. Imaging methods, such as flow cytometry and fluorescence dye labelling, are experimental methods that allow to measure these aspects. Flow cytometry technology allows scientists to measure the fluorescent intensity of single examined cells and quantify the total number of cells within an assay or tissue. Absolute cell counts, can indeed be obtained by combining information about cells' concentration. Moreover, fluorescence molecules, such as green fluorescent protein (GFP), carboxyfluorescein succinimidyl ester (CFSE), or CellTrace Violet labelling dye, can be used to label small molecules or proteins and thus monitor cells (Lemieszek *et al.*, 2022; Lyons, 2000). Cell tracking dyes are used to monitor the proliferative status of cells and to trace multiple cells' generations. In particular, CFSE labelling is widely used to monitor the proliferative responses in complex populations in vitro (Hawkins *et al.*, 2007). After a division event, the daughter cells of a CFSE labelled cell show a reduced CFSE intensity. At each round of cell division, CFSE intensity is halved and thus daughter cells can be distinguished, even if with a limitation in the depth of the cell's family tree.

In this Chapter, we examine how a small flux of cells may be amplified through a sequence of compartments, based on stochastic rules governing the fates of individual cells. We calculate the distributions of the number of product cells per progenitor cell, and of the number of rounds of division that separates them. Our particular focus is on how these distributions depend on the number of compartments that might represent spacial location or cells' phenotype. Rather than focusing on probability distributions of numbers of cells, previous theoretical work on compartmental models has used linear systems of ordinary differential equations for mean quantities (Colijn & Mackey, 2005; Whichard *et al.*, 2010; Zierler, 1981). Deterministic compartmental models are widely used in pharmaco-dynamics, as well as in mathematical ecology (Matis & Wehrly, 1994), immunology (Eftimie *et al.*, 2016) and epidemiology (Capasso, 2008). In Chapter 6 of this PhD thesis, a deterministic model for the exhaustion-differentiation

process and cells reinvigoration is defined. In this Chapter, instead, we focus on the theory of branching processes to model cellular dynamics across compartments, where cells in the same compartment are assumed to be homogeneous and behave identically. Within the field of stochastic analysis, the theory of branching processes has been widely used in cellular dynamics (Kimmel & Axelrod, 2002). For instance, the classic Galton-Watson model, which was originally developed to study the extinction of family surnames (Watson & Galton, 1875), has been successfully applied to quantify the progeny of a cell (Kimmel & Axelrod, 2002). The theory of branching processes can answer questions related to the limiting behaviour (e.g. probability of extinction versus unlimited growth) of cell populations. A natural generalisation is the multi-type branching process, where individuals are not all of the same type. These models can effectively represent cells changing their spatial location (de la Higuera *et al.*, 2019), or their phenotype (Nordon *et al.*, 2011) over time. Among studies in literature, we highlight the seminal work by Matis (1970), proposing a stochastic compartmental model of cell dynamics, where cells in the same compartment are assumed to be homogeneous and behave identically.

In this Chapter, a compartmental stochastic model is defined; specifically, we treat the journey from a single progenitor cell (depicted in green in Figure 3.2) to a population of product cells as a realisation of a branching process with multiple types (Antal & Krapivsky, 2011; Luria & Delbrück, 1943). We study the mean and probability distribution of several random variables of interest, related to cell number and replication. In particular, in Section 3.1 a mathematical model for cell differentiation among a sequence of compartments is defined. Given the sequence of compartments C_i and the events cells can undergo independently in each compartment, we proceed by defining, in Sections 3.2 and 3.3, two different random variables of interest. Their first and second moments as well as their probability distribution are studied. With the aim of linking this theoretical framework to natural biology and cell development processes, we relate the considered random variables to the size of cell population as well as to the number of cell cycle events that the cells undergo. In particular, Section 3.2 introduces the random variable $\mathbf{R}$ for the *family size*. This represents the number of product cells that are descended from one progenitor cell (depicted in green

in Figure 3.2), for one realisation of the stochastic process. Consequently, we denote by N the *amplification factor*, that is the mean number of product cells per progenitor, $N = \mathbb{E}(\mathbf{R})$. A realisation of the stochastic process via "direct" differentiation, $M = 1$, is depicted in Figure 3.4. If in Section 3.2 individual cells are characterised by the compartment each belongs to, in Section 3.3 we introduce the concept of generation. The progenitor initial cell is said to be in generation 0. Whenever a cell in generation n divides, the result is two cells in generation $n+1$. Thus, the population of product cells in the last compartment is heterogeneous: made up of cells of different generations $n \in \mathbb{N}_0$. In particular, we define the random variable $\mathbf{G}$ for the generation number of a randomly-selected product cell and we calculate D, the mean age of product cells, measured in number of generations from the progenitor. A realisation of the stochastic process via "direct" differentiation, $M = 1$, is depicted in Figure 3.14. In Section 3.4 the relationship between the two expected values, N and D, for the number of cells and for the number of divisions, respectively, is investigated. Here, we focus on how the relationship between N and D varies with the number of compartments in the sequence. For Sections 3.2, 3.3 and 3.4, the case of "direct" differentiation ($M = 1$) is considered at first. Secondly, in Sections 3.2.6, 3.3.2 and 3.4.3, the arguments and results are extended to the multi-compartments case of $M > 1$.

Finally, we accommodate asymmetric division event in our mathematical framework. Asymmetric division (see Figure 3.1) is subject of recent research: when this type of division happens one daughter cell remains in the mother's compartment while the other moves to the next compartment (Barile *et al.*, 2020; Böttcher *et al.*, 2018; Gerlach *et al.*, 2013; Pham *et al.*, 2015; Shahriyari & Komarova, 2013; Stiehl & Marciniak-Czochra, 2017; Werner *et al.*, 2015; Yang *et al.*, 2015). From the point of view of Markov processes, an asymmetric division event may be seen as unnatural, in that division and change of type are supposed to be simultaneous. From the biological point of view, on the other hand, the idea is more natural: the mother's intracellular and cell-surface proteins will not be exactly evenly partitioned between the two daughters, who may experience different conditions during the process of cell division (Borsa *et al.*, 2019; Chang *et al.*, 2007). The asymmetric division event brings complexity in the model,

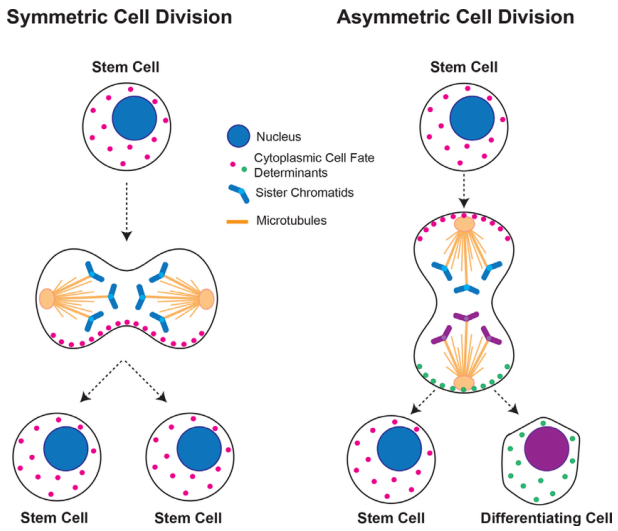


Figure 3.1: On the left self-renewal cell division: the two daughters cells receive identical genetic information as well as the same cell fate determinants. They thus have identical cell fate, such as the mother. On the right asymmetric division: a cell divides to give rise to two daughter cells with distinct cell fates. Figure reprinted with permission (RightsLink service) from Zion & Chen (2021).

thus, for Sections 3.2, 3.3 and 3.4, we start illustrating the analysis and the results for the simpler case where cellular events are restricted to death, self-renewal and differentiation, setting the asymmetric division rate to zero. Later, in Sections 3.2.7, 3.3.3 and 3.4.4, we extend the performed calculations for random variable $\mathbf{R}$ and $\mathbf{G}$ as well as the relationship between N and D, to include asymmetric division. Moreover, in Section 3.5, we show how the developed theoretical methods can be applied to thymus development, a biological system in which asymmetric cell division may play a role (Pham *et al.*, 2015).

3.1 Mathematical model

We propose a stochastic model of cell proliferation, death and differentiation across a sequence of compartments. We consider cells belonging to the same compartment as sharing the same spatial location, or biological role, i.e. function, or simply the same set of cell-surface attributes (the same phenotype). We aim

at investigating how a population of cells performing a specific role is maintained, via a sequence of compartments, from a progenitor source.

We consider a sequence of M compartments, depicted in Figure 3.2, so that, for $i = 1, \ldots, M$, cells that exit compartment C_i enter compartment C_{i+1}. Individual cells in compartment C_i may die or divide, via self-renewal or asymmetric division, or make the transition to the next compartment, meaning that they change phenotype or differentiate. We assume cells exiting the last compartment C_M cannot die, further divide, or differentiate more. Product cells, depicted in orange in Figure 3.2, are thus an accumulative pool and may represent a specific cell type population which can be experimentally measured.

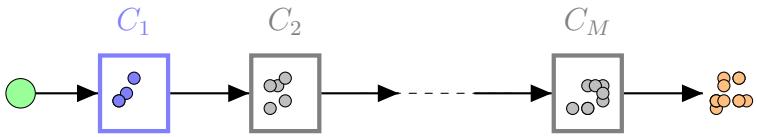


Figure 3.2: An initial cell, in green, enters the sequence of M compartments from progenitor cells (blue compartment C_1) to maturation state (in orange) via a sequence of intermediate compartments (in grey).

In this Chapter, we restrict ourselves to counting cells' number and generation, ignoring both inter-event times and the total time taken for progeny to disappear from all intermediate compartments. The product population of cells, depicted in orange in Figure 3.2, has exited the sequence of M compartments; we focus here in analysing the product population when all cells left previous intermediate compartments, either dying or moving forward in the differentiation chain. The reader should refer to Chapters 4 and 5 for a time-dependent analysis of a similar stochastic compartmental model for cell differentiation.

Starting from a single progenitor cells, we are interested in the ultimate fate of the system; thus, we proceed as in the theory of discrete-time branching processes, by defining relationships between random variables using probability generating functions. We refer the reader to Section 2.1.4 for the definition and the main properties of generating functions. Following the fundamental assumption of the theory of branching processes (Harris, 1963; Kimmel & Axelrod, 2002), we assume each cell in a given compartment follows the same rules and behaves

independently. We refer the reader to Section 2.2.3 for a mathematical overview of branching processes. As depicted in Figure 3.3, a cell in compartment C_i can undergo four, biologically meaningful, cellular events. In particular, each cell, independently, may die, exit the compartment or divide. When a division occurs, daughter cells might both belong to the same compartment as the mother (typically referred to as *self-renewal*), or one daughter cell might belong to the same compartment as the mother while the other belongs to the next compartment (*asymmetric division*) (Barile *et al.*, 2020; Zhang *et al.*, 2015). By definition, each event of a continuous-time Markov process is characterised by an exponential waiting time with a probability. We associate, to each of these events in compartment C_i, a probability, namely $p_d(i)$ for death, $p_e(i)$ for differentiation, $p_b(i)$ for self-renewal, $p_a(i)$ for asymmetric division, so that $p_{\mathrm{d}}(i)+p_{\mathrm{b}}(i)+p_{\mathrm{a}}(i)+p_{\mathrm{e}}(i) = 1$. We assume that the balance of probabilities between fates depends on the com-

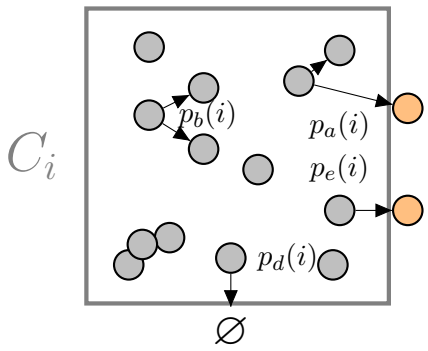


Figure 3.3: Each cell in compartment C_i is characterised by different events: death, self-renewal, direct differentiation and asymmetric division. Those happens with probabilities $p_{\mathrm{d}}(i), p_{\mathrm{b}}(i), p_{\mathrm{e}}(i), p_{\mathrm{a}}(i)$ respectively such that $p_{\mathrm{d}}(i)+p_{\mathrm{b}}(i)+p_{\mathrm{e}}(i)+p_{\mathrm{a}}(i) = 1$.

partment but each cell in a compartment chooses its fate independently. Also, in order to ensure extinction as the ultimate fate of the population of cells in the chain of compartments (blue and grey in Figure 3.2) we assume that

$$p_{\mathrm{d}}(i) + p_{\mathrm{e}}(i) > p_{\mathrm{b}}(i), \qquad \text{for all } i = 1, \ldots, M\,. \tag{3.1}$$

This corresponds to the subcritical assumption within the theory of branching processes described in Kimmel & Axelrod (2002). Note that cell differentiation

has the same effect as death on the population in the compartment because exited cells play no further role in the dynamics of that compartment. Moreover, asymmetric division events happening in compartment C_i do not impact the population balance of that compartment where the division takes place. Rearranging the terms of Equation (3.1), we get

$$2p_{\mathrm{b}}(i) + p_{\mathrm{a}}(i) < 1, \qquad \text{for all } i = 1, \dots, M\,.$$

In the case of no asymmetric division, that is $p_{\mathrm{a}}(i) = 0$ for $i = 1, \dots, M$, the condition ensuring that the mean number of cells descending from a progenitor cell is finite, reads

$$p_{\mathrm{b}}(i) < 1/2\,.$$

In this Chapter, we will define the random variables $\mathbf{R}$ and $\mathbf{G}$ in terms of the events' probabilities $p_{\mathrm{b}}(i), p_{\mathrm{d}}(i), p_{\mathrm{e}}(i), p_{\mathrm{a}}(i)$. However, by definition of discrete-time Markov chains, we can relate them to corresponding event rates. Let μ_i be the death rate, a_i and λ_i be the asymmetric and self-renewal division rate respectively and ν_i be the cell differentiation rate; then it holds

$$p_{\mathrm{d}}(i) = \frac{\mu_i}{\mu_i + \nu_i + a_i + \lambda_i}, \qquad p_{\mathrm{b}}(i) = \frac{\lambda_i}{\mu_i + \nu_i + a_i + \lambda_i},$$
$$p_{\mathrm{a}}(i) = \frac{a_i}{\mu_i + \nu_i + a_i + \lambda_i}, \qquad p_{\mathrm{e}}(i) = \frac{\nu_i}{\mu_i + \nu_i + a_i + \lambda_i}\,.$$

In the analysis carried out in this Chapter we only considered population of cells for large values of time t and thus we could neglect differences in the parameter rates of cells in each of the M compartments of the sequence; for this reason we define cellular rates in terms of probabilities. On the contrary, the time-dependent analysis in Chapters 4 and 5 has been carried out in terms of event rates to take into account differences in compartment dynamics.

The following Sections are organised as follow: in Section 3.2, a stochastic model is formulated such that $\mathbf{R}$ is the total number of product cells, through the sequence of compartments, starting from a single progenitor cell. Secondly, in Section 3.3, we define a discrete-time multi-type branching process initiated by a single progenitor cell. We define $\mathbf{Z}_n(i)$ as the number of cells in generation n and in compartment C_i of the process and we study the random variable $\mathbf{G}$ describing

the generation number of a randomly-selected product cell. For both Sections 3.2 and 3.3 the analysis is carried out for the case of $M = 1$ and $p_a(1) = 0$ at first and then results are generalised to a sequence of compartments as well as to include the asymmetric division event, that is $p_a(i) \neq 0$.

3.2 Counting product cells

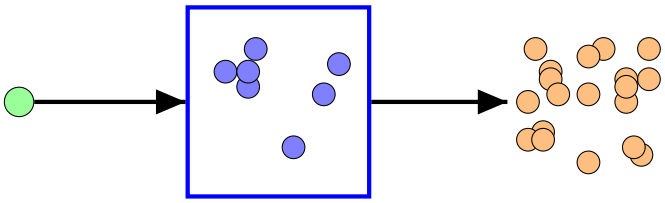


Figure 3.4: One compartment. A single progenitor cell (green) enters the compartment. Each (blue) cell in the intermediate compartment is descended from the progenitor. Each, independently, may die, divide, or exit. Cells (orange) that have exited the compartment do not die or divide further. The random variable **R** is the number of orange cells remaining when there are no remaining blue cells.

Let **R** be the random variable describing the number of cells that exit the sequence of M compartments, descended from one cell entering the first compartment C_1. In this Section, the random variable **R** and its density function are studied. We start restricting ourselves to the case of "direct" differentiation, that is $M = 1$, depicted in Figure 3.4. The population of product cells (in orange) is maintained via a single compartment and it results from the sum over realisations of families, each founded by a single progenitor cell, depicted in green. Moreover, we assume that only three types of single-cell events contribute to the creation of a family of product cells: self-renewal, death and differentiation. This means that, for now, we set $p_a(1) = 0$ for the unique compartment in the sequence. Sections 3.2.1 to 3.2.5 report the main results obtained for the random variable **R** for the simpler case of only self-renewal division and "direct" differentiation ($M = 1$). Sections 3.2.6 and 3.2.7 extend our results to the case of $M > 1$ and of $p_a(i) \neq 0$ for $i = 1, \ldots, M$ respectively.

In Sections 3.2.1 to 3.2.5, when working with only one compartment, we simplify the notation and drop the compartment-index $i = 1$. For example, we denote event probabilities $p_{\rm b}(1), p_{\rm d}(1), p_{\rm e}(1)$ with $p_{\rm b}, p_{\rm d}, p_{\rm e}$ such that $p_{\rm b} + p_{\rm d} + p_{\rm e} = 1$.

3.2.1 R and its probability generating function

Let $\mathbf{R}$ be the random variable describing the number of cells exiting compartment C_1, descended from one cell entering that compartment. We compute the first and second order moments of $\mathbf{R}$ and we analyse q_k defined as the probability distribution of $\mathbf{R}$:

$$q_k = \mathbb{P}(\mathbf{R} = k), \qquad k = 0, 1, 2, \dots .$$

The distribution q_k can be found by a first-step argument: if the first event occurring after the green progenitor enters the blue compartment is cell death, then the population becomes extincted and the process ends; whereas if the first event is cell division (self-renewal) then the two daughter cells, independently, follow the same rules as their mother cell. Therefore, q_0 satisfies the quadratic equation

$$q_0 = \underbrace{p_{\rm d}}_{\text{death}} + \underbrace{p_{\rm b}\, q_0^2}_{\text{self-renewal}} . \tag{3.2}$$

Equation (3.2) can be also derived making use of the total expectation law

$$\sum_{s \in \{\rm d,e,b\}} p_s\, \mathbb{P}(\mathbf{R} = 0 \,|\, \text{first event is } s) = p_{\rm d} 1 + p_{\rm e} 0 + p_{\rm b}\, q_0^2.$$

Note that Equation (3.2) is a quadratic equation with two solutions. Because q_0 is a probability, it can not be greater than one; thus, we take the solution in $[0, 1]$ given by

$$q_0 = \frac{1 - \Delta}{2p_{\rm b}} = \frac{2p_{\rm d}}{1 + \Delta} \qquad \text{where} \qquad \Delta^2 = 1 - 4p_{\rm d}\, p_{\rm b}. \tag{3.3}$$

Next we have,

$$\begin{aligned} q_1 &= p_{\rm e} + p_{\rm b}\, 2q_0\, q_1, \qquad \text{so} \qquad q_1 = \frac{p_{\rm e}}{\Delta}, \\ q_2 &= p_{\rm b}\, 2q_0\, q_1 + p_{\rm b} q_1^2 . \end{aligned} \tag{3.4}$$

We may find further q_k using $q_k = p_\mathrm{b}\left(q_k q_0 + q_{k-1}q_1 + \ldots + q_1 q_{k-1} + q_0 q_k\right)$ so

$$q_k = \frac{p_\mathrm{b}}{\Delta}\sum_{i=1}^{k-1} q_i q_{k-i}\,, \qquad k \geq 2.$$

In this simple case the distribution q_k can be computed by hand; however, in general, it is useful to compute q_k via a probability generating function approach (Allen, 2010; Wilf, 2005). This method gives the terms of the distribution q_k as the coefficients of a power series. It works for more complicated stochastic processes and can be used here to find a general expression of q_k in a quicker way. For example, if $p_\mathrm{a} > 0$, we will show in Section 3.2.7 how the generating function approach is needed to describe the distribution function q_k. We let the probability generating function of $\mathbf{R}$ be

$$\phi(z) = \mathbb{E}(z^{\mathbf{R}}) = q_0 + q_1 z + q_2 z^2 + \cdots\,, \quad z \in [0,1]$$

with z a symbolic argument and such that $\phi(1) = 1$. The probability generating function satisfies a quadratic equation (Steel, 2001; Wilf, 2005):

$$\begin{aligned}\phi(z) &= \sum_{s\in\{\mathrm{d,e,b}\}} p_s \mathbb{E}(z^{\mathbf{R}} \mid \text{first event is } s)\\ &= p_\mathrm{d} z^0 + p_\mathrm{e} z^1 + p_\mathrm{b}\phi^2(z).\end{aligned}$$

Taking the sign of the square root that yields $\phi(1) = 1$,

$$\phi(z) = \frac{1 - (1 - 4p_\mathrm{b}p_\mathrm{d} - 4p_\mathrm{b}p_\mathrm{e}z)^{1/2}}{2p_\mathrm{b}}. \tag{3.5}$$

Using (3.5) and applying the Newton's generalised binomial series, we find

$$q_k = \left(\frac{p_\mathrm{b}}{\Delta}\right)^{k-1}\left(\frac{p_\mathrm{e}}{\Delta}\right)^{k} c_{k-1}\,, \qquad k \geq 1, \tag{3.6}$$

where $c_0 = 1$ and

$$c_k = \frac{(2k)!}{k!(k+1)!}.$$

The c_k are known as the Catalan numbers (Singmaster, 1978) arising in various counting problems in combinatorial mathematics. Examples of q_k distribution are shown in Figure 3.5 for two different choices of p_b and p_e.

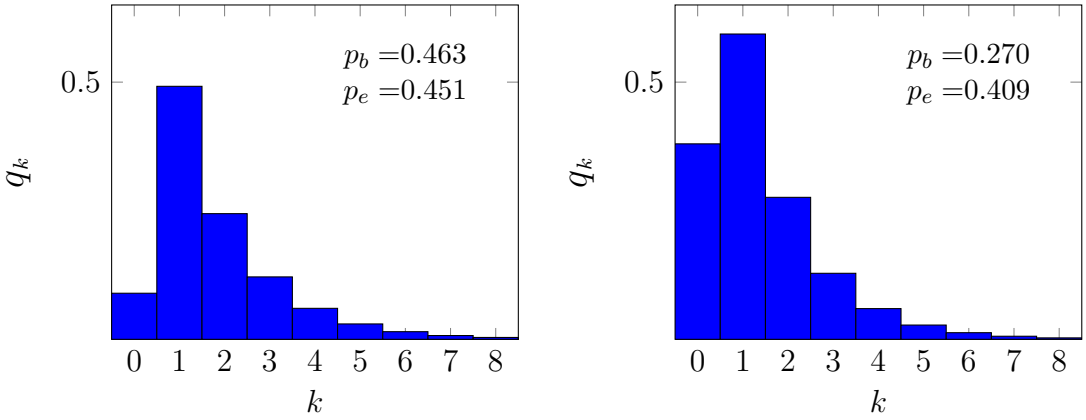


Figure 3.5: The quantity q_k is the probability that k cells exit the compartment C_1, given that one cell enters. Results, using Equation (3.6), are shown for two different choices of p_b and p_e taken from the work of Sawicka *et al.* (2014) about thymocytes maturation. On the left, we used $p_\mathrm{b} = 0.463$ and $p_\mathrm{e} = 0.451$ as computed from single positive CD4 (SP4) thymocytes rates and on the right, we used $p_\mathrm{b} = 0.270$ and $p_\mathrm{e} = 0.409$ as computed from single positive CD8 (SP8) thymocytes rates reported in Sawicka *et al.* (2014).

The generating function $\phi(z)$ of Equation (3.5) contains all the information of the process defined by $\mathbf{R}$. Nevertheless, it is useful to compute the first and the second moments of random variable $\mathbf{R}$. The expected value of $\mathbf{R}$ satisfies

$$\mathbb{E}(\mathbf{R}) = \sum_{s\in\{\mathrm{d,e,b}\}} p_s \,\mathbb{E}(\mathbf{R}\,|\text{ first event is } s) = p_\mathrm{d}0 + p_\mathrm{e}1 + p_\mathrm{b}2N,$$

so rearranging, we get

$$N = \mathbb{E}(\mathbf{R}) = \frac{p_\mathrm{e}}{1-2p_\mathrm{b}}. \tag{3.7}$$

The quantity $N = \mathbb{E}(\mathbf{R})$ is the "amplification factor": the mean number of product cells per progenitor. If there is a constant influx J of progenitor cells, then there is a constant outflux $N\,J$ of product cells. The size of a population of cells is easily measurable (for example via flow cytometry) and it is usually quantified during experiments in cell biology. Thus, the quantity N allows us to create a bridge between our mathematical framework and experimental biology.

In addition to the first moment of $\mathbf{R}$, we can compute the variance V =

$\mathrm{var}(\mathbf{R})$. The second derivative of (3.5) is

$$\phi''(z) = 2p_\mathrm{b}p_\mathrm{e}^2(\Delta^2 - 4p_\mathrm{b}p_\mathrm{e}z)^{-3/2}\,.$$

Thus, by definition,

$$\begin{aligned} V &= \phi''(1) + N - N^2 \\ &= \frac{2p_\mathrm{b}}{1-2p_\mathrm{b}}N^2 + N - N^2\,. \end{aligned} \tag{3.8}$$

3.2.2 Parameterisation in terms of N

Given that $p_\mathrm{d} + p_\mathrm{b} + p_\mathrm{e} = 1$, we can recover any of the probabilities p_d, p_b or p_e given the other two. Consequently, we may parameterise the compartment in terms of any two, linearly independent, combinations of p_d, p_b and p_e.

Given the concreteness of N in biological experiments, we will, on occasions, make use of its expression in Equation (3.7) and write all parameters as a function of p_d and N:

$$p_\mathrm{b} = \frac{N-1+p_\mathrm{d}}{2N-1} \qquad \text{and} \quad p_\mathrm{e} = \frac{N(1-2p_\mathrm{d})}{2N-1}. \tag{3.9}$$

In addition, we can use (3.9) and express the variance $V = \mathrm{var}(\mathbf{R})$ using N itself along with p_d. Using (3.9), expression (3.8) can be rewritten as

$$V = 2\frac{N-1+p_\mathrm{d}}{N(1-2p_\mathrm{d})}N^3 + N - N^2\,.$$

One can observe that the standard deviation of $\mathbf{R}$ is proportional to $N^{3/2}$ as $N \to +\infty$. Figure 3.8 reports (in red, for $M = 1$) the standard deviation $\sqrt{V}$ as a function of the amplification factor N.

Moreover, Figure 3.16 reports the quantity N, that characterises the population of cells exiting compartment C_1, as functions of p_b and p_d. In particular, each blue line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of N. Note that, if $p_\mathrm{d} > \frac{1}{2}$ then $N < \frac{1}{2}$ always. The triangular part of the parameter space corresponding to $N > 1$ is at bottom right. Also, if p_d is fixed, $\lim_{p_\mathrm{b}\to 0} N = 1 - p_\mathrm{d}$.

3.2.3 Recurrence relation

We showed how either a first-step argument or a generating function approach can be used to find an expression of the distribution $q_k = \mathbb{P}(\mathbf{R} = k)$. By definition of probability generating function, we can determine the distribution q_k of $\mathbf{R}$ by repeatedly differentiating the probability generating function of $\mathbf{R}$, as

$$q_k = \frac{1}{k!}\frac{d^{(k)}}{d\, z^{(k)}}\,\phi(z)\bigg|_{z=0}.$$

However, this requires the calculation of the kth derivative of the probability generating function which is not always straightforward. Moreover, numerical differentiation is inaccurate for large values of k and thus, in practice, it is convenient to generate q_k via a recurrence relation. We rewrite the probability generating function in Equation (3.5) as

$$2p_{\mathrm{b}}\phi(z) = 1 - w(z), \qquad \text{where} \quad w^2(z) = 1 - 4p_{\mathrm{b}}p_{\mathrm{d}} - 4p_{\mathrm{b}}p_{\mathrm{e}}z.$$

Differentiating $\phi(z)$ we get

$$w(z)\phi'(z) = p_{\mathrm{e}},$$

and thus, $\phi(z)$ satisfies the differential equation

$$w^2(z)\phi'(z) + 2p_{\mathrm{e}}p_{\mathrm{b}}\phi(z) - p_{\mathrm{e}} = 0.$$

We recall $\phi(z)$ is a generating function and that, by definition, $\phi(z)$ and its derivative can be expressed as power series,

$$\phi(z) = \sum_{k=0}^{+\infty} q_k z^k \quad \text{and} \quad \phi'(z) = \sum_{k=0}^{+\infty}(k+1)q_{k+1}z^k\,.$$

Matching terms proportional to z^k yields the recurrence relation

$$q_{k+1} = \frac{2k-1}{k+1}\frac{2p_{\mathrm{b}}p_{\mathrm{e}}}{1-4p_{\mathrm{b}}p_{\mathrm{d}}}q_k. \tag{3.10}$$

Equation (3.10) is a 2-term recurrence for the distribution q_k.

3.2.4 Asymptotic behaviour

Given a generating function $f(z) = \sum_{k=0}^{+\infty} f_k\, z^k$, one may analyse its analytic properties in order to obtain asymptotic information about the sequence f_k. Following Flajolet & Sedgewick (2009), we can interpret the algebraic object of generating functions as an analytic object, that is a map $f : \mathbb{C} \to \mathbb{C}$ of the complex plane to itself. When $f(z)$ has a simple description such as a singularity of square root type, we can apply the method of singularity analysis and exploit the correspondences between properties of the function $f(z)$ singular at an isolated point and the asymptotic behaviour of its coefficients f_k of the series expansion. The probability generating function $\phi(z)$ in Equation (3.5) is a rational function of square root type; thus, its singular expansion involves fractional powers and the method of singularity analysis developed by Flajolet & Sedgewick (2009) can be applied.

Given a function $f(z) = \sum_{k=0}^{+\infty} f_k\, z^k$, we denote with $[z^k]f(z)$ the coefficient f_k of z^k in the power series. If $f(z)$ has a singularity in $z = \sigma$, then via the scaling rule of Taylor expansion

$$[z^k]f(z) = \sigma^{-k}[z^k]f(\sigma\, z) = \sigma^{-k}[z^k]g(z)\,,$$

where $g(z) = f(\sigma\, z)$ has a singularity in $z = 1$. Rational functions involve, near a singularity σ, fractional powers of the form $\left(1 - \frac{z}{\sigma}\right)^{-\alpha}$ and one can apply the method for singularity analysis developed by Flajolet & Sedgewick (2009). In particular, if $g(z) = (1 - z)^{-\alpha}$, for a general $\alpha \in \mathbb{R}$ or $\alpha \in \mathbb{C}$, Theorem 6.1 in (Flajolet & Sedgewick, 2009, Chapter 6) proves that

$$[z^k](1 - z)^{-\alpha} \sim \frac{k^{\alpha-1}}{\Gamma(\alpha)} \left(1 + \frac{\alpha(\alpha - 1)}{2k} + O\left(\frac{1}{k^2}\right)\right)\,.$$

Thus, for a function with singularity in $z = \sigma$,

$$[z^k]\left(1 - \frac{z}{\sigma}\right)^{-\alpha} \sim \frac{\sigma^{-k}}{\Gamma(\alpha)}\, k^{\alpha-1}\,. \tag{3.11}$$

Recall that if $M = 1$ and $p_{\mathrm{a}} = 0$, the probability generating function of $\mathbf{R}$ is given by Equation (3.5),

$$\phi(z) = \frac{1 - \sqrt{\Delta^2 - 4p_{\mathrm{b}}p_{\mathrm{e}}z}}{2p_{\mathrm{b}}}\,,$$

that is a rational function with a single singularity in $\sigma = \frac{\Delta^2}{4p_b p_e}$. We can then use the results of the asymptotic analysis of $g(z) = (1-z)^{-\alpha}$, with $\alpha = -1/2$, to derive the asymptotic behaviour of the series' coefficients of $\phi(z)$. In particular, we write

$$\phi(z) = \frac{1}{2p_b} - \frac{\sqrt{p_e}}{\sqrt{p_b}}\sqrt{\frac{\Delta^2}{4p_b p_e} - z}\,,$$

and, following Flajolet & Sedgewick (2009), we find

$$[z^k]\phi(z) \sim -\frac{\sqrt{p_e}}{\sqrt{p_b}}\left(\frac{\Delta^2}{4p_b p_e}\right)^{-k}\frac{k^{-3/2}}{\Gamma(-1/2)}\,.$$

Making the Gamma function explicit, we get

$$[z^k]\phi(z) \sim \frac{\sqrt{p_e}}{2\sqrt{\pi p_b}}\left(\frac{4p_b p_e}{\Delta^2}\right)^k k^{-3/2} = \frac{p_\mathrm{e}}{\sqrt{\pi}\Delta}\gamma^{k-1}k^{-3/2}\,,$$

where $\sigma = 1/\gamma$,

$$\gamma = \frac{4p_\mathrm{b}p_\mathrm{e}}{1 - 4p_\mathrm{b}p_\mathrm{d}}. \tag{3.12}$$

We can thus conclude that, for $k \to +\infty$,

$$q_k \sim \frac{p_\mathrm{e}}{\sqrt{\pi}\Delta}\gamma^{k-1}k^{-3/2}\,. \tag{3.13}$$

This is in agreement with the observation made by Flajolet & Sedgewick (2009): square root singularities universally translate into an asymptotic estimate with exponent $-\frac{3}{2}$. Moreover, as underlined by Flajolet & Sedgewick (2009), if $g(z) = (1-z)^{-\alpha}$, one can apply the binomial expansion for a general α, and carry out an asymptotic analysis of the resulting binomial coefficient by means of Stirling's formula.

An example of the long behaviour of distribution q_k is depicted in Figure 3.6 for thymocytes SP4 and SP8 population, in light and dark green respectively (Sawicka *et al.*, 2014).

Equation (3.13) outlines the presence of two règimes in the behaviour of q_k for $N \gg 1$: when k is small enough that $\gamma^k \approx 1$, q_k is governed by the power law; then, for greater k, by the geometric γ^k. Also, from Equation (3.13) we get

$$\lim_{k\to+\infty} \log\left(\frac{q_{k+1}}{q_k}\right) = \log\gamma - \frac{3}{2}\log\left(1+\frac{1}{k}\right).$$

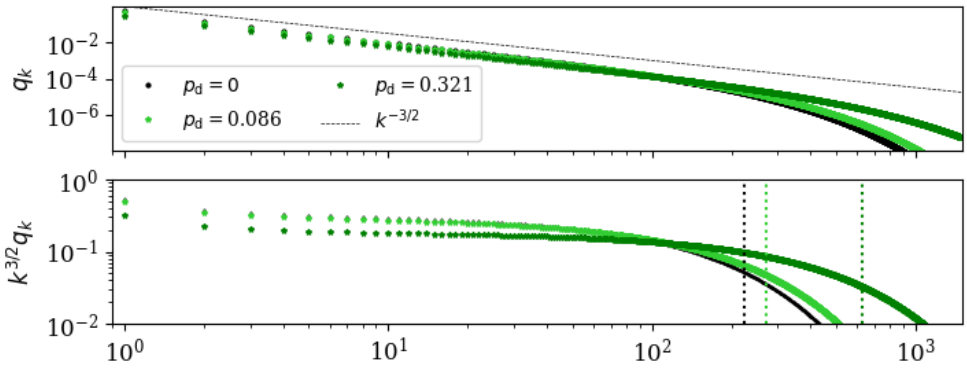


Figure 3.6: The probability that the number of product cells is k, logarithmic scales, with and without death. The dashed line is the power law $q_k = k^{-3/2}$. The lower panel plots $k^{3/2}q_k$ in the same two cases. The vertical dotted lines, at $k = 6N^2/(1-2p_\mathrm{d})$, indicate where the behaviour of q_k begins to deviate from the power law. The parameter values, calculated using (3.9) so that $N = 6.1$, are taken from the work of Sawicka *et al.* (2014) about thymocytes maturation. In particular, we consider SP4 (in light green) and SP8 (in dark green), single positive CD4 and CD8, respectively. Black dots show q_k distribution for $p_\mathrm{d} = 0$.

Making use of Equations (3.9), one can write γ in terms of p_d and N, and find

$$\gamma = 1 - \frac{1-2p_\mathrm{d}}{4N^2 + 1 - 4N - 2p_\mathrm{d} + 4Np_\mathrm{d}}\,.$$

Then, if $N \gg 1$

$$\lim_{\substack{k\to+\infty \\ N\to+\infty}} \log\left(\frac{q_{k+1}}{q_k}\right) = -\frac{1-2p_\mathrm{d}}{4N^2} - \frac{3}{2}\frac{1}{k}\,, \tag{3.14}$$

where we used that, for $|x| < 1$, $\log(1+x) = x + O(x^2)$ and $\log(1-x) = -x + O(x^2)$. Thus, the decrease in q_k as a function of k is primarily due to factor $k^{-3/2}$ when $(1-2p_\mathrm{d})k < 6N^2$; thereafter, for $k > \frac{6N^2}{1-2p_\mathrm{d}}$, it is due to the factor γ^k.

3.2.5 Analysis of cell's clonality

A population of product cells is said to have a high degree of *clonality*, if it is dominated by a few large families (cells originated from the same progenitor). Biologically this is relevant when looking at gene mutations, random alterations

of nucleotides in DNA sequences that can occur during DNA replication (the construction of a new DNA filament). When a mutation happens, it is then propagated into all descendants of the mutant cell. It has been observed that every round of division brings with it a risk of mutation and that a high degree of clonality may increase the risk of cancerous mutations becoming established (Lyne *et al.*, 2021; Wainscoat & Fey, 1990). Mathematically, in a population of cells made up of multiple realisations of $\mathbf{R}$, we can understand the dominance of large families of cells by evaluating k_{50}, the lowest value of k such that half of cells are part of a family of fewer than k cells. Given $N = \mathbb{E}(\mathbf{R})$, we define k_{50} as the value of the index k such that

$$\sum_{k=0}^{k_{50}} k\, q_k = \frac{N}{2}\,. \tag{3.15}$$

Solving Equation (3.15) and finding an exact expression of k_{50} is not straightforward. We can, however, determine a lower bound for k_{50} by making use of the inequality

$$q_k < \frac{p_e}{\sqrt{\pi}\Delta}\gamma_1^{k-1}k^{-3/2}\,,$$

derived from Equation (3.13). We recall, as outlined in Section 3.2.4, that for k small enough, we have $\gamma^k \approx 1$ and thus q_k is governed by the factor $k^{-3/2}$. Thus, it holds

$$\frac{N}{2} = \sum_{k=0}^{k_{50}} k\, q_k < \frac{p_e}{\sqrt{\pi}\Delta}\sum_{k=1}^{k_{50}} k^{-1/2}\,. \tag{3.16}$$

We look for an upper and lower bound for the sum $\sum_{k=1}^{k_{50}} k^{-1/2}$ by making use of Riemann sums (Oberbroeckling, 2021). Since the function $f(k) = k^{-1/2}$ is monotonic decreasing in $(0, +\infty)$ as a function of k, then

$$\sum_{k=1}^{k_{50}} k^{-1/2} \leq f(1) + \int_1^{k_{50}} k^{-1/2}dk\,,$$

and we can find an upper bound for our sum of interest,

$$\sum_{k=1}^{k_{50}} k^{-1/2} \leq 1 + \int_1^{k_{50}} k^{-1/2}dk = 1 + 2\sqrt{k_{50}} - 2 = 2\sqrt{k_{50}} - 1 \leq 2\sqrt{k_{50}}\,.$$

Applying the inequality to (3.16), we have

$$\frac{N}{2} < \frac{2p_e}{\sqrt{\pi}\Delta}\sqrt{k_{50}}\,,$$

that is

$$k_{50} > \frac{N^2\Delta^2\pi}{16p_e^2} = \frac{\pi}{16}\frac{\Delta^2}{(1-2p_{\mathrm{d}})^2}(2N-1)^2\,,$$

where we made use of (3.9) to express p_{e} in terms of N and p_{d}.

If $N \gg 1$, the quantity k_{50} is proportional to N^2. Moreover, Figure 3.7 shows how k_{50} is an increasing function of the death probability p_{d} and of amplification factor N. For example, if $N = 10$ and $p_{\mathrm{d}} = 0$, then $k_{50} = 83$ and the analytical bound is $k_{50} > 71$; if $N = 100$ and $p_{\mathrm{d}} = 0$, then $k_{50} = 9000$ and the analytical bound is $k_{50} > 7775$. Consequently, if $M = 1$ (that is progenitor cells directly differentiate into product cells), a large mean number of product cells N can be reached only via a high value k_{50}; likewise, it is high the risk of mutations.

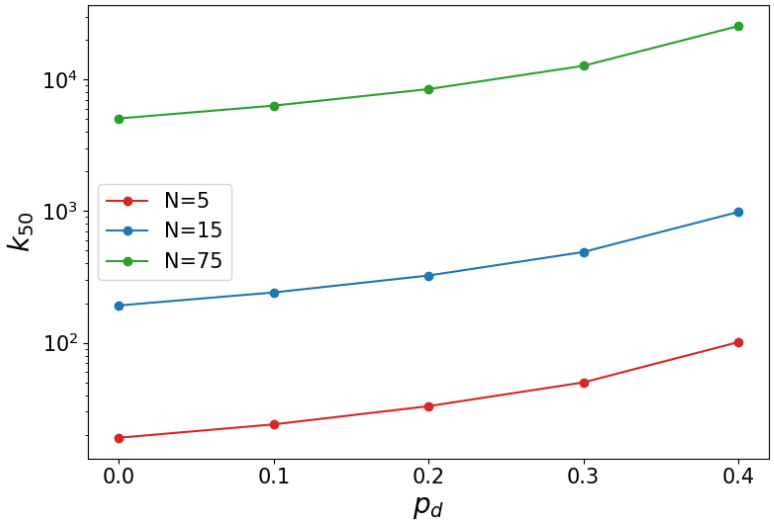


Figure 3.7: We consider three different values of the *amplification factor* N, that is red for $N = 5$, blue for $N = 15$ and green for $N = 75$. Using (3.10), we compute the analytical bound k_{50} for increasing values of $p_{\mathrm{d}} \in [0, 0.1, 0.2, 0.3, 0.4]$. Note the log scale of the y-axis.

3.2.6 Generalisation to a sequence of M compartments

In biology a large population of product cells is often maintained via a sequence of intermediate states, or compartments (MacLean *et al.*, 2018). For example, in the context of immune cells differentiation, the maturation and development of T cells in the thymus pass through a sequence of intermediate stages (Yates, 2014). In this Section, we assume $M > 1$ and we derive properties of product cells, investigating how they depend on the number of intermediate compartments M. Since we consider now the case of at least one intermediate compartment in the sequence, so $M > 1$, we find it helpful to define either "compartment" or "total" variables. In particular, we define the random variable $\mathbf{R}_i$ for the number of cells exiting compartment C_i descended from one cell that enters compartment C_i. We still denote with $\mathbf{R}$ the random variable representing the number of product cells (depicted in orange in Figure 3.2) exiting the final compartment C_M, given that one cell enters compartment C_1. Following the sequence of compartments, the product cells exiting compartment C_M have undergone M differentiation events between the progenitor and the product type phenotype.

R and its probability generating function

Let $Q_k(M)$ be the probability that the number of product cells, descended from a single progenitor via M intermediate compartments, is equal to k; that is

$$Q_k(M) = \mathbb{P}(\mathbf{R} = k), \qquad k = 0, 1, 2, \dots. \tag{3.17}$$

In this case of a sequence of M compartments we cannot apply a first-step argument to directly compute the distribution $Q_k(M)$, but a generating function approach is required. The probability generating function of $\mathbf{R}$, $\Phi_M(z)$, is defined as

$$\Phi_M(z) = \mathbb{E}(z^{\mathbf{R}}) = Q_0(M) + zQ_1(M) + z^2Q_2(M) + \cdots, \tag{3.18}$$

with $\Phi_M(1) = 1$. Let the probability generating function of R_i be

$$\phi_i(z) = \frac{1 - (\Delta_i^2 - 4p_{\mathrm{b}}(i)p_{\mathrm{e}}(i)z)^{1/2}}{2p_{\mathrm{b}}(i)}, \qquad i = 1, \dots, M, \tag{3.19}$$

where $\Delta_i^2 = 1 - 4p_{\mathrm{d}}(i)p_{\mathrm{b}}(i)$. As in Section 3.2.1 for $M = 1$, we can define the *amplification factor* N_i of each compartment

$$N_i = \mathbb{E}(\mathbf{R}_i) = \phi_i'(z)|_{z=1} = \frac{p_{\mathrm{e}}(i)}{1 - 2p_{\mathrm{b}}(i)}\,, \tag{3.20}$$

as the mean number of cells exiting compartment C_i for each cell entering compartment C_i. The probability generating function for the total process can be written as the compositions of the probability generating functions of each intermediate compartment C_i (Harris, 1963; Kimmel & Axelrod, 2002; Wilf, 2005). That is, for a sequence of M compartments, the probability generating function (3.17) of the "total" random variable $\mathbf{R}$ can be written as

$$\Phi_M(z) = \phi_1(\phi_2(\cdots\phi_M(z))) = \phi_1(\chi_M(z)). \tag{3.21}$$

The function $\chi_M(z)$ has been defined as needed for the calculations of the second moment of $\mathbf{R}$ (see Equation (3.23)) as well as for the asymptotic coefficient γ_M (see Equation (3.26)). Of course, when $M = 1$ we fall in the case of "direct" differentiation with no intermediate compartments, detailed in Section 3.2.1; then $Q_k(1) = q_k$ and $\Phi_1(z) = \phi(z)$.

We denote with N the expected value of $\mathbf{R}$; that is the overall amplification factor (from one single progenitor cell), given by

$$N = \prod_{i=1}^{M} N_i.$$

Moreover, in addition to the first moment, one can compute the second moment of $\mathbf{R}$,

$$\mathrm{var}(\mathbf{R}) = \Phi_M''(1) + N - N^2\,. \tag{3.22}$$

Recall $\Phi_M(z) = \phi_1(\chi_M(z))$, thus $\Phi_M'(z) = \phi_1'(\chi_M(z))\chi_M'(z)$ and

$$\Phi_M''(1) = \phi_1''(1)\left(\chi_M'(1)\right)^2 + \phi_1'(1)\chi_M''(1),$$

where $\Phi_M'(z) = \frac{d}{dz}\Phi_M(z)$. If we assume identical compartments, that is $\phi_i(z) = \phi(z)$, $i = 1, \ldots, M$, and if, in each compartment C_i, $N_i = N^{1/M}$ is independent

of i and $p_{\mathrm{d}}(i) = 0$, we can make further comments on the distribution of $\mathbf{R}$. In this case, indeed,

$$\phi'(1) = N^{1/M}, \qquad \phi''(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} N^{3/M},$$

and

$$\Phi''_M(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} \left(N^{3/M} N^{2(1-1/M)}\right) + N^{1/M} \chi''_M(1).$$

We find

$$\Phi''_1(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} N^3, \quad \Phi''_2(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} \left(N^{5/2} + N^2\right), \quad \Phi''_3(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} \left(N^{7/3} + N^2 + N^{5/3}\right)$$

and, for a general M,

$$\Phi''_M(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}} N^{2+1/M} \sum_{j=0}^{M-1} N^{-j/M}. \tag{3.23}$$

Thus, the variance of $\mathbf{R}$ is proportional to $N^{2+\frac{1}{M}}$ as $N \to +\infty$ (see Figure 3.8). In Section 3.4.3, we show how by assuming identical compartments, we can reduce the number of parameters and show that these assumptions are optimal from the perspective of minimising the mean number of divisions per cell.

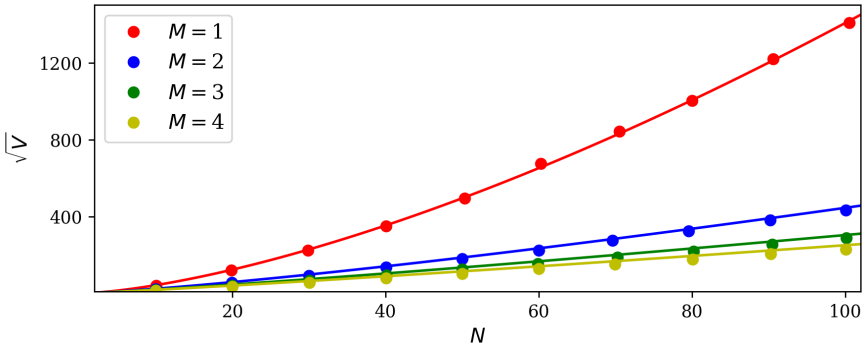


Figure 3.8: The standard deviation of $\mathbf{R}$ as a function of the $N = \mathbb{E}(\mathbf{R})$, with different numbers of compartments. The lines use the formula (3.22); each line corresponds to one value of M. The dots are obtained as averages over 10^4 numerical realisations.

Recurrence relation

As done in Section 3.2.3 for the case $M = 1$, we now look for a recurrence to generate values of $Q_k(M)$. The calculation of an analytical recurrence relation for $Q_k(M)$, for the general case of a sequence of M compartments, is not straightforward. In Appendix A, we report the calculations for the case $M = 2$. This method is inspired by the work of Mathar (2021) and it can be, in principle, generalised for longer sequences of compartments, $M > 2$; we reserve to do that in future work.

For $M = 2$ and $p_{\mathrm{a}}(1) = p_{\mathrm{a}}(2) = 0$, we find the three-term recurrence relation,

$$\begin{aligned} &a_2(k+1)(k+2)Q_{k+2}(2) + a_1(2k+1)(2k+2)Q_{k+1}(2) \\ &\quad + a_0\left(64k^2 - 4\right)Q_k(2) = 0\,, \end{aligned} \tag{3.24}$$

where

$$\begin{aligned} a_2 &= \Delta_2^2\left(\Delta_1^2(4p_{\mathrm{e}}(1)\,r_b - \Delta_1^2) - 16p_{\mathrm{e}}^2(1)p_{\mathrm{b}}(1)p_{\mathrm{d}}(2)\,r_b\right)\,, \\ a_1 &= \left(\Delta_1^4 - 4p_{\mathrm{e}}(1)\,r_b\Delta_1^2 + 4p_{\mathrm{e}}^2(1)\,r_b^2(1 - 2\Delta_2^2)\right)p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)\,, \\ a_0 &= p_{\mathrm{e}}^2(1)p_{\mathrm{e}}^2(2)p_{\mathrm{b}}^2(1)\,, \end{aligned}$$

where $r_b = p_{\mathrm{b}}(1)/p_{\mathrm{b}}(2)$. For more technical details about the derivation of (3.24), we refer the reader to Appendix A. In Figure 3.9, we report $Q_k(1)$ and $Q_k(2)$ for three different scenarios regarding the amplification factors N_1 and N_2. Moreover, Figure B.1 reports the behaviour of $Q_k(M)$, for $M = 1, 2, 3$, for large values of k; distributions $Q_k(1)$ and $Q_k(2)$ (in green and red, respectively) diverge for larger values of k.

In addition, we can study the dominance of large families of cells and estimate k_{50} for the case $M > 1$. For example, if $M = 2$ and $N = 10$ then the minimum k_{50} value is 33; whereas if $M = 2$ and $N = 100$ it is 1010. For $M = 3$, the minimum k_{50} values are 25 and 528 for $N = 10$ and $N = 100$ respectively. This indicates how k_{50} is decreasing with the number of compartments. In particular, Figure 3.10 reports the *clonality* index k_{50} for the case $M = 2$, computed using the recurrence (3.24). Here, we focus on three different scenarios with fixed total amplification factor $N = 75$ and varying ratios of N_1/N_2. We consider two compartments having equal amplification factor, that is $N_1/N_2 = 1$; the first

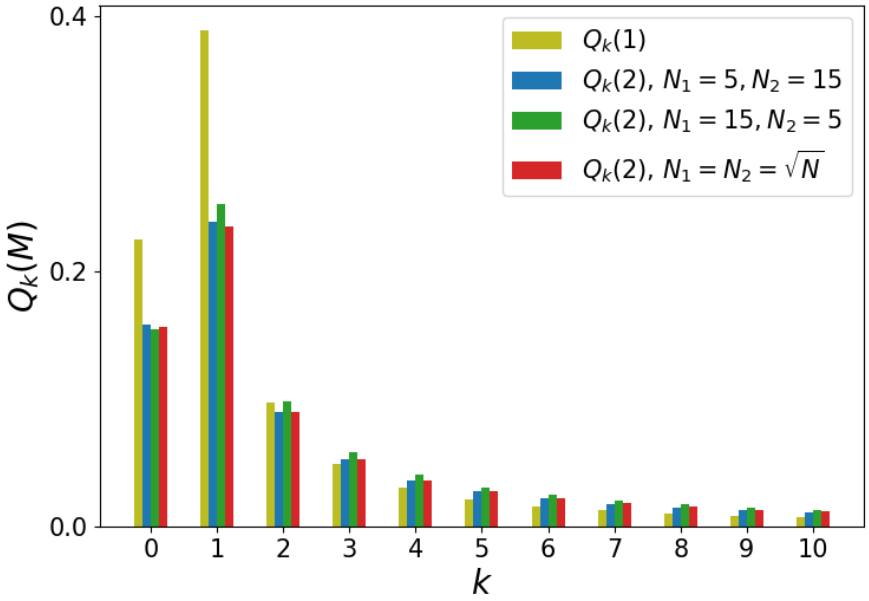


Figure 3.9: We set $N = 75$ and plot $Q_k(M)$ as a function of k for $M = 1, 2$. For $M = 2$ we consider three different scenarios, two compartments with the same amplification factor (in blue) and the second compartment with greater (in red) and smaller (in green) amplification than the first one.

compartment having a lower amplification than the second one, that is $N_1/N_2 = 1/3$; and vice-versa, $N_1/N_2 = 3$. In agreement with the case $M = 1$, one can observe that k_{50} is an increasing function of $p_{\mathrm{d}}(1)$ and $p_{\mathrm{d}}(2)$. In particular, we observe a higher variability in the value of k_{50} if the second compartment has a higher amplification factor than the first, that is $N_1 = 5$ and $N_2 = 15$ (blue lines). Most importantly, we observe that, given a fixed value of $p_{\mathrm{d}}(2)$, a variation of $p_{\mathrm{d}}(1) \in [0, 0.4]$ has a greater impact on the value of k_{50} (dotted lines).

Asymptotic behaviour

Given definition (3.18), the coefficients of the probability generating function $\Phi_M(z)$ are the elements of the distribution $Q_k(M)$. In this Section, we apply the method developed by Flajolet & Sedgewick (2009) to analyse the long tail and the asymptotic behaviour for $k \to +\infty$ of the distribution $Q_k(M)$. We proceed by

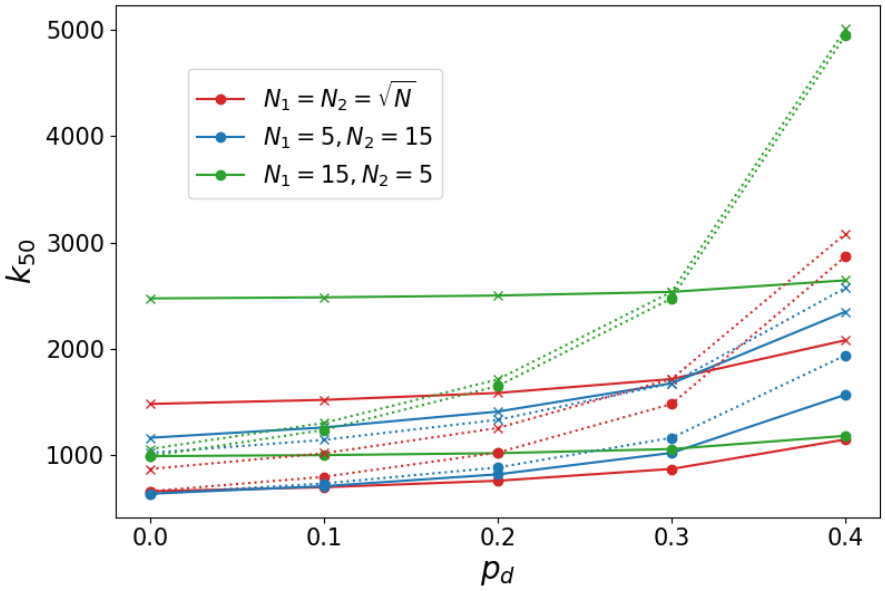


Figure 3.10: We consider the total amplification factor $N = 75$ as fixed, and vary N_1, N_2. In particular, in red we have two compartments with the same amplification factor, in blue we consider the case of the second compartment having a greater amplification than the first one, and in green the vice-versa. Using (3.24), we compute k_{50} for increasing values of $p_{\mathrm{d}}(1)$ or $p_{\mathrm{d}}(2)$. In particular, dotted lines shows the value of k_{50} for increasing $p_{\mathrm{d}}(1)$ as in x-axis and $p_{\mathrm{d}}(2) = 0$ (circle) or $p_{\mathrm{d}}(2) = 0.3$ (cross). Continuous lines, shows the vice-versa: $p_{\mathrm{d}}(2)$ increases as in x-axis and $p_{\mathrm{d}}(1) = 0$ (circle) or $p_{\mathrm{d}}(1) = 0.3$ (cross). Parameters $p_{\mathrm{b}}(i), p_{\mathrm{e}}(i)$, for $i = 1, 2$, are chosen appropriately to get the desired amplification factor using (3.9).

locating the singularities of the function $\Phi_M(z)$ and find the long-term behaviour of the coefficients of its series expansion denoted as $[z^n]\Phi_M(z)$. Details about how the singularity of $\Phi_M(z)$ results from the most outer root of the function, are shown in Appendix B for $M = 2$ and $M = 3$ as a generalisation of the case $M = 1$ detailed in Section 3.2.4.

In case of a sequence of M compartments, the probability generating function of $\mathbf{R}$ is a composition of functions $\Phi_M(z) = \phi_1(\phi_2(\dots\phi_M(z)))$ where each $\phi_i(z)$ is a square root function of the form of Equation (3.19). The asymptotic

analysis of $Q_k(M)$ can be directly based on the general discussion of composition of singularities of $\phi_1(z), \phi_2(z), \ldots, \phi_M(z)$. Since each $\phi_i(z)$ has a square root type singularity and the outer functions are analytic in the singularity of the inner function, the coefficients of the asymptotic estimate still have the form $k^{-3/2}$ (Flajolet & Sedgewick, 2009). That is an interesting feature of the distribution of **R**: for large k we observe a universal behaviour

$$Q_k(M) \sim \gamma_M^k k^{-3/2}, \qquad \text{as} \quad k \to +\infty\,, \tag{3.25}$$

where γ_M is determined by locating the square-root singularity of $\Phi_M(z)$.

Moreover, Figure 3.11 shows that the distribution of **R** narrows as the number of compartments increases: here $k^{3/2}Q_k(M)$ is plotted as a function of k, with $p_{\mathrm{d}}(i) = 0$ for $i = 1, \ldots, M$. In all three cases shown, the mean number of product cells, N, is equal to 25. In terms of the underlying dynamics, we see that individual realisations may yield numbers of product cells many times greater than N, but, within the same amplification factor N, such realisations become rarer as the number of intermediate compartments is increased.

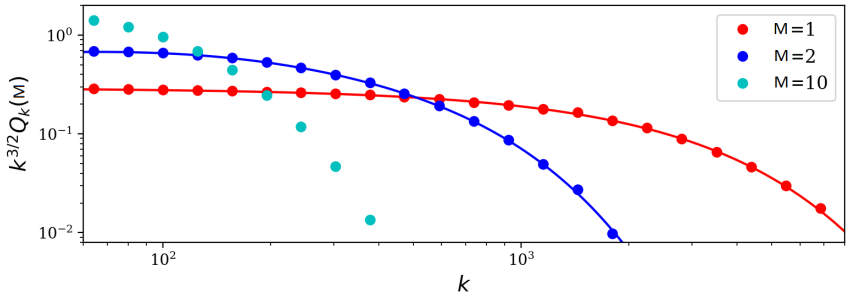


Figure 3.11: Plot of $k^{3/2}Q_k(M)$ as a function of k, with logarithmic scales, for $M = 1$, $M = 2$ and $M = 10$. The solid lines are the exact results, computed using (3.10) and (3.24). The dots are averages obtained from Gillespie realisations. Parameter values, chosen using (3.9) with $N = 25$, are $M = 1$: $p_{\mathrm{d}} = 0$, $p_{\mathrm{b}} = 0.4898$. $M = 2$: $p_{\mathrm{d}}(1) = p_{\mathrm{d}}(2) = 0$, $p_{\mathrm{b}}(1) = p_{\mathrm{b}}(2) = 0.4444$ and $N_1 = N_2 = 5$. $M = 10$: $p_{\mathrm{d}}(i) = 0$, $p_{\mathrm{b}}(i) = 0.2158$ and $N_i = 1.38$ for each $i = 1, \ldots, 10$.

The reader may also note from (3.25) the two different regimes of the distribution $Q_k(M)$: for k small enough the terms of distribution are driven by the

universal term $k^{-3/2}$; for larger k the behaviour is driven by the power γ_M^k. The constant $\gamma_M < 1$, driving the behaviour of $Q_k(M)$ for large k, may be determined by locating the square root singularity of $\Phi_M(z)$ (Flajolet & Sedgewick, 2009; Greene & Knuth, 1990; Knuth & Wilf, 1989). If $M = 1$ we found the constant γ as in Equation (3.12); for consistency of notation, we denote it as $\gamma_1 = 4p_{\mathrm{b}}(1)p_{\mathrm{d}}(1)/\Delta_1^2$. Then, using $\Phi_2(z) = \phi_1(\phi_2(z))$, we have that γ_2 satisfies $4p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)\phi_2(1/\gamma_2) = \Delta_1^2$. In general, making use of (3.21) we can write

$$(1 - 2p_{\mathrm{b}}(1)\Phi_M(z))^2 = \Delta_1^2(1 - \gamma_1\chi_M(z)). \tag{3.26}$$

Since the singularity of $\chi_M(z)$ is close to one, we can approximate $\chi_M(z)$ by its Taylor expansion in $z = 1$. Recall that, by definition, $\chi_M(z)$ is a probability generating function, and thus $\chi_M(1) = 1$ and $\chi'_M(1) = N/N_1$. Thus, we get

$$\begin{aligned} 1 - \gamma_1\chi_M(z) &\sim 1 - \gamma_1\left(1 - (1-z)\frac{N}{N_1}\right) \\ &= \left(\gamma_1\left(\frac{N}{N_1} - 1\right) + 1\right)\left(1 - \frac{\gamma_1\frac{N}{N_1}}{\gamma_1\left(\frac{N}{N_1} - 1\right) + 1}z\right). \end{aligned}$$

This gives the approximation

$$\gamma_M \sim \left(1 + \frac{1 - \gamma_1}{\gamma_1 N/N_1}\right)^{-1}.$$

Now, if $N_1, N \gg 1$ then $1 - \gamma_1 \sim \frac{1}{4N_1^2}$ and

$$\gamma_M \sim 1 - \frac{1 - 2p_{\mathrm{d}}(1)}{4N_1 N},$$

in agreement with what we observed in Equation (3.14) for $M = 1$. Moreover, under the hypothesis of identical compartments, $p_{\mathrm{d}}(i) = 0$ and $N_i = N^{1/M}$ independent,

$$\gamma_M = 1 - \frac{1}{4}\frac{1}{N^{1+1/M}} + \frac{1}{16}\frac{1}{N^{2(1+1/M)}} + O\left(\frac{1}{N^{3(1+1/M)}}\right).$$

3.2.7 Generalisation to asymmetric division

In this Section, we accommodate asymmetric division in our framework by introducing it as a fourth type of event, that is $p_{\mathrm{a}}(i) > 0$, for $i = 1, \dots, M$. From the mathematical point of view, asymmetric division introduces complexity in the model and the calculation. For ease of notation, in this Section and in the following where asymmetric division is considered (Sections 3.3.3 and 3.4.4), we denote the functions and quantities with the corresponding symbol used in Sections 3.2.1- 3.2.5, but with a tilde on top, e.g. $\tilde{\phi}$ instead of ϕ. Moreover, as in previous Section, we omit the index $i = 1$ for the case of $M = 1$.

If $M = 1$, we compute the probability generating function of the random variable $\mathbf{R}$ and compute its distribution $\tilde{q}_k$. Moreover, we find an expression for the recurrence relation of $\tilde{q}_k$ and its asymptotic behaviour. The universal behaviour observed in (3.25) for large k still holds. Similar analysis performed in Section 3.2.6 is carried out here for $M = 2$ tackling the complexity introduced by the asymmetric division event.

R and its probability generating function Following the same argument as before, we start looking at the distribution $\tilde{Q}_k(M)$ of the random variable $\mathbf{R}$. If $M = 1$, we denote $\tilde{Q}_k(1) = \tilde{q}_k$ and observe that $\tilde{q}_0 = \mathbb{P}(\mathbf{R} = 0)$ still satisfies (3.3) but now (3.4) becomes $\Delta\, \tilde{q}_1 = p_{\mathrm{e}} + p_{\mathrm{a}}\tilde{q}_0$ and

$$\tilde{q}_k = \frac{p_{\mathrm{b}}}{\Delta} \sum_{i=1}^{k-1} \tilde{q}_i \tilde{q}_{k-i} + \frac{p_{\mathrm{a}}}{\Delta} \tilde{q}_{k-1}\,, \quad k \geq 2.$$

The complexity introduced by the second addend $\frac{p_{\mathrm{a}}}{\Delta}\tilde{q}_{k-1}$ prevents us to directly derive the corresponding of Equation 3.6. As for the case of a sequence of compartments, a generating function approach is required. Let $\tilde{\phi}(z)$ be the probability generating function of the random variable $\mathbf{R}$; it satisfies

$$\tilde{\phi}(z) = p_{\mathrm{d}} + p_{\mathrm{b}}\tilde{\phi}^2(z) + p_{\mathrm{e}} z + p_{\mathrm{a}} z \tilde{\phi}(z)\,,$$

and its solution is given by

$$\tilde{\phi}(z) = \frac{1 - p_{\mathrm{a}} z - ((1 - p_{\mathrm{a}} z)^2 - 4p_{\mathrm{b}}p_{\mathrm{d}} - 4p_{\mathrm{b}}p_{\mathrm{e}} z)^{1/2}}{2p_{\mathrm{b}}}\,. \tag{3.27}$$

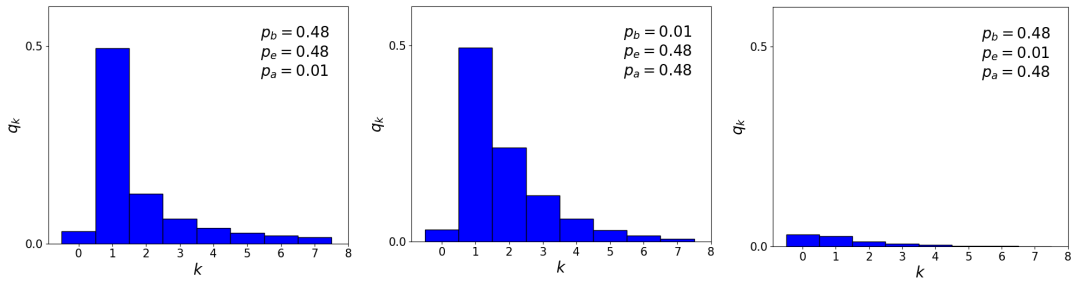


Figure 3.12: The quantity $\tilde{q}_k$ is the probability that k cells exit the compartment C_1, given that one cell enters. Results, using (3.27), are shown for three different choices of p_b, p_a and p_e, and $p_\mathrm{d} = 0.01$.

The well known Binomial Theorem in Elementary Algebra, can be generalised to real exponents (Arfken & Weber, 2016). Let $\alpha \in \mathbb{R}$, then

$$(1+x)^\alpha = \sum_{k=0}^{+\infty} \binom{\alpha}{k} x^k \,,$$

which converges for $|x| < 1$. We can apply the generalised Binomial Theorem to the square root of Equation (3.27),

$$\begin{aligned}
&\sqrt{(1-p_a z)^2 - 4p_b p_d - 4p_b p_e z} \\
&\quad = \sqrt{1-4p_b p_d}\left(1 - \frac{2p_a + 4p_b p_e}{1-4p_b p_d} z + \frac{p_a^2}{1-4p_b p_d} z^2\right)^{1/2} \\
&\quad = \Delta \sum_{k=0}^{+\infty} \binom{1/2}{k} \left(-\frac{2p_a + 4p_b p_e}{\Delta^2} z + \frac{p_a^2}{1-4p_b p_d} z^2\right)^k \\
&\quad = \Delta \sum_{k=0}^{+\infty} \binom{1/2}{k} \left(-\frac{2p_a + 4p_b p_e}{\Delta^2} z\right)^k \sum_{j=0}^{k} \binom{k}{j} \left(\frac{-p_a^2}{2p_a + 4p_b p_e} z\right)^j \,.
\end{aligned}$$

Substituting into the probability generating function we get

$$\begin{aligned}
\tilde{\phi}(z) &= \frac{1-p_a z}{2p_b} - \frac{\Delta}{2p_b}\sum_{k=0}^{+\infty}\binom{1/2}{k}\left(\sum_{j=0}^{k}\binom{k}{j}\left(\frac{-p_a^2}{2p_a+4p_bp_e}\right)^j z^j\right)\left(-\frac{2p_a+4p_bp_e}{\Delta^2}\right)^k z^k \\
&= \frac{1-p_a z}{2p_b} + \sum_{k=0}^{+\infty} c_{k-1}\frac{p_b^{k-1}}{\Delta^{2k-1}}\left(\frac{p_a}{2p_b}+p_e\right)^k\left(\sum_{j=0}^{k}\binom{k}{j}\left(\frac{-p_a^2}{2p_a+4p_bp_e}\right)^j z^j\right) z^k \\
&= \frac{1-p_a z}{2p_b} + \sum_{k=0}^{+\infty} c_{k-1}\frac{p_b^{k-1}}{\Delta^{2n}}\Delta^{k+1}\left(q_1+\frac{p_a}{2p_b}\right)^k\left(\sum_{j=0}^{k}\binom{k}{j}\left(\frac{-p_a^2}{2p_a+4p_bp_e}\right)^j z^j\right) z^k \\
&= \frac{1-p_a z}{2p_b} + \sum_{k=0}^{+\infty} a_k\left(\sum_{j=0}^{k}\binom{k}{j}\left(\frac{-p_a^2}{2p_a+4p_bp_e}\right)^j z^j\right) z^k\,,
\end{aligned}$$

where $a_k = c_{k-1}\frac{p_b^{k-1}}{\Delta^{2k}}\Delta^{k+1}\left(\tilde{q}_1+\frac{p_a}{2p_b}\right)^k$ and $c_k = -\frac{1}{2}\binom{1/2}{k+1}(-4)^{k+1}$ denotes the Catalan numbers. The coefficients of the generating function $\tilde{\phi}(z)$ are the $\tilde{q}_k$ we are interested in, that is,

$$\begin{aligned}
\tilde{q}_0 &= \frac{1-\Delta}{2p_b}\,,\\
\tilde{q}_1 &= \frac{p_a(1-\Delta)+2p_bp_e}{2\Delta p_b}\,,\\
\tilde{q}_k &= \sum_{i=0}^{\lfloor k/2\rfloor}\binom{k-i}{i}b^i a_{k-i}, \quad \text{for} \quad k\geq 2 \quad \text{with} \quad b=-\frac{p_a^2}{2p_a+4p_bp_e}\,,
\end{aligned}$$

where $\tilde{q}_0$ agrees with Equation (3.3) as we observed at the beginning of this Section. In order to make the expression of $\tilde{q}_k$ more explicit, we expand the term a_{k-i} as

$$a_{k-i} = c_{k-i-1}\frac{\Delta}{p_b}\left(\frac{2\Delta}{2p_b\tilde{q}_1+p_a}\right)^i\left(\frac{2p_b\tilde{q}_1+p_a}{2\Delta}\right)^k.$$

Thus, for $k\geq 2$,

$$\begin{aligned}
\tilde{q}_k &= \frac{\Delta}{p_b}\sum_{i=0}^{\lfloor k/2\rfloor} c_{k-i-1}\binom{k-i}{i}\left(\frac{-2p_a^2\Delta}{(2p_a+4p_bp_e)(2p_b\tilde{q}_1+p_a)}\right)^i\left(\frac{2p_b\tilde{q}_1+p_a}{2\Delta}\right)^k \\
&= \frac{\Delta}{p_b}\left(\frac{2p_b\tilde{q}_1+p_a}{2\Delta}\right)^k\sum_{i=0}^{\lfloor k/2\rfloor}\frac{1}{k-i}\binom{2k-2i-1}{k-i}\binom{k-i}{i}\left(\frac{-2p_a^2\Delta}{(2p_a+4p_bp_e)(2p_b\tilde{q}_1+p_a)}\right)^i.
\end{aligned} \tag{3.28}$$

Examples of $\tilde{q}_k$ are shown in Figure 3.12 for three different choices of $p_\mathrm{b}, p_\mathrm{a}$ and p_e.

The probability generating function $\tilde{\phi}(z)$ contains all the information of the random variable $\mathbf{R}$. Nevertheless, it is useful to compute the expected value of $\mathbf{R}$

$$\tilde{N} = \mathbb{E}(\mathbf{R}) = \frac{p_\mathrm{e} + p_\mathrm{a}}{1 - 2p_\mathrm{b} - p_\mathrm{a}}\,. \tag{3.29}$$

Figure 3.19 reports the quantity $\tilde{N}$ that characterises the population of cells exiting compartment C_1, as functions of p_b and p_d. In particular, we set $p_\mathrm{a} = 0.2$ so that each blue line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of $\tilde{N}$ given by (3.29). The part of the parameter space corresponding to $\tilde{N} > 1$ is at bottom right. Note that if $p_\mathrm{a} = 0$, $\tilde{N} = N$ as in (3.7).

Moreover, if $p_e = 0$, that is cells can die and self-renew, but they exit the compartment only via asymmetric division, Equation (3.27) is replaced by

$$\tilde{\phi}(z) = \frac{1 - p_\mathrm{a} z - ((1 - p_\mathrm{a} z)^2 - 4p_\mathrm{b} p_\mathrm{d})^{1/2}}{2p_\mathrm{b}}\,,$$

and Equation (3.28) simplifies into

$$\tilde{q}_k = \frac{\Delta}{p_b}\left(\frac{p_a}{2\Delta^2}\right)^k \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{1}{k-i}\binom{2k-2i-1}{k-i}\binom{k-i}{i}\left(-\Delta^2\right)^i, \quad \text{for } k \geq 2$$

with

$$\tilde{q}_0 = \frac{1-\Delta}{2p_b} \quad \text{and} \quad \tilde{q}_1 = \frac{p_a}{\Delta}\tilde{q}_0\,.$$

Recurrence relation As outlined in Section 3.2.3, in some situations it is convenient to generate $\tilde{q}_k$ via a recurrence relation. We rewrite (3.27) as

$$2p_\mathrm{b}\tilde{\phi}(z) = 1 - p_\mathrm{a} z - \tilde{w}(z), \qquad \text{where} \quad \tilde{w}^2(z) = \Delta - (2p_\mathrm{a} + 4p_\mathrm{b}p_\mathrm{e})z + p_\mathrm{a}^2 z^2.$$

Thus, $\tilde{\phi}(z)$ satisfies the differential equation

$$\tilde{w}^2(z)\tilde{\phi}'(z) + \tilde{w}'(z)\tilde{w}(z)\tilde{\phi}(z) + c(z) = 0$$

with $c(z) = (p_\mathrm{a}^2 - p_\mathrm{a} + 2p_\mathrm{a}p_\mathrm{b}p_\mathrm{e} - 4p_\mathrm{b}p_\mathrm{e})z + \Delta^2 - p_\mathrm{a} - 2p_\mathrm{b}p_\mathrm{e}$. As in Section 3.2.3, we make use of the fact that $\tilde{\phi}(z)$ is a generating function and it can be written

as a power series of z. Matching terms proportional to z^k yields the recurrence relation

$$\Delta^2(k+2)\tilde{q}_{k+2} = (2k+1)(p_\mathrm{a} + 2p_\mathrm{b}p_\mathrm{e})\tilde{q}_{k+1} - (k-1)p_\mathrm{a}^2\tilde{q}_k. \tag{3.30}$$

Note that, since $\tilde{w}^2(z)$ is a second-order polynomial in z, the recurrence relation (3.30) for $\tilde{q}_k$ with $M = 1$ is a three-term recurrence; whereas, if $p_\mathrm{a} = 0$, q_k is given by the two-term recurrence (3.10).

Asymptotic behaviour The probability generating function $\tilde{\phi}(z)$ contains information about the long tail of the distribution $\tilde{q}_k$. We follow the approach of Flajolet & Sedgewick (2009) and proceed as in Section 3.2.4 by locating the singularities of the probability generating function

$$\tilde{\phi}(z) = \frac{1 - p_a z - \sqrt{(1-p_a z)^2 - 4p_b p_d - 4p_b p_e z}}{2p_b}\,.$$

Note that, as for $\phi(z)$, $\tilde{\phi}(z)$ is a square root function. However, now the squared polynomial is a second-order polynomial, and thus $\tilde{\phi}(z)$ has two singularities at z_1, z_2 such that

$$p_a z^2 - 2(p_a + 2p_b p_e)z + \Delta^2 = 0\,.$$

Then,

$$\tilde{\phi}(z) = \frac{1 - p_a z - \sqrt{(z_2 - z)(z_1 - z)}}{2p_b} = \frac{1 - p_a z}{2p_b} - \frac{\sqrt{z - z_2}}{2p_b}\sqrt{z_1 - z}\,.$$

Following Flajolet & Sedgewick (2009) and without loss of generality, one can assume z_1 to be the dominant singularity ($z_1 \lesssim 1$ and $z_2 \gtrsim 0$). Since $\tilde{\phi}(z)$ is analytic in $z_1 - z_2$, we singularly expand the term $\sqrt{z_1 - z}$ around $z = z_1$ and analytically expand the factor $\frac{\sqrt{z-z_2}}{2p_b}$ around the point $z = z_1$. By linking, as in Flajolet & Sedgewick (2009), the coefficients of the power series with the location of the singularities, the following estimate holds

$$\begin{aligned}[z^k]\tilde{\phi}(z) &\sim -\frac{\sqrt{z_1 - z_2}}{2p_b}\sqrt{z_1 - z} \sim \frac{\sqrt{z_1 - z_2}}{2p_b} z_1^{-k}\frac{k^{-3/2}}{\Gamma(-1/2)} \\ &= \frac{\sqrt{z_1 - z_2}}{4\sqrt{\pi}p_b}\left(\frac{1}{z_1}\right)^k k^{-3/2}\,.\end{aligned}$$

Thus, as $k \to +\infty$,

$$\tilde{q}_k \sim \tilde{\gamma}^k k^{-3/2}, \tag{3.31}$$

where $\tilde{\gamma} = 1/z_1$ satisfies

$$(1 - 4p_\mathrm{b}p_\mathrm{d})\tilde{\gamma}^2 - (2p_\mathrm{a} + 4p_\mathrm{b}p_\mathrm{e})\tilde{\gamma} + p_\mathrm{a}^2 = 0.$$

We would like to underline, one more time, the universal behaviour $k^{-3/2}$ for small-k. In Figure 3.13, we report a comparison between the long-term behaviour of $\tilde{q}_k$ in the purely asymmetric case (in yellow), where cells can exit the compartment only via asymmetric division ($p_\mathrm{e} = 0$) and in the self-renewal case with $p_\mathrm{a} = 0$ (in green).

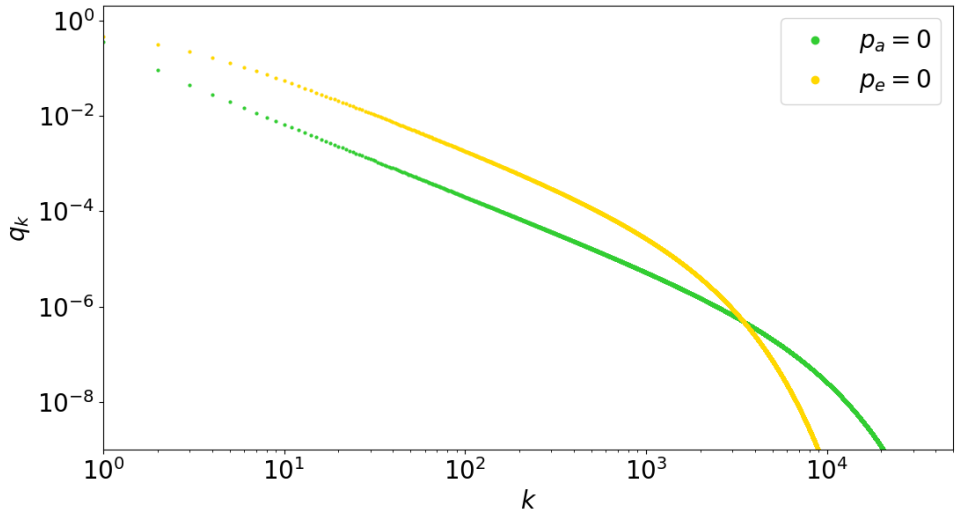


Figure 3.13: Comparing exit with and without asymmetric division. In green, the distribution q_k for the self-renewal case (3.6) with $p_\mathrm{a} = 0$. In yellow, the purely asymmetric case where we set $p_\mathrm{a} > 0$ and $p_\mathrm{e} = 0$, using (3.30). In both cases, $N = \tilde{N} = 25$ and $p_\mathrm{d} = 0.25$.

Generalisation to a sequence of $M = 2$ compartments We consider two non-identical compartments, i.e. $\tilde{\phi}_1(z) \neq \tilde{\phi}_2(z)$ and, as in (3.21), we define, by composition of generating functions,

$$\tilde{\Phi}_2(z) = \tilde{\phi}_1(\tilde{\phi}_2(z))\,,$$

the probability generating function of the random variable $\mathbf{R}$ for the number of cells exiting the second compartment C_2, for one cell entering C_1; in particular,

$$\tilde{\Phi}_2(z) = \frac{1 - p_a(1)\tilde{\phi}_2(z) - \sqrt{(1 - p_a(1)\tilde{\phi}_2(z))^2 - 4p_b(1)p_d(1) - 4p_b(1)p_e(1)\tilde{\phi}_2(z)}}{2p_b(1)}.$$

We can observe that $\tilde{\Phi}_2(z)$ and the probability generating function $\Phi_2(z)$ for the self-renewal case are both nested square root functions of the form,

$$p_3(z) + \sqrt[r]{p_1(z) + \sqrt{p_2(z)}},$$

with $p_1(z), p_2(z), p_3(z)$ polynomials. As detailed in Appendix A, if $p_{\mathrm{a}}(1) = p_{\mathrm{a}}(2) = 0$, then $p_1(z) = \Delta_1^2 - 2\frac{p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}$ is a constant; whereas in this scenario where asymmetric division is included, $\tilde{p}_1(z)$ is a z-dependent function. This brings complexity in the recurrence relation for $\tilde{Q}_k(2) = \mathbb{P}(\mathbf{R} = k)$. Following what done before in this Section for the case $M = 1$, we denote, for $i = 1, 2$,

$$\tilde{w}_i^2(z) = 1 - 4p_{\mathrm{b}}(i)p_{\mathrm{d}}(i) - (2p_{\mathrm{a}}(i) + 4p_{\mathrm{b}}(i)p_{\mathrm{e}}(i))z + p_{\mathrm{a}}^2(i)z^2.$$

Then, one can show that

$$2p_{\mathrm{b}}(1)\tilde{\Phi}_2(z) = H(z) - G(z),$$

where $H(z) = 1 - \dfrac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)} + \dfrac{p_{\mathrm{a}}(1)p_{\mathrm{a}}(2)}{2p_{\mathrm{b}}(2)}z + \dfrac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)}\tilde{w}_2(z)$ and

$$\begin{aligned} G(z) = \Bigg[& \frac{p_{\mathrm{a}}^2(1)p_{\mathrm{a}}^2(2)}{2p_{\mathrm{b}}^2(2)}z^2 + \left(\frac{p_{\mathrm{a}}(2)}{p_{\mathrm{b}}(2)}(p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1) - \frac{p_{\mathrm{a}}^2(1)}{p_{\mathrm{b}}^2(2)}(p_{\mathrm{a}}(2) + p_{\mathrm{b}}(2)p_{\mathrm{e}}(2))\right) z \\ & + \left(\frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)} - \frac{p_{\mathrm{a}}^2(1)(1 - p_{\mathrm{a}}(2)z)}{2p_{\mathrm{b}}^2(2)}\right)\tilde{w}_2(z) \\ & + \Delta_1^2 + \frac{p_{\mathrm{a}}^2(1)}{2p_{\mathrm{b}}^2(2)}(1 - 2p_{\mathrm{d}}(2)p_{\mathrm{b}}(2)) - \frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}\Bigg]^{1/2}. \end{aligned}$$

The reader may notice the different nature of the two functions $H(z)$ and $G(z)$: the first one is of first order in $\tilde{w}_2(z)$, whereas $G(z)$ is a square root function of $\tilde{w}_2(z)$. Thus, in order to generate the distribution $\tilde{Q}_k(2)$, one needs to compute two recurrences, namely h_k for $H(z)$ and g_k for $G(z)$. This strategy (inspired by

the work of Mathar (2021)) leads to a three-term recurrence for $H(z)$ and a six-term recurrence for $G(z)$. We report in Appendix A.2 the detailed mathematical calculations. Here, we just want to highlight the main two differences between the self-renewal case (analysed in Sections 3.2.3 and 3.2.6) and the one where asymmetric division is included. First of all, the functions $w_i(z)$ and $\tilde{w}_i(z)$ are of first and second order respectively. For the case $M = 1$, this leads to a two (if $p_a(1) = p_a(2) = 0$) and three-term recurrence (if $p_a(1), p_a(2) > 0$). As we can see in Appendix A.2, this difference in the degree of $w_i(z)$ and $\tilde{w}_i(z)$ is inherited also by the case of a sequence of two compartments and the six-term recurrence of the function $G(z)$. Secondly, and most importantly, there is a difference in the term $\tilde{p}_3(z)$. For $p_a(i) > 0$, for $i = 1, 2$, the probability generating function is such that $\tilde{p}_3(z)$ is a function of z and this implies the need of two recurrence function of separate nature for $\tilde{Q}_k(2)$; namely h_k and g_k as detailed in Appendix A.2.

3.3 Tracking age of product cells

In this Section, we focus on the age of product cells, rather than on the number of product cells reached at the end of the realisation. During its life, a cell undergoes mitosis (i.e. cellular nuclear division occurring when a cell divides to produce two daughter cells) a certain number of times that depends on its division rate. During this process the cellular DNA is replicated and this comes with an intrinsic risk of errors and onset of mutations in the daughters cell. In our mathematical model of cell differentiation, we assume the number of divisions a cell has undergone is strictly correlated with its age. With the aim of understanding how the product population is influenced by the differentiation via one or multiple compartments, we analyse the differentiation process in terms of number of divisions that cells have undergone during the sequence of compartments from a progenitor to product phenotype.

The initial cell, depicted in green in Figure 3.14, is said to be in generation zero and a cell that differentiate maintains its generation number. Daughter cells of the progenitor cell are said to be in generation 1 and daughter cells of a cell in generation n are in generation $n + 1$. The product population is consequently divided into classes by their generation number (see Figure 3.14 for $M = 1$). We

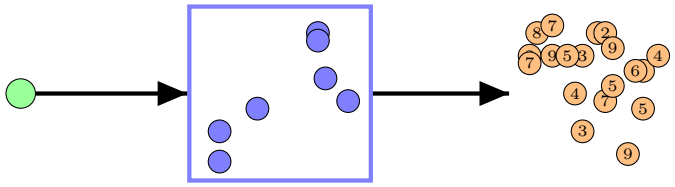


Figure 3.14: One-compartment system, $M = 1$. The green progenitor cell on the left is the founder of the population. The progenitor cell is said to be in generation 0. Blue cells may divide before they either die or leave the compartment. Thus, we may classify the set of product (orange) cells according to generation (number of divisions from the progenitor cell). Whenever a cell in generation n divides, the result is two cells in generation $n + 1$. Whereas differentiation event retains generation number. The final state of the process is a population of orange cells, each with its own generation number.

consider the random variable $\mathbf{G}$ for the generation number of a randomly-selected product cell, and its mean value $D = \mathbb{E}(\mathbf{G})$ that represents the mean age of the product cell population, measured in number of generations from the progenitor. As done in Section 3.2 for the random variable $\mathbf{R}$, we start by analysing a simple case and then generalise our observations: Section 3.3.1 reports the main results for the random variable $\mathbf{G}$ for the case of self-renewal division ($p_a = 0$) and "direct" differentiation, that is $M = 1$; Section 3.3.2 and Section 3.3.3 extend our results to consider a sequence of M compartments and the asymmetric event ($p_a \neq 0$) respectively.

3.3.1 G and its probability generating function

We let $\mathbf{G}$ be the random variable representing the generation number of a randomly-selected product cell. One realisation of the process is reported in Figure 3.15. To define the random variable $\mathbf{G}$, we begin by defining two simple random variables $\mathbf{U}$ and $\mathbf{V}$ with state space $\{0, 2\}$ and $\{0, 1\}$, respectively. Let

$$\mathbb{P}(\mathbf{U} = 0) = 1 - p_b, \ \mathbb{P}(\mathbf{U} = 2) = p_b, \ \mathbb{P}(\mathbf{V} = 0) = 1 - p_e, \text{ and } \mathbb{P}(\mathbf{V} = 1) = p_e.$$

If $M = 1$, we recall the random variables of a standard discrete-time branching process (Harris, 1963; Kimmel & Axelrod, 2002; Stirzaker, 2005); whereas later

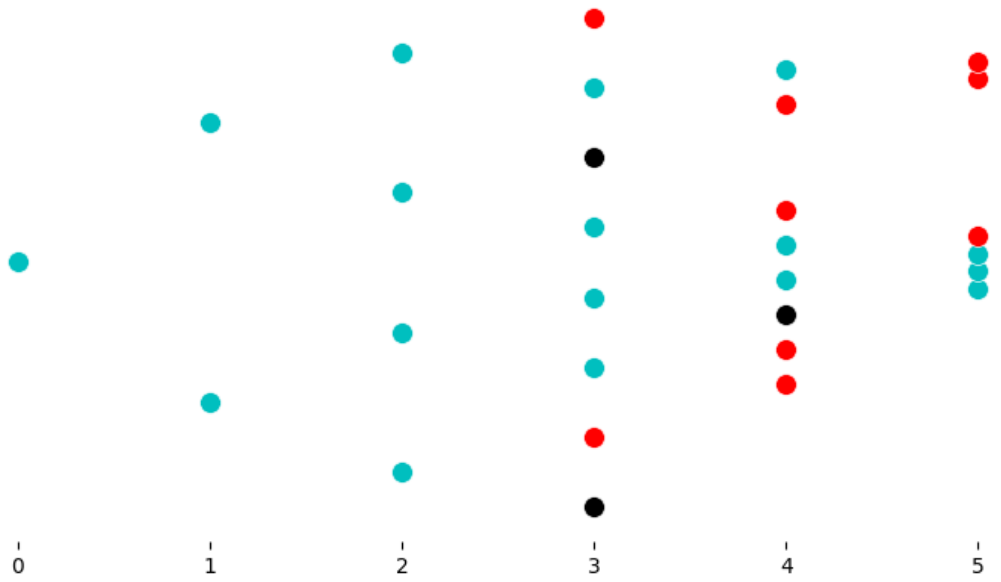


Figure 3.15: One realisation, $M = 1$, showing generation numbers from left to right, with $\mathbf{Z}_0 = 1$. Cyan cells divide, red cells exit, black cells die. In this realisation $\mathbf{Y}_0 = 0$, $\mathbf{Y}_1 = 0$, $\mathbf{Y}_2 = 0$, $\mathbf{Y}_3 = 2$, $\mathbf{Y}_4 = 4$ and $\mathbf{Y}_5 = 3$. Thus, we have $\mathbf{R} = 9$. We set parameters $p_\mathrm{b} = 0.45$, $p_\mathrm{d} = 0.15$.

in Section 3.3.2, where a sequence of compartments is analysed, we will consider a multi-type branching process with immigration. Let us introduce $\mathbf{Z}_0 = 1$ and

$$\mathbf{Z}_{n+1} = \sum_{j=1}^{\mathbf{Z}_n} \mathbf{U}_j, \qquad n = 0, 1, 2, \ldots, \tag{3.32}$$

where, for each j, $\mathbf{U}_j$ is an independent copy of $\mathbf{U}$ representing the number of daughter cells from one cell. The random variable $\mathbf{Z}_n$, defined in (3.32), represents the number of cells in generation n, whatever their fate. Moreover, since cells can exit the compartment, we define $\mathbf{Y}_n$ as the number of product cells in generation n

$$\mathbf{Y}_n = \sum_{j=1}^{\mathbf{Z}_n} \mathbf{V}_j, \qquad n = 0, 1, 2, \ldots, \tag{3.33}$$

where each $\mathbf{V}_j$ is an independent copy of $\mathbf{V}$. We can relate the mean values of

the random variables $\mathbf{Z}_n$ and $\mathbf{Y}_n$ as

$$\mathbb{E}(\mathbf{Y}_n) = p_{\mathrm{e}}\mathbb{E}(\mathbf{Z}_n) = p_{\mathrm{e}}(2p_{\mathrm{b}})^n. \tag{3.34}$$

For $M = 1$, the random variables $\mathbf{R}$ and $\mathbf{G}$ are defined so that the following equations hold

$$\mathbf{R} = \sum_{n=0}^{+\infty} \mathbf{Y}_n \qquad \text{and} \qquad \mathbb{P}\left(\mathbf{G} = n\right) = \frac{1}{N}\mathbb{E}(\mathbf{Y}_n). \tag{3.35}$$

We recall that in our framework the extinction condition (3.1), that is $2p_{\mathrm{b}} < 1$, holds. Accordingly, all cells in the sequence of compartments eventually die or reach the final compartment of product cells, that is $\mathbb{E}(\mathbf{Z}_n) \to 0$ and $\mathbb{E}(\mathbf{Y}_n) \to 0$, as $n \to +\infty$.

To keep track of the generation number of cells exiting compartment C_1, we introduce $\xi(z)$ the probability generating function of the random variable $\mathbf{G}$. By definition, we can find $\xi(z)$ from Equation (3.35),

$$\begin{aligned}
\xi(z) &= \frac{1}{N}\sum_{n=0}^{+\infty} \mathbb{E}(\mathbf{Y}_n) z^n \\
&= \frac{1}{N}\left(p_{\mathrm{e}} + \sum_{n=1}^{+\infty} \mathbb{E}(\mathbf{Y}_n) z^n\right) \\
&= \frac{p_{\mathrm{e}}}{N(1-2p_{\mathrm{b}}z)} \\
&= \frac{1-2p_{\mathrm{b}}}{1-2p_{\mathrm{b}}z}\,,
\end{aligned} \tag{3.36}$$

where in the last step we used (3.7) for the average number of product cells N. Given Equation (3.36), we can differentiate twice and find the average generation number in the product cell population, D, and the variance of $\mathbf{G}$. In particular, we get

$$D = \mathbb{E}(\mathbf{G}) = \frac{p_{\mathrm{e}}}{N}\sum_{n=1}^{+\infty} n(2p_{\mathrm{b}})^n = \frac{2p_{\mathrm{b}}}{1-2p_{\mathrm{b}}} \quad \text{and} \quad \mathrm{var}(\mathbf{G}) = D(D+1)\,. \tag{3.37}$$

Figure 3.16 reports the quantity D that characterises the population of cells exiting compartment C_1, as functions of p_{b} and p_{d}. In particular, each red line is

the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of D. Recall that, if p_d is fixed, $\lim_{p_\mathrm{b}\to 0} N = 1 - p_\mathrm{d}$. Similarly, we can observe that $\lim_{p_\mathrm{b}\to 0} D = 0$. Moreover, we have

$$\lim_{p_\mathrm{b}\to\frac{1}{2}} \frac{D}{N} = \frac{2}{1-2p_\mathrm{d}}\,.$$

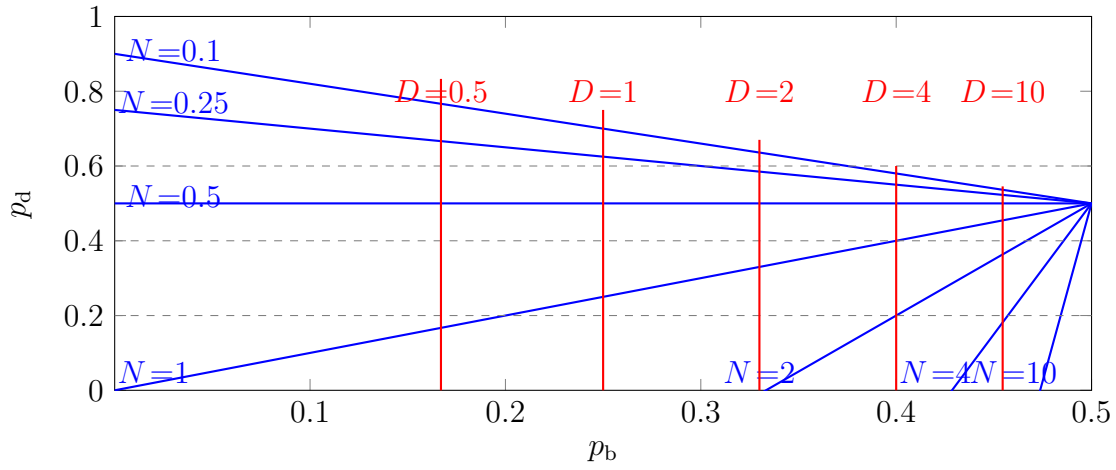


Figure 3.16: Lines of constant D (red) and lines of constant N (blue) in the part of the plane representing possible parameter values. The two quantities characterising the population of cells exiting a compartment are given, as functions of p_b and p_d by (3.7) and (3.37), respectively. Each blue line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of N. Each red line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of D. The triangular part of the parameter space corresponding to $N > 1$ is at bottom right.

3.3.2 Generalisation to a sequence of M compartments

In this Section, we consider the case of a sequence of $M > 1$ compartments. We assume that, for all $i = 1, \ldots, M$, cells exiting compartment C_i via differentiation retain their generation number. We define a multi-type branching process with immigration, of which one realisation for $M = 2$ is illustrated in Figure 3.17. Using the same notation as in Section 3.2.6, cells exiting compartment C_M are said to be product cells (depicted in orange in Figure 3.2).

3. WHY ARE CELL POPULATIONS MAINTAINED VIA MULTIPLE INTERMEDIATE COMPARTMENTS?

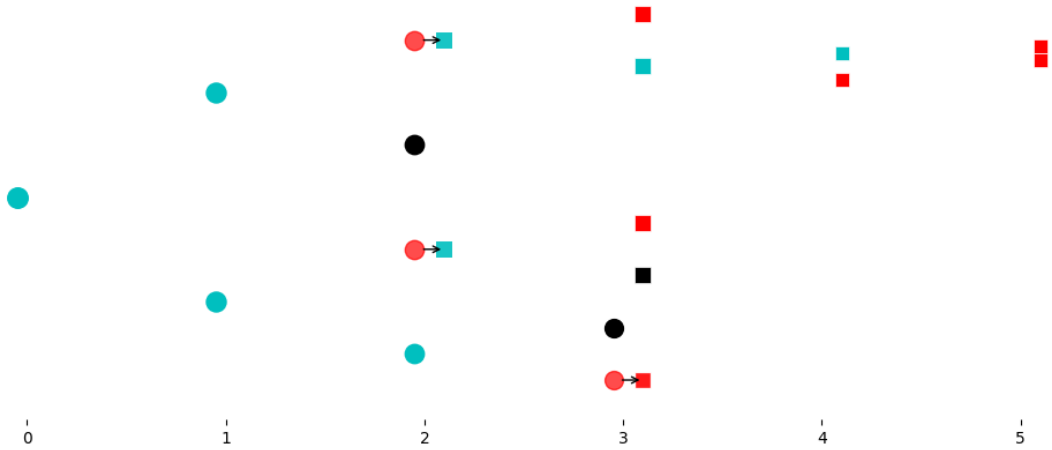


Figure 3.17: One realisation, showing generation numbers from left to right. Cells in the first compartment are shown as circles; cells in the second compartment are shown as squares. Cyan cells divide, red cells exit, black cells die. Arrows indicate transition from the first to the second compartment. In this realisation $\mathbf{Y}_0(1) = 0$, $\mathbf{Y}_1(1) = 0$, $\mathbf{Y}_2(2) = 2$, $\mathbf{Y}_3(1) = 1$, $\mathbf{Y}_4(1) = 0$ and $\mathbf{Y}_5(1) = 1$; $\mathbf{Y}_0(2) = 0$, $\mathbf{Y}_1(2) = 0$, $\mathbf{Y}_2(2) = 0$, $\mathbf{Y}_3(2) = 3$, $\mathbf{Y}_4(2) = 1$ and $\mathbf{Y}_5(2) = 2$. Thus $\mathbf{R} = 6$. Here $M = 2$ and we set parameters $p_\mathrm{b}(1) = p_\mathrm{b}(2) = 0.45$, $p_\mathrm{d}(1) = p_\mathrm{d}(2) = 0.15$.

To analyse the differentiation process via a sequence of compartments with respect to the generation number of the cells, we start by defining, for $i = 1, \ldots, M$, the random variables $\mathbf{U}(i)$ and $\mathbf{V}(i)$, with state space $\{0, 2\}$ and $\{0, 1\}$, respectively:

$$\mathbb{P}\left(\mathbf{U}(i) = 0\right) = 1 - p_\mathrm{b}(i), \qquad \mathbb{P}\left(\mathbf{U}(i) = 2\right) = p_\mathrm{b}(i)$$

$$\text{and} \quad \mathbb{P}\left(\mathbf{V}(i) = 0\right) = 1 - p_\mathrm{e}(i), \quad \mathbb{P}\left(\mathbf{V}(i) = 1\right) = p_\mathrm{e}(i).$$

Consequently, we can define two important random variables, $\mathbf{Z}_n(i)$ and $\mathbf{Y}_n(i)$, typical in branching processes:

- for $n \geq 0$ and $1 \leq i \leq M$, $\mathbf{Z}_n(i)$ is the number of generation n cells in compartment C_i, whatever their fate;
- for $n \geq 0$ and $1 \leq i \leq M$, $\mathbf{Y}_n(i)$ is the number of generation n cells that exit compartment C_i. That is, $\mathbf{Y}_n(i) \leq \mathbf{Z}_n(i)$.

Then, for cells in generation zero,

$$\mathbf{Z}_0(1) = 1 \quad \text{and} \quad \mathbf{Z}_0(i) = \mathbf{Y}_0(i-1), \qquad i = 2, \ldots, M\,,$$

and, for $n = 0, 1, \ldots$, we have $\mathbf{Z}_{n+1}(1) = \sum_{j=1}^{\mathbf{Z}_n(1)} \mathbf{U}_j(1)$ and

$$\mathbf{Z}_{n+1}(i) = \mathbf{Y}_{n+1}(i-1) + \sum_{j=1}^{\mathbf{Z}_n(i)} \mathbf{U}_j(i), \quad i = 2, \ldots, M\,,$$

where, for each j, $\mathbf{U}_j(i)$ is an independent copy of $\mathbf{U}(i)$. Also, for $n = 0, 1, \ldots$,

$$\mathbf{Y}_n(i) = \sum_{j=1}^{\mathbf{Z}_n(i)} \mathbf{V}_j(i), \qquad i = 1, \ldots, M\,.$$

where, for each j, $\mathbf{V}_j(i)$ is an independent copy of $\mathbf{V}(i)$.

When working with a sequence of compartments we find helpful to define either "compartment" or "total" variables. In particular, we denote with $\mathbf{G}_i$ the "compartment" random variable for the generation number of a randomly-selected cell exiting compartment C_i for a cell entering that compartment. Its probability distribution is defined as

$$\mathbb{P}\left(\mathbf{G}_i = n\right) = \frac{1}{N_i}\mathbb{E}(\mathbf{Y}_n(i))\,,$$

where $N_i = \mathbb{E}(\mathbf{R}_i)$ is the *amplification factor* defined in Equation (3.20). For each $\mathbf{G}_i$, we introduce its probability generating function:

$$\xi_i(z) = \frac{p_{\mathrm{e}}(i)}{N_i}\sum_{n=1}^{+\infty}(2z\,p_{\mathrm{b}}(i))^n = \frac{1-2p_{\mathrm{b}}(i)}{1-2p_{\mathrm{b}}(i)z}, \quad i = 1, \ldots, M\,. \tag{3.38}$$

Single-compartment quantities can be defined as well: we let D_i be the average increase in the generation number in compartment C_i,

$$D_i = \mathbb{E}(\mathbf{G}_i) = \frac{2p_{\mathrm{b}}(i)}{1-2p_{\mathrm{b}}(i)} \qquad i = 1, \ldots, M. \tag{3.39}$$

The corresponding "total" variables are denoted with $\mathbf{R}$ and $\mathbf{G}$. In particular, we express the number of product cells $\mathbf{R}$ in terms of the number of cells, of all generations, exiting the last compartment of the sequence,

$$\mathbf{R} = \sum_{n=0}^{+\infty} \mathbf{Y}_n(M).$$

Also, the probability distribution of the generation number of a randomly-selected product cell, denoted with $\mathbf{G}$, is defined as

$$\mathbb{P}(\mathbf{G} = n) = \frac{1}{N}\mathbb{E}(\mathbf{Y}_n(M)). \tag{3.40}$$

Recall in Section 3.2.6 we defined, for a sequence of M compartments, N as the mean number of product cells for every progenitor cell; now we define D as the average generation number of a product cell. It holds

$$N = \mathbb{E}(\mathbf{R}) = N_1 N_2 \dots N_M \qquad \text{and} \quad D = \mathbb{E}(\mathbf{G}) = D_1 + D_2 + \cdots D_M. \tag{3.41}$$

Moreover, we denote with $\Xi(z)$ the probability generating function of $\mathbf{G}$. In particular, $\Xi(z) = \sum_{n=0}^{+\infty} \mathbb{P}(\mathbf{G} = n) z^n$ and it can be expressed as product of the single-compartment functions

$$\Xi_M(z) = \xi_1(z)\xi_2(z)\cdots\xi_M(z)\,, \tag{3.42}$$

where, for each $i = 1, \dots, M$, $\xi_i(z)$ has been defined in (3.38).

Figure 3.18 shows the distribution of $\mathbf{G}$, defined as in (3.40), for $M = 1, 2, 3$. Here, we assumed equal compartments, that is $p_{\mathrm{a}}(i) = p_{\mathrm{a}}$, $p_{\mathrm{d}}(i) = p_{\mathrm{d}}$ for $i = 1, \dots, M$, and we consider two different scenarios, $p_{\mathrm{a}} = 0$ and $p_{\mathrm{a}} = 0.3$. Note that $\mathbb{P}(\mathbf{G} = n)$ narrows as the number of intermediate compartments increases and as p_{d} decreases. Moreover, in case of only self-renewal division, that is $p_{\mathrm{a}} = 0$, the distribution of $\mathbf{G}$ has a longer tail and it is thus more likely to observe product cells that underwent a higher number of divisions.

3.3.3 Generalisation to asymmetric division

In Section 3.2.7, asymmetric division has been accommodated in our framework to study the number of product cells; here, we include asymmetric division in the cells' generations analysis. In Section 3.3.1 we defined $\mathbf{G}$ via some simpler random variables $\mathbf{U}$, $\mathbf{V}$; now with the inclusion of asymmetric division, we need to define a third variable to consider cells that are increasing both the generation and the compartment index at the same time, i.e. the daughter cell exiting C_i when an asymmetric division event happens in C_i.

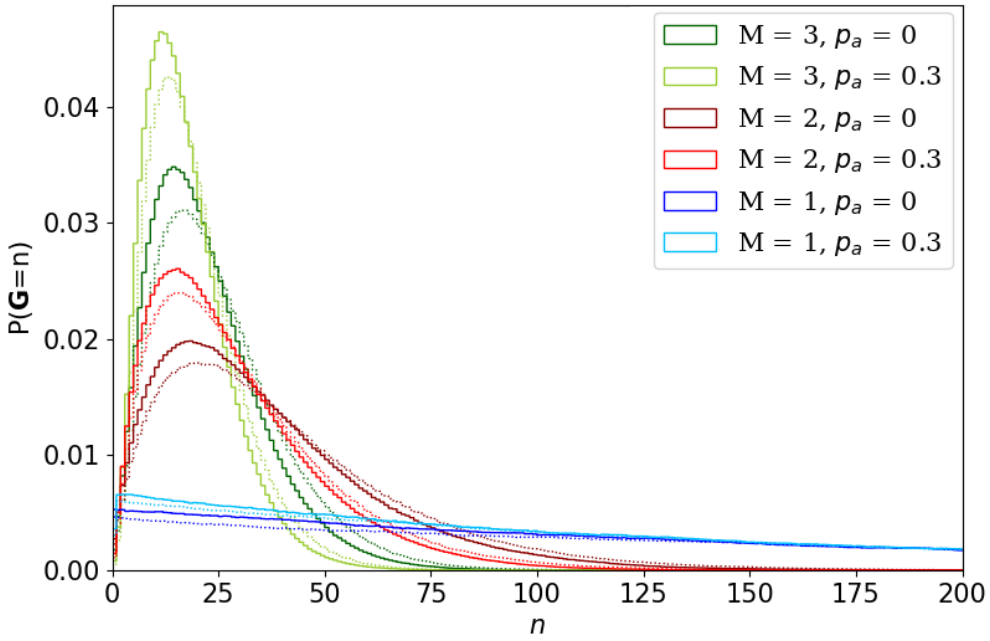


Figure 3.18: The probability distribution of the random variable $\mathbf{G}$, the generation number in the product cell population. One, two and three compartments have been shown, in green, red and blue respectively. In all cases, $N = 100$ and we assume equal compartments, that is $p_{\mathrm{d}}(i) = p_{\mathrm{d}}$, $p_{\mathrm{b}}(i) = p_{\mathrm{b}}$, $p_{\mathrm{e}}(i) = p_{\mathrm{e}}$ and $p_{\mathrm{a}}(i) = p_{\mathrm{a}}$, for $i = 1, \ldots, M$. Solid lines represents the distribution for $p_{\mathrm{d}} = 0$ and two different choices of probability of asymmetric division, $p_{\mathrm{a}} = 0$ (darker colour) and $p_{\mathrm{a}} = 0.3$ (lighter colour). Dotted lines represents the distribution for $p_{\mathrm{d}} = 0.05$ and two different choices of probability of asymmetric division, $p_{\mathrm{a}} = 0$ (darker colour) and $p_{\mathrm{a}} = 0.3$ (lighter colour).

We assume to have a sequence of M compartments and, following arguments in Section 3.3.2, we define for $i = 1, \ldots, M$, "compartment" random variables $\mathbf{U}(i)$, $\mathbf{V}(i)$ and $\mathbf{W}(i)$ with state space $\{0, 1, 2\}$, $\{0, 1\}$, $\{0, 1\}$ respectively:

$$\mathbb{P}\left(\mathbf{U}(i) = 0\right) = 1 - p_{\mathrm{b}}(i) - p_{\mathrm{a}}(i), \quad \mathbb{P}\left(\mathbf{U}(i) = 1\right) = p_{\mathrm{a}}(i), \quad \mathbb{P}\left(\mathbf{U}(i) = 2\right) = p_{\mathrm{b}}(i),$$

$$\mathbb{P}\left(\mathbf{V}(i) = 0\right) = 1 - p_{\mathrm{e}}(i), \qquad \mathbb{P}\left(\mathbf{V}(i) = 1\right) = p_{\mathrm{e}}(i),$$

$$\text{and} \qquad \mathbb{P}\left(\mathbf{W}(i) = 0\right) = 1 - p_{\mathrm{a}}(i), \qquad \mathbb{P}\left(\mathbf{W}(i) = 1\right) = p_{\mathrm{a}}(i)\,.$$

Similarly to Section 3.3.2 for the case of self-renewal division, we introduce $\mathbf{Z}_n(i)$ for the number of cells in generation n in compartment C_i, whatever their fate. We have

$$\begin{aligned}
&\mathbf{Z}_0(1) = 1, \quad \mathbf{Z}_0(i) = \mathbf{Y}_0(i-1), \qquad i = 2, \dots, M \\
&\mathbf{Z}_{n+1}(i) = \mathbf{Y}_{n+1}(i-1) + \sum_{j=1}^{\mathbf{Z}_n(i)} \mathbf{U}_j(i), \qquad i = 2, \dots, M, \quad n = 0, 1, 2, \dots,
\end{aligned}$$

where, for each j, $\mathbf{U}_j(i)$ is an independent copy of $\mathbf{U}(i)$. Also, definition (3.33) is replaced by

$$\mathbf{Y}_n(i) = \sum_{j=1}^{\mathbf{Z}_n(i)} \mathbf{V}_j(i) + \sum_{j=1}^{\mathbf{Z}_{n-1}(i)} \mathbf{W}_j(i), \qquad i = 1, \dots, M, \quad n = 1, 2, \dots,$$

where each $\mathbf{V}_j(i)$, $\mathbf{W}_j(i)$ are an independent copy of $\mathbf{V}(i)$ and $\mathbf{W}(i)$ respectively. A generalisation of Equation (3.34), for the mean values of $\mathbf{Y}_n(i)$, for $i = 1, \dots, M$, is

$$\begin{aligned}
\mathbb{E}(\mathbf{Y}_0(i)) &= p_{\mathrm{e}}(i) \\
\mathbb{E}(\mathbf{Y}_n(i)) &= p_{\mathrm{e}}(i)\, \mathbb{E}(\mathbf{Z}_n(i)) + p_{\mathrm{a}}(i)\, \mathbb{E}(\mathbf{Z}_{n-1}(i)) \\
&= p_{\mathrm{e}}(i)(2p_{\mathrm{b}}(i) + p_{\mathrm{a}}(i))^n + p_{\mathrm{a}}(i)(2p_{\mathrm{b}}(i) + p_{\mathrm{a}}(i))^{n-1}, \quad n \geq 1\,.
\end{aligned}$$

Note that condition (3.1) ensures the extinction of the process thus, for $k \geq 1$ and $i = 1, \dots, M$,

$$\lim_{n \to +\infty} \mathbb{P}\left(\mathbf{Z}_n(i) = k\right) = 0 \quad \text{and} \quad \lim_{n \to +\infty} \mathbb{P}\left(\mathbf{Y}_n(i) = k\right) = 0\,.$$

Recall that, for ease of notation, we denote the asymmetric variables and quantities with the corresponding symbol used in Sections 3.3.1 and 3.3.2, but with a tilde on top, e.g. $\tilde{N}$ where N. Given the "compartment" variable $\mathbf{G}_i$ we can define its probability generating function as

$$\begin{aligned}
\tilde{\xi}_i(z) &= \frac{1}{\tilde{N}_i} \sum_{n=0}^{+\infty} \mathbb{E}(\mathbf{Y}_n(i)) z^n = \frac{1}{\tilde{N}_i} \left[p_{\mathrm{e}}(i) + \sum_{n=1}^{+\infty} \mathbb{E}(\mathbf{Y}_n(i)) z^n \right] \\
&= \frac{p_{\mathrm{e}}(i) + p_{\mathrm{a}}(i) z}{\tilde{N}_i (1 - (2p_{\mathrm{b}}(i) + p_{\mathrm{a}}(i)) z)}\,.
\end{aligned} \tag{3.43}$$

Differentiating once, one can compute the expected value of $\mathbf{G}_i$ for the case of asymmetric division,

$$\tilde{D}_i = \mathbb{E}(\mathbf{G}_i) = \frac{1}{p_{\mathrm{e}}(i) + p_{\mathrm{a}}(i)} \frac{p_{\mathrm{a}}(i) + p_{\mathrm{e}}(i)(2p_{\mathrm{b}}(i) + p_{\mathrm{a}}(i))}{1 - 2p_{\mathrm{b}}(i) - p_{\mathrm{a}}(i)} \tag{3.44}$$

where we used that $1 = p_{\mathrm{d}}(i) + p_{\mathrm{a}}(i) + p_{\mathrm{b}}(i) + p_{\mathrm{e}}(i)$ for $i = 1, \dots, M$. Also, the variance of $\mathbf{G}_i$ can be computed from Equation (3.43),

$$\mathrm{var}(\mathbf{G}_i) = \frac{2p_{\mathrm{a}}(i) + p_{\mathrm{e}}(i)}{p_{\mathrm{e}}(i)} \tilde{D}_i \left(\tilde{D}_i + 1\right).$$

Equation (3.42) still holds. Thus, the probability generating function of $\mathbf{G}$ is given by the product of the single-compartment generating functions

$$\tilde{\Xi}_M(z) = \tilde{\xi}_1(z)\, \tilde{\xi}_2(z) \cdots \tilde{\xi}_M(z).$$

In particular, if $M = 2$ we let $\tilde{N} = \tilde{N}_1 \tilde{N}_2$ and we have

$$\begin{aligned}
\tilde{\Xi}_2(z) = \tilde{\xi}_1(z)\, \tilde{\xi}_2(z) &= \left(\frac{1}{\tilde{N}_1} \sum_{n=0}^{+\infty} \mathbb{E}(\mathbf{Y}_n(1)) z^n\right) \left(\frac{1}{\tilde{N}_2} \sum_{n=0}^{+\infty} \mathbb{E}(\mathbf{Y}_n(2)) z^n\right) \\
&= \frac{p_{\mathrm{e}}(1)p_{\mathrm{e}}(2)}{\tilde{N}} + \frac{p_{\mathrm{e}}(1)}{\tilde{N}} \left(p_{\mathrm{e}}(2) + \frac{p_{\mathrm{a}}(2)}{2p_{\mathrm{b}}(2) + p_{\mathrm{a}}(2)}\right) \frac{(2p_{\mathrm{b}}(2) + p_{\mathrm{a}}(2))z}{1 - (2p_{\mathrm{b}}(2) + p_{\mathrm{a}}(2))z} \\
&\quad + \frac{p_{\mathrm{e}}(2)}{\tilde{N}} \left(p_{\mathrm{e}}(1) + \frac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(1) + p_{\mathrm{a}}(1)}\right) \frac{(2p_{\mathrm{b}}(1) + p_{\mathrm{a}}(1))z}{1 - (2p_{\mathrm{b}}(1) + p_{\mathrm{a}}(1))z} \\
&\quad + \frac{1}{\tilde{N}} \left(p_{\mathrm{e}}(1) + \frac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(1) + p_{\mathrm{a}}(1)}\right) \left(p_{\mathrm{e}}(2) + \frac{p_{\mathrm{a}}(2)}{2p_{\mathrm{b}}(2) + p_{\mathrm{a}}(2)}\right) \\
&\quad \times \sum_{n=1}^{+\infty} \left(\sum_{k=1}^{n-1} (2p_{\mathrm{b}}(1) + p_{\mathrm{a}}(1))^k (2p_{\mathrm{b}}(2) + p_{\mathrm{a}}(2))^{n-k}\right).
\end{aligned}$$

If $M = 1$, then $\tilde{D} = \tilde{D}_i$ and one can plot $\tilde{D}$ in the plane p_{b}-p_{d}, as in Figure 3.19. Since $p_{\mathrm{a}} \neq 0$, we restrict to the part of the plane bounded by

$$2p_{\mathrm{b}} + p_{\mathrm{a}} < 1 \quad \text{and} \quad p_{\mathrm{d}} < 1 - p_{\mathrm{b}} - p_{\mathrm{a}},$$

corresponding to the extinction condition in (3.1) and condition $p_{\mathrm{e}} > 0$ derived from (3.44), respectively. We recall the reader that Figure 3.19 reports also N as a function of $p_{\mathrm{b}}, p_{\mathrm{d}}$ summarising the effects of asymmetric division, compared to Figure 3.16.

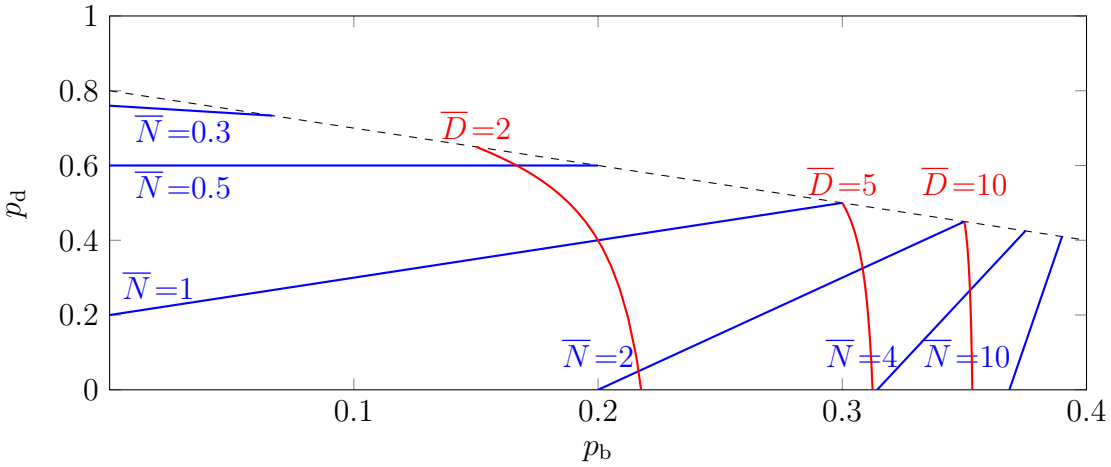


Figure 3.19: We set asymmetric division probability to $p_a = 0.2$. Lines of constant $\tilde{D}$ (red) and lines of constant $\tilde{N}$ (blue) in the part of the plane, delimited by the dashed line, representing possible parameter values. The two quantities characterising the population cells exiting a compartment, as functions of p_b and p_d, (3.29) and (3.44). Each blue line is the set of pairs (p_b, p_d) corresponding to the indicated value of $\tilde{N}$. Each red line is the set of pairs (p_b, p_d) corresponding to the indicated value of $\tilde{D}$. The part of the parameter space corresponding to $N > 1$ is at bottom right.

3.4 The relationship between events' probabilities, N and D

Given our compartmental mathematical system, we can express mean quantities N and D as a function of probabilities p_b, p_d as well as find a relationship between these two quantities. This can be extremely useful from an experimental point of view: design of experiments in cell biology often allows to easily quantify population of cells and their replicative age, that is N and D respectively.

3.4.1 Parameterisation in terms of N and D

If $p_a(i) = 0$, then $p_d(i) + p_b(i) + p_e(i) = 1$ and we can recover any of the probabilities $p_d(i)$, $p_b(i)$ or $p_e(i)$ given the other two. We can generalise (3.9) and express

probabilities of compartment C_i in terms of N_i itself along with $p_{\mathrm{d}}(i)$:

$$p_{\mathrm{b}}(i) = \frac{N_i - 1 + p_{\mathrm{d}}(i)}{2N_i - 1} \qquad \text{and} \quad p_{\mathrm{e}}(i) = \frac{N_i(1 - 2p_{\mathrm{d}}(i))}{2N_i - 1}. \tag{3.45}$$

Moreover, we may express events' probabilities in terms of N_i and D_i. In particular,

$$p_{\mathrm{b}}(i) = \frac{1}{2}\frac{D_i}{D_i + 1} \qquad \text{and} \qquad p_{\mathrm{e}}(i) = \frac{N_i}{D_i + 1}.$$

These relationships enable model parameters, $p_{\mathrm{b}}(i)$ and $p_{\mathrm{e}}(i)$, to be determined from experimentally-measurable quantities, $N_i = \mathbb{E}(\mathbf{R}_i)$ and $D_i = \mathbb{E}(\mathbf{G}_i)$. Moreover, we can rearrange Equation (3.8) and express the variance of the "total" variable $\mathbf{R}$ in terms of N and D,

$$V = \mathrm{var}(\mathbf{R}) = N^2(D - 1) + N\,.$$

If considering a sequence of compartments, e.g. $M = 10$, we can visually represent the number of cells of generation n exiting compartment C_i in a heat map. Figure 3.20 shows how the number of cells in generation n varies according to the compartment C_i; we consider equal compartments and we set $p_{\mathrm{d}} = 0.1$ and $p_{\mathrm{a}} = 0$.

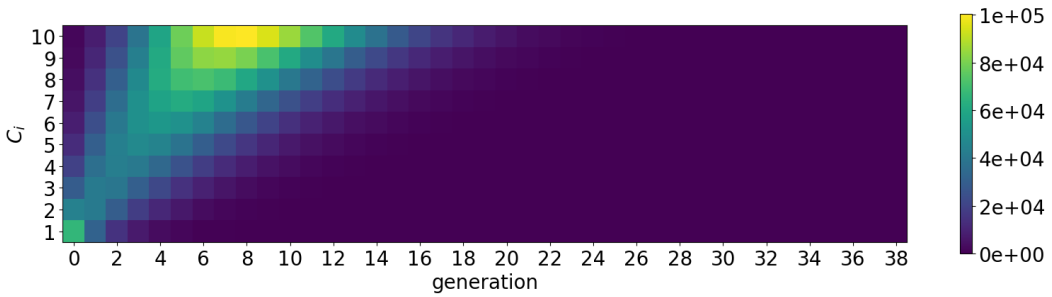


Figure 3.20: Heat map for the number of cells of generation n exiting compartment C_i for a sequence of $M = 10$ identical compartments. We set $N = 10$, $p_{\mathrm{d}} = 0.1$ and $p_{\mathrm{a}} = 0$.

3.4.2 Relationship between N and D

If $M = 1$ and $p_a = 0$, we can combine (3.7) and (3.37) and get the following linear relationship between D and N:

$$D = \frac{2N - 1}{1 - 2p_d} - 1\,. \tag{3.46}$$

A similar equation holds for "compartments" quantities D_i and N_i. We observe that, given $N > 1$, the minimum possible value of D is found when $p_d = 0$. That is

$$D_{\min} = 2(N - 1).$$

Figure 3.21, on the left, shows the linear relationship (3.46) between N and D for increasing values of p_d. We show in Section 3.4.3 that, for $M \geq 1$, D is a decreasing function of p_d.

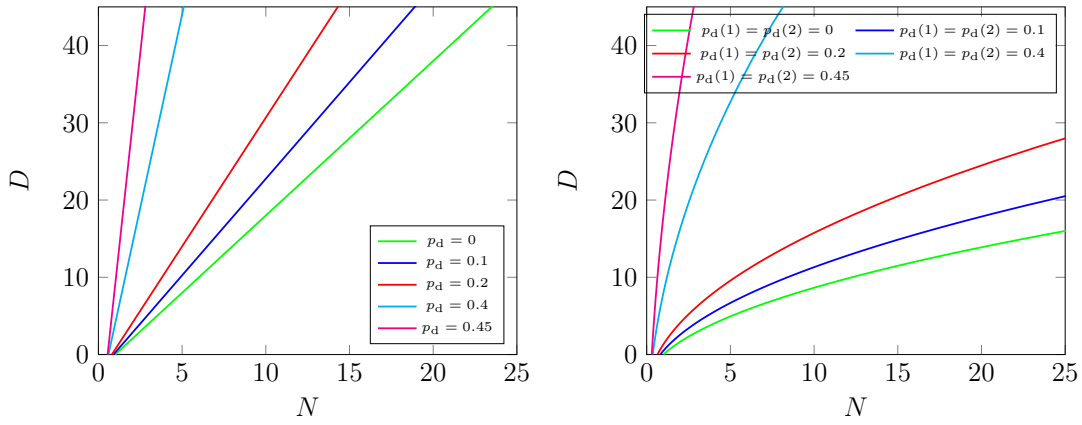


Figure 3.21: The average generation number of product cells, as a function of the mean number of exiting cells. Left: $M = 1$. Given (3.46), N and D are linearly related. Right: $M = 2$, with parameters chosen so that $N_1 = N_2$. Given a value of N, D is lower when $M = 2$ (proportional to $\sqrt{N}$ as $N \to +\infty$) than when $M = 1$ (proportional to N as $N \to +\infty$).

3.4.3 A sequence of compartments and minimisation of D

Every round of division brings with it a risk of mutation, senescence or exhaustion (Frank, 2007), thus an excessive clonality may increase the risk of cancerous

mutations becoming established (Wainscoat & Fey, 1990). The compartmental model that has been developed in this Chapter may help in understanding how should a sequence of M compartments be constructed in order to yield a given amplification of progenitor to product cells and, at the same time, minimise the average number of divisions. Mathematically, given the mean number of product cells N resulting from a sequence of compartments, we seek to minimise D, given by Equation (3.41). In order to do that, we make use of single-compartment quantities N_i in Equation (3.20) and write (3.39) as

$$D_i = \alpha_i N_i - \beta_i, \quad \text{where} \quad \alpha_i = \frac{2}{1-2p_{\mathrm{d}}(i)} \quad \text{and} \quad \beta_i = \frac{2-2p_{\mathrm{d}}(i)}{1-2p_{\mathrm{d}}(i)}.$$

Consequently, if $N \geq 1$, the mean number of divisions over the whole sequence of compartments is given by

$$D = \sum_{i=1}^{M}(\alpha_i N_i - \beta_i) = \sum_{i=1}^{M}\left(\frac{2N_i}{1-2p_{\mathrm{d}}(i)} - \frac{2-2p_{\mathrm{d}}(i)}{1-2p_{\mathrm{d}}(i)}\right). \tag{3.47}$$

Equation (3.47) is an increasing function of $p_{\mathrm{d}}(i)$. Indeed, by differentiating (3.47) with respect to $p_{\mathrm{d}}(i)$, we get

$$\frac{\partial D}{\partial p_d(i)} = \frac{4N_i - 2}{(1-2p_{\mathrm{d}}(i))^2}.$$

That is,

$$\frac{\partial D}{\partial p_d(i)} > 0 \quad \Leftrightarrow \quad N_i > \frac{1}{2}.$$

Since $N_i \geq 1$, the average generation number D is always an increasing function of $p_{\mathrm{d}}(i)$. Thus, the minimum D is observed if $p_{\mathrm{d}}(i) = 0$ for all compartments C_i, for $i = 1, \ldots, M$. One can use the Lagrange multiplier method to impose the constraint $N = N^*$ and find the compartments' arrangement that minimise the average generation of product cells. Let

$$\begin{aligned}
&\mathcal{L}(p_{\mathrm{d}}(1), \ldots, p_{\mathrm{d}}(M), p_{\mathrm{b}}(1), \ldots, p_{\mathrm{b}}(M), \Lambda) \\
&\quad = D - \Lambda(N - N^*) \\
&\quad = \sum_{i=1}^{M} \frac{2p_{\mathrm{b}}(i)}{1-2p_{\mathrm{b}}(i)} - \Lambda\left(\prod_{i=1}^{M} \frac{1-p_{\mathrm{b}}(i)-p_{\mathrm{d}}(i)}{1-2p_{\mathrm{b}}(i)} - N^*\right).
\end{aligned}$$

3. WHY ARE CELL POPULATIONS MAINTAINED VIA MULTIPLE INTERMEDIATE COMPARTMENTS?

We compute the gradient of the Lagrangian, $i = 1, \dots, M$

$$\left(\frac{\partial \mathcal{L}(\boldsymbol{p_d}, \boldsymbol{p_d}, \Lambda)}{\partial p_{\rm b}(i)}, \frac{\partial \mathcal{L}(\boldsymbol{p_d}, \boldsymbol{p_d}, \Lambda)}{\partial p_{\rm d}(i)}, \frac{\partial \mathcal{L}(\boldsymbol{p_d}, \boldsymbol{p_d}, \Lambda)}{\partial \Lambda} \right) = \left(\frac{2}{(1-p_{\rm b}(i))^2} \left(1 - \Lambda \frac{N^*}{\alpha_i N_i} \right), \frac{\Lambda}{1-2p_{\rm b}(i)}, N - N^\star \right) .$$

where we denoted with $\boldsymbol{p_d} = (p_{\rm d}(1), \dots, p_{\rm d}(M))$ and $\boldsymbol{p_d} = (p_{\rm b}(1), \dots, p_{\rm b}(M))$. By setting it to zero we find

$$1 - \Lambda \frac{N^*}{\alpha_i N_i} = 0, \quad i = 1, \dots, M .$$

Thus, the minimum average generation D is reached if the following conditions are satisfied

$$\alpha_1 N_1 = \alpha_2 N_2 = \cdots = \alpha_M N_M. \tag{3.48}$$

We define the arithmetic and geometric means of the parameters α_i:

$$\bar{\alpha} = \frac{1}{M} \sum_{i=1}^{M} \alpha_i \quad \text{and} \quad \tilde{\alpha} = \left(\prod_{i=1}^{M} \alpha_i \right)^{1/M} ,$$

so that we can re-write condition (3.48) for the optimal values of N_i as

$$\alpha_i N_i = N^{1/M} \tilde{\alpha} , \qquad \text{for each} \quad 1 \le i \le M.$$

The corresponding minimum value of D is then given by

$$D_{\rm min} = \sum_{i=1}^{M} (\alpha_i N_i - \beta_i) = M \left(\tilde{\alpha} N^{1/M} - \frac{1}{2} \bar{\alpha} - 1 \right), \tag{3.49}$$

which is an increasing function of each of the $p_{\rm d}(i)$, for $1 \le i \le M$ and a decreasing function of the number of M compartment in the sequence. We can observe that if $p_{\rm d}(i)$ does not depend on i, then N_i is also independent of i. That is, if the death probability does not vary from compartment to compartment, then the optimal arrangement of division rates is such that each compartment has the same amplification factor $N_i = N^{1/M}$. Then, we have

$$D_{\rm min} = \frac{2M}{1-2p_{\rm d}} \left(N^{1/M} - 1 + p_{\rm d} \right) .$$

In Figure 3.21 we plot the average generation number, D, as a function of the mean number of exiting cells N for $M = 1$ (on the left) and $M = 2$ (on the right) and two identical compartments, that is $N_1 = N_2$. The green lines show cases where there is no cell death; note that for $p_{\mathrm{d}}(1) = p_{\mathrm{d}}(2) = 0$ we have the lowest value of D. A comparison between direct differentiation $M = 1$ and the case of one intermediate compartment $M = 2$, given the same value of N, already shows the difference between a single compartment and multiple compartments sequence. We can observe how if $M = 1$, N and D are linearly related, whereas for $M = 2$, D is proportional to the square root of N, according to expression (3.49).

3.4.4 Generalisation to asymmetric division

When generalising to the case $p_{\mathrm{a}}(i) > 0$, we have $p_{\mathrm{d}}(i) + p_{\mathrm{b}}(i) + p_{\mathrm{e}}(i) + p_{\mathrm{a}}(i) = 1$ and we can still recover anyone of $p_{\mathrm{d}}(i)$, $p_{\mathrm{b}}(i)$, $p_{\mathrm{e}}(i)$ or $p_{\mathrm{a}}(i)$ given the other three. In this case, however, we have an additional degree of freedom and thus a third variable to express events' probabilities of compartments C_i is needed. In terms of $\tilde{N}_i$, $\tilde{D}_i$ and $p_{\mathrm{a}}(i)$ we have

$$p_{\mathrm{b}}(i) = \frac{\tilde{N}_i\left(\tilde{D}_i(1-p_{\mathrm{a}}(i)) - p_{\mathrm{a}}(i)\right) - p_{\mathrm{a}}(i)}{2\tilde{N}_i(\tilde{D}_i+1)}, \qquad p_{\mathrm{e}}(i) = \frac{\tilde{N}_i - p_{\mathrm{a}}(i)\tilde{D}_i}{\tilde{D}_i+1}$$

$$\text{and} \qquad p_{\mathrm{d}}(i) = \frac{\tilde{N}_i\left(2 + \tilde{D}_i(1+p_{\mathrm{a}}(i)) - 2\tilde{N}_i - p_{\mathrm{a}}(i)\right) + p_{\mathrm{a}}(i)}{2\tilde{N}_i(\tilde{D}_i+1)}.$$

Moreover, if $M = 1$, we may choose $\tilde{N}$, p_{a} and p_{d} as the three independent parameters, so that (3.46) is replaced by

$$\tilde{D} = \frac{2\tilde{N}-1}{1+p_{\mathrm{a}}-2p_{\mathrm{d}}}\left(1+\frac{p_{\mathrm{a}}}{\tilde{N}}\right) - 1.$$

3.5 Discussion

In this Chapter, the importance of having a sequence of compartments in maintaining a population of product cells is explored. Using theoretical arguments, we show why a sequence of multiple compartments is often observed in biology. In

our model, individual cells in a compartment may die or divide (via self-renewal or asymmetric division), or transition to the next compartment, meaning that they differentiate and change phenotype. In our stochastic approach, all population properties are deduced from a complete understanding of the possible progeny of a single progenitor. We assume that each cell in a given compartment, independently, chooses its fate according to compartment-specific division, transition and death probabilities. We do not consider inter-event times, but each realisation is a sequence of events that ultimately results in extinction of cells in the intermediate compartments, with only product cells surviving. We let $Q_k(M)$ be the probability that the number of product cells, descended from a single progenitor via a sequence of M compartments, is equal to k. Using probability generating functions, we study the distribution $Q_k(M)$ as well as its behaviour for large values of k. An interesting feature we found is the universality of the large-k behaviour of $Q_k(M)$, for M compartments and $p_{\mathrm{a}}(i) = 0$ or $p_{\mathrm{a}}(i) > 0$,

$$Q_k(M) \sim \gamma_M^k k^{-3/2}, \quad k \to +\infty\,,$$

where γ_M has to be determined by locating the square root singularity of the probability generating function $\Phi_M(z)$.

Moreover, we relate the events probabilities with two important quantities N and D that can be often measured experimentally, via flow cytometry and CellTrace Violet or CFSE fluorescent dyes (Lemieszek *et al.*, 2022; Lyons, 2000). In case of direct differentiation and only one compartment, these two experimentally measurable quantities have both a long-tailed distribution. The analysis outlined in this Chapter shows how a sequence of compartments allows to achieve the amplification factor N of progenitor to product cells required in a specific biological framework while avoiding excessive clonality and minimising the average number of divisions D.

The analysis performed in this Chapter, has been carried out for a simpler case at first. In particular, we consider direct differentiation, that is $M = 1$, and only self-renewal division, that is $p_{\mathrm{a}} = 0$. Later we extend the results for the random variable $\mathbf{R}$ and $\mathbf{G}$ to consider a sequence of compartments as well as to include the asymmetric division event, $p_{\mathrm{a}}(i) > 0$. Additional details have been provided in the appendices: the recurrence relations to obtain the probability that k cells exit

from one or two compartments are given in Appendix A; calculations (Flajolet & Sedgewick, 2009) to obtain the asymptotic behaviour of distribution $Q_k(M)$ for large values of k are given in Appendix B.

The asymmetric division event introduces complexity in the mathematical model as division and change of cell type are happening simultaneously in the defined Markov process. However, asymmetric division plays a role in many biological situations and thus, in Sections 3.2.7, 3.3.3, 3.4.4, we show how to accommodate our theoretical methods to asymmetric division and how these can be applied to a biological system in which asymmetric cell division may play a role.

Thymus generates cells of the immune system and has a great variety of spanning cells of different types. In particular, it is responsible for the maturation process of T cells. We refer the reader to Section 1 for more biological details about thymocytes development and CD4, CD8 T cells formation. Here, we apply our compartmental model to the negative selection process of T cells, extending the deterministic model developed by Pham *et al.* (2015) to study thymocytes development from the DN3a to SP transition. We set up a model of five compartments and consider SP cells as the product cells population; we compute the distributions of the two biologically significant random variables in our stochastic model: the number of product cells in a family founded by one progenitor and the generation number of a product SP cell. Within our compartmental framework, cells behave independently and choose their fate following compartment-specific events. In particular, Pham *et al.* (2015) suggested a possible role for the asymmetric division event during the β-selection step, that, in this case, is represented by C_1, C_2 and C_3 compartments, namely DN3a-pre and DN3a-post. It has been experimentally observed that cells have already divided at least once when they undergo β-selection; thus, we assume a first compartment of DN3a cells that can only die or divide asymmetrically, namely DN3a-pre (Pham *et al.*, 2015). In particular, DN3a-pre cells have a higher probability to die than divide asymmetrically; thus, given (3.1) and $p_{\mathrm{a}}(1) + p_{\mathrm{d}}(1) = 1$, we have

$$p_{\mathrm{b}}(1) = p_{\mathrm{e}}(1) = 0, \quad p_{\mathrm{a}}(1) < 1/2 \quad \text{and} \quad p_{\mathrm{a}}(1) < p_{\mathrm{d}}(1)\,.$$

3. WHY ARE CELL POPULATIONS MAINTAINED VIA MULTIPLE INTERMEDIATE COMPARTMENTS?

Cells of all the subsequent populations can, independently, die, divide (via self-renewal) or differentiate to the next thymocytes development stage. The final fate of all cells is either death or SP stage. We define probabilities of death, division and differentiation, $p_{\mathrm{b}}(i)$, $p_{\mathrm{d}}(i)$, $p_{\mathrm{e}}(i)$ for $i = 2, 3, 4, 5$ such that $p_{\mathrm{b}}(i) + p_{\mathrm{d}}(i) + p_{\mathrm{e}}(i) = 1$. In addition, Pham *et al.* (2015) suggest that during the β-selection cells are more likely to die, that is $p_{\mathrm{d}}(i) > p_{\mathrm{b}}(i)$ for $i = 1, 2, 3$; whereas, cells in the following development stages are more likely to divide, that is $p_{\mathrm{b}}(i) > p_{\mathrm{d}}(i)$ for $i = 4, 5$. Here, we study the distribution of $\mathbf{R}$ and $\mathbf{G}$ in two different scenarios defined by different values of the asymmetric division probability, $p_{\mathrm{a}}(1) = 0.1$ and $p_{\mathrm{a}}(1) = 0.45 < 1/2$. In particular, we consider parameter values as in Table 3.1 and show how the amplification factor as well as the average generation of product cells are affected. We refer the reader to Section 3.2 for the analytical details about the methods applied to $\mathbf{R}$. Also, DP cells classification by their generation is done following Section 3.3.

Parameter	DN3a-pre	DN3a-post	DN3b	DN4	DP
$p_{\mathrm{b}}(i)$	0	0.25	0.25	0.45	0.45
$p_{\mathrm{e}}(i)$	0	0.3	0.3	0.3	0.3
$p_{\mathrm{d}}(i)$	0.55 or 0.9	0.45	0.45	0.25	0.25
$p_{\mathrm{a}}(i)$	0.45 or 0.1	0	0	0	0
N_i	9/11 or 1/9	0.6	0.6	3	3
D_i	20/11 or 10/9	1	1	9	9

Table 3.1: Parameter values for the five-compartment model of thymocyte development. We consider two possible scenarios to show how asymmetric division during β-selection step, can influence subsequent sub-populations of thymocytes development (Pham *et al.*, 2015). In particular, we consider $p_{\mathrm{a}}(1) = 0.45$ and $p_{\mathrm{a}}(1) = 0.1$. Values of N_i and D_i are computed using (3.20), (3.29) and (3.39), (3.44).

If $p_{\mathrm{a}}(1) = 0.1$, the resulting SP product population has a smaller average family size per single progenitor DN3a-pre cell, compared to the case of $p_{\mathrm{a}}(1) = 0.45$. This difference results from the higher or lower death probability of the first compartment, where the difference in the asymmetric events probabilities is

introduced. On the contrary, when computing the distribution of the generation number $\mathbf{G}$ for a product SP cell, we see a small difference between the two scenarios $p_a(1) = 0.1$ and $p_a(1) = 0.45$. This can be explained as, in both cases, the majority of cells that make the transition from DN3a-pre to DN3a-post do so in the first generation. In both cases, cells reach the product SP stage after few division and an older SP cell with $\mathbf{G} > 100$ is very rare. In particular, with parameter values as in Table 3.1 we get

$$\begin{aligned} N = 0.36 \quad \text{and} \quad D = 21.1, \quad \text{if} \quad p_a(1) = 0.1\,, \\ N = 2.65 \quad \text{and} \quad D = 21.9, \quad \text{if} \quad p_a(1) = 0.45\,. \end{aligned}$$

Figure 3.22 shows the distribution of random variable $\mathbf{R}$ and $\mathbf{G}$ for scenario $p_a(1) = 0.1$ (in blue) and $p_a(1) = 0.45$ (in green), respectively.

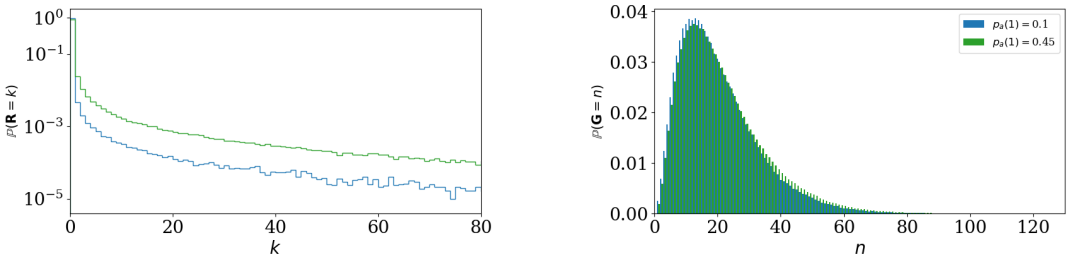


Figure 3.22: Numerical results for two cases of the five-compartment mathematical model for T cell development in the thymus from the DN3a to the SP stage (Pham *et al.*, 2015). The histograms show the distributions of family sizes (on the left) and of cell generation number (on the right) in the population of product SP cells. The difference between the two cases is the first compartment, where only death and asymmetric division have non-zero probabilities. We set parameter values as in Table 3.1.

As mentioned, asymmetric cellular division is a subject of recent research. However, other events might be of biological interest. For example, it is possible to consider a fifth event, the symmetric division, where both daughter cells exit their mother's compartment at birth (Derényi & Szöllősi, 2017; Zhang *et al.*, 2015); or to incorporate "de-differentiation" of cells moving backward in the hierarchy, for biological processes where cells revert their phenotype to the previous

state (Zhang *et al.*, 2015; Zhou *et al.*, 2019). Both symmetric division and de-differentiation events open the door to a wide range of possible mathematical models and discussion; in this PhD thesis, we partially tackle them in Chapter 5 where a general stochastic and time-dependent compartmental model is defined.

A limitation of the analysis outlined in this Chapter is the fact that the population of "product" cells is analysed at the end of the stochastic realisation when cells from intermediate compartments have died out or differentiated to subsequent compartments. Information about the time-dependent distribution of the number of exiting cells and of the generation of cells over time is lacking. Compartmental stochastic models with time-dependent random variables are defined in Chapters 4 and 5 of this thesis. In particular, Chapter 4 reports a generalisation of the time-independent analysis performed in this Chapter; however, the increased in the complexity as well as the timeline of this thesis result in a restricted study which would require further analysis and development of *ad hoc* mathematical methods.

Chapter 4

The importance of time in cell differentiation process

In Chapter 3, we consider the compartmental model depicted in Figure 3.2 and analyse the system in discrete-time; that is we ignore inter-event times as well as the total time taken for cells to proceed through the sequence of intermediate compartments leaving it either by death event or differentiation to the product state (depicted in orange in Figure 3.2). Product cells exiting the last compartment C_M are thought of as an accumulative pool for modelling purposes and, thus, they cannot die, further divide or differentiate. We define the random variable $\mathbf{R}$ for the total number of product cells in one realisation and we analyse its mean $N = \mathbb{E}(\mathbf{R})$ as well as its variance and its distribution q_k; that is defined as, for $k = 0, 1, 2, \ldots$, $q_k = \mathbb{P}(\mathbf{R} = k)$ the probability there are k cells in the final state of product orange cells. Moreover, we divide the population of product cells into classes by their generation number. The random variable $\mathbf{G}$ represents the generation number of a randomly-selected product cell, and its mean value $D = \mathbb{E}(\mathbf{G})$ is the mean age of the product cell population, measured in number of generations from the initial progenitor.

In this Chapter, instead, we focus on a time-dependent model. Indeed, differences in the times cells reach the final state of the sequence of compartments and differentiate to the product state, might be biologically relevant. Different types of cellular differentiation processes might require either a rapid response, such as during the generation of skin fibroblasts to repair damage resulting from

a cut or wound, or a long-term response. For example, the differentiation process of mesenchymal stem cells has been observed to be extremely slow, even if the physiological and morphologial reasons are not well understood (Kim *et al.*, 2014); also, hematopoietic stem cells (HSCs) are characterised by a slow cycle which is probably meant to reduce the metabolic stress and ensure a long-term maintenance of the subsequent differentiated stem cells (Tower, 2012).

Given a certain average number of product cells N that can be measured in an experiment, we wonder how this quantity influence the dynamics of cells through compartments over time as well as the number of cells reaching the product state at time t. The population of product cells is found to be heterogeneous with respect to both the time cells differentiate to the final state as well as the generation number cells are characterised with when reaching that state. Hence, in this Chapter we aim to investigate the time it takes for cells to differentiate to the product state and we define time-dependent random variables representing the number of cells in each compartment at time t. The time-dependent compartmental model, defined in Section 4.1, is a continuous-time Markov chain representing a sequence of M compartments. Cells in each compartment can, independently, divide, die or differentiate to the next compartment; cellular events are defined as per cell rates but, by definition of a continuous-time Markov process with exponential waiting times, can be clearly related to corresponding probabilities of death, division and differentiation. This same approach is also applied in Chapter 5, where we define a general stochastic model of a sequence of compartments and analyse the impacts of different cellular events. Section 4.2 describes the mean behaviour of the system of cells dividing, dying or exiting across a sequence of compartments over time, in terms of a continuous-time Markov chain. In Section 4.3, time-dependent probability generating functions for random variables $\mathbf{C}_i(t)$ are defined and the corresponding system of differential equation is explicitly solved for $M = 1$ and $M = 2$ (see Appendix C). Finally, in Section 4.4, our focus is a number of summary statistics related to the time cells exit the sequence of compartments. We are aware that the work outlined in this Chapter, is still preliminary and would require further analysis as mentioned in Section 4.5. An analytical expression for the case of $M > 2$ as well as a study about product cells generation number with respect to time are missing and left for future work.

4.1 Model definition

We propose a stochastic model, depicted in Figure 4.1 for cell differentiation across an ordered sequence of M compartments towards a product state. We assume cells of the same compartment behave independently of each other and are characterised by three possible fates: death, division via self-renewal or differentiation. Note that, differently than in Chapter 3, we restrict to the case with no asymmetric division event ($p_a = 0$). Moreover, as in Chapter 3, we assume cells exiting the last compartment C_M cannot die, further divide (via self-renewal) or differentiate. These product cells, depicted in orange in Figure 4.1, are thus an accumulative pool and may represent a specific cell type population which can be experimentally measured.

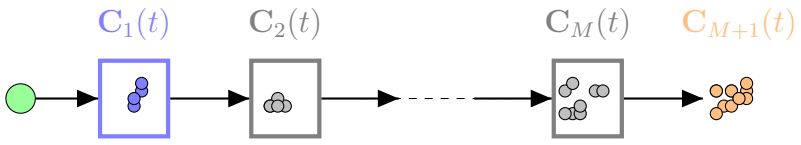


Figure 4.1: An initial cell, in green, enters the sequence of M compartments from progenitor cells (blue compartment C_1) to maturation state (in orange) via a sequence of intermediate compartments (in grey). The system is described by a continuous-time Markov chain (CTMC) $\{(\mathbf{C}_1(t), \mathbf{C}_2(t), \dots, \mathbf{C}_M(t), \mathbf{C}_{M+1}(t)) : t \geq 0\}$, with state space $\{0, 1, \dots\}^{M+1}$.

If in Chapter 3 we work in terms of a discrete-time Markov chain, we now describe the model in terms of a continuous-time Markov chain (CTMC) $\{(\mathbf{C}_1(t), \mathbf{C}_2(t), \dots, \mathbf{C}_M(t), \mathbf{C}_{M+1}(t)) : t \geq 0\}$, with state space $\{0, 1, \dots\}^{M+1} = \mathbb{N}_0^{M+1}$. The random variable $\mathbf{C}_i(t)$ represents the number of cells in compartment C_i at time t, for $i = 1, \dots, M$, and $\mathbf{C}_{M+1}(t)$ represents the number of product cells (cells exiting compartment C_M) at time t; note that since product cells can not die or differentiate further, $\mathbf{C}_{M+1}(t)$ represents the number of product cells accumulated until time t. We may want to remind the reader that, given a sequence of M compartments, the necessity of defining an additional random variable $\mathbf{C}_{M+1}(t)$ is arising from the time-dependent analysis that is carried out in this Chapter. With the aim of describing how cells differentiate to the product state over time, we find

convenient to define $\mathbf{C}_{M+1}(t)$ describing the number of product cells over time (cells exiting compartment C_M). Using notation of Chapter 3, one could denote the random variable $\mathbf{C}_{M+1}(t)$ as $\mathbf{R}(t)$, representing the cumulative quantity of product cells up to time t. In this case, we have

$$\lim_{t\to+\infty} \mathbf{R}(t) = \mathbf{R}\,.$$

We assume that cells behave independently and define cellular events of self-renewal, death and differentiation to the next compartment as transitions across states. In particular, for $i = 1, \ldots, M$, a cell in compartment C_i can divide via self-renewal at rate λ_i, die at rate μ_i and differentiate to compartment C_{i+1} at rate ν_i. That is, given state $(n_1, \ldots, n_M, n_{M+1}) \in \mathbb{N}_0^{M+1}$, possible cellular events are represented as follows:

(E1) Self-renewal (cellular division where both daughter cells remain in the same compartment as the mother) can occur in any compartment C_i, with per cell rate λ_i, for $i = 1, \ldots, M$,

$$(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_{M+1}) \xrightarrow{\lambda_i n_i} (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_{M+1}).$$

(E2) Differentiation to the subsequent compartment can occur with per cell rate ν_i, for $i = 1, \ldots, M$,

$$(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_{M+1}) \xrightarrow{\nu_i n_i} (n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1} + 1, \ldots, n_{M+1}).$$

(E3) Cells can die in any compartment C_i with per cell rate μ_i, for $i = 1, \ldots, M$,

$$(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_{M+1}) \xrightarrow{\mu_i n_i} (n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots, n_{M+1}).$$

In the following of this Chapter we will denote $k_i = \mu_i + \nu_i - \lambda_i$ and the sum of the rates as $S_i = \mu_i + \lambda_i + \nu_i$ for $i = 1, \ldots, M$. As in Chapter 3, we assume, that extinction is the ultimate fate of the population of cells in the intermediate compartments; that is, for $i = 1, \ldots, M$,

$$\mu_i + \nu_i > \lambda_i\,.$$

Thus, each realisation is a sequence of events that ultimately results in extinction of cells in the intermediate compartments, with only product cells surviving. The system defined by the CTMC reaches the absorbing state $(n_1, \ldots, n_M, n_{M+1}) = (0, \ldots, 0, n^\star_{M+1})$, with $n^\star_{M+1} \geq 0$. By definition, product cells are an accumulative pool that is, the random variable $\mathbf{C}_{M+1}(t)$ can only increase over time. Whereas, random variables $\mathbf{C}_i(t)$, for $i = 1, \ldots, M$, can either increase or decrease over time, as both division as well as death and differentiation events are modelled. However, for large time t, their values tend to zero as system approaches the absorbing state.

Finally, we would like to underline that cellular events in the time-independent model of Chapter 3 are defined in terms of event probabilities; whereas, the mathematical analysis carried out here as well as in Chapter 5 is performed in terms of cell rates. We might justify this since, when analysing the behaviour of the system for large values of t as in Chapter 3, differences in the parameter rates of cells in each of the M compartments of the sequence can be neglected. In this respect, one can define the compartmental model via probabilities with the aim of normalising compartments rates. Whereas, when including time in the compartmental analysis (as in this Chapter), the values of event rates might become relevant. Indeed, differences in the rates of cell death, division and differentiation identify slower and faster dynamics in a compartment. Of course, one could relate these three event rates to the compartment probabilities $p_{\mathrm{d}}(i)$, $p_{\mathrm{b}}(i)$ and $p_{\mathrm{e}}(i)$, respectively. This is natural since the defined model is a continuous-time birth-death-migration Markov process with exponential waiting times, thus

$$p_{\mathrm{d}}(i) = \frac{\mu_i}{\mu_i + \nu_i + \lambda_i}, \quad p_{\mathrm{b}}(i) = \frac{\lambda_i}{\mu_i + \nu_i + \lambda_i} \quad \text{and} \quad p_{\mathrm{e}}(i) = \frac{\nu_i}{\mu_i + \nu_i + \lambda_i}\,.$$

We define a general compartmental model and analyse the case of "direct differentiation" and one intermediate compartment, at first. Overall, the aim would be to extend the performed analysis to $M > 1$ compartments; however, due to the complexity of the mathematical analysis, the analytical work outlined in this Chapter is restricted to the case $M = 1, 2$. In some cases, numerical considerations are done for $M > 2$ but, in general, we leave for future work an analytical analysis of the case of a sequence of several intermediate compartments.

4.2 Time and mean populations' behaviour

We first describe the dynamics of the process and study the mean number of cells in each compartment, $\mathbb{E}(\mathbf{C}_i(t))$, which obeys the following system of differential equations

$$\begin{aligned}
\frac{d\,\mathbb{E}(\mathbf{C}_1(t))}{dt} &= -(\mu_1+\nu_1-\lambda_1)\mathbb{E}(\mathbf{C}_1(t))\,,\\
\frac{d\,\mathbb{E}(\mathbf{C}_i(t))}{dt} &= \nu_{i-1}\,\mathbb{E}(\mathbf{C}_{i-1}(t))\\
&\quad -(\mu_i+\nu_i-\lambda_i)\,\mathbb{E}(\mathbf{C}_i(t)), \quad i\in\{2,...,M\}\,,\\
\frac{d\,\mathbb{E}(\mathbf{C}_{M+1}(t))}{dt} &= \nu_M\,\mathbb{E}(\mathbf{C}_M(t))\,,
\end{aligned} \tag{4.1}$$

where $\mathbb{E}(\mathbf{C}_i(t))$ represents the expectation of the random variable $\mathbf{C}_i(t)$. We consider the initial value problem of System (4.1) with $\mathbb{E}(\mathbf{C}_1(0))=1$ and $\mathbb{E}(\mathbf{C}_i(0))=0$ for $i=2,\ldots,M+1$. Variables of this system of differential equations are time-dependent and the solution of the initial value problem $(\mathbb{E}(\mathbf{C}_1(t)),\ldots,\mathbb{E}(\mathbf{C}_{M+1}(t)))^T$ describes, in a quantitative way, the differentiation process of cells towards the product state over time. For a general solution of System (4.1) we refer the reader to Equation (5.5) where parameters for symmetric division event needs to be set to zero that is, $s_i=0$ for all $i=1,\ldots,M$.

Dynamics of the mean number of product cells are influenced by parameter rates of intermediate compartments. For example, Figure 4.2 reports cellular dynamics when $M=2$ for three different illustrative scenarios: one where compartments C_1 and C_2 have equally fast rates, and two scenarios where one compartment has faster or slower rates than the other one. Parameter rates are chosen so that, in all three scenarios, corresponding birth, death and differentiation probabilities in compartments are constant; in particular, $p_{\mathrm{b}}(i)=0.4$, $p_{\mathrm{e}}(i)=0.5$ and $p_{\mathrm{d}}(i)=0.1$, for $i=1,2$. We can observe that, for large values of t, the number of cells in a compartment is not affected by differences in the parameter rates. Since $p_{\mathrm{b}}(i)<p_{\mathrm{d}}(i)+p_{\mathrm{e}}(i)$, the number of cells in compartments C_1 and C_2 tends to zero and cells accumulate in the product state reaching the same steady state. This is in agreement with what outlined in Chapter 3: the mean number of product cells for large time t is determined by the so called *amplification factor* N. However, transient dynamics differ and product cells take a bit

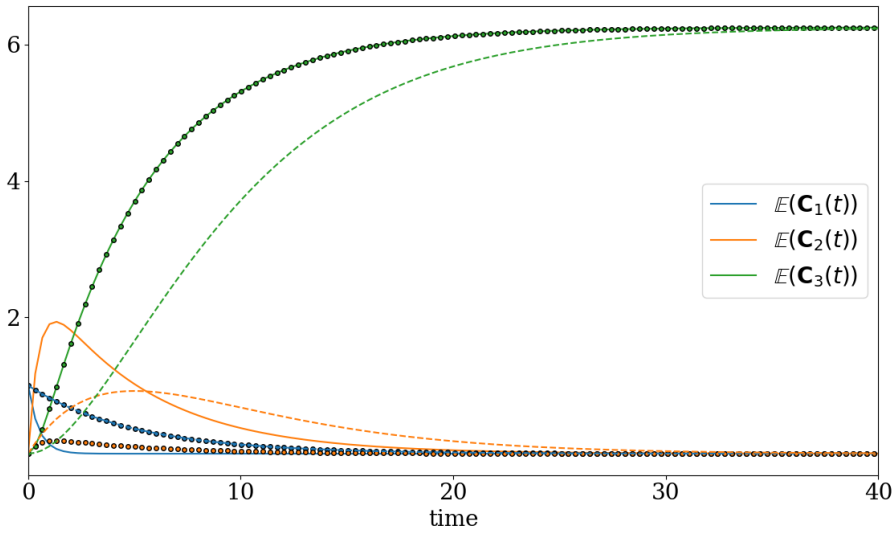


Figure 4.2: Plot of the solution for the mean number of cells if $M = 2$ with progenitor, intermediate and product compartment in blue, orange and green, respectively. We set three different illustrative scenarios: C_1 faster than C_2 (solid line), C_1 and C_2 equally fast (dots), and C_2 faster than C_1 (dashed line). The fastest compartment has parameters $\lambda_i = 4$, $\nu_i = 5$, $\mu_i = 1$, the slowest $\lambda_i = 0.4$, $\nu_i = 0.5$, $\mu_i = 0.1$, for $i = 1, 2$. In case of equal compartments we set parameter values $\lambda_i = 0.4$, $\nu_i = 0.5$, $\mu_i = 0.1$ for $i = 1, 2$. Note that for all the three cases, $N_1 = N_2$.

longer to accumulate if compartment C_2 has faster rates than the C_1. Moreover, cells tend to leave earlier the first compartment C_1 if that is the faster compartment. Also, in the latter case, the maximum number of cells in compartment C_2 is reached before extinction.

In addition, we can define the half-time t_{50}^M as the average time for half of the population of product cells to reach that state; that is

$$t_{50}^M = \left\{ t \,:\, \mathbb{E}(\mathbf{C}_{M+1}(t)) = \frac{1}{2} \lim_{t \to +\infty} \mathbb{E}(\mathbf{C}_{M+1}(t)) \right\},$$

where $\mathbb{E}(\mathbf{C}_{M+1}(t))$ is the solution of System (4.1) for the mean number of product cells. Figure 4.3 reports values of t_{50}^M numerically calculated for $M = 1, 2, 3$.

4. THE IMPORTANCE OF TIME IN CELL DIFFERENTIATION PROCESS

As detailed below, even if the same amplification is maintained for the three cases of a sequence of $M = 1, 2, 3$ compartments, dynamics and corresponding half-times t_{50}^M differ. Values of t_{50}^M can be computed as a solution of a transcendental equation derived from the expression of $\mathbb{E}(\mathbf{C}_{M+1}(t))$.

In the remaining of this Section, we report the calculations about cellular dynamics for $M = 1, 2, 3$. If $M = 1$, we let $\mathbf{C}_1(t)$ and $\mathbf{C}_2(t)$ be the random variables of the CTMC and $\mathbb{E}(\mathbf{C}_1(t))$ and $\mathbb{E}(\mathbf{C}_2(t))$ their expected values for cells in the progenitor and product state, respectively; System (4.1) is reduced to

$$\begin{aligned} \frac{d\mathbb{E}(\mathbf{C}_1(t))}{dt} &= -(\mu_1 + \nu_1 - \lambda_1)\mathbb{E}(\mathbf{C}_1(t))\,, \\ \frac{d\mathbb{E}(\mathbf{C}_2(t))}{dt} &= \nu_1 \mathbb{E}(\mathbf{C}_1(t))\,, \end{aligned} \tag{4.2}$$

with initial conditions $\mathbb{E}(\mathbf{C}_1(0)) = 1$, $\mathbb{E}(\mathbf{C}_2(0)) = 0$. Denoting $k_1 = \mu_1 + \nu_1 - \lambda_1 > 0$, the solution of System (4.2) reads

$$\mathbb{E}(\mathbf{C}_1(t)) = \mathrm{e}^{-k_1 t}\,, \quad \mathbb{E}(\mathbf{C}_2(t)) = \frac{\nu_1}{k_1}\left(1 - \mathrm{e}^{-k_1 t}\right)\,.$$

In this case, t_{50}^1 satisfies

$$\mathbb{E}(\mathbf{C}_2(t_{50}^1)) = \frac{1}{2}\frac{\nu_1}{k_1}\,,$$

that is

$$t_{50}^1 = \frac{1}{k_1}\log 2\,. \tag{4.3}$$

Similarly, if $M = 2$, the analogous of System (4.1) reads

$$\begin{aligned} \frac{d\mathbb{E}(\mathbf{C}_1(t))}{dt} &= -k_1 \mathbb{E}(\mathbf{C}_1(t))\,, \\ \frac{d\mathbb{E}(\mathbf{C}_2(t))}{dt} &= \nu_1 \mathbb{E}(\mathbf{C}_1(t)) - k_2 \mathbb{E}(\mathbf{C}_2(t))\,, \\ \frac{d\mathbb{E}(\mathbf{C}_3(t))}{dt} &= \nu_2 \mathbb{E}(\mathbf{C}_2(t))\,, \end{aligned}$$

with initial conditions $\mathbb{E}(\mathbf{C}_1(0)) = 1$, $\mathbb{E}(\mathbf{C}_2(0)) = \mathbb{E}(\mathbf{C}_3(0)) = 0$ and $k_i = \mu_i + \nu_i - \lambda_i$ for $i = 1, 2$. Then,

$$\mathbb{E}(\mathbf{C}_1(t)) = e^{-k_1 t}, \quad \mathbb{E}(\mathbf{C}_2(t)) = \frac{\nu_1}{k_2 - k_1}\left(e^{-k_1 t} - e^{-k_2 t}\right)\,,$$

$$\mathbb{E}(\mathbf{C}_3(t)) = \frac{\nu_1\nu_2}{k_1k_2}\left(1 + \frac{k_1 e^{-k_2 t}}{k_2 - k_1} - \frac{k_2 e^{-k_1 t}}{k_2 - k_1}\right).$$

One can clearly see that $\lim_{t\to+\infty} \mathbb{E}(\mathbf{C}_3(t)) = \frac{\nu_1\nu_2}{k_1k_2}$ and thus, the average half-time t^2_{50} satisfies the transcendental equation

$$\log\frac{k_2}{k_2 - k_1} = k_1\, t^2_{50} + \log\frac{1}{2}\left(1 + \frac{2k_1\, \mathrm{e}^{-k_2 t^2_{50}}}{k_2 - k_1}\right).$$

Similarly, if $M = 3$, we have

$$\mathbb{E}(\mathbf{C}_4(t)) = \frac{\nu_1\nu_2\nu_3}{k_1\, k_2\, k_3}\left(1 - \frac{k_2k_3\, \mathrm{e}^{-k_1 t}}{(k_2 - k_1)(k_3 - k_1)} - \frac{k_1k_2\, \mathrm{e}^{-k_3 t}}{(k_3 - k_1)(k_3 - k_2)} + \frac{k_1k_3\, \mathrm{e}^{-k_2 t}}{(k_2 - k_1)(k_3 - k_2)}\right)$$

and the average half-time t^3_{50} satisfies

$$\frac{1}{2} = \frac{k_2k_3\mathrm{e}^{-k_1 t^3_{50}}}{(k_2 - k_1)(k_3 - k_1)} + \frac{k_1k_2\, \mathrm{e}^{-k_3 t^3_{50}}}{(k_3 - k_1)(k_3 - k_2)} - \frac{k_1k_3\, \mathrm{e}^{-k_2 t^3_{50}}}{(k_2 - k_1)(k_3 - k_2)}.$$

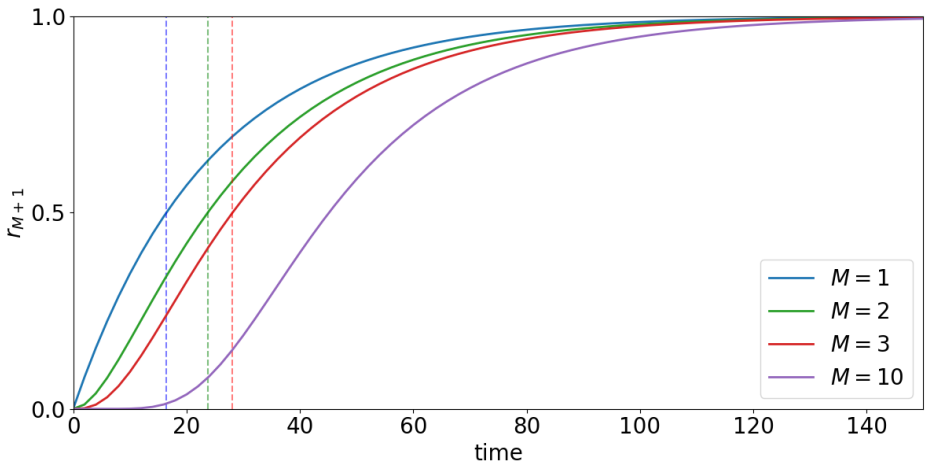


Figure 4.3: Normalised average population of product cells $r_{M+1} = \dfrac{\mathbb{E}(\mathbf{C}_{M+1}(t))}{\lim_{t\to+\infty}\mathbb{E}(\mathbf{C}_{M+1}(t))}$ for $M = 1, 2, 3, 10$ in blue, green, red and purple, respectively. We set $N = 10$, $S_i = 1$ and $p_{\mathrm{d}}(i) = \frac{\mu_i}{S_i} = 0.1$, for $i = 1, \ldots, M$. Dotted vertical lines show the half-time t^M_{50}.

The analytical t^1_{50} and the numerical solutions of the transcendental equations for the half-time for $M = 2$ and $M = 3$ are reported in Figure 4.3 (vertical dashed

lines). Moreover, we depict the dynamics of the population of product cells for $M = 1, 2, 3, 10$ in blue, green, red and purple, respectively. In particular, the ratio $r_{M+1} = \dfrac{\mathbb{E}(\mathbf{C}_{M+1}(t))}{\lim_{t\to+\infty}\mathbb{E}(\mathbf{C}_{M+1}(t))}$ is reported to quantify the accumulation of product cells over time. In the Figure, we set a fixed amplification factor $N = 100$, $S_i = 1$ and $p_{\mathrm{d}}(i) = \frac{\mu_i}{S_i} = 0.1$, for $i = 1, \dots, M$; then division and differentiation rates of intermediate compartments are deduced according to (3.45), so

$$\lambda_i = \frac{N^{1/M} - 1 + \mu_i}{2N^{1/M} - 1}$$

for $i = 1, \dots, M$; numerical values, rounded at the second decimal place, used in Figure 4.3 are reported in Table 4.1.

	λ_1	λ_2	λ_3	$\cdots$	λ_{10}
$M = 1$	0.48				
$M = 2$	0.42	0.42			
$M = 3$	0.38	0.38	0.38		
$M = 10$	0.24	0.24	0.24	$\cdots$	0.24

Table 4.1: Numerical value (rounded at the second decimal place) used in Figure 4.3. For all M, we set $S_i = 1$ and $\mu_i = 0.1$, for $i = 1, \dots, M$.

We can observe that, when setting the same amplification factor $N = \prod_{i=1}^{M} N_i$ for a sequence of $M = 1, 2, 3$ or 10 compartments, the time dynamics are adjusted accordingly, generating faster or slower dynamics. For this choice of parameter values we find $t_{50}^1 = 16$ for the case of direct differentiation; whereas, when including an intermediate compartment the time to reach half of the population of product cells increases, as $t_{50}^2 = 23$; and when $M = 3$ we find $t_{50}^3 = 28$. That is, when $M = 1$ more cells reach the product state at earlier times compared to the case of one, two or ten intermediate compartments; however, we can observe that the time by which all cells differentiate to the final compartment (and $N = 10$ is reached) does not differ significantly between the four cases.

4.3 Time-dependent probability generating function

The probability generating function of a random variable contains all the information about the associated random variable and can be used to find key properties of its distribution. In this Section, we will employ a probability generating function approach to analyse the population of product cells over time. We consider one initial progenitor cell that is, $\mathbf{C}_1(0) = 1$ and $\mathbf{C}_2(0) = \mathbf{C}_3(0) = \cdots = \mathbf{C}_{M+1}(0) = 0$, and we assume that for a small time step Δt only one or none event (death, division, differentiation) can happen for each compartment C_i, for $i = 1, \ldots, M$. Given that the defined model is a CTMC with multiple random variables, we define time-dependent joint probabilities

$$\begin{aligned} p_{n_i,\ldots,n_{M+1}}(t) = &\mathbb{P}\left(\mathbf{C}_i(t) = n_i, \mathbf{C}_{i+1}(t) = n_{i+1}, \ldots, \mathbf{C}_{M+1}(t) = n_{M+1} \,|\right. \\ &\left. \mathbf{C}_i(0) = 1, \mathbf{C}_{i+1}(0) = \cdots = \mathbf{C}_{M+1}(0) = 0\right), \end{aligned} \tag{4.4}$$

with $n_j \in \mathbb{N}^+$, for $j = i, \ldots, M+1$ and for $i = 1, \ldots, M+1$. For each i, the probability generating function can be defined

$$F_i(x_i, x_{i+1}, \ldots, x_M, x_{M+1}, t) = \sum_{n_i,\ldots,n_{M+1}=0}^{+\infty} p_{n_i,\ldots,n_{M+1}}(t)\, x_i^{n_i}\, x_{i+1}^{n_{i+1}} \ldots x_{M+1}^{n_{M+1}}.$$

Each function F_i contains all the joint probabilities $p_{n_i,\ldots,n_{M+1}}(t)$ and is a function of $M + 3 - i$ variables, where $|x_j| < 1$ for $j = i, \ldots, M+1$ and $t \in [0, +\infty)$. Probabilities $p_{n_i,\ldots,n_{M+1}}(t)$ can be deduced by solving the system of differential equations defined by the time derivatives of the probability generating functions. The following holds, for $i = 1, \ldots, M$,

$$\begin{aligned} \frac{\partial}{\partial t} F_i(x_i, \ldots, x_{M+1}, t) = \mu_i + \nu_i F_{i+1}(x_{i+1}, \ldots, x_{M+1}, t) - S_i F_i(x_i, \ldots, x_{M+1}, t) \\ + \lambda_i F_i^2(x_i, \ldots, x_{M+1}, t), \end{aligned} \tag{4.5}$$

where $S_i = \mu_i + \lambda_i + \nu_i$ and with initial conditions $F_i(x_i, \ldots, x_{M+1}, 0) = x_i$ for $i = 1, \ldots, M+1$. Moreover, without loss of generality, we can assume that for all times $t \in [0, +\infty)$, $F_{M+1}(x_{M+1}, t) = x_{M+1}$; this because, product cells exited compartment C_M can only accumulate without further dying, dividing or differentiating. Thus, $\frac{\partial}{\partial t} F_{M+1}(x_{M+1}, t) = 0$.

4. THE IMPORTANCE OF TIME IN CELL DIFFERENTIATION PROCESS

We may want to observe that, in spite of the dependency of the probability generating function F_i on $M+3-i$ variables, its analytical solution can be found by solving a system of partial differential equation where only the time derivative is considered. In general, however, solving System 4.5 is a challenging task since the right-hand-side of each differential equation has a time-dependent inhomogeneous term. Only two cases are different. Indeed, by model definition, the differential equation for $i = M+1$ is trivial; moreover, if $i = M$, the corresponding differential equation is simplified as the inhomogeneous term is constant over time. The resulting differential equation is then a standard Riccati equation. This is exploited in the cases of $M = 1$ and $M = 2$, as detailed below,

We start considering the case of direct differentiation ($M = 1$). We assume $(\mathbf{C}_1(0), \mathbf{C}_2(0)) = (1, 0)$ and that, for a small time step Δt, only one or none event (death, division, differentiation) can happen:

$$(\mathbf{C}_1(\Delta t), \mathbf{C}_2(\Delta t)) = \begin{cases} (1,0) & \text{with probability } 1 - \Delta t\,(\mu_1 + \lambda_1 + \nu_1) \\ (2,0) & \text{with probability } \Delta t\,\lambda_1 \\ (0,0) & \text{with probability } \Delta t\,\mu_1 \\ (0,1) & \text{with probability } \Delta t\,\nu_1 \end{cases} .$$

We define, $0 \le x_1, x_2 \le 1$, the time-dependent probability generating functions

$$F_1(x_1, x_2, t) = \sum_{n_1, n_2 = 0}^{+\infty} \mathbb{P}\left(\mathbf{C}_1(t) = n_1, \mathbf{C}_2(t) = n_2 \,|\, \mathbf{C}_1(0) = 1, \mathbf{C}_2(0) = 0\right) x_1^{n_1} x_2^{n_2} ,$$

$$F_2(x_2, t) = \sum_{n_2 = 0}^{+\infty} \mathbb{P}\left(\mathbf{C}_2(t) = n_2 \,|\, \mathbf{C}_2(0) = 1\right) x_2^{n_2} .$$

where $n_j \in \mathbb{N}^+$, for $j = 1, 2$. Thus, for a small time step Δt,

$$\begin{aligned} F_1(x_1, x_2, t + \Delta t) = F_1(x_1, x_2, t)(1 - \Delta t\,(\mu_1 + \lambda_1 + \nu_1)) + \mu_1 \Delta t + \nu_1 \Delta t\, x_2 \\ + \lambda_1 \Delta t\, F_1^2(x_1, x_2, t) . \end{aligned}$$

Here, we make use that $F_2(x_2, t) = x_2$ for all $t > 0$, as we consider product cells as an accumulating pool, i.e. product cells can not die, divide or differentiate further. Consequently, the probability generating function satisfies the partial differential equation

$$\frac{\partial}{\partial t} F_1(x_1, x_2, t) = \mu_1 + \nu_1 x_2 - (\mu_1 + \nu_1 + \lambda_1)\, F_1(x_1, x_2, t) + \lambda_1 F_1^2(x_1, x_2, t) . \quad (4.6)$$

with initial condition $F_1(x_1, x_2, 0) = x_1$ and $F_2(x_2, 0) = x_2$. For a solution of (4.6), we refer the reader to Section 4.3.1.

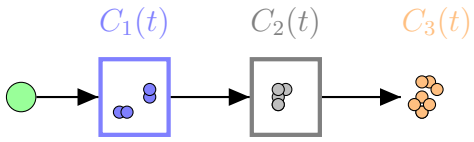


Figure 4.4: An initial cell, in green, enters the sequence of compartments from progenitor cells (blue compartment C_1) to maturation (in orange) via one intermediate compartment C_2.

Similarly, if $M = 2$ we consider a sequence of two compartments C_1 and C_2, depicted in Figure 4.4, and analyse the population of cells exiting compartment C_2. Random variables $\mathbf{C}_i(t)$, for $i = 1, 2, 3$, represent the number of cells in compartment C_1, C_2 and the number of product cells at time t, respectively. One can define the time-dependent joint probabilities $p_{n_1,n_2,n_3}(t)$, $p_{n_2,n_3}(t)$ and $p_{n_3}(t)$ as in (4.4), for $n_j \in \mathbb{N}^+$ for $j = 1, \ldots, 3$. Then, for $0 \leq x_1, x_2, x_3 \leq 1$, the time-dependent probability generating functions can be defined

$$
\begin{aligned}
F_1(x_1, x_2, x_3, t) &= \sum_{n_1,k_2,k_3=0}^{+\infty} p_{n_1,n_2,n_3}(t)\, x_1^{n_1} x_2^{n_2} x_3^{n_3}\,, \\
F_2(x_2, x_3, t) &= \sum_{n_2,n_3=0}^{+\infty} p_{n_2,n_3}(t)\, x_2^{n_2} x_3^{n_3}\,, \\
F_3(x_3, t) &= \sum_{k_3=0}^{+\infty} p_{n_3}(t)\, x_3^{n_3}\,.
\end{aligned}
$$

Since product cells exiting compartment C_2 can only accumulate over time $F_3(x_3, t) = x_3$ for all $t > 0$; then, for a small time step Δt, we have

$$
\begin{aligned}
F_1(x_1, x_2, x_3, t + \Delta t) &= F_1(x_1, x_2, x_3, t)(1 - \Delta t\, S_1) + \mu_1 \Delta t + \nu_1 \Delta t\, F_2(x_2, x_3, t) \\
&\quad + \lambda_1 \Delta t\, F_1^2(x_1, x_2, x_3, t)\,, \\
F_2(x_2, x_3, t + \Delta t) &= F_2(x_2, x_3, t)(1 - \Delta t\, S_2) + \mu_2 \Delta t + \nu_2 \Delta t\, x_3 \\
&\quad + \lambda_2 \Delta t\, F_2^2(x_2, x_3, t)\,,
\end{aligned}
$$

where $S_i = \mu_i + \lambda_i + \nu_i$ for $i = 1, 2$. Thus, the following system of differential equations is satisfied

$$\begin{aligned} \frac{\partial F_1(x_1, x_2, x_3, t)}{\partial t} &= \mu_1 + \nu_1 F_2(x_2, x_3, t) - S_1 F_1(x_1, x_2, x_3, t) + \lambda_1 F_1^2(x_1, x_2, x_3, t)\,, \\ \frac{\partial F_2(x_2, x_3, t)}{\partial t} &= \mu_2 + \nu_2\, x_3 - S_2 F_2(x_2, x_3, t) + \lambda_2 F_2^2(x_2, x_3, t)\,, \end{aligned} \tag{4.7}$$

with initial conditions $F_1(x_1, x_2, x_3, 0) = x_1$, $F_2(x_2, x_3, 0) = x_2$, $F_3(x_3, 0) = x_3$,

In the following two Sections, we consider the case of $M = 1$ and $M = 2$ and we show how to compute the solution of Equation (4.6) and System (4.7), respectively.

4.3.1 Direct differentiation, $M = 1$

In this Section, we assume $M = 1$ and compute the solution $F_1(x_1, x_2, t)$ of Equation (4.6). When looking at the case of direct differentiation, for the sake of notation, we will omit the compartment-index for events' rates λ_1, μ_1, ν_1 and, thus, denote $\lambda_1 = \lambda$, $\mu_1 = \mu$, $\nu_1 = \nu$ and $S = \mu + \lambda + \nu$. By the method of separation of variables, the solution of System (4.6) is

$$F_1(x_1, x_2, t) = \frac{ab\left(\mathrm{e}^{-\lambda(b-a)t} - 1\right) + x_1\left(a - b\,\mathrm{e}^{-\lambda(b-a)t}\right)}{a\,\mathrm{e}^{-\lambda(b-a)t} - b + x_1\left(1 - \mathrm{e}^{-\lambda(b-a)t}\right)}\,, \tag{4.8}$$

where $a < b$ are the two roots of $r(F_1) = \lambda F_1^2 - S F_1 + \mu + \nu\, x_2$,

$$a, b = \frac{S}{2\lambda}\left(1 \pm \sqrt{S^2 - 4\lambda(\mu + \nu\, x_2)}\right)\,.$$

Given (4.8), one can look at the long-term behaviour of the probability generating function

$$\lim_{t \to +\infty} F_1(x_1, x_2, t) = \frac{-ab + x_1 a}{-b + x_1} = a\,.$$

Here, we would like to observe that the limit for $t \to +\infty$ of $F_1(x_1, x_2, t)$ is finite and equals the probability generating function in Equation (3.5) from the time-independent analysis of Chapter 3. This can be explained because for large time t the system tends to reach the absorbing state $(0, n_2^\star)$, with $n_2^\star \geq 0$; that is cells leave intermediate compartments and either die or accumulate to the product

state. Thus, for this case of $M = 1$, $\mathbf{C}_2(t \to +\infty)$ corresponds to the random variable $\mathbf{R}$ that we defined and analysed in Chapter 3.

By definition, the probability generating function is a power series representation of the probability density function of the respective random variable $\mathbf{C}_1(t)$. Thus, one can compute the probability $p_n(t)$ of having n cells in the first compartment as

$$p_n(t) = \mathbb{P}\left(\mathbf{C}_1(t) = n\right) = \frac{1}{n!}\frac{\partial^{(n)}}{\partial x_1^{(n)}} F_1(x_1, 1, t)\Big|_{x_1=0}.$$

We see how the probability density $p_n(t)$ depends on the marginal distribution $F_1(x_1, 1, t)$. Setting $x_2 = 1$ in Equation (4.8), we find

$$F_1(x_1, 1, t) = \frac{x_1\left(\lambda - (\mu+\nu)\mathrm{e}^{-kt}\right) - (\mu+\nu)\left(1-\mathrm{e}^{-kt}\right)}{x_1\left(1-\mathrm{e}^{-kt}\right)\lambda - (\mu+\nu-\lambda\mathrm{e}^{-kt})}, \tag{4.9}$$

where we denote $k = \mu + \nu - \lambda$. Now we observe that $F_1(x_1, 1, t)$ is a rational function of the form $(c_1(t)\,x_1 + c_2(t))/(c_3(t)\,x_1 + c_4(t))$ where $c_i(t)$, for $i = 1, 2, 3, 4$, are exponential functions in terms of the time variable t. In particular,

$$\begin{aligned} c_1(t) &= \lambda - (\mu+\nu)\mathrm{e}^{-kt}, \quad & c_2(t) &= -(\mu+\nu)\left(1-\mathrm{e}^{-kt}\right), \\ c_3(t) &= \lambda\left(1-\mathrm{e}^{-kt}\right), \quad & c_4(t) &= -(\mu+\nu-\lambda\mathrm{e}^{-kt}). \end{aligned}$$

Clearly, $p_0(t) = \mathbb{P}\left(\mathbf{C}_1(t) = 0\right) = F_1(0, 1, t)$. Differentiating once Equation (4.9) we get

$$\frac{\partial}{\partial x_1} F_1(x_1, 1, t) = \frac{c_1(t)c_4(t) - c_2(t)c_3(t)}{(c_3(t)x_1 + c_4(t))^2},$$

and thus,

$$p_1(t) = \frac{(\lambda - (\mu+\nu))^2\,\mathrm{e}^{-kt}}{(\mu+\nu-\lambda\,\mathrm{e}^{-kt})^2}.$$

In general, we can express the nth derivative as

$$\frac{\partial^{(n)}}{\partial x_1^{(n)}} F_1(x_1, 1, t) = \frac{(-1)^{n-1}\,n!\,c_3(t)^{n-1}(c_1(t)c_4(t) - c_2(t)c_3(t))}{(c_3(t)x_1 + c_4(t))^{n+1}}.$$

The probability of having n cells in compartment C_1 at time t can then be easily derived as

$$p_n(t) = \frac{(-1)^{n-1}\,c_3(t)^{n-1}(c_1(t)c_4(t) - c_2(t)c_3(t))}{(c_4(t))^{n+1}}. \tag{4.10}$$

For the defined stochastic process $\{(\mathbf{C}_1(t), \mathbf{C}_2(t)) : t \geq 0\}$, states $\{(0, n_2^\star)\}$, with $n_2^\star \geq 0$, are absorbing states. Cells indeed tend to die or exit compartment C_1 and accumulate as product cells. Thus, extinction of cells in C_1 is certain that is, $\lim_{t\to+\infty} p_0(t) = 1$. However, the expected time until absorption might be large and one could examine the dynamics of the process prior this state is reached. One can wonder what is the size of the non-extinct population at a large time t. From a mathematical perspective it is, thus, interesting to compute the quantity $(p_n(t))/(1-p_0(t))$, for $n = 1, 2, \ldots$, that is the conditional probability of n cells being in compartment C_1 at time t, given that absorption has not occurred up until then. In particular, a study of the asymptotic behaviour of this quantity when $t \to +\infty$ can be done. The defined stochastic process $\{(\mathbf{C}_1(t), \mathbf{C}_2(t)) : t \geq 0\}$, can be seen as a birth-and-death process with death rate $\mu+\nu$ and birth rate λ; the stationary distribution of such a process is well studied by Allen (2010). Under our assumption, $\lambda < \mu + \nu$, it is known that the stationary probability distribution is given by a geometric distribution with parameter $\frac{\lambda}{\mu+\nu}$; this is, indeed, confirmed by our findings,

$$\begin{aligned}
\lim_{t\to+\infty} \frac{p_n(t)}{1-p_0(t)} &= \lim_{t\to+\infty} (-1)^{n-1} \left(\frac{c_3(t)}{c_4(t)}\right)^{n-1} \frac{c_1(t)c_4(t) - c_3(t)c_2(t)}{c_4(t)(c_4(t) - c_2(t))} \\
&= \lim_{t\to+\infty} (-1)^{n-1} \left(\frac{c_3(t)}{c_4(t)}\right)^{n-1} \frac{p_1(t)}{1-p_0(t)} \\
&= \lim_{t\to+\infty} (-1)^{n-1} \left(\frac{\lambda(1-\mathrm{e}^{-kt})}{-(\mu+\nu-\lambda\,\mathrm{e}^{-kt})}\right)^{n-1} \frac{k}{\mu+\nu-\lambda\,\mathrm{e}^{-kt}} \\
&= \left(\frac{\lambda}{\mu+\nu}\right)^{n-1} \frac{k}{\mu+\nu}
\end{aligned}$$

4.3.2 One intermediate compartment, $M = 2$

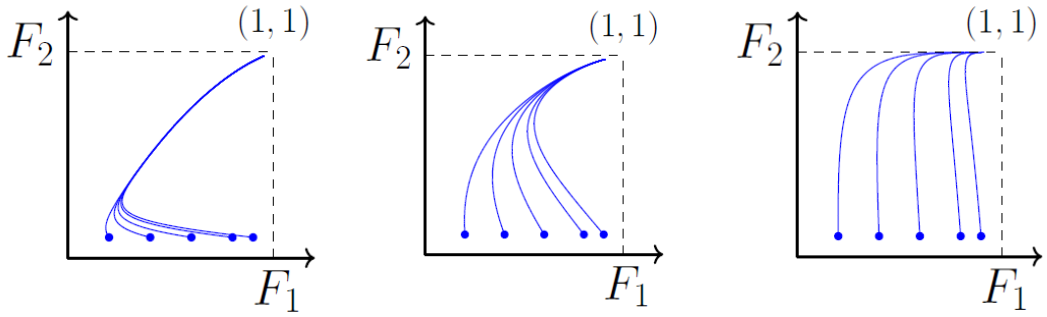
In this Section, we consider the case of one intermediate compartment ($M = 2$) and we will see how, including just one intermediate compartment in the differentiation chain, the mathematical description of the time-dependent process become more complex. The probability generating functions $F_1(x_1, x_2, x_3, t)$ and $F_2(x_2, x_3, t)$ satisfy System (4.7), and we report in Figure 4.5 the discretised solution for three different illustrative scenarios and with different initial conditions

(x_1, x_2) for $x_1 \in \{0.2, 0.4, 0.6, 0.8, 0.9\}$ and $x_2 = 0.1$, and for $x_3 = 1$. In particular, similar to Figure 4.2, we consider from left to right the case where compartment C_1 has faster rates that compartment C_2, the two compartments C_1 and C_2 have equally fast rates, and the case where compartment C_1 has slower rates than the last one. Since boundary conditions of probability generating functions need to be satisfied, we know that $F_1(1, 1, 1, t) = F_2(1, 1, t) = 1$ and thus state $(1, 1)$ is a steady state. Moreover, we can observe in Figure 4.5 that state $(1, 1)$ is always reached, regardless the starting point (x_1, x_2), represented by the blue dots. Thus, state $(1, 1)$ is an asymptotically stable steady state. Population of cells in intermediate compartments C_1 and C_2 tends to the extinction state but the transient dynamics vary according to compartments' events rates: on the left, C_1 has faster rates than C_2 and population of cells quickly leave the first compartment; in this case, regardless the initial condition x_1, the dynamics are mainly driven by the second compartment and reach a stable manifold. On the contrary, if C_2 has faster rates than C_1, the first compartment slowly contribute to the population of product cells.

We now focus on solving System (4.7) in the general case, for x_1, x_2 and x_3 in $[0, 1]$. Since the considered System is defined in terms of derivatives with respect to time, we will omit, for simplicity of notation, dependencies of variables x_1, x_2, x_3 and, thus, denote the two probability generating functions as $F_1(t)$ and $F_2(t)$ in the remaining of this Section. System (4.7) consists of two Riccati equations, but with different characteristics. The function $F_2(t)$ satisfies a differential equation with constant coefficients of the type of Equation (4.6); its solution is given by

$$F_2(t) = \frac{a_2(x_2 - b_2) + b_2(a_2 - x_2)\mathrm{e}^{-\lambda_2(b_2-a_2)t}}{x_2 - b_2 + (a_2 - x_2)\mathrm{e}^{-\lambda_2(b_2-a_2)t}} = \frac{d_1 + d_2\,\mathrm{e}^{-\gamma_2 t}}{d_3 + d_4\,\mathrm{e}^{-\gamma_2 t}}\,, \tag{4.11}$$

with $a_2 < b_2$ and $a_2, b_2 = (S_2 \pm \sqrt{S_2^2 - 4\lambda_2(\mu_2 + \nu_2\, x_3)})/2\,\lambda_2$. In the second step, we short the notation and denote $\gamma_2 = \lambda_2(b_2 - a_2)$ and $d_1 = a_2(x_2 - b_2)$, $d_2 = b_2(a_2 - x_2)$, $d_3 = x_2 - b_2$ and $d_4 = a_2 - x_2$. If the differential equation for $F_2(t)$ has a known solution, the differential equation for $F_1(t)$ has time-dependent coefficients and specific solving methods are required. Indeed, considering System (4.7) and



(a) C_1 has faster rates than C_2.

(b) C_1, C_2 have equally fast rates.

(c) C_2 has faster rates than C_1.

Figure 4.5: Transient dynamics resulting from System (4.7) with different initial conditions (x_1, x_2), shown as blue dots, and $x_3 = 1$. In particular, we set $x_1 \in \{0.2, 0.4, 0.6, 0.8, 0.9\}$ and $x_2 = 0.1$. In all cases the solution reaches state $(1, 1)$, since $F_1(1, 1, 1, t) = F_2(1, 1, t) = 1$. The fastest compartment has parameter values $\lambda_i = 4$, $\nu_i = 5$, $\mu_i = 1$, the slowest $\lambda_i = 0.4$, $\nu_i = 0.5$, $\mu_i = 0.1$, for $i = 1, 2$ as in subplots (a) and (c). In subplot (b) we consider equal compartments with parameter values $\lambda_i = 0.4$, $\nu_i = 0.5$, $\mu_i = 0.1$ for $i = 1, 2$.

substituting Equation (4.11), one needs to solve a first order differential equation with non-constant coefficients,

$$\begin{aligned}\frac{\partial F_1(t)}{\partial t} &= \mu_1 + \nu_1 F_2(t) - S_1 F_1(t) + \lambda_1 F_1^2(t) \\ &= \mu_1 + \nu_1 \frac{d_1 + d_2\, \mathrm{e}^{-\gamma_2 t}}{d_3 + d_4\, \mathrm{e}^{-\gamma_2 t}} - S_1 F_1(t) + \lambda_1 F_1^2(t)\,. \end{aligned} \tag{4.12}$$

The quadratic and first order coefficients, λ_1 and S_1 respectively, are constant; whereas, the inhomogeneous term, $\mu_1 + \nu_1 F_2(t)$, is time-dependent. A change of variables may transform the first order Riccati equation into a linear second order equation with non-constant coefficients for which more solving methods are available in the literature (Pala & Ertas, 2017; Sugai, 1960). Following the usual substitution method for Riccati equations, we introduce a new function $f_1(t)$, such that $F_1(t) = -\frac{f_1'(t)}{\lambda_1 f_1(t)}$. Thus, $f_1(0) = 1$, $f_1'(0) = -\lambda_1 x_1$ and

$$f_1''(t) + S_1 f_1'(t) + \lambda_1 \left(\mu_1 + \nu_1 F_2(t)\right) f_1(t) = 0\,, \tag{4.13}$$

where we denoted with $f_1'(t)$ and $f_1''(t)$ the first and second derivative of $f_1(t)$ with respect to the independent variable t.

Equation (4.13) is a differential equation with an exponential coefficient, which is a function of the independent variable; thus, the change of variable $\tau = \mathrm{e}^{-\frac{\gamma_2}{2}t}$ is required before applying the Frobenius' method to solve the equation (Boyce & di Prima, 2009). As detailed in Section 2.5, this method exploits the idea that the solution of the given differential equation has a power series expansion whose coefficients need to be determined. In Appendix C.1, we show that the solution of Equation (4.13), and consequently of Equation (4.12), is a linear combination of Gaussian hypergeometric functions with different parameter values; moreover, the mathematical steps to obtain such hypergeometric solution $F_1(t)$ are reported.

In addition, given the complexity of the solution of Equation (4.13), we report in Appendix C.2 the calculations required to solve the equation for a simpler case. There, we assume the solution of $F_2(t)$ is a simple exponential with parameter γ (instead of (4.11)); moreover, we let $\mu_1 = 0$ and $S_1 = \lambda_1 + \nu_1 = 1$. Under this restrictions, the resulting solution is a linear combination of Bessel functions of first order. Since this case is more tractable by hand, the solution is explicitly reported in Appendix C.2.

As done above for the case $M = 1$, one can compute the long-term behaviour of the probability generating function $F_1(t)$ that is, the limit for $t \to +\infty$. Given Equation (4.11), it holds

$$\lim_{t\to+\infty} F_2(x_2, x_3, t) = F_2^\star = \frac{S_2 - \sqrt{S_2^2 - 4\lambda_2\mu_2 - 4\lambda_2\nu_2\, x_3}}{2\lambda_2},$$

and we can observe that $F_2^\star = a_2$, where a_2 is the smaller root of $r(F_2)$. Consequently, for $F_1(t)$ we have

$$\lim_{t\to+\infty} F_1(x_1, x_2, x_3, t) = F_1^\star = \frac{S_1 - \sqrt{S_1^2 - 4\lambda_1\mu_1 - 4\lambda_1\nu_1 F_2^\star}}{2\lambda_1}.$$

Again, we would like to observe that the limit for $t \to +\infty$ of $F_2(x_2, x_3, t)$ is finite and equals the probability generating function in Equation (3.5) from the time-independent analysis in Chapter 3. Also, the limit of function $F_1(x_1, x_2, x_3, t)$ for large values of t depends on the limiting value of $F_2(x_2, x_3, t)$. This, can be generalised to a sequence of M compartments as $\lim_{t\to+\infty} F_1(x_1, \ldots, x_{M+1}, t)$

corresponds to $\Phi_M(z)$ as defined in Equation (3.21) during the time-independent analysis. The described behaviour can be observed also from Figure 4.6 which has been done by setting the right-hand-side of the partial differential equation for $F_1(x_1, x_2, x_3, t)$ to zero (see Equation 4.7); so that the steady state of the system could be found.

For a birth-and-death process with parameters λ and μ respectively, it is well known that the probability generating function tends to 1 if $\lambda < \mu$ (extinction is an absorbing state) and to $\frac{\mu}{\lambda}$ if $\lambda > \mu$ (the system reach extinction with nonzero probability) (Allen, 2010). In our case, the stochastic process in each compartment is a birth-and-death process with migration and immigration. When looking at the case $M = 2$, we have $r(F_1) = \lambda_1 F_1^2 - S_1 F_1 + \mu_1 + \nu_1 F_2$ and graphing the quadratic function on the right-hand-side can provide information about the long-term behaviour $\lim_{t\to+\infty} F_1(x_1, x_2, x_3, t)$. In particular, in Figure 4.6 we report $r(F_1)$ and consider two different sets of parameter values: $x_1 = 0.7, x_2 = 0.4, x_3 = 0.8$, on the left, and $x_1 = 0.4, x_2 = 0.7, x_3 = 0.1$, on the right plot. On the x-axis, the function F_1. Colours represent different times: to represent the limiting behaviour as $t \to +\infty$, we consider five different increasing times $t \in \{0, 1, 4, 10, 50\}$, and compute the function F_2 at those fixed times; the resulting $r(F_1)$ computed for $t \in \{0, 1, 4, 10, 50\}$ is reported in the plot. The blue vertical dashed line indicates the limit $\lim_{t\to+\infty} F_1(x_1, x_2, x_3, t)$. We can observe how, for $t \in \{0, 1, 4, 10, 50\}$ the function $r(F_1)$ is always a parabola with one root bigger than 1 and one smaller. The latter is the one we consider as within the boundary limits of a probability generating function. Despite the similar behaviour, given by the quadratic structure, it is clear how as time changes, solution F_2 is changing as well and function F_1 adjusts so that, as time increases, the smaller root of $r(F_1)$ always approaches the limiting value given by $\lim_{t\to+\infty} F_1(x_1, x_2, x_3, t)$. This is in agreement with the fact that we assumed $\lambda_i < \mu_i + \nu_i$, for $i = 1, 2$, so that extinction of intermediate compartments is certain. In addition, we can observe that, if $x_2 < x_3$ (as on the left plot) the root of $r(F_1)$ is increasing with time, approaching the limit from the left. Also, as $x_3 \to 1$, then the limiting value $F_1^\star$ tends to 1. Whereas, on the second case the smaller root of $r(F_1)$ decreases over time and the limit is reached from the right.

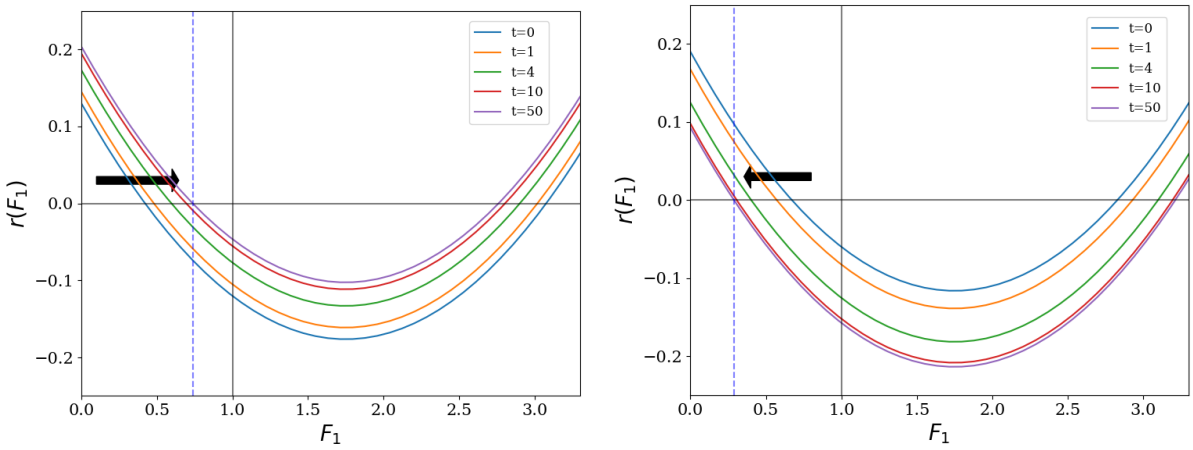


Figure 4.6: Plot of $r(F_1)$ for two different values of x_1, x_2, x_3 and five different time points $t \in \{0, 1, 4, 10, 50\}$. The blue vertical dashed line indicates the limit $\lim_{t\to+\infty} F_1(x_1, x_2, x_3, t)$. In particular, on the left we consider the case of $x_1 = 0.7, x_2 = 0.4, x_3 = 0.8$ $(x_3 > x_1 > x_2)$; on the right $x_1 = 0.4, x_2 = 0.7, x_3 = 0.1$ that is, $x_2 > x_3$ and $x_2 > x_1$.

4.4 First, last and exiting times

In Chapter 3, we define a time-independent model of M compartments with total amplification factor N discarding information about inter-event times as well as the time cells exit the last compartment C_M reaching the product state. Thus, questions such as when, starting from one progenitor cell, does the first or the last differentiation event towards the product state happens remain unanswered. Moreover, one can investigate the overall distribution of times cells exit compartment C_M differentiating to the product state. In Section 4.3, we focus on solving, analytically, the system of differential equations for probability generating functions for $M = 1, 2$. Here, we define some summary statistics of interest. An analytical solution is computed for the case $M = 1$ and numerical simulations are run for $M \geq 1$ to highlight the impact, with respect to time, of sequence of compartments on the cellular differentiation process.

4.4.1 When will the first product cell arrive?

In this Section, we focus on the first cellular differentiation event towards the product state. In particular, for a sequence of M compartments we define

$$T_p^M = \inf\{t \geq 0 : \mathbf{C}_{M+1}(t) \neq 0 \,|\, \mathbf{C}_1(0) = 1, \mathbf{C}_2(0) = \cdots = \mathbf{C}_{M+1}(0) = 0\}\,,$$

as the random variable representing the time to observe the first product cell that is, the first cell exiting compartment C_M, starting from one progenitor cell.

Consequently, one could analyse the probability distribution of T_p^M as well its expected value $\mathbb{E}(T_p^M)$. We start observing that, since death event is included in the stochastic model, single cells can die before reaching the product state; mathematically, we have for every $t \geq 0$,

$$\mathbb{P}\left(\mathbf{C}_1(t) = \cdots = \mathbf{C}_{M+1}(t) = 0 \,|\, \mathbf{C}_1(0) = 1, \mathbf{C}_2(0) = \cdots = \mathbf{C}_{M+1}(0) = 0\right) > 0\,,$$

and $\mathbb{E}(T_p^M) = +\infty$. Thus, we are interested in computing the restricted time

$$\mathbb{E}(T_p^M \, \mathbb{1}_{\{T_p^M < +\infty\}}) = \mathbb{E}(T_p^M \,|\, T_p^M < +\infty)\, \mathbb{P}\left(T_p^M < +\infty\right), \tag{4.14}$$

where $\mathbb{1}_{\{T_p^M < +\infty\}}$ is the indicator function defined as

$$\mathbb{1}_{\{T_p^M < +\infty\}} = \begin{cases} 1, & \text{if} \quad T_p^M < +\infty, \\ 0, & \text{if} \quad T_p^M = +\infty. \end{cases}$$

We start analysing the case of direct differentiation, $M = 1$. The random variable T_p^1 represents the time to observe a first cell exiting compartment C_1, starting from one cell in that compartment; it is defined as

$$T_p^1 = \inf\{t \geq 0 : \mathbf{C}_2(t) \neq 0 \,|\, \mathbf{C}_1(0) = 1, \mathbf{C}_2(0) = 0\}\,.$$

From Equation (4.8), we can set $x_1 = 1$, $x_2 = 0$ and find the marginal distribution $F_1(1, 0, t) = \mathbb{P}\left(\mathbf{C}_2(t) = 0\right)$. Given that $\mathbb{P}\left(\mathbf{C}_2(t) \neq 0\right) = 1 - \mathbb{P}\left(\mathbf{C}_2(t) = 0\right)$, we can differentiate $F_1(1, 0, t)$ with respect to time and find the probability density of the random variable T_p^1,

$$\frac{d}{dt}\mathbb{P}\left(T_p^1 < t\right) = \frac{\lambda(1-a)(1-b)(a-b)^2\, \mathrm{e}^{-\lambda(b-a)t}}{\left(1 - b + (a-1)\, \mathrm{e}^{-\lambda(b-a)t}\right)^2}\,,$$

where $a < b$ are the roots of $r(F_1) = \lambda F_1^2 - S F_1 + \mu$. This is the probability density function when $\mathbb{E}(T_p^1)$ is finite. By definition of the expected value of a random variable, the restricted mean time $\mathbb{E}(T_p^1 \, \mathbb{1}_{\{T_P^1 < +\infty\}})$ to get the first product cell can be then computed as

$$\begin{aligned}\mathbb{E}(T_p^1 \, \mathbb{1}_{\{T_p^1<+\infty\}}) &= \int_0^{+\infty} t \, \frac{\lambda(1-a)(1-b)(b-a)^2 \, \mathrm{e}^{-\lambda(b-a)t}}{\left(1-b+(a-1)\,\mathrm{e}^{-\lambda(b-a)t}\right)^2} dt \\ &= \lambda(1-a)(1-b)(b-a)^2 \int_0^{+\infty} \frac{t \, \mathrm{e}^{-\lambda(b-a)t}}{\left(1-b+(a-1)\mathrm{e}^{-\lambda(b-a)t}\right)^2} dt \, .\end{aligned}$$

Computing and evaluating the integral, we get

$$\mathbb{E}(T_p^1 \, \mathbb{1}_{\{T_p^1<+\infty\}}) = \frac{1}{\lambda} \log\left(\frac{b-a}{b-1}\right) .$$

We apply Equation (4.14) and find the conditioned mean time to observe the first product cell

$$\mathbb{E}(T_p^1 \,|\, T_p^1 < +\infty) = \frac{\mathbb{E}(T_p^1 \, \mathbb{1}_{\{T_p^1<+\infty\}})}{1-a} = \frac{1}{\lambda(1-a)} \log\left(\frac{b-a}{b-1}\right) .$$

Similarly, when $M = 2$ we let

$$T_p^2 = \inf\{t \geq 0 : \mathbf{C}_3(t) \neq 0 \,|\, \mathbf{C}_1(0) = 1, \mathbf{C}_2(0) = 0, \mathbf{C}_3(0) = 0\} \, ,$$

be the random variable representing the time to observe the first cell exiting compartment C_2 (first product cell), starting from one cell in compartment C_1. The probability distribution of T_p^2 is given by $-\frac{d}{dt}F_1(1,1,0,t)$. However, the complexity of computing a solution of System (4.7) reflects also on the case where $x_1 = x_2 = 1$ and $x_3 = 0$. Given the preliminary nature of this work, we leave as future work the calculation of an analytical expression of $F_1(1,1,0,t)$. Here, we restrict in considering the simplified case of $\lambda_1 = \lambda_2 = 0$ and $\mu = \mu_1 = \mu_2$, $\nu = \nu_1 = \nu_2$. Solving System (4.7), it is easy to find

$$\begin{aligned}F_2(1,0,t) &= \frac{\mu}{k} + \frac{\nu}{k}\mathrm{e}^{-kt} \\ F_1(1,1,0,t) &= 1 - \left(\frac{\nu}{k}\right)^2 + \left(\frac{\nu}{k}\right)\mathrm{e}^{-kt} + \frac{\nu^2}{k} t \, \mathrm{e}^{-kt} \, .\end{aligned} \tag{4.15}$$

Differentiating with respect to time we get

$$-\frac{d}{dt}F_1(1,1,0,t) = \nu^2 t \, \mathrm{e}^{-kt} \, .$$

Thus we can conclude that the probability distribution of the random variable $T_p^2 \, \mathbb{1}_{\{T_p^2<+\infty\}}$ follows a Gamma distribution. This can be explained as, in this case where division is absent, cells either die or move through the sequence of compartments via differentiation. We find,

$$T_p^2 \, \mathbb{1}_{\{T_p^2<+\infty\}} \sim \text{Gamma}\left(2, \frac{k^3}{\nu^2}\right) .$$

Moreover, the restricted mean time can be computed as

$$\mathbb{E}(T_p^2 \, \mathbb{1}_{\{T_p^2<+\infty\}}) = \int_0^{+\infty} t \left(-\frac{d}{dt} F_1(1,1,0,t)\right) dt$$

and

$$\mathbb{E}(T_p^2 \,|\, T_p^2 < +\infty) = \frac{\mathbb{E}(T_p^2 \, \mathbb{1}_{\{T_p^2<+\infty\}})}{1 - F_1^\star} .$$

These analytical considerations, can be supported by numerical results that could highlight interesting properties about the time cells differentiate through a sequence of M compartments. A total of 10^5 numerical simulations has been run to analyse the probability distribution of T_p^M, for $M = 1, 3, 10$. In particular, we report in blue, in Figure 4.7, the probability distribution for the case $M = 1$,

$$\mathbb{P}\left(T_p^1 = t\right), \quad t = 0, 1, 2, \ldots .$$

Figure 4.8 reports in blue the probability distribution $\mathbb{P}\left(T_p^M = t\right)$, $t = 0, 1, 2, \ldots$, for the case $M = 3$ (on the top) and $M = 10$ (bottom). With the white blended blue we depict the probability distribution of the time a first cell differentiates, exiting the first compartment C_1 of a sequence of $M = 3$ and $M = 10$ compartments, on the top and bottom plots, respectively. Differences in these distribution as well as with the distribution for $M = 1$, can be justified because in all the three cases we consider the same amplification factor $N = 10$. In this way we could keep consistency with respect to the population of product cells and analyse how the differences in the number of intermediate compartments reflect on compartments' events probabilities and the differentiation time. Looking at T_p^M for $M = 1, 3, 10$, we can observe that its probability distribution resemble a Gamma distribution with different shape parameters. A non zero rate of division should clearly make an impact on the dynamics of population of cells compared to the

simplified System (4.15); however, this result might hint that the overall distribution could still be approximated by a Gamma distribution. As M increases the probability distributions of T_p^M tends to skew to the right having a longer tail. This supports the natural thought that, given an total amplification factor N, the time to observe a first product cell (cell exiting compartment C_M) increases with the number of compartments in the sequence. Moreover, one can note that, in a sequence of M compartments, the time between the first differentiation event (first cell exiting compartment C_1) and the first differentiation event to the product state (first cell exiting C_M) increases as well with the number of intermediate compartments. Further mathematical studies are required to describe in analytical way this observed trend; for example one could think about how to approximate the solution of System (4.7), so that the resulting differentiated function reflects the numerical probability distribution function that has been found.

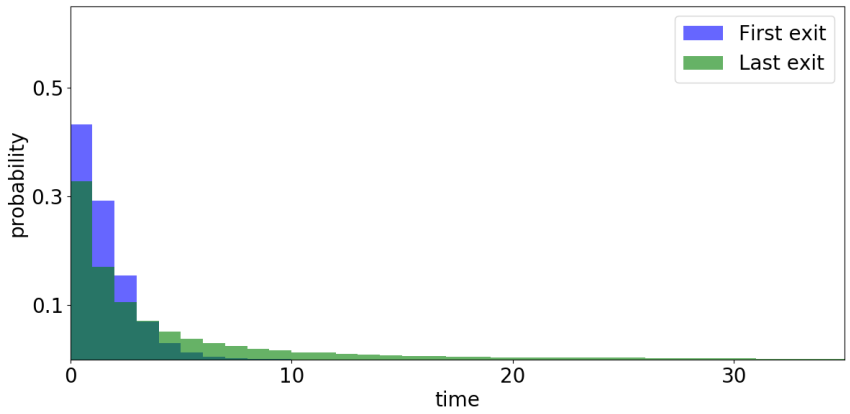


Figure 4.7: Plot of the probability distribution of the time a first cell exit (first product cell) and the last cell exit (last product cell) compartment C_M for $M = 1$, given one initial cell in that compartment. We set $N = 10$ and $p_{\mathrm{d}}(1) = \frac{\mu_1}{S_1} = 0.1$.

4.4.2 When will the last cell exit?

In this Section, we focus on the last differentiation event towards the product state. First, we want to observe that, within our compartmental model, the

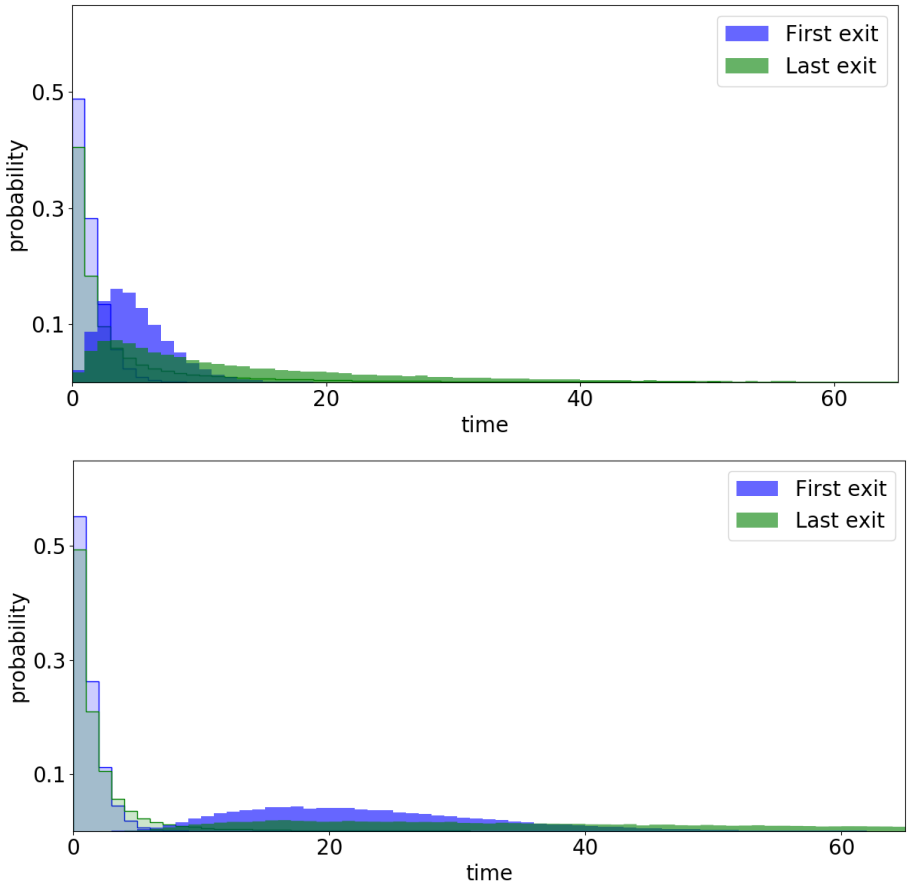


Figure 4.8: Plot of the probability distribution of the time a first cell exit (first product cell) and the last cell exit (first product cell) compartment C_M for $M = 3, 10$ (from top to bottom). Also, we report, in white blended colours, the probability distribution of the time the first (in blue) and the last (in green) cell exit compartment C_1, differentiating to the following one, for a sequence of $M = 3$ and $M = 10$ compartments. We set $N = 10$ and, for $i = 1, \dots, M$, $p_{\mathrm{d}}(i) = 0.1$.

dynamics of cells in compartment C_i, for $i = 2, \dots, M$ can be seen as a birth-and-death process with immigration (happening at rate ν_{i-1} from the previous compartment) (Allen, 2010). Different of course is the first compartment of progenitor cells. Indeed, we do not consider an external source of cells arriving to the first compartment and if $\mathbf{C}_1(t) = 0$ for a time $t = \tau$, then $\mathbf{C}_1(t) = 0$ for all $t > \tau$. We recall that in our system we assume, $\mu_i + \nu_i > \lambda_i$ for $i = 1, \dots, M$, so

that $\lim_{t\to+\infty}\mathbb{E}(\mathbf{C}_i(t))=0$ for $i=1,\dots,M$ and cells accumulate in the product state, $\lim_{t\to+\infty}\mathbb{E}(\mathbf{C}_{M+1}(t))\geq 0$. In order to quantify the time it takes for cells to traverse the sequence of compartments and leave the system, one can consider an initial progenitor cell and look at the time the last cell leaves (by death or differentiation) the sequence of M compartments $C_1,\dots,C_M$, before the system reaches the steady state $(0,\dots,0,n^\star_{M+1})$, with $n^\star_{M+1}\geq 0$. We let

$$T_0^M=\inf\{t\geq 0:\mathbf{C}_1(t)=\cdots=\mathbf{C}_M(t)=0\,|\,\mathbf{C}_1(0)=1,\,\mathbf{C}_2(0)=\cdots=\mathbf{C}_M(0)=0\}$$

be the random variable for the time to observe the first M compartments empty that is, the system reaches the steady state $(0,\dots,0,n^\star_{M+1})$, with $n^\star_{M+1}\geq 0$. As observed for the random variable T_p^M, since death event is included in the stochastic model, single cells can die before reaching the product state. When looking at the case of a non-zero population of product cells, we are interested in the case $\mathbf{C}_{M+1}(T_0^M)>0$; thus, one might consider the random variable $T_0^M\,\mathbb{1}_{\{T_0^{M+1}=+\infty\}}$ where $\mathbb{1}_{\{T_0^{M+1}=+\infty\}}$ is the indicator function defined as

$$\mathbb{1}_{\{T_0^{M+1}=+\infty\}}=\begin{cases}1, & \text{if}\quad T_0^{M+1}=+\infty,\\ 0, & \text{if}\quad T_0^{M+1}<+\infty.\end{cases}$$

In the case of direct differentiation ($M=1$), the time to observe $\mathbf{C}_1(t)=0$ is equivalent to the time to reach the absorbing state in a linear birth-and-death process (without immigration). We let

$$T_0^1=\inf\{t\geq 0:\mathbf{C}_1(t)=0\,|\,\mathbf{C}_1(0)=1\}\,,$$

be the random variable for the first time we observe $\mathbf{C}_1(t)=0$, i.e. all cells in compartment C_1 either have died or differentiated to the product state.

As mentioned in Section 4.3.1, setting $x_1=0$ and $x_2=1$ in Equation (4.8), gives $p_0(t)=\mathbb{P}\left(\mathbf{C}_1(t)=0\right)$ that is, the probability there are no more cells in compartment C_1 at time t. By differentiating $p_0(t)=\mathbb{P}\left(\mathbf{C}_1(t)=0\right)$ with respect to time, one can find the probability distribution of the first compartment to reach extinction,

$$\frac{d}{dt}\mathbb{P}\left(T_0^1<t\right)=\frac{(\mu+\nu)k^2\,\mathrm{e}^{-k\,t}}{((\mu+\nu)-\lambda\mathrm{e}^{-k\,t})^2}\,,$$

where $k = \mu + \nu - \lambda$. Indeed, given $x_1 = 0$ and $x_2 = 1$, it holds $\lambda(b-a) = \mu + \nu - \lambda = k$ where a, b are the roots of function $r(F_1)$. Similarly, for $M = 2$, one could set $x_1 = x_2 = 0$ and $x_3 = 1$ and solve System (4.7). The probability distribution of the first and second compartments to reach extinction, is then given by $\dfrac{d}{dt}F_1(0, 0, 1, t)$.

Moreover, when looking at summary statistics related to the system reaching the absorbing state $(0, \ldots, 0, n^\star_{M+1})$, we find interesting to consider the time the last differentiation event to the product state takes place, conditioned on at least one cell reaching the product state, $n^\star_{M+1} > 0$. Let us imagine, for example, to have a sequence of $M = 2$ compartments. We start at time $t = 0$ with only one cell in the first compartment, $\mathbf{C}_1(0) = 1$ and $\mathbf{C}_2(0) = \mathbf{C}_3(0) = 0$, and observe that at time $\tau_1 > 0$, the following state of the system

$$\mathbf{C}_1(\tau_1) = 2, \quad \mathbf{C}_2(\tau_1) = 1, \quad \mathbf{C}_3(\tau_1) = 10\,.$$

If now, at time τ_1, the cell in compartment C_2 differentiates to the product state and afterwards, at times τ_2 and τ_3, the two cells in C_1 die, then the system would reach the absorbing state $(0, 0, 11)$ at time τ_3. However, what we would like to investigate here is τ_1, as the time the last differentiation event to the product state happens. Note that, in case of $\mu_i = 0$ for $i = 1, \ldots, M$, the distribution given by τ_1 equals the distribution of the random variable T_0^M.

In order to analyse this summary statistic, we run 10^5 numerical simulations. Figure 4.7 reports in green the probability distribution of time to observe the last cell leaving compartment C_1 and differentiating to the product state, before the system reaches the absorbing state. Similarly, we report in Figure 4.8 in green the probability distribution of time to observe the last cell leaving compartment C_M and differentiating to the product state, for $M = 3$ and $M = 10$ on the top and bottom plots, respectively. As noted in previous Section for T_p^M, we depict in white blended green the probability distribution of the time the last cell in C_1 differentiates, exiting the first of a sequence of $M = 3$ and $M = 10$ compartments, on the top and bottom plots, respectively. Differences in these distribution as well as with the distribution for $M = 1$, can be justified because for the three choices of M we consider the same amplification factor $N = 10$. In this way, we could keep consistency with respect to the population of product cells and

analyse how the differences in the number of intermediate compartments reflect on compartments' events probabilities and the differentiation time. If for $M = 1$, the time to observe the last differentiation event follows an exponential distribution, when adding more compartments in the sequence, the distribution is skewed to the right with a shape parameter depending on the number of compartments M. For consistency we restrict, for $M = 10$, the x-axis to a time $t = 60$; however, we want to underline that the probability distribution has a very long tail and, as a result of the 10^5 simulations, probability approaches zero only for larger times, $t > 250$.

4.4.3 Exiting time

Finally, in this Subsection, we consider the distribution of times cells differentiate among compartments before the system reaches the absorbing state. In Section 4.2 we show how to compute $\mathbb{E}(\mathbf{C}_i(t))$ the mean number of cells in compartment C_i, for $i = 1, \ldots, M$, at time t. Then, given that the defined model is a CTMC where cells in compartment C_i behave independently, the probability of observing a differentiation event of a cell in compartment C_i at time t is given by the differentiation rate times the number of cells present at that time in the compartment. That is,

$$\rho_i(t) = \nu_i \mathbb{E}(\mathbf{C}_i(t)) \,.$$

In particular, we can look at the population of cells in the last compartment C_M and analyse the distribution of the time cells exit C_M, that is $i = M$. Then, if $M = 1$,

$$\rho_1(t) = \nu_1 \, \mathbb{E}(\mathbf{C}_1(t)) = \nu_1 \mathrm{e}^{-k_1 \, t} = \nu_1 \sum_{n=0}^{+\infty} n \, p_n(t) \,,$$

with $p_n(t) = \mathbb{P}\left(\mathbf{C}_1(t) = n\right)$ as in Equation (4.10). In this case of $M = 1$, we do have an analytical expression of both the mean number of cells $\mathbb{E}(\mathbf{C}_1(t))$ and the probability $\mathbb{P}\left(\mathbf{C}_1(t) = n\right)$. Whereas, for $M > 1$ only an analytical expression of $\mathbb{E}(\mathbf{C}_i(t))$, given by Equation (5.5), is found. However, a corresponding probability mass function $p_n^M(t) = \mathbb{P}\left(\mathbf{C}_M(t) = n\right)$ can be always defined. For example, for

$M = 2$ one would have

$$\nu_2 \, \mathbb{E}(\mathbf{C}_2(t)) = \nu_2 (\mathrm{e}^{-k_1 t} - \mathrm{e}^{-k_2 t}) = \nu_2 \sum_{n=0}^{+\infty} n \, p_n^2(t) \, .$$

It is left for the future, the problem to find this probability mass function, whose analytical expression would be extremely interesting to analyse.

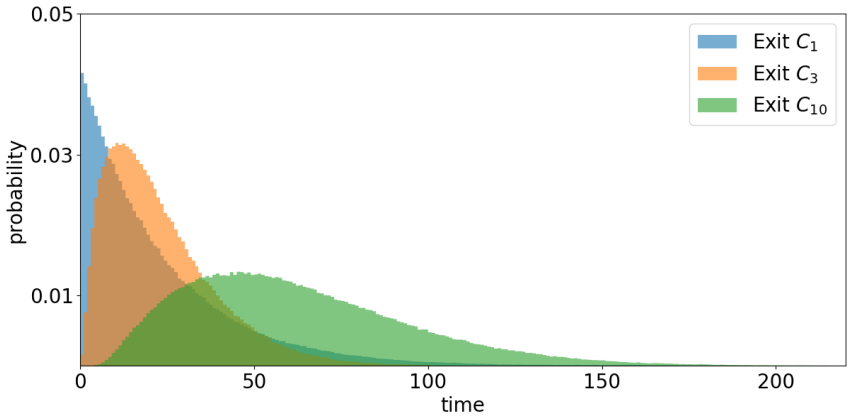


Figure 4.9: Plot of the distribution of the differentiation times towards a product state (time cells exit compartment C_M via differentiation event) for $M = 1, 3, 10$, starting from one progenitor cell in C_1. In all three cases, we set $N = 10$ and, for $i = 1, \ldots, M$, $p_{\mathrm{d}}(i) = \frac{\mu}{S_i} = 0.1$.

Here, we perform a numerical analysis by running 10^5 Gillespie simulations where we track when each differentiation event happens. Figure 4.9 reports the probability distributions of the time differentiation events in compartments C_1, C_3 and C_{10} happen (in blue, orange and green) for a sequence of $M = 1, 3, 10$ compartments, respectively. As expected, we note that for $M = 1$, the distribution follows an exponential with parameter k_1. Exiting times to the product state for a sequence of $M > 1$ compartments are shifted to the right. In particular, at around time $t = 12$, a differentiation event towards the product state for $M = 3$ is the most probable; whereas a differentiation event to the product state of a cell in compartment C_{10} is likely to happen even at larger times and it is unlikely at very early times, for $t < 5$.

So overall, even if the three considered sequences of $M = 1, 3$ and 10 compartments are characterised by the same amplification factor N (and the same probability of death among different C_i, for $i = 1, \ldots, M$) cells arrive to the product-final state at different times. Also, when $M = 1$, by differentiating the probability generating function of the process, we do find a closed expression for the time of a first cell differentiates to the product state. Whereas, if $M > 1$, the nature of the differential equation for $F_1(x_1, \ldots, x_{M+1}, t)$ drives an increase in the complexity: for this PhD thesis, instead of finding a closed formula, we carried out numerical simulations to understand the key aspects of the process.

4.5 Discussion

In this Chapter, a time-dependent stochastic model for cell division, death and differentiation through a sequence of M compartments has been developed. We focus on understanding how cells proceed in the sequence of compartments over time, reaching the product state changes at different times, and how these observations are impacted by differences in the number of compartments in the sequence. At each time step, the number of cells in compartment C_i, for $i = 1, \ldots, M$, is represented by the random variable $\mathbf{C}_i(t)$; the number of product cells at time t is represented by the random variable $\mathbf{C}_{M+1}(t)$. The mean number of cells at each time step is computed and a probability generating function approach is followed With this, we aim to characterise the distribution of exiting cells and find $\mathbb{P}(\mathbf{C}_{M+1}(t) = n)$ at time t that is the probability the number of product cells, descended from a single progenitor via M intermediate compartments, is equal to n at time t. The calculation of such distribution is challenging since it requires the knowledge, at each time step, of the number of cells in all previous intermediate compartments. Thus, a second stochastic approach is proposed to compute summary statistics of interest, such as the first and the last time a cell of compartment C_M differentiate to the product state, starting from one cell in the progenitor state and conditioned on having an non-zero population of product cells.

Overall, the work presented in this Chapter is preliminary work setting up the space for future work. The analytical analysis has been carried out only for

$M = 1$ and $M = 2$; and an exact solution of System (4.5) for $M > 2$ requires further analytical studies as the complexity of dealing with multi-compartments is already evident from the case $M = 2$. Indeed, when $M = 3$, the probability generating function $F_3(x_3, x_4, t)$ can be directly computed as in Equation (4.8); then, $F_2(x_2, x_3, x_4, t)$ can be solved in terms of hypergeometric functions, applying the Frobenius method as we did for the case $M = 2$. Consequently, one need to deal with a differential equation with respect to time for $F_1(x_1, x_2, x_3, x_4, t)$ where the time-dependent coefficient of the inhomogeneous term is a linear combination of hypergeometric functions. An exact solution of such differential equation involves complex calculations; alternatively, one might think to approximate $F_2(x_2, x_3, x_4, t)$ with a sum of exponential (i.e. Taylor expansion) and, thus, solve a simplified differential equation for $F_1(x_1, x_2, x_3, x_4, t)$. Similar to what we did when we considered a simplified function, $F_2(t) \sim \mathrm{e}^{\gamma t}$, for the case $M = 2$ (see Section 4.3.2 and Appendix C.2). Generalisation of analytical methods for the case of $M > 2$ is out of this PhD thesis and is the aim of future work.

In the time-analysis of the compartmental model outlined in this Chapter, we have not considered the generation number of cells differentiating through the sequence of compartments and eventually reach the product state. We believe rather that investigating how the generation number of product cells changes over time and as a function of the number of intermediate compartments in the sequence, is a crucial point, interesting both from a mathematical and biological perspective. Indeed, as mentioned in Chapter 3, in many circumstances a high number of rounds of cell divisions brings a risk of mutation becoming established. Already in Chapter 3 we analyse the differentiation process in terms of the number of divisions that cells have undergone during the sequence of compartments; in particular, we consider the time the system reached the absorbing state and analyse the population of product cells. This has been done by definition of the random variable $\mathbf{G}$ for the generation number of a randomly-selected product cell as in Equation (3.35) for a sequence of M compartments.

In the framework on the time-dependent analysis performed in this Chapter, we can investigate how the generation number varies over time, by defining the random variable $\mathbf{G}(t)$ for the generation number of a randomly-selected product

cell at time t for a sequence of M compartments; such a mathematical investigation requires more calculations and definition of specific random variables. For example, one could consider, similarly to Section 3.3.1, random variables describing the number of cells of generation n in each compartment at time t as well as a random variable describing the number of product cells of generation n at time t. In particular, for $M = 1$, we can define $\mathbf{Z}_n(t)$ as the random variable describing the number of cells of generation n in compartment C_1 at time t and $\mathbf{Y}_n(t)$ as the random variable describing the number of product cells of generation n at time t. In order to apply a generating function approach, we find necessary to limit the generation space and, thus, set a boundary to the generation number that is tracked. In particular, we could assume that only cells up to a generation G are distinguished and cells of generation $n \geq G$ belong to the same class. Under this assumption, random variables $\mathbf{Z}_0(t), \ldots, \mathbf{Z}_G(t)$ and $\mathbf{Y}_0(t), \ldots, \mathbf{Y}_G(t)$ need to be defined and, consequently, the joint probabilities

$$
\begin{aligned}
&p_{m_n,\ldots,m_G,l_n,\ldots,l_G}(t) \\
&\quad = \mathbb{P}\big(\mathbf{Z}_n(t) = m_n, \ldots, \mathbf{Z}_G(t) = m_G, \mathbf{Y}_n(t) = l_n, \ldots, \mathbf{Y}_G(t) = l_G \,| \\
&\quad \mathbf{Z}_n(0) = 1, \mathbf{Z}_{n+1}(0) = \cdots = \mathbf{Z}_G(0) = 0, \mathbf{Y}_n(0) = \cdots = \mathbf{Y}_G(0) = 0\big),
\end{aligned}
$$

with $m_j, l_j \geq 0$ for $j = n, \ldots, G$. Then, we can define the time-dependent probability generating function $\xi_{1,n}$ and $\xi_{2,n}$ for random variables $\mathbf{Z}_n(t)$ and $\mathbf{Y}_n(t)$, respectively. We have

$$
\xi_{1,n}(z_n, \ldots, z_G, y_n, \ldots, y_G, t) = \sum_{\substack{m_n,\ldots,m_G, \\ l_n,\ldots,l_G=1}}^{+\infty} p_{m_n,\ldots,m_G,l_n,\ldots,l_G}(t) z_n^{m_n} \cdots z_G^{m_G}\, y_n^{l_n} \cdots y_G^{l_G}
$$

and $\xi_{2,n}(y_n, \ldots, y_G, t) = y_n$, for all $n = 0, \ldots, G$, since product cells do not have dynamics, by definition. Note that function $\xi_{1,n}$ is a function of $2G - 2n + 3$ variables and we are interested in its partial derivative with respect to time. In

particular, when $M = 1$ and $G = 2$, the following differential equations hold

$$\begin{aligned}
\frac{\partial}{\partial t}\xi_{1,0}(z_0, z_1, z_2, y_0, y_1, y_2, t) &= -S\,\xi_{1,0}(z_0, z_1, z_2, y_0, y_1, y_2, t) + \mu \\
&\quad + \lambda\,\xi_{1,1}^2(z_1, z_2, y_1, y_2, t) + \nu y_0\,, \\
\frac{\partial}{\partial t}\xi_{1,1}(z_1, z_2, y_1, y_2, t) &= -S\,\xi_{1,1}(z_1, z_2, y_1, y_2, t) + \mu + \lambda\,\xi_{1,2}^2(z_2, y_2, t) + \nu y_1\,, \\
\frac{\partial}{\partial t}\xi_{1,2}(z_2, y_2, t) &= -S\,\xi_{1,2}(z_2, y_2, t) + \mu + \lambda\,\xi_{1,2}^2(z_2, y_2, t) + \nu y_2\,.
\end{aligned}$$

with $S = \lambda + \mu + \nu$ the sum of the rates of compartment C_1. Given system's structure, the last equation for the highest generation needs to be solved at first. The right hand side is a quadratic equation with an inhomogeneous term constant with respect to time, thus we can use (4.8) and find

$$\xi_{1,2}(z_2, y_2, t) = \frac{\alpha_2\beta_2(\mathrm{e}^{-\lambda(\beta_2-\alpha_2)t} - 1) + z_2(\alpha_2 - \beta_2\mathrm{e}^{-\lambda(\beta_2-\alpha_2)t})}{a_2\mathrm{e}^{-\lambda(\beta_2-\alpha_2)t} - \beta_2 + z_2(1 - \mathrm{e}^{-\lambda(\beta_2-\alpha_2)t})}\,,$$

where $\alpha_2 < \beta_2$ roots of the second order polynomial $\mu+\nu y_2-(\lambda+\mu+\nu)\,\xi_{1,2}(z_2, y_2, t)+\lambda\,\xi_{1,2}^2(z_2, y_2, t) = 0$. Solutions of $\xi_{1,1}$ and $\xi_{1,0}$ come by time-integration of the corresponding differential equation that is now a linear first order differential equation with time-dependent inhomogeneous term. Thus, the desired probabilities could be computed but at a cost of complex and tedious calculations. We reserve for future work the mathematical analysis, following a probability generating function approach, for cells generation number within a sequence of M compartments.

Nevertheless, one can make use of numerical simulations to investigate how the generation number of a randomly-selected product cell changes over time, by analysing the distribution of the random variable $\mathbf{G}(t)$. In particular, we can consider $\mathbb{P}(\mathbf{G}(t) = n)$ for different numbers of intermediate compartments M. Figure 4.10 reports a heat map for the distribution of $\mathbf{G}(t)$ for $M = 3$ (on top) and $M = 10$ (on bottom). Here, we fix an average number of product cells $N = 100$ descended from a single progenitor cell as well as $p_{\mathrm{d}}(i) = \frac{\mu_i}{S_i} = 0.1$ for all compartments C_i, $i = 1, \ldots, M$ so that probabilities $p_{\mathrm{b}}(i) = \frac{\lambda_i}{S_i}$ and $p_{\mathrm{e}}(i) = \frac{\nu_i}{S_i}$ are computed as in (3.45). One can note that the presence of many intermediate compartments is advantageous for the generation number of product cells. Indeed, Figure 4.10 shows that, given N, it takes more for cells to exit a longer sequence of compartments and reaching the product state, but they do it with a lower

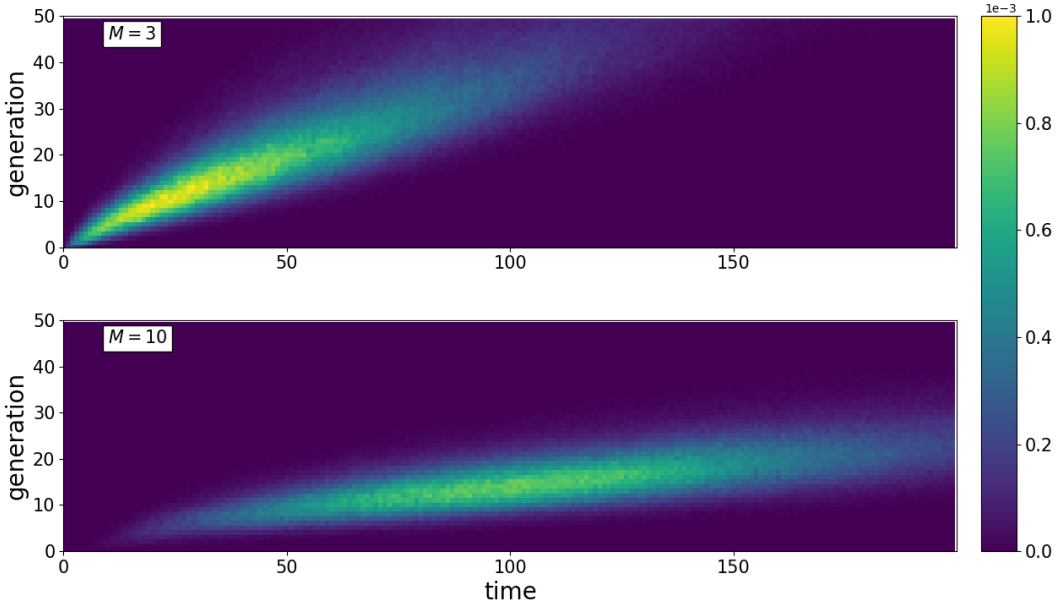


Figure 4.10: Heat map of $\mathbb{P}(\mathbf{G}(t) = n)$ for $M = 3$ (on the top) and $M = 10$ (bottom). On the x-axis time, on the y-axis generation number. In both cases, we set $N = 100$ and $p_{\mathrm{d}}(i) = \frac{\mu_i}{S_i} = 0.1$ for all compartments C_i, $i = 1, \ldots, M$. Probabilities $p_{\mathrm{b}}(i) = \frac{\lambda_i}{S_i}$ and $p_{\mathrm{e}}(i) = \frac{\nu_i}{S_i}$ are computed as in (3.45).

generation number. Indeed, when $M = 10$ most of the cells reach the product state around time $t = 100$ but with a maximum generation number of 20.

Beyond the future research that we mentioned above, a natural prosecution of this study would consist in the inclusion of different cellular events. For example, in Chapter 5, we define a time-dependent compartmental model where the symmetric division event (when a cell in compartment C_i divides, the two daughter cells both belongs to compartment C_{i+1}) as well as the reverse transition (a cell de-differentiate moving from compartment C_i to C_{i-1}) are considered. A study of distributions $\mathbb{P}(\mathbf{C}_{M+1}(t) = k)$ and $\mathbb{P}(\mathbf{G}(t) = n)$ could be extended when these two types of event are included in the model definition.

Chapter 5

Stochastic journeys of cell progenies through compartments: self-renewal, symmetric and asymmetric division

Cellular division, death and differentiation are essential ongoing processes during all stages of life. A well-known example is embryonic development, where non-specialised cells give rise to functional and differentiated cells (Evans & Kaufman, 1981; Martin, 1981). In Chapter 1, we give an overview of cell differentiation process in human body (Barile *et al.*, 2020; Michor *et al.*, 2005; Murray *et al.*, 2011; Sawicka *et al.*, 2014; Zhang *et al.*, 2015) with a specific focus on the dysfunctional exhaustion process (Chen *et al.*, 2019; Wherry & Kurachi, 2015).

Recent experimental techniques in cell biology have made it possible to track individual cell states (or compartments) and the progeny of a single cell (Hodzic, 2016; Perié *et al.*, 2014). In recent decades, the development of mathematical models in biology has significantly increased our ability to gain a quantitative understanding of cellular fate and cell population dynamics. Deterministic models are typically easier to analyse; they do not incorporate randomness and allow one to describe the dynamics over time of populations of hosts, cells or molecules. However, when interested in tracking single-cell behaviour, stochastic fluctuations

arising from complex cellular interactions and the molecular events which regulate cellular fate, must be considered.

Here, similar to Chapters 3 and 4, we consider a sequence of $M+1$ compartments and model the process by which, from a stem-like (progenitor) state (compartment C_1), cells can undergo differentiation/migration events across adjacent compartments, potentially leading to the rise of a terminally differentiated product population (compartment C_{M+1}). Cells in each compartment can divide, die or transit to adjacent compartments (e.g representing potentially reversible phenotype change). In contrast to Chapters 3 and 4, in the stochastic model defined in this Chapter, differentiation event can be either considered to be reversible (e.g cancer cell mutations) or arising from a mutation event that has a negligible probability of being reversed (e.g embryonic cell development).

Moreover, here, additional cell events are considered. Indeed, advances in genetic labelling (Klein & Simons, 2011) have confirmed the existence of different types of division, which affect the flexibility of a cell pool to expand or contract. Different division events can drive a range of biological processes according to whether the less differentiated pool expands by one new cell (self-renewal), stays the same size (asymmetric division, where one of the daughter cells changes phenotype), or shrinks (symmetric division, where both daughter cells change phenotype) (Barile *et al.*, 2020). In the time-dependent mathematical model defined in this Chapter, we consider two additional cellular events compared to previous Chapters: a reversible differentiation event, and the symmetric division event.

The Chapter is organised as follows. In Section 5.1 we describe the dynamics of cells dividing, dying or exiting across a sequence of compartments over time, in terms of a continuous-time Markov chain. The mean behaviour of the system is analysed in Section 5.1.1 and results are applied also in Section 4.2 as System 4.1 results as a simplified case of what considered here. In Section 5.1.2, we make use of the probability generating function approach and we study the proliferative potential of the system by quantifying the number of cells within the genealogy of a single progenitor cell. In both Sections, we consider either the situation where differentiation events can be reversible, or a mathematically more simple *irreversible model*, where differentiation to the next compartment cannot

be reversed. In Section 5.1.3, our focus is a number of summary statistics related to a single cell which we track over time. Finally, in Section 5.2, we propose a set of numerical results inspired by biological applications to illustrate our approach and methods, and to highlight the impact that asymmetric and symmetric division can have on the fate of a population of cells arising from a progenitor.

5.1 Stochastic compartmental model

We propose a stochastic model of cell division, death and differentiation (or migration) across an ordered sequence of compartments. Cells in a given compartment may represent a spatial location, a common phenotype, or share some common characteristics. As in Chapters 3 and 4, we consider a sequence of compartments C_i, $i \in \{1, \dots, M+1\}$, which cells, behaving independently of each other, follow.

Our general stochastic model considers a number of cellular events inspired by some recently proposed mathematical models (Barile *et al.*, 2020; de la Higuera *et al.*, 2019; Sawicka *et al.*, 2014; Zhang *et al.*, 2015). If in Chapter 3 we work with event probabilities defining cells transitions between compartments, here and in Chapter 4 we assume that each of these cell events takes place at a certain per cell rate. Of course, since the defined model is a continuous-time birth-death-migration Markov process with exponential waiting times, one could relate the defined event rates to the corresponding compartment probabilities. Particular situations of interest arise from setting some of these rates equal to zero, so some events are not allowed to happen, as we illustrate in three different case studies in Section 5.2. Each cell in a given compartment, C_i, can divide, die or exit to one of the two adjacent compartments. When a division occurs, daughter cells might both belong to the same compartment as the mother (this event is referred to as *self-renewal*), both daughter cells might instantaneously move to the next compartment (*symmetric division*), or one daughter cell might belong to the same compartment as the mother, and the other will belong to the next compartment (*asymmetric division*) (Barile *et al.*, 2020; Zhang *et al.*, 2015).

The stochastic model, depicted in Figure 5.1, is a continuous-time Markov chain (CTMC) $\mathcal{X} = \{(\mathbf{C}_1(t), \mathbf{C}_2(t), \dots, \mathbf{C}_{M+1}(t)) : \ t \geq 0\}$, where the random variable $\mathbf{C}_i(t)$ represents the number of cells in compartment C_i at time t, with

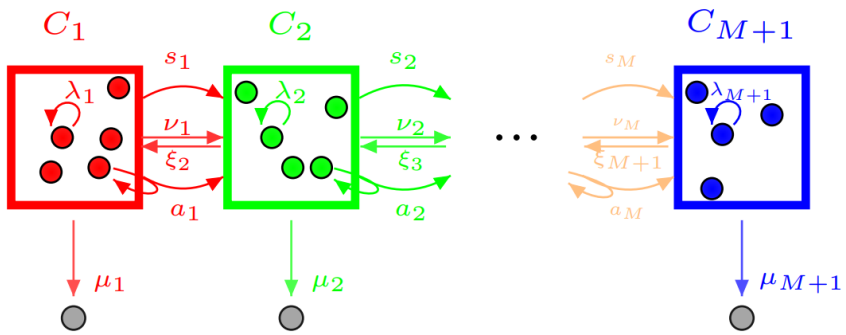


Figure 5.1: General stochastic model of cell division, death and differentiation across an ordered sequence of compartments. Grey cells represent cellular death. Self-renewal events take place with rate λ, symmetric division events with rate s, and asymmetric division events with rate a. Differentiation events happen with rate ν (forward) or ξ (backward). Death events have per cell rate μ. All rates are assumed to be real positive numbers.

state space given by $\{0,1,2,\ldots\}^{M+1} = \mathbb{N}_0^{M+1}$. Cellular events, labelled E1 to E5 below, represent transitions across states $(n_1,\ldots,n_{M+1}) \in \mathbb{N}_0^{M+1}$ as follows:

(E1) Self-renewal (cellular division where both daughter cells remain in the same compartment as the mother) can occur in any compartment C_i, with per cell rate λ_i, for $i \in \{1,\ldots,M+1\}$,

$$(n_1,\ldots,n_{i-1},n_i,n_{i+1},\ldots,n_{M+1}) \xrightarrow{\lambda_i n_i} (n_1,\ldots,n_{i-1},n_i+1,n_{i+1},\ldots,n_{M+1}).$$

(E2) Symmetric division (cellular division where both daughter cells instantaneously move to the next compartment) occurs in compartment C_i with per cell rate s_i, for $i \in \{1,\ldots,M\}$,

$$(n_1,\ldots,n_{i-1},n_i,n_{i+1},\ldots,n_{M+1}) \xrightarrow{s_i n_i} (n_1,\ldots,n_{i-1},n_i-1,n_{i+1}+2,\ldots,n_{M+1}).$$

(E3) Asymmetric division (cellular division where one of the daughter cells remains in the same compartment as the mother, while the other goes to the next compartment) occurs in compartment C_i with per cell rate a_i, for $i \in \{1,\ldots,M\}$,

$$(n_1,\ldots,n_{i-1},n_i,n_{i+1},\ldots,n_{M+1}) \xrightarrow{a_i n_i} (n_1,\ldots,n_{i-1},n_i,n_{i+1}+1,\ldots,n_{M+1}).$$

(E4) Differentiation (or migration) between adjacent compartments can occur with per cell rates ν_i, for $i \in \{1, \dots, M\}$ and ξ_i, for $i \in \{2, \dots, M+1\}$,

$$(n_1, \dots, n_{i-1}, n_i, n_{i+1}, \dots, n_{M+1}) \xrightarrow{\nu_i n_i} (n_1, \dots, n_{i-1}, n_i - 1, n_{i+1} + 1, \dots, n_{M+1}),$$
$$(n_1, \dots, n_{i-1}, n_i, n_{i+1}, \dots, n_{M+1}) \xrightarrow{\xi_i n_i} (n_1, \dots, n_{i-1} + 1, n_i - 1, n_{i+1}, \dots, n_{M+1}).$$

(E5) Cells can die in any compartment C_i with per cell rate μ_i, $i \in \{1, \dots, M+1\}$,

$$(n_1, \dots, n_{i-1}, n_i, n_{i+1}, \dots, n_{M+1}) \xrightarrow{\mu_i n_i} (n_1, \dots, n_{i-1}, n_i - 1, n_{i+1}, \dots, n_{M+1}).$$

We note that cells in the last compartment C_{M+1} cannot symmetrically or asymmetrically divide, or differentiate to the next compartment. All per cell rates are assumed to be positive real numbers. Moreover, we would like to underline the reader that, when $s_i = \xi_i = 0$ for all $i = 1, \dots, M$ and $\xi_{M+1} = \lambda_{M+1} = \mu_{M+1} = 0$ we fall back into the hypothesis considered in previous two Chapters.

5.1.1 Mean number of cells in each compartment over time

We first describe the dynamics of the process and study the mean number of cells in each compartment, $\mathbb{E}(\mathbf{C}_i(t))$, which obeys the following system of differential equations (Matis, 1970)

$$\begin{aligned}
\frac{d\,\mathbb{E}(\mathbf{C}_1(t))}{dt} &= -(\mu_1 + \nu_1 + s_1 - \lambda_1)\mathbb{E}(\mathbf{C}_1(t)) + \xi_2 \mathbb{E}(\mathbf{C}_2(t)), \\
\frac{d\,\mathbb{E}(\mathbf{C}_i(t))}{dt} &= (\nu_{i-1} + a_{i-1} + 2s_{i-1})\,\mathbb{E}(\mathbf{C}_{i-1}(t)) - (\mu_i + \nu_i + \xi_i + s_i - \lambda_i)\,\mathbb{E}(\mathbf{C}_i(t)) \\
&\quad + \xi_{i+1}\,\mathbb{E}(\mathbf{C}_{i+1}(t)), \quad i \in \{2, ..., M\}, \qquad (5.1) \\
\frac{d\,\mathbb{E}(\mathbf{C}_{M+1}(t))}{dt} &= (\nu_M + a_M + 2s_M)\,\mathbb{E}(\mathbf{C}_M(t)) \\
&\quad - (\mu_{M+1} + \xi_{M+1} - \lambda_{M+1})\mathbb{E}(\mathbf{C}_{M+1}(t)),
\end{aligned}$$

where $\mathbb{E}(\mathbf{C}_i(t))$ represents the expectation of the random variable $\mathbf{C}_i(t)$. These equations constitute a homogeneous first-order linear system of ODEs with constant coefficients, which can be written more succinctly in matrix form as follows

$$\frac{d\mathbf{C}(t)}{dt} = \mathbf{A}\,\mathbf{C}(t), \qquad (5.2)$$

where $\mathbf{C}(t) = (\mathbb{E}(\mathbf{C}_1(t)), \mathbb{E}(\mathbf{C}_2(t)), \ldots, \mathbb{E}(\mathbf{C}_M(t)), \mathbb{E}(\mathbf{C}_{M+1}(t)))^T$,

$$\mathbf{A} = \begin{pmatrix} -K_1 & \xi_2 & 0 & \cdots & 0 \\ L_1 & -K_2 & \xi_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & L_{M-1} & -K_M & \xi_{M+1} \\ 0 & \cdots & 0 & L_M & -K_{M+1} \end{pmatrix}, \tag{5.3}$$

and

$$\begin{aligned} K_1 &= \mu_1 + \nu_1 + s_1 - \lambda_1, \\ K_i &= \mu_i + \nu_i + s_i + \xi_i - \lambda_i, \quad L_{i-1} = \nu_{i-1} + a_{i-1} + 2s_{i-1}, \ i \in \{2, \ldots, M\}, \\ K_{M+1} &= \mu_{M+1} + \xi_{M+1} - \lambda_{M+1}, \quad L_M = \nu_M + a_M + 2s_M. \end{aligned}$$

Note that if $s_i = 0$ for $i = 1, \ldots, M$ and $\xi_i = 0$ for $i = 2, \ldots, M+1$ then $K_i = \mu_i + \nu_i - \lambda_i$; recalling the k_i defined in Chapter 4, we then have $K_i = k_i$.

The initial value problem of Equation (5.2) with $\mathbf{C}_0 = \mathbf{C}(0)$ has a unique solution (Allen, 2010, Theorem 4.1)

$$\mathbf{C}(t) = e^{\mathbf{A}t}\mathbf{C}_0, \tag{5.4}$$

where $e^{\mathbf{A}t}$ represents the matrix exponential,

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2!} + \mathbf{A}^3\frac{t^3}{3!} + \ldots = \sum_{i=0}^{+\infty} \frac{(\mathbf{A}t)^i}{i!}.$$

System (5.2) admits $\lim_{t\to+\infty} \mathbf{C}(t) = \mathbf{0}_{M+1}$ (column vector of zeros) as an asymptotic solution. It is exponentially stable (all solutions of the system from any initial conditions converge to extinction) if and only if each eigenvalue of $\mathbf{A}$ has a negative real part (see, for example, Corollaries 3.5 and 3.6 of the work of Teschl (2012)).

We note that, in certain biological applications, some of the rates in Figure 5.1 will be zero, and thus, the analysis of such systems would simplify. For instance, differentiation may be irreversible (Barile *et al.*, 2020; Sawicka *et al.*, 2014), so that $\xi_i = 0$ for $i \in \{2, \ldots, M+1\}$ as shown in Figure 5.2. We will refer to this scenario as the *irreversible model*. In this case, and if one considers a single

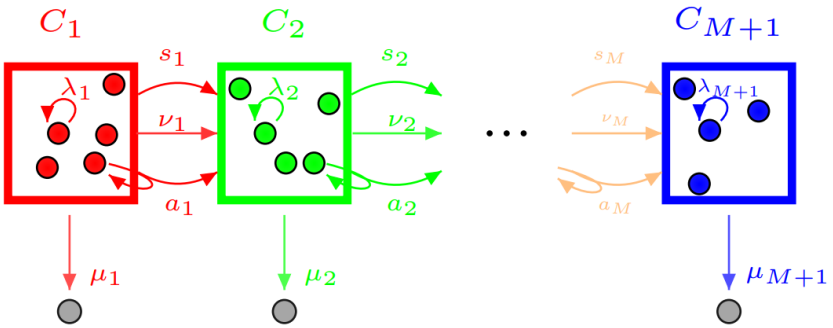


Figure 5.2: Irreversible stochastic model of cell division, death and forward differentiation across an ordered sequence of compartments. Grey cells represent cell death happening at per cell rate μ_i. For a compartment C_i, self-renewal events take place with rate λ_i, symmetric division events with rate s_i, and asymmetric division events with rate a_i. Differentiation events happen with rate ν_i. All rates are assumed to be real positive numbers.

progenitor cell starting in compartment C_1 at time $t = 0$, $\mathbf{C}(0) = (1, 0, \ldots, 0)^T$, it is possible to obtain the mean behaviour of the system in terms of the following expressions

$$\mathbb{E}(\mathbf{C}_i(t)) = \begin{cases} e^{-K_1 t}, & i = 1, \\ \left(\prod_{l=1}^{i-1} L_l\right) \sum_{j=1}^{i} e^{-K_j t} \prod_{\substack{m=1 \\ m \neq j}}^{i} (K_m - K_j)^{-1}, & i \in \{2, \ldots, M+1\} \end{cases} \tag{5.5}$$

The expression above is only well-defined if $K_i \neq K_j$ for all pairs (i, j) with $i \neq j$. If this is not the case, alternative analytic solutions can be found. For example, if $K_i = K_j$ for all $i, j \in \{1, \ldots, M+1\}$, Equation (5.5) simplifies to

$$\mathbb{E}(\mathbf{C}_i(t)) = \left(\prod_{l=1}^{i-1} L_l\right) \frac{t^{i-1}}{(i-1)!} e^{-K_i t}, \quad t \geq 0. \tag{5.6}$$

Thus, it is clear that in the irreversible model $\lim_{t \to +\infty} \mathbb{E}(\mathbf{C}_i(t)) = 0$ when $K_i > 0$ $\forall i \in \{1, \ldots, M+1\}$. This is consistent since $\{-K_i,\ i \in \{1, \ldots, M+1\}\}$ are the eigenvalues of $\mathbf{A}$ in this case.

For specific biological applications such as those considered in Section 5.2 (Sawicka *et al.*, 2014; Zhang *et al.*, 2015), it is of interest to quantify the cumulative

number of cells, on average, that exit compartment C_M and arrive to the final compartment, C_{M+1}, starting with a single progenitor cell in compartment C_1. To this end, one can set in the irreversible model $\lambda_{M+1} = \mu_{M+1} = 0$, so that cells exiting compartment C_M accumulate and can be counted. Then, for the last compartment $K_{M+1} = 0$. We may underline the reader that these are the assumptions made in Chapters 3 and 4. In this case, Equation (5.5) leads to

$$\mathbb{E}(\mathbf{C}_{M+1}(t)) = \left(\prod_{l=1}^{M} L_l\right)\left[\sum_{j=1}^{M} e^{-K_j t}\prod_{\substack{m=1\\m\neq j}}^{M}(K_m - K_j)^{-1} + \prod_{m=1}^{M} K_m^{-1}\right], \tag{5.7}$$

which is well-defined for $K_i \neq K_j$ for all $i, j \in \{1, \ldots, M\}$. From here, we have

$$\lim_{t\to+\infty} \mathbb{E}(\mathbf{C}_{M+1}(t)) = \prod_{i=1}^{M} \frac{L_i}{K_i}. \tag{5.8}$$

Interestingly, this limit holds also if $K_i = K_j$ for all $i, j \in \{1, \ldots, M\}$ and $K_{M+1} = 0$. Then, one can obtain

$$\mathbb{E}(\mathbf{C}_{M+1}(t)) \;=\; \prod_{i=1}^{M} \frac{L_i}{K_i} - \sum_{j=1}^{M} \mathbb{E}(\mathbf{C}_j(t)) \prod_{i=j}^{M} \frac{L_i}{K_i}.$$

Under population extinction conditions ($K_i > 0$ for all $i \in \{1, \ldots, M\}$, so that the population of cells in intermediate compartments $\{1, \ldots, M\}$ dies out and only product cells remain in C_{M+1} at late times), $\lim_{t\to+\infty} \mathbb{E}(\mathbf{C}_i(t)) = 0$ for all $i \in \{1, \ldots, M\}$, and thus Equation (5.8) also holds.

A particular feature of this system is that cells behave independently from each other. This means that the dynamics of the genealogy of a set of K progenitor cells in compartment C_1 at time $t = 0$, can be analysed as K independent stochastic processes. Thus, in Section 5.1.2 we consider a number of summary statistics of interest related to the genealogy of a single cell starting in a given compartment (typically compartment C_1).

5.1.2 The genealogy of a single progenitor cell

For a cell starting in compartment C_i, we can define $\mathbf{g}_i$ to be the random variable representing the total number of cells in the genealogy of this cell. Cells in the

genealogy are the daughters, granddaughters, etc. of the progenitor cell, which originate from division events (either self-renewal, asymmetric or symmetric) in any compartment over time, not including the progenitor cell. It is a summary statistic of the process which quantifies the proliferative potential of a single cell in compartment C_i, and its offsprings. For example, in Figure 5.3, we represent a particular realisation of the stochastic process, where $\mathbf{g}_1 = 8$.

The mean number of cells in the genealogy of a progenitor cell, $m_i = \mathbb{E}(\mathbf{g}_i)$, for any initial compartment of interest $i \in \{1, \ldots, M+1\}$, can be obtained via first-step arguments by conditioning on the next event that occurs in the stochastic process. This approach leads to the following system of equations

$$\begin{aligned} K_1 m_1 &= L_1 m_2 + 2(\lambda_1 + a_1 + s_1), \\ K_i m_i &= L_i m_{i+1} + \xi_i m_{i-1} + 2(\lambda_i + a_i + s_i), \quad i \in \{2, \ldots, M\}, \\ K_{M+1} m_{M+1} &= \xi_{M+1} m_M + 2\lambda_{M+1}. \end{aligned}$$

The system above can be expressed in matrix form via the column vectors $\mathbf{m} = (m_1, \ldots, m_{M+1})^T$ and $\mathbf{b} = (2(\lambda_1 + a_1 + s_1), \ldots, 2(\lambda_M + a_M + s_M), 2\lambda_{M+1})^T$, as follows

$$\mathbf{J}\,\mathbf{m} = \mathbf{b}, \tag{5.9}$$

with a tri-diagonal coefficient matrix

$$\mathbf{J} = \begin{pmatrix} K_1 & -L_1 & 0 & 0 & \cdots & 0 \\ -\xi_2 & K_2 & -L_2 & 0 & \cdots & 0 \\ 0 & -\xi_3 & K_3 & -L_3 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\xi_M & K_M & -L_M \\ 0 & \cdots & 0 & 0 & -\xi_{M+1} & K_{M+1} \end{pmatrix}. \tag{5.10}$$

One can exploit the tri-diagonal structure of $\mathbf{J}$ to obtain an explicit or recursive solution of this system. In particular, by following a Gaussian forward-elimination backward-substitution approach, such as the Thomas algorithm (Conte & De Boor, 2017; Thomas, 1949), one can obtain the recursive equations

$$m_{M+1} = \rho_{M+1}, \quad m_i = \rho_i - \gamma_i m_{i+1}, \quad i \in \{1, \ldots, M\}, \tag{5.11}$$

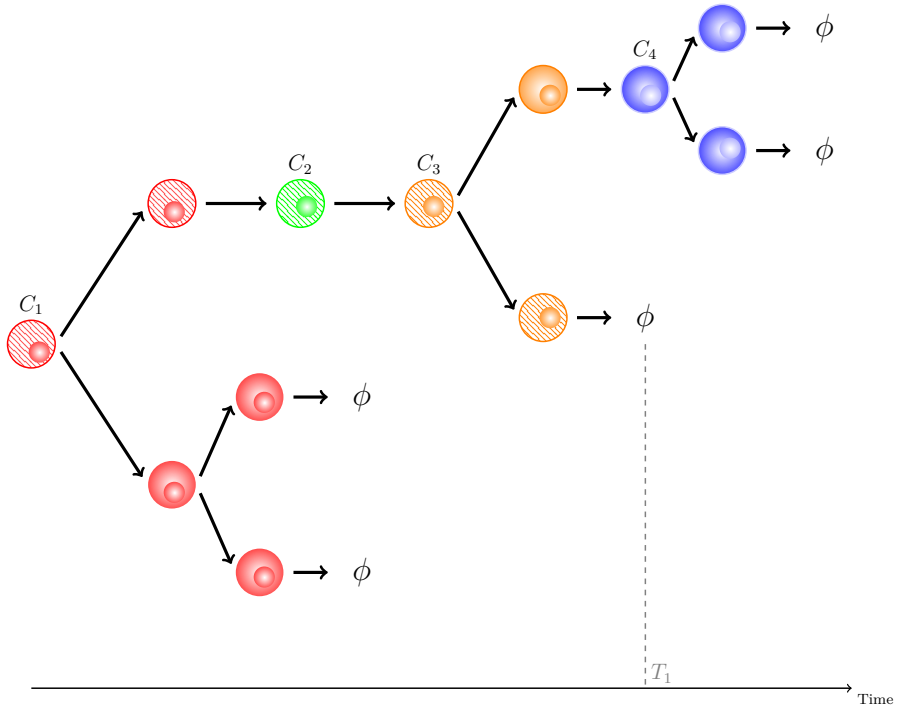


Figure 5.3: A realisation of the stochastic process following the genealogy of a single progenitor cell which starts in compartment C_1. The cell tracked (see Section 5.1.3) is depicted as striped. For each cell, the colour indicates the compartment where it is at any given time. Here, the tracked cell dies in C_3 (brown), while its genealogy continues up to C_4. In this example, $\mathbf{g}_1 = 8 = \mathbf{g}_1(1) + \mathbf{g}_1(2) + \mathbf{g}_1(3) + \mathbf{g}_1(4) = 4 + 0 + 2 + 2$.

where $\gamma_1 = -K_1^{-1}L_1$, $\rho_1 = 2K_1^{-1}(\lambda_1 + s_1 + a_1)$, and

$$\begin{aligned} \gamma_i &= -\frac{L_i}{K_i + \xi_i\gamma_{i-1}}, \quad i \in \{2, \ldots, M\}, \\ \rho_i &= \frac{2(\lambda_i + s_i + a_i) + \xi_i\rho_{i-1}}{K_i + \xi_i\gamma_{i-1}}, \quad i \in \{2, \ldots, M+1\}. \end{aligned}$$

This recursive scheme leads to the explicit solution

$$m_i = \sum_{j=i}^{M+1} (-1)^{j-i}\rho_j \left(\prod_{l=i}^{j-1}\gamma_l\right), \quad i \in \{1, \ldots, M+1\}, \tag{5.12}$$

where $\prod_{l=i}^{i-1}\gamma_l = 1$. A condition on the parameters arises during the implementation of the recursive scheme,

$$K_1 > 0, \quad K_i + \xi_i\gamma_{i-1} > 0, \quad i \in \{2,\dots,M+1\},$$

so that the mean values $m_1,\dots,m_{M+1}$ are finite and non-negative, for all $i \in \{1,\dots,M+1\}$. This ensures that the number of cells in the genealogy of a progenitor cell is finite with probability one, $\mathbb{P}(\mathbf{g}_i < +\infty) = 1$, since $m_i = \mathbb{E}(\mathbf{g}_i) = \mathbb{E}(\mathbf{g}_i|\mathbf{g}_i < +\infty)\mathbb{P}(\mathbf{g}_i < +\infty) + \mathbb{E}(\mathbf{g}_i|\mathbf{g}_i = +\infty)\mathbb{P}(\mathbf{g}_i = +\infty)$.

In the irreversible case, where $\xi_i = 0$ for $i \in \{2,\dots,M+1\}$, the solution above simplifies to

$$m_i = 2\sum_{j=i}^{M+1}\frac{\lambda_j + a_j + s_j}{K_j}\left(\prod_{l=i}^{j-1}\frac{L_l}{K_l}\right), \quad i \in \{1,\dots,M+1\}, \tag{5.13}$$

where for $j = M+1$, we set $a_{M+1} = s_{M+1} = 0$. For $i = M+1$, the empty product above is equal to one, so that $m_{M+1} = \frac{2\lambda_{M+1}}{\mu_{M+1}-\lambda_{M+1}}$. In this irreversible model the condition on the parameters for finite and non-negative solutions becomes $K_i > 0$ for all $i \in \{1,\dots,M+1\}$. Direct inspection of Figure 5.2 shows that the condition $K_i > 0$ avoids unlimited accumulation of cells in compartment C_i.

Probability generating function

Let us now go beyond the mean number of cells in a given compartment or genealogy. For the irreversible model we can consider the probability generating function of $\mathbf{g}_i$,

$$\Upsilon_i(z) = \mathbb{E}(z^{\mathbf{g}_i}) = \sum_{k=0}^{+\infty}\mathbb{P}(\mathbf{g}_i = k)z^k.$$

We recall that the variable $\mathbf{g}_i$ counts the cells in the genealogy of a progenitor cell starting in compartment C_i, arising from division events (self-renewal, asymmetric and symmetric division), and not including the progenitor cell itself. If one wants to consider the progenitor cell in the genealogy, one can define $\mathbf{s}_i \equiv \mathbf{g}_i + 1$, and denote the new generating function by $\Omega_i(z) = \mathbb{E}(z^{\mathbf{s}_i})$. The total expectation law over all possible events implies that

$$\mathbb{E}(z^{\mathbf{s}_i}) = \sum_{E_j}\mathbb{E}(z^{\mathbf{s}_i} \mid event\, E_j)\mathbb{P}(E_j), \tag{5.14}$$

with $E_j \in \{$*death, differentiation, self-renewal, asymmetric division, symmetric division*$\}$. This leads to

$$\begin{aligned}\Omega_i(z) &= \frac{\mu_i}{S_i}z + \frac{\nu_i}{S_i}z\,\Omega_{i+1}(z) + \frac{\lambda_i}{S_i}z(\Omega_i(z))^2 + \frac{a_i}{S_i}z\,\Omega_i(z)\Omega_{i+1}(z) \\ &\quad + \frac{s_i}{S_i}z(\Omega_{i+1}(z))^2,\end{aligned}$$

with $S_i = \mu_i + \nu_i + \lambda_i + a_i + s_i$. Since $\mathbf{s}_i = \mathbf{g}_i + 1$, one can write

$$\Upsilon_i(z) = \frac{\Omega_i(z)}{z},$$

so that

$$\lambda_i z^2\Upsilon_i^2(z) + (a_i z^2\Upsilon_{i+1}(z) - S_i)\Upsilon_i(z) + (\nu_i z + s_i z^2\Upsilon_{i+1}(z))\Upsilon_{i+1}(z) + \mu_i = 0. \quad (5.15)$$

We note that these probability generating functions are in agreement with the mean values obtained above. In particular, by differentiating and setting $z = 1$, we have

$$(S_i - a_i - 2\lambda_i)\Upsilon_i'(1) = 2(a_i + \lambda_i + s_i) + (\nu_i + a_i + 2s_i)\Upsilon_{i+1}'(1), \quad (5.16)$$

which can be solved recursively, leading to

$$\mathbb{E}(\mathbf{g}_i) = \Upsilon_i'(1) = 2\sum_{j=i}^{M+1}\frac{a_j + \lambda_j + s_j}{K_j}\left(\prod_{l=i}^{j-1}\frac{L_l}{K_l}\right), \quad (5.17)$$

in agreement with Equation (5.13). We also note that when $i = M + 1$, the number of cells in the genealogy of a single progenitor cell in final compartment C_{M+1} arises from a linear birth-and-death process, which in discrete-time has death probability, $p_d = \frac{\mu_{M+1}}{\mu_{M+1}+\lambda_{M+1}}$, and birth probability, $p_b = \frac{\lambda_{M+1}}{\mu_{M+1}+\lambda_{M+1}}$. Then, a first-step argument for the random variable $\mathbf{s}_{M+1}$ leads to

$$\Omega_{M+1}(z) = p_d z + p_b z\Omega_{M+1}^2(z) = z\phi_{M+1}(\Omega_{M+1}(z)),$$

with $\phi_{M+1}(z) = p_d + p_b z^2$. We can then write

$$\Upsilon_{M+1}(z) = \frac{\Omega_{M+1}(z)}{z} = \frac{z\,\phi_{M+1}(\Omega_{M+1}(z))}{z} = \phi_{M+1}(z\,\Upsilon_{M+1}(z)),$$

which has solution

$$\Upsilon_{M+1}(z) = \frac{S_{M+1} - \sqrt{S_{M+1}^2 - 4\lambda_{M+1}\mu_{M+1}z^2}}{2\lambda_{M+1}z^2}.$$

This allows one to solve Equation (5.15) recursively via backwards substitution. We finally note that by differentiating and setting $z = 1$, and since $\lambda_{M+1} < \mu_{M+1}$, one gets

$$\mathbb{E}(\mathbf{g}_{M+1}) = \Upsilon'_{M+1}(1) = \frac{2\lambda_{M+1}}{\mu_{M+1} - \lambda_{M+1}},$$

in agreement with Equation (5.13).

Compartmental analysis of the genealogy

Cells in the genealogy of a single progenitor can belong, in principle, to different compartments, as shown in Figure 5.3. For a progenitor cell starting in compartment C_i, those compartments C_j, $j \in \{1, 2, \ldots, M+1\}$ with greater proliferative potential will contribute more to $\mathbf{g}_i$. The proliferative potential of compartment j depends on the parameters $\lambda_j, a_j, s_j, \mu_j, \nu_j, \xi_j$, but also on the number of cells in the genealogy arriving into that compartment. It is of interest then to write $\mathbf{g}_i = \sum_{j=1}^{M+1} \mathbf{g}_i(j)$, with $\mathbf{g}_i(j)$ the number of cells in compartment C_j which belong to the genealogy of the progenitor cell from compartment C_i. For example, for the stochastic realisation in Figure 5.3, $\mathbf{g}_1 = \mathbf{g}_1(1) + \mathbf{g}_1(2) + \mathbf{g}_1(3) + \mathbf{g}_1(4) = 4 + 0 + 2 + 2 = 8$.

One can follow similar arguments to the ones in Section 5.1.2 to compute the mean quantities $m_i(j) \equiv \mathbb{E}(\mathbf{g}_i(j))$. In particular, for an initial compartment C_i, a first-step argument yields the following equations

$$\begin{aligned} K_i m_i(i) &= 2\lambda_i + a_i + L_i\, m_{i+1}(i) + \xi_i m_{i-1}(i), \\ K_i m_i(i+1) &= 2s_i + a_i + L_i\, m_{i+1}(i+1) + \xi_i m_{i-1}(i+1), \\ K_i m_i(j) &= L_i m_{i+1}(j) + \xi_i m_{i-1}(j), \quad j \notin \{i, i+1\}, \end{aligned}$$

where we implicitly set $\xi_1 = 0$, and $\forall j \in \{1, 2, \ldots, M+1\}$, $m_{M+2}(j) = m_{M+1}(M+2) = m_0(j) = 0$ for notational convenience. Making use of a recursive approach one can show that for any $j \in \{1, \ldots, M+1\}$,

$$m_{M+1}(j) = \rho_{M+1}(j), \quad m_i(j) = \rho_i(j) - \gamma_i m_{i+1}(j), \quad i \in \{1, \ldots, M\},$$

where $\gamma_1 = -K_1^{-1}L_1$, $\rho_1(j) = K_1^{-1}d_{(1,j)}$, and

$$\begin{aligned}\gamma_i &= -\frac{L_i}{K_i + \xi_i\gamma_{i-1}}, \qquad i \in \{2,\dots,M\},\\ \rho_i(j) &= \frac{d_{(i,j)} + \xi_i\rho_{i-1}(j)}{K_i + \xi_i\gamma_{i-1}}, \quad i,j \in \{2,\dots,M+1\},\end{aligned}$$

with

$$d_{(i,j)} = \begin{cases} 2\lambda_i + a_i, & \text{if} \quad j = i,\\ 2s_i + a_i, & \text{if} \quad j = i+1,\\ 0, & \text{otherwise.}\end{cases}$$

This recursive scheme leads to the solution

$$m_i(j) = \sum_{k=i}^{M+1}(-1)^{k-i}\rho_k(j)\left(\prod_{p=i}^{k-1}\gamma_p\right), \tag{5.18}$$

where $\prod_{p=i}^{i-1}\gamma_p = 1$. This expression simplifies further for the irreversible model, since $\xi_i = 0$ for all $i \in \{2,\dots,M+1\}$. In this instance, $m_i(j) = 0$ whenever $i > j$, and we can write

$$\begin{aligned}m_i(i) &= K_i^{-1}(2\lambda_i + a_i),\\ m_i(j) &= K_{j-1}^{-1}\left(\prod_{p=i}^{j-2}K_p^{-1}L_p\right)\left(d_{(j-1,j)} + d_{(j,j)}K_j^{-1}L_{j-1}\right), \quad j \in \{i+1,...,M+1\},\end{aligned}$$

for any $i \in \{1,...,M\}$. Finally, we note that $\prod_{p=i}^{i-1}K_p^{-1}L_p = 1$, and $m_{M+1}(M+1) = \frac{2\lambda_{M+1}}{K_{M+1}}$.

5.1.3 Single-cell analysis

In the previous section we have analysed summary statistics of the genealogy of a progenitor cell. We now turn to a number of summary statistics related to the lifespan of a single cell, extending the single-cell analysis originally proposed by de la Higuera *et al.* (2019), which was focused on T cell migration. In what follows we assume that no symmetric or asymmetric division occurs, so that $s_i = a_i = 0$ for all $i \in \{1,\dots,M+1\}$. We propose to track the dynamics of a

single cell, starting in compartment C_i, until it dies. In any given compartment, this cell can divide (and, since both daughter cells are indistinguishable, we keep tracking one of them), move to the next or previous compartment, or die. The dynamics can then be represented by the CTMC $\mathcal{Y} = \{\mathbf{Y}(t) : \ t \geq 0\}$ over the state space $\{C_1, C_2, \ldots, C_{M+1}, \emptyset\}$, where $\mathbf{Y}(t)$ represents the state of the cell at time t; the cell is either in some compartment, C_j, or it has died (state $\emptyset$). A schematic representation of the process $\mathcal{Y}$ is given in Figure 5.4, and a particular realisation of this stochastic process is shown in Figure 5.3, where the tracked cell is shown with stripes.

We study the stochastic process with the following summary statistics:

- the lifespan, T_i, of the cell, starting in compartment C_i,

$$T_i \ = \ \inf\{t \geq 0 : \ \mathbf{Y}(t) = \emptyset \,|\, \mathbf{Y}(0) = C_i\},$$

  which quantifies the survival potential of cells in the system depending on their initial state,

- the number of divisions carried out by the cell during its lifespan, D_i, which quantifies the proliferative capacity of cells according to their initial state, and

- the probability of the cell to die in a given compartment C_j; that is, $\beta_i(j) = \mathbb{P}(\mathbf{Y}(T_i - \Delta t) = C_j)$ for a small enough Δt.

Lifespan of a single cell

Let T_i be the lifespan of a single cell starting in compartment C_i, $i \in \{1, \ldots, M+1\}$,

$$T_i \ = \ \inf\{t \geq 0 : \mathbf{Y}(t) = \emptyset \,|\, \mathbf{Y}(0) = C_i\},$$

and consider $\tau_i = \mathbb{E}(T_i)$, its average lifespan. By conditioning on the next event that the cell can undergo in this process, we obtain the following recursive equa-

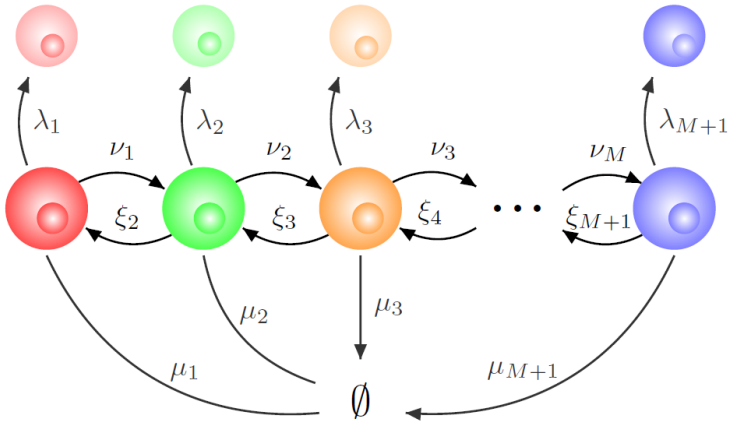


Figure 5.4: Representation of the process $\mathcal{Y}$ to follow the fate of a single cell.

tions

$$\begin{aligned}(\mu_1+\nu_1)\tau_1 &= \nu_1\tau_2+1,\\ (\mu_i+\nu_i+\xi_i)\tau_i &= \nu_i\tau_{i+1}+\xi_i\tau_{i-1}+1, \quad i\in\{1,\dots,M\},\\ (\mu_{M+1}+\xi_{M+1})\tau_{M+1} &= \xi_{M+1}\tau_M+1.\end{aligned}$$

These equations have the general solution

$$\tau_i=\sum_{k=i}^{M+1}(-1)^{k-i}\bar{\rho}_k\left(\prod_{p=i}^{k-1}\bar{\gamma}_p\right),$$

where $\prod_{p=i}^{i-1}\bar{\gamma}_p=1$. We have made use of the notation $\bar{\gamma}_1=-\bar{K}_1^{-1}\nu_1$, $\bar{\rho}_1=\bar{K}_1^{-1}$, with

$$\begin{aligned}\bar{\gamma}_i &= -\frac{\nu_i}{\bar{K}_i+\xi_i\bar{\gamma}_{i-1}}, \quad i\in\{2,\dots,M\},\\ \bar{\rho}_i &= \frac{1+\xi_i\bar{\rho}_{i-1}}{\bar{K}_i+\xi_i\bar{\gamma}_{i-1}}, \quad i\in\{2,\dots,M+1\},\end{aligned}$$

where $\bar{K}_i=\mu_i+\nu_i+\xi_i$. A simpler solution is obtained in the irreversible model ($\xi_i=0$ for all compartments), given by

$$\tau_i=\sum_{k=i}^{M+1}\frac{1}{\mu_k+\nu_k}\left(\prod_{p=i}^{k-1}\frac{\nu_p}{\mu_p+\nu_p}\right), \quad i\in\{1,\dots,M+1\},$$

where $\prod_{p=i}^{i-1} \frac{\nu_p}{\mu_p+\nu_p} = 1$, and we set $\nu_{M+1} = 0$ for notational convenience.

A similar approach allows one to compute the Laplace-Stieltjes transform of T_i, and any of its higher order moments (Castro *et al.*, 2018). For example, in the irreversible model, the second order moment of the lifespan of a cell starting in compartment C_i is given by

$$\mathbb{E}[T_i^2] = \sum_{j=i}^{M+1} R_j \left(\prod_{r=i}^{j-1} \frac{\nu_r}{\mu_r + \nu_r} \right), \quad i \in \{1, \ldots, M\},$$

where $\prod_{p=i}^{i-1} \frac{\nu_r}{\mu_r+\nu_r} = 1$, $R_i = \frac{2(\nu_i \tau_{i+1}+1)}{(\mu_i+\nu_i)^2}$, and $R_{M+1} = \frac{2}{\mu_{M+1}^2}$. We note that if the cell starts in the product compartment C_{M+1}, $\mathbb{E}(T_{M+1}^2) = 2\mu_{M+1}^{-2}$ since $T_{M+1} \sim \text{Exp}(\mu_{M+1})$.

Number of divisions during the lifespan of a single cell

We denote by D_i the number of division events performed by the tracked cell during its lifespan, starting in compartment C_i. One can compute its average value, $\eta_i = \mathbb{E}(D_i)$, via first-step arguments, which lead to

$$\begin{aligned}
(\lambda_1 + \mu_1 + \nu_1)\eta_1 &= \lambda_1(\eta_1 + 1) + \nu_1\eta_2, \\
(\lambda_i + \mu_i + \nu_i + \xi_i)\eta_i &= \lambda_i(\eta_i + 1) + \nu_i\eta_{i+1} + \xi_i\eta_{i-1}, \quad i \in \{1, \ldots, M\}, \\
(\lambda_{M+1} + \mu_{M+1} + \xi_{M+1})\eta_{M+1} &= \lambda_{M+1}(\eta_{M+1} + 1) + \xi_{M+1}\eta_M,
\end{aligned}$$

with solution

$$\eta_i = \sum_{k=i}^{M+1} (-1)^{k-i} \tilde{\rho}_k \left(\prod_{p=i}^{k-1} \bar{\gamma}_p \right), \tag{5.19}$$

where $\prod_{p=i}^{i-1} \bar{\gamma}_p = 1$, $\tilde{\rho}_1 = \bar{K}_1^{-1}\lambda_1$, and

$$\tilde{\rho}_i = \frac{\lambda_i + \xi_i\tilde{\rho}_{i-1}}{\bar{K}_i + \xi_i\bar{\gamma}_{i-1}}, \quad i \in \{2, \ldots, M+1\},$$

with $\bar{K}_i$ and $\bar{\gamma}_i$ defined as above. In the irreversible model, this expression simplifies to

$$\tau_i = \sum_{k=i}^{M+1} \frac{\lambda_k}{\mu_k + \nu_k} \left(\prod_{p=i}^{k-1} \frac{\nu_p}{\mu_p + \nu_p} \right), \quad i \in \{1, \ldots, M\},$$

where $\prod_{p=i}^{i-1} \frac{\nu_p}{\mu_p+\nu_p} = 1$. We note that in this case $\eta_{M+1} = \lambda_{M+1}\mu_{M+1}^{-1}$ since $D_{M+1} \sim$ Geometric $\left(\frac{\lambda_{M+1}}{\mu_{M+1}+\lambda_{M+1}}\right)$.

These division events can occur at any time during the lifespan of the cell, which is visiting different compartments over time. Thus, one can quantify the proliferation potential of the cell during its eventual visit to each compartment by considering $D_i = \sum_{j=1}^{M+1} D_i(j)$, where $D_i(j)$ is the number of divisions which occur exactly in compartment C_j during the lifespan of the cell, which started in compartment C_i. The average values $\eta_i(j) \equiv \mathbb{E}(D_i(j))$ can be computed (again) with first-step arguments. For instance, in the irreversible model, one has

$$\begin{aligned} \eta_i(i) &= \frac{\lambda_i}{\mu_i+\nu_i}, \\ \eta_i(j) &= \frac{\lambda_j}{\mu_j+\nu_j}\left(\prod_{k=i}^{j-1}\frac{\nu_k}{\mu_k+\nu_k}\right), \quad j \geq i+1, \end{aligned}$$

for $i \in \{1,\ldots,M+1\}$ and $\eta_{M+1}(M+2) = 0$. We note that the expression above is consistent with the interpretation that, in the irreversible model, $D_i(j) \sim$ Geometric $\left(\frac{\lambda_j}{\lambda_j+\nu_j+\mu_j}\right)$ conditioned on the arrival of the cell to compartment C_j. In general, one can write

$$\begin{aligned} \eta_i(j) &= \mathbb{E}(D_i(j)) = \mathbb{E}(D_i(j) \mid \textit{cell ever visits } C_j)\,\mathbb{P}(\textit{cell ever visits } C_j) \\ &+ \mathbb{E}(D_i(j) \mid \textit{cell never visits } C_j)\,\mathbb{P}(\textit{cell never visits } C_j), \end{aligned}$$

and, since $\mathbb{E}(D_i(j) \mid \textit{cell never visits } C_j) = 0$, the quantity $\eta_i(j)$ accounts for the probability of the cell not visiting this compartment.

More generally, one can go beyond the computation of mean values and consider the probability distribution of D_i, defined by $\omega_i(n) \equiv \mathbb{P}(D_i = n)$, the probability that a single cell starting in compartment C_i divides exactly n times before it dies or leaves the system, for any integer $n \geq 0$. This can be computed recursively (following our previous arguments). One can show that

$$\omega_i(n) = \sum_{k=i}^{M+1}(-1)^{k-i}\hat{\rho}_k(n)\left(\prod_{j=i}^{k-1}\hat{\gamma}_j\right), \quad i \in \{1,\ldots,M+1\},\ n \geq 0,$$

where $\prod_{j=i}^{i-1} \hat{\gamma}_j = 1$. We have introduced the notation $\hat{\rho}_1(n) = (\lambda_1\omega_1(n-1) + \mu_1\mathbb{1}_{n=0})\hat{K}_1^{-1}$, $\hat{\gamma}_1 = -\nu_1\hat{K}_1^{-1}$, and

$$\hat{\rho}_i(n) = \frac{\lambda_i\omega_i(n-1) + \mu_i\mathbb{1}_{n=0} + \xi_i\hat{\rho}_{i-1}(n)}{\hat{K}_i + \xi_i\hat{\gamma}_{i-1}}, \quad \hat{\gamma}_i = \frac{-\nu_i}{\hat{K}_i + \xi_i\hat{\gamma}_{i-1}},$$

with $\hat{K}_i = \mu_i + \lambda_i + \xi_i + \nu_i$. We note that $\omega_i(-1) = 0$ and that $\mathbb{1}_{n=0}$ is the indicator function, equal to 1 if $n = 0$, and zero otherwise. Thus, the probabilities $\omega_i(n)$ can be computed recursively for increasing values of n, since $\omega_i(n)$ depends on $\omega_i(n-1)$.

Cellular death

Finally, one can compute the probability of the tracked cell to die in a given compartment, which we denote by

$$\beta_i(j) \;=\; \mathbb{P}(\mathbf{Y}(T_i - \Delta t) = C_j),$$

for small enough Δt, and for any $i, j \in \{1, \ldots, M+1\}$. Once again, a first-step argument leads to the following recursive relationship

$$\begin{aligned}
(\mu_1 + \nu_1)\beta_1(j) &= \nu_1\beta_2(j) + \mu_1\mathbb{1}_{j=1}, \\
(\mu_i + \nu_i + \xi_i)\beta_i(j) &= \nu_i\beta_{i+1}(j) + \xi_i\beta_{i-1}(j) + \mu_i\mathbb{1}_{i=j}, \\
(\mu_{M+1} + \xi_{M+1})\beta_{M+1}(j) &= \xi_{M+1}\beta_M(j) + \mu_{M+1}\mathbb{1}_{j=M+1},
\end{aligned}$$

with solution

$$\beta_i(j) \;=\; \sum_{k=i}^{M+1}(-1)^{k-i}\bar{\rho}_k(j)\left(\prod_{p=i}^{k-1}\bar{\gamma}_p\right) \quad i, j \in \{1, \ldots, M+1\},$$

where $\bar{\rho}_1(j) = \mu_1\bar{K}_1^{-1}\mathbb{1}_{j=1}$ for $j \in \{1, \ldots, M+1\}$, and

$$\bar{\rho}_i(j) = \frac{\mu_i\mathbb{1}_{i=j} + \xi_i\bar{\rho}_{i-1}(j)}{\bar{K}_i + \xi_i\bar{\gamma}_{i-1}}, \quad i \in \{2, \ldots, M+1\},\ j \in \{1, \ldots, M+1\}.$$

For the irreversible model, this simplifies to

$$\beta_i(i) \;=\; \frac{\mu_i}{\mu_i + \nu_i}, \quad \beta_i(j) \;=\; \frac{\mu_j}{\mu_j + \nu_j}\prod_{k=i}^{j-1}\frac{\nu_k}{\mu_k + \nu_k}, \quad j \in \{i+1, \ldots, M+1\},$$

for $i \in \{1, \ldots, M+1\}$.

We conclude this analysis with a note. A particular advantage of deriving analytical expressions for these summary statistics is that they allow one to explicitly compute sensitivities, $\partial\beta_i(j)/\partial\theta$, or elasticities, $(\partial\beta_i(j)/\partial\theta)/(\beta_i(j)/\theta)$, with respect to model parameters of interest, θ (see Definition 2.22). This can be rather useful when considering a complex model (with many parameters), as illustrated in Section 5.2.3. A local sensitivity analysis of this kind provides a quantification of the impact that small perturbations of model parameters can have on a given summary statistics of interest. This is especially relevant when a subset of parameter values are being estimated from experimental data sets, and thus, will have inherent uncertainties.

5.2 Results

We illustrate our methods with three case studies. In Section 5.2.1, we implement the methods from Section 5.1.1 and Section 5.1.2 to explore the impact of asymmetric and symmetric division events, for the specific case of $M = 3$ compartments. We perform sensitivity analysis for the probabilities of self-renewal, asymmetric and symmetric division. The impact of asymmetric and symmetric division is further analysed in Section 5.2.2, where we consider hematopoietic stem cells, in light of recent experimental data and a mathematical model proposed by Barile *et al.* (2020). Finally, in Section 5.2.3 we apply the single-cell analysis from Section 5.1.3 to an existing model of thymic T cell development (Sawicka *et al.*, 2014).

5.2.1 Asymmetric and symmetric division: the case of four compartments

Let us consider the case $M = 3$, for illustrative purposes, where the last compartment, C_4, of product cells does not involve any death, division or differentiation events ($\mu_4 = \lambda_4 = \xi_4 = 0$), to represent the terminal accumulation of cells in it. The same assumption has been done in Chapters 3 and 4 for a sequence of M compartments; in this case, it allows us to quantify the number of cells that

exit the system formed by the first three compartments, which can be of biological interest in processes such as thymic development (Sawicka *et al.*, 2014). We choose $\mu_i = 1 = \mu$ for all $i \in \{1, 2, 3\}$, so that the unit of time for the system is the mean lifetime of a cell. We now want to study the impact of asymmetric and symmetric division on the dynamics, and thus, set $\nu_1 = \nu_2 = \nu_3 = 1/2$ in the irreversible model.

Cells can divide in each compartment at a rate H; this *division rate* represents self-renewal with probability p_{SR}, asymmetric division with probability p_{AD}, and symmetric division with probability p_{SD}. This is equivalent to setting, with the notation introduced in Section 5.1, $\lambda_i = p_{SR}H$, $s_i = p_{SD}H$ and $a_i = p_{AD}H$, for $i \in \{1, 2, 3\}$. We choose $H = 0.9 < 1.0 = \mu$, so that the system has significant proliferative potential, and focus on the following scenarios of interest:

Only-SR. $p_{SR} = 1.0$, $p_{SD} = 0$, $p_{AD} = 0$.

Dominant-SR. $p_{SR} = 0.8$, $p_{SD} = 0.1$, $p_{AD} = 0.1$.

Dominant-SD. $p_{SR} = 0.1$, $p_{SD} = 0.8$, $p_{AD} = 0.1$.

Dominant-AD. $p_{SR} = 0.1$, $p_{SD} = 0.1$, $p_{AD} = 0.8$.

Our aim in what follows is to explore the impact that asymmetric or symmetric division has on the dynamics of the system (dominant-SR/SD/AD scenarios), compared to the situation where only self-renewal takes place (only-SR scenario). In Figure 5.5 we plot the mean number of cells, $\mathbb{E}(\mathbf{C}_i(t))$, in compartments $i \in \{1, 2, 3, 4\}$ for each scenario. For compartments C_i with $i \in \{1, 2, 3\}$, and since $K = K_1 = K_2 = K_3$ and $L = L_1 = L_2 = L_3$, one can directly use Equation (5.6). For compartment C_{M+1}, with $M = 3$, one has

$$\mathbb{E}(\mathbf{C}_{M+1}(t)) = \left(\frac{L}{K}\right)^M - \sum_{j=1}^{M} \mathbb{E}(\mathbf{C}_j(t)) \left(\frac{L}{K}\right)^{M+1-j}.$$

In Figure 5.5, we consider initial conditions $\mathbf{C}_1(0) = 10^2$, $\mathbf{C}_2(0) = \mathbf{C}_3(0) = \mathbf{C}_4(0) = 0$, representing 10^2 initial cells in the first compartment and no cells in the other compartments. We observe that an exponential decay in the number of cells in C_1 is followed by sequential increases in the subsequent compartments

until a steady state number of cells is achieved in the product state of compartment C_4. Interestingly, the dynamics of the system is faster once symmetric or asymmetric division is considered (dominant-SR/SD/AD scenarios compared to only-SR): the decay in compartment C_1 is quicker and the steady state is reached faster. The fastest dynamics is observed for the dominant-SD case, where symmetric division is more likely, and the two daughters of a cell move directly to the next compartment. We note that symmetric or asymmetric division does not only affect the dynamics but also the total average number of cells exiting the system in Figure 5.5 (*i.e.,* reaching the product state in compartment C_4). In particular, $\lim_{t\to+\infty} \mathbb{E}(\mathbf{C}_4(t))$ is significantly larger when asymmetric and (especially) symmetric division can occur. We note that, importantly, the per cell division rate, H, is equal in all four scenarios. This suggests that, in this type of systems, asymmetric or symmetric division (compared to self-renewal proliferation) facilitates the generation of a larger population of product cells with the same overall *proliferative capacity*; that is, in the only-SR scenario, a larger number of divisions would be required in each compartment for enough cells to escape death and differentiate to the next compartment, and to eventually reach C_4.

Our comments above are consistent with the results shown in Figure 5.6, where we plot the mean number of cells, $m_1(j)$, in the genealogy of a single cell starting in C_1, belonging to compartments C_j, $j \in \{1, 2, 3, 4\}$, for all four scenarios. In the only-SR scenario, the mean number of cells in the genealogy of the progenitor cell decreases monotonically across the sequence of compartments, $m_1(1) > m_1(2) > m_1(3) > m_1(4)$. We note that $m_1(4) = 0$ can be explained since it only accounts for progeny cells which arrive into compartment C_4 as a direct result of cell proliferation, and no symmetric or asymmetric division is considered in the Only-SR scenario. We also stress here that this monotonic decrease happens even though the division and differentiation rates are equal in all compartments $j \in \{1, 2, 3\}$. This can be explained by the fact that some cells in the genealogy will die before reaching compartments C_2 or C_3. On the other hand, and as discussed for Figure 5.5, scenario only-SR leads to the largest mean progeny, $m_1 = m_1(1) + m_1(2) + m_1(3) + m_1(4)$. This suggests that the only-SR scenario is an inefficient way to reach a desired population size of product cells. Asymmetric and (especially) symmetric division events significantly reduce

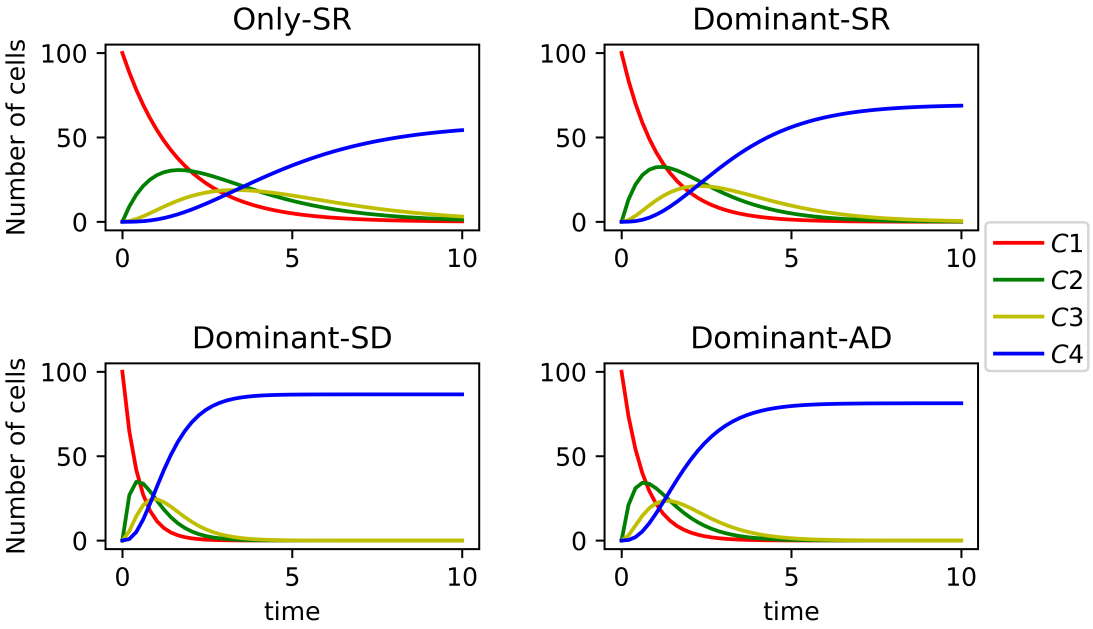


Figure 5.5: Dynamics of the mean number of cells, $\mathbb{E}(\mathbf{C}_i(t))$, in compartments $i \in \{1, 2, 3, 4\}$ for each of the four scenarios considered in this section.

the number of descendants from a single progenitor cell in all compartments, while maximising the number of product cells (see Figure 5.5). In particular, the dominant-SD scenario is characterised by the highest total mean number of product cells, $\mathbb{E}(\mathbf{C}_4(+\infty))$, as well as the smallest progeny size, m_1, while leading to the largest progeny, $m_1(4)$, in the product compartment.

Finally, the fastest dynamics observed in scenarios with symmetric division, as well as the reduced progeny observed from a single progenitor in C_1, imply that this kind of systems can go from uncontrolled cellular growth to population extinction for late times, by increasing the number of symmetric division events. We explore this further in the next section looking at a particular case study.

5.2.2 Hematopoietic stem cells: self-renewal, asymmetric and symmetric division

We consider the model proposed in (Barile *et al.*, 2020, Figure 6A) for the gradual differentiation of hematopoietic stem cells (HSCs) which are responsible for the

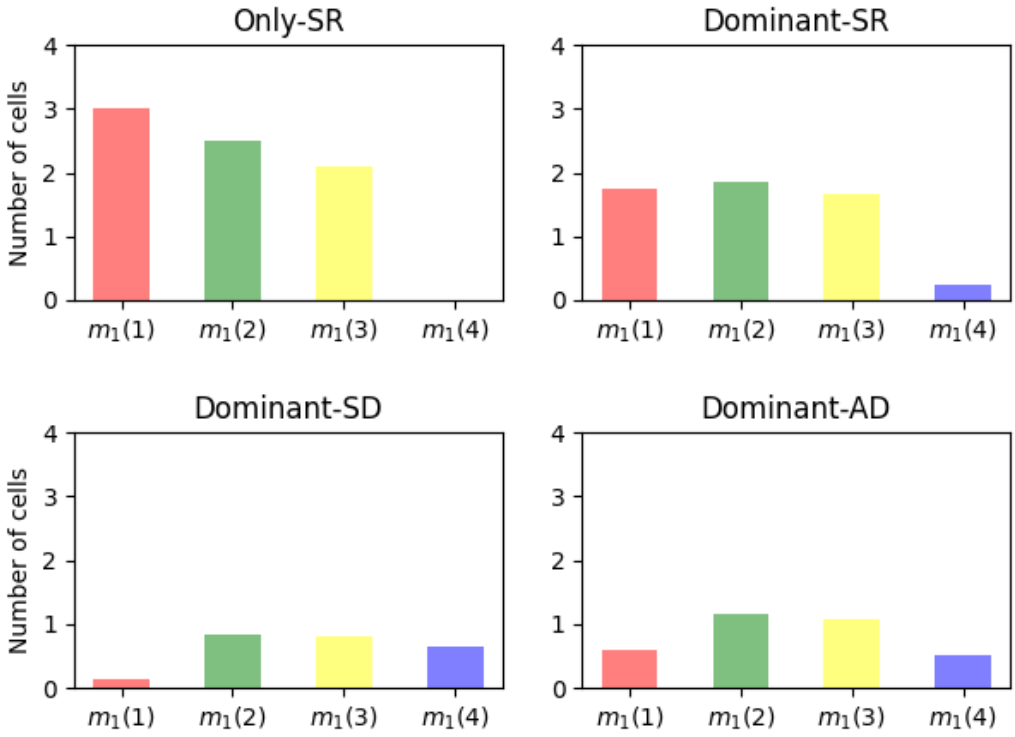


Figure 5.6: Mean number of cells $m_1(j)$ in the genealogy of a single cell starting in C_1, belonging to compartments C_j, $j \in \{1, 2, 3, 4\}$, for all four scenarios.

production of all blood cells (Seita & Weissman, 2010). In order to maintain such an heterogeneous population, HSCs are characterised by the ability to self-renew and differentiate. Moreover, HSCs need to act continuously and rapidly to either replace short-lived blood cells or respond to hematopoietic stress arising from events such as bleeding and toxin spotting (Wilson *et al.*, 2009). Recent advances in flow cytometry and single-cell analysis have shown that HSC cells are a small population compared to the many other cell types that they can generate in the blood. Despite their importance, the molecular mechanisms involved in hematopoietic stem cell maintenance remain unclear.

In order to study simultaneously HSC proliferation and differentiation, Barile *et al.* (2020) propose a novel mathematical model inferred by cell-cycle dependent labelling and HCS fate mapping data. Cells at each state can undergo five dif-

ferent processes: self-renewal, asymmetric cell division, symmetric cell division, direct differentiation and cell death. This leads to the following compartmental sequence, reported in (Barile *et al.*, 2020, Figure 6A): HSC1 $\rightarrow$ HSC2 $\rightarrow$ MPP1+2 $\rightarrow$ MPP3 $\rightarrow$ HPC1 consisting of 5 different compartments (that is $M = 4$). These represent two stages of hematopoietic stem cells (HSC1 and HSC2), two stages of multipotent progenitor cells (MPP1+2 and MPP3), and a last stage of hematopoietic progenitor cells (HPC1).

	K_i	L_i
HSC1	-0.0046197	0.016497
HSC2	0.0017357	0.007847
MPP1+2	0.0044844	0.032834
MPP3	0.01556	0.16113
HPC1	0.0293	0

Table 5.1: Parameter values obtained from (Barile *et al.*, 2020, Figure 6), where they set $a_i = s_i = 0$ for all $i \in \{1, \dots, 5\}$. All parameters reported in this section have time units of $days^{-1}$.

The model proposed by Barile *et al.* (2020) corresponds to the irreversible model (*i.e.*, $\xi_i = 0$ for $i \in \{2, \dots, 5\}$) shown in Figure 5.2, when we set $M = 4$ and consider the parameter values in Table 5.1, directly obtained from (Barile *et al.*, 2020, Figure 6). Recall $K_i \equiv \mu_i + \nu_i + s_i - \lambda_i$; we note that the *net loss rate* for HSC1 cells, $K_1 = -4.6197 \times 10^{-3}$, is negative. Thus the growth of the HSC1 compartment is unbounded, as can be observed in (Barile *et al.*, 2020, Figure S1 H). This is also shown in Figure 5.7, where we simulate the system given by Equation (5.2) for parameters in Table 5.1. We note that the parameter calibration performed by Barile *et al.* (2020) predicted relatively negligible symmetric and asymmetric division rates for most of the compartments. Thus, for simplicity, they assumed $s_i = a_i = 0$ for all $i \in \{1, \dots, 5\}$, as considered in Table 5.1 and Figure 5.7. We now want to explore, for this system, the potential role that symmetric and asymmetric division could play. To that end, we perform a sensitivity analysis on the parameters (s_i, a_i), for $i \in \{1, \dots, 5\}$. In particular, we consider four different scenarios:

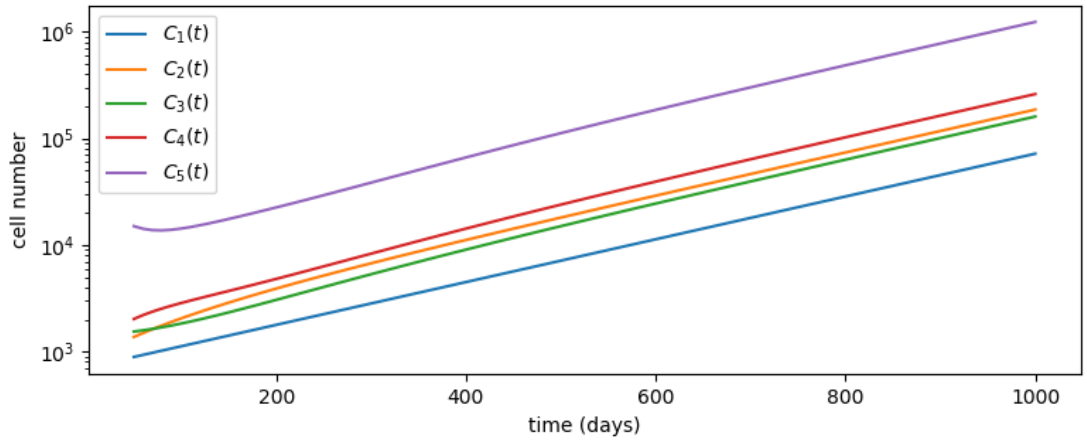
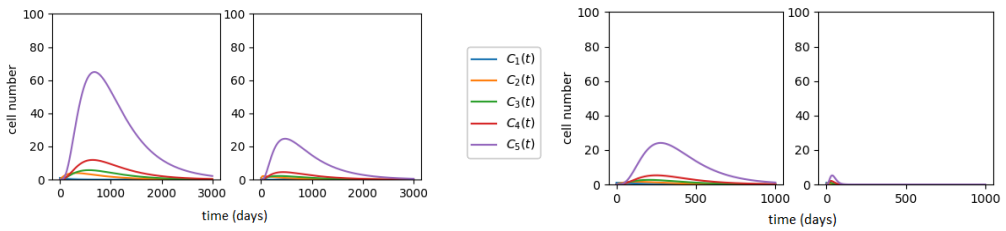


Figure 5.7: Dynamics of the ODEs System (5.2) and parameters from Table 5.1, corresponding to the only-SR scenario. We set initial conditions $(\mathbf{C}_1(0), \mathbf{C}_2(0), \mathbf{C}_3(0), \mathbf{C}_4(0), \mathbf{C}_5(0)) = (890, 1370, 1540, 2020, 1.5 \times 10^4)$, taken from (Barile *et al.*, 2020, Figure S1 H).

- **Only-SR.** All parameters as in Table 5.1, where $s_i = a_i = 0$ for $i \in \{1, \ldots, 5\}$ and division events correspond to self-renewal, as reported by Barile *et al.* (2020), and shown in Figure 5.7.
- **Symm1.** Symmetric division rate $s_1 \in \{10^{-2}, 10^{-1}\}$ for compartment HSC1, is added to the rates in Table 5.1, with $s_i = 0$ for $i \in \{2, \ldots, 5\}$, and $a_i = 0$ for $i \in \{1, \ldots, 5\}$.
- **SymmAll.** Equal symmetric division rates $s_i \in \{10^{-2}, 10^{-1}\}$ for $i \in \{1, \ldots, 5\}$ are added to the rates in Table 5.1, with $a_i = 0$ for $i \in \{1, \ldots, 5\}$.
- **AsymmAll.** Equal asymmetric division rates $a_i \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, $i \in \{1, \ldots, 5\}$ are added to the rates in Table 5.1. Symmetric division only occurs in compartment HSC1, $s_1 = 5 \times 10^{-3}$, chosen so that population extinction is guaranteed.

The net loss rate for compartment i, $K_i \equiv \mu_i + \nu_i + s_i - \lambda_i$, does not depend on the asymmetric division rate, a_i, but it does depend on the symmetric rate, s_i. Thus, just by tuning the symmetric division rate, s_1, of HSC1 cells (Symm1 scenario), we can drastically change the dynamics of the entire system: from infinite cell growth

(Figure 5.7) to population extinction (Figure 5.8a). Furthermore, as shown in Figure 5.8b, if all the compartments are characterised by a non-zero symmetric division rate (SymmAll scenario), then the resulting population size is smaller but the overall dynamics faster. The change in the asymptotic behaviour of the system, from growth to extinction, and the smaller number of HSC cells observed in the SymmAll scenario compared to the Symm1 one, can be explained by the transitions across compartments generated by symmetric division events. When a symmetric division event occurs in compartment C_i, two cells in the subsequent compartment C_{i+1} are generated, while the number of cells in compartment C_i decreases by one. Thus, symmetric division events speed up the transition of cells to subsequent compartments and, at the same time, deplete cells from the compartment where the division took place.



(a) Symm1 scenario for $s_1 = 10^{-2}$ (left) and $s_1 = 10^{-1}$ (right).

(b) SymmAll scenario for $s_i = 10^{-2}$ (left) and $s_i = 10^{-1}$ (right) with $i \in \{1, \ldots, 4\}$.

Figure 5.8: Dynamics of ODEs System (5.2) with initial conditions $(\mathbf{C}_1(0), \mathbf{C}_2(0), \mathbf{C}_3(0), \mathbf{C}_4(0), \mathbf{C}_5(0)) = (1, 0, 0, 0, 0)$. Parameter values as in Table 5.1, except for symmetric division rates. (a) Scenario Symm1, where the symmetric division rate s_1 is positive only in the HSC1 compartment. (b) Scenario SymmAll, where the symmetric division rate s_i is positive and identical for all compartments.

In Figure 5.9 we explore the impact of increasing the asymmetric division rates, by considering the AsymmAll scenario, where those rates are positive and identical for all compartments. We note that, since asymmetric division rates do not affect the asymptotic qualitative behaviour of the system (*i.e.*, growth versus extinction), we set $s_1 = 5 \times 10^{-3}$, so that population extinction is guaranteed. When comparing the AsymmAll, Symm1 and SymmAll scenarios, it is clear that increasing the asymmetric division rates leads to a greater number of cells

across compartments. This is due to the fact that asymmetric division events increase the subsequent compartments without depleting the compartment where the event takes place. In practice, for situations where population extinction is guaranteed at late times (Figure 5.9), increasing the asymmetric division rates delays the time when extinction occurs. We note Figure 5.9 differs from Figure 5.6, but they are not in contradiction. In Figure 5.6, the division rate H is kept constant, while the probability for each kind of division event is not, but given by (p_{SR}, p_{SD}, p_{AD}). In Figure 5.9, we increase the asymmetric division rate in each compartment instead, effectively increasing the division rate H, and thus leading to a greater population size over time in the system.

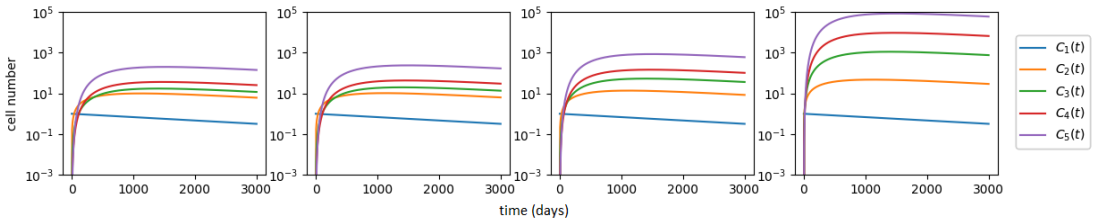


Figure 5.9: Dynamics of the ODEs System (5.2) and parameters values as in Table 5.1, except for $s_1 = 5 \times 10^{-3}$ and asymmetric division rates. AsymmAll scenario with equal asymmetric division rates $a_i \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ in all compartments, and initial conditions $(\mathbf{C}_1(0), \mathbf{C}_2(0), \mathbf{C}_3(0), \mathbf{C}_4(0), \mathbf{C}_5(0)) = (1, 0, 0, 0, 0)$.

Studies suggest that very low numbers of HSC cells (HSC1 and HSC2) are able to maintain a continuous stream of differentiating cells generating a huge number of mature blood cells (Busch *et al.*, 2015; Seita & Weissman, 2010). During haematopoiesis, HSCs cells slowly replace short-living MPP cells; this heterogeneous population no longer possess self-renewal ability but still retain differentiation potential (Seita & Weissman, 2010). We note here that the parameter values estimated by Barile *et al.* (2020) result in an almost zero net loss rate for HSC2s and MPP1+2 cells; that is, $K_2 \approx K_3 \approx 0$. This agrees with the hypothesis that the self-renewal rate of HSC2 and MPP1+2 is sufficient to maintain, alone, the populations of more differentiated cells (e.g MPP3, HPC1), with minimal input from HSC1 cells. Thus, it is pertinent to study the genealogy of a single HSC

(HSC1 or HSC2), and study how symmetric and asymmetric division events influence it. To do this, we implement Equation (5.13) for the irreversible model for $i \in \{1, 2\}$, $j \in \{i, \dots, 5\}$ and compute $m_i(j)$, the mean number of cells within the genealogy of a single HSC (HSC1 or HSC2) in subsequent compartments C_j, $j \in \{i, \dots, 5\}$. The results are shown in Figure 5.10, where we plot $m_i(j)$ for the Symm1, SymmAll and AsymmAll scenarios for $i = 1, 2$. We note that the Symm1 scenario is not considered in Figure 5.10, since changes in the symmetric division rate, s_1, do not affect $m_2(j)$. First, we observe that the mean number of cells in the genealogy of a single HSC1 progenitor across compartments, $m_1(j)$, increases for increasing values of j (that is, for more differentiated cells) regardless of the scenario. Indeed, most cells within the genealogy of a single HSC1 progenitor belong to the last HPC compartment, consistent with the dynamics observed in Figure 5.8 and Figure 5.9. However, the trend of $m_1(j)$ is drastically different for scenario AsymmAll compared to scenarios Symm1 or SymmAll, for increasing values of the corresponding rate (s_1 in Symm1, s_k, $k \in \{1, \dots, M+1\}$ in SymmAll, and a_k, $k \in \{1, \dots, M+1\}$ in AsymmAll). In the AsymmAll scenario, $m_1(j)$ is an increasing function of the asymmetric division rate a_k, whereas in the Symm1 and SymmAll scenarios, $m_1(j)$ is a decreasing function of the symmetric division rate. This agrees with the dynamics shown in Figure 5.8, where increasing values of the symmetric division rate prevent population growth, guaranteeing extinction at late times. An increase of the asymmetric division rate leads to significant production of MPP and HPC cells (see $m_1(4)$ and $m_1(5)$ in Figure 5.10) within the genealogy, and thus, it could potentially play a role in situations of hematopoietic stress. Similar behaviour can be observed for $m_2(j)$. In agreement with $m_1(j)$ in Figure 5.10, we observe a decrease of $m_2(j)$ for scenario SymmAll and an increase of $m_2(j)$ in scenario AsymmAll, as a function of the corresponding division rate.

We now compare the Symm1 and SymmAll scenarios. We observe that when the symmetric division rate equals 10^{-2}, the number of cells in the genealogy of a single HSC1 progenitor differs by almost one order of magnitude between the Symm1 and SymmAll scenarios; that is, symmetric division happening in all compartments (SymmAll) results in smaller progeny from an initial HSC1 cell in subsequent compartments, compared to the Symm1 scenario. When increasing

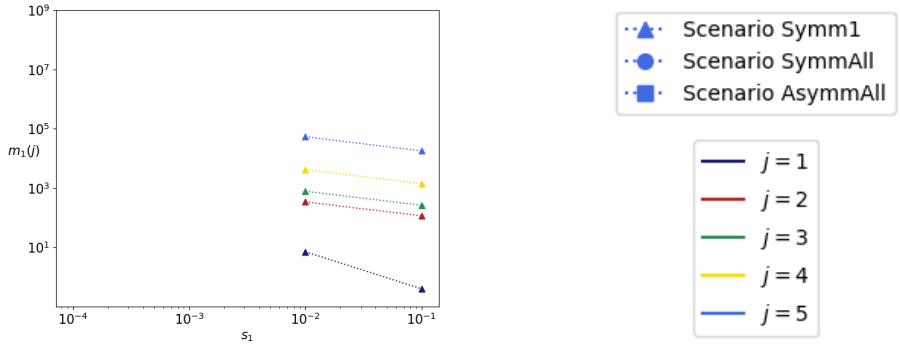
the symmetric division rate to 10^{-1}, the difference between the number of cells, $m_1(j)$, within the genealogy for the Symm1 and SymmAll scenarios increases further. As mentioned also in Chapter 3, excessive clonality may increase the risk of cancerous mutations becoming established. In particular, clonal hematopoiesis has been observed in older mice and humans (Bowman *et al.*, 2018) and it is strictly linked to the accumulation of mutations in the HPC population and an increasing risk of leukemia. Thus, our results suggest that symmetric division could be a possible way to control cell differentiation and limit mutation accumulation in the hematopoietic system.

5.2.3 Tracking a thymocyte during its development

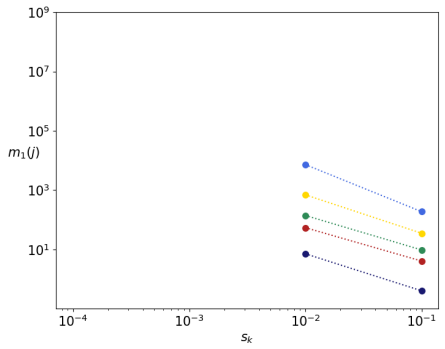
We consider now the model proposed in (Sawicka *et al.*, 2014, Model 2), and shown in Figure 5.11, of T cell thymic development. The reader can find more details about thymic development and how double negative (DN) thymocytes differentiate to eventually become CD4 or CD8 cells in Section 1.1. In particular, DN cells differentiate to pre-selection DP thymocytes (pre-DP); this population represents the first compartment in the model of Sawicka *et al.* (2014) and will contain an initial number of cells (initial condition, $C_1(0)$). Pre-DPs undergo maturation in the thymus and progress to the double positive stage (post-DP), where thymocytes express both CD4 and CD8 co-receptors. Post-DP cells that are positively selected transition to the single positive (SP) stage, where they can express either the CD4 or CD8 co-receptor. Some of these cells will then reach the periphery as (single) CD4 or CD8 SP cells.

We exploit this particular model to illustrate the applicability of our single-cell analysis developed in Section 5.1.3. This model also allows us to show how our methods can be easily adapted to different network topology of compartments. In this case, it is a compartmental bifurcation, rather than a linear sequence of compartments.

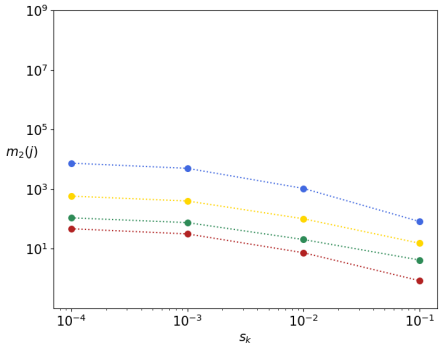
First, it is of interest to quantify cell fate in this system; that is, to estimate the percentage of pre-DP thymocytes that are predicted to die in each of the compartments during development, and the percentage that successfully reach the periphery instead (either as a CD4 or CD8 SP cell). It is clear that our



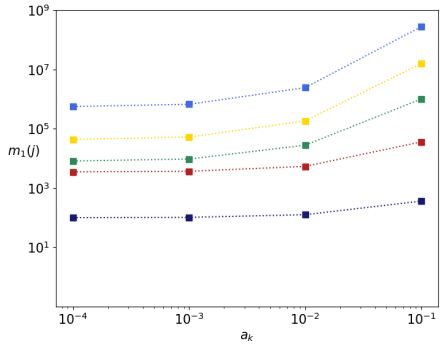
(a) HSC1 progenitor for Scenario Symm1



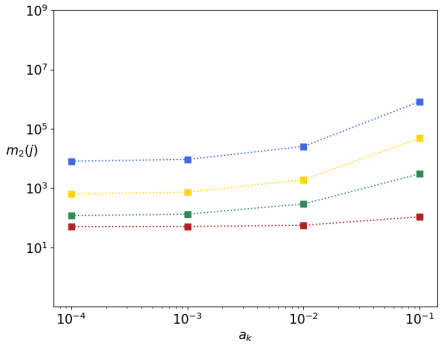
(b) HSC1 progenitor for Scenario SymmAll



(c) HSC2 progenitor for Scenario SymmAll



(d) HSC1 progenitor for Scenario AsymmAll



(e) HSC2 progenitor for Scenario AsymmAll

Figure 5.10: Mean number, $m_i(j)$, of cells in the genealogy for an (a, b, d) HSC1 ($i = 1$) or (b, e) HSC2 ($i = 2$) progenitor, in compartments $j \in \{i, \dots, 5\}$. In each scenario, we vary the corresponding rate (s_1 in Symm1, s_k, $k \in \{1, \dots, M\}$ in SymmAll, and a_k, $k \in \{1, \dots, M+1\}$ in AsymmAll), where only values leading to finite $m_i(j)$ are considered.

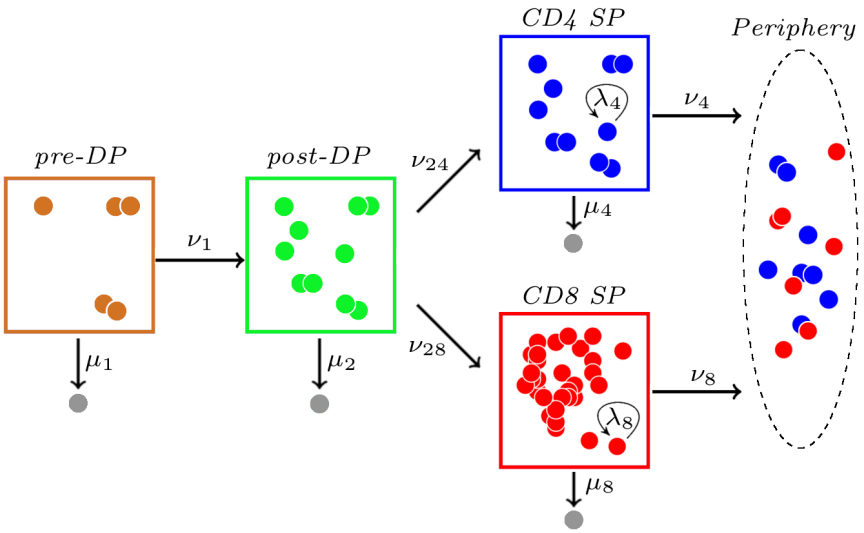


Figure 5.11: Thymic development model proposed by Sawicka *et al.* (2014). Grey cells represent cell death.

arguments in Section 5.1.3 can be easily adapted to quantify this. In particular, one can slightly redefine the probabilities $\beta_i(j)$ in Section 5.1.3, with $i = 1$ (i.e. a single pre-DP thymocyte being tracked), as

$$\begin{aligned}
\beta_1(1) &= \text{probability that the pre-DP thymocyte dies in the pre-DP compartment,}\\
\beta_1(2) &= \text{probability that the pre-DP thymocyte dies in the post-DP compartment,}\\
\beta_1(4) &= \text{probability that the pre-DP thymocyte dies in the CD4 SP compartment,}\\
\beta_1(8) &= \text{probability that the pre-DP thymocyte dies in the CD8 SP compartment,}\\
\beta_1(4P) &= \text{probability that the pre-DP thymocyte reaches the periphery as a CD4 SP cell,}\\
\beta_1(8P) &= \text{probability that the pre-DP thymocyte reaches the periphery as a CD8 SP cell.}
\end{aligned}$$

Similar solutions to those in Section 5.1.3 can be obtained by incorporating the

compartmental bifurcation in the first-step analysis, leading to

$$\begin{aligned}
\beta_1(1) &= \frac{\mu_1}{\mu_1+\nu_1},\\
\beta_1(2) &= \frac{\nu_1}{\mu_1+\nu_1}\frac{\mu_2}{\mu_2+\nu_{24}+\nu_{28}},\\
\beta_1(4) &= \frac{\nu_1}{\mu_1+\nu_1}\frac{\nu_{24}}{\mu_2+\nu_{24}+\nu_{28}}\frac{\mu_4}{\mu_4+\nu_4},\\
\beta_1(8) &= \frac{\nu_1}{\mu_1+\nu_1}\frac{\nu_{28}}{\mu_2+\nu_{24}+\nu_{28}}\frac{\mu_8}{\mu_8+\nu_8},\\
\beta_1(4P) &= \frac{\nu_1}{\mu_1+\nu_1}\frac{\nu_{24}}{\mu_2+\nu_{24}+\nu_{28}}\frac{\nu_4}{\mu_4+\nu_4},\\
\beta_1(8P) &= \frac{\nu_1}{\mu_1+\nu_1}\frac{\nu_{28}}{\mu_2+\nu_{24}+\nu_{28}}\frac{\nu_8}{\mu_8+\nu_8}.
\end{aligned}$$

These analytical expressions allow us to perform a local sensitivity analysis by computing partial derivatives with respect to model parameters. For example, we have

$$\frac{\partial\beta_1(8P)}{\partial\nu_{24}} = \frac{\partial\beta_1(8P)}{\partial\mu_2} = -\frac{\nu_1\,\nu_{28}\,\mu_8}{(\mu_1+\nu_1)(\mu_2+\nu_{24}+\nu_{28})^2(\mu_8+\nu_8)}.$$

The proliferation potential of thymocytes during thymic development directly depends on them reaching the CD4 SP or CD8 SP compartments, where they are able to divide, before they exit to the periphery. Thus, the average number of divisions initiated by a single pre-DP thymocyte during its thymic development journey, $\eta_1 = \eta_1(4) + \eta_1(8)$, is given by

$$\begin{aligned}
\eta_1(4) &= \frac{\nu_1\nu_{24}}{(\mu_1+\nu_1)(\mu_2+\nu_{24}+\nu_{28})}\frac{\lambda_4}{\mu_4+\nu_4},\\
\eta_1(8) &= \frac{\nu_1\nu_{28}}{(\mu_1+\nu_1)(\mu_2+\nu_{24}+\nu_{28})}\frac{\lambda_8}{\mu_8+\nu_8}.
\end{aligned}$$

Finally, the average lifespan of a pre-DP cell during thymic development (i.e. the mean time until it dies or it reaches the periphery) is given by

$$\tau_1 = \frac{1}{\mu_1+\nu_1}\left[\frac{\nu_1}{\mu_2+\nu_{24}+\nu_{28}}\left(\frac{\nu_{24}}{\mu_4+\nu_4}+\frac{\nu_{28}}{\mu_8+\nu_8}+1\right)+1\right].$$

We consider parameter values in Table 5.2, selected from (Sawicka *et al.*, 2014, Section 3.2). Our methods enable one to compute the average lifespan of a pre-DP cell during the development process (until it dies or reaches the periphery),

which corresponds to $\tau_1 = 2.84$ days. During its lifetime, a cell may undergo differentiation and proliferation, before dying in one of the compartments without ever reaching the periphery, or reaching the periphery either as a CD4 or CD8 cell. We plot the predicted cell fate probabilities in Figure 5.12. The most likely outcome corresponds to cell death, especially during the early stages (pre-DP and post-DP compartments). This agrees with existing knowledge that the bulk of negative selection occurs during the DP stages of development (Yates, 2014). Once a cell reaches the CD4 SP compartment, it is more likely for this cell to reach the periphery than to die in that compartment, while these probabilities are comparable in the CD8 SP case.

Rate	μ_1	ν_1	μ_2	ν_{24}	ν_{28}	λ_4	λ_8	μ_4	μ_8	ν_4	ν_8
Value	0.263	0.137	1.369	0.07	0.054	0.216	0.093	0.04	0.11	0.21	0.14

Table 5.2: Parameter values from (Sawicka *et al.*, 2014, Section 3.2), in units $days^{-1}$.

In Table 5.3, we present the elasticities (i.e. normalised derivatives) of the probabilities $\beta_1(j)$, $j \in \{1, 2, 4, 8, 4P, 8P\}$, with respect to model parameters. This can be of particular relevance when parameters have been estimated experimentally with some uncertainty, so that one can assess the impact of perturbations in these values on model outputs. As expected, the division rates, λ_4, λ_8, do not affect the probability of death of the tracked cell starting as pre-DP. Moreover, death rates $\mu_1, \mu_2, \mu_4, \mu_8$ positively contribute to the probability of the cell dying in the corresponding compartment $j \in \{1, 2, 4, 8\}$, while negatively contributing to the probability of the cell dying in other compartments. It is also worth noting that, for $j \in \{2, 4, 8\}$, the probability $\beta_1(j)$ of the cell dying in that compartment is mainly affected by the differentiation rate into compartment j; that is, $\nu_1, \nu_{24}, \nu_{28}$, respectively. This can be understood since we are tracking a cell starting in compartment $i = 1$ and following the journey of this pre-DP cell developing across the sequence of compartments. Thus, the probability of dying in a compartment $j \neq 1$ is mainly driven by the differentiation rates of previous compartments.

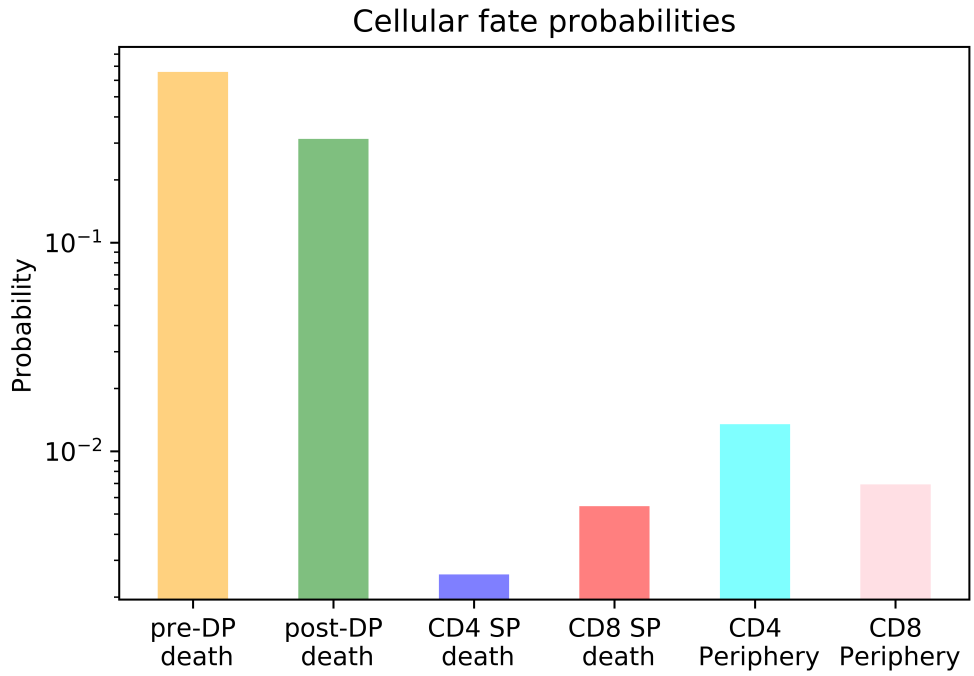


Figure 5.12: Probabilities of a single pre-DP cell to die in each of the compartments (pre-DP, post-DP, CD4 SP, or CD8 SP) before reaching the periphery, or to reach the periphery as a CD4 or CD8 SP cell. In particular, $\beta_1(1) = 0.6575$, $\beta_1(2) = 0.3140$, $\beta_1(4) = 0.0026$, $\beta_1(8) = 0.0055$, $\beta_1(4P) = 0.0135$ and $\beta_1(8P) = 0.0069$.

$\frac{\partial\beta_i(j)}{\partial\theta}/\frac{\beta_i(j)}{\theta}$	μ_1	ν_1	μ_2	ν_{24}	ν_{28}	λ_4	λ_8	μ_4	μ_8	ν_4	ν_8
$\beta_1(1)$	0.34	−0.34	0	0	0	0	0	0	0	0	0
$\beta_1(2)$	−0.66	0.66	0.08	−0.04	−0.04	0	0	0	0	0	0
$\beta_1(4)$	−0.66	0.66	−0.92	0.96	−0.04	0	0	0.84	0	−0.84	0
$\beta_1(8)$	−0.66	0.66	−0.92	−0.04	0.96	0	0	0	0.56	0	−0.56
$\beta_1(4P)$	−0.66	0.66	−0.92	0.96	−0.04	0	0	−0.16	0	0.16	0
$\beta_1(8P)$	−0.66	0.66	−0.92	−0.04	0.96	0	0	0	−0.44	0	0.44

Table 5.3: Elasticities for the probabilities $\beta_1(j)$, $j \in \{1, 2, 3, 8, 4P, 8P\}$, with respect to parameter $\theta \in \{\mu_1, \nu_1, \mu_2, \nu_{24}, \nu_{28}, \lambda_4, \lambda_8, \mu_4, \mu_8, \nu_4, \nu_8\}$. That is, the ratio $(\partial\beta_1(j)/\partial\theta)/(\beta_i(j)/\theta)$ of the cell fate probabilities (rows) with respect to model parameters (columns).

The average number of division events performed by a single pre-DP cell is $\eta_1 = \eta_1(4) + \eta_1(8) = 0.0139 + 0.0046 = 0.0185$. This implies that out of 10^2 pre-DP cells entering the thymic development process only around 2 cells are expected to be produced by cell division from the original cells, when visiting the CD4 SP or CD8 SP compartments. These small values are directly related to the small probabilities of reaching these compartments at all, so that the cell can actually divide. Our results here are in agreement with the conclusions and insights from (Sawicka *et al.*, 2014), where the authors make use of a deterministic model to conclude that thymic development is a stringent process characterised by an extremely low success rate.

5.3 Discussion

In this Chapter, we have presented a general stochastic model for cell division, death, and differentiation across a sequence of compartments. Cells can divide, die or exit to adjacent compartments over time. We have derived analytical expressions for the mean number of cells in each compartment over time under different scenarios of interest (e.g irreversible model, where differentiation cannot be reversed) and studied the genealogy of a single progenitor cell in terms of the probability generating function and summary statistics appropriately defined. The single-cell analysis allows one to track the journey of a cell during its lifetime across the system of compartments. We have then calculated its lifetime, its proliferation potential and the probability of the different cell fates. We have also presented a number of case studies to illustrate the applicability of our techniques and the impact that model parameters can have on the corresponding predictions.

Our results highlight the significant role that symmetric and asymmetric division events can play in these systems, when compared to self-renewal. In particular, we have shown how symmetric division events can significantly affect the dynamics of the system, potentially moving it from unlimited growth to extinction. Increasing the asymmetric division rates cannot change the system's asymptotic behaviour. Still, it can delay population extinction by increasing the number of cells arising over time across compartments. On the other hand, for a fixed division rate H and different probabilities for each kind of division event

(self-renewal, p_{SR}; symmetric division, p_{SD}; asymmetric division, p_{AD}), compartmental systems where symmetric and asymmetric division is more likely lead to smaller cell populations and faster dynamics towards extinction or steady state, compared to systems where self-renewal is the dominant division event. Interestingly, increasing the probability of symmetric or asymmetric division leads to smaller genealogies of a single progenitor cell, while maximising the size of the fully differentiated (or product) population.

A particular limitation of our approach, common also to Chapter 3 and 4, is that it relies on cells behaving independently from each other, which is directly related to the fact that the corresponding system of ODEs for the average behaviour of the process is linear. This allows one to implement techniques from the theory of branching processes, and makes the single-cell analysis implementation feasible (where one can follow a single-tracked cell, while neglecting the dynamics of other cells). On the other hand, cellular interactions might not be negligible in some situations (e.g under competition for resources, where logistic growth-type models might be more appropriate). Relaxing this particular assumption is, thus, the aim of future work.

Beyond the techniques applied in this Chapter and the future research that mentioned above, a natural prosecution of the studies carried out for this PhD thesis would consist in the inclusion of the different cellular events considered here, within the probability generating function method applied in Chapters 3 and 4. For example, a study of distributions $\mathbb{P}\left(\mathbf{C}_{M+1}(t) = k\right)$ and $\mathbb{P}\left(\mathbf{G}(t) = n\right)$ could be extended to include the additional cellular events considered in this Chapter.

Chapter 6

A deterministic approach to CD8$^+$ T cell exhaustion dynamics

As described in greater detail in Chapter 1, during an acute infection, once näive CD8$^+$ T cells encounter their specific antigen, they initiate a program that drives them to expand in large numbers; they differentiate into cytotoxic effector cells, which are able to control and, eventually, clear infection. However, in the case of a chronic infection, there are substantial alterations of the T cell activation and differentiation process, potentially leading T cells to a dysfunctional exhausted state described in details in Section 1.2.

In this Chapter, we present the main results of the work carried out as part of the PhD thesis, in collaboration with the Salk Institute for Biological Studies in La Jolla (California, USA). Professor Susan Kaech and a post-doctoral scientist in her group, Dr. Thomas Mann, are addressing, from an experimental point of view, the question of how CD8$^+$ T cells recognise a "chronic" antigen during early time point of a primary infection; a secondary immune response or a longer infection window have been not considered and might be part of future research. In particular, they focus on how the duration of antigen signalling in the first hours of antigen presentation could affect CD8$^+$ T cell differentiation into the dysfunctional exhausted state. As detailed in Section 1.2, exhaustion is a progressive and heterogeneous process involving several differentiation steps and cell phenotypes arising very soon after antigen presentation. Thus, a crucial point Professor Kaech and Dr. Mann would like to investigate is the role

of both progenitor effector and progenitor exhausted cells in the generation of a terminally exhausted population of cells. Note that we used, as common in the literature and done in previous Chapters, the terminology "progenitor cells" to indicate cells of an early phenotypes.

Professor Susan Kaech and Dr. Thomas Mann are making use of novel optogenetic approaches to carry out experiments in which CD8$^+$ T cells are stimulated for different lengths of time, over a maximum of 48 hours, to examine the exhaustion phenotypes that develop in this time window. The experiment has been carried out *in vitro* with CD19 CAR T cells, genetically modified T cells able to express a chimeric antigen receptor (CAR) targeted against the CD19 antigen. Activated CAR T cells have been exposed to different duration of antigen stimulation, and flow cytometry data have been collected after 48 hours to identify the different populations of CAR T cells, from "functional" to "exhausted" phenotypes, present in the samples. Specific questions that have been addressed include i) understanding when CD8$^+$ T cells make the decision to change their intra-cellular state and differentiate, ii) understanding how the presence of antigen (in chronic viral infection or cancer) influences the dynamics of exhaustion, iii) determining which exhaustion states become irreversible.

Within this PhD research project several deterministic mathematical models for CD8$^+$ T cell dynamics, representing different possible pathways of cell differentiation to exhausted states, have been developed with the aim to unravel some of these complex biological questions. Data sets from Dr. Mann and the developed mathematical models are brought together with Bayesian inference to quantify exhaustion process, investigate CD8$^+$ T cell differentiation and the mechanisms of T cells responding to the presence of antigen (viral or tumour).

Details of the optogenetic experiment carried out by Dr. Mann are given in Section 1.2.2; Section 6.1 aims to provide the reader with an overview of the experimental setting and the biological process that our mathematical models should characterise. Given the novelty of the topic, a conceptualisation of the mathematical model structures representing different hypotheses regarding the heterogeneity of CD8$^+$ T cell exhaustion has been carried out at the beginning of this PhD project. In Section 6.2, different mathematical models are defined as an example of the performed study about CD8$^+$ T cell trajectories during the

exhaustion-differentiation process. Section 6.3 reports the mathematical analysis for a reduced mathematical model, where only four cell states are considered; a study of parameter space and Bayesian inference have also been carried out.

Bayesian techniques aim to validate a model and finally make predictions of parameter values and cell dynamics, which in this case would be related to antigen stimulation of different duration. However, given experimental biological measurements, it is not always possible to identify an appropriate mathematical model; this depends, among other things, on the dimension of parameter space, on parameters dependencies with model variables and, importantly, on the measurability of model outputs. Indeed, some populations of cells cannot be univocally targeted and identified during measurements due to a lack of biological details; also, their observation and measurement might not be possible given current limitations in the analysis techniques. In this case, more biological as well as mathematical assumptions must be made. The mathematical analysis of the deterministic models developed in this Chapter has been affected by the novelty of exhaustion topic: already published experimental data sets were lacking and, hence, the need to rely on ongoing experiments. A consultation with our immunology collaborators was, thus, crucial. Unfortunately, due to the COVID-19 pandemic, the performed optogenetic experiments resulted only in preliminary data sets; thus, model validation and prediction could not be performed within the timeline of this PhD thesis.

When using deterministic mathematical models, one assumes that the cell population size is large enough so that random fluctuations can be ignored. Deterministic models are, thus, widely used in the limit of large population sizes, but they tend to become inaccurate if the number of cells of a given population becomes small. For example, when the number of functional effector cells drops to a low value, this population might die out due to fluctuations. In such situations, stochastic models might become essential to describe the dynamics of these cells. In Sections 6.4 and 6.5, we take the reader through some considerations and limitations of the current mathematical model, as well as through some thoughts for future work on mathematical modelling for $CD8^+$ T cell exhaustion process.

6.1 The optogenetic setup

A summary of the experiments performed by Prof. Kaech and Dr. Mann from the Salk Institute of Biological Studies in La Jolla, California, is as follows. They aim to explore how, in an *in vitro* system, CD8$^+$ CAR T cells differentiate to an exhausted phenotype over different duration of antigen stimulation. Dr. Mann has established an optogenetic system for using blue light to perform a precise and high-throughput investigation of how engineered T cell signalling dynamics influence cell fate to an effector or exhausted state. The reader may refer to Section 1.2 for further details about the exhaustion-differentiation process and in particular to Section 1.2.2 for a more extensive description of the experiments. Here, we only provide an overview of the experimental scenario relevant for the mathematical modelling.

Dr. Mann has made use of an optogenetically controllable version of a chimeric antigen receptor, named "OptoCAR", that is sensitive to blue light and, thus, allows to control signal dynamics in the T cell receptor pathway. The engineered receptor can transduce antigen signal when the cell is in the dark, but signalling is inhibited when the cell is exposed to blue light (see Figure 1.6 in Chapter 1). Each experimental repeat is run with a 96 well plate; a fixed volume of sample is placed in each well at the beginning of the optogenetic experiment. The abundance of two populations of exhausted cells was quantified by flow cytometry analysis after 48 hours. The experimental data sets used in this Chapter consist of six repeats carried out between August and December 2020.

In particular, only active cells showing the specific antigen receptor, are placed in the wells and optogenetically stimulated; a fixed sample volume containing 25×10^3 cells is placed in each well at the beginning of the optogenetic experiment. Given the structure of the optoPlate-96, reported in Figure 1.7, different wells can receive different duration of signal within subsequent 48 hours. At the end of the experiments, cells are collected and analysed via flow cytometry. This method, widely used in biology, allows to identify population of cells by the cell-surface markers they express, *i.e.,* proteins attached to the cellular membrane. In particular, for this study, populations characterised as TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$ have been quantified. Indeed, as detailed in Section 1.2, PD-1, TIM-3

inhibitory receptors and the transcription factor TCF1 have a clear expression in the exhausted T cell population (Mann & Kaech, 2019; Wherry *et al.*, 2007).

The data sets used for the mathematical models, defined and analysed in this Section, consist of the percentage of TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$ cells over the entire population of transduced live CD8$^+$ OptoCAR$^+$ T cells, *i.e.,* cells showing the OptoCAR at sufficient level and which express the corresponding markers. They represent populations of reversible and terminal differentiated cells, characterised as TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$, respectively. Moreover, some beads have been added after the optogenetic experiment and before the flow cytometry analysis in order to quantify the loss of cells during the staining procedure with fluorescence markers. The collected number of beads out of the initial amount represents a correction factor for the total number of cells in each well. Those corrected ratios have been used as measured experimental outputs for our mathematical models.

At the beginning of the experiment, the 25×10^3 cells present in a well are placed in the dark; that is, antigen signalling is activated. Then, light is switched on at different times. Each well is characterised by a different *dark delay*, duration of the antigen stimulation happening in the dark before light is switched on. In particular, we denote by $l_{dd} \in \boldsymbol{d}$, the length of the dark delay in hours, with

$$\begin{aligned}\boldsymbol{d} = \big(&0,\ 1.33,\ 2.67,\ 4,\ 5.33,\ 6.67,\ 8,\ 9.33,\ 10.67,\ 12,\ 13.33,\ 14.67,\ 16,\\ &17.33,\ 18.67,\ 20,\ 21.33,\ 22.67,\ 24,\ 25.33,\ 26.67,\ 28,\ 29.33,\ 30.67,\ 32,\\ &33.33,\ 34.67,\ 36,\ 37.33,\ 38.67,\ 40,\ 41.33,\ 42.67,\ 44,\ 45.33,\ 48\big)\,.\end{aligned}$$

Figure 6.1 provides an example of a dark delay of 20 *hours*: cells have been placed in the dark (antigen signalling is perceived) for the first 20 *hours*, then light has been switched on and signal interrupted.

For every optogenetic experiment and length of dark delays the ratios, over the total live cell population, of TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$ exhausted cells have been quantified at the end of 48 *hours*. These quantified ratios of cells can be thought of as a function of the dark delay cells have been experiencing; that is, the duration of the antigen signal before light has been switched on and, thus, the LOVTRAP domain unbound (see Section 1.2.2). An example of a specific experiment run (August 2020) is reported in Figure 6.2.

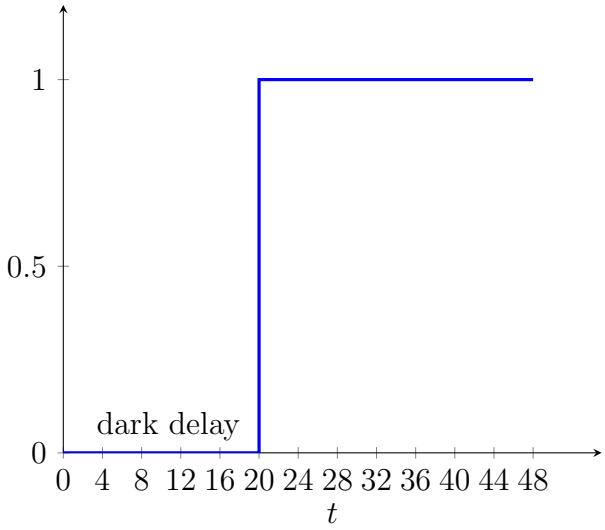


Figure 6.1: Example of a dark delay of 20 *hours* within the total 48 *hours* of the optogenetic experiment. Cells have been placed in wells in the dark and lit with blue light only after 20 hours. Thus, for the first 20 *hours* cells are stimulated, then from time $t = 20$ to $t = 48$ *hours* they are not.

6.2 Mathematical models of T cell exhaustion-differentiation

In this Section, we aim to investigate the population dynamics of CD8$^+$ T cells during the exhaustion-differentiation process using experimental data provided by Dr. Mann. In particular, the experiment performed by Dr. Mann has a total duration of six days, during which cells are activated, transduced, selected and, finally, optogenetically stimulated. Our mathematical analysis of T cell dynamics focuses on the last two days of the entire process, when cells have been placed in the opto-96 plate and optogenetically stimulated. Cell counts in each well at the beginning of the optogenetic experiment are the initial conditions of our mathematical models. At the beginning of the optogenetic experiment, cells have been already exposed to the antigen-peptide (during the first 4 days) and, thus, they cannot be considered näive cells anymore. However, an analysis of their heterogeneity has not been performed before the start of the experiment. Consequently, only an estimate of the population cell counts has been provided

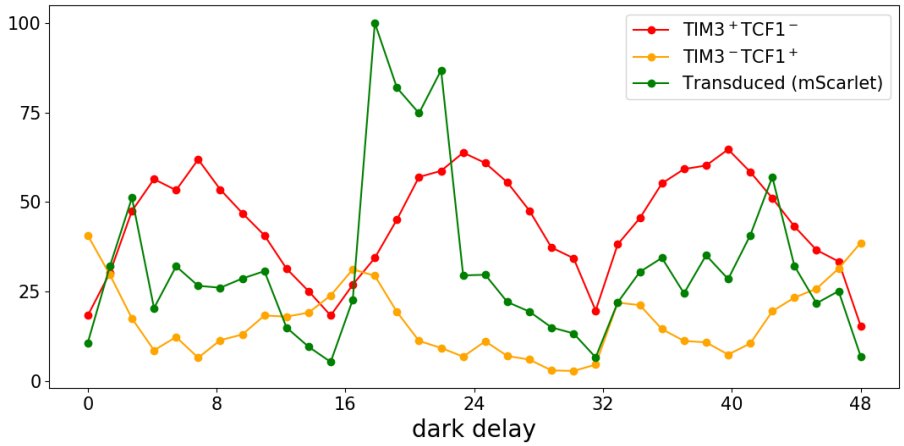


Figure 6.2: Percentages of TIM-3$^-$ TCF1$^+$ (in orange) and TIM-3$^+$ TCF1$^-$ (in red) over the total number of transduced and live cells at time $t = 48$ *hours*. These markers characterise a reversible and a terminally exhausted population, respectively. On the x-axis the dark delay; that is, the duration of dark time (antigen signal) cells have been experiencing from 0 to a maximum of 48 *hours*. In green, the transduced cell count, at time $t = 48$ *hours*, normalised over the 36 different considered dark delays.

by our experimental collaborators.

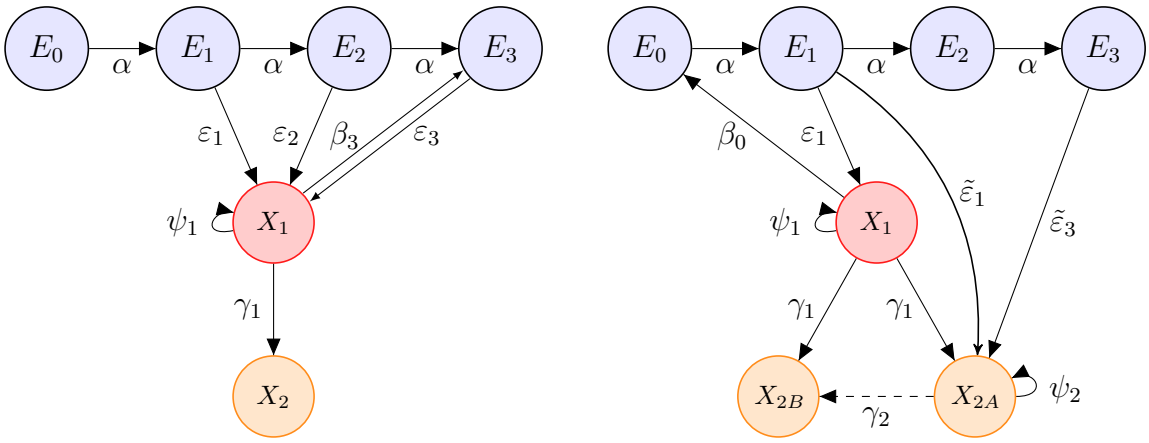
For this PhD project, several mathematical models have been considered in order to study a broad class of effector states (from a less to a more terminally differentiated) and their possible transition to an exhausted phenotype. Each cell state represented in the mathematical model is thought to be characterised by flow cytometry markers; as detailed in Chapter 1, these markers define the states of a cell according to the proteins and receptors that are expressed on its external membrane.

Two examples of the mathematical models defined in collaboration with the biological expertise of Dr. Mann from Salk Institute are shown in Figure 6.3. We model the exhaustion-differentiation process within a compartmental structure where cells with a specific phenotype belong to a compartment and they can undergo compartment-specific events (Feliciangeli *et al.*, 2022; Höfer *et al.*, 2016; Johnston *et al.*, 2007; Thomas-Vaslin *et al.*, 2008; Zhang *et al.*, 2015). Here, four

states of effector cells, from a less to a more differentiated state, namely E_0, E_1, E_2,E_3, are considered. Two or three exhausted states are defined in Model 1 (on the left) and Model 2 (right), respectively. In particular, Model 1 is characterised by a reversible exhausted state, X_1, and by a terminally exhausted state, X_2. Cells belonging to state X_1, characterised by surface marker expression TIM-3$^-$ TCF1$^-$, are still able to undergo self-renewal and revert to a functional effector state. Whereas, cells of state X_2, characterised as TIM-3$^+$ TCF1$^-$, have a fixed exhausted transcriptional profile. In Model 2 (on the right of Figure 6.3), two different states of terminally exhausted cells are considered following the findings of Hudson *et al.* (2019) and Beltra *et al.* (2020). Very little is known about these two cell populations but recent studies agree on the existence of an intermediate terminally exhausted population, still able to proliferate, and a second terminal population undergoing little to no division and poorly contributing to viral control. In both models, cells of all states can die via apoptosis; death rates have not been graphically represented in Figure 6.3 in order to not overload the figure. In particular, we assume all effector cells die at rate μ_E, exhausted cells die at rate μ_1, and μ_2 for reversible, terminal states of Model 1, or μ_1, μ_{2A} and μ_{2B} for reversible and the two terminal subsets of Model 2, respectively.

In both acute and chronic infections, differentiation into terminal states (effector or exhausted) is linked to cell division: in order to acquire new functions and become more specialised, cells undergo rounds of division and change their phenotype. However, despite its fundamental importance, the role of proliferation during the differentiation process is still not completely clarified. Experimentally, fluorescence dyes, such as green fluorescent protein (GFP) labelling dye, CFSE (carboxyfluorescein succinimidyl ester) dye, or CellTrace Violet labelling dye, can be used to measure the number of times cells divide (Lemieszek *et al.*, 2022; Lyons, 2000); indeed, their fluorescence intensity is reduced by a half at each division event. In this way, different effector states could be defined according to the number of divisions cells underwent during the length of the experiment. In our simplified model, these "effector-proliferative" states (determined by the number of divisions) correspond to effector-differentiation states: the longer the antigen signal is present, the more cells proliferate and differentiate to more terminal effector states. That is, each round of division brings a cellular maturation to an

effector state, with a greater antigen exposure: when a cell divides it moves further in the chain of the CD8$^+$ T cell differentiation process. Mathematically, we let $i \in \mathbb{N}^+$ and describe the transition of cells from a more stem-like (progenitor) state, E_i, to a more differentiated state, E_{i+1}, as a conveyor belt process (Dingli *et al.*, 2007; Marciniak-Czochra *et al.*, 2009; Stiehl & Marciniak-Czochra, 2011; Weekes *et al.*, 2014). In particular, when a cell in state E_i divides, both daughter cells belong to state E_{i+1}, for $i = 0, 1, 2$. The division rate is denoted by α. Cells in state E_3 are thought of as terminally differentiated effector cells and, thus, not able to divide and differentiate further.



(a) Model 1 (b) Model 2

Figure 6.3: Representation of two mathematical models (Model 1 on the left and Model 2 on the right) for cell proliferation, death and differentiation from a progenitor effector state to a terminally exhausted state.

Besides the number of exhausted states, Model 1 and 2, shown in Figure 6.3, differ in the pathway to exhaustion. In Model 1, all effector states, E_i, but the progenitor, E_0, can differentiate to the reversible exhausted state X_1 at a rate ε_i, for $i = 1, 2, 3$. Cells of the reversible exhausted state X_1 can self-renew at rate ψ_1, revert back only to a terminal effector phenotype E_3 at rate β_3, or move to the terminally exhausted state X_2 at rate γ_1. In Model 2, instead, we assume that only early differentiated effector cells of state E_1, or terminally differentiated effector cells of state E_3 can move to the reversible exhausted state X_1; also,

reversible cells in X_1 can self-renew at rate ψ_1 (as in Model 1), but they can revert back only to the progenitor effector state E_0. Moreover, effector cells in state E_1 and E_3 can commit to the terminally exhausted state X_{2A} at rate $\tilde{\varepsilon}_1$ and $\tilde{\varepsilon}_3$, respectively. Terminally exhausted cells in state X_{2A} move to state X_{2B} at rate γ_2, and exhausted cells in X_1 move to a terminal state at rate γ_1. In order to model this transition, we introduce the probability p such that, when differentiating to a terminal state, cells in X_1 move to X_{2A} with probability p and to X_{2B} with probability $1-p$.

In both Model 1 and Model 2, we consider the proliferation rates to be antigen-dependent via a time-dependent function $f(t)$, $f : [0, +\infty) \to \{0, 1\}$ defined as

$$f(t) = \begin{cases} 1, & \text{for time } t \leq l_{dd}, \\ 0, & \text{for time } t > l_{dd}, \end{cases} \tag{6.1}$$

for $l_{dd} \in \boldsymbol{d}$ the dark delay; that is, when cells are in the dark, the LOVTRAP domain is bound and the signal transduced, thus $f(t) = 1$. Other functions could have been used to model the presence of the antigen signal according to the binding/unbinding state of the LOVTRAP domain; however, since there are not enough details to discriminate between a step-wise form of $f(t)$ and a more complex function, *e.g.,* a Hill function, we decided to avoid additional parameters and keep the model simpler by choosing a step signal function.

Given the considerations above for Model 1 (shown on the left in Figure 6.3), we let $E_0(t)$, $E_1(t)$, $E_2(t)$, $E_3(t)$, $X_1(t)$ and $X_2(t)$ be variables representing the size of the respective populations as a function of time. Assuming linear proliferation, differentiation and death processes, the dynamics, for $t \geq 0$, of $E_0(t)$, $E_1(t)$, $E_2(t)$,

$E_3(t)$, $X_1(t)$ and $X_2(t)$ are described by the following equations

$$\begin{aligned}
\frac{dE_0(t)}{dt} &= -(\alpha f(t) + \mu_E)E_0(t)\,, \\
\frac{dE_1(t)}{dt} &= 2\alpha f(t)E_0(t) - (\alpha f(t) + \mu_E + \varepsilon_1)E_1(t)\,, \\
\frac{dE_2(t)}{dt} &= 2\alpha f(t)E_1(t) - (\alpha f(t) + \mu_E + \varepsilon_2)E_2(t)\,, \\
\frac{dE_3(t)}{dt} &= 2\alpha f(t)E_2(t) - (\varepsilon_3 + \mu_E)E_3(t) + \beta_3 X_1(t)\,, \\
\frac{dX_1(t)}{dt} &= \psi_1 f(t)X_1(t) + \varepsilon_1 E_1(t) + \varepsilon_2 E_2(t) + \varepsilon_3 E_3(t) - (\gamma_1 + \beta_3 + \mu_1)X_1(t)\,, \\
\frac{dX_2(t)}{dt} &= \gamma_1 X_1(t) - \mu_2 X_2(t)\,.
\end{aligned}$$

Initial conditions are set to

$$\begin{aligned}
&E_0(0), E_1(0), E_2(0), E_3(0), X_1(0), X_2(0) \\
&\qquad = (35\%\, N^\star, 25\%\, N^\star, 20\%\, N^\star, 10\%\, N^\star, 5\%\, N^\star, 5\%\, N^\star)\,,
\end{aligned}$$

where we denoted by $N^\star = 25 \times 10^3$, the total number of live and transduced cells placed in each well at the beginning of the optogenetic experiment. We remind the reader that cells have been experienced antigen during the first four days of entire experiment and, thus, they already show some effector functions. However, in the absence of experimental determination of the initial cell numbers, the reported percentages above have been agreed with the experimental collaborators based on their intuition. Similarly, for Model 2 we let $E_0(t)$, $E_1(t)$, $E_2(t)$, $E_3(t)$, $X_1(t)$, $X_{2A}(t)$ and $X_{2B}(t)$ be variables representing the size of the respective populations as a function of time and we assume linear proliferation, differentiation and death

processes. Thus, for $t \geq 0$

$$\begin{aligned}
\frac{dE_0(t)}{dt} &= -(\alpha f(t) + \mu_E)E_0(t) + \beta_0 X_1(t)\,, \\
\frac{dE_1(t)}{dt} &= 2\alpha f(t)E_0(t) - (\alpha f(t) + \mu_E + \varepsilon_1 + \tilde{\varepsilon}_1)E_1(t)\,, \\
\frac{dE_2(t)}{dt} &= 2\alpha f(t)E_1(t) - (\alpha f(t) + \mu_E)E_2(t)\,, \\
\frac{dE_3(t)}{dt} &= 2\alpha f(t)E_2(t) - (\tilde{\varepsilon}_3 + \mu_E)E_3(t)\,, \\
\frac{dX_1(t)}{dt} &= \varepsilon_1 E_1(t) + (\psi_1 f(t) - \gamma_1 - \beta_1 - \mu_1)X_1(t)\,, \\
\frac{dX_{2A}(t)}{dt} &= p\,\gamma_1 X_1(t) + \tilde{\varepsilon}_1 E_1(t) + \tilde{\varepsilon}_3 E_3(t) + (\psi_2 f(t) - \gamma_2 - \mu_{2A})X_{2A}(t)\,, \\
\frac{dX_{2B}(t)}{dt} &= (1-p)\,\gamma_1 X_1(t) + \gamma_2 X_{2A}(t) - \mu_{2B} X_{2B}(t)\,,
\end{aligned}$$

with initial conditions

$$\begin{aligned}
&E_0(0), E_1(0), E_2(0), E_3(0), X_1(0), X_{2A}(0), X_{2B}(0) \\
&\quad = (35\%\, N^\star, 25\%\, N^\star, 20\%\, N^\star, 10\%\, N^\star, 5\%\, N^\star, 2.5\%\, N^\star, 2.5\%\, N^\star)\,.
\end{aligned}$$

Again, $N^\star = 25 \times 10^3$ denotes the total number of live and transduced cells placed in each well at the beginning of the optogenetic experiment, and the percentages represent an estimate, provided by our experimental collaborator, of the number of cells in each state.

As mentioned in Section 6.1, the experimental data used for the mathematical models consist of the percentage, over the entire population of live CD8$^+$ OptoCAR$^+$ T cells, of exhausted populations identified by TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$ markers. The ODE systems defined for Model 1 (and Model 2) describe the dynamics of the population size for each of the six (or seven) cell states. Thus, experimental data can be compared to model output via their corresponding summary statistics, *i.e.,* the percentage of reversible and terminally exhausted cells over the total number of live and transduced cells at the collection time, after $48h$ of optogenetic experiment. In particular, we denote, for Model 1 and Model 2, the total number of cells present in the system at time t by $N_t(t) = \sum_{j=0}^{3} E_j(t) + X_1(t) + X_2(t)$, where, for Model 2, we let

$X_2(t) = X_{2A}(t) + X_{2B}(t)$. Thus, we consider the following summary statistics

$$S_1 = \frac{X_1(t = 48h)}{N_t(t = 48h)} \quad \text{and} \quad S_2 = \frac{X_2(t = 48h)}{N_t(t = 48h)}.$$

as the percentages of exhausted cells in the system after $48h$ of optogenetic experiment.

6.2.1 Bayesian model selection

With the purpose of investigating possible pathways of the CAR T cell differentiation at early time points in a chronic infection, we define several mathematical models with different cell states and transitions. Here, we focus on Model 1 and Model 2 previously defined in this Section (see Figure 6.3) and apply Bayesian model selection; this method, introduced in Section 2.3 (Algorithm 3), can inform hypotheses relating to the different pathways provided by Model 1 and Model 2. Given the preliminary experimental data generated by Dr. Mann, one can, thus, determine which of the model pathway for $CD8^+$ T cell is more likely.

The reader can find a summary of the optogenetic experiment performed by Dr. Mann in Section 6.1 and a detailed description in Section 1.2.2. Flow cytometry data from a single experiment with different simulated durations of dark delay were used to compute summary statistics of interest, *i.e.* $S_1 = X_1(t)/N_t(t)$ and $S_2 = X_2(t)/N_t(t)$ with $t = 48$h. We denote Model 1 and Model 2 with $\mathcal{M}_m$ for $m = 1, 2$; then, for each model we can compute a distance measure to compare summary statistics S_1 and S_2 of the two models with the experimental data, measured at time $t = 48$h. In particular, for model $\mathcal{M}_m$, we used a generalisation of the Euclidean distance given by

$$\delta(\mathcal{M}_m) = \Bigg[\sum_{l_{dd} \in \boldsymbol{d}} \left(\frac{X_1(t = 48h; \mathcal{M}_m, l_{dd})}{N_t(t = 48h; \mathcal{M}_m, l_{dd})} - data_{X_1}(l_{dd}) \right)^2 + \left(\frac{X_2(t = 48h; \mathcal{M}_m, l_{dd})}{N_t(t = 48h; \mathcal{M}_m, l_{dd})} - data_{X_2}(l_{dd}) \right)^2 \Bigg]^{1/2},$$

where $l_{dd} \in \boldsymbol{d}$ is the length of the considered dark delay in hours.

Given the uncertainty in parameters values, at the first iteration of the model selection algorithm, parameters were sampled from uniform prior distributions of

wide interval discussed with the experimental counterpart; in particular, parameter $p \sim Unif(0,1)$ and all the others are assumed as in Table 6.1. The model (hypothesis) selection algorithm was run for a sample of size $n_{sim} = 10^5$. Making use of Bayes theorem we define, for $m = 1, 2$, the relative probability of Model $\mathcal{M}_m$ being accepted with a distance threshold $\varepsilon^\star$ as follows

$$\pi(\mathcal{M}_m \,|\, \varepsilon = \varepsilon^*) = \frac{f(\mathcal{M}_m \,|\, \varepsilon = \varepsilon^*)}{f(\mathcal{M}_1 \,|\, \varepsilon = \varepsilon^*) + f(\mathcal{M}_2 \,|\, \varepsilon = \varepsilon^*)},$$

where $f(\mathcal{M}_m)$ is a frequency, namely the number of accepted parameter sets for Model $\mathcal{M}_m$. The acceptance criterion of the ABC algorithm is to accept only the results with a distance equal or less than the chosen threshold ε^*.

Parameter	μ_E	μ_1	μ_2	α	ψ_1	ε_1	ε_2	ε_3	$\tilde{\varepsilon}_1$	$\tilde{\varepsilon}_3$	γ_1	γ_2	μ_{2A}	μ_{2B}	β_0	β_3	ψ_2
Prior	10^r	10^r	10^r	10^a	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r	10^r

Table 6.1: Summary of the prior distribution for each of the parameters in the two mathematical models. Here $r \sim Unif(-3, 0)$ and $a \sim Unif(-3, 1)$.

Results of the computed relative probability, for increasing values of $\varepsilon^\star$, are given in Figure 6.4. From the figure, it can be seen that, as $\varepsilon^\star$ decreases, the relative probability of Model 1 tends to zero and that of Model 2 tends to one. However, for large enough values of the threshold $\varepsilon^\star$, the relative probability of both hypotheses converges to 0.5. This means that, for large $\varepsilon^\star$, the two hypothesis are equally likely hypothesis given the experimental data. However, for small distance thresholds, the relative probability of Model 2 is higher; hence, we conclude that Model 2 is more plausible than Model 1, and can better explain the experimental data sets. Moreover, we can observe that, for small $\varepsilon^\star$, the number of accepted parameters grows, for increasing values of $\varepsilon^\star$, in a similar manner for both hypothesis; whereas, for greater values of the distance threshold, Model 2 (depicted by a dashed line) is performing slightly better. A similar Bayesian selection technique can be applied to the different mathematical models proposed, during this PhD project, to investigate possible pathways of the CAR T cell differentiation process at early time points of a chronic infection. For the purpose of this thesis, we restrict to illustrate the potential of the technique for Model 1 and Model 2 defined in this Section. Future work, dependent on final

experimental data, will be to run the developed algorithm considering a greater number of possible models.

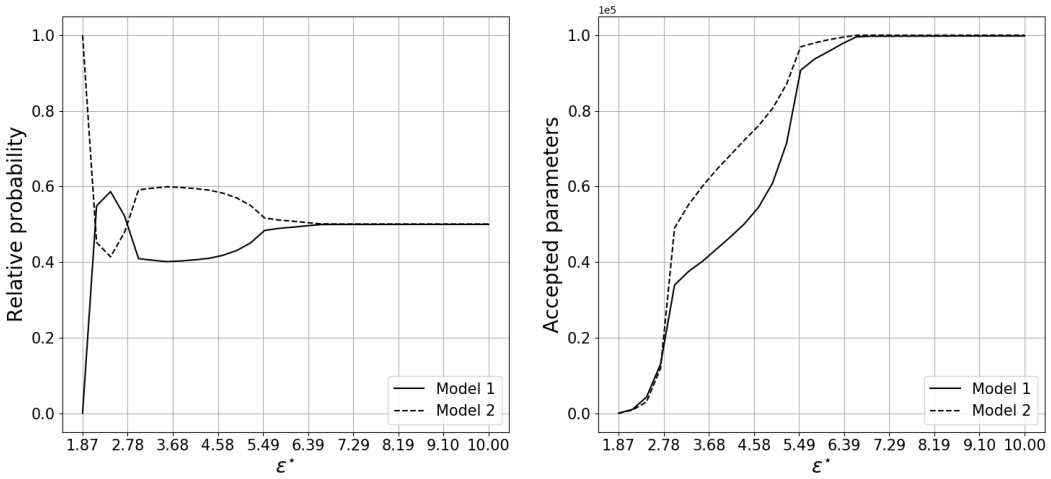


Figure 6.4: On the left, the relative probability for the two different models depending on the distance threshold $\varepsilon^\star$. On the right, the frequency of the two tested models, the number of accepted parameters for Model 1 and Model 2 as a function of the distance threshold $\varepsilon^\star$. Note the scientific notation on the right plot.

6.3 A reduced mathematical model

Given that the experimental data used for the mathematical model quantify reversible and terminally exhausted cells only, an extensive and detailed mathematical model of the exhaustion-differentiation process and the heterogeneous population involved might not be informative. Indeed, Model 1 and Model 2 defined in Section 6.2 aim to represent the dynamics of six and seven populations of T cells, respectively. However, the experimental data used quantify two cell populations (reversible and terminally exhausted cells only) at a unique time point, $t = 48$ *hours*. Moreover, only the ratio (not absolute count), over the total number of live cells after the 48 *hours* of optogenetic experiment, of reversible and terminally exhausted cells has been measured. Thus, in this Section, we propose

a reduced compartmental mathematical model, where only two states of effector cells and two states of exhausted cells are represented. We find this restricted set of cell populations to be the minimum number able to address our original questions about the heterogeneity of the exhaustion-differentiation process during antigen stimulation.

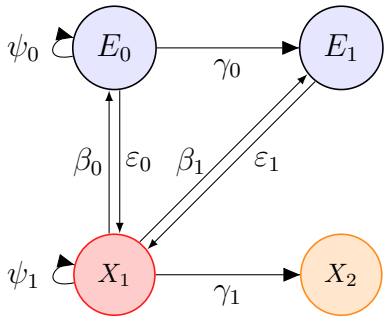


Figure 6.5: Representation of the cell states and transitions which define the toy mathematical model with a reduced number of states; here, progenitor and terminal effector cells (namely E_0 and E_1) as well as reversible and terminally exhausted CD8$^+$ T cells are represented. Cells in state E_0 and X_1 can proliferate at rates $\alpha_i + \psi_i$, for $i = 0, 1$ respectively, where α_i is the homeostatic contribution, and ψ_i is the antigen-dependent contribution. Following the suggestion of the experimental collaborators from Salk Institute of Biological Studies, we assume $\varepsilon_0, \varepsilon_1$ and β_0, β_1 to be antigen-dependent. Cells from each state can die at rates μ_E, μ_1 and μ_2 for effector, reversible and terminally exhausted state, respectively.

If in Section 6.2 four different effector differentiation states are modelled and the differentiation state of an effector cell has been identified by the number of divisions undergone; here, a toy model with only two effector states, and where proliferation and differentiation of effector cells are represented as two different events (happening at different rate) is defined. A representation of the toy mathematical model is shown in Figure 6.5. We consider progenitor effector cells, E_0, that can proliferate and differentiate, giving rise to terminal effector cells E_1. Effector cells E_0, E_1 can differentiate to the reversible exhausted state X_1. These reversible exhausted cells can slowly proliferate or commit to the terminally exhausted state X_2, or revert back to the effector state E_i, for $i = 0, 1$. Here, we assume that both progenitor effector and reversible exhausted cells can proliferate,

as well as differentiate to terminal effector and exhausted cells (at rate γ_0 and γ_1), respectively. In particular, we assume that cells can proliferate both in the absence and in the presence of antigen: we define an homeostatic proliferation rate (α_0 and α_1 for effector and exhausted cells, respectively), as well as an antigen-dependent rate (ϕ_0 and ϕ_1 for effector and exhausted cells, respectively). Moreover, we assume that differentiation rates to a terminal effector or exhausted state (γ_0 and γ_1, respectively), as well as the rates at which effector cells become exhausted (ε_0 and ε_1) are antigen-dependent, *i.e.,* they do not take place in the absence of antigen stimulation. Finally, we assume reinvigoration of exhausted cells of state X_1 (to a progenitor or terminal effector state) happens at rates β_0 and β_1, respectively, and in an antigen-dependent fashion, *i.e.,* those transitions are possible only in the absence of antigen.

As in Section 6.2, the function $f(t)$ represents the modulation of the antigen-dependent rates in response to antigen presence; we remember that $f(t)$, $f : [0, +\infty) \to \{0, 1\}$ is defined as

$$f(t) = \begin{cases} 1, & \text{for time } t \leq l_{dd}, \\ 0, & \text{for time } t > l_{dd}, \end{cases}$$

for $l_{dd} \in \boldsymbol{d}$ the dark delay; that is, when cells are in the dark, the LOVTRAP domain is bound and the signal transduced, thus, $f(t) = 1$.

In the following Sections, we define variables to represent the size of the four populations shown in Figure 6.5; we derive the corresponding ODE system and compute its solution to describe the dynamics over time of cells in states E_0, E_1, X_1, X_2. Moreover, we perform a steady state analysis to investigate the homeostatic state reached by the system. Furthermore, we carry out an identifiability analysis, following the method of Castro & de Boer (2020), as well as a global sensitivity analysis (Sobol index); this is a preliminary analysis before we perform parameter inference. Finally, we have considered six different experiment repeats and we used an Approximate Bayesian Calculation (ABC) method (see Section 2.3 for mathematical details) to infer the parameter distributions in a likelyhood free approach.

When considering a single experiment repeat, we can observe that experimental data sets show an oscillatory trend, over dark delays, of the normalised ratio,

measured at time 48 *hours*, of reversible and terminally exhausted cells. This is shown for example in Figure 6.2. This behaviour, may suggest that the death, division and proliferation rates of effector and exhausted cells are a function of the dark delay; that is, they vary with the duration of the antigen signal. As detailed in Section 6.1, cells have been stimulated for different signal durations. In particular, they were subjected to an initial period of dark delay $l_{dd} \in \boldsymbol{d} \in \mathbb{R}^{36}$ and then light has been switched on for the remaining time of the experiment's duration (total of 48 hours). For the sake of simplicity and in order to investigate the impact of the presence of antigen in the model, we consider a restricted subset of seven different dark delays. In particular, we set

$$l_{dd} \in \big(0,\, 8,\, 16,\, 24,\, 32,\, 40,\, 48\big)$$

hours and run Bayesian analysis with the aim of quantifying the impact of persistent antigen stimulation on parameter values. For this, in Section 6.3.4 parameter trends over dark delays have been computed.

6.3.1 Model analysis: dynamics and steady states

In the toy model shown in Figure 6.5 four different T cell populations are considered; for $t \geq 0$, effector and exhausted populations are defined as,

$E_0(t) =$ "the number of progenitor effector CD8$^+$ T cells",
$E_1(t) =$ "the number of terminally differentiated effector CD8$^+$ T cells",
$X_1(t) =$ "the number of reversible exhausted CD8$^+$ T cells",
$X_2(t) =$ "the number of terminally exhausted CD8$^+$ T cells".

Assuming linear proliferation, differentiation and death processes, the time evolution, for $t \geq 0$, of $E_0(t)$, $E_1(t)$, $X_1(t)$, $X_2(t)$, is described by the following

system of equations

$$\begin{aligned}
\frac{dE_0(t)}{dt} &= (\alpha_0 + (\psi_0 - \varepsilon_0)f(t) - \gamma_0 - \mu_E)E_0(t) + \beta_0(1 - f(t))X_1(t)\,,\\
\frac{dE_1(t)}{dt} &= \gamma_0 E_0(t) - (\varepsilon_1 f(t) + \mu_E)E_1(t) + \beta_1(1 - f(t))X_1(t)\,,\\
\frac{dX_1(t)}{dt} &= \varepsilon_0 f(t)E_0(t) + \varepsilon_1 f(t)E_1(t) + (\alpha_1 - \mu_1)X_1(t)\\
&\quad + ((\psi_1 - \gamma_1)f(t) - (\beta_0 + \beta_1)(1 - f(t)))X_1(t)\,,\\
\frac{dX_2(t)}{dt} &= \gamma_1 f(t)X_1(t) - \mu_2 X_2(t)\,.
\end{aligned} \tag{6.2}$$

A general solution of System (6.2), as a function of $f(t)$, requires solving a linear time-varying system of differential equations. We report the solution for two particular cases, $f(t) \equiv 0$ for all $t \in [0, 48]$; that is, cells are constantly lit (LOVTRAP domain unbound and signal never transduced); and $f(t) \equiv 1$ for all $t \in [0, 48]$; that is, cells are kept in the dark for the entire optogenetic experiment (LOVTRAP domain unbound and signal always transduced).

Let $(E_0(0),\, E_1(0),\, X_1(0),\, X_2(0)) = (E_0^0,\, E_1^0,\, X_1^0,\, X_2^0)$ be the initial condition of the system. When $f(t) = 0$ for all $t \in [0, 48]$, the dynamics of exhausted cells is described by

$$X_1(t) = X_1^0\, \mathrm{e}^{-(\mu_1+\beta_0+\beta_1-\alpha_1)t}, \quad X_2(t) = X_2^0\, \mathrm{e}^{-\mu_2 t}\,.$$

Moreover, for effector cells we have

$$\begin{aligned}
E_0(t) &= \frac{\beta_0\, X_1^0}{\gamma_0 + \mu_E + \alpha_1 - \alpha_0 - \mu_1 - \beta_0 - \beta_1}\left[\mathrm{e}^{-(\mu_1+\beta_0+\beta_1-\alpha_1)t} - \mathrm{e}^{-(\gamma_0+\mu_E-\alpha_0)t}\right]\\
&\quad + E_0^0\, \mathrm{e}^{-(\gamma_0+\mu_E-\alpha_0)t}\,,\\
E_1(t) &= \frac{X_1^0}{\mu_E + \alpha_1 - \mu_1 - \beta_0 - \beta_1}\left(\beta_1 + \frac{\gamma_0\beta_0}{\gamma_0 + \mu_E + \alpha_1 - \alpha_0 - \mu_1 - \beta_0 - \beta_1}\right)\\
&\quad \times \left(\mathrm{e}^{-(\mu_1+\beta_0+\beta_1-\alpha_1)t} - \mathrm{e}^{-\mu_E t}\right)\,.
\end{aligned}$$

We assume all parameters are non-negative, and that $\mu_E + \alpha_1 \neq \mu_1 + \beta_0 + \beta_1$ and $\gamma_0 + \mu_E + \alpha_1 \neq \alpha_0 + \mu_1 + \beta_0 + \beta_1$, to ensure the existence of a solution.

If cells are constantly stimulated, that is, $f(t) \equiv 1$ for all $t \in [0, 48]$, the dynamics of effector cells is described by

$$\begin{aligned}
E_0(t) &= E_0^0\, \mathrm{e}^{-(\mu_E+\gamma_0+\varepsilon_0-\alpha_0-\psi_0)t},\\
E_1(t) &= \frac{\gamma_0 E_0^0}{\gamma_0 + \varepsilon_0 + \varepsilon_1 - \alpha_0 - \psi_0}\left(\mathrm{e}^{-(\mu_E+\gamma_0+\varepsilon_0-\alpha_0-\psi_0)t} - \mathrm{e}^{-(\mu_E+\varepsilon_1)t}\right) + E_1^0 \mathrm{e}^{-(\mu_E+\varepsilon_1)t}\,,
\end{aligned}$$

and for the two populations of exhausted cells

$$
\begin{aligned}
X_1(t) = {} & \frac{E_0^0\left(\mathrm{e}^{-(\mu_E+\gamma_0+\varepsilon_0-\alpha_0-\psi_0)t}-\mathrm{e}^{-(\mu_E+\varepsilon_1)t}\right)}{\mu_1+\gamma_1+\alpha_0+\psi_0-\alpha_1-\psi_1-\mu_E-\gamma_1}\left(\varepsilon_0+\frac{\varepsilon_1\gamma_0}{\gamma_0+\varepsilon_0+\varepsilon_1-\alpha_0-\psi_0}\right) \\
& + \frac{\varepsilon_1\left(\mathrm{e}^{-(\mu_E+\varepsilon_1)t}-\mathrm{e}^{-(\mu_1+\gamma_1-\alpha_1-\psi_1)t}\right)}{\mu_1+\gamma_1-\alpha_1-\psi_1-\mu_E-\varepsilon_1}\left(E_1^0+\frac{\gamma_0E_0^0}{\gamma_0+\varepsilon_0+\varepsilon_1-\alpha_0-\psi_0}\right) \\
& + X_1^0\mathrm{e}^{-k_1t}\,, \\
X_2(t) = {} & \frac{\gamma_1E_0^0}{\mu_1+\gamma_1+\alpha_0+\psi_0-\alpha_1-\psi_1-\mu_E-\gamma_1}\left(\varepsilon_0+\frac{\varepsilon_1\gamma_0}{\gamma_0+\varepsilon_0+\varepsilon_1-\alpha_0-\psi_0}\right) \\
& \times\left(\frac{\mathrm{e}^{-(\mu_E+\gamma_0+\varepsilon_0-\alpha_0-\psi_0)t}-\mathrm{e}^{-\mu_2t}}{\mu_2+\alpha_0+\psi_0-\mu_E-\gamma_0-\varepsilon_0}-\frac{\mathrm{e}^{-(\mu_1+\gamma_1-\alpha_1-\psi_1)t}-\mathrm{e}^{-\mu_2t}}{\mu_2+\alpha_1+\psi_1-\mu_1-\gamma_1}\right) \\
& + \frac{\varepsilon_1\gamma_1}{\mu_1+\gamma_1-\alpha_1-\psi_1-\mu_E-\varepsilon_1}\left(E_1^0+\frac{\gamma_0E_0^0}{\gamma_0+\varepsilon_0+\varepsilon_1-\alpha_0-\psi_0}\right) \\
& \times\left(\frac{\mathrm{e}^{-(\mu_E+\varepsilon_1)t}-\mathrm{e}^{-\mu_2t}}{\mu_2-\mu_E-\varepsilon_1}-\frac{\mathrm{e}^{-(\mu_1+\gamma_1-\alpha_1-\psi_1)t}-\mathrm{e}^{-\mu_2t}}{\mu_2+\alpha_1+\psi_1-\mu_1-\gamma_1}\right) \\
& + \frac{\gamma_1X_1^0}{\mu_2+\alpha_1+\psi_1-\mu_1-\gamma_1}(\mathrm{e}^{-(\mu_1+\gamma_1-\alpha_1-\psi_1)t}-\mathrm{e}^{-\mu_2t})+X_2^0\mathrm{e}^{-\mu_2t}\,.
\end{aligned}
$$

We assume all parameters are non-negative and that $\gamma_0+\varepsilon_0+\varepsilon_1 \neq \alpha_0+\psi_0$, $\mu_1+\gamma_1+\alpha_0+\psi_0 \neq \alpha_1+\psi_1+\mu_E+\gamma_1$, $\mu_1+\gamma_1 \neq \alpha_1+\psi_1+\mu_E+\varepsilon_1$ and $\mu_2 \neq \mu_E+\varepsilon_1$ $\mu_2+\alpha_0+\psi_0 \neq \mu_E+\gamma_0+\varepsilon_0$, $\mu_2+\alpha_1+\psi_1 \neq \mu_1+\gamma_1$, so that a solution exists.

Here, the steady state solution of System (6.2) is computed. At steady state, the number of cells in each of the four populations is no longer changing with time. Steady states of a general system of ODEs can be found by setting the right hand side of each of the ODEs to zero and solving the resulting equations simultaneously. Given the *in vitro* experimental setting, we are modelling a close system where cells are placed in the wells at day 4 and measured at day 6, without any external influx of cells. Since we do not include an influx of cells, the steady state of the System (6.2) is given by $P^\star = (E_0^\star, E_1^\star, X_1^\star, X_2^\star) = (0,0,0,0)$ for a general choice of $f(t)$. In order to study the linear stability of the steady state, the Jacobian matrix of System (6.2) is calculated. In particular, we consider again the two cases $f(t) \equiv 0$ (signal OFF) and $f(t) \equiv 1$ (signal ON). We denote with J_{OFF} the Jacobian matrix of the system where $f(t) \equiv 0$ for all times $t \in [0, 48]$

hours,

$$J_{OFF} = \begin{pmatrix} \alpha_0 - \gamma_0 - \mu_E & 0 & \beta_0 & 0 \\ \gamma_0 & -\mu_E & \beta_1 & 0 \\ 0 & 0 & \alpha_1 - \mu_1 - \beta_0 - \beta_1 & 0 \\ 0 & 0 & 0 & -\mu_2 \end{pmatrix}.$$

Since the system is linear, J_{OFF} is also the Jacobian matrix of the steady state $P^\star = (0, 0, 0, 0)$, and one can study its stability. Matrix J_{OFF} has four eigenvalues,

$$e_1 = \alpha_0 - \gamma_0 - \mu_E, \quad e_2 = -\mu_E, \quad e_3 = \alpha_1 - \mu_1 - \beta_0 - \beta_1, \quad e_4 = -\mu_2 \,.$$

Therefore, the steady state $P^\star = (0, 0, 0, 0)$ is stable if and only if

$$\alpha_0 < \gamma_0 + \mu_E, \quad \text{and} \quad \alpha_1 < \mu_1 + \beta_0 + \beta_1 \,.$$

Similarly, we denote with J_{ON} the Jacobian matrix of the system where $f(t) \equiv 1$ for all times $t \in [0, 48]$ *hours*,

$$J_{ON} = \begin{pmatrix} \alpha_0 + \psi_0 - \varepsilon_0 - \gamma_0 - \mu_E & 0 & 0 & 0 \\ \gamma_0 & -(\varepsilon_1 + \mu_E) & 0 & 0 \\ \varepsilon_0 & \varepsilon_1 & \alpha_1 + \psi_1 - \mu_1 - \gamma_1 & 0 \\ 0 & 0 & \gamma_1 & -\mu_2 \end{pmatrix}.$$

Since the system is linear, J_{ON} is also the Jacobian matrix of the steady state $P^\star = (0, 0, 0, 0)$, and one can study the stability. Matrix J_{ON} has four eigenvalues,

$$e_1 = \alpha_0 + \psi_0 - \varepsilon_0 - \gamma_0 - \mu_E, \quad e_2 = -(\varepsilon_1 + \mu_E), \quad e_3 = \alpha_1 + \psi_1 - \mu_1 - \gamma_1, \quad e_4 = -\mu_2 \,.$$

Therefore, the steady state $P^\star = (0, 0, 0, 0)$ is stable if and only if

$$\alpha_0 + \psi_0 < \varepsilon_0 + \gamma_0 + \mu_E, \quad \text{and} \quad \alpha_1 + \psi_1 < \mu_1 + \gamma_1 \,.$$

For both cases of $f(t) \equiv 0$ for all $t \in [0, 48]$ and $f(t) \equiv 1$ for all $t \in [0, 48]$, the steady state $P^\star = (0, 0, 0, 0)$ is stable in case of decreasing exponentials in the solutions of $E_0(t)$, $E_1(t)$, $X_1(t)$ and $X_2(t)$. Otherwise, both effector and exhausted population of cells would grow without limits.

6.3.2 Structural identifiability analysis

In order to understand if it is possible to estimate all model parameters, given the model's structure shown in Figure 6.5, we carry out a structural identifiability analysis, following the method proposed by Castro & de Boer (2020). This method is a rigorous but simple way to test the identifiability of a linear (or nonlinear) model with constant parameters. For mathematical details we refer to Section 2.4. The method involves writing each of the four ODEs in System (6.2) as functionally independent terms. Details of the method applied to the model shown in Figure 6.5 and defined by the differential System (6.2) are given in Appendix D.

As detailed in Appendix D, the model defined in System (6.2) with four state variables, $E_0(t)$, $E_1(t)$, $X_1(t)$, $X_2(t)$, and thirteen parameters, is not structurally identifiable given that only populations X_1 and X_2, for reversible and terminally exhausted cells, can be set as observable (experimentally measured) in the analysis. Indeed, the preliminary experimental data sets, used for mathematical modelling, consider reversible and terminally exhausted cells after the 48 *hours* of the optogenetic experiment only.However, a recent work of Massonis *et al.* (2021) shows that having a parameter changing over time can improve the observability and, thus, the identifiability of the model. This can be explained as follows: a symmetry between constant parameters can result in changes in one parameter being compensated by changes in the second one; however, this is not the case if only one of these two parameters varies over time. Moreover, identifiability can also be improved by including populations of effector cells E_0 and E_1 as two more observable outputs, *i.e.,* experimentally measured effector cells. Indeed, if these two populations of progenitor and terminally differentiated effector cells could be measured at the end of the optogenetic experiment, then the defined model would become structurally identifiable (see Appendix D for details).

Conscious of the model's limitations, we found meaningful, for the educational purposes of this PhD thesis, to consider the model in Figure 6.5 and carry out sensitivity analysis and parameter inference based on the provided experimental data, *i.e.,* the two exhausted cell states $X_1(t)$ and $X_2(t)$ measured at $t = 48$ *hours* from the beginning of the optogenetic experiment.

6.3.3 Global sensitivity analysis

In this Section, we investigate which of the thirteen parameters of the mathematical model defined in System (6.2) are most influential to model output or any derived summary statistics. We denote by $\boldsymbol{\theta}$ the vector of thirteen parameters to be estimated,

$$\boldsymbol{\theta} = \left(\mu_E,\, \mu_1,\, \mu_2,\, a_0,\, a_1,\, \psi_0,\, \psi_1, \varepsilon_0,\, \varepsilon_1,\, \beta_0,\, \beta_1,\, \gamma_0,\, \gamma_1\right),$$

and we apply the Sobol method described in Section 2.4.

The Sobol sensitivity analysis was carried out using the SALib package in Python, to find the total-order Sobol index for each of the parameters in $\boldsymbol{\theta}$. The algorithm takes the vector $\boldsymbol{\theta}$ as input and studies how varying parameters in ranges affect model output. The range $[10^{-3}, 1]\, h^{-1}$, reflecting a uniform prior distribution, is assigned to each parameter in $\boldsymbol{\theta}$.

When considering the model output, we need to remember that the experimental data quantify the percentages (over the total number of cells in the system) of reversible and terminally exhausted cells after 48 *hours* of a optogenetic experiment. We remind the reader that the experiment focuses on understanding how the development of exhausted cells is influenced by different durations of antigen stimulus. Thus, we focus on the terminally exhausted population and investigate the parameters' influence on the summary statistics given by the percentage of terminally exhausted cells at the end of the optogenetic experimental time course, namely $\frac{X_2(t)}{E_0(t)+E_1(t)+X_1(t)+X_2(t)}$ at $t = 48$ *hours*. As in Section 6.2 and previously in this Section, we consider a step signal function $f(t)$ defined by (6.1); we denote with $l_{dd} \in \boldsymbol{d}$ the dark delay; that is, the time cells spend in the dark before light is switched on (and, thus, signal interrupted). In this Section, we aim to perform a global sensitivity analysis under different conditions and, for the sake of simplicity, we restrict ourselves to analyse the output changes under three different durations of antigen stimulation. We consider the case the signal is never transduced, the case it is transduced for half of the considered time window, and the case that is always transduced; that is $l_{dd} \in \left(0,\, 22.67,\, 48\right)$. The results of this analysis are plotted in Figure 6.6 for $l_{dd} = 0$ *hours* of signals (on the top plot), $l_{dd} = 22.67$ *hours* of signal (the lower left plot) and $l_{dd} = 48$

hours of stimulation (in the lower right). From Figure 6.6, it can be seen that different stimulation durations have an impact on which parameter influences the normalised ratio of terminally exhausted cells most. In case of absence of signal (top plot), the model output is mainly influenced by the death rate of effector cells, μ_E, and the homeostatic growth rate of reversible exhausted cells, α_1. If the Sobol index of μ_E is decreasing over time, the growth rate of X_1 if increasing; this can be explained because effector cells are the main population at the start of the optogenetic experiment, but, in the absence of antigen signal, they poorly proliferate and their effect on exhausted cells is limited. Thus, the dynamics of terminally exhausted cells with respect to the total cell population are mainly driven, for later time points, by the homeostatic proliferation rate of reversible cells in X_1. Also, we might have expected a greater influence of the death rate of terminally exhausted cells, μ_2. Indeed, one can observe the normalised ratio of terminally exhausted cells X_2 over the total cell population greatly influenced by μ_2 only for early time points; this can be explained as follows: if not stimulated by antigen, reversible cells X_1 do not differentiate to a terminally exhausted state and thus, this latter population dies out already at early time points.

On the contrary, if cells are stimulated for a few hours, ε_1 is the most influential parameter (see two lower plots in Figure 6.6). In particular, if LOVTRAP domain is bound only for the first 22 *hours* (lower left plot), one can observe how, as soon as light is switched on, parameters influence on the normalised ratio of terminally exhausted cells X_2 is slowly drifting to the case of $l_{dd} = 0\,h$ (top plot). Indeed, rates α_1, μ_E and β_i, for $i = 0, 1$ (the homeostatic proliferation rate of reversible exhausted cells, the effector death rate and the reinvigoration rates, respectively) are subjected to an increase in their importance over time window of absence of antigen stimulation, *i.e.,* interval [22.67, 48] *hours*.

We may want to stress to the reader that the plots of Figure 6.6 report the Sobol index of all model parameters; however, for the three dark delays considered, some parameters have a very low total Sobol index value. We can assume that a parameter with mean total order Sobol index less than 0.15 is of little influence to the considered model output. These parameters include, for instance, the death rate of reversible exhausted cells μ_1, as well as the reverse transition rates β_0 and β_1.

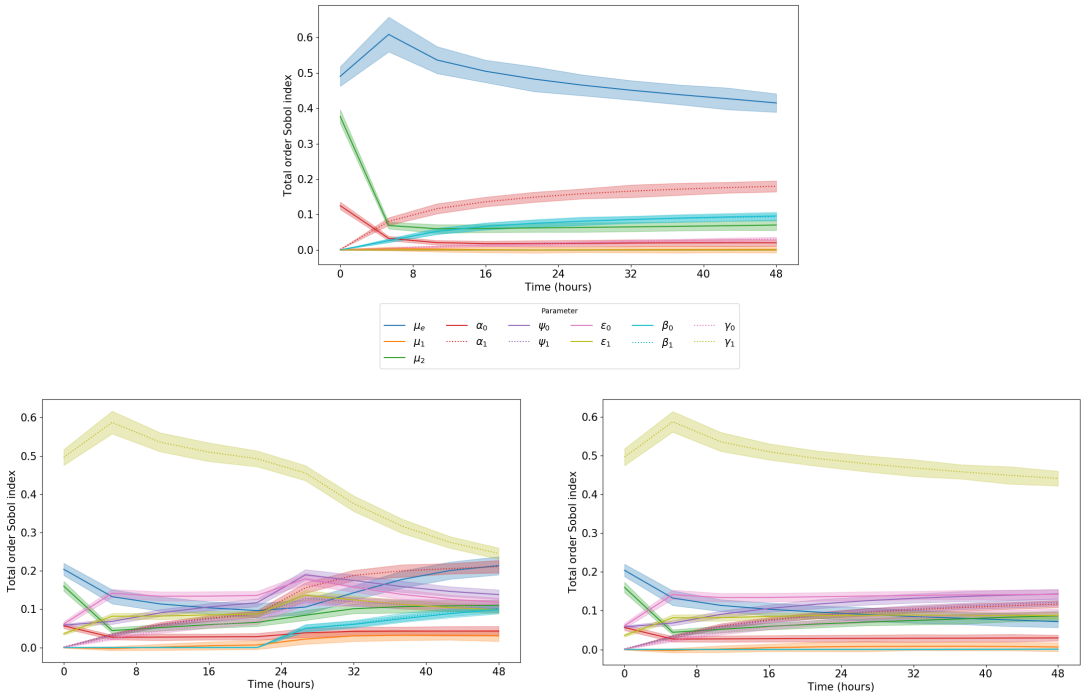


Figure 6.6: Means and 95% confidence intervals of the total Sobol indices for the parameters of the model shown in Figure 6.5. The ratio of terminally exhausted cells over the total number of cells in the system at the end of the optogenetic experiment; that is $\frac{X_2(t)}{E_0(t)+E_1(t)+X_1(t)+X_2(t)}$ for $t = 48$ *hours*, has been considered as model output. From left to right, cells have been exposed to $0, 22.67, 48$ *hours* of stimulation within the optogenetic experiment; a step function $f(t)$ has been used.

6.3.4 Bayesian parameter inference

In this Section, we consider the toy model shown in Figure 6.5 and employ an Approximate Bayesian Calculation method to predict the posterior distribution of model parameters. Technical details of this method are discussed in Section 2.3. Starting from uniform prior distributions for model parameters, the aim is to narrow down parameters ranges and, thus, identify a biological timescale of exhaustion. Moreover, we make use of a Bayesian approach to investigate if any of the considered cell events is influenced by the presence of antigen. Indeed, as outlined in Section 6.1, the output of six different repeats of the optogenetic

experiment suggests an oscillatory trend of the normalised ratio of reversible and terminally exhausted cells. Figure 6.2 shows an example for a single repeat.

With the aim of studying a possible oscillatory behaviour of parameters values over the different dark delays, we start making use of an ABC method to infer parameter values for any of the considered dark delays. For this we characterise the experimental data sets as $data(j, l_{dd})$, resulting from experiment $j \in \{1, 2, 3, 4, 5, 6\}$ and initial dark delay l_{dd}. For simplicity, we restrict the analysis to a meaningful subset of $\boldsymbol{d} \in \mathbb{R}^{36}$: we consider seven dark delays equally distributed among the 48 *hours* of the experimental duration, $l_{dd} \in \big(0,\ 8,\ 16,\ 24,\ 32,\ 40,\ 48\big)$ *hours*.

As outlined previously in this Section, given the experimental data sets and the mathematical model structure, the model is not structurally identifiable. Thus, in this Section, we carry out a Bayesian analysis assuming the homeostatic proliferation rate of progenitor effector cells and the death rate of reversible exhausted cells to be measurable. Such model, with parameters α_0 and μ_1 fixed, is structurally identifiable and, thus, an informative ABC analysis might be performed. Parameters α_0 and μ_1 could be experimentally estimated, and it is room for future work to perform the proposed analysis with more refined values of the two fixed parameters. Alternatively, additional experimental outputs for the population of effector cells can be considered and, in this case, a distance comprising cell states E_0 and E_1 could be applied. Overall, we would like to remind the reader that the aim of this Section is to learn the technique in view of reproducing the same analysis when further experimental data or parameter estimates could be generated. Here, we assume $\alpha_0 = 0.08$ h^{-1} and $\mu_1 = 0.1$ h^{-1} and estimate the parameters in $\hat{\boldsymbol{\theta}}$ via the ABC algorithm,

$$\hat{\boldsymbol{\theta}} = \big(\alpha_1, \psi_0, \psi_1, \mu_E, \varepsilon_0, \varepsilon_1, \beta_0, \beta_1, \gamma_0, \gamma_1, \mu_2\big)\,.$$

Prior distributions for the parameters in $\hat{\boldsymbol{\theta}}$ correspond to the uniform distributions in the same ranges considered for the global sensitivity analysis; that is, $[10^{-3}, 1]\, h^{-1}$. In this way, the prior distributions of all parameters are defined in the intervals tested by the Sobol algorithm. Simulated summary statistics for the populations of exhausted cells, at time point $t = 48$ *hours*, were generated from

the System (6.2), and compared with ratios of exhausted cells provided by experimental data sets. The algorithm was run 10^7 times and only the top 10^2 results where considered. For each dark delay considered, statistics of the six experimental repeats have been put together and a generalised Euclidean distance has been used to compare simulated and experimental statistics for $j \in \{1, 2, 3, 4, 5, 6\}$; that is

$$\delta(l_{dd}, n_{sim}) = \Bigg[\left(\left(\frac{X_1(t = 48h; l_{dd})}{N_t(t = 48h; l_{dd})} - \mu_{X_1}(l_{dd}) \right) / \sigma_{X_1}(l_{dd}) \right)^2 + \left(\left(\frac{X_2(t = 48h; l_{dd})}{N_t(t = 48h; l_{dd})} - \mu_{X_2}(l_{dd}) \right) / \sigma_{X_2}(l_{dd}) \right)^2 \Bigg]^{1/2}, \quad (6.3)$$

where μ_{X_i} and σ_{X_i} are the mean and the standard deviation of the percentage of exhausted cells X_i, for $i = 1, 2$, over the six experiment repeats. We note that we consider $l_{dd} \in \big(0,\ 8,\ 16,\ 24,\ 32,\ 40,\ 48\big)$ *hours*.

Initial conditions for the number of cells in each of the four states have been fixed according to the experimental conditions, and aligned with the mathematical model reported in Section 6.2. In particular, we let $N^\star = 25 \times 10^3$ as the total number of live and transduced cells, *e.g.,* cells present in our system, and set

$$(E_0(0), E_1(0), X_1(0), X_2(0)) = (0.5\, N^\star, 0.4\, N^\star, 0.05\, N^\star, 0.05\, N^\star)\,.$$

Posterior distributions for each of the parameters in $\hat{\boldsymbol{\theta}}$ and for $l_{dd} \in \big(0,\ 8,\ 16,\ 24,\ 32,\ 40,\ 48\big)$ *hours* have been computed via the ABC algorithm. Figure 6.7 displays an example, for $l_{dd} = 24$ *hours*, of the probability histograms of the uniform sample priors (in red) and of the posterior distributions (in green) for each of the parameters in $\hat{\boldsymbol{\theta}}$. One can notice that for some parameters the Bayesian inference method provides very little learning: distributions of parameters α_1, ψ_1 and ε_1 are slightly shifted to the left compared to the uniform distribution. Moreover, the Bayesian parameterisation did not reveal any new information for parameters ψ_0, ε_0 and γ_0. Whereas, a significant learning occurs for parameters μ_E, μ_2, β_0, β_1, γ_1, whose posterior distribution is shifted to the right around one order of magnitude $[-1, 0]$.

Results from the global sensitivity analysis (Sobol index) indicate that parameter γ_1 is one with the strongest influence on the summary statistics $\frac{X_2(t=48h)}{N_t(t=48h)}$

model output for $l_{dd} = 24$ *hours* (see Figure 6.6). The exhaustion rate from reversible to terminally exhausted cells, γ_1, is characterised by an order of magnitude between -1 and 0; that is, two times higher than other parameters relative to X_1, such as α_1, ε_1 and ψ_1.

Similarly, posterior distributions for dark delays $l_{dd} \in \big(0,\ 8,\ 16,\ 24,\ 32,\ 40,\ 48\big)$ have been computed; overall, given the high number of parameters compared to the number of experimental measurements, the Bayesian learning is small for many of the parameters when compared to their uniform prior distributions. Thus, a quantification of all the individual parameter values in the model by inferring their posterior distributions was not possible.

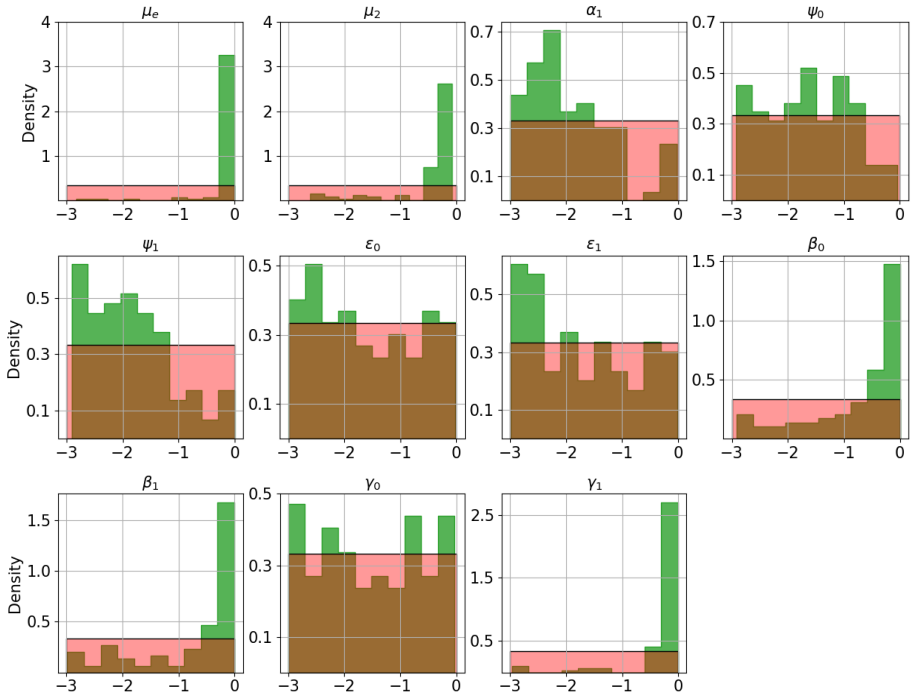


Figure 6.7: Posterior distribution (in green) for the parameters in $\hat{\boldsymbol{\theta}}$ of the mathematical model depicted in Figure 6.5. Posterior distributions, for $l_{dd} = 24$, have been computed considering the best 10^2 parameter sets resulting from ABC analysis with 10^7 simulations; distance (6.3) has been used. On the x-axis, $\log_{10}(param)$ $hours^{-1}$. Prior distributions are shown in red.

Posterior distributions of parameters in $\hat{\boldsymbol{\theta}}$ computed for $l_{dd} \in \big(0,\ 8,\ 16,\ 24,\ 32,\ 40,\ 48\big)$ *hours* can be used to investigate their dependence on the duration of

antigen stimulation. Figure 6.8 displays the trend of parameters in $\hat{\boldsymbol{\theta}}$ according to the results of the posterior distribution computed with the Approximate Bayesian analysis for the toy model defined by System (6.2). One can observe that, as suggested from the learning on the posterior distributions, for many of the model parameters there is not a clear trend over the different dark delays. However, parameters β_0 and β_1 clearly show a decrease in their order of magnitude as cells are stimulated for longer time. This can be explained given that β_0 and β_1 are the rates at which exhausted cells X_1 can differentiate back to an effector phenotype; however, when cells are stimulated by the antigen for longer time, exhausted cells are less likely to revert to an effector state, either a progenitor E_0 or a terminally differentiated type E_1. Overall, we should mention the wide 95% credible interval that characterises many of the posterior distributions for parameters in $\hat{\boldsymbol{\theta}}$. Thus, we find that our conclusions about the existence of a trend in parameters over the different duration of dark delay are not very strong.

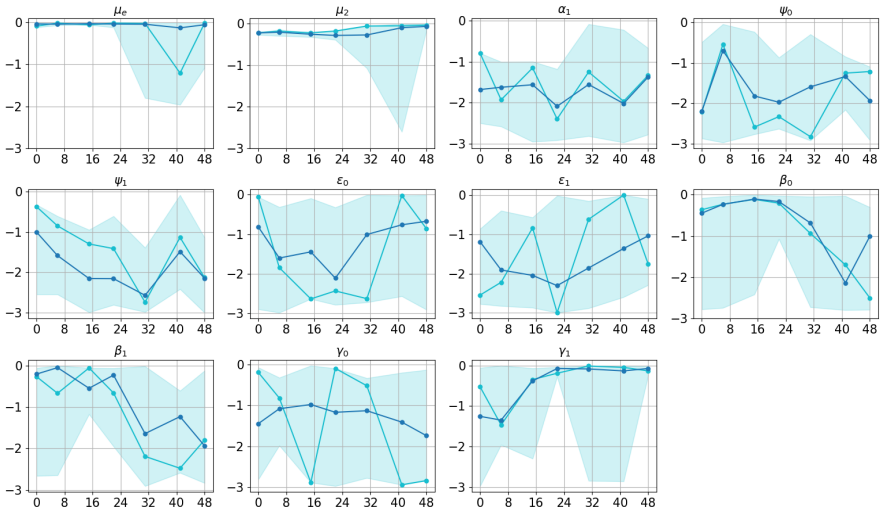


Figure 6.8: Trend of parameter values over duration of stimulation (dark delay). We report the point-wise median (in blue) and best (in light blue) values over the best 10^2 parameters set resulting from the ABC analysis with 10^7 simulations; the shaded area shows the 95% credible interval of model simulations. On the y-axis order of magnitude of the parameters, with $\hat{\boldsymbol{\theta}} \in [10^{-3}, 1]^{11}$. On the x-axis the considered duration of dark delay.

6.4 Model justification and limitations

In this Section, we justify the structure of the defined mathematical models based on other published studies present in the literature about CD8$^+$ T cell exhaustion and describe the many limitations of the proposed models. Understanding the origins of exhaustion as well as devising ways of preventing it are both areas of intense focus nowadays, as they bring promise of improvement in cancer immunotherapy (McLane *et al.*, 2019; Tabana *et al.*, 2021).

Current mathematical models do not focus on an antigen-driven differentiation process towards terminal exhaustion (Eftimie *et al.*, 2016). They mainly consider effector cells (either endogenous or CARs) and their dynamics with respect to tumour cells or antigen; memory or näive cells dynamics sometimes are included as well. Exhausted cells, as a cell state, do not contribute to the immune-induced death of cancer or chronically infected cells. Hence, exhaustion of CD8$^+$ T cell is modelled similarly to cell death: effector cells disappear from the system and no pool of exhausted cells is generated. In other words, the immune system is investigated mainly to show how näive, memory or effector cells behave with respect to tumour cells (Owens & Bozic, 2021; Serre *et al.*, 2016). Exhaustion of CD4$^+$ T cells has been modelled in a similar way; see, for instance, Reference Makhlouf *et al.* (2020). A different approach is followed by Wodarz *et al.* (1998): dynamics of näive, memory and effector T cells, as well as virus infected cells are modelled. In order to accommodate exhaustion into the model, they define a dummy parameter and study its regime for the infection to persist or be eradicated. Similarly, Stromberg & Antia (2012) describe the density of CD8$^+$ T cell via a partial differential equation in terms of time and a dummy variable which represents the exhaustion level as a continuous variable.

Recent developments in cell engineering methods kicked off *in vitro* and *in vivo* studies about the persistence and efficacy of engineered T cells (chimeric antigen receptor cells, also known as CARs) to enhance the immune system fighting cancer. Following hand in hand the biological results are the mathematical studies representing CARs dynamics in tumour micro-environment; in this regard, the work of Sahoo *et al.* (2020) is worth a mention as pioneer. Moreover, interactions

between CAR T cells and endogenous T cells have been recently mathematically investigated by Kimmel *et al.* (2020, 2021).

From a mathematical perspective, most of the models representing the immune cell response to antigen, under a chronic infection or in the presence of cancer, are deterministic, *i.e.,* comprised of ordinary differential equations (Eftimie *et al.*, 2016). Here, the interactions between immune cells, *i.e.,* B cells, dendritic cells, or näive, effector, memory $CD8^+$ T cells, and tumour cells are investigated (Bocharov, 1998; Kuznetsov *et al.*, 1994; León-Triana *et al.*, 2021; Moore & Li, 2004; Nikolopoulou *et al.*, 2018; Owens & Bozic, 2021; Serre *et al.*, 2016; Valentinuzzi & Jeraj, 2020; Wodarz *et al.*, 1998). For example, Boer & Perelson (2013) suggest a predator-prey type model for the response of näive, memory and activated effector cells to an exponentially growing pathogen, *i.e.,* a bacteria.

Even though remarkable progress has been made in recent years in characterising the exhausted state from an immunological perspective (Chen *et al.*, 2019; Mo *et al.*, 2021; Toettcher *et al.*, 2013; Weber *et al.*, 2021; Wherry *et al.*, 2007), a deep understanding of the molecular and genetic mechanisms, as well as the influence that stimulus duration and antigen load have on cell differentiation and proliferation is still lacking. As mentioned above, many models of cellular immune responses to infection (acute or chronic) are structured with deterministic system (of differential equations) as predator-prey models in ecology. Here, the proliferation rate of immune cells is proportional to the density of antigen or tumour cells in a linear or logistic fashion. For example, Johnson *et al.* (2011) developed a virus-host interaction model to describe, during an acute or chronic infection, the interplay between uninfected cells, virus-infected (or tumour) cells and $CD8^+$ T cells characterised by their level of exhaustion. Again, only dynamics of functional effector cells, not of exhausted cells, are modelled: exhaustion and death are equivalently brought together as exit from the effector functional state.

T cell responses, upon antigen-induced activation, is promoted by a range of different stimuli and signals: T cell receptor, the co-stimulatory signal from antigen presenting cells and the signal coming from circulating cytokines (see

Section 1.1 for more details). Several studies, focus on analysing, in a quantitative way, how these signals are integrated and how this modulates the strength of the T cell response. Studies of Marchingo *et al.* (2016, 2014) are pivotal in modelling T cell stimuli integration and the consequent regulation of the division process. Also, we would like to mention the importance of proliferation in the activation and differentiation process during a chronic infection. The work of Boer *et al.* (2001) focuses on the quantification of the major parameters, such as proliferation, of the CD8$^+$ cells during lymphocytic choriomeningitis virus (LCMV) infection. Moreover, the work of Antia *et al.* (2003) is a pioneer in proliferation program studies; here, the question of an antigen-independent proliferation program for CD8$^+$ T cell is investigated in an acute framework. Their findings suggest that a continuous exposure to antigen, in a short window after infection, might subsequently modify the cellular program established at the first encounter.

The aforementioned mathematical models in the literature do not consider the heterogeneous spectrum of effector CD8$^+$ T cell during their differentiation into more terminally differentiated effector states or an exhausted state, both progenitor or more terminal ones. On the contrary, as part of this PhD project, we focus, in collaboration with the Salk Institute of Biological Studies, on the development of deterministic mathematical models describing dynamics of progenitor and terminally differentiated effector and exhausted cells. In particular, with the purpose of investigating possible pathways of the T cell exhaustion process during chronic infection or cancer, we define, in Sections 6.2 and 6.3, mathematical models differing in the number of differentiation states; so that different biological characterisation has been assigned to both effector and exhausted states. As a clear biological understanding of the process is still lacking, preliminary work to conceptualise model structures representing different hypotheses regarding heterogeneity of CD8$^+$ T cell exhaustion has been carried out. Not many experimental studies have been published about the exhaustion-differentiation process; several, and sometimes controversial, hypotheses have been made about the differentiation pathway of näive CD8$^+$ T cells to an effector and exhausted state, as well as about the number of key differentiation states and their biological meaning.

During the entire length of this PhD, a collaboration with Prof. Kaech and Dr. Mann from the Salk Institute of Biological Studies has been kept vivid, and an extensive literature research about the state of the art of the exhaustion process from a biological point of view has been carried out. In parallel, mathematical methods to address model analysis and parameter inference have been learnt. These same techniques could be applied to more refined model structures that might be conceptualised as part of future work with the advance of experimental results, *i.e.,* measurement of more time points, definition of markers for other cell populations of interest, analysis of transcription factors and RNA-seq data, as well as estimation of proliferation rates. Conscious of the limitations of our models, in the remaining part of this section, we go through them, suggesting possible routes forward.

A first limitation, valid for all models defined in Sections 6.2 and 6.3, is the low number of measured outputs, both in terms of time points and in terms of cell populations characterised by specific biological functions. Indeed, the provided experimental output consists of one measurement, at the end of the optogenetic experiment, of two exhausted populations, characterised as TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$. Effector and exhaustion markers have been considered during the flow analysis, such as TIM-3, TCF1, PD-1, TOX (see Section 1.1 for more details); however, only two types of exhausted states, and no quantification of effector cells, have been provided as measured output. This is the case, since a clear agreement about a characterisation of effector differentiation states via specific flow cytometry markers has not been established (McLane *et al.*, 2019; Winkler & Bengsch, 2020). In addition, the heterogeneity of the initial state of the system has not been quantified: a total number of $N^\star = 25 \times 10^3$ cells has been placed in the system, but a quantification with respect to the presence of different effector or exhausted states was missing. Hence, only an estimate, made by the experimental collaborators, has been used as initial condition of the cell states defined in the mathematical models.

The restriction on output data, is accompanied by a related important limitation: the defined models have a high number of parameters representing cell events and transitions between multiple effector states or to exhausted states. This high dimensional parameter space comes with a few measured experimental

data sets (measurement of two populations at one time point). The implemented ABC algorithm is considered computationally efficient (Toni & Stumpf, 2010; Toni *et al.*, 2009) but, given the large number of variables and parameters in each of the mathematical models, we observe rather little learning, in the Bayesian inference, for some of the parameters. As stated in Section 6.3.2, model identifiability (concerning uniqueness of the model parameters determined from experimental observations) comes with a quantification of additional cell populations (*e.g.,* effector states) or with the estimation of parameter values (*e.g.,* death or proliferation). However, the implementation of a more efficient algorithm, such as ABC-SMC, might help improving parameter inference results (Toni & Stumpf, 2010).

In this regard, we remind the reader that the performed analysis of parameter space (ABC, Sobol index and identifiability) would have benefited from the availability of intermediate time points during the optogenetic experiment. At the time of writing this thesis, intermediate time points have not been measured yet, due to technical limitations: the staining procedure performed before the flow cytometry analysis does not result in constant antibody staining across samples. Thus, when planning to compare cell data sets at different time points, experiments should be started at different times and then measured simultaneously. A quantification of further populations of effector and exhausted cells, as well as the measurement of consecutive time points, can be considered part of future work for the experimental plan.

In addition, the experimental data that we used for the modelling quantify TIM-3$^-$ TCF1$^+$ and TIM-3$^+$ TCF1$^-$ population of cells (namely progenitor reversible and terminally differentiated exhausted cells) in terms of their ratio normalised over the total number of cells. This has implications for the mathematical modelling. In particular, when considering the Bayesian analysis and model selection, we have been computing a generalised Euclidean distance between the output summary statistics, *i.e.,* ratio of reversible and terminally exhausted cells over the total population of live and stained cells present at the end of the optogenetic experiment (that is, time $t = 48$ *hours*). A better comparison between experimental and simulated output would have been possible if considering absolute values, *i.e.,* number of cells in each population.

Proliferation and cell expansion have a crucial importance in a T cell response, and in giving rise to several cell phenotypes (Bresser *et al.*, 2022; Wherry *et al.*, 2007). Differentiation to an effector or exhausted phenotype occurs at the same time as extensive proliferation (Antia *et al.*, 2003; Buchholz *et al.*, 2013; Gerlach *et al.*, 2013; Kaech & Ahmed, 2001; Plambeck *et al.*, 2022). However, the specific role of proliferation in an exhaustion framework has not been clarified yet. In this Chapter, two different hypotheses are considered. In Model 1 and 2 of Section 6.2, cell division is linked to cell differentiation as happening with the same antigen-dependent transition, *i.e.,* when a cell divides both of the two daughters belong to the more differentiated effector state; whereas in the model defined in Section 6.3 (see Figure 6.5) less differentiated effector cells can undergo both homeostatic and antigen-dependent proliferation (at rate α_0 and ψ_0, respectively); moreover, they are able to differentiate to a terminal effector state in an antigen-dependent fashion (at rate γ_0). However, at the time of writing this thesis, experiments tracking cell proliferation, *i.e.,* number of cell divisions, in the context of optogenetic experiments, have not been performed. Hence, we leave for future work any further investigation about modelling techniques to study the role of cell division in the effector population during the exhaustion-differentiation process. For example, the use of the green fluorescence protein enables measurement of cell divisions; if this is a simple and limited approach, novel studies enhance the tracking possibilities: Bresser *et al.* (2022) develop a genetic-tracing approach that allows for the measurement of the number division of cells in a specific pool, even over extensive rounds of division, without the limitation of fluorochrome decay.

In all the models defined in Sections 6.2 and 6.3, we quantify the influence of antigen on the cells' behaviour for a signal density function, $f(t) \in [0, 1]$. In particular, we let $f(t)$ be a step function with value 0 if the antigen signal is absent (cells in the light) and value 1 in its presence (cells in the dark). Smoother functions such as an hyperbolic tangent function can be considered for the definition of $f(t)$; indeed, a sigmoid function might better represent the transient re-binding in a short period of the two parts of the LOVTRAP domain. Moreover, one could refine this simple way of modelling the antigen signal perceived by T cells for a more appropriate approach, such as modelling antigen dynamics.

One can question the choice of antigen-dependent rates. For example, in the model with four cell states outlined in Section 6.3, we make antigen-dependent the following events (Kaech & Cui, 2012; Wherry & Kurachi, 2015): proliferation of effector and exhausted cells; differentiation rates from effector to exhausted states; differentiation from the reversible to the terminally exhausted state. However, experimental findings might suggest differently; the oscillatory behaviour of ratios of exhausted populations over dark delay's duration reported in Figure 6.2, might imply that the death rate of exhausted cells is also varying in an antigen-dependent fashion. This consideration could be included in the model's structure by letting the death rates depend on the signal function $f(t)$.

Constraints and dependencies between parameters can be also considered. In all the models, cell proliferation, differentiation and death processes have been modelled in a linear fashion; however, more complex dependencies between growth and differentiation might best apply to the exhaustion framework, *i.e.,* a non-linear or logistic growth for effector and/or reversible exhausted cells. In addition, one can include biologically meaningful constrains between parameters in the model structure. For example, in the two models reported in Section 6.2, exhaustion-differentiation rates, namely ε_1, ε_2, ε_3, have been uniformly and independently sampled. However, an interesting question, still open in the exhaustion field, is whether early differentiated effector cells commit to an exhausted phenotype more or less likely than more terminally differentiated effector cells. Indeed, as mentioned in Section 1.2, progenitor exhausted cells have large similarities, in terms of their behaviour and expressed surface markers, with memory precursor cells arising in an acute infection. Thus, an exhausted phenotype might arise directly from less differentiated effector cells which have experienced only an antigen stimulus of short duration. On the other hand, differentiation of CD8$^+$ T cells to more terminal effector states is an intrinsic characteristic of antigen exposure; hence, terminally differentiated effector cells might play a main role in maintaining the pool of exhausted cells. This can be investigated, for example, by defining two different model structures with constrains, $\varepsilon_1 < \varepsilon_2 < \varepsilon_3$ or $\varepsilon_1 > \varepsilon_2 > \varepsilon_3$, for increasing and decreasing rates, respectively. A model selection analysis should then follow to determine which model structure is more likely given the experimental data.

Overall, the mathematical analysis performed in this Section is based on well known and powerful techniques to analyse model structure, establish its sensitivity, and infer parameter values. However, the mathematical output relies on experimental data sets and the results reported in this Chapter leave room for further considerations and improvements requiring additional mathematical studies, as well as experiment refinement.

6.5 Discussion

Upon acute infection, T cell responses are generally able to lead to a large number of effector cells, which can eliminate the infection, as well as develop immunological memory to protect against future re-encounters of the same pathogen. In contrast to acute infections, in chronic infections and cancer, the long-term behaviour of most effector cells is to become exhausted, and the formation of the memory pool is impaired (McLane *et al.*, 2019). However, in both cases T cells differentiate into a heterogeneous population, with each subset of cells characterised by different phenotypes and functions (Gerlach *et al.*, 2013; Kaech *et al.*, 2002a; Sallusto *et al.*, 1999).

A key pursuit in $CD8^+$ T cell immunology is to untangle the heterogeneous cell differentiation process to exhaustion and understand how acute and chronic infections differ from very early times, and lead to an exhausted progenitor phenotype to arise. Understanding the $CD8^+$ T cell response to a chronic antigen stimulus has major implications. It will help clarify the complex process of TCR response upon activation, as well as the intra-cellular signalling mechanism. Moreover, it might contribute to fight against chronic infections and tumours; indeed, immunotherapies received a lot of attention in the medical field and, recently, many different therapeutic approaches have been developed. For more details we refer the reader to Chapter 1 as well as to the review works of Waldman *et al.* (2020) and Raskov *et al.* (2021). Due to the advent of novel experimental techniques significant progress has been made in understanding the $CD8^+$ T cell response and the differentiation process; as advances in the field one could mention that TCR activation signalling, transcriptomic and epigenetic landscape have been studied at the single cell level (Chen *et al.*, 2019; Heinzel *et al.*, 2018; Hudson *et al.*, 2019);

or that the proliferation program and drug interaction in exhaustion fate have been investigated (Boer & Perelson, 2013; Hashimoto *et al.*, 2018; Marchingo *et al.*, 2014; Wherry & Kurachi, 2015). However, challenges remain still open as these studies clarify, only partially, the rules governing the dynamics of immune cells. They do not, for example, tell us how the cell response is specified at early time points after antigen encounter, and whether it can be subsequently modified by the persistency or load of antigen. In addition, a characterisation of several populations emerging during the exhaustion-differentiation process is still missing; finally, single cell chromatin observations have not been related to the population behaviour and a comprehensive framework of the reinvigoration process of exhausted CD8$^+$ T cells has not been developed yet.

As part of the QuanTII, Horizon 2020 Marie Skłodowska-Curie Action project, the process of CD8$^+$ T cell exhaustion has been studied, from a mathematical perspective, during this PhD; in this Chapter, possible trajectories of the T cell differentiation pathway to an exhausted state have been investigated. In particular, a collaboration with the Salk Institute of Biological Studies (La Jolla, USA) has been established for this doctoral study; hence, we focus on modelling *in vitro* dynamics of CD8$^+$ CAR T cells at early time points of the performed optogenetic experiments. We develop and analyse deterministic mathematical models representing the differentiation process to exhaustion, as characterised by several differentiation steps, from an early effector to a progenitor reversible exhausted phenotype to a terminally exhausted one. Unfortunately, at the start of the SARS-CoV-2 pandemic, the laboratory at Salk Institute was closed for several months and, consequently, the experiments were delayed. Moreover, this interruption to experimental plans has been accompanied by a change in the experimental methods. Indeed, Prof. Kaech and Dr. Mann have modified the experimental aims to investigate changes in some protein kinase enzymes arising in a short window after the infection. These studies, that might point out the beginning of an exhaustion pathway, are ongoing at the moment of writing this thesis and, thus, cannot be considered for the mathematical modelling.

Moreover, as better detailed in Section 6.4, the defined deterministic models come with some limitations. Given experimental data, it is not always possible to identify the correct underlying model: this depends, among other things, on

the dimension of parameter space, on parameter dependencies with model variables and, importantly, on the measurability of model outputs. The developed deterministic models aim to analyse a still unknown and novel topic, hence, the need to rely on experiments carried out by collaborators and less on biological literature. However, the experimental data used for parameter inference were still preliminary, as experiments were still ongoing; this brought uncertainty in the reliability of the quantified model output and, consequently, in parameter inference. In addition, the experimental data that have been used consist of only one time point and only two exhausted cell populations. The consequence of limited experimental data is quite clear when looking at the modelling results. We are aware that, given the current outputs of the modelling analysis, estimates of parameters representing cell rates could not be considered reliable; and, equally, could not be applied to deduce and forecast future cell behaviour. However, upon availability of more extensive experimental data, the same mathematical techniques, learnt during the doctoral studies and applied in this Chapter, could be easily adapted to different model structures. In this context, upon model validation with reliable experimental data, the same mathematical methods could be indeed used to estimate cell rates and make informative predictions about cell dynamics. In particular, conscious of current model limitations, we could consider, as part of future work, the refinement of exhaustion models with the advanced experimental results, *i.e.,* measurement of more time points, definition of markers for other populations of cells, analysis of transcription factors, RNA-seq data, or estimation of proliferation rates.

Deterministic models developed in this Chapter rely on experimental data, and a different method could be considered to model the process of $CD8^+$ T cell exhaustion from another mathematical perspective. Stochastic models might become significant in describing the evolution of cell dynamics, when random fluctuations in population sizes cannot be ignored. This can be the case when considering only a small number of cells modelled in the system, and we are not in the limit of large population size. Many questions regarding differentiation fate and timelines are difficult to determine experimentally; however, it might be possible to analyse some of their properties and the corresponding summary statistics of a stochastic mathematical model.

6. A DETERMINISTIC APPROACH TO CD8$^+$ T CELL EXHAUSTION DYNAMICS

With the aim of mathematically modelling the exhaustion-differentiation process, the same theoretical methods, outlined in Chapters 3, 4 and 5 for a compartmental model, could be applied to stochastic models representing näive, effector and exhausted cells. Indeed, as pointed out in Chapter 1, the exhaustion process can be considered a special example of differentiation, to a dysfunctional, and not functional-specific, state.

One can consider the techniques developed in Chapter 3 for a time-independent compartmental model and apply a probability generating function approach to characterise the probability distribution of a stochastic process for T cell exhaustion. For example, one can assume to have three exhausted populations representing precursor, progenitor and terminally exhausted cells; and set up a model of a sequence of three compartments, C_1, C_2 and C_3, respectively. Cellular events in each compartment can be associated with probabilities and one can define a random variable representing the number of cells reaching the terminally exhausted state, descended from one precursor cell of compartment C_1. We can proceed as in Chapter 3: define the corresponding probability generating function and find a recursive relationship to describe the probability density of the corresponding random variable, $\mathbf{R}$. Similarly, a time-dependent compartmental model as outlined in Chapter 5 could be defined for the process of T cell exhaustion.

In both Chapter 3 and 5, we mention that asymmetric division (cell division event characterised by one daughter cell remaining in the mother's compartment, while the other moving to the next compartment) is a subject of recent research. We show how it can play a role in thymus development (Pham *et al.*, 2015) and hematopoiesis (see Sections 3.5, 5.2.3 and 5.2.2, respectively) and it might be relevant for exhaustion development, as well. Indeed, current assumptions on memory precursor-cell formation include bifurcation models, where effector and memory fates segregate early following asymmetric division; this event is responsible to unequally distribute Tbet and Eomes transcription factors (Dolina *et al.*, 2021; Pipkin *et al.*, 2010; Rao *et al.*, 2010; Zehn *et al.*, 2022). Theoretical work outlined in Sections 3.2.7 and 3.3.3 or Section 5.1.2 could be applied to a compartmental model for the exhaustion process. Such a model relies on compartment definition according to the characterised and measured populations. It will be part of future work and, hence, beyond the scope of this thesis.

As outlined in Section 1.2, the mechanisms underlying the exhaustion process have not been clearly defined yet. The landscape is quite complex as, on one side, the persistency of antigen drives cells to an exhausted state (lacking memory potential) and, on the other side, similarities in expressed molecules between exhausted and memory cells have been observed (Angelosanto *et al.*, 2012; Wherry & Kurachi, 2015). Thus, a mathematical study of the interplay between näive, effector, memory T cells and the exhausted state could be important in shedding light onto the question if exhausted cells necessarily arise from terminally differentiated effector cells or may also arise from the memory pool. As in Chapter 5, we could define a stochastic model at the population or single-cell level as a multi-variate Markov process for the $CD8^+$ T cell exhaustion-differentiation process. Since the analytical treatment of such models is typically complex, we could follow some analytical methods to study summary statistics of interest, as an alternative to the computational approach of running Gillespie simulations, as well as to the classic analytical study of the Kolmogorov equations. For more mathematical details about probability theory and stochastic processes, we refer the reader to Sections 2.1 and 2.2. Moreover, a sensitivity analysis can be carried out to quantify how the considered stochastic descriptors are influenced by the model's parameters.

When analysing statistics at the population level, one needs to take into account the total number of cells over time, as well as the possible interactions between cells states. Using a single-cell approach, we lose information about population sizes but we lighten the calculation and we could compute different summary statistics to outline basic model properties that are then reflected at the population level dynamics. This approach has been applied in mathematical biology before by de la Higuera *et al.* (2019); Gómez-Corral *et al.* (2020); López-García *et al.* (2018). The model depicted in Figure 6.9, for instance, represents the näive N, effector E and memory M cell state of a $CD8^+$ T cell as well as the exhausted state X. We propose, as part of future work, to track the dynamics of a single cell, starting in state i, until the cell eventually reaches the exhausted state or dies. In this framework, one can define the continuous-time Markov chain $\Upsilon = \{\mathcal{Y}(t) : t \geq 0\}$ with $\mathcal{Y}(t) \geq 0$ for all $t \geq 0$, over the state space $S_{\mathcal{Y}} = \{N, E, M, X\} \cup \{\emptyset\}$, where the state $\{\emptyset\}$ represents the death of the cell.

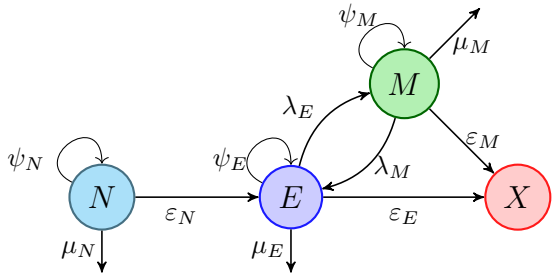


Figure 6.9: State diagram for the continuous-time Markov chain $\Upsilon = \{\mathcal{Y}(t) : t \geq 0\}$ with $\mathcal{Y}(t) \geq 0$ for all $t \geq 0$, over the state space $S_{\mathcal{Y}} = \{N, E, M, X\} \cup \{\emptyset\}$, representing näive, effector, memory and exhausted CD8$^+$ T cells, respectively. The state $\{\emptyset\}$ represents the death of the cell. Näive, effector and memory cells can die, divide or differentiate. Effector and memory can differentiate to the exhausted state X.

Summary statistics of the defined stochastic process could be then analysed following first-step arguments. For example, we could define:

- the probability of the cell, starting in state i for $i = \{N, E, M, X\}$, to differentiate into the exhausted state before dying;
- the "functional"-span T_i of the cell starting in state i, T_i, which, since exhaustion is a dysfunctional state, quantifies the functional potential of cells, before reaching the exhaustion state, in the system depending on their initial location. Note that $T_X = 0$; moreover, since the tracked cell can die, given

  $$T_i = \inf\{t \geq 0 : \mathcal{Y}(t) = X \,|\, \mathcal{Y}(0) = i\}, \quad \text{for} \quad i \in \{N, E, M, X\}$$

  we could be actually interested in the restricted time $T_i \, \mathbb{1}_{\{T_i < +\infty\}}$ where $\mathbb{1}_{\{T_i < +\infty\}}$ is the indicator function defined as

  $$\mathbb{1}_{\{T_i < +\infty\}} = \begin{cases} 1, & \text{if} \quad T_i < +\infty, \\ 0, & \text{if} \quad T_i = +\infty, \end{cases}$$

  and
- the number of divisions carried out by the cell during its "functional"-span, which quantifies the proliferation potential of cells in the system.

Overall, the results of a modelling strategy for the process of $CD8^+$ T cell exhaustion can be enhanced by further mathematical analysis, as well as the development of novel experimental techniques. The cell division rate, the transcriptomic and epigenetic profile, and a characterisation of the different population of cells arising from the differentiation process are examples of still open questions in biology. Recent works, such as the one of Bresser *et al.* (2022), might innovate the biological and mathematical research in the field of exhaustion: the genetic-tracing method they developed and applied during an acute infection could be used to determine the replicative history of terminal effector and exhausted cells.

Appendix A

Recurrence relation for the probabilities $Q_k(M)$

This Appendix is meant to extend the work outlined in Section 3.2.3 and provide more details about the observations made in Sections 3.2.6 and 3.2.7; here, the recurrence relations for $Q_k(2)$ for both the cases of $p_{\mathrm{a}}(i) = 0$ and $p_{\mathrm{a}}(i) > 0$, for $i = 1, 2$, are derived. Moreover, we hint how to generalise the method for $M > 2$.

A.1 Recurrence relation in the case $p_{\mathrm{a}}(i) = 0$

We set $M = 2$ and consider the general case of two non-identical compartments, that is $\phi_1(z) \neq \phi_2(z)$. The probability generating function of the process is $\Phi_2(z) = \phi_1(\phi_2(z)) = \sum_{k=0}^{+\infty} Q_k(2) z^k$. More explicitly,

$$\Phi_2(z) = \frac{1 - \sqrt{\Delta_1^2 - 2\frac{p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)} + 2\frac{p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}\sqrt{\Delta_2^2 - 4p_b(2)p_e(2)z}}}{2p_b(1)},$$

that is a rational function with a nested root of the type,

$$p_3(z) + \sqrt[r]{p_1(z) + \sqrt{p_2(z)}}, \tag{A.1}$$

where $r = 2$, $p_1(z) = \Delta_1^2 - 2\frac{p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}$ constant and $p_2(z) = \Delta_2^2 - 4p_b(2)p_e(2)z$ with a constant coefficient upfront. The third addend $p_3(z) = \frac{1}{2p_{\mathrm{b}}(1)}$ is constant and does not contribute to the recurrence since it does not influence the z^k-dependent

terms of the power series of $\Phi_2(z)$. Given that $p_1(z)$ and $p_2(z)$ are polynomials, we can follow the method of Mathar (2021) and derive a P-finite recurrence for $Q_k(2)$. For simplicity of notation, we denote, similar to Sections 3.2.3 and 3.2.7, $w_i^2(z) = 1 - 4p_b(i)p_d(i) - 4p_b(i)p_e(i)z$ for $i = 1, 2$. Then, $\Phi_2(z)$ and its derivatives read

$$\begin{aligned}
&2p_b(1)\Phi_2(z) = 1 - w_1, \\
&\Phi_2'(z) = \frac{p_e(1)}{w_1}\phi_2'(z) = \frac{p_e(1)p_e(2)}{w_1 w_2}, \\
&\Phi_2''(z) = \frac{2p_e(1)p_e^2(2)}{w_1^3 w_2^3}\left(p_b(2)w_1^2 + p_e(1)p_b(1)w_2\right),
\end{aligned}$$

where w_1 and w_2 are shorthand for $w_1(\phi_2(z))$ and $w_2(z)$. Since the probability generating function $\Phi_2(z)$ is D-finite (see Definition (A.1) and Stanley (1980) for more details), we can follow Mathar (2021) and assume there exist polynomials $P_2(z), P_1(z)$ and $P_0(z)$ such that

$$P_2(z)\Phi_2''(z) + P_1(z)\Phi_2'(z) + P_0(z)(1 - 2p_b\Phi_2(z)) = 0\,. \tag{A.2}$$

We substitute the expression of $\Phi_2(z)$ and its derivatives and multiply by $w_1^3 w_2^3$,

$$\begin{aligned}
2p_e(1)p_e^2(2)\left(p_b(2)w_1^2 + p_b(1)p_e(1)w_2\right) P_2(z) + \\
+ p_e(1)p_e(2)w_2^2 w_1^2\, P_1(z) + w_2^3 w_1^4\, P_0(z) = 0.
\end{aligned} \tag{A.3}$$

We substitute $w_1^2 = \Delta_1^2 - 2p_e(1)r_b + 2p_e(1)r_b w_2$ where we denoted $r_b = \frac{p_b(1)}{p_b(2)}$ and note that all terms of Equation (A.3) are proportional to w_2^j with $j = 1, \ldots, 5$. We note that if j is odd, then $w_2^j \propto (\Delta_2 - 4p_b(2)p_e(2))^{1/2}$, otherwise, if j is even, $w_2^j \propto (\Delta_2 - 4p_b(2)p_e(2))$. Instead of solving Equation (A.3), we follow Mathar (2021) and ask for the components which depend on $w_2(z)$ and $w_2^2(z)$ (that are j odd or even, respectively) to be individually zero. This results in a system of two equations

$$\begin{cases} \alpha_1 P_2(z) + \alpha_2 P_1(z) + \alpha_3 P_0(z) = 0 \\ \alpha_4 P_2(z) + \alpha_5 P_1(z) + \alpha_6 P_0(z) = 0 \end{cases},$$

where we denote with α_i, for $i = 1, 2, 3$ the coefficients of A.3 proportional to $w_2(z)$, and for $i = 4, 5, 6$ the terms proportional to $w_2^2(z)$. One can think of $(P_2(z), P_1(z), P_0(z))$ as a three dimensional vector orthogonal to $(\alpha_1, \alpha_2, \alpha_3)$ and

$(\alpha_4, \alpha_5, \alpha_6)$. So that we can find $(P_2(z), P_1(z), P_0(z))$ as the cross product of the two α vectors,

$$\begin{aligned} P_2(z) &= \alpha_2\alpha_6 - \alpha_3\alpha_5\,, \\ P_1(z) &= \alpha_3\alpha_4 - \alpha_1\alpha_6\,, \\ P_0(z) &= \alpha_1\alpha_5 - \alpha_2\alpha_4\,. \end{aligned}$$

The solution of this linear system gives $P_2(z), P_1(z)$ and $P_0(z)$ as a second order, first order and constant polynomials in terms of z, respectively. Namely, $P_2(z) = P_{2,0} + P_{2,1}z + P_{2,2}z^2$, $P_1(z) = P_{1,0} + P_{1,1}z$ and $P_0(z) = P_{0,0}$ where

$$\begin{aligned} P_{0,0} &= -2p_{\rm b}(1)p_{\rm e}^2(1)p_{\rm e}^2(2)\,, \\ P_{1,1} &= 32p_{\rm b}(1)p_{\rm e}^2(1)p_{\rm e}^2(2)\,, \\ P_{1,0} &= \frac{p_{\rm e}(2)}{r_b}\left(\Delta_1^2 - 2p_{\rm e}(1)r_b\right)^2 - 8p_{\rm e}(2)\left(p_{\rm e}^2(1)r_b - 4\frac{p_{\rm b}^2(2)p_{\rm d}(2)}{r_b}\right)\,, \\ P_{2,2} &= P_{1,1}\,, \\ P_{2,1} &= 2\frac{p_{\rm e}(2)}{r_b}\left(\Delta_1^2 - 2p_{\rm e}(1)r_b\right)^2 - 64p_{\rm b}(1)p_{\rm e}^2(1)p_{\rm e}^2(2)\,, \\ P_{2,0} &= -\frac{\Delta_2^2}{2p_{\rm b}(1)}\Bigg(1 - 4p_{\rm e}(1)r_b \\ &\qquad +8p_{\rm b}(1)p_{\rm d}(1)\left(-1 + 2p_{\rm b}(1)p_{\rm d}(1) + 2p_{\rm e}(1)r_b + 2\frac{p_{\rm e}^2(1)p_{\rm d}(2)r_b}{p_{\rm d}(1)}\right)\Bigg)\,. \end{aligned}$$

Recall that $\Phi_2(z)$ is a probability generating function and thus, by definition,

$$\Phi_2(z) = \sum_{k=0}^{+\infty} Q_k(2)z^k, \qquad \Phi_2'(z) = \sum_{k=0}^{+\infty}(k+1)Q_{k+1}(2)z^k,$$
$$\text{and} \quad \Phi_2''(z) = \sum_{k=0}^{+\infty}(k+2)(k+1)Q_{k+2}(2)z^k\,.$$

We substitute in (A.2) and we match terms proportional to z^k. We thus derive a three-terms recurrence relation for the two compartments case

$$\begin{aligned} P_{2,0}\,(k+2)(k+1)\,Q_{k+2}(2) + \left[P_{2,1}k^2 + (P_{1,0}+P_{2,1})k + P_{1,0}\right]Q_{k+1}(2) \\ +\left[P_{2,2}k^2 + (P_{1,1}-P_{2,2})k + P_{0,0}\right]Q_k(2) = 0\,. \end{aligned} \tag{A.4}$$

Substituting the coefficients $P_{j,i}$ for $j = 0, 1, 2$ and $i = 0, \ldots, j$, we get

$$\begin{aligned}
&\Delta_2^2 \left(\Delta_1^2(4p_{\mathrm{e}}(1)r_b - \Delta_1^2) - 16p_{\mathrm{e}}^2(1)p_{\mathrm{b}}(1)p_{\mathrm{d}}(2)r_b\right)(k+1)(k+2)Q_{k+2}(2) \\
&\quad + \left(\Delta_1^4 - 4p_{\mathrm{e}}(1)r_b\Delta_1^2 + 4p_{\mathrm{e}}^2(1)r_b^2(1 - 2\Delta_2^2)\right)p_{\mathrm{b}}(2)p_{\mathrm{e}}(2) \\
&\quad \times (2k+1)(2k+2)Q_{k+1}(2) + p_{\mathrm{e}}^2(1)p_{\mathrm{e}}^2(2)p_{\mathrm{b}}^2(1)\left(64k^2 - 4\right)Q_k(2) = 0.
\end{aligned}$$

If $p_{\mathrm{d}}(1) = p_{\mathrm{d}}(2) = 0$, then $\Delta_1 = \Delta_2 = 1$ and (A.4) takes the simpler form

$$\begin{aligned}
&(4p_{\mathrm{e}}(1)r_b - 1)(k+1)(k+2)Q_{k+2}(2) \\
&\quad - p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)(4p_{\mathrm{e}}^2(1)r_b^2 + 4p_{\mathrm{e}}(1)r_b - 1)(2k+1)(2k+2)Q_{k+1}(2) \\
&\quad + 4p_{\mathrm{b}}^2(2)p_{\mathrm{e}}^2(1)p_{\mathrm{e}}^2(2)r_b^2(16k^2 - 1)Q_k(2) = 0.
\end{aligned}$$

That is the two-compartments recurrence relation that gives $Q_k(2) = \mathbb{P}(\mathbf{R} = k)$ for the case of $M = 2$.

A.1.1 Recurrence for the general case of $M > 2$

For the general case of $M > 2$ compartments, we haven't found an explicit recurrence. However, we can hint that this method could still be implemented, even if with long and complicated calculations. We start recalling the definitions of *D-finite* and *algebraic* (Stanley, 1980).

Definition A.1. A formal power series $f(x) \in \mathbb{C}[[x]]$ is said to be *D-finite* (or differentiably finite) if $f(x)$ and all its derivatives $f^{(d)}(x) = \frac{d^d}{dx^d}f(x)$, for $d \geq 1$, span a finite-dimensional subspace of $\mathbb{C}[[x]]$. Thus, there exist finitely many polynomials $P_d(x), \ldots, P_0(x)$ non all zero, and a polynomial $S(x)$, such that

$$P_d(x)f^{(d)}(x) + P_{d-1}(x)f^{(d-1)}(x) + \cdots + P_1(x)f^{(1)}(x) + P_0(x)f(x) + S(x) = 0\,.$$

Definition A.2. A formal power series $f(x) \in \mathbb{C}[[x]]$ is said to be *algebraic* if $f(x)$ and its powers span a finite-dimensional subspace of $\mathbb{C}[[x]]$; that is, there exists polynomials $R_n(x), \ldots, R_0(x)$ non all zero such that

$$R_n(x)f^n(x) + R_{n-1}(x)f^{n-1}(x) + \cdots + R_1(x)f(x) + R_0(x) = 0\,.$$

Moreover, we recall the definitions of $\Phi_M(z)$ and $\chi_M(z)$ in Section 3.2.6 and observe that, since the probability generating function $\phi_1(z)$ is D-finite and algebraic, the composition $\phi_1(\chi_M(z))$ is D-finite (Stanley, 1980). Thus, by definition, there exist polynomials $P_i(z)$, for $i = 0, \ldots, d$, and $S(z)$ such that

$$P_d(z)\Phi_M^{(d)} + P_{d-1}(z)\Phi_M^{(d-1)} + \cdots + P_1(z)\Phi'_M(z) + P_0(z)\Phi_M(z) + S(z) = 0\,,$$

Note that when $M = 2$ this simplifies into Equation (A.2). In this case, $\phi(z)$ is algebraic and $\Phi_2(z) = \phi_1(\phi_1(z))$ is D-finite and $d = 2$. We reserve for future work, out of the scope of this PhD thesis, to apply the method described in Mathar (2021) and find a general recurrence relation for $Q_k(M)$, for $M > 2$. Open questions one needs to answer are the order of the differential equation of $\Phi_M(z)$, that is the value of d, as well as the degree of each polynomial $P_i(z)$ with respect to z.

A.2 Recurrence relation in the case of asymmetric division

We set $M = 2$ and consider the general case of two non-identical compartments, that is $\tilde{\phi}_1(z) \neq \tilde{\phi}_2(z)$. The probability generating function of the process is $\tilde{\Phi}_2(z) = \tilde{\phi}_1(\tilde{\phi}_2(z)) = \sum_{k=0}^{+\infty} \tilde{Q}_k(2)z^k$, that is,

$$\tilde{\Phi}_2(z) = \frac{1 - p_a(1)\tilde{\phi}_2(z) - \sqrt{(1 - p_a(1)\tilde{\phi}_2(z))^2 - 4p_b(1)p_d(1) - 4p_b(1)p_e(1)\tilde{\phi}_2(z)}}{2p_b(1)}\,.$$

Note that $\tilde{\Phi}_2(z)$ is a nested square root function of the type in (A.1) as for the self-renewal case of $p_{\mathrm{a}}(i) = 0$. However, the inclusion of the asymmetric division event brings complexity into the calculations for the recurrence relation, because $\tilde{p}_1(z)$ and $\tilde{p}_3(z)$ in (A.1) are not constant addends but functions of z. We can rearrange the terms of $\tilde{\Phi}_2(z)$ and write it as

$$2p_{\mathrm{b}}(1)\tilde{\Phi}_2(z) = H_1(z) - H_2(z)\,,$$

where

$$
\begin{aligned}
H_1(z) &= 1 - \frac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)} + \frac{p_{\mathrm{a}}(1)p_{\mathrm{a}}(2)}{2p_{\mathrm{b}}(2)}z + \frac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)}\tilde{w}_2(z)\,,\\
H_2^2(z) &= \frac{p_{\mathrm{a}}^2(1)p_{\mathrm{a}}^2(2)}{2p_{\mathrm{b}}^2(2)}z^2 + \left(\frac{p_{\mathrm{a}}(2)}{p_{\mathrm{b}}(2)}(p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1) - \frac{p_{\mathrm{a}}^2(1)}{p_{\mathrm{b}}^2(2)}(p_{\mathrm{a}}(2) + p_{\mathrm{b}}(2)p_{\mathrm{e}}(2))\right)z\\
&\quad + \left(\frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)} - \frac{p_{\mathrm{a}}^2(1)(1 - p_{\mathrm{a}}(2)z)}{2p_{\mathrm{b}}^2(2)}\right)\tilde{w}_2(z)\\
&\quad + \Delta_1^2 + \frac{p_{\mathrm{a}}^2(1)}{2p_{\mathrm{b}}^2(2)}(1 - 2p_{\mathrm{d}}(2)p_{\mathrm{b}}(2)) - \frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}\,,
\end{aligned}
$$

and we denoted

$$\tilde{w}_i^2(z) = 1 - 4p_{\mathrm{b}}(i)p_{\mathrm{d}}(i) - (2p_{\mathrm{a}}(i) + 4p_{\mathrm{b}}(i)p_{\mathrm{e}}(i))z + p_{\mathrm{a}}^2(i)z^2.$$

The different form of functions $H_1(z) = \sum_{k=0}^{+\infty} h_k z^k$ and $H_2(z) = \sum_{k=0}^{+\infty} g_k z^k$ results in different recurrence relations. In particular, the function $H_1(z)$ satisfies

$$2p_{\mathrm{b}}(2)H_1(z) = 2p_{\mathrm{b}}(2) - p_{\mathrm{a}}(1) + p_{\mathrm{a}}(1)p_{\mathrm{a}}(2)z + p_{\mathrm{a}}(1)\tilde{w}_2(z)\,.$$

We compute its first derivative,

$$2p_{\mathrm{b}}(2)H_1'(z) = p_{\mathrm{a}}(1)p_{\mathrm{a}}(2) + \frac{1}{2}p_{\mathrm{a}}(1)\frac{\tilde{w}_2'(z)}{\tilde{w}_2(z)}\,,$$

and assume, following Mathar (2021), there exist polynomials $P_1(z), P_0(z)$ and $S(z)$ such that:

$$P_1(z)H_1'(z) + P_0(z)H_1(z) + S(z) = 0\,.$$

After some algebra we find

$$
\begin{aligned}
P_1(z) &= 2\tilde{w}_2^2(z)\,,\\
P_0(z) &= -(\tilde{w}_2^2(z))'\,,\\
S(z) &= \frac{(2p_{\mathrm{b}}(2) - p_{\mathrm{a}}(1)(1 - p_{\mathrm{a}}(2))z)\,(\tilde{w}_2^2(z))' - 2p_{\mathrm{a}}(1)p_{\mathrm{a}}(2)\tilde{w}_2^2(z)}{2p_{\mathrm{b}}(2)}\,.
\end{aligned}
$$

Given that $\deg(P_1(z)) = 2$ and $\deg(P_0(z)) = 1$, we can write

$$2P_{1,0}(k+2)h_{k+2} + \big(2P_{1,1}(k+1) - P_{0,0}\big)h_{k+1} + (2P_{1,2}k - P_{0,1})h_k = 0\,,$$

where we denoted with $P_{j,i}$ the coefficients of order i of polynomial $P_j(z)$. Finally, we get the following three-term recurrence

$$\Delta_2^2(k+2)h_{k+2} + \big(p_a(2) + 2p_b(2)p_e(2)\big)(2k+1)h_{k+1} + p_a^2(2)(k-1)h_k = 0\,. \quad \text{(A.5)}$$

A recurrence relation for the function $H_2(z)$ is instead less straightforward to obtain. Following Mathar (2021), we assume there exist polynomials $P_2(z), P_1(z), P_0(z)$ and $S(z)$ such that

$$P_2(z)H_2''(z) + P_1(z)H_2'(z) + P_0(z)H_2(z) + S(z) = 0\,. \quad \text{(A.6)}$$

We substitute in (A.6) the first and second order derivatives of $H_2(z)$ and we ask for the components proportional to $\tilde{w}_2(z)$ and $\tilde{w}_2^2(z)$ (that are odd or even powers of $\tilde{w}_2(z)$, respectively) to be individually zero (Mathar, 2021). This results in a system of three equations which straightforwardly gives $S(z) = 0$. After some algebra, one can find polynomials $P_2(z), P_1(z), P_0(z)$ such that

$$P_2(z) = \sum_{i=0}^{6} P_{2,i}\, z^i, \quad P_1(z) = \sum_{i=0}^{4} P_{1,i}\, z^i, \quad \text{and} \quad P_0(z) = \sum_{i=0}^{3} P_{0,i}\, z^i$$

where, for $j = 0, 1, 2$, $P_{j,i}$ are the constant coefficients of the corresponding polynomial $P_j(z)$. Recall that $H_2(z)$ is a generating function and thus $H_2(z) = \sum_{k=0}^{+\infty} g_k z^k$,

$$H_2'(z) = \sum_{k=0}^{+\infty}(k+1)g_{k+1}z^k, \quad \text{and} \quad H_2''(z) = \sum_{k=0}^{+\infty}(k+2)(k+1)g_{k+2}z^k\,.$$

Equating terms proportional to z^k, we find the following recurrence relation

$$\begin{aligned}(k+6)(k+5)g_{k+6} + \left[P_{2,1}k^2 + (9P_{2,1} + P_{1,0})k + (20P_{2,1} + 5P_{1,0})\right] g_{k+5} + \\ + \left[P_{2,2}k^2 + (7P_{2,2} + P_{1,1})k + (12P_{2,2} + 4P_{1,1} + P_{0,0})\right] g_{k+4} + \\ + \left[P_{2,3}k^2 + (5P_{2,3} + P_{1,2})k + (6P_{2,3} + 3P_{1,2} + P_{0,1})\right] g_{k+3} + \\ + \left[P_{2,4}k^2 + (3P_{2,4} + P_{1,3})k + (2P_{2,4} + 2P_{1,3} + P_{0,2})\right] g_{k+2} + \\ + \left[P_{2,5}k^2 + (P_{2,5} + P_{1,4})k\right] g_{k+1} + k(k-1)P_{2,6}\, g_k = 0\,.\end{aligned}$$

This six-terms recurrence for $H_2(z)$, together with the recurrence for $H_1(z)$ of Equation (A.5), give the recurrence for $\tilde{Q}_k(2) = \mathbb{P}(\mathbf{R} = k)$ for the general case of two compartments with $p_a(i) > 0$, for $i = 1, 2$.

Appendix B

Asymptotic behaviour of distribution $Q_k(M)$

This Appendix is meant to extend the work outlined in Section 3.2.4 where the asymptotic behaviour of $Q_k(M)$ is analysed for $M = 1$, $p_a = 0$; here, we provide more details about the observations made in Sections 3.2.6 and 3.2.7, for the case $M > 1$, $p_a = 0$ and $p_a > 0$, respectively. We follow the method of singularity analysis developed by Flajolet & Sedgewick (2009). The idea is that by locating the isolated singularities of a function, the long-term behaviour of its coefficients of the series expansion can be found. The asymptotic analysis of $Q_k(M)$ can be directly based on the general discussion of composition of singularities of $\phi_1(z), \phi_2(z), \ldots, \phi_M(z)$. When considering the composition of two functions analytic at the origin and with an isolated singularity, one can list three different cases according to the values of their radii of convergence: *supercritical case* if the singularity is driven by the external function; *subcritical case* if the singularity is driven by the internal function; and the *critical case* in case of a confluence of singularities (Flajolet & Sedgewick (2009)).

Two compartments, $M = 2$. We consider the case of two compartments and, for simplicity of notation, we assume them to be identical, that is $\phi_1(z) = \phi_2(z)$, so that we can omit the compartment-index i. We aim to find the long-term

behaviour of the coefficients of the power series of the function

$$\Phi_2(z) = \phi(\phi(z)) = \frac{1 - \sqrt{\Delta^2 - 2p_e + 2p_e\sqrt{\Delta^2 - 4p_b p_e z}}}{2p_b}\,.$$

The inner function $\phi(z)$ has singularity in $\sigma = \frac{\Delta^2}{4p_b p_e} = \frac{1}{\gamma_1}$, as outlined in Equation (3.12) in Section 3.2.4. Whereas the outer function $\phi(\phi(z))$ has a singularity in $z = \sigma_2$ such that

$$\Delta^2 - 2p_e + 2p_e\sqrt{\Delta^2 - 4p_b p_e z} = 0\,,$$

which gives us

$$\sigma_2 = \frac{\Delta^2}{4p_b p_e} - \frac{1}{4p_b p_e}\left(\frac{2p_e - \Delta^2}{2p_e}\right)^2 \approx 1\,. \tag{B.1}$$

Since $\sigma_2 < \sigma$ we are in the *supercritical case*, that is the singularity of the function $\Phi_2(z)$ is driven by the external function. In particular, $\Phi_2(z)$ has a singularity in σ_2 as in Equation (B.1), but the inner function $\phi(z)$ is analytic in it. We can then Taylor expand the inner function $\phi(z)$ in σ_2 until the first order and substitute it in the expression of $\Phi_2(z)$,

$$\begin{aligned}
\Phi_2(z) &= \frac{1 - \sqrt{\Delta^2 - 4p_b p_e(\phi(\sigma_2) + \phi'(\sigma_2)(z - \sigma_2))}}{2p_b} \\
&= \frac{1}{2p_b} - \frac{\sqrt{\Delta^2 - 4p_b p_e(\phi(\sigma_2) + \sigma_2\phi'(\sigma_2))}}{2p_b}\sqrt{1 - \frac{4p_b p_e\phi'(\sigma_2)}{\Delta^2 - 4p_b p_e(\phi(\sigma_2) - \sigma_2\phi'(\sigma_2))}\,z} \\
&= \frac{1}{2p_b} - \frac{\sqrt{4p_b p_e\sigma_2\phi'(\sigma_2)}}{2p_b}\sqrt{1 - \frac{1}{\sigma_2}\,z}\,,
\end{aligned}$$

where, in the last step, we used that $\phi(\sigma_2) = \frac{\Delta^2}{4p_b p_e}$. We let $Q_k(2) = [z^k]\Phi_2(z)$ and apply Equation (3.11) to find the asymptotic coefficients of the generating function $\Phi_2(z)$,

$$\begin{aligned}
Q_k(2) &\sim \frac{\sqrt{4p_b p_e\sigma_2\phi'(\sigma_2)}}{4p_b\sqrt{\pi}}k^{-3/2}\sigma_2^{-k} \\
&= \sqrt{\frac{p_e^3\sigma_2}{8p_b\pi(2p_e - \Delta^2)}}\left(\frac{1}{\sigma_2}\right)^k k^{-3/2}\,,
\end{aligned} \tag{B.2}$$

given that $\phi'(z)\big|_{z=\sigma_2} = \dfrac{2p_e^2}{2p_e - \Delta^2}$. The behaviour of $Q_k(2)$ for large value of k, is depicted in red in Figure B.1. Here, we set $N = 25$ and consider equal compartments, that is $p_d(i) = 0.1/2$ for $i = 1, 2$.

Three compartments, $M = 3$ If we add a second intermediate compartment, so that $M = 3$, we need to analyse the singularities of a function with three nested square roots. Assuming, for simplicity of notation, the compartments to be identical, that is $\phi_1(z) = \phi_2(z) = \phi_3(z) = \phi(z)$,

$$\Phi_3(z) = \phi(\phi(\phi(z))) = \frac{1 - \sqrt{\Delta^2 - 4p_b p_e \phi(\phi(z))}}{2p_b}\,.$$

Let σ_3 satisfy $\Delta^2 - 4p_b p_e \phi(\phi(\sigma_3)) = 0$, we find

$$\sigma_3 = \frac{\Delta^2}{4p_b p_e} - \frac{1}{16 p_b p_e^3}\left[\left(\frac{2p_e - \Delta^2}{2p_e}\right)^2 - \Delta^2 + 2p_e\right]^2 \approx 1\,.$$

Since $\sigma_3 < \sigma_2 < \sigma = \frac{\Delta^2}{4p_b p_e}$ we fall in the *supercritical case*, that is the inner function $\phi(\phi(z))$ is analytic in σ_3. We denote $\hat{\phi}_2(z) = \phi(\phi(z))$ and $\hat{\phi}_2'(z)$ its derivative with respect to z. We substitute $\hat{\phi}_2(z)$ first order Taylor approximation in $\Phi_3(z)$,

$$\Phi_3(z) = \frac{1}{2p_b} - \frac{1}{2p_b}\sqrt{\left(\Delta^2 - 4p_b p_e \hat{\phi}_2(\sigma_3) + 4p_b p_e \sigma_3 \hat{\phi}_2'(\sigma_3)\right) - 4p_b p_e \hat{\phi}_2'(\sigma_3)\, z}\,.$$

Thus,

$$Q_k(3) \sim \frac{\sqrt{\Delta^2 - 4p_b p_e \hat{\phi}_2(\sigma_3) + 4p_b p_e \sigma_3 \hat{\phi}_2'(\sigma_3)}}{4p_b\sqrt{\pi}}\,\gamma_3^k\, k^{-3/2} \tag{B.3}$$

where

$$\gamma_3 = \frac{4p_{\mathrm{b}} p_{\mathrm{e}} \hat{\phi}_2'(\sigma_3)}{\Delta^2 - 4p_b p_e \hat{\phi}_2(\sigma_3) + 4p_b p_e \sigma_3 \hat{\phi}_2'(\sigma_3)}\,.$$

The behaviour of $Q_k(3)$ for large value of k, is depicted in blue in Figure B.1 where $N = 25$ and consider equal compartments, that is $p_{\mathrm{d}}(i) = 0.1/3$ for $i = 1, 2, 3$.

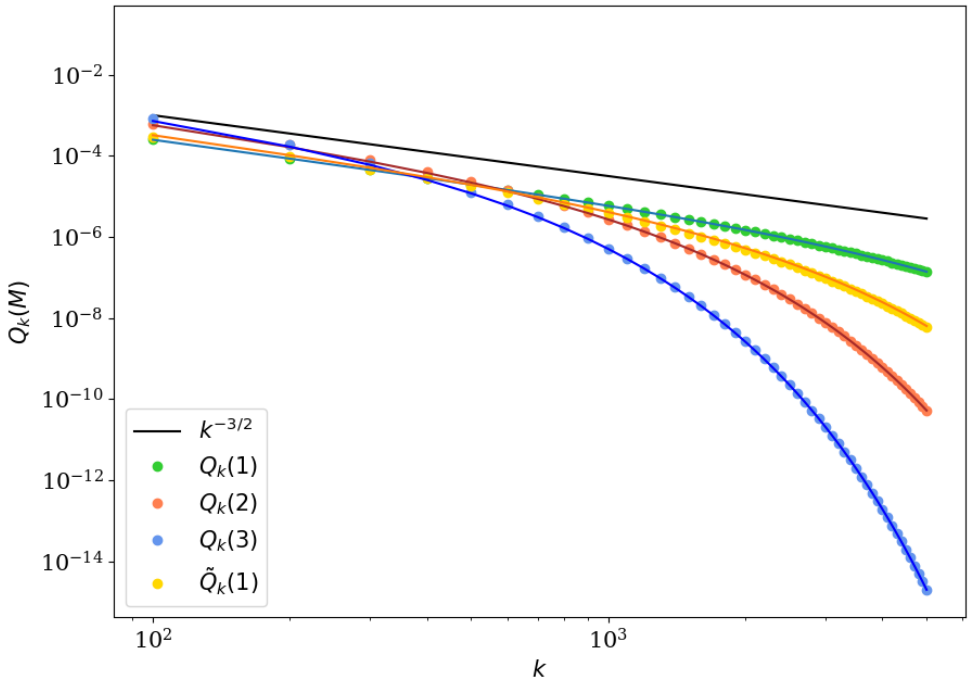


Figure B.1: Asymptotic behaviour of $Q_k(M)$ for $M = 1, 2, 3$ in green, red and blue, respectively. Yellow line shows $\tilde{Q}_k(1)$ distribution. We set $N = 25$ and consider equal compartments, $p_{\mathrm{d}}(i) = 0.1/M$ for $i = 1, \ldots, M$ and $p_{\mathrm{a}}(1) = 0.2$ for the asymmetric case ($p_{\mathrm{a}}(i) = 0$ otherwise). Straight lines are obtained using Equations (3.13), (B.2) and (B.3) for $M = 1, 2, 3$ respectively. Yellow line for the asymptotic case with $M = 1$ is given by Equation (3.31). Dots are obtained from compositions of generating function with Mathematica.

Appendix C

Solution of a Riccati system of two equations

In this appendix, we focus on the mathematical steps necessary to solve System (4.7), defined in Section 4.3 for a sequence of two compartments, and find the probability generating function $F_1(x_1, x_2, x_3, t)$; this System involves a differential equation of Riccati type and, thus, special considerations are required. Moreover, we consider a less complex case where the function F_2 is a simple exponential; for this, a complete analytical solution can be reported.

C.1 Solution of System (4.7)

As in Section 4.3.2, since we are interested in the time dependency of the probability generating functions, we will keep the notation short and denote $F_1(x_1, x_2, x_3, t)$, $F_2(x_2, x_3, t)$ with $F_1(t), F_2(t)$, respectively. As outlined in Section 4.3.2 for $M = 2$, solution of the differential equation for $F_2(x_2, x_3, t) = F_2(t)$ can be derived by applying the integration by parts method. The corresponding differential equation, indeed, is of the same form of the differential equation resulting from the case of direct differentiation $M = 1$. Solution is given by Equation (4.11) which reads

$$F_2(t) = \frac{d_1 + d_2\, \mathrm{e}^{-\gamma_2 t}}{d_3 + d_4 \mathrm{e}^{-\gamma_2 t}}\,.$$

with $d_1 = a_2\,(x_2 - b_2)$, $d_2 = b_2\,(a_2 - x_2)$, $d_3 = x_2 - b_2$ and $d_4 = a_2 - x_2$, where $a_2 < b_2$ are the two roots of $r(F_2) = \lambda_2 F_2^2 - S_2\, F_2 + \mu_2 + \nu_2 x_3$; that is

C. SOLUTION OF A RICCATI SYSTEM OF TWO EQUATIONS

$a_2, b_2 = \left(S_2 \pm (S_2^2 - 4\lambda_2(\mu_2 + \nu_2 x_3))^{1/2}\right)/2\,\lambda_2$.

We now consider $F_1(t)$. The differential equation for $F_1(t)$ is a Riccati equation, i.e. a first order differential equation, with non-constant coefficients. In particular, the quadratic and first order coefficients, λ_1, S_1 respectively, are constant; whereas the inhomogeneous term, $\mu_1 + \nu_1 F_2(t)$, is an exponential function of the independent variable t. Following the usual substitution method for Riccati equation, we transform the first order equation into a linear second order equation with non-constant coefficients for which more solving methods are available in the literature (Pala & Ertas, 2017; Sugai, 1960). In particular, we introduce a new function $f_1(t)$, such that $F_1(t) = -\frac{f_1'(t)}{\lambda_1 f_1(t)}$. Thus, $f_1(0) = 1$, $f_1'(0) = -\lambda_1 x_1$ and

$$f_1''(t) + S_1 f_1'(t) + \lambda_1[\mu_1 + \nu_1 F_2(t)]\, f_1(t) = 0\,. \tag{C.1}$$

where we denoted with $f_1'(t)$ and $f_1''(t)$ the first and second derivative of $f_1(t)$ with respect to the independent variable t. Equation (C.1) is a second order differential equation where the coefficient of the unknown function $f_1(t)$ has an exponential form. Differential equation of this type are usually solved via a change of variable and application of the Frobenius' method (Boyce & di Prima, 2009). In particular, we define $\tau = \mathrm{e}^{-\frac{\gamma_2}{2}t}$ so that

$$\begin{aligned}
\frac{df_1(t)}{dt} &= -\frac{\gamma_2}{2}\tau\frac{df_1(\tau)}{d\tau}\,,\\
\frac{d^2 f_1(t)}{d\,t^2} &= -\frac{\gamma_2^2\tau}{4}\left(\frac{df_1(\tau)}{d\tau} + \tau\frac{d^2 f_1(\tau)}{d\,\tau^2}\right)\,.
\end{aligned}$$

With an abuse of notation, we denote with $f_1'(\tau)$ and $f_1''(\tau)$ the first and second derivatives of $f_1(\tau)$ with respect to the independent variable τ. If $d_3 + d_4\tau^2 \neq 0$, that is $x_2 \neq \frac{b_2 - a_2\mathrm{e}^{-\gamma_2 t}}{1+\mathrm{e}^{-\gamma_2 t}}$, we get

$$\left(p_{2,2} + p_{2,4}\tau^2\right)\tau^2 f_1''(\tau) + \left(p_{1,1} + p_{1,3}\tau^2\right)\tau f_1'(\tau) + \left(p_{0,0} + p_{0,2}\tau^2\right) f_1(\tau) = 0\,. \tag{C.2}$$

Here, to keep the notation simple, we make use of the fact that the coefficient of the ith derivative of $f_1(\tau)$ is a polynomial with respect to the new independent variable τ. Thus, we denote with $p_{i,j}$ the coefficient of the term τ^j of the derivative

$f_1^{(i)}(\tau)$. In particular,

$$
\begin{aligned}
&p_{2,4} = d_4\frac{\gamma_2^2}{4}, \quad p_{2,2} = d_3\frac{\gamma_2^2}{4}, \quad p_{2,3} = p_{2,1} = p_{2,0} = 0;\\
&p_{1,3} = d_4\frac{\gamma_2}{2}\left(\frac{\gamma_2}{2} - S_1\right), \quad p_{1,1} = d_3\frac{\gamma_2}{2}\left(\frac{\gamma_2}{2} - S_1\right), \quad p_{1,2} = p_{1,0} = 0;\\
&p_{0,2} = \nu_1 d_2, \quad p_{0,0} = \nu_1 d_4, \quad p_{0,1} = 0\,.
\end{aligned}
$$

Following the Frobenius' method, we assume $f_1(\tau) = x^r \sum_{n=0}^{+\infty} a_n x^n$, for $r \in \mathbb{R}$. Substituting $f_1(\tau)$ and its derivatives into Equation (C.2) we get

$$
\begin{aligned}
&\sum_{n=0}^{+\infty} a_n(r+n)(r+n-1)\big(p_{2,2}x^{r+n} + p_{2,4}x^{r+n+2}\big) + \sum_{n=0}^{+\infty} a_n(r+n)\big(p_{1,1}x^{r+n} + p_{1,3}x^{r+n+2}\big)\\
&\quad + p_{0,0}\sum_{n=0}^{+\infty} a_n x^{r+n} + p_{0,2}\sum_{n=0}^{+\infty} a_n x^{r+n+2} = 0\,.
\end{aligned}
$$

Rearranging the terms, we find

$$
\begin{aligned}
&\big(r(r-1)p_{2,2} + rp_{1,1} + p_{0,0}\big)a_0 x^r + \big(r(r+1)p_{2,2} + (r+1)p_{1,1} + p_{0,0}\big)a_1 x^{r+1}\\
&\quad + \sum_{n=2}^{+\infty}\Big[\big((r+n-2)(r+n-3)p_{2,4} + (r+n-2)p_{1,3} + p_{0,2}\big)a_{n-2}\\
&\quad + \big((r+n)(r+n-1)p_{2,2} + (r+n)p_{1,1} + p_{0,0}\big)a_n\Big]x^{r+n} = 0\,. \qquad \text{(C.3)}
\end{aligned}
$$

When setting the first coefficient of the series to zero, one obtains the so called *indicial equation*,

$$
p_{2,2}r^2 + (p_{1,1} - p_{2,2})r + p_{0,0} = 0\,. \qquad \text{(C.4)}
$$

Equation (C.4) is a quadratic polynomial in r and, thus, has two real roots, r_1, r_2,

$$
r_{1,2} = \frac{S_1 \pm \sqrt{S_1 - 4\lambda_1(\mu_1 + \nu_1 d_1/d_3)}}{\gamma_2} = \frac{S_1}{\gamma_2} \pm \alpha\,,
$$

where $\alpha = (S_1 - 4\lambda_1(\mu_1 + \nu_1 d_1/d_3))^{1/2}/\gamma_2$. Since $r_1 - r_2$ is not an integer, we have two linearly independent Frobenius-type solutions (Boyce & di Prima, 2009),

$$
f_1(\tau) = c_1\,\tau^{r_1}\sum_{n=0}^{+\infty} a_n(r_1)\tau^n + c_2\,\tau^{r_2}\sum_{n=0}^{+\infty} a_n(r_2)\tau^n\,, \qquad \text{(C.5)}
$$

where $a_n(r_1), a_n(r_2)$ are given by the recurrence relation in Equation (C.3). In particular, we set $a_0(r) = 1$ and obtain

$$a_n(r) = -\frac{(r+n-2)(r+n-3)p_{2,4} + (r+n-2)p_{1,3} + p_{0,2}}{(r+n)(r+n-1)p_{2,2} + (r+n)p_{1,1} + p_{0,0}}\, a_{n-2}(r)\,.$$

Thus, we have $a_1(r) = a_3(r) = \cdots = a_n(r) = 0$ for all $n = 2m+1$ with $m \in \mathbb{N}^+$, and otherwise,

$$a_{2m}(r) = -\frac{(r+2m-2)(r+2m-3)p_{2,4} + (r+2m-2)p_{1,3} + p_{0,2}}{(r+2m)(r+2m-1)p_{2,2} + (r+2m)p_{1,1} + p_{0,0}}\, a_{2m-2}(r)\,.$$

Substituting r_1 and r_2 we get

$$a_{2m}(r_1) = -\frac{a_{2m-2}(r_1)}{2p_{2,2}m(m-\alpha)}\Bigg(p_{2,4}m^2 - 4p_{2,4}\alpha m - 4p_{2,4}(2m-\alpha-1) - d_4\lambda_1\nu_1\left(\frac{d_1}{d_3} - \frac{d_2}{d_4}\right)\Bigg)$$

and

$$a_{2m}(r_2) = -\frac{a_{2m-2}(r_2)}{2p_{2,2}m(m-\alpha) - d_3\gamma S_1\alpha}\Bigg(p_{2,4}m^2 + 4p_{2,4}\alpha m - 4p_{2,4}(2m+\alpha-1) - d_4\lambda_1\nu_1\left(\frac{d_1}{d_3} - \frac{d_2}{d_4}\right)\Bigg)\,.$$

We can now substitute $a_n(r_1)$ and $a_n(r_2)$ in Equation (C.5) and find the general solution $f_1(\tau)$ as the sum of two independent solutions; in particular, $f_1(\tau)$ is the sum of two linearly independent Gaussian hypergeometric function ${}_2F_1(a,b,c;\tau)$ that we computed with Mathematica. By substituting back $\tau = \mathrm{e}^{-\frac{\gamma_2}{2}t}$ and computing the first derivative of $f_1(t)$ with respect to time t, we find $F_1(t) = -\frac{f_1'(t)}{\lambda_1 f_1(t)}$ where constants c_1, c_2 of (C.5) are determined from the initial conditions $f_1(0) = 1$ and $f_1'(0) = -\lambda_1 x_1$. The resulting close formula for $F_1(t)$ is a linear combination of Gaussian hypergeometric functions with different parameters. However, due to its long expression we do not report it in this PhD thesis. Instead, we find more interesting to show the reader the calculations required to obtain the probability generating function $F_1(t)$ for a simplified case.

C.2 A simplified case

In this Section, instead of the actual solution of $F_2(t)$ reported in Equation (4.11), we consider $\tilde{F}_2(t) = \mathrm{e}^{-\gamma t}$ as a simple exponential of parameter γ. Moreover, we set ourselves into the case of $\mu_1 = 0$ so that no death event is occurring in the first compartment, and we assume $S_1 = \lambda_1 + \nu_1 = 1$. The equivalent of Equation (4.13) is now

$$f_1''(t) + S_1\, f_1'(t) + \lambda_1 \nu_1 \tilde{F}_2(t)\, f_1(t) = 0\,.$$

Since the coefficient of $f_1(t)$ is time-dependent, the application of the Frobenius's method is still required. As before, we let $\tau = \mathrm{e}^{-\frac{\gamma}{2}t}$ and obtain a differential equation with respect to τ,

$$p_{2,2}\tau^2 f_1''(\tau) + p_{1,1}\tau f_1'(\tau) + p_{0,2}\tau^2 f_1(\tau) = 0\,,$$

where $p_{2,2} = \frac{\gamma^2}{4}$, $p_{1,1} = \frac{\gamma}{2}\left(\frac{\gamma}{2} - 1\right)$, $p_{0,2} = \lambda_1\nu_1$. Note that, in this simplified assumptions, the coefficients of the function $f_1(\tau)$ as well as of its first and second order derivative are of degree two, one and two, respectively. Now, Equation (C.3) becomes

$$\begin{aligned}&\big(p_{2,2}r(r-1) + p_{1,1}r\big)a_0 + \big(p_{2,2}r(r+1) + p_{1,1}(r+1)\big)a_1 \\ &\quad + \sum_{n=2}^{+\infty}\Big[\big(p_{2,2}(r+n)(r+n-1) + p_{1,1}(r+n)\big)a_n + p_{0,2}a_{n-2}\Big]x^{r+n} = 0\,.\end{aligned} \tag{C.6}$$

The roots of the indicial equation are now

$$r_1 = 0, \quad r_2 = \frac{2}{\gamma}\,,$$

which are distinct and such that $r_1 - r_2$ is not an integer. Thus, we still have two linearly independent Frobenius-type solutions such that the general solution is given by

$$f_1(\tau) = c_1 \sum_{n=0}^{+\infty} a_n(0)\tau^n + \tau^{2/\gamma}\sum_{n=0}^{+\infty} a_n(2/\gamma)\tau^n\,.$$

From (C.6), we can write the following recurrence relation for the coefficients $a_n(r)$,

$$a_n(r) = \frac{-\lambda_1\nu_1}{\frac{\gamma}{2}(r+n)\left(\frac{\gamma}{2}(r+n) - 1\right)}a_{n-2}(r), \quad a_0(r) = 1\,.$$

Thus, $a_1 = a_3 = \cdots = a_n = 0$ for all odd indices $n = 2m+1$; otherwise, substituting the values of r_1, r_2, we get

$$a_{2m}(0) = \Gamma\left(1-\frac{1}{\gamma}\right)\frac{(-1)^m(\lambda_1\nu_1)^m}{\gamma^{2m}m!\left(m-\frac{1}{\gamma}\right)!} \quad \text{and} \quad a_{2m}(2/\gamma) = \Gamma\left(1+\frac{1}{\gamma}\right)\frac{(-1)^m(\lambda_1\nu_1)^m}{\gamma^{2m}m!\left(m+\frac{1}{\gamma}\right)!}\,.$$

Putting all together and substituting back $\tau = \mathrm{e}^{-\gamma/2\,t}$, we get

$$\begin{aligned} f_1(t) = {} & c_1\,\Gamma\left(1-\frac{1}{\gamma}\right)\gamma^{-\frac{1}{\gamma}}(\lambda_1\nu_1)^{\frac{1}{2\gamma}}\mathrm{e}^{-t/2}J\left(-\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right) \\ & + c_2\,\Gamma\left(1+\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}}(\lambda_1\nu_1)^{-\frac{1}{2\gamma}}\mathrm{e}^{-t/2}J\left(\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right)\,. \end{aligned}$$

Here, $J(k,x)$ denotes the Bessel function of first kind of order k,

$$J(k,x) = \sum_{n=0}^{+\infty}\frac{(-1)^n}{n!\,\Gamma(n+k+1)}\left(\frac{x}{2}\right)^{2n+k}\,.$$

Moreover, we recall that

$$\frac{d}{dx}J(k,x) = J(k-1,x) - \frac{k}{x}J(k,x)\,.$$

Thus, the first derivative of $f_1(t)$ results

$$\begin{aligned} f_1'(t) = {} & -\frac{1}{2}f_1(t) + c_1\,\Gamma\left(1-\frac{1}{\gamma}\right)\gamma^{-\frac{1}{\gamma}}(\lambda_1\nu_1)^{\frac{1}{2\gamma}}\,\mathrm{e}^{-t/2}\left[J\left(-\frac{1}{\gamma}-1,\frac{2\sqrt{\lambda_1\nu_1}\,\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right)\right. \\ & \left. + \frac{\mathrm{e}^{\gamma/2\,t}}{2\sqrt{\lambda_1\nu_1}}J\left(-\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}\,\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right)\right] + c_2\,\Gamma\left(1+\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}}(\lambda_1\nu_1)^{-\frac{1}{2\gamma}}\,\mathrm{e}^{-t/2} \\ & \times\left[J\left(\frac{1}{\gamma}-1,\frac{2\sqrt{\lambda_1\nu_1}\,\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right) - \frac{\mathrm{e}^{\gamma/2\,t}}{2\sqrt{\lambda_1\nu_1}}J\left(\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}\,\mathrm{e}^{-\gamma/2\,t}}{\gamma}\right)\right]. \end{aligned}$$

By imposing initial conditions $f_1(0) = 1$ and $f_1'(0) = -\lambda_1 x_1$, we find constants c_1, c_2 as

$$c_1 = \frac{1}{\beta_1}(1-\beta_2 c_2) \quad \text{and} \quad c_2 = -\frac{\lambda_1 x_1\beta_1 + \beta_3}{\beta_1(\beta_2\beta_3+\beta_4)},$$

where we denote

$$\begin{aligned}
\beta_1 &= \Gamma\left(1-\frac{1}{\gamma}\right)\gamma^{-\frac{1}{\gamma}}(\lambda_1\nu_1)^{\frac{1}{2\gamma}}J\left(-\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right),\\
\beta_2 &= \Gamma\left(1+\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}}(\lambda_1\nu_1)^{-\frac{1}{2\gamma}}J\left(\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right),\\
\beta_3 &= \Gamma\left(1-\frac{1}{\gamma}\right)\gamma^{-\frac{1}{\gamma}}(\lambda_1\nu_1)^{\frac{1}{2\gamma}}\left[-\frac{1}{2}+\frac{1}{2\sqrt{\lambda_1\nu_1}}J\left(-\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right)\right.\\
&\qquad \left.+J\left(-\frac{1}{\gamma}-1,\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right)\right],\\
\beta_4 &= \Gamma\left(1+\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}}(\lambda_1\nu_1)^{-\frac{1}{2\gamma}}\left[-\frac{1}{2}-\frac{1}{2\sqrt{\lambda_1\nu_1}}J\left(\frac{1}{\gamma},\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right)\right.\\
&\qquad \left.+J\left(\frac{1}{\gamma}-1,\frac{2\sqrt{\lambda_1\nu_1}}{\gamma}\right)\right].
\end{aligned}$$

All the terms of the solution $F_1(t) = -f_1'(t)/(\lambda_1 f_1(t))$ are, hence, explicitly determined.

Appendix D

Identifiability analysis for a four state mathematical model for the CD8$^+$ T cell exhaustion process

In this appendix, the structural identifiability analysis summarised in Section 6.3.2 for the toy model depicted in Figure 6.5, is extended. The analysis is based on the method developed by Castro & de Boer (2020) and described in Section 2.4. As detailed in Section 6.1, the provided experimental data consists of the percentages, over the entire population of transduced live, of TCF1$^-$ TIM3$^+$ and TCF1$^+$ TIM3$^-$ cells, representing reversible and terminal exhausted cells, respectively. Given the model structure reported in Figure 6.5, these correspond to X_1 and X_2 populations; hence, the two exhausted populations X_1 and X_2 can be considered to be observable.

In this Appendix, to keep the notation simpler, we omit the time-dependency in system variables $E_0(t)$, $E_1(t)$, $X_1(t)$ and $X_2(t)$ and we denote the time-dependent transitions as $\varepsilon_i\, f(t) = \tilde{\varepsilon}_i$, $\beta_i\, f(t) = \tilde{\beta}_i$, $\psi_i\, f(t) = \tilde{\psi}_i$ for $i = 0, 1$ as well as $\gamma_1\, f(t) = \tilde{\gamma}_1$.

We define thirteen scaling factors, namely $v_1, \ldots, v_{13}$, one for each parameter of the model, α_0, ψ_0, γ_0, μ_E, ε_0, β_0, ε_1, β_1, α_1, ψ_1, γ_1, μ_1, γ_1, μ_2. Also, we let u_1, u_2 be two additional scaling factors for the latent state variables E_0, E_1, respectively. The functionally independent functions of model defined in (6.2)

are

$$f_{1,1} = (\alpha_0 - \gamma_0 - \mu_E + \tilde{\phi}_0 - \tilde{\varepsilon}_0)E_0\,, \tag{D.1}$$
$$f_{1,2} = (\beta_0 - \tilde{\beta}_0)X_1\,, \tag{D.2}$$
$$f_{2,1} = \gamma_0 E_0\,, \tag{D.3}$$
$$f_{2,2} = -(\tilde{\varepsilon}_1 + \mu_E)E_1\,, \tag{D.4}$$
$$f_{2,3} = (\beta_1 - \tilde{\beta}_1)X_1\,, \tag{D.5}$$
$$f_{3,1} = \tilde{\varepsilon}_0 E_0\,, \tag{D.6}$$
$$f_{3,2} = \tilde{\varepsilon}_1 E_1\,, \tag{D.7}$$
$$f_{3,3} = (\alpha_1 - \tilde{\gamma}_1 - \mu_1 + \tilde{\phi}_1 - \beta_0 - \beta_1 + \tilde{\beta}_0 + \tilde{\beta}_1)X_1\,, \tag{D.8}$$
$$f_{4,1} = \tilde{\gamma}_1 X_1\,, \tag{D.9}$$
$$f_{4,2} = -\mu_2 X_2\,. \tag{D.10}$$

One need to equate each term to its scaled form, solve the corresponding system of identifiability equations, and finally derive the scaling constants u_i, v_j for $i = 1, \ldots, 13$ and $j = 1, 2$. For example, from Equation (D.2),

$$(\beta_0 - \tilde{\beta}_0)v_6 X_1 = (\beta_0 - \tilde{\beta}_0)X_1 \quad \Rightarrow \quad v_6 = 1\,,$$

since the other terms cancel out. Similarly from Equation (D.5) we get $v_8 = 1$. Moreover, given that X_1 and X_2 are assumed to be observable, terms in Equations (D.9) and (D.10) lead to $v_{11} = v_{13} = 1$. Whereas, since effector populations E_0 and E_1 are not measured at the end of the optogenetic experiment, Equation (D.3) results into

$$\gamma_0 v_3 E_0 u_1 = \gamma_0 E_0 \quad \Rightarrow \quad v_3 u_1 = 1\,.$$

Similarly, from Equations (D.6) and (D.7) one obtains $v_5 u_1 = 1$ and $v_7 u_2 = 1$, respectively. Since parameters β_0, β_1 and $\tilde{\gamma}_1$ are found to be measurable, Equation (D.8) simplifies into

$$\alpha_1 v_9 + \tilde{\phi}_1 v_{10} - \mu_1 v_{12} = \alpha_1 + \tilde{\phi}_1 - \mu_1\,.$$

The latter leads to an infinite number of combinations of the scaling factors satisfing the invariance condition, as all lying in the such defined manifold. Also,

given the identifiable pairs derived from Equations (D.3) and (D.6), once can simplify Equation (D.1) and obtain the manifold

$$(\alpha_0 v_1 + \tilde{\phi}_0 v_2 - \mu_E v_4) u_1 = \alpha_0 + \tilde{\phi}_0 - \mu_E \,. \tag{D.11}$$

Finally, given the identifiable pair $u_2 v_7 = 1$ derived from Equation (D.7), Equation (D.4) leads to the identifiable pair $v_4 u_2 = 1$.

As a result of the analysis, factors v_1, v_2, v_3, v_4, v_5, v_7, v_9, v_{10}, v_{12} as well as u_1, u_2 cannot be solved individually; this means that the corresponding parameters cannot be determined univocally and are thus unidentifiable.

However, following Massonis *et al.* (2021), we notice, for example, that the symmetry between α_0, $\tilde{\phi}_0$ and μ_E, given by Equation D.11, is broken due to the time-dependency of parameter ϕ_0. Indeed, if $\phi_0 f(t)$ is varying with time and $\alpha_0 + \mu_E$ is constant, the latter sum cannot compensate the changes of $\tilde{\phi}_0$. Similar considerations can be done for the manifold defined by α_1, $\tilde{\phi}_1$ and μ_1, whose symmetry is broken given that $\tilde{\phi}_1$ varies with time. What is left is the symmetry between parameters α_0, μ_E and the one between parameters α_1, μ_1; thus, if one among α_0 and μ_E as well one among α_1 and μ_1 is measurable, then parameters α_0, $\tilde{\phi}_0$, μ_E, α_1, $\tilde{\phi}_1$, μ_1 are all identifiable.

Moreover, one can notice that, if cellular population E_0 and E_1 are both observable (that is effector populations E_0 and E_1 are measured by experimental methods), then $u_1 = u_2 = 1$ and the identifiable pairs are resolved,

$$v_3 = v_5 = \frac{1}{u_1}, \quad \text{and} \quad v_4 = v_7 = \frac{1}{u_2} \,.$$

In this case, parameters μ_E, γ_0, ε_0 and ε_1 are identifiable. If so, given the line defined by $\alpha_0 + \mu_E$ and $\tilde{\phi}_0$, we have that also α_0 and $\tilde{\phi}_0$ result identifiable and the symmetry between α_0 and μ_E is solved as well. Similar considerations applies for the line defined by $\alpha_1 + \mu_1$ and $\tilde{\phi}_1$ since we can consider parameters α_1 and $\tilde{\phi}_1$ as identifiable. Consequently, all the groups involving scaling variables have been decoupled and the model could be said to be structurally identifiable.

In conclusion, the toy model (6.2) with the provided experimental data is not totally identifiable as symmetries between parameters cannot be completely decoupled. However, under the assumption that experimental collaborators can

provide measurements of effector population of cells, E_0 and E_1, at the end of the 48 hours of the optogenetic experiment, the model depicted in Figure 6.5 is structurally identifiable. Alternatively, the same conclusion of a structurally identifiable model can be reached if one between the homeostatic proliferation rate and the death rate of effector and exhausted cells (one between α_0 and μ_E as well one between α_1 and μ_1) is experimentally measured or estimated from the literature.

References

Abdel-Hakeem, M.S., Manne, S., Beltra, J.C., Stelekati, E., Chen, Z., Nzingha, K., Ali, M.A., Johnson, J.L., Giles, J.R., Mathew, D., Greenplate, A.R., Vahedi, G. & Wherry, E.J. (2021). Epigenetic scarring of exhausted T cells hinders memory differentiation upon eliminating chronic antigenic stimulation. *Nature Immunology*, **22**, 1008–1019.

Alfei, F., Kanev, K., Hofmann, M., Wu, M., Ghoneim, H.E., Roelli, P., Utzschneider, D.T., von Hoesslin, M., Cullen, J.G., Fan, Y., Eisenberg, V., Wohlleber, D., Steiger, K., Merkler, D., Delorenzi, M., Knolle, P.A., Cohen, C.J., Thimme, R., Youngblood, B. & Zehn, D. (2019). TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection. *Nature*, **571**, 265–269.

Allen, J.L. & Stein, F.M. (1964). On solution of certain Riccati differential equations. *The American Mathematical Monthly*, **71**, 1113–1115.

Allen, L.J. (2010). *An Introduction to Stochastic Processes with applications to biology*. CRC press, USA, 2nd edn.

Ando, S., Perkins, C.M., Sajiki, Y., Chastain, C., Valanparambil, R.M., Wieland, A., Hudson, W.H., Hashimoto, M., Ramalingam, S.S., Freeman, G.J., Ahmed, R. & Araki, K. (2023). mTOR regulates T cell exhaustion and PD-1–targeted immunotherapy response during chronic viral infection. *The Journal of Clinical Investigation*, **133**.

Angelosanto, J.M., Blackburn, S.D., Crawford, A. & Wherry, E.J. (2012). Progressive loss of memory T cell potential and commitment to exhaustion during chronic viral infection. *Journal of Virology*, **86**, 8161–8170.

Antal, T. & Krapivsky, P. (2011). Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, **2011**, P08018.

Antia, R., Bergstrom, C.T., Pilyugin, S.S., Kaech, S.M. & Ahmed, R. (2003). Models of CD8+ responses: 1. what is the antigen-independent proliferation program. *Journal of Theoretical Biology*, **221**, 585–598.

Arfken, G.B. & Weber, H.J. (2016). *Mathematical methods for physicists*. Elsevier.

Bailey, N.T.J. (1964). *Elements of Stochastic Processes With Applications to the Natural Sciences*, vol. 25. John Wiley & Sons.

Bailis, W. & Pear, W.S. (2014). *The Molecular Basis of T Cell Development and How Epigenetic/Transcriptional Deregulation Leads to T-ALL*. Springer.

Barber, D.L., Wherry, E.J., Masopust, D., Zhu, B., Allison, J.P., Sharpe, A.H., Freeman, G.J. & Ahmed, R. (2006). Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature*, **439**, 682–687.

Barile, M., Busch, K., Fanti, A.K., Greco, A., Wang, X., Oguro, H., Zhang, Q., Morrison, S.J., Rodewald, H.R. & Höfer, T. (2020). Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. *doi.org/10.1101/2020.08.06.239186*.

Beltra, J.C., Manne, S., Abdel-Hakeem, M.S., Kurachi, M., Giles, J.R., Chen, Z., Casella, V., Ngiow, S.F., Khan, O., Huang, Y.J., Yan, P., Nzingha, K., Xu, W., Amaravadi, R.K., Xu, X., Karakousis, G.C., Mitchell, T.C., Schuchter, L.M., Huang, A.C. & Wherry, E.J. (2020). Developmental relationships of four exhausted $CD8^+$ T cell subsets reveals underlying transcriptional and epigenetic landscape control mechanisms. *Immunity*, **52**, 825–841.

Berika, M., Elgayyar, M.E. & El-Hashash, A.H. (2014). Asymmetric cell division of stem cells in the lung and other systems. *Frontiers in Cell and Developmental Biology*, **2**, 33.

REFERENCES

Blackburn, S.D., Shin, H., Freeman, G.J. & Wherry, E.J. (2008). Selective expansion of a subset of exhausted CD8$^+$ T cells by αPD-L1 blockade. *Proceedings of the National Academy of Sciences*, **105**, 15016–15021.

Blank, C.U., Haining, W.N., Held, W., Hogan, P.G., Kallies, A., Lugli, E., Lynn, R.C., Philip, M., Rao, A., Restifo, N.P., Schietinger, A., Schumacher, T.N., Schwartzberg, P.L., Sharpe, A.H., Speiser, D.E., Wherry, E.J., Youngblood, B.A. & Zehn, D. (2019). Defining "T cell exhaustion". *Nature Reviews Immunology*, **19**, 665–674.

Bocharov, G.A. (1998). Modelling the dynamics of LCMV infection in mice: Conventional and exhaustive CTL responses. *Journal of Theoretical Biology*, **192**, 283–308.

Boer, R.J.D. & Perelson, A.S. (2013). Quantifying T lymphocyte turnover. *Journal of Theoretical Biology*, **327**, 45–87.

Boer, R.J.D., Oprea, M., Antia, R., Murali-Krishna, K., Ahmed, R. & Perelson, A.S. (2001). Recruitment times, proliferation, and apoptosis rates during the CD8$^+$ T-cell response to lymphocytic choriomeningitis virus. *Journal of Virology*, **75**, 10663–10669.

Borsa, M., Barnstorf, I., Baumann, N.S., Pallmer, K., Yermanos, A., Gräbnitz, F., Barandun, N., Hausmann, A., Sandu, I., Barral, Y. & Oxenius, A. (2019). Modulation of asymmetric cell division as a mechanism to boost CD8+ T cell memory. *Science Immunology*, **4**.

Böttcher, M.A., Dingli, D., Werner, B. & Traulsen, A. (2018). Replicative cellular age distributions in compartmentalized tissues. *Journal of The Royal Society Interface*, **15**.

Bowman, R.L., Busque, L. & Levine, R.L. (2018). Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell*, **22**, 157–170.

Boyce, W.E. & di Prima, R.C. (2009). *Elementary Differential Equations and Boundary Value Problems*. John Wiley & Sons.

Bresser, K., Kok, L., Swain, A.C., King, L.A., Jacobs, L., Weber, T.S., Perié, L., Duffy, K.R., de Boer, R.J., Scheerenand, F.A. & Schumacher, T.N. (2022). Replicative history marks transcriptional and functional disparity in the CD8+ T cell memory pool. *Nature Immunology*, **23**, 791–801.

Buchholz, V.R., Flossdorf, M., Hensel, I., Kretschmer, L., Weissbrich, B., Gräf, P., Verschoor, A., Schiemann, M., Höfer, T. & Busch, D.H. (2013). Disparate individual fates compose robust CD8+ T cell immunity. *Science*, **340**, 630–634.

Bugaj, L.J. & Lim, W.A. (2019). High-throughput multicolor optogenetics in microwell plates. *Nature Protocols*, **14**, 2205–2228.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Höfer, T. & Rodewald, H.R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, **518**, 542–546.

Caldwell, C., Johnson, C.E., Balaji, V.N., Balaji, G.A., Hammer, R.D. & Kannan, R. (2017). Identification and validation of a PD-L1 binding peptide for determination of PD-L1 expression in tumors. *Scientific Reports*, **7**, 1–11.

Capasso, V. (2008). *Mathematical Structures of Epidemic Systems*, vol. 97. Springer Science & Business Media.

Castro, M. & de Boer, R.J. (2020). Testing structural identifiability by a simple scaling method. *PLoS Computational Biology*, **16**, e1008248.

Castro, M., López-García, M., Lythe, G. & Molina-París, C. (2018). First passage events in biological systems with non-exponential inter-event times. *Scientific Reports*, **8**, 1–16.

Chang, J.T., Palanivel, V.R., Kinjyo, I., Schambach, F., Intlekofer, A.M., Banerjee, A., Longworth, S.A., Vinup, K.E., Mrass, P., Oliaro, J., Killeen, N., Orange, J.S., Russell, S.M., Weninger,

W. & Reiner, S.L. (2007). Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science*, **315**, 1687–1691.

Chen, L. & Flies, D.B. (2013). Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nature*, **13**, 227–242.

Chen, Z., Ji, Z., Ngiow, S.F., Manne, S., Cai, Z., Huang, A.C., Johnson, J., Staupe, R.P., Bengsch, B., Xu, C., Yu, S., Kurachi, M., Herati, R.S., Vella, L.A., Baxter, A.E., Wu, J.E., Khan, O., Beltra, J.C., Giles, J.R., Stelekati, E., McLane, L.M., Lau, C.W., Yang, X., Berger, S.L., Vahedi, G., Ji, H., & Wherry, E.J. (2019). TCF-1-centered transcriptional network drives an effector versus exhausted CD8 T cell-fate decision. *Immunity*, **51**, 840–855.

Chow, A., Perica, K., Klebanoff, C.A. & Wolchok, J.D. (2022). Clinical implications of T cell exhaustion for cancer immunotherapy. *Nature Reviews Clinical Oncology*, **19**, 775–790.

Chu, H.H., Chan, S.W., Gosling, J.P., Blanchard, N., Tsitsiklis, A., Lythe, G., Shastri, N., Molina-París, C. & Robey, E.A. (2016). Continuous effector $CD8^+$ T cell production in a controlled persistent infection is sustained by a proliferative intermediate population. *Immunity*, **45**, 159–171.

Chung, H.K., McDonald, B. & Kaech, S.M. (2021). The architectural design of CD8+ T cell responses in acute and chronic infection: Parallel structures with divergent fates. *Journal of Experimental Medicine*, **218**.

Ciofani, M. & Zúñiga-Pflücker, J.C. (2007). The thymus as an inductive site for T lymphopoiesis. *Annual Review Cell Development Biology*, **23**, 463–493.

Colijn, C. & Mackey, M.C. (2005). A mathematical model of hematopoiesis - I. Periodic chronic myelogenous leukemia. *Journal of Theoretical Biology*, **237**, 117–132.

Conte, S.D. & De Boor, C. (2017). *Elementary numerical analysis: an algorithmic approach*. SIAM.

Csilléry, K., Blum, M.G., Gaggiotti, O.E. & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in ecology & evolution*, **25**, 410–418.

Day, C.L., Kaufmann, D.E., Kiepiela, P., Brown, J.A., Moodley, E.S., Reddy, S., Mackey, E.W., Miller, J.D., Leslie, A.J., DePierres, C., Mncube, Z., Duraiswamy, J., Zhu, B., Eichbaum, Q., Altfeld, M., Wherry, E.J., Coovadia, H.M., Goulder, P.J.R., Klenerman, P., Ahmed, R., Freeman, G.J. & Walker, B.D. (2006). PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression. *Nature*, **443**, 350–354.

de la Higuera, L., López-García, M., Castro, M., Abourashchi, N., Lythe, G. & Molina-París, C. (2019). Fate of a naive T cell: A stochastic journey. *Frontiers in Immunology*, **10**, 194.

den Braber, I., Mugwagwa, T., Vrisekoop, N., Westera, L., Mögling, R., de Boer, A.B., Willems, N., Schrijver, E.H.R., Spierenburg, G., Gaiser, K., Mul, E., Otto, S.A., Ruiter, A.F.C., Ackermans, M.T., Miedema, F., Borghans, J.A.M., de Boer, R.J. & Tesselaar, K. (2012). Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, **36**, 288–297.

Derényi, I. & Szöllősi, G.J. (2017). Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations. *Nature Communications*, **8**, 1–8.

Dinarello, C.A. (2007). Historical insights into cytokines. *European Journal of Immunology*, **37**, S34–S45.

Dingli, D., Traulsen, A. & Pacheco, J.M. (2007). Compartmental architecture and dynamics of hematopoiesis. *PloS one*, **2**, e345.

Dolina, J.S., Braeckel-Budimir, N.V., Thomas, G.D. & Salek-Ardakani, S. (2021). CD8+ T cell exhaustion in cancer. *Frontiers in Immunology*, **12**, 715234.

REFERENCES

Dreiwi, H., Feliciangeli, F., Castro, M., Lythe, G., Molina-París, C. & López-García, M. (2021). A stochastic model of cell proliferation and death across a sequence of compartments. *arXiv preprint arXiv:2111.11753*.

Duffy, K.R. & Hodgkin, P.D. (2012). Intracellular competition for fates in the immune system. *Trends in cell biology*, **22**, 457–464.

Eftimie, R., Gillard, J.J. & Cantrell, D.A. (2016). Mathematical models for immunology: Current state of the art and future research directions. *Bulletin of mathematical biology*, **78**, 2091–2134.

Erban, R., Chapman, S.J. & Maini, P.K. (2007). A practical guide to stochastic simulations of reaction-diffusion processes. *arXiv preprint arXiv:0704.1908*.

Evans, C.M. & Jenner, R.G. (2013). Transcription factor interplay in T helper cell differentiation. *Briefings in functional genomics*, **12**, 499–511.

Evans, M. & Kaufman, M. (1981). Establishment in culture of pluripotent cells from mouse embryos. *Nature*.

Feliciangeli, F., Dreiwi, H., López-García, M., Castro, M., Molina-París, C. & Lythe, G. (2022). Why are cell populations maintained via multiple compartments? *Journal of the Royal Society Interface*, **19**, 20220629.

Flajolet, P. & Sedgewick, R. (2009). *Analytic combinatorics*. Cambridge University Press.

Frank, S.A. (2007). *Dynamics of Cancer Incidence, Inheritance, and Evolution*. Princeton University Press.

Freiling, G. (2002). A survey of nonsymmetric Riccati equations. *Linear algebra and its applications*, **351**, 243–270.

Gao, Z., Feng, Y., Xu, J. & Liang, J. (2022). T-cell exhaustion in immune-mediated inflammatory diseases: New implications for immunotherapy. *Frontiers in Immunology*, **13**, 977394.

Gerlach, C., Rohr, J.C., Perié, L., van Rooij, N., van Heijst, J.W., Velds, A., Urbanus, J., Naik, S.H., Jacobs, H., Beltman, J.B., de Boer, R.J. & Schumacher, T.N.M. (2013). Heterogeneous differentiation patterns of individual CD8+ T cells. *Science*, **340**, 635–639.

Gillespie, D. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, **22**, 403–434.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, **81**, 2340–2361.

Gray, S.M., Amezquita, R.A., Guan, T., Kleinstein, S.H. & Kaech, S.M. (2017). Polycomb repressive complex 2-mediated chromatin repression guides effector CD8+ T cell terminal differentiation and loss of multipotency. *Immunity*, **46**, 596–608.

Greene, D.H. & Knuth, D.E. (1990). *Mathematics for the Analysis of Algorithms*. Birkhauser, Boston-Basel-Stuttgart.

Grelaud, A., Robert, C.P., Marin, J.M., Rodolphe, F. & Taly, J.F. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, **4**, 317 – 335.

Grimmett, G. & Welsh, D. (2014). *Probability: an introduction*. Oxford University Press.

Gross, D., Shortle, J.F., Thompson, J.M. & Harris, C.M. (2008). *Fundamentals of Queueing Theory*. John Wiley & Sons.

Gutcher, I. & Becher, B. (2007). APC-derived cytokines and T cell polarization in autoimmune inflammation. *The Journal of clinical investigation*, **117**, 1119–1127.

Gómez-Corral, A. & López-García, M. (2018). Perturbation analysis in finite LD-QBD processes and applications to epidemic models. *Numerical linear algebra with applications*, **25**, e2160.

REFERENCES

Gómez-Corral, A., López-García, M., Lopez-Herrero, M.J. & Taipe, D. (2020). On first-passage times and sojourn times in finite QBD processes and their applications in epidemics. *Mathematics*, **8**, 1718.

Harris, T.E. (1963). *The theory of branching processes*. Springer-Verlag, Berlin.

Hashimoto, M., Kamphorst, A.O., Im, S.J., Kissick, H.T., Pillai, R.N., Ramalingam, S.S., Araki, K. & Ahmed, R. (2018). CD8 T cell exhaustion in chronic infection and cancer: Opportunities for interventions. *Annual Review of Medicine*, **69**, 301–318.

Hawkins, E.D., Hommel, M., Turner, M.L., Battye, F.L., Markham, J.F. & Hodgkin, P.D. (2007). Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nature protocols*, **2**, 2057–2067.

He, R., Hou, S., Liu, C., Zhang, A., Bai, Q., Han, M., Yang, Y., Wei, G., Shen, T., Yang, X., Xu, L., Chen, X., Hao, Y., Wang, P., Zhu, C., Ou, J., Liang, H., Ni, T., Zhang, X., Zhou, X., Deng, K., Chen, Y., Luo, Y., Xu, J., Qi, H., Wu, Y. & Ye, L. (2016). Follicular CXCR5-expressing $CD8^+$ T cells curtail chronic viral infection. *Nature*, **537**, 412–416.

Heinzel, S., Marchingo, J.M., Horton, M.B. & Hodgkin, P.D. (2018). The regulation of lymphocyte activation and proliferation. *Current Opinion in Immunology*, **51**, 32–38.

Hodzic, E. (2016). Single-cell analysis: Advances and future perspectives. *Bosnian journal of basic medical sciences*, **16**, 313.

Höfer, T., Barile, M. & Flossdorf, M. (2016). Stem-cell dynamics and lineage topology from *in vivo* fate mapping in the hematopoietic system. *Current Opinion in Biotechnology*, **39**, 150–156.

Huang, A.C., Orlowski, R.J., Xu, X., Mick, R., George, S.M., Yan, P.K., Manne, S., Kraya, A.A., Wubbenhorst, B., Dorfman, L., D'Andrea, K., Wenz, B.M., Liu, S., Chilukuri, L., Kozlov, A., Carberry, M., Giles, L., Kier, M.W., Quagliarello, F., McGettigan, S., Kreider, K., Annamalai, L., Zhao, Q., Mogg, R., Xu, W., Blumenschein, W.M., Yearley, J.H., Linette, G.P., Amaravadi, R.K., Schuchter, L.M., Herati, R.S., Bengsch, B., Nathanson, K.L., Farwell, M.D., Karakousis, G.C., Wherry, E.J. & Mitchell, T.C. (2019). A single dose of neoadjuvant PD-1 blockade predicts clinical outcomes in resectable melanoma. *Nature medicine*, **25**, 454–461.

Hudson, W.H., Gensheimer, J., Hashimoto, M., Wieland, A., Valanparambil, R.M., Li, P., Lin, J.X., Konieczny, B.T., Im, S.J., Freeman, G.J., Leonard, W.J., Kissick, H.T. & Ahmed, R. (2019). Proliferating transitory T cells with an effector-like transcriptional signature emerge from PD-1^+ stemlike $CD8^+$ T cells during chronic infection. *Immunity*, **51**, 1043–1058.

Im, S.J., Hashimoto, M., Gerner, M.Y., Lee, J., Kissick1, H.T., Burger, M.C., Shan, Q., Hale, J.S., Lee, J., Nasti, T.H., Sharpe, A.H., Freeman, G.J., Germain, R.N., Nakaya, H.I., Xue, H.H. & Ahmed, R. (2016). Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. *Nature*, **537**, 417–421.

Janeway, C.A., Travers, P., Walport, M. & Capra, J. (1999). *Immunobiology: The Immune System in Health and Disease*. Garland Science Publishing.

Johnson, P.L.F., Kochin, B.F., McAfee, M.S., Stromnes, I.M., Regoes, R.R., Ahmed, R., Blattman, J.N., & Antia, R. (2011). Vaccination alters the balance between protective immunity, exhaustion, escape, and death in chronic infections. *Journal of Virology*, **85**, 5565–5570.

Johnston, M.D., Edwards, C.M., Bodmer, W.F., Maini, P.K. & Chapman, S.J. (2007). Mathematical modeling of cell population dynamics in the

colonic crypt and in colorectal cancer. *Proceedings of the National Academy of Sciences*, **104**, 4008–4013.

Kaech, S.M. & Ahmed, R. (2001). Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naïve cells. *Nature immunology*, **2**, 415–422.

Kaech, S.M. & Cui, W. (2012). Transcriptional control of effector and memory CD8+ T cell differentiation. *Nature Review Immunology*, **12**, 749–761.

Kaech, S.M., Hemby, S., Kersh, E. & Ahmed, R. (2002a). Molecular and functional profiling of memory CD8+ T cell differentiation. *Cell*, **111**, 837–851.

Kaech, S.M., Wherry, E.J. & Ahmed, R. (2002b). Effector and memory T-cell differentiation: Implications for vaccine development. *Nat Rev Immunol.*.

Kasmani, M.Y., Zander, R., Chung, H.K., Chen, Y., Khatun, A., Damo, M., Topchyan, P., Johnson, K.E., Levashova, D., Burns, R., Lorenz, U.M., Tarakanova, V.L., Joshi, N.S., Kaech, S.M. & Cui, W. (2022). Clonal lineage tracing reveals mechanisms skewing CD8+ T cell fate decisions in chronic infection. *Journal of Experimental Medicine*, **220**, e20220679.

Khan, O., Giles, J.R., McDonald, S., Manne1, S., Ngiow, S.F., Patel, K.P., Werner, M.T., Huang, A.C., Alexander, K.A., Wu, J.E., Attanasio, J., Yan, P., George, S.M., Bengsch, B., Staupe, R.P., Donahue, G., Xu, W., Amaravadi, R.K., Xu, X., Karakousis, G.C., Mitchell, T.C., Schuchter, L.M., Jonathan Kaye1 and, S.L.B. & Wherry, E.J. (2019). TOX transcriptionally and epigenetically programs $CD8^+$ T cell exhaustion. *Nature*, **571**, 211–218.

Kim, H.D., Park, S., Jeong, S., Lee, Y.J., Lee, H., Kim, C.G., Kim, K.H., Hong, S.M., Lee, J.Y., Kim, S., Kim, H.K., Min, B.S., Chang, J.H., Ju, Y.S., Shin, E.C., Song, G.W., Hwang, S. & Park, S.H. (2020). 4-1BB delineates distinct activation status of exhausted tumor-infiltrating CD8+ T cells in hepatocellular carcinoma. *Hepatology*, **71**, 955–971.

Kim, K.M., Choi, Y.J., Hwang, J.H., Kim, A.R., Cho, H.J., Hwang, E.S., Park, J.Y., Lee, S.H. & Hong, J.H. (2014). Shear stress induced by an interstitial level of slow flow increases the osteogenic differentiation of mesenchymal stem cells through TAZ activation. *PloS one*, **9**, e92427.

Kimmel, G.J., Locke, F.L. & Altrock, P.M. (2020). Response to CAR T cell therapy can be explained by ecological cell dynamics and stochastic extinction events. *bioRXiv preprint bioRXiv:10.1101/717074*.

Kimmel, G.J., Locke, F.L. & Altrock, P.M. (2021). The roles of T cell competition and stochastic extinction events in chimeric antigen receptor T cell therapy. *Proceedings of the Royal Society B*, **288**, 20210229.

Kimmel, M. & Axelrod, D.E. (2002). *Branching processes in biology*. Springer, USA, 9th edn.

Klein, A. & Simons, B. (2011). Universal patterns of stem cell fate in cycling adult tissues. *Development*, **138**, 3103–3111.

Knuth, D.E. & Wilf, H.S. (1989). A short proof of Darboux's lemma. *Applied Mathematics Letters*, **2**, iii–iv.

Kulkarni, V.G. (2017). *Modeling and analysis of stochastic systems*. Taylor & Francis Group.

Kumar, A., Chamoto, K., Chowdhury, P.S. & Honjo, T. (2020). Tumors attenuating the mitochondrial activity in T cells escape from PD-1 blockade therapy. *eLife*, **9**, e52330.

Kuznetsov, V., Makalkin, I., Taylor, M. & Perelson, A. (1994). Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis. *Bulletin of Mathematical Biology*, **56**, 295–321.

Lemieszek, M.B., Findlay, S.D. & Siegers, G.M. (2022). CellTrace™ Violet flow cytometric assay to assess cell proliferation. In *Cancer Cell Biology: Methods and Protocols*, 101–114, Springer.

León-Triana, O., Pérez-Martínez, A., Ramírez-Orellana, M. & Pérez-García, V.M. (2021). Dual-target CAR-Ts with on- and off-tumour activity may override immune suppression in solid cancers: A mathematical proof of concept. *Cancers*, **13**, 703.

Li, J., He, Y., Hao, J., Ni, L. & Dong, C. (2018). High levels of Eomes promote exhaustion of anti-tumor CD8+ T cells. *Frontiers Immunology*, **9**, 2981.

Luria, S.E. & Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, **28**, 491.

Lyne, A.M., Laplane, L. & Perié, L. (2021). To portray clonal evolution in blood cancer, count your stem cells. *Blood*, **137**, 1862–1870.

Lyons, A.B. (2000). Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution. *Journal of immunological methods*, **243**, 147–154.

López-García, M., Nowicka, M., Bendtsen, C., Lythe, G., Ponnambalam, S. & Molina-París, C. (2018). Quantifying the phosphorylation timescales of receptor–ligand complexes: a Markovian matrix-analytic approach. *Open biology*, **8**, 180126.

MacLean, A.L., Hong, T. & Nie, Q. (2018). Exploring intermediate cell states through the lens of single cells. *Current Opinion in Systems Biology*, **9**, 32–41.

Makhlouf, A.M., El-Shennawy, L. & Elkaranshawy, H.A. (2020). Mathematical modelling for the role of CD4+ T cells in tumor-immune interactions. *Computational and Mathematical Methods in Medicine*, **2020**.

Manescu, I.B., Manu, D.R., Serban, G.M. & Dobreanu, M. (2020). Variability of ex-vivo stimulated T-cells secretory profile in healthy subjects. *Revista Romana de Medicina de Laborator*, **28**, 75–89.

Mann, T.H. & Kaech, S.M. (2019). Tick-TOX, it's time for T cell exhaustion. *Nature Immunology*, **20**, 1092–1094.

Marchingo, J., Prevedello, G., Kan, A., Heinzel, S., Hodgkin, P. & Duffy, K. (2016). T-cell stimuli independently sum to regulate an inherited clonal division fate. *Nature Communications*, **7**, 1–12.

Marchingo, J.M., Kan, A., Sutherland, R.M., Duffy, K.R., Wellard, C.J., Belz, G.T., Lew, A.M., Dowling, M.R., Heinzel, S. & Hodgkin, P.D. (2014). Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science*, **346**, 1123–1127.

Marciniak-Czochra, A., Stiehl, T., Ho, A.D., Jäger, W. & Wagner, W. (2009). Modeling of replicative senescence in hematopoietic development. *Aging (Albany NY)*, **1**, 723.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, **78**, 7634–7638.

Massonis, G., Banga, J.R. & Villaverde, A.F. (2021). Structural identifiability and observability of compartmental models of the COVID-19 pandemic. *Annual reviews in control*, **51**, 441–459.

Mathar, R.J. (2021). P-finite recurrences from generating functions with roots of polynomials. *arXiv preprint arXiv:2109.02112*.

Matis, J. & Wehrly, T. (1994). Compartmental models of ecological and environmental systems. vol. 12 of *Handbook of Statistics*, 583–613, Elsevier.

Matis, J.H. (1970). *Stochastic compartmental analysis: model and least squares estimation from time series data*. Texas A&M University.

McLane, L.M., Abdel-Hakeem, M.S. & Wherry, E.J. (2019). CD8 T cell exhaustion during chronic viral infection and cancer. *Annual Review of Immunology*, **37**, 457–495.

REFERENCES

Michor, F., Hughes, T.P., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C.L. & Nowak, M.A. (2005). Dynamics of chronic myeloid leukaemia. *Nature*, **435**, 1267–1270.

Miller, B.C., Sen, D.R., Abosy, R.A., Bi, K., Virkud, Y.V., LaFleur, M.W., Yates, K.B., Lako, A., Felt, K., Naik, G.S., Manos, M., Gjini, E., Kuchroo, J.R., Ishizuka, J.J., Collier, J.L., Griffin, G.K., Maleri, S., Comstock, D.E., Weiss, S.A., Brown, F.D., Panda, A., Zimmer, M.D., Manguso, R.T., Hodi, F.S., Rodig, S.J., Sharpe, A.H. & Haining, W.N. (2019). Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade. *Nature Immunology*, **20**, 326–336.

Mo, F., Yu, Z., Li, P., Oh, J., Spolski, R., Zhao, L., Glassman, C.R., Yamamoto, T.N., Chen, Y., Golebiowski, F.M., Hermans, D., Majri-Morrison, S., Picton, L.K., Liao, W., Ren, M., Zhuang, X., Mitra, S., Lin, J.X., Gattinoni, L., Powell, J.D., Restifo, N.P., Garcia, K.C. & Leonard, W.J. (2021). An engineered IL-2 partial agonist promotes CD8+ T cell stemness. *Nature*, **597**, 544–548.

Montagna, W., Kligman, A. & Carlisle, K. (1992). Overview of skin. *Atlas of Normal Human Skin*, 3–5.

Moore, H. & Li, N.K. (2004). A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction. *Journal of Theoretical Biology*, **227**, 513–523.

Muegge, K. & K.Durum, S. (1990). Cytokines and transcription factors. *Cytokine*, **2**, 1–8.

Mueller, D.L., Jenkins, M.K. & Schwartz, R.H. (1989). Clonal expansion versus functional clonal inactivation: a costimulatory signalling pathway determines the outcome of T cell antigen receptor occupancy. *Annual Review of Immunology*, **7**, 445–480.

Murphy, G.M. (1960). *Ordinary differential equations and their solutions*. Newyork: Van Nostrand.

Murray, P.J., Walter, A., Fletcher, A.G., Edwards, C.M., Tindall, M.J. & Maini, P.K. (2011). Comparing a discrete and continuum model of the intestinal crypt. *Physical Biology*, **8**.

Mushegian, A.R. (2020). Are there 10^{31} virus particles on earth, or more, or fewer? *Journal of bacteriology*, **202**, e00052–20.

Nikolopoulou, E., Johnson, L.R., Harris, D., Nagy, J.D., Stites, E.C. & Kuang, Y. (2018). Tumour-immune dynamics with an immune checkpoint inhibitor. *Letters in Biomathematics*, **5**, S137–S159.

Nordon, R.E., Ko, K.H., Odell, R. & Schroeder, T. (2011). Multi-type branching models to describe cell differentiation programs. *Journal of Theoretical Biology*, **277**, 7–18.

Norris, J. (1997). *Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Oberbroeckling, L.A. (2021). *Programming Mathematics Using MATLAB*. Academic Press.

Ogawa, M. (1993). Differentiation and proliferation of hematopoietic stem cells. *Blood*, **81**, 2844–2853.

Owens, K. & Bozic, I. (2021). Modeling CAR T-Cell therapy with patient preconditioning. *Bulletin of Mathematical Biology*, **83**, 1–36.

Pala, Y. & Ertas, M.O. (2017). An analytical method for solving general Riccati equation. *International Journal of mathematical and computational sciences*, **11**, 125–130.

Paley, M.A., Kroy, D.C., Pamela M. Odorizzi and, J.B.J., Dolfi, D.V., Barnett, B.E., Bikoff, E.K., Robertson, E.J., Lauer, G.M., Reiner, S.L. & Wherry, E.J. (2012). Progenitor and terminal subsets of

CD8+ T cells cooperate to contain chronic viral infection. *Science*, **338**, 1220–1225.

Pauken, K.E. & Wherry, E.J. (2015). Overcoming T cell exhaustion in infection and cancer. *Trends in Immunology*, **36**, 265–276.

Pauken, K.E., Sammons, M.A., Odorizzi, P.M., Manne, S., Godec, J., Khan, O., Drake, A.M., Chen, Z., Sen, D.R., Kurachi, M., Barnitz, R.A., Bartman, C., Bengsch, B., Huang, A.C., Schenkel, J.M., Vahedi, G., Haining, W.N., Berger, S.L. & Wherry, E.J. (2016). Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science*, **354**.

Perié, L., Hodgkin, H.P.D., Naik, S.H., Schumacher, T.N., de Boer, R.J. & Duffy, K.R. (2014). Determining lineage pathways from cellular barcoding experiments. *Cell reports*, **6**, 617–624.

Pham, K., Shimoni, R., Charnley, M., Ludford-Menting, M.J., Hawkins, E.D., Ramsbottom, K., Oliaro, J., Izon, D., Ting, S.B. & Reynolds, J. (2015). Asymmetric cell division during T cell development controls downstream fate. *Journal of Cell Biology*, **210**, 933–950.

Philip, M., Fairchild, L., Sun, L., Horste, E.L., Camara, S., Shakiba, M., Scott, A.C., Viale, A., Lauer, P., Merghoub, T., Hellmann, M.D., Wolchok, J.D., Leslie, C.S. & Schietinger, A. (2017). Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature*, **545**, 452–456.

Pipkin, M.E., Sacks, J.A., Cruz-Guilloty, F., Lichtenheld, M.G., Bevan, M.J. & Rao, A. (2010). Interleukin-2 and inflammation induce distinct transcriptional programs that promote the differentiation of effector cytolytic T cells. *Immunity*, **32**, 79–90.

Plambeck, M., Kazeroonian, A., Loeffler, D., Kretschmer, L., Salinno, C., Schroeder, T., Busch, D.H., Flossdorf, M. & Buchholz, V.R. (2022). Heritable changes in division speed accompany the diver-

sification of single t cell fate. *Proceedings of the National Academy of Sciences*, **119**, e2116260119.

Plumlee, C.R., Sheridan, B.S., Cicek, B.B. & Lefrançois, L. (2013). Environmental cues dictate the fate of individual CD8+T cells responding to infection. *Immunity*, **39**, 347–356.

Rao, P.R.P. (1962). The Riccati differential equation. *The American Mathematical Monthly*, **69**, 995–996.

Rao, R.R., Li, Q., Odunsi, K. & Shrikant, P.A. (2010). The mTOR kinase determines effector versus memory CD8+ T cell fate by regulating the expression of transcription factors T-bet and Eomesodermin. *Immunity*, **32**, 67–78.

Raskov, H., Orhan, A., Christensen, J.P. & Gögenur, I. (2021). Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. *British Journal of Cancer*, **124**, 359–367.

Reya, T., Morrison, S.J., Clarke, M.F. & Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. *Nature*, **414**, 105–111.

Robert, P.A., Kunze-Schumacher, H., Greiff, V. & Krueger, A. (2021). Modeling the dynamics of T-cell development in the thymus. *Entropy*, **23**, 437.

Sahoo, P., Yang, X., Abler, D., Maestrini, D., Adhikarla, V., Frankhouser, D., Cho, H., Machuca, V., Wang, D., Barish, M., Gutova, M., Branciamore, S., Brown, C.E. & Rockne, R.C. (2020). Mathematical deconvolution of CAR T-cell proliferation and exhaustion from real-time killing assay data. *Journal of the Royal Society Interface*, **17**, 20190734.

Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. (1999). Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature*, **401**, 708–712.

REFERENCES

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (2008). *Global Sensitivity Analysis. The primer*. John Wiley & Sons.

Sawicka, M., Stritesky, G., Reynolds, J., Abourashchi, N., Lythe, G., Molina-París, C. & Hogquist, K. (2014). From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection. *Frontiers in Immunology*, **5**, 19.

Schietinger, A., Philip, M., Krisnawan, V.E., Chiu, E.Y., Delrow, J.J., Basom, R.S., Lauer, P., Brockstedt, D.G., Knoblaugh, S.E., Hammerling, G.J., Schell, T.D., Garbi, N. & Greenberg, P.D. (2016). Tumor-specific T cell dysfunction is a dynamic antigen-driven differentiation program initiated early during tumorigenesis. *Immunity*, **45**, 389–401.

Scott, A.C., Dündar, F., Zumbo, P., Chandran, S.S., Klebanoff, C.A., Shakiba, M., Trivedi, P., Menocal, L., Appleby, H., Camara, S., Zamarin, D., Walther, T., Snyder, A., Femia, M.R., Comen, E.A., Wen, H.Y., Hellmann, M.D., Anandasabapathy, N., Liu, Y., Altorki, N.K., Lauer, P., Levy, O., Glickman, M.S., Kaye, J., Betel, D., Philip, M. & Schietinger, A. (2019). TOX is a critical regulator of tumour-specific T cell differentiation. *Nature*, **571**, 270–274.

Seita, J. & Weissman, I.L. (2010). Hematopoietic stem cell: Self-renewal versus differentiation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **2**, 640–653.

Seo, H., Chen, J., González-Avalos, E., Samaniego-Castruita, D., Das, A., Wang, Y.H., López-Moyado, I.F., Georges, R.O., Zhang, W., Onodera, A., Wu, C.J., Lu, L.F., Hogan, P.G., Bhandoola, A. & Rao, A. (2019). TOX and TOX2 transcription factors cooperate with NR4A transcription factors to impose $CD8^+$ T cell exhaustion. *Proceedings of the National Academy of Sciences*, **116**, 12410–12415.

Serre, R., Benzekry, S., Padovani, L., Meille, C., André, N., Ciccolini, J., Barlesi, F., Muracciole, X. & Barbolosi, D. (2016). Mathematical modeling of cancer immunotherapy and its synergy with radiotherapy. *Cancer Research*, **76**, 4931–4940.

Shahriyari, L. & Komarova, N.L. (2013). Symmetric vs. asymmetric stem cell divisions: an adaptation against cancer? *PloS one*, **8**, e76195.

Sharpe, A.H. & Pauken, K.E. (2018). The diverse functions of the PD-1 inhibitory pathway. *Nature Reviews Immunology*, **18**, 153–167.

Shin, H. & Wherry, E.J. (2007). CD8 T cell dysfunction during chronic viral infection. *Current Opinion in Immunology*, **19**, 408–415.

Singer, A., Adoro, S. & Park, J. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nature Reviews Immunology*, **8**, 788–801.

Singmaster, D. (1978). An elementary evaluation of the Catalan numbers. *The American Mathematical Monthly*, **85**, 366–368.

Sobol, I.M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, **55**, 271–280.

Stanley, R.P. (1980). Differentiably finite power series. *European journal of combinatorics*, **1**, 175–188.

Steel, J.M. (2001). *Stochastic Calculus and Financial Applications*, vol. 1. Springer.

Stiehl, T. & Marciniak-Czochra, A. (2011). Characterization of stem cells using mathematical models of multistage cell lineages. *Mathematical and Computer Modelling*, **53**, 1505–1517.

Stiehl, T. & Marciniak-Czochra, A. (2017). Stem cell self-renewal in regeneration and cancer: insights from mathematical modeling. *Current Opinion in Systems Biology*, **5**, 112–120.

REFERENCES

Stirzaker, D. (2005). *Stochastic processes and models*. Oxford University Press.

Stromberg, S.P. & Antia, R. (2012). On the role of CD8 T cells in the control of persistent infections. *Biophysical Journal*, **103**, 1802–1810.

Sugai, I. (1960). Riccati's nonlinear differential equation. *The American Mathematical Monthly*, **67**, 134–139.

Tabana, Y., Moon, T.C., Siraki, A., Elahi, S. & Barakat, K. (2021). Reversing T-cell exhaustion in immunotherapy: a review on current approaches and limitations. *Expert Opinion on Therapeutic Targets*, **25**, 347–363.

Teixeira, L.K., Fonseca, B.P.F., de Abreu, A.V., Barboza, B.A., Robbs, B.K., Bozza, P.T. & Viola, J.P.B. (2005). IFN-γ production by $CD8^+$ T cells depends on NFAT1 transcription factor and regulates Th differentiation. *The Journal of Immunology*, **175**, 5931–5939.

Teschl, G. (2012). *Ordinary differential equations and dynamical systems*, vol. 140. American Mathematical Society, USA.

Thomas, L.H. (1949). Elliptic problems in linear difference equations over a network. *Watson Sci. Comput. Lab. Rept*, **1**, 71.

Thomas-Vaslin, V., Altes, H.K., de Boer, R.J. & Klatzmann, D. (2008). Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *Journal Immunology*, **180**, 2240.

Thommen, D.S., Koelzer, V.H., Herzig, P., Roller, A., Trefny, M., Dimeloe, S., Kiialainen, A., Hanhart, J., Schill, C., Hess, C., Prince, S.S., Wiese, M., Lardinois, D., Ho, P.C., Klein, C., Karanikas, V., Mertz, K.D., Schumacher, T.N. & Zippelius, A. (2018). A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in nonsmall-cell lung cancer treated with PD-1 blockade. *Nature Medicine*, **24**, 994–1004.

Toettcher, J.E., Weiner, O.D. & Lim, W.A. (2013). Using optogenetics to interrogate the dynamic control of signal transmission by the Ras/Erk module. *Cell*, **155**, 1422–1434.

Toni, T. & Stumpf, M. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, **26**, 104.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187.

Tower, J. (2012). Stress and stem cells. *Wiley Interdisciplinary Reviews: Developmental Biology*, **1**, 789–802.

Utzschneider, D.T., Legat, A., Marraco, S.A.F., Carrié, L., Luescher, I., Speiser, D.E. & Zehn, D. (2013). T cells maintain an exhausted phenotype after antigen withdrawal and population reexpansion. *Nature Immunology*, **14**, 603–610.

Utzschneider, D.T., Charmoy, M., Chennupati, V., Pousse, L., Ferreira, D.P., Calderon-Copete, S., Danilo, M., Alfei, F., Hofmann, M., Wieland, D., Pradervand, S., Thimme, R., Zehn, D. & Held, W. (2016). T cell factor 1 - expressing memory-like CD8+ T cells sustain the immune response to chronic viral infections. *Immunity*, **45**, 415–427.

Utzschneider, D.T., Gabriel, S.S., Chisanga, D., Gloury, R., Gubser, P.M., Vasanthakumar, A., Shi, W. & Kallies, A. (2020). Early precursor T cells establish and propagate T cell exhaustion in chronic infection. *Nature Immunology*, **21**, 1256–1266.

Valentinuzzi, D. & Jeraj, R. (2020). Computational modelling of modern cancer immunotherapy. *Physics in Medicine & Biology*, **65**, 24TR01.

Wainscoat, J. & Fey, M. (1990). Assessment of clonality in human tumors: a review. *Cancer Research*, **50**, 1355–1360.

Waldman, A.D., Fritz, J.M. & Lenardo, M.J. (2020). A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology*, **20**, 651–668.

Watson, H.W. & Galton, F. (1875). On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **4**, 138–144.

Weber, E.W., Parker, K.R., Sotillo, E., Lynn, R.C., Anbunathan, H., Lattin, J., Good, Z., Belk, J.A., Daniel, B., Klysz, D., Malipatlolla, M., Xu, P., Bashti, M., Heitzeneder, S., Labanieh, L., Vandris, P., Majzner, R.G., Qi, Y., Sandor, K., Chen, L.C., Prabhu, S., Gentles, A.J., Wandless, T.J., Satpathy, A.T., Chang, H.Y. & Mackall, C.L. (2021). Transient rest restores functionality in exhausted CAR-T cells through epigenetic remodeling. *Science*, **372**, eaba1786.

Weekes, S.L., Barker, B., Bober, S., Cisneros, K., Cline, J., Thompson, A., Hlatky, L., Hahnfeldt, P. & Enderling, H. (2014). A multi-compartment mathematical model of cancer stem cell driven tumor growth dynamics. *Bulletin of mathematical biology*, **76**, 1762–1782.

Weinberg, R.A. (2013). *The biology of cancer*. Garland Science.

Werner, B., Beier, F., Hummel, S., Balabanov, S., Lassay, L., Orlikowsky, T., Dingli, D., Brümmendorf, T.H. & Traulsen, A. (2015). Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *eLife*, **4**, e08687.

Wherry, E. & Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nature Reviews Immunology*, **15**, 486–499.

Wherry, E.J. (2011). T cell exhaustion. *Nature Immunology*, **12**, 492–499.

Wherry, E.J., Teichgräber, V., Becker, T.C., Masopust, D., Kaech, S.M., Antia, R., von Andrian, U.H. & Ahmed, R. (2003). Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nature Immunology*, **4**, 225–234.

Wherry, E.J., Ha, S.J., Kaech, S.M., Haining, W.N., Sarkar, S., Kalia, V., Subramaniam, S., Blattman, J.N., Barber, D.L. & Ahmed, R. (2007). Molecular signature of $CD8^+$ T cell exhaustion during chronic viral infection. *Immunity*, **27**, 670–684.

Whichard, Z.L., Sarkar, C.A., Kimmel, M. & Corey, S.J. (2010). Hematopoiesis and its disorders: a systems biology approach. *Blood*, **115**, 2339–2347.

Wilf, H.S. (2005). *generatingfunctionology*. CRC press.

Wilson, A., Laurenti, E. & Trumpp, A. (2009). Balancing dormant and self-renewing hematopoietic stem cells. *Current Opinion in Genetics & Development*, **19**, 461–468.

Winkler, F. & Bengsch, B. (2020). Use of mass cytometry to profile human t cell exhaustion. *Frontiers in Immunology*, **10**, 3039.

Wodarz, D., Klenerman, P. & Nowak, M.A. (1998). Dynamics of cytotoxic T-lymphocyte exhaustion. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **265**, 191–203.

Wu, T., Ji, Y., Moseman, E.A., Xu, H.C., Manglani, M., Kirby, M., Anderson, S.M., Handon, R., Kenyon, E., Elkahloun, A., Wu, W., Lang, P.A., Gattinoni, L., McGavern, D.B. & Schwartzberg, P.L. (2016). The TCF1-Bcl6 axis counteracts type I interferon to repress exhaustion and maintain T cell stemness. *Science Immunology*, **1**, eaai8593.

Yang, J., Plikus, M.V. & Komarova, N.L. (2015). The role of symmetric stem cell divisions in tissue homeostasis. *PLoS Computational Biology*, **11**, e1004629.

Yates, A. (2014). Theories and quantification of thymic selection. *Frontiers in Immunology*, **5**, 13.

Zajac, A., Blattman, J., Murali-Krishna, K., Sourdive, D., Suresh, M., Altman, J. & Ahmed, R. (1998). Viral immune evasion due to persistence of activated T cells without effector function. *The Journal of experimental medicine*, **188**, 2205–2213.

Zehn, D., Thimme, R., Lugli, E., de Almeida, G.P. & Oxenius, A. (2022). "Stem-like" precursors are the fount to sustain persistent CD8+ T cell responses. *Nature Immunology*, **23**, 836–847.

Zhang, H., Hou, W., Henrot, L., Schnebert, S., Dumas, M., Heusèle, C. & Yang, J. (2015). Modelling epidermis homoeostasis and psoriasis pathogenesis. *Journal of The Royal Society Interface*, **12**.

Zhang, Z., Liu, S., Zhang, B., Qiao, L., Zhang, Y. & Zhang, Y. (2020). T cell dysfunction and exhaustion in cancer. *Frontiers in cell and developmental biology*, **8**, 17.

Zhou, D., Luo, Y., Dingli, D. & Traulsen, A. (2019). The invasion of de-differentiating cancer cells into hierarchical tissues. *PLoS Computational Biology*, **15**.

Zierler, K. (1981). A critique of compartmental analysis. *Annual review of biophysics and bioengineering*, **10**, 531–562.

Zion, E. & Chen, X. (2021). Breaking symmetry: the asymmetries in epigenetic inheritance. *The biochemist*, **43**, 14–19.