ULB FACULTÉ DES SCIENCES
UNIVERSITÉ LIBRE DE BRUXELLES

VUB VRIJE UNIVERSITEIT BRUSSEL

Statistical biophysics of hematopoiesis and growing cell populations

Thesis submitted by Nathaniel Vincent MON PÈRE
in fulfilment of the requirements of the PhD Degree in Science (ULB - "Docteur en Science") and in Sciences (VUB – "Doctor in de Wetenschappen")
Academic year 2020-2021

Supervisors: Professor Tom LENAERTS (Université libre de Bruxelles)
Machine Learning Group

and Professor Sophie DE BUYL (Vrije Universiteit Brussel)
Applied Physics Research Group

Thesis jury:

Dominique MAES (Vrije Universiteit Brussel, Chair)
Bortolo MOGNETTI (Université libre de Bruxelles, Secretary)
Lendert GELENS (KU Leuven)
Benjamin WERNER (Queen Mary University of London)

Statistical biophysics of hematopoiesis and growing cell populations

Nathaniel Vincent Mon Père

Contents

Foreword

This thesis has been four years in the making, and as such I would like to express some thanks to the people who helped make it a reality.

Many thanks to all of my wonderful colleagues, who made "going to work" exciting, which for some reason always felt like cheating; as well as all of my friends who, in their incredible kindness, tried really hard not to condemn me for my "fake job".

I am very grateful to Tom for trusting me to chase whatever crazy ideas I had, and for suffering the whims of my easily distracted brain. His help and encouragement in finding the research that interests me kept me motivated after every setback.

Sophie's supervision came at a time when I felt most nostalgic for the physics of my days as a student, and as such her help was both welcome and invaluable. Her enthusiasm for diving into a field neither of us was particularly familiar with was inspiring, and her meticulous approach proved to be the perfect fit for my rambling thought processes.

Finally, it is unlikely this thesis would exist if it weren't for my mom, in whose footsteps I've unconsciously been following the past 9 years. In helping me discover the joy of solving problems as a child, she instilled in me the love for mathematics that coaxed me into science as an adult. Though I never willfully decided what I wanted to be when I would grow up, her example has taught me that it's fine never to settle as long as I'm doing what I love. I'll let her know when I decide to grow up.

1. Introduction

> *Sometimes science is a lot more art than science, Morty.*
>
> — Rick Sanchez, *Rick & Morty*

Portraying the natural world around us through mathematics is a powerful tool for advancing our understanding of it. The formalization and rigorous application of this approach played an immense role in the scientific revolution pioneered by Renaissance scholars such as Copernicus and Newton, and would ultimately lead to the birth of modern physics and countless other fields derived from it. As our collective understanding of mathematics and the problems to which it has been applied grow, so also does the complexity of the systems under consideration. However, with this increasing complexity comes a reduced ability to obtain useful information: "difficult" problems are those which contain too many "moving parts", so that finding their solution is either unfeasible considering the time it would take to solve them, or pointless if their dependence on initial conditions is too strong.

The realm of stochastics has provided a suitable workaround for this issue, allowing problems that are difficult to solve exactly to be reduced to simpler forms, though at the expense of perfect information. This has in the past proven to be immensely powerful, with the achievements of statistical physics – relating the behavior of singular molecules and atoms to the empirically observed laws of thermodynamics – as a proud testament to its success. It is therefore no surprise that the 20th century saw the application of stochastic models across countless fields of study, from economics to human behavior, although accomplishments in such areas have been perhaps more modest. We cannot, for example, predict as accurately as we might like the outcome of an election, or the

influence of a global pandemic on the economy. What makes these systems harder to predict than those of classical physics is a question which merits its own discussion, though a simple answer is that their complexity makes them far more chaotic, resulting in behavior which is observed to be significantly more stochastic.

In contrast, employing stochastic mathematical models in biology may seem less extreme, it being merely a natural progression to apply the techniques honed through the study of the physical reality to the realm of life. However the issues described above are intensely present in biological systems as well. Over a century of investigations in cell biology have taught us that they are in fact tremendously complex machines. The comparatively "simple" physical laws describing the behavior of the atoms from which they are constructed – and which an aspiring biophysicist might be primed to apply – are only present in the same way that electrons are present in the circuit boards of a computer: They are fragments of a larger whole, which in itself acts as a component of another encompassing object, which in fact only serves the purpose of yet another whole, and so forth. The scope of such a system is daunting, however the possibility of this Russian-doll type of encapsulation simultaneously provides an opportunity: if we can course-grain the full picture to only consider certain "emergent" objects as the fundamental particles of the system, we might obtain new empirical "laws" which can be used to extract information. Still, this can be a challenging task. While, for example, the atom functions as an excellent model for the underlying interactions of its constituent bosons and fermions, there is no such simple model for a cell. Instead, we must choose the relevant properties of the desired fundamental "unit" based on the problems at hand. And given that we ignore so much of the underlying reality, these properties will often be stochastic in nature. In this sense the biophysical approach to a set of questions is as much choosing a model as it is solving it. Impressive as some models may be, the answer might very well be "42" and nobody would quite know what to do with it.

This thesis examines the application of mathematical models to two different systems in biology, both related to the behavior of human (or more generally mammalian) cells,

but both ultimately requiring different sets of questions and models. Though the mathematical tools applied are similar, there is little overlap in the models themselves, and as such these are treated in separate parts.

Part I covers an investigation of hematopoiesis, the process by which precursor cells of the blood are cultivated and matured in the bone marrow. It is essential to enable mammalian physiology, from providing oxygen-carrying erythrocytes to ensuring regular upkeep and preservation of the immune system. The general mechanism follows a pyramidal architecture, with rare slowly acting multipotent stem cells seeding more differentiated progenitors through successive levels of maturation. Obtaining a quantitative understanding of key aspects of this system can provide valuable insights and testable predictions concerning the origin and dynamics of various blood-related diseases such as anemia, hemochromatosis, leukemias, and other. However, *in vivo* bone marrow studies pose significant challenges and *in vitro* studies often provide only limited predictive power, as the hierarchical landscape of differentiation can rapidly lead to non-trivial dynamics. Such a system is on the other hand well fit to the application of mathematical and computational techniques relying only on a few basic assumptions and parameters. In this context three separate research questions are posed and investigated, respectively discussed in Chapters 5, 6, and 7.

The first question concerns an attempt to gain insight in the dynamics of the rare blood disorder paroxysmal nocturnal hemoglobinuria (PNH). It is an acquired blood disorder, clinically characterized by hemolysis (destruction of red blood cells) and a high risk of thrombosis (obstructive blood clotting). These symptoms are caused by a population of blood cells presenting a deficiency in several identifying surface proteins, the lack of which elicits an activation of the complement immune system and consequently their premature destruction. The origin of this defect has been traced to a somatic mutation in the *PIGA* gene, which furthermore is required to occur within the hematopoietic stem cells (HSCs) for the disease to present. However, to date the question of how this mutant clone expands in size to contribute significantly to hematopoiesis remains under debate. One hypothesis posits the existence of a selective advantage of *PIGA* mutated cells

due to an immune mediated attack on normal HSCs, however, the evidence supporting this hypothesis is inconclusive. An alternative explanation attributes clonal expansion to neutral drift, in which case selection neither favors nor inhibits the expansion of *PIGA* mutated HSCs. In this chapter the implications of the neutral drift model are examined, both in terms of its likelihood as well as its predicted dynamics. This is done by modeling the dynamics of the HSC compartment as a stochastic Moran process, and numerically evolving a Markov chain for the probabilities of all possible outcomes. We find predictions of the model to agree surprisingly well with the known incidence of the disease, as well as the average age at diagnosis. Furthermore, we observe that the predicted dynamical variation of *PIGA* mutated HSC clones in the model qualitatively matches what has been observed in human trials, and can be made to quantitatively fit the observed clonal expansion rates by introducing a coupling between the stem cell compartment and the blood, as in reality one would expect the extreme loss of cells due to hemolysis to cause a compensating increase in HSC production.

The second research question entails an attempt to better quantify the properties of the hematopoietic stem cell pool – as such knowledge can greatly benefit the construction of specialized models such as those of Chapter 5 and Chapter 7 – by analyzing patterns of somatic mutation accumulation within the population. As HSCs divide, they continuously acquire novel mutations at random positions in the genome. As most of these are selectively neutral – conferring neither an advantage nor a disadvantage to their carriers – their stochastic dynamics follow the same probabilistic trajectory as derived for the *PIGA* mutation in Chapter 5, so that the Moran model can (with some alterations) be similarly applied to obtain the dynamics for any number of mutant clones. However, the possible outcomes for the system of multiple of stochastically evolving clones form a highly complex state space, so that structured reductions of this space must be analyzed. In particular, we investigate the predicted distributions for the number of mutations found in a single cell – the mutational burden – as well as the number of variants found at a particular frequency – the variant allele frequency (VAF). These predictions are then applied to a dataset containing high resolution mutational information on a

sample of 89 human HSCs taken from a single patient, allowing us to estimate certain fundamental quantities such as the mutation rate per cell division and the rate of cell divisions. These results furthermore highlight the wealth of information encoded in the somatic mutational landscape, and the usefulness of the stochastic approach.

Finally, in Chapter 7 we turn our attention to the full hematopoietic system, with the question of what cellular dynamics can be expected to occur given our current understanding of its structure. A handful of models to this end have been proposed previously, however these have typically been introduced to describe either a particular observed phenomenon or a proposed gene regulation network, whereas a general characterization of how the system behaves – both under normal circumstances as well as under stress – remains lacking. Such a model could prove useful in understanding the dynamics of blood disorders which are typically non-trivial to predict – even qualitatively – due to both the hematopoietic system's pyramidal structure as well as its complex feedback networks. In order to characterize the possible dynamics following different types of perturbations, we investigate a generalized model of hematopoiesis which represents the system as a sequence of compartments covering all maturation stages from stem cells to committed progenitors, in which cells can both self-renew and differentiate. As the system's plasticity is driven by feedback networks which adapt the cell production to ongoing requirements, we design the model to transparently show the effect of different feedback types on the overall dynamics. We find that the dynamic character of the system following a transient perturbation depends on the balance between the altered differentiation and self-renewal rates, where a simultaneous adaptation of both rates is required in order for the system to maintain stability. Furthermore, we show that under continuous disruption – as found in certain hematopoietic disorders – compartment cell numbers may evolve to new equilibria, implying that chronic illnesses can lead to distorted size distributions of progenitor cell populations in the bone marrow.

In Part II of this thesis we will move away from specific tissues, and instead consider the mechanical properties of proliferating cell populations. Various cell types exists with

some form of locomotive capability, their recruitment occurring for a wide variety of tasks such as tissue regeneration, vascularization, and wound healing. Recently it has been found that such motility also occurs in cancer, as a differentiational pathway associated with cellular changes from an epithelial form to such a motile phenotype has been related to an increased metastasis risk. The application of stochastic models of motion can be useful in investigating such phenomena, however, while many models of self-driven particles (often referred to as active matter) such as cells have been studied, little has been shown for systems in which the particles simultaneously proliferate, effectively reducing the space available to them. Such models would be highly useful though, as increasingly complex culturing experiments investigating motile proliferating cell types require a robust mathematical framework to extract useful information. To quantitatively answer the question of how proliferative crowding influences motility, we approach the problem of growth in a confined space from first principles. Starting from the basic Langevin equation for Brownian motion, we introduce a dependence on the cell density through the system's diffusion coefficient and its relation to the moving particle's mean free path. By coupling the density to an appropriate growth curve – such as the logistic function – we show how apparent sublinear diffusion occurs as a result of the increasing particle density. As this model represents the influence of dynamic crowding on a ballistically moving particle, it can be coupled with any other models of active motion given by a Langevin representation, and may therefore be used to model proliferating particles with widely different locomotive capabilities. In fact, the notion of actively moving "agents" in a dynamically crowding environment is not limited to cells, and may find applications in widely different topics as well; from ecological models involving populations of animals which both migrate and reproduce, to human behavior in crowded city streets.

Published works

At the time of writing, some of the work presented in this thesis has been published, or is currently under review, in peer reviewed journals in the following articles:

- Mon Père, N., Lenaerts, T., Pacheco, J. M., & Dingli, D. (2018). Evolutionary dynamics of paroxysmal nocturnal hemoglobinuria. *PLoS computational biology*, 14(6), e1006133.
  - documents the methods and results described in Chapter 5.
- Mon Père, N. V., Lenaerts, T., Pacheco, J. M., & Dingli, D. (2020). Multistage feedback driven compartmental dynamics of hematopoiesis. *bioRxiv*.
  - presents the model and results discussed in Chapter 7, It is currently under review in iScience, and available in preprint on bioRxiv.

Furthermore, Chapters 6 and 9 are both currently being prepared as manuscripts for submission:

- Moeller, M.E., Mon Père, N.V., Werner, B., Huang, W. Dynamics of genetic diversification in normal somatic tissue. (in preparation)
  - Part of this manuscript deals with the results of Chapter 6.
- Mon Père, N.V., De Buyl, P, De Buyl, S. Crowded motion in proliferating populations. (in preparation)
  - This manuscript deals with the methods and results of Chapter 9.

Télévie funding

I would like to gratefully acknowledge the funding from the Télévie organisation, who made this research possible. They took a chance on a project which is considerably closer to fundamental research than is customary for this grant, for which I am very thankful. While the trajectory of this project ended up straying somewhat farther away from the original cancer-related questions than initially intended, I truly believe that a greater understanding of the fundamental processes regulating the human body and underlying tumorigenesis will ultimately prove indispensable in the long run.

Part I.

Population dynamics of hematopoiesis

2. Hematopoiesis: the factory for blood

Can't you recognize the human in the inhuman?

— Ray Bradbury, *The Martian Chronicles*

Blood plays a critical role in the machinery of the human body [70]. The expansive network of vessels and capillaries through which it is pumped spans the entirety of the body, and as such many responsibilities related to this interconnectivity are performed by the blood. This includes the transportation of oxygen and nutrients required for various metabolic processes to all tissues, as well as the removal of any resulting waste products; providing numerous messenger functionalities through the transport of hormones and other signaling factors; and performing immunological functions such as the detection and destruction of foreign cells and materials. The laundry list of tasks relegated to the blood is expansive, making it perhaps not so surprising to find its composition to contain a broad collection of cells, highly diverse in both function and phenotype. While there are different ways to categorize these so called *hemocytes*, usually either by morphology or by lineage, hematologists generally classify three major blood cell types: erythrocytes (red blood cells) – which primarily exist to transport oxygen; thrombocytes (platelets) – which cause clotting in reaction to bleeding; and leukocytes (white blood cells) – which are the cells of the immune system, and can themselves be densely partitioned into various subtypes which fulfill different functions within the body's natural defense system. Interestingly, while these different cell types vary drastically in morphology and function, they share a close familial relationship as products of the same cell production factory. Specifically, all mature blood cells – irrespective of their type – originated through successive cell divisions from the same small population of blood specific stem

cells in the bone marrow, not unlike how during the earliest stages of human development embryonic cells are tasked with developing the complex and diverse landscape of cells found throughout the body. The process by which all such hematopoietic cells are cultivated and primed for a particular type is referred to as hematopoiesis. The need for such a cell production factory is evident, as most mature blood cells are highly transient, with many leukocytes living for as little as a few hours up to a handful of days, and erythrocytes remaining for only 100-120 days in circulation before undergoing programmed cell death. The generality of the hematopoietic system and how it spans such a vast landscape of cell types is quite remarkable, and perhaps alludes to the mark of an evolutionarily crafted system.

2.1. A brief history of hematopoiesis

As with many fields of research, our picture of hematopoiesis has changed throughout the history of its study, though the bones of our understanding were introduced over a century ago. While the idea of hematopoiesis as a hierarchical system can be traced back to turn of the 20th Century, the earliest concepts of a differentiation process by which cells change from an unspecific template state to a highly specified functional state first emerged in the field of embryogenesis in the 1860s. It was the eminent German biologist Ernst Haeckel who introduced the term *stem cell*, though in a slightly different context, referring to a primordial unicellular organism at the root of an intricately branching evolutionary tree – the Stammbaum – wherein all multicellular life would trace their ancestry [41]. Only later did he use the term to denote an individual's first embryonic cell which would catalyze an ontogeny (the transformative development from embryo to adult) that recapitulated the species' evolution, the theory known as Haeckel's biogenetic-law. While not quite carrying the meaning of the word today, its captivating imagery proved appealing, leading to the term's adoption by others in the field. It was the germ-plasm theory of August Weismann [89] which sparked a revolution in envisioning a differentiation process during ontogeny. In it, Weismann hypothesized that early embryogenesis was characterized by germ cells – carefully protected cells which

passed genetic information on to later offspring – and somatic cells – which make up the rest of the body. In this theory the former could change into the latter, but not vice versa, with such a conversion occurring during cell-division. This not only materialized the notion that hereditary information cannot (easily) be passed on by experience, effectively banishing ideas related to Lamarckian inheritance, as the germline cells would pass on genetic information uninfluenced by the somatic cells; it also seeded the more familiar concept of an omnipotent undifferentiated cell type which can provide differentiated offspring. Though Weismann himself never used the term stem cell the name was soon applied by others, and it was in the wake of this theory that the idea became popularized in the field of hematology at the turn of the century. The major cell components of the blood had already been identified, and the hypothesis that these could share a common ancestor was appealing – especially since it had already been deduced that they originated in the same tissues – however a debate endured between the so-called 'unitarians' and 'dualists', on whether there could be a single ancestor cell type for al hemocytes, or different stem cells for the lymphocyte and leukocyte lineages [89]. A prominent figure in the field was Arthur Pappenheim, a strong adherent of the unitarian hypothesis, conceptualized the stem cell as an embryonic multipotent cell, meaning it had the potential to differentiate into a variety of different specialized cell types. Working initially on blood formation in amphibians, Pappenheim became known for his adept use of cell trees – illustrations of the various (often hypothetical) stages of differentiation of a cell (Figure 2.1) – the widespread use of which marked a growing interest in uncovering hematopoietic lineages. This theoretical curiosity was somewhat relieved by the development of simple tissue culturing methods in the early 20th century, which could show the existence of cell differentiation in tissue growth. However, indisputable proof for the stem cell concept remained lacking, and it wasn't until after the second world war that another breakthrough occurred. In the post-war 1950's great efforts were being made in search of treatments for radiation sickness, which was found to significantly impact the lymph nodes and bone marrow. A promising avenue of investigation came in the form of transfusion and transplantation experiments performed on

mice, where it was shown that expected lethal irradiation doses could be mitigated by grafting hematopoietic tissue from healthy specimens [46]. Such insight led James Till and Ernst McCulloch in 1961 to the discovery that injecting an appropriate amount of marrow cells from a normal donor into an irradiated host could trigger the formation of small proliferative cell colonies in the spleen [131]. They later identified these as fully clonal outgrowths comprising many histologically recognizable differentiated cell types, each originating from a single injected cell [11]. Originally dubbed *colony-forming units*, these cells were found to have the capabilities of long-term self-renewal as well as multipotent differentiation – the two properties generally constituting the definition of a stem cell at the time – making it clear that these were indeed the elusive stem cells which had been hypothesized half a decade ago. Alongside their experimental findings Till and McCulloch also provided one of the first mathematical models of cell production in hematopoiesis [133], which consisted of a system of three differentiation compartments – corresponding to stem cells, early progenitor cells, and the fully differentiated end products – with cell migration through the compartments modeled as a birth-death process. Despite the fact that there was no truly effective method for identifying multipotent cells other than testing their colony forming ability, the following years showed impressive advances in the empirical mapping of hematopoietic lineages, combining such culturing techniques with studies using chromosomal markers as well as some cleverly designed genetic tracing (e.g. the observation of coexisting cell populations with differing alleles of the enzyme glucose-6-phosphate dehydrogenase) [132]. However, it was the development of surface marker detection methods – a technique used to identify cells by classifying the variety of proteins attached to their surface membrane – which proved to be invaluable for mapping lineage descent, and eventually led to the isolation and identification of single hematopoietic stem cells [124]. By the end of the 20th century, discoveries of numerous surface markers and increasingly detailed detection methods had given rise to a relatively modern map of hematopoiesis (Figure 2.3a), with lineage paths and branching points in the differentiation landscape that remain in use today. And yet, the development of transcriptome measurement technologies – providing insight into the

expression of a cell's genome – and their effective large-scale applications in the past decade have conferred a new perspective on the differentiation process, questioning the validity of the discrete progenitor model altogether, and introducing what may turn out to be a new paradigm of hematopoietic differentiation.

2.2. Cellular differentiation

Before entering into the specifics of hematopoiesis, it is useful to take a step back and briefly review what the process of differentiation exactly means from a modern point of view, as it forms the cornerstone of the hematopoietic system.

2.2.1. Providing variety and specification

While it is perhaps somewhat of a cliché to say that cells are "the building blocks of life", it is undeniably a fitting analogy for their role in the complex multicellular organisms that make up the biological world around us. Indeed, not only are all the plants and animals we see in our daily lives "merely" complex machines constructed from millions or billions or even trillions of cells acting in some form of unison, the countless microscopic organisms found in the sea and air and even our bodies – pretty much any corner of the earth – are either *made up of* or simply *are* cells. In fact, most definitions of *life* constitute the cell as its base unit [110]. Poignant as the "building block" metaphor may be, a cell is rather more complicated than the average Lego piece, particularly because it is itself a machine executing most of the processes generally associated with biological function. In the human body for example, the metabolic processes for the construction of amino acids and proteins, the production of ATP through the Krebs cycle, the execution of DNA replication and cell division, and countless other processes are all performed within the cells themselves. In fact, for many of the tasks that are associated with specific tissues or organs, it is the cells in their makeup that constitute the actual workforce. This observation carries a significant implication: different tissues serve different purposes and as such their constituent cells are required to perform widely varying tasks. For this

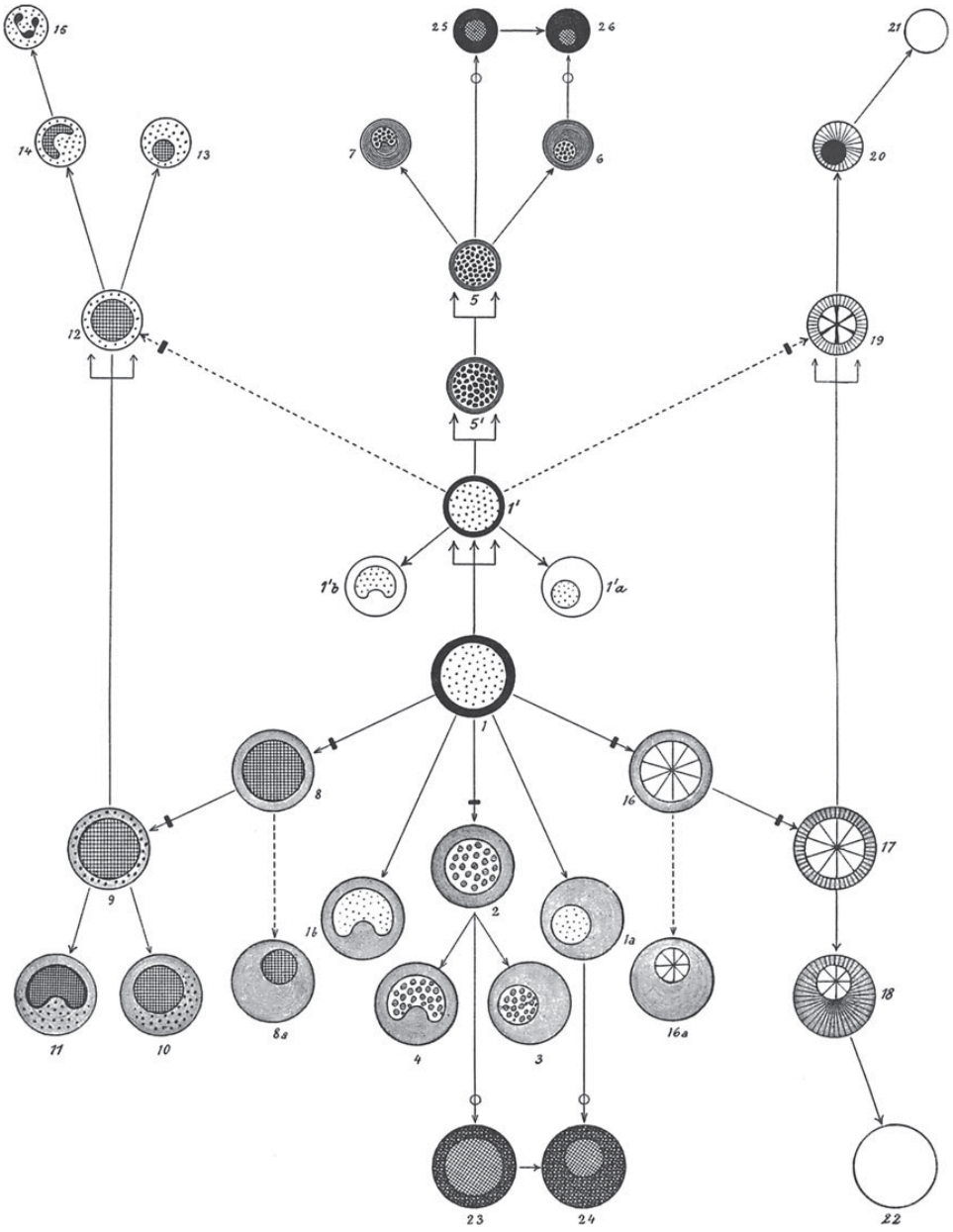


Figure 2.1.: *A cell tree of the hematopoietic system by Artur Pappenheim [108], with at the center the hematopoietic stem cell. Differentiation trajectories are depicted through the arrows connecting various (hypothetical) progenitor states, with the fully differentiated cells as the endpoints of each lineage.*

reason it is no surprise that upon observation, cells of a single organism are found to be differentially specialized to their particular function. This specialization presents itself in different ways, with variations found in shape and size, activity of metabolic processes, response to chemical stimulants, and even physical capabilities such as replication or self-driven motion. While this rich variation of cell types was originally described by visual characteristics under a microscope, it was later found that cells of a particular function present similar assortments of proteins externally attached to their cell membrane [9]. The classification of these has proven to be an effective method for identifying (and classifying) cell types, and has in this context led to them commonly being referred to as *surface markers*. But while the existence of different cell types within a single organism makes sense from a functional perspective, it simultaneously raises the question of how this variation can occur at all, given that we know the blueprint for all cells to be the same. In other words, how do such extreme variations between cells in the same organism arise, even though they all (with some exceptions) carry exact copies of that organism's DNA? The answer is surprisingly simple: different parts of the genome (the collection of all genes described by the DNA) are "activated" for different cell types. Thus each cell carries the instructions for all necessary functions in the body, however only certain parts of it are accessible – referred to as the genes which are *expressed* – specifically those related to the cell's current function. From this perspective the pattern of expression of a cell ultimately determines its type, meaning that measurements of this quantity (through techniques collectively referred to as *transcriptomics*) describe another valuable method for distinguishing cell types.

The question remains how such different expression profiles arise in the first place, given that patterns of gene expression are generally maintained by cells after a division, and that even complex organisms such as ourselves begin their lives as a single cell (the *zygote*). To obtain the rich variation of cells throughout different tissues and organs, a process must therefore exist which facilitates the progression (potentially facilitated by cell divisions) from one cell type to another. This process is called *differentiation*, and generally follows a one-directional hierarchical architecture (Figure 2.3a). At the

top is an undifferentiated cell – the so-called *stem cell* – which has the the potential to develop into a wide variety of cell types, perhaps even all of them (as must clearly be the case for the zygote), a characteristic described as *potency*. If it embarks on the path of differentiation, it progresses through a number of intermediate cell types – often termed progenitors – which no longer possess the complete potency of their original state, but still maintain a number of *cell fates* (i.e. specialized cell types) to choose from. Cells characterized by such a reduced potency are often referred to as being *pluripotent*, as opposed to their previously *totipotent* (or *omnipotent*) stem cell state. Finally, after a number of such branching choices have been made, the cell reaches a fully differentiated state, which is the functionally specialized cell type discussed earlier. This state (often referred to as the *mature* cell) is considered to be final in most human cells *in vivo*, though there are organisms in which the differentiation process has been shown to be reversible on a large scale, for example in plants or animals with regenerative properties [128]. The entire process of differentiation was famously illustrated (Figure 2.2) by C. H. Waddington [3] as a pebble rolling down a hill, whereby various grooves in the surface could cause it to end up in different locations at the bottom corresponding to different cell fates.

2.2.2. Transitional states and cell fate

While the differentiation process is clearly necessary during *ontogenic growth* (the development of an organism from the zygote to its adult state) it occurs in many systems of the body throughout adulthood as well, such as the colon [10], the skin [17], and perhaps most famously: the blood. As each of these ecosystems requires its own set of specialized cells, they present different trajectories of maturation. While mapping these so-called cell *lineages* forms an important part of understanding such systems, an equally pertinent question is what the mechanisms underlying these epigenetic changes are.

The oldest models of differentiation pictured various distinct cellular phenotypes – identified by their visually distinguishable characteristics – which maturing cells would

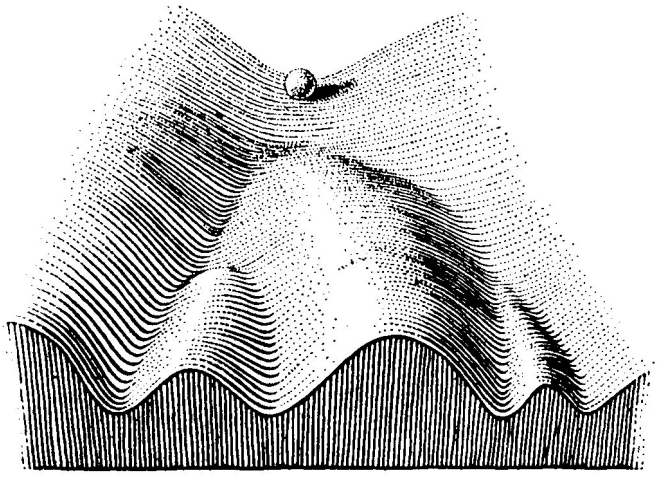


Figure 2.2.: C. H. Waddington's epigenetic landscape. *As a visual metaphor for the differentiation process, Waddington depicted a cell as a pebble being pulled by gravity down a hill. As it rolls downward, it encounters a complex topological structure with various bumps pushing it along different trajectories, ultimately leading to distinct locations at the foot of the hill, corresponding to distinct cell fates.*

transition into and out of in discrete steps (see for example Figure 2.1); however, more modern methods of surface marker detection and most notably genome expression measurements (transcriptomics) paint a more muddled picture. If the notion of discrete transition states were true, one might expect to observe distinct patterns of expressed genes shared by groups of cells, corresponding to their separate states. Instead, single cell transcriptomics have revealed high degrees of heterogeneity among cells, even those similarly classified by surface marker methods, implying that progression along a differentiation pathway may have a more continuous character [137, 100]; a concept which is difficult to reconcile with the idea of distinct states.

While this new paradigm is currently an important topic of debate [76], it does not answer the question of what occurs at branching points in a pathway, where a choice must be made between two ultimately different cell fates. In fact, the mechanisms which underlie this decision making and their emergent behavioral patterns form another subject of ongoing research [96]. It is well-known that cell signaling plays an important role in

this process [54, 113]. Depending on the external biochemical signals they receive, cells may initiate specific differentiation programs – sequentially performed epigenetic changes to their expression patterns – which are driven by so-called *gene regulatory networks*, the molecular processes underlying these alterations. The elucidation of such networks related to specific differentiation paths is an important branch of investigation[80], and the enforcement of particular cell fates through exposure to known signaling factors can be performed *in vitro* [113] for various differentiation pathways. It is however still unclear to what extent stochasticity plays a role in the decision process [84]. While the classical view following Waddington's interpretation (Figure 2.2) entails a smooth and (mostly) deterministic transition, a competing view suggests that the competition of multiple active gene regulation networks results in a type of transition state, whereby stochastic effects cause highly variable expression profiles [96].

2.2.3. The role of cell divisions in differentiation

Since its theoretical conception cellular differentiation has been associated with cell divisions [89], in the sense that (more) differentiated cells arise as daughter cells after a division. While morphological changes and commitment to a particular cell fate have indeed been associated with the cell cycle [47], the continuous landscape of transcriptional patterns discussed in the previous section complicates this picture, indicating that cells may continuously undergo development along certain lineages [137]. Nevertheless, the notion of differentiation accompanying divisions remains useful in describing progression through established progenitor states [95, 35], and is furthermore important in the behavior of stem cells. Since stem cell populations exist in the body throughout adult life, their own numbers must be replenished given their loss of cells to differentiation. The first mechanism to ensure this is *asymmetric* division, whereby a stem cell will divide with one of its daughter cells maintaining the stem cell phenotype while the other initiates differentiation [73, 62]. Alternatively the stem cell may divide *symmetrically* [47], in which case the daughter cells end up identical, both having entered differentiation or remained stem cells, with the latter process typically referred to as *self-renewal*.

2.3. Hematopoietic lineages

Hematopoiesis is perhaps the most (or certainly the longest) studied differentiation system in the body. As a result of this longstanding interest, its differentiational landscape – i.e. the various possible lineages along which maturing cells transition from stem cells to fully differentiated hemocytes – has been mapped out in extensive detail. Figure 2.3a shows the classical "branching tree" picture of hematopoiesis prevalent by the early 2000's, whereby the distinct states were categorized throughout the 80's and 90's by both surface marker classifications as well as *in vitro* experiments probing lineage potential [95]. The earliest branch following the stem cells occurs with the myeloid and lymphoid lineages, the latter providing the B- and T-cells of the adaptive immune system (which mature mostly in the lymph nodes), with the former branching further into macrophages, granulocytes, and erythrocytes. As discussed in the previous section, novel insights into the DNA expression profiles of cells at various points in the differentiation pathways have painted a somewhat different paradigm, as shown in Figure 2.3b, where the discrete states are replaced by a continuous shaping of the cells' development. Finally, experiments of the past decade have shown the traditional lineages to be less rigorous than initially assumed, with the occurrence of various alternative trajectories through the pictured developmental landscape [102, 76], leading to the notion of a more fluid formalism, as depicted in Figure 2.3c.

2.4. Hematopoietic stem cells

With a clear picture of the hematopoietic differentiation landscape in mind, we may now turn our attention to the root of this tree: the hematopoietic stem cells (HSCs). As has already been touched upon in Sections 2.1 and 2.2, these cells play an incredibly important role in bodily functioning, both during development and throughout adulthood, and are to this end endowed with some very unique capabilities. Having already covered a historical perspective of the stem cell concept, let us here consider a more modern definition.

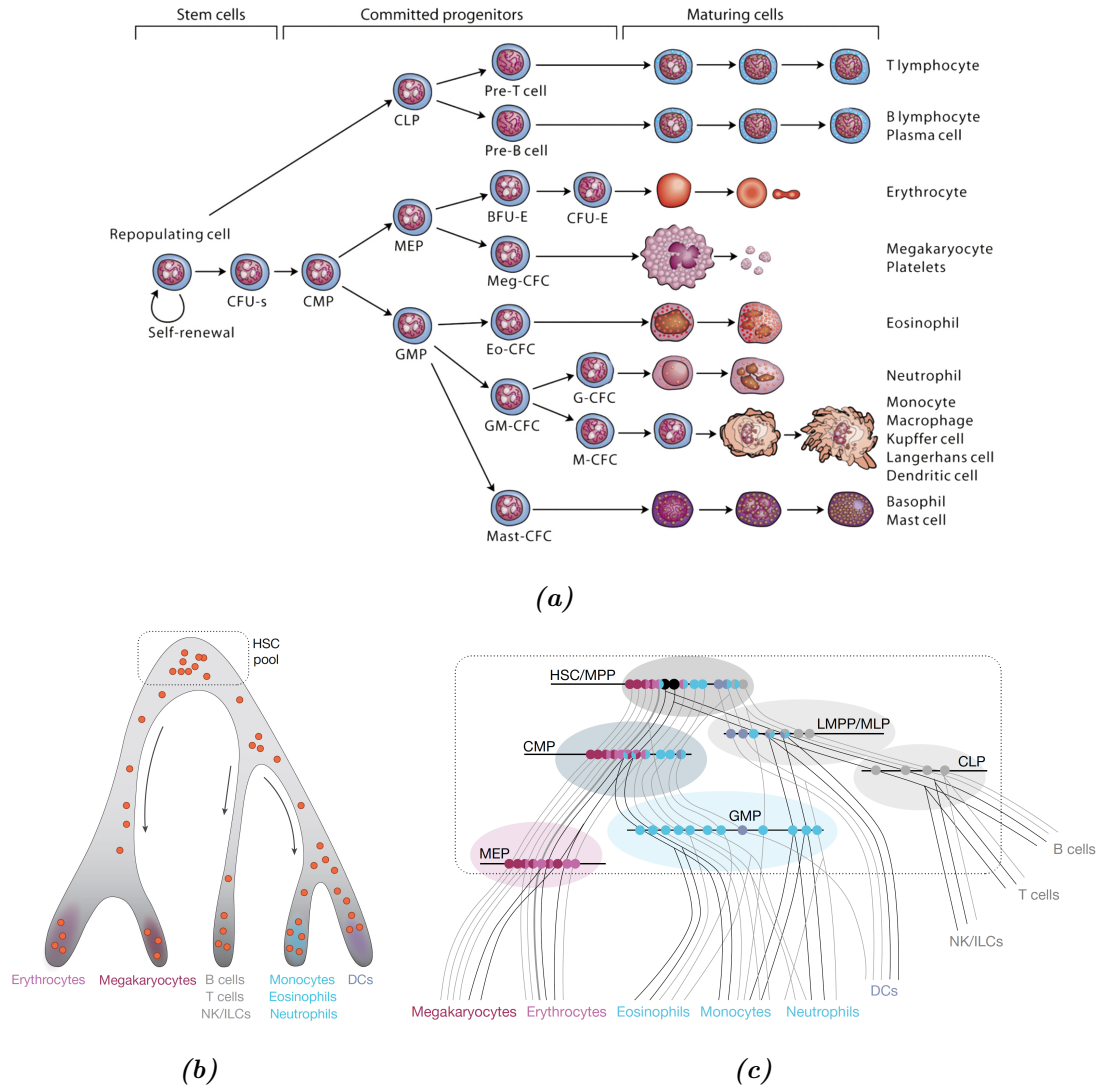


Figure 2.3.: Visualization of the hematopoietic lineages under different paradigms. *(a) Branching tree structure obtained by the early 2000's summarizing the different lineages and progenitor states identified over the previous decades. Reproduced from [95]. (b) Alternate depiction of the hematopoietic lineages under the continuous differentiation paradigm. Reproduced from [76]. (c) Depiction of the system as a fluid trajectory structure, as suggested by Laurenti and Göttgens [76]. Reproduced from [76].*

Stem cells are in their basest form defined by two characteristics: the ability to self-renew – i.e. to maintain their stem cell state after a division – and their potency (Section 2.2) – i.e. the variety of differentiational fates available to them. As previously mentioned, numerous differentiation hierarchies exist in the body which oversee the maintenance of different tissues, each thus having an associated stem cell type whose potency encompasses its constituent cells. From this perspective, the stem cell is a cell with the ability to – by itself – reconstruct the architecture of its associated tissue, such as for example the colony forming units discovered by Till and McCulloch [131] which could successfully reconstitute the bone marrow. This definition is somewhat complicated by an existing heterogeneity amongst such cells, where the larger fraction fails to maintain such a colony indefinitely [42]. Such cells are dubbed *short term* stem cells (ST-HSCs), as opposed to the long term (LT-)HSCs which can sustain reconstitution for at least a lifetime, though it appears even these do not possess infinite divisional potential [95].

Extensive study of such LT-HSCs has found them to be not only extremely rare, but also highly passive in terms of divisions, which much of their number residing in a *quiescent* state outside of the cell cycle [25]. However, as opposed to senescence – a state in which a the cell cycle can no longer be entered – quiescence appears to be reversible in response to normal physiological stimuli. The functional relevance of this phenomenon remains debated; originally assumed to be a reaction to adverse circumstances such as nutrient depletion [25], it has alternatively been proposed that quiescent cells aid in increasing the longevity of stem cell niches by serving as replacements for damaged active HSCs [81], however this question has ultimately remained unanswered.

2.5. Differentiational tissues accumulate mutations

It was previously mentioned that hematopoietic cells – while potentially differing in their expression – still share identical copies of the underlying genome, however, this is not entirely true. The mechanisms by which the DNA is copied and distributed before and during mitosis (cell division) are imperfect, and occasionally errors will occur resulting in slight alterations of the genome of a daughter cell. While such mutations are rare –

typically estimated on the order of 10^{-8} to 10^{-10} times per single basepair [74, 144, 87, 20] – given that there are 6×10^9 basepairs one would expect there to be on average at least one up to a handful of mutations to occur per cell division. This is furthermore enhanced by the fact that external effects such as radiation and even oxidation [5] can cause DNA damage as well. While mutations – sometimes also referred to as *variants*, as they constitute a variant *allele* for the gene in which they reside – in the germline are famously the drivers of evolution, in somatic tissue (cells which are not part of the germline) the notion of a selective advantage in a cell compared to its peers does not carry the same implication, and it is instead associated with undesirable consequences such as cancer [53]. Still, it appears that most mutations are neutral – as is similarly the case on the scale of the organism – meaning that they will not affect the functionality of a cell in a significant way. In fact, detrimental mutations are likely to result in a cell's premature expiration if they inhibit a cell's normal functioning, while neutral mutations will persist and be passed on to future offspring. The result is that cells in the body accumulate mutations over time [74], though the rate at which this mutational burden (the total number of mutations in a cell or cell population) grows generally varies for different tissues [93, 92].

As mutations are acquired by cells, their persistence in the population becomes a highly stochastic process, given that their extinction or perseverance is entirely dependent on the offspring of their carriers. This can result in interesting patchworks of mutations, which may present unique properties in hierarchical systems such as hematopoiesis. In the past decade, thanks to advanced sequencing techniques it has become possible for such mutational landscapes to be observed [93, 92, 77], which presents an opportunity for new forms of statistical and mathematical analyses.

2.6. Open questions and perspectives

Although current knowledge of the hematopoietic system is extensive, there are clearly numerous open questions which remain debated or even unaddressed. We have already touched upon the the current ambiguity surrounding the nature of differentiational states

2.2.2: which model of cell development and lineage commitment best describes the differentiation process will likely become more clear in the coming decade. Similarly, the functional characteristics of HSCs continue to be mapped, and it remains to be seen whether the current model of their behavior persists.

Importantly, while our qualitative understanding of the hematopoietic process – i.e. mappings of the various lineages and the mechanistic underpinnings of their existence – has made great strides in the past decades, a quantitative picture of the cell dynamics lags behind, perhaps in part due to the extreme difficulty of observations *in vivo*. Indeed, a decade ago most direct methods of investigation required experimentation *in vitro* or invasive procedures *in vivo*, with no real observations of HSC or progenitor behavior under normal hematopoiesis. Furthermore, even equipped with basic presumptions concerning the dynamics of marrow cells *in vivo*, there was little clear consensus on numerical values (or even orders of magnitude) for even the simplest quantities involved in this picture; such as the total number of active and dormant HSCs, their symmetric and asymmetric division rates during both normal and perturbed hematopoiesis, and even their relative contributions to the overall production of mature hemocytes. Fortunately more recent developments have begun to address these questions. An ongoing boom in experimental advancements has provided exciting opportunities for obtaining data closer to the quantities of interest, with two methods in particular showing promise in paving the way for obtaining quantitative data of unperturbed hematopoiesis *in vivo*. The first is unsurprisingly the advancement in high coverage deep genome sequencing techniques [117], which can be used to quantify the number and distribution of somatic mutations with a group of cells; the second is the development of heritable genetic markers which can be used to trace clonal descent, thus depicting the *in vivo* dynamics of maturing blood cells [129, 23]. While such methods provide a wealth of novel data, the interpretation of their results is not trivial, and calls for the application of expressly devised statistical methods and mathematical models.

3. Mathematical tools

They're drills that shoot lasers. They're totally believable and cool.

— Stan Marsh/Toolshed, *South Park*

The incredible complexity of the systems we wish to study will require the use of abstractions of various underlying processes. For example, instead of directly modeling mechanistic processes such as the cells' differentiation cycle or their complex interactions with hormones and growth factors, we take them as statistical black boxes that output probabilistic quantities. In this chapter we therefore first introduce the stochastic concepts and models which are applied throughout this part.

3.1. Stochastic processes

A modern treatment of probability theory involves the careful introduction of a *probability space* equipped with a σ-algebra and an accompanying *probability measure*. This level of rigor is in truth not required for the concepts developed here, and may ultimately distract from the applied nature of this thesis. We will therefore take a less specialized approach, generally assuming that the random variables and functions we introduce can be more rigorously defined if so desired. In this manner the prerequisite knowledge for this chapter is kept at a minimum, though it is assumed the reader has a basic understanding of the concepts of set theory, random variables, and the rules of probability. For a more expansive treatment of the processes discusses here the interested reader is referred to the classic book by Feller [44], whereas for more rigorous derivations most modern introductory texts suffice, such as for example [45].

3.1.1. The Bernoulli process

The simplest stochastic process to which we will appeal in this thesis – and from which all the following processes can be conveniently derived – is the Bernoulli process. It is in essence the mathematical formulation of successively performed coin tosses, with the added bonus that the coin is allowed to be unfair – i.e. preferring either heads or tail – as long as the unfairness is consistent with each toss. Stating this in formal terms, the Bernoulli process describes *the independent repetition of an experiment which has exactly two outcomes, each with the same fixed probability of occurring at each repetition.* Denoting the possible outcomes of a single experiment (typically called a Bernoulli trial) as *success* (S) and *failure* (F), the result can be visualized as a sequence of length n (the number of trials performed) with each element in the sequence being one of the two possible outcomes, as shown in Figure 3.1. Given that the Bernoulli trial has only two possible outcomes, its probability distribution can be written concisely as $\mathbb{P}\{S\} = p$ and $\mathbb{P}\{F\} = 1 - p$, with p thus being the probability of a success. Perhaps the most interesting quantity related to this process is the probability of k successes occurring after n repetitions. In order to obtain this, one can take all possible sequences of n trials in which there are k successes, and sum over their respective probabilities. Since each "correct" sequence has the same probability $p^k(1-p)^{n-k}$, the resulting distribution is given by

$$\mathbb{P}\Big\{k \text{ times } S \mid n \text{ trials}, \mathbb{P}\{S\} = p\Big\} \equiv B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

The distribution $B(k; n, p)$ associated with the probabilities of all possible k is generally known as the *binomial distribution*. Its mean and variance can be shown to be

$$\mathrm{E}[k] = np \tag{3.2}$$

$$\mathrm{Var}[k] = np(1-p) \tag{3.3}$$

Another important result of this description is the probability of obtaining k successes or failures in a row, which we can see to be p^k and $(1-p)^k$ respectively.

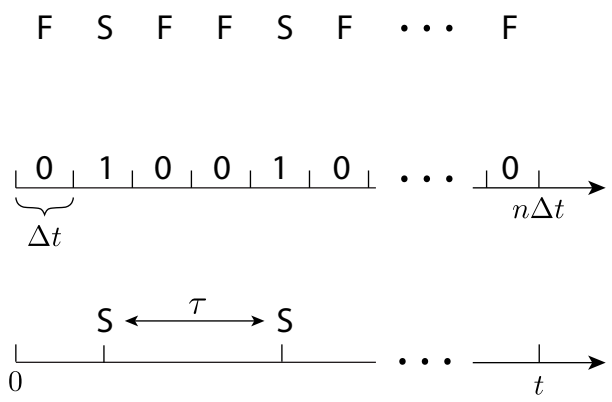


Figure 3.1.: From Bernoulli to Poisson processes. *A sequence of Bernoulli trials (top) can also be applied to the occurrence of a stochastic event in discrete time (middle). In the limit of infinitesimal time steps the Poisson process is obtained (bottom).*

Timed process on a grid

While the picture of repeatedly tossing a coin is a useful way to introduce the Bernoulli process, let us look at another application that will prove useful in the following sections. Envision a system in which a particular event can occur at any point in time, lets say for example a coffee being prepared in the office. We might be interested in constructing a model for the likelihood of coffee's being prepared at various points in time, or the total number of coffees prepared in a day. If we divide the time of a workday into finite increments – for example every half an hour – and assume that the probability of a coffee being prepared during each time increment is always the same, then this becomes an application of the Bernoulli process. Indeed, a coffee being made in a time increment can be interpreted as a success and an unused coffee machine as a failure (Figure 3.1), so that the total number of coffees made is given by the above derived binomial distribution – k being the number of coffees and n the number of time increments in a day. Unfortunately, a problem might arise if it is possible for two coffees be made in a single increment. To avoid this we can imagine taking this time step small enough – the time it takes to prepare a single coffee for example – so that this possibility disappears.

Besides the question of how many coffees are made in a day, we might also be interested in how much time passes in between preparations – typically referred to as the *waiting time* or *inter-arrival times*. If a coffee was just prepared in the previous time step, we can think of this waiting time τ as given by the number k of consecutive failures occurring in the next times, up until the first success. With Δt the length of the time step, the waiting time is then given by $\tau = k \times \Delta t$. We can thus write for the distribution of τ:

$$\mathbb{P}\{\tau = k\Delta t\} = p(1-p)^k \tag{3.4}$$

Interestingly, we can see that this result holds true for any point in time, irrespective of whether a coffee was just made or not. This follows from our very first assumption: that all coin tosses occur independently with the same probabilities of success and failure. In this context of successive steps in time such a process is referred to as having *independent increments*. This leads in our case to a very important property of the system known as *memorylessness* or the *Markov* property, which states that the system's past history – i.e. anything that might have occurred previously – cannot influence possible futures any more than its current state. Or in other terms: if we know the current state of the system, we have all possible information about its (probabilistic) future, and no additional information about the past can improve upon this.

3.1.2. The Poisson process

While modeling the probabilistic occurrence of specific events in time as a Bernoulli process certainly works (to an extent), it is rarely done, for the simple reason that there is a much better model for this particular type of problem – one that manages to sidestep the issue we encountered related to multiple events occurring in a time step: the Poisson process. Concretely, it describes the incidence of stochastically occurring events in *continuous time* (Figure 3.1), rather than on some discrete grid, while maintaining the property of "independence of tosses" (we will see later more specifically what this means in the context of continuous time).

To arrive at the Poisson process, let us first return to our example of coffee being prepared at random times in the office. Whereas in the previous section we modeled the problem by considering a discrete grid of time points, we can attempt to improve on this model by moving to a description that is continuous in time. In fact, we have already hinted at an approach that can be taken to this end: In order to solve the issue of potentially multiple events occurring in a single time step, we noted that reducing the length of this increment reduces the likelihood of such an occurrence. While this may work fine for our example of preparing a coffee – we could conveniently take the time step as the time it takes for a single preparation – it is not such an elegant solution in general; some applications might require an immense number of discretizations, which would make the binomial distribution computationally difficult to solve, or even worse, the system of interest may allow for events to occur simultaneously. On the other hand, it turns out that taking an *infinitesimally* small time step – and thus an infinite number of discretizations – results in a much simpler description.

It is worth nothing that while we will derive the distributions related to the Poisson process by taking the limit of the discrete time picture, they can also be found through other methods which do not require the existence of the various limits performed. In this sense our current approach might be considered somewhat less rigorous, however it provides – in my opinion – a clear understanding of what it means to model something as a Poisson process, and highlights the assumptions being made when applying it to real systems.

The probability rate

In the previous section, when dividing the time axes into discrete steps we made use of the fact that in each time increment there is a finite probability p for an event to occur. However, what was somewhat glossed over is the notion that this probability depends on the size of the chosen time step. Indeed, the likelihood of a coffee being made in a time frame of 2 minutes is of course much smaller than in a span of 2 hours. For the moment, we will remain agnostic as to the exact form of this dependence $p(\Delta t)$, and

instead make use of our intuition that the probability decreases for smaller time steps, which tells us that $p(\Delta t)$ becomes infinitely small as Δt goes to 0. Thus in light of our desired move to a continuous time frame, we introduce the *probability rate* λ

$$\lambda = \lim_{\Delta t \to 0} \frac{p(\Delta t)}{\Delta t} \tag{3.5}$$

which is – as opposed to the time step and its associated success probability – a finite quantity; though we will at this point simply assume this limit exists (for a more rigorous treatment the reader is referred to for example [45]).

The exponential distribution

Recall that in the discrete picture we found the distribution of waiting times by first considering the probability of k consecutive failures – given by $[1 - p(\Delta t)]^k$ – which amounts to a time $t_f = k \times \Delta t$ wherein not a single event occurs. We can look at what this quantity becomes in the continuous picture by writing $k = t_f/\Delta t$ and taking the limit $\Delta t \to 0$:

$$\mathbb{P}\{\text{no } S \text{ in } t_f\} = \lim_{\Delta t \to 0} \left[1 - p(\Delta t)\right]^{t_f/\Delta t} \tag{3.6}$$

Noting that in this limit we can rewrite the now infinitesimal probability per time step as $p(\Delta t) = \lambda \Delta t$, (3.6) can be easily simplified by looking at the logarithm:

$$\log \mathbb{P}\{\text{no } S \text{ in } t_f\} = \lim_{\Delta t \to 0} \frac{t_f}{\Delta t} \log\left[1 - \lambda \Delta t\right] \tag{3.7}$$

$$= \lim_{\Delta t \to 0} \frac{t_f}{\Delta t} \left(-\lambda \Delta t - \frac{1}{2}(\lambda \Delta t)^2 + \dots \right) \tag{3.8}$$

so that we finally obtain

$$\mathbb{P}\{\text{no } S \text{ in } t_f\} = e^{-\lambda t_f} \tag{3.9}$$

Since this quantity represents the probability of an event *not* having occurred in the time t_f, it can be interpreted as the *right tail distribution* of the waiting time, or in other words, $1 - \mathbb{P}\{\text{no } S \text{ in } t\}$ gives the probability of the next event having occurred in t:

$$\mathbb{P}\{\tau < t\} = 1 - e^{-\lambda t} = F(t) \tag{3.10}$$

This is the cumulative distribution function $F(t)$ of the waiting time τ: as t increases so does the likelihood of the next event having occurred. In fact, this is actually the concrete dependence $p(\Delta t) = F(\Delta t)$ which we could not specify earlier when defining the probability rate in (3.5), and it is now easy to check that this limit is indeed true.

Because we are working in a continuous-time picture, there is no finite probability associated with the waiting time having the exact value $\tau = t$ such as (3.4) for the discrete picture. However, we can look at the waiting time's probability density function $f(t)$ – which can be integrated over an interval $[t_0, t_1]$ to obtain the probability $\mathbb{P}\{\tau \in [t_0, t_1]\}$ – by taking the derivative of $F(t)$:

$$f(t) = \lambda e^{-\lambda t} \tag{3.11}$$

This is the exponential distribution, which to an extent forms the basis of countless probabilistic models in biology. We can interpret it as the continuous-time analogue of (3.4), where (with some abuse of notation) $\mathbb{P}\{\tau = t\} = f(t)dt$. It is clear from how we arrived at it that the only assumptions made are those of a constant probability rate – equivalent to the fixed probability of a success in the coin toss – and independent increments – though now these increments are infinitesimally small.

It is fairly simple to show through integration that the mean and variance are given by

$$\mathrm{E}[\tau] = \frac{1}{\lambda} \tag{3.12}$$

$$\mathrm{Var}[\tau] = \frac{1}{\lambda^2} \tag{3.13}$$

The Poisson distribution

Now that we have a continuous analogue of the distribution of waiting times, we might wonder if we can find a similar counterpart to the binomial distribution (3.1), which would then assign probabilities to the number of events occurring in a chosen span of time. It turns out there is, and we will derive it in a similar manner as we did the

exponential distribution: by taking the limit of infinitesimal time steps.

In a discrete picture with time step Δt, we have that $B(k;n,p)$ provides the probability for k occurrences in a time $t = n\Delta t$. First we note that from (3.1), dividing the probabilities of successive k's provides a recursive method for calculating the binomial distribution:

$$B(k;n,p) = \frac{np-(k-1)p}{k(1-p)} B(k-1;n,p) \tag{3.14}$$

Thus we might investigate the limit of infinitesimal time steps and see if a pattern emerges. Now, when taking the limit $\Delta t \to 0$ we have that simultaneously $n \to \infty$, however their product remains fixed: $\lim(n\Delta t) = t$. Given that in the same limit $p(\Delta t) = \lambda \Delta t$ (3.5), we have $np = \lambda n \Delta t = \lambda t$. Thus we may write

$$\begin{aligned}\lim_{\Delta t\to 0} B(k;n,p) &= \lim_{\Delta t\to 0}\left(\frac{\lambda t-(k-1)\lambda\Delta t}{k(1-\lambda\Delta t)} B(k-1;n,p)\right) \\ &= \frac{\lambda t}{k}\lim_{\Delta t\to 0} B(k-1;n,p)\end{aligned}$$

Finally, we note that we have already found the first probability in this sequence for $k = 0$, as this is the probability of no event occurring in t, given by (3.9):

$$\lim_{\Delta t\to 0} B(0;n,p) = e^{-\lambda t} \tag{3.15}$$

From this we can successively construct

$$\begin{aligned}\lim_{\Delta t\to 0} B(1;n,p) &= \lambda t e^{-\lambda t} \\ \lim_{\Delta t\to 0} B(2;n,p) &= \frac{(\lambda t)^2}{2} e^{-\lambda t} \\ &\vdots\end{aligned}$$

Introducing the notation $P(k;\lambda) = \lim_{\Delta t\to 0} B(k;n,p)$, we see by induction that

$$P(k;\lambda) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{3.16}$$

which is the famously known *Poisson distribution* $P(k;\lambda)$, describing the probability of k events occurring in a time t, and constituting our continuous-time analogue of the

binomial distribution. The mean and variance can be shown to be

$$\mathrm{E}[k] = \lambda t \tag{3.17}$$

$$\mathrm{Var}[k] = \lambda t \tag{3.18}$$

This is an interesting result. Recall that we introduced λ as the probability rate of an event occurring. We now find that for independent increments it can be also interpreted as the average number of events occurring in a unit of time.

Combining Poisson processes

A useful property is that it turns out to be very easy to combine separate independent Poisson processes into a single description. Returning to our example of coffees in the office, we might also be interested in how often a tea is prepared. Assuming these events occur entirely independently from the coffees – perhaps the coffee drinkers and the tea drinkers are two separate crowds – it is straightforward to construct a process for the occurrence of coffee *or* tea being prepared. In particular, given two independent Poisson processes with events P and R and rates φ and ρ, the joint stochastic process for either of their respective events occurring is again a Poisson process with rate $\lambda = \varphi + \rho$. This can be seen from moving to the rate: The probability of either P or R occurring in a time step dt is $\mathbb{P}_{dt}\{P \cup R\} = \mathbb{P}_{dt}\{P\} + \mathbb{P}_{dt}\{R\} - \mathbb{P}_{dt}\{P \cap R\}$; using the independence of the two processes to write $\mathbb{P}_{dt}\{P \cap R\} = \mathbb{P}_{dt}\{P\}\mathbb{P}_{dt}\{R\}$ and taking the rate of this as in (3.5) this becomes

$$\begin{aligned}
\lambda &= \lim_{dt\to 0} \frac{\mathbb{P}_{dt}\{P\} + \mathbb{P}_{dt}\{R\} - \mathbb{P}_{dt}\{P\}\mathbb{P}_{dt}\{R\}}{dt} \\
&= \lim_{dt\to 0} \frac{\mathbb{P}_{dt}\{P\}}{dt} + \lim_{dt\to 0} \frac{\mathbb{P}_{dt}\{R\}}{dt} \\
&= \varphi + \rho
\end{aligned} \tag{3.19}$$

where the joint term $\mathbb{P}_{dt}\{P\}\mathbb{P}_{dt}\{R\}$ goes to zero in the limit as it is of order dt^2. It is clear that the same notion holds for more than two processes, so that for any number of independent Poisson processes with rates φ_i the process which describes any of their occurrences is again a Poisson process with rate $\lambda = \sum_i \varphi_i$. It can furthermore be

shown that for such a joint process with events P_i, given the occurrence of any event, the probability of it being a specific P_j is $\varphi_j / \sum_i \varphi_i$. A proof for this is found Appendix A.1.

Poisson everything?

The Poisson process is a powerful tool for modeling stochastic processes, used ubiquitously in models across many domains. It's core assumption however – namely that the process is memoryless – can be quite limiting in practice. For example, in the office coffee system memorylessness implies that the probability of a preparation is always constant, even immediately after the machine has been used. But some information about human behavior can tell us this is not entirely true: someone who has just made a coffee is – barring a catastrophic spilling accident – not likely to make another immediately after. In fact, some individuals may have a more regular schedule, where their desire for coffee abates following a consumption and only returns after some amount of time. Specifically, this means that the success probability $p(t)$ from which we derived the rate in (3.5) depends on what occurred in the past, and the process is therefore no longer memoryless. In fact, we will encounter this exact problem later when discussing the occurrence of cell divisions in a population. On the other hand, in the absence of specific information related to the system's memory it remains incredibly useful to apply the Poisson model, in large part due to the fact that its single parameter, the rate, is typically an easily measurable quantity in a system – i.e. the average number of occurrences in a unit time. In this sense memorylessness or exponentially distributed waiting times forms a nice "first guess" to construct a model, though its limitations must be kept in mind.

3.2. Markov Chains

While the Bernoulli and Poisson processes described above allowed us to derive the powerful probability distributions which provide "perfect information" – in the sense that they contain the probabilities for any possible outcome of a random variable – the

system of interest may in many cases be too complex to directly obtain a function for the probabilities of all possible outcomes. In this section we will examine a more general method for obtaining such probabilities nonetheless, provided we have some knowledge of how the system can change over time, and we can identify some form of memorylessness to exploit.

3.2.1. The state space

We have already used the term "state" somewhat haphazardly throughout the previous sections, however at this point it is useful to formalize what is exactly meant by it. When referring to the state of a system, we are actually referring to the result of a particular observation of that system. For example, when tossing a coin one might go to observe which side of the coin is facing up; in this context there are only two states in which the system could possibly be: heads or tails. On the other hand, after tossing the coin five times, one could have also counted how many times the coin landed on heads; this is a different question, and has a different set of states associated with it, namely the integers from 0 to 5. Importantly, the possible states of a system can be visualized as a discrete set of elements, whereby at any given point in time an observation of the system would return one such element, as shown in Figure 3.2. Thus we can in general define a state space to be the set of all possible states – i.e. all possible values *of a particular observation* – of the system. Formally, we might say the states S_i are elements of the space $\mathcal{S}$, with $i \in 1, \ldots, N$ the number of all possible states in the system. Now envision a system with state space $\mathcal{S}$, that starts in some state $S_0 \in \mathcal{S}$, and can move to different states as times passes; though since we are working in a stochastic picture, there is no certainty as to what state the system will be in at a future time. This notion of the system moving probabilistically through possible states is known as a *Markov chain*. Although the state of the system at a later time is uncertain, we might still construct a probability distribution over all possible states at a given time t, though to do this we must furthermore require that these states be disjunct – i.e. the system can only be in

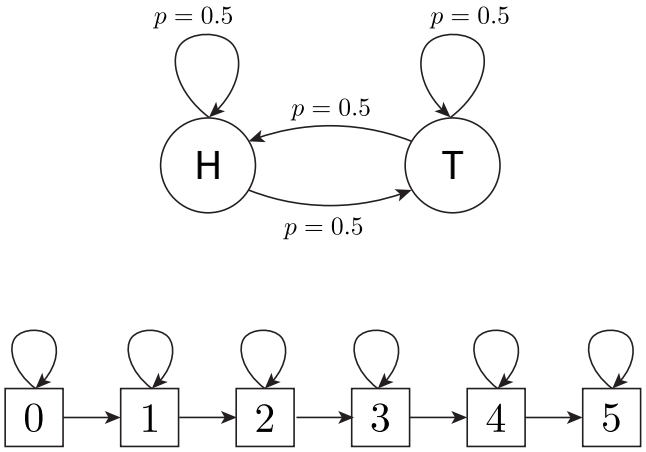


Figure 3.2.: *A Markov chain description for the repeated coin toss experiment. One possible state space describes the result of the most recent toss (top): the current state of the system can be heads (H) or tails (T), while after each toss the system transitions (arrows) to either the other state with probability* $p = 0.5$ *or the same state with probability* $p = 0.5$*. Another state space might be the number of heads that occurred after repeated tosses (bottom): the current state is then an integer, and with each toss either increases by 1 or remains the same.*

one state at a time – so that their probability distribution is correctly normalized:

$$\sum_{S_i \in \mathcal{S}} \mathbb{P}\{S_i \text{ at } t\} = 1 \tag{3.20}$$

3.2.2. Markov transition probabilties

If the system's evolution in time is memoryless – i.e. the probabilities of being in future states only depend on the current state – we can apply a powerful trick, which involves writing the probability of the system being in a specific state at a future time t through the probability of its transition via another state at an intermediate time $t' < t$:

$$\mathbb{P}\{S_i \text{ at } t \mid S_0 \text{ at } t_0\} = \sum_{S_j \in \mathcal{S}} \mathbb{P}\big\{S_i \text{ at } t \mid S_j \text{ at } t'\big\}\mathbb{P}\big\{S_j \text{ at } t' \mid S_0 \text{ at } t_0\big\} \tag{3.21}$$

This is actually a (Markovian) version of an important identity in probability theory known as the *Chapman-Kolmogorov* equation, where $\mathbb{P}\{S_i \text{ at } t \mid S_j \text{ at } t'\}$ is the probability of the system being in state S_i at time t, given that it is in state S_j at time t'. At first glance this statement may not seem all that useful, however we will see that for most systems we are interested in, short time transition probabilities are the quantities we can deduce from first principles, which we will in turn use to find the state space probability distribution at a later time. Note that it is the memoryless (or Markovian) property which allows us to assume such a unique *transition probability* exists in the first place: if this were not fulfilled the state of the system at times before t' could influence this probability, and thus we would not be able to know its value without knowledge of this history.

3.2.3. Discrete time Markov chains

Let us, as before, first look at the time evolution in a discrete picture, which could be the passing of time in finite increments Δt, or simply the state of the system changing due to discrete events, such as for example the tossing of a coin (Section 3.1.1). A realization of the system can thus be seen as a sequence of states S_t, each existing in $\mathcal{S}$. If we can somehow figure out the transition probabilities related to a single step $p_{j,i;\Delta t} = \mathbb{P}\{S_i \text{ at } t + \Delta t \mid S_j \text{ at } t\}$, given a starting state S_0 we can find the probability of being in a particular state far away in the sequence by repeated calculation of

$$\mathbb{P}\{S_i \text{ at } t + \Delta t \mid S_0 \text{ at } t_0\} = \sum_j p_{j,i;\Delta t} \mathbb{P}\{S_j \text{ at } t \mid S_0 \text{ at } t_0\} \tag{3.22}$$

One might colloquially say we are *evolving* the probability distribution of the state space over time, by first using $\mathbb{P}\{S_0 \text{ at } t_0\} = 1$ to find $\mathbb{P}\{S_i \text{ at } t + \Delta t\}$, which we use to find $\mathbb{P}\{S_i \text{ at } t + 2\Delta t\}$, and so forth. This will prove to be a powerful tool in finding the dynamics of a system, since even if we cannot obtain a concrete function for the probability distribution of states at arbitrary times – such as those we found for the Bernoulli, exponential, and Poisson distributions – we can still compute them iteratively.

3.2.4. Continuous time Markov chains

We are of course also interested in the continuous-time picture, where the results are no longer influenced by the size of the chosen time step. By now we know this can be achieved by carefully taking the limit of the time step $\Delta t \to 0$. Thus from the single step transition probability $p_{i,j;\Delta t}$ we introduce the *transition rate* in the same manner as before (Section 3.1.2):

$$p_{i,j} = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{S_j \text{ at } t + \Delta t \mid S_i \text{ at } t\}}{\Delta t}$$

Taking a simplified notation for the state probabilities $P_i(t) = \mathbb{P}\{S_i \text{ at } t \mid S_0 \text{ at } t_0\}$ we first rewrite (3.22) as a difference, and then take the limit:

$$\lim_{\Delta t \to 0} \frac{P_i(t + \Delta t) - P_i(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\sum_{j \neq i} p_{j,i;\Delta t} P_j(t) - (1 - p_{i,i;\Delta t}) P_i(t) \right] \tag{3.23}$$

Furthermore noting that we may write

$$1 - p_{i,i;\Delta t} = \sum_{j \neq i} p_{i,j;\Delta t}$$

since the probability of not transitioning to the same state $S_i \to S_i$ must be equal to the sum over all probabilities of transitioning to other states $S_i \to S_j \neq S_i$ (this follows from 3.20), we obtain in the limit

$$\frac{\mathrm{d}P_i(t)}{\mathrm{d}t} = \sum_{j \neq i} \Big[p_{j,i} P_j(t) - p_{i,j} P_i(t) \Big] \tag{3.24}$$

This is known as a *master equation*, and provides us with a recipe for evolving the probability distribution of the state space in continuous time.

3.2.5. Non-discrete state spaces

So far we have assumed the state space to be discrete, in other words the collection of states forms a countable set, and we have somewhat carefully chosen our examples to fit this assumption. But it is easy to imagine a system where this is not the case. Take for example a Poisson process where we are interested in measuring the time that has

elapsed after a specified number of events occurring. In this picture the system evolves discretely – imagine a simple counter which increases by 1 every time an event occurs – but our states are found from summing the waiting times τ, which are in $\mathbb{R}$ so that the space of all possible states becomes $\mathbb{R}$, which is an uncountable set. In general we cannot define a probability distribution on such a set (there would be no way of taking the sum in (3.20)), though for some spaces we could construct a probability density, as discussed previously. However, generalizing the master equation from the discrete to a continuous state space – similar to how we moved from discrete to continuous time – is more complicated without additional knowledge of the space's properties. We will on the other hand see a particular example of this later, specifically in the *Fokker-Planck equation* [112].

3.3. Stochastic population dynamics with Markov chains

Now that we have covered some basic mathematical tools for constructing and analysing stochastic systems, we will briefly look at two basic models used to describe the stochastic dynamics of populations, which rely on the probabilistic concepts we have introduced in the previous section. Their introduction forms the basis of the modeling done in Chapters 4, 5, and 6, and thus a careful overview of their properties and assumptions will benefit us later. For a more in depth discussion of these models the reader is referred to [44] (for the birth-death process) and [43] (for the Moran model).

3.3.1. The birth-death process

Envision a population of individuals with the ability to reproduce and to die, and imagine that we are interested in characterizing how the total size of the population changes over time. This description is general enough that it could apply to many real world systems, one particular example which we might be interested in being a population of cells: a cell dividing causes a new cell to be added to the population, while a cell dying causes it to be removed. If there is some stochasticity involved – i.e. individual cells

divide and die randomly in time – the size of the population at a future time cannot be known, though given knowledge of the exact stochastic processes involved, its probability distribution might be obtained. Thus, constructing a model for this system amounts to choosing a stochastic process for the occurrence of births and deaths by individuals. The simplest choice has only two assumptions: that the probability of an event (birth or death) occurring for a single individual is the same at any point in time, and that each individual acts independently of the others – i.e. the probability of a single individual reproducing or dying is not influenced by the death or reproduction of other individuals. Mathematically, the former we know to be characterized by independent increments, which we have seen to be the Poisson process, and the latter implies we may combine the separate processes (see Section 3.1.2) for each individual in the manner discussed earlier. Concretely, for every individual we have a probability rate β for it reproducing – known as the birth rate – and a rate δ for it dying – known as the death rate – so that in each infinitesimal time step $dt = \lim \Delta t \to 0$ there is a probability βdt of it reproducing and δdt of it dying, just as we showed in Section 3.1.2. Now consider the entire population; because separate individuals act independently we may take all of their reproductions into a single Poisson process, and do the same for all of their deaths. Then the probabilities of *any* birth or death occurring in the population in dt are given by $n\beta dt$ and $n\delta dt$ (with n the size of the population at that time). Since there can still only be at most a single event occurring in the infinitesimal time step (see Appendix A.1), we can construct a Markov chain to describe the process, as shown in Figure 3.3. With the natural numbers $\mathbb{N}$ as the state space, in any infinitesimal time step the population must do one of three things: increase by 1 if a birth occurs, decrease by one if a death occurs, or remain the same if neither happens. Thus we have the transition rates:

$$\begin{cases} p_{n,n-1} = \lim\limits_{dt\to 0} \dfrac{\mathrm{d}\mathbb{P}\{n-1 \text{ at } t+dt \mid n \text{ at } t\}}{\mathrm{d}t} = n\delta \\ p_{n,n} = \lim\limits_{dt\to 0} \dfrac{\mathrm{d}\mathbb{P}\{n \text{ at } t+dt \mid n \text{ at } t\}}{\mathrm{d}t} = 1 - n(\beta+\delta) \\ p_{n,n+1} = \lim\limits_{dt\to 0} \dfrac{\mathrm{d}\mathbb{P}\{n+1 \text{ at } t+dt \mid n \text{ at } t\}}{\mathrm{d}t} = n\beta \end{cases} \tag{3.25}$$

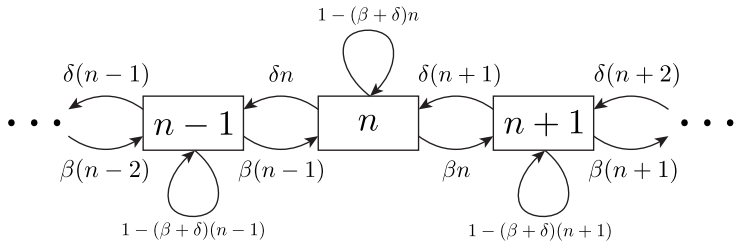


Figure 3.3.: *Markov chain visualization of the birth-death process.*

with all other $p_{j,i} = 0$. With these we can construct the master equation (3.24) for the probabilities $P_n(t)$ of the population having size $n \in \mathbb{N}$:

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = (n-1)\beta P_{n-1}(t) - n(\beta+\delta)P_n(t) + (n+1)\delta P_{n+1}(t) \tag{3.26}$$

Solving this differential equation analytically is clearly not easily done for all states $n \in \mathbb{N}$, though a numeric approach is certainly possible. Furthermore, we can use this expression to find the time evolution of the expected value of n

$$N(t) = \langle n(t) \rangle = \sum_{n=1}^{\infty} nP_n(t) \tag{3.27}$$

by noting that multiplying (3.26) by n and summing over $n = 1, 2, \ldots, \infty$ we obtain

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = (\beta - \delta)N(t) \tag{3.28}$$

Thus for some known initial size n_0 the expected size after a time t is given by

$$N(t) = n_0 e^{(\beta-\delta)t} \tag{3.29}$$

which is the exponential growth we would expect if we had ignored stochasticity in the first place.

3.3.2. The Moran process

In many applications of population modeling we are interested in characterizing the variation of a population's size with respect to that of another population with which it is competing. For example two types of predatorial animals hunting the same prey,

or two variant alleles of the same gene existing within a single group of individuals (as we will see in Chapters 4 and 5). While such problems might be most famous from a Darwinist perspective where selective advantages drive the dynamics, stochasticity remains a compelling force if selection is minimal. The Moran model was one of the first models proposed to study competition in the absence of selection. In its original form it describes the following: Given two populations A and B with respective sizes of n_A and n_B individuals, in each time step a random individual is chosen from the combined population $A \cup B$ to reproduce and another is chosen to die. In this manner the size of the total population $N_{A\cup B} = n_A + n_B$ remains constant in time, while n_A and n_B will vary stochastically. The model is similar to the birth-death model in that the individuals are granted only the two abilities of reproduction and death, however instead of characterizing the rates at which these occur in time the Moran model can be seen as a discrete Markov chain, with time measured in the number of simultaneously occurring birth and death events. Because the total size of the population is constant we need only know n_A to know the state of the system, since $n_B = N_{A\cup B} - n_A$, so that we may construct a state space for $N_A \in \mathcal{S}$ of the form $\mathcal{S} = 0, 1, 2, \ldots, N_{A\cup B}$. To evolve the probability distribution $P_n[i]$ of these states we can find the transition probabilities for the single time step, noting that again only a select few possible transitions can occur:

(i) The reproducing individual is not in A, the dying individual is in A

- state transition $n \to n - 1$
- occurs with probability $\frac{N-n}{N}\frac{n}{N}$

(ii) The reproducing individual is in A, the dying individual is not in A

- state transition $n \to n + 1$
- occurs with probability $\frac{n}{N}\frac{N-n}{N}$

(iii) Reproducing and dying individuals are both in or not in A

- state transition $n \to n$
- occurs with probability $\frac{n}{N}\frac{n}{N} + \frac{N-n}{N}\frac{N-n}{N}$

which leads to the transition probabilities:

$$\begin{cases} p_{n,n-1} = \dfrac{N-n}{N}\dfrac{n}{N} = p_n \\ p_{n,n+1} = \dfrac{n}{N}\dfrac{N-n}{N} = p_n \\ p_{n,n} = \left(\dfrac{n}{N}\right)^2 + \left(\dfrac{N-n}{N}\right)^2 = 1 - 2p_n \end{cases} \tag{3.30}$$

with again all other transitions $p_{i,j} = 0$. It is worth noting the existence of two absorbing states – meaning if the system enters this state it can no longer leave – at $n = 0$ and $n = N$, which correspond to either of the subpopulations going extinct.

3.4. summary

In this chapter we have introduced and discussed a number of stochastic processes which we will lean upon to construct models in the upcoming chapters. We have seen how many complicated processes involving a random outcome can be derived from the simple *Bernoulli* experiment (Section 3.1.1), which amounts to the tossing of a coin. By interpreting a succession of such trials as a timed process on a grid, and then taking the number of trials to infinity, we obtained the *Poisson process* (Section 3.1.2), which describes the likelihood of occurrences of a stochastic event in continuous time. While the Poisson process provides powerful tools for modeling random variables, it was discussed how the implicit assumption of *memorylessness* can be a limiting factor in its applicability, and must be taken into consideration when constructing or interpreting more complex models.

In Section 3.3.1 we introduced the *birth-death* process, which is a simple model for the stochastic growth of a population, constructed by taking the births and deaths of its members to occur as separate Poisson processes. Finally, in Section 3.3.2 we briefly discussed an alternative approach to population dynamics – the *Moran model* – which characterizes a competition between two stochastically varying populations, with the additional requirement that the sum of their sizes remains constant.

These two models of population dynamics will prove especially useful in the following Chapters, where we will investigate their application to the dynamics of hematopoietic stem cells.

4. Hematopoietic stem cells: a neutral stochastic population

I'm suspicious of people who are certain.

— Josiah Bancroft, *Senlin Ascends*

Stem cells form the basis of hematopoiesis. Indeed, our entire picture of hemocyte production hinges on the notion that the 10^{11} new cells required daily by the blood originated not so many divisions ago in a multipotent stem cell. Thus any model of hematopoiesis, be it qualitative or quantitative, must contain at least some basic principles describing HSC behavior. In fact, given the evidently vital role the stem cells play in facilitating a recovery from life-threatening disruptions, as well as the numerous blood disorders that have been linked to complications in their operation, one might argue that a detailed model of HSC functioning is essential to understanding the hematopoietic process. On the other hand, it is not surprising that observing these properties has proven particularly difficult. While *in vitro* experimentation has uncovered many of the capabilities of HSCs, their behavior *in vivo* remains to a large degree hidden. Nevertheless, with computational and mathematical models acting as a bridge between *in vitro* characterization and *in vivo* behavior, certain basic depictions of the system can be obtained and tested. In fact, one of the first mathematical models describing *in vivo* HSC dynamics was by Till and McCulloch themselves [133] – only a few years after their discovery of HSCs in mice – in which they apply a birth-death process (Section 3.3.1) to characterize the growth of the clonal spleen colonies described in their previous work. Since then, such models have played an important role in interpreting experimental findings and extrapolating from novel discoveries [12, 114, 37, 30, 91, 90, 125]. In this

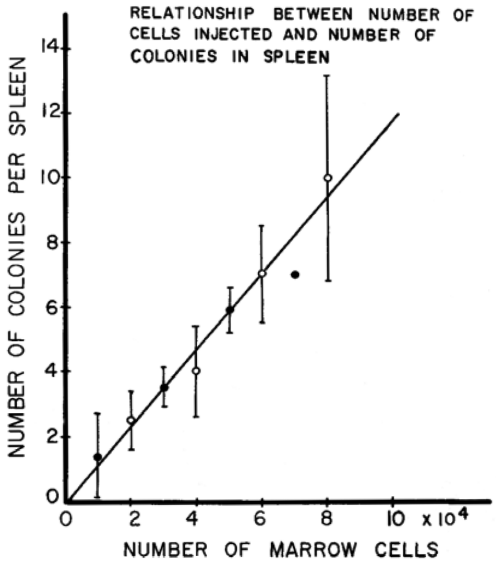


Figure 4.1.: Rarity of HSCs in mice. *Reproduced from Till and McCulloch 1961. Given that each colony is founded by a single HSC, one can estimate* 1 *in* 10^4 *mouse marrow cells to be a hematopoietic stem cell.*

chapter we will go over what biological characteristics might be relevant for modeling the dynamics of an HSC population, attempt to formulate what mathematical principles underlie their quantitative description, and establish what simplifying assumptions must be made in the process. In particular, our focus will lie on characterizing the stochasticity of the dynamics; on the one hand because the influence of the inherent randomness in the HSC population appears to be non-negligible in certain circumstances, and on the other because the predicted stochastic fluctuations can be a powerful tool in verifying a model's premise and assumptions. A deliberate and conscious introduction of the mathematical formulations of our biological concepts will prove useful in order to avoid the pitfalls of applying popular stochastic models naively without careful consideration of their underlying assumptions.

4.1. The importance of stochasticity

From the seminal paper in which Till and McCulloch demonstrated for the first time experimentally the existence of hematopoietic stem cells in adult mice [131] it was already apparent that these cells were rare, even within the confined niches of the bone marrow (Figure 4.1). This fact only solidified as HSC identification methods in humans became increasingly available. Even today, where our more detailed picture of "stemness" hints at subtler ambiguities concerning what constitutes a stem cell (see Section 2.4) [137, 101, 102], it remains undisputed that the total number of active (i.e. non-dormant) HSCs in an individual at a given point in time is relatively small, though the exact order of magnitude remains a topic of some debate [32, 78, 16, 76]. This carries an important implication for the clonal dynamics within an HSC pool: If some degree of randomness exists within the population dynamics – which data certainly seems to suggest [122, 133, 97, 137] – many emergent quantities will be more susceptible to stochastic fluctuations the smaller that population is. Indeed, in Chapter 5 we will see how from a simple model of stochastic divisional dynamics it can be shown that mutants arising in a small stem cell pool will have a realistic chance of expanding. Assessing the impact of these mutations can be important, since while their majority ends up being harmless – occurring in regions of little import or having no bearing on the particular function of the cell – it is no secret that occasionally a more malignant defect slips through the cracks. Cancers are of course the best known example of somatically acquired diseases, and though a large selective advantage is generally associated with a fully malignant cell population, a handful of driver mutations are typically required to achieve this state [138], whereas the earliest stages of oncogenesis can be subject to larger degrees of stochasticity [53, 146] (more [REF]s?). But there are other diseases originating from somatic mutations as well – such as *paroxysmal nocturnal hemoglobinuria*, studied in detail in Chapter 5 – in which stochasticity of the clonal dynamics plays an important role. Thus obtaining even a basic quantitative picture of mutational acquisition and the probabilities associated with clonal expansion in the HSC pool can be useful for improving our understanding of hematopoietic disorders. Moreover, acquiring a somewhat general picture of the stochas-

ticity involved in HSC dynamics will also prove incredibly useful for interpreting data coming from more recent *in vivo* experimental observations – as will be seen in Chapter 6 – allowing for more rigorous tests of our model assumptions and more direct methods of estimating the parameter values involved.

4.2. Assumptions for stochastic HSC dynamics

Consider an adult population of hematopoietic stem cells, slow in its activity, yet consistently providing a steady stream of multipotent progenitors embarking along the various paths of differentiation, as well as replenishing its own numbers through symmetric divisions. Although all of the HSCs in the population must trace their divisional ancestry back to a single zygote, each division in a cell's past has had the potential to add somatic mutations to its genome. In fact, the occurrence of such mutations (or *variants*) happens surprisingly quickly, with recent estimates suggesting on average 1.14 novel mutations are acquired per genome per cell division (without triggering apoptosis or other cellular self-repair mechanisms) in HSCs [145], leading to a rich diversity of somatic mutations within the population. A new mutation arising from a division is initially unique, however if the cell by which it is carried divides this mutation is passed on to both daughter cells, at which point it exists twice in the population. This notion establishes the idea of a single variant forming a *clone* – the term referring to a set of cells which share a particular mutation in the population. This clone, essentially tagged by the mutation which sets it apart from its ancestors, increases in size whenever one of its constituent cells divides symmetrically within the HSC pool, or decreases if it instead differentiates. Thus, given the assumption of stochasticity in its divisions, the expectation is for a clone to fluctuate randomly in size like a stock market price; sometimes oscillating around a long term average, other times drifting upwards or downwards for a time. The importance of such random drift cannot be underestimated, as it constitutes a mechanism for a somatic variant to expand – no matter how unlikely – within a cell population. Conversely, while oscillations around a value may not drastically influence the outcome of the hematopoietic process as a whole, the character of such fluctuations will (in theory)

conform to some predictable probability distributions, which in turn are determined by the underlying principles driving the dynamics. If experimental observation permits, such fluctuations can be measurable, thus providing a powerful method for testing the presumed behavior.

A useful model might tell us something about either of these things – the probability of events occurring, or the character of the fluctuations resulting from this randomness. The important characteristic highlighted here is that the fundamental quantity of interest is *the number of cells in a clone over time*. In this sense any model of HSC dynamics must in essence convey the mechanisms by which a genetic clone changes in size. We have already mentioned symmetric division and differentiation as drivers of such change, however, given what we know of HSCs and cell biology in general, there are other processes which can achieve the same result. Apoptosis or other forms of cell death can remove a cell from a population, while senescence could render a cell to be considered equivalently irrelevant for certain purposes [51]. The creation of new genetic clones occurs by mutations [87], which can only happen during cell divisions, unless extreme cases such as radiation are taken into account. Finally, it has been shown that differentiation from the stem cell state can in some cases occur without a cell division [52]. In order to construct a model, we must therefore decide which processes are most relevant to account for, and which may be ignored given current biological knowledge or in light of the proposed goal.

We will first adopt two simplifying assumptions which facilitate the construction of the models used in the following chapters:

- We take the supposition that differentiation away from the stem cell state occurs alongside a division, even though, as mentioned above, evidence does exist for differentiation programs which precede division.
- We ignore the effects of cell death and senescence.

While both assumptions may appear quite limiting, they can always be lifted if necessary, their respective influence on the conclusions drawn here by themselves providing questions of interest for future work. In this simplified picture the cell divisions act as

the drivers of change, causing both the arrival of new clones as well as their variability in size over time. We can now characterize the different ways in which these events alter the current state of the system. As shown in Figure 4.2, a cell can undergo one of three types of divisions: the cell can divide symmetrically, where the two daughter cells both either *(i)* remain stem cells or *(ii)* differentiate; or the cell can divide asymmetrically, in which case *(iii)* one of the daughter cells maintains the stem cell type while the other is differentiated. Each of these divisions alters the state of the system in a different manner:

(i) Symmetric self renewal

- The size of the total population increases by 1.
- The sizes of any clones (i.e. mutations) to which the dividing cell belongs increase by 1.
- Both daughter cells can acquire new mutations according to some mutation probability.

(ii) Symmetric differentiation

- The size of the total population decreases by 1.
- The sizes of any clones to which the dividing cell belongs decrease by 1.

(iii) Asymmetric division

- The size of the total populations is unchanged.
- The sizes of any subclones to which the dividing cell belongs are unchanged.
- The daughter cell which remains in the population can acquire new mutations according to some mutation probability.

Given these possible changes to the system, two things remain to be determined in order to simulate a population's history: the rate at which each of these divisions occurs and the rate at which mutations accumulate during each division.

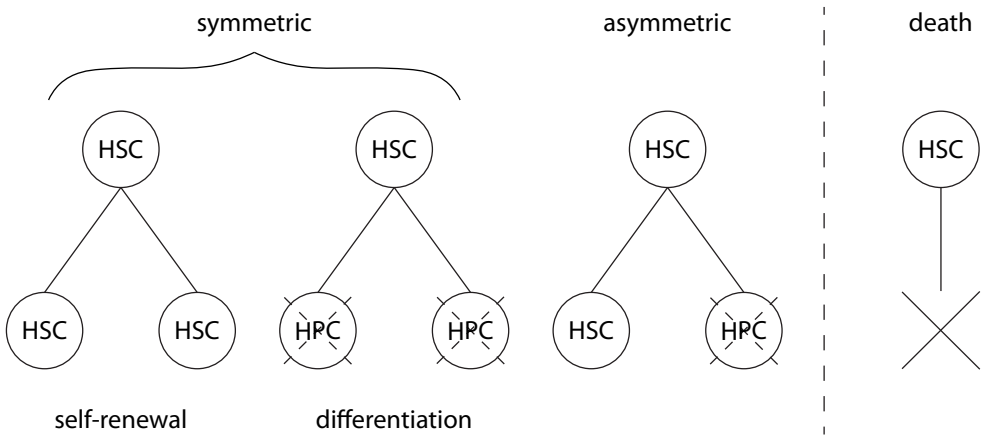


Figure 4.2.: *Possible events a single hematopoietic stem cell may undergo. After a symmetric self-renewal (i) both daughter cells are again stem cells (HSC), while a symmetric differentiation (ii) results in both daughter cells becoming progenitors (HPC), meaning that they are effectively removed from the population. An asymmetric division (iii) results in one HSC and one HPC. Finally an HSC may also undergo apoptosis or senescence, however these are not taken into account in the current model.*

4.2.1. Mutation rate

While multiple types of mutations can occur during the copying of a DNA strand, any variation will be passed on to progeny in future divisions and can thus serve as the first ancestor to a new clone. To develop a model for mutations acquisition we envision a cell division as a procedure where each of the $n = 3 \times 10^9$ nucleotides in the human genome must be copied, and each has the same probability of an error occurring. Thus each base pair in the daughter cell is a coin flip with a very high probability of success, but nonetheless a finite chance of failure. This is none other than the Bernouilli process described in Section 3.1.1, with the probability for m failures in the n trials given by the binomial distribution. Since the number of trials n is incredibly high, we can substitute it by the more computationally friendly Poisson distribution to which the binomial converges. As shown in Section 3.1.2 it is characterized by a single parameter μ, which is also the distribution's expected value. In other words, given the average number of mutations occurring after a single cell division, we immediately obtain the underlying probability distribution for the more general stochastic process.

4.2.2. Division rate

It is tempting to model the occurrence of divisions as a Poisson process as well, and this is ultimately what we will do. However, it is worth considering what reasons exist for rejecting this model, in part because it is important to be aware of how our models deviate from the reality, but also because the Poisson rate is so ubiquitous in biological models that one risks applying it without a conscious realization of its underlying assumptions.

The Poisson process assumes the time between events is exponentially distributed (see Chapter 3). This implies that an event is equally like to occur at any point in time, independently of the system's past history (the Markovian assumption). Thus in a cell population model, a cell's probability of dividing immediately after a division is the same as it would be at any later point in time. But this is untrue in most realistic scenarios. Most cells typically move through the cell cycle according to some biological clock, and while the time of each cycle may be subject to stochastic noise, it is by no means uni-

formly distributed [140, 18]. So applying this assumption of time independent division probabilities actually results in overestimating the randomness of the real system. Furthermore, allocating Poisson distributed divisions to each cell individually implies that all cells divide independently from each other, another strong supposition which may not always be the case. In spite of all this the Poisson model can still be surprisingly applicable, even in situations where this independence appears to be violated. While an in depth mathematical treatment of this topic falls out of the scope of this thesis, we can already form an intuitive picture of when time-independence can be approximately assumed. The heart of the argument is this: while time independence breaks down if the system has memory of its past, if this memory is finite – i.e. only a certain amount of time is remembered – a process which occurs on a much larger timescale will not be affected by it. Or in more mathematical terms: if any correlations of a system's state with its past are short livid, events which occur sporadically on a larger timescale are approximately independent in time. In the case of hematopoietic stem cells we can identify such a decoupling of timescales: While (long term) HSCs indeed undergo divisions according to a cell cycle, they spend most of their time in the non-proliferative quiescent state (see Section 2.4), only entering the cell cycle sporadically. The rate at which HSCs do end up dividing has been estimated on periods of once per few weeks up to as slow as once per year [76], implying that time-independence may indeed be an acceptable approximation. As for independence between different cells, this is an assumption we must take in absence of evidence to the contrary.

4.3. Modeling the stochastic dynamics of a mutant clone

4.3.1. A birth-death model (is not sufficient)

Let us now return to our population of stem cells. Ignoring for the moment the accumulation of new mutations, the only events which change the system are symmetric self-renewal – which increases the total population and any clones to which the cell may belong by 1 – and symmetric differentiation – which decreases these by 1. If we

assume these divisions all occur independently, then the times between them are exponentially distributed (i.e. Poisson distributed events). This is simply the birth-death process discussed in Section 3.3.1, where "births" are caused by self-renewal and "deaths" by differentiation, both occurring at distinct Poisson rates β and δ. Considering that throughout adulthood the size of the stem cell pool N remains (mostly) constant, we could set $\beta = \delta$, which means that on average we would expect there to be an equal number of self-renewals as there are differentiations. If we then consider different competing subpopulations, each would randomly evolve according to a different trajectory, with observable competitive dynamics as a result. However, there is a subtle issue with this approach. Specifically, that fixing the expected value of the size of the population $E(N)$ does not actually fix the size of the population. Thus, in a system which follows these dynamics it is perfectly plausible for the total population to change in size (even though the average for many such systems is fixed), and more importantly there is no force acting to push the system back to the expected value if it does vary. This does not fit with what we know of the hematopoietic stem cell pool. While fluctuations on cell numbers are occasionally observed in hemocyte populations [107], there is clearly an equilibrium number which is actively maintained, and while the HSC population is more difficult to observe in this respect, both mathematical arguments [32] as well as modern estimates of this number suggest this holds true (for the most part) at the stem level [78, 51]. We will solve this problem in the following manner. Recall that we may recast the independently occurring Poisson processes as a single combined Poisson process equipped with a probability p that determines which of the two occurs during each combined event (Section 3.1.2). In the birth-death model – with p the probability of a division resulting in differentiation – our best effort is to set $p = 0.5$, however this is clearly not enough to ensure an attracting equilibrium. A generalized method for introducing this is therefore to loosen the requirement of fixed self-renewal and differentiation rates, and allow p to depend on the current state of the system. The existence of a stable equilibrium around $N = E(N)$ then implies $p = 0.5$ (or $\beta = \delta$) only if this equilibrium is reached, while if $N > E(N)$ the next division is more likely to be a differentiation

($p > 0.5$), and if $N < E(N)$ a self-renewal ($p < 0.5$). The variation of p and its relationship with the system's deviation from equilibrium presents a new question in itself. Identifying a relationship $p(N)$ could involve either constructing a quantifiable model of the underlying processes which causes the mean-reverting behavior of N, or performing statistical analyses of time-resolved data to quantify HSC pool size fluctuations in single individuals. While both approaches would constitute a valuable study by themselves, they are somewhat out of the scope of the current research, especially since the simplest case scenario happens to coincide with one of the most famously studied models in the field of mathematical population dynamics: Taking $p = 1$ (certain differentiation) if $N < E(N)$ and $p = 0$ (certain self-renewal) if $N > E(N)$, each differentiation is always followed by a self-renewal, and vice versa – this is (nearly) exactly the Moran process discussed in Section 3.3.2 (although in the Moran dynamics the birth and death events are assumed to occur simultaneously).

4.3.2. A Moran model

Considering the above described assumptions naturally lead to the Moran model, we begin our treatment with this approach. Indeed, assuming differentiation and self-renewal events always come in pairs, there is mathematically no difference in taking them together in a single event. Furthermore, since the occurrence of division events are the only way for the system to change state, it is reasonable to consider measuring the time in number of divisions first and worry about moving to the real time picture later.
We first consider the stochastic dynamics of a single subclone $\mathcal{K}$ of size k within the total population of size N_{HSC}. After each Moran event (one self-renewal and one differentiation) the size of the total population remains unchanged, however the size of the subclone depends on the membership of the cells which divided. From 3.3.2 we know it may have moved to the states $k-1$ or $k+1$, or remained in state k according to the

transition probabilities:

$$\begin{cases} p_{k,k-1} = \dfrac{k}{N_{HSC}}\left(1 - \dfrac{k}{N_{HSC}}\right) = p_k \\ p_{k,k+1} = \dfrac{k}{N_{HSC}}\left(1 - \dfrac{k}{N_{HSC}}\right) = p_k \\ p_{k,k} = 1 - (p_{k,k-1} + p_{k,k+1}) = 1 - 2p_k \end{cases} \tag{4.1}$$

From these a master equation can be constructed denoting how the probability P_k of each state changes with a division:

$$P_k[T+1] = p_{k-1,k}P_{k-1}[T] + p_{k,k}P_k[T] + p_{k+1,k}P_{k+1}[T] \tag{4.2}$$

which for the transition probabilities given in 4.1 results in

$$P_k[T+1] - P_k[T] = p_{k-1}P_{k-1}[T] - 2p_kP_k[T] + p_{k+1}P_{k+1}[T] \tag{4.3}$$

4.3.3. Moving to real time

While measuring time in the number of division events that have occurred is useful for stating the dynamics in terms of the transition probabilities, we are still interested in obtaining a model which evolves in real time. Fortunately our careful introduction of division events occurring as a Poisson process in Section 4.2.2 allows us to easily extend the Moran model to a time based picture. Indeed, the independence of time increments means the Markovian property holds for an infinitesimal time step $\mathrm{d}t$, for which we may also write state transition probabilities. With a Poisson rate of Moran events ρ per single cell, and the assumption that all cells divide independently, we may take the total rate of events in the population as $N\rho$ (see Section 3.1.2). Thus we obtain the infinitesimal time transition probabilities:

$$\begin{cases} \mathbb{P}\{k+1, t+\mathrm{d}t \mid k, t\} = N\rho\, p_k \mathrm{d}t \\ \mathbb{P}\{k-1, t+\mathrm{d}t \mid k, t\} = N\rho\, p_k \mathrm{d}t \end{cases} \tag{4.4}$$

This allows us to write the master equation as a variation of the states in real time:

$$\frac{1}{N\rho}\frac{\mathrm{d}P_k(t)}{\mathrm{d}t} = p_{k-1}P_{k-1}(t) - 2p_kP_k(t) + p_{k+1}P_{k+1}(t) \tag{4.5}$$

It is worth emphasizing that the Poisson rate ρ for Moran events is not the same as the division rate per cell, as it describes the occurrence of two simultaneous symmetric divisions – one a self-renewal and the other a differentiation. We might then simply think of the total division rate as 2ρ, though doing this we must keep in mind an important property of this Moran based system: The occurrence of divisions irrespective of their type (self-renewal of differentiation) is not a Poisson process. Indeed, we saw in Section 4.3.1 that a true Poisson process of this form is the birth-death process, and it is the addition of memory (albeit very short term) that led us to the Moran model. By taking two division events simultaneously, we have a total number of division events $d = 2X$ with $X \sim \text{Pois}(\rho)$. Comparing this to a Poisson process $Y \sim \text{Pois}(2\rho)$ we find the same mean, however for the variance we find

$$\text{Var}(d) = \left\langle (2X)^2 \right\rangle - \langle 2X \rangle = 4(\rho + \rho^2) - 4\rho^2 = 4\rho \tag{4.6}$$

which is twice the variance of $Y \sim \text{Pois}(2\rho)$. Thus, while we will sometimes refer to 2ρ as *the total symmetric division rate*, we must take care not to apply it as if it described a Poisson process, since it is by no means a probability rate in the sense of (3.5).

4.3.4. The diffusion approximation

While the set of differential equations in (4.5) provide a useful method for evolving the system, they can become unwieldy if the population size N is large, since there are $N + 1$ coupled equations to be solved. Given that the active HSC pool in an adult human consists of somewhere between 5×10^2 and 10^6 cells [32, 78], we may find that despite their relative scarcity – which ensures the importance of stochasticity – their total population size can be considered large for the purpose of this computation. There is however an approximation which can be made to simplify the problem. Consider first

a rescaling of the states

$$\begin{aligned} &\mathcal{K} = 0, 1, \ldots, N \quad \rightarrow \quad \mathcal{F} = 0, 1/N, 2/N, \ldots, 1 \\ &k \mapsto f_k = k/N \end{aligned} \tag{4.7}$$

which conforms to simply moving from the picture of a clone's absolute size k to its frequency in the population. The larger the population size is the closer the states $f \in \mathcal{F}$ are together, so that if N goes to infinity $\mathcal{F}$ becomes a continuous subset of the real line $\mathcal{X} = [0, 1]$. Thus we might take this as an approximation for large populations. In Section 3.2.1 we briefly touched upon uncountable state spaces, noting that there is some trickiness in taking the limit from a discrete space, which is what we are attempting to do here. Fortunately, this particular type of problem has been well studied, and the correct solution is known.

Here we will briefly summarize the main aspects of this derivation and its result, and for an in depth treatment the interested reader is referred to [43] for the case specific to the Moran model, or [112] for a more rigorous and general approach.

In this picture our state space is the continuous subset of the real line $\mathcal{X} = [0, 1] \subset \mathbb{R}$. Now we define the transition $p(x', t'|x, t)$ as the probability to go from state $x \in \mathcal{X}$ to $x' \in \mathcal{X}$ in the time interval $t' - t$. Because our system is still Markovian, we can write the following property for this continuous space:

$$p(x_1, t_1|x_0, t_0) = \int_{-\infty}^{\infty} dx\, p(x_1, t_1|x, t) p(x, t|x_0, t_0) \tag{4.8}$$

From this property a general time evolution of the state probabilities can be obtained. The entire derivation is somewhat involved and can be found in any relevant reference ([112] is highly recommended) so it is left out here, but – using the shorthand $p(x, t) = p(x, t|x_0, t_0)$ – the result is

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial [A(x,t)p(x,t)]}{\partial x} + \frac{1}{2}\frac{\partial^2 [B(x,t)p(x,t)]}{\partial x^2} \tag{4.9}$$

which is the famous *Fokker-Planck equation* (or the *forward Kolmogorov equation*, de-

pending on who you ask), with

$$A(x,t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int d(\Delta x)\, \Delta x\, p(x + \Delta x, t + \Delta t | x, t) \tag{4.10}$$

$$B(x,t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int d(\Delta x)\, (\Delta x)^2 p(x + \Delta x, t + \Delta t | x, t). \tag{4.11}$$

The integrals in $A(x,t)$ and $B(x,t)$ are the first and second moments of the displacement Δx for the infinitesimal time step $\Delta t \to 0$. Thus constructing the Fokker-Planck equation for a system essentially comes down to finding these moments. For our current problem – i.e. the probability of a clone achieving a particular frequency – a trick to finding these is to assume these moments can be expanded in the small timestep:

$$\langle \Delta x \rangle_{\Delta t} = \langle \Delta x \rangle_\tau \, \Delta t + \vartheta(\Delta t^2) \tag{4.12}$$

$$\left\langle (\Delta x)^2 \right\rangle_{\Delta t} = \left\langle (\Delta x)^2 \right\rangle_\tau \Delta t + \vartheta(\Delta t^2) \tag{4.13}$$

where $\langle \rangle_\tau$ is the average in the discrete-time picture. Denoting $x = k/N$ we have

$$\langle \Delta x \rangle_\tau = -\frac{1}{N} \lambda x(1-x) + \frac{1}{N} \lambda x(1-x) = 0 \tag{4.14}$$

and

$$\left\langle (\Delta x)^2 \right\rangle_\tau = \frac{1}{N^2} \lambda x(1-x) + \frac{1}{N^2} \lambda x(1-x) \tag{4.15}$$

so that we obtain

$$A(x,t) = 0 \tag{4.16}$$

$$B(x,t) = \frac{2\lambda}{N^2} x(1-x) \tag{4.17}$$

With $\lambda = N\rho$ the event occurrence rate, the Fokker-Planck equation then becomes:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\rho}{N} \frac{\partial^2 [x(1-x)p(x,t)]}{\partial x^2}, \tag{4.18}$$

Given that this continuous picture is an approximation, it is fair to wonder to what extent it deviates from the true system. It has been shown that this approach is the least accurate near the boundaries $x = 0$ and $x = 1$, which is not entirely surprising: the states $k = 0$ and $k = N$ in the true system are absorbing states, whereas in the approximation the system can approach these much closer without fixating. This particular behavior will not be relevant for our purposes, and the reader is referred to [43] for more information.

5. Evolutionary dynamics of paroxysmal nocturnal hemoglobinuria

The fact that clonal dynamics can occur even in the absence of selection is an important realization, as it has implications for the impact the acquisition of malignant mutations can have on the body. A prime example of this can be found in the hematopoietic disorder *paroxysmal nocturnal hemoglobinuria*, or PNH. With the advent of DNA sequencing techniques this long studied disease was found to be strikingly simple in nature, caused by a single point mutation in a gene known as *PIGA* [15], and occurring in the hematopoietic stem cell pool. As it is an acquired disease, it was for a long time assumed this mutation must have some selective advantage for its associated clone to reach the sizes found in patients [85]. However, an alternative explanation – first proposed by Dingli et al. [31] – is that expansion of the PNH clone could be simply due to neutral drift, or in other words: plain old bad luck. In this Chapter we use the techniques developed in Chapter 4 to investigate the implications of this hypothesis and how it relates to the available data, starting from the initial work done by Dingli et al. and expanding on it to obtain new insights.

5.1. Paroxysmal nocturnal hemoglobinuria

PNH originates from any function breaking mutation of the gene *PIGA* found on the X-chromosome [15, 85]. It encodes a protein required for the biosynthesis of GPI (glycophosphatidylinositol), a phospholipid used to anchor various proteins to the cell surface of hemocytes [59, 14]. Among such proteins are CD55 and CD59, both serving as

regulators of cell-to-cell interactions and acting to protect the bearer from complement mediated lysis (destruction due to the immune system). Thus a lack or reduced abundance of these on a cell's surface (here referred to as the GPI- phenotype) results in its destruction in circulation. Clinical PNH occurs when a large fraction of hemocytes share the GPI- phenotype, leading to severe intravascular hemolysis and a plethora of symptoms, the most common of which are anemia (decreased amount of erythrocytes in circulation), venous thrombosis (blood clotting in veins), and hemoglobinuria (excess of hemoglobin content in urine).

Because *PIGA* resides on the X-chromosome and is subject to X-linked inactivation, a single somatic mutation is sufficient in both males and females to disrupt the GPI production pathway [85]. Thus from a probabilistic standpoint one can consider its X-linked nature to be causative for the disease's existence. Still, the appearance of a single *PIGA* mutant in the hemocyte precursors does not directly lead to PNH, as a clone of sufficient size is required to produce clinical symptoms; in fact, *PIGA* mutations can be found in the blood at low frequency in healthy adults [7]. Furthermore, the observation that in a single individual a large *PIGA* clone appears across all hematopoietic lineages [19, 99] implies such a clone must originate early in the hematopoietic hierarchy, a notion which is strengthened by our dynamic picture of hematopoiesis: If a *PIGA* mutant arises late in the maturation process, its clonal offspring will soon be washed out of the system by new cells from earlier less committed stages [143]. Thus it is generally accepted that clinical PNH must be initiated by a *PIGA* mutation arising in the hematopoietic stem cell pool. However, what occurs after the first *PIGA* mutation arises remains debated. Because clone sizes can be extremely large in patients – ranging up to 90% of all blood cells [123] – it was initially assumed there must be some selective advantage to the GPI- variant in the bone marrow, however what mechanism this would be proved difficult to find. So far it has been shown that *PIGA* deficient cells are not more resistant to apoptosis [67], nor do they exhibit a proliferative advantage compared to normal cells [88], and neither their replication rate [8] nor their mutation rate is altered [6]. In light of this it has been proposed that a selected advantage of the GPI- phenotype is extrinsic to the cells

themselves, and is instead mediated by an immune attack on normal GPI+ cells [86]. While this hypothesis is supported by some evidence [24, 49, 48], it is unable to explain two key observations. Firstly, *PIGA* and the GPI- phenotype are ubiquitously expressed in the body, and an explanation as to why an immune attack against the GPI anchor would be restricted to the HSC population is lacking. Secondly, a significant fraction of patients with PNH undergo spontaneous reduction and even extinction of the GPI- clone [60], which is difficult to reconcile with the proposed autoimmune selection. A second selection-based hypothesis postulates the occurrence of additional mutations in one or more other genes, which when appearing simultaneously with the *PIGA* variant in the same cell confer a fitness advantage to the GPI- phenotype. Indeed, several case reports of coexisting mutations involving *PIGA* are available, including two patients with a mutation in HMGA2 [64], one patient with a concomitant *JAK2V617F* mutation [127], a mutant *NRAS* [98], and more recently a patient with PNH and concomitant *BCR-ABL* fusion in the same cell population [69]. However, these cases appear to be the exception rather than the rule, and there is little evidence for a conferred fitness advantage in any of them. Furthermore, the arrival of a second key mutation in an existing clone required for its expansion would be an extremely rare event if the clone is still small [135, 34], while deep sequencing of diagnosed patients has revealed that a significant portion of PNH clones do not carry additional mutations apart from that in *PIGA* [118].

Given these observations Dingli et al. showed that with a small enough stem cell population and the simple assumption of stochastic dynamics, this appeal to selection may not be necessary, as the disease's low prevalence can be perfectly explained by attributing the large clone sizes to stochastic outliers [31]. They demonstrated this by directly simulating Moran like dynamics: By sampling state transitions from the appropriate probability distributions, a large ensemble of stochastic trajectories was used to extract the expected prevalence of the disease under neutral drift, which was shown to be of the same order as found in a real world population. Here we will reexamine their model with the more powerful Markov chain method described in Chapter 4, which will

allow us to investigate a number of additional properties and predictions of the model, such as the occurrence of PNH during childhood, the spontaneous loss of a clone due to neutral drift, and the expansion rate of PNH clones. Furthermore, we can apply it to estimate the likelihood of coexistence of multiple distinct PNH clones (i.e. arising separately from different mutating cell divisions) in the HSC population, a phenomenon that has been observed in some patients. In such cases often the two clones present different levels of severity, e.g. one with complete deficiency of GPI anchored proteins and another with partial deficiency of GPI [59, 115]. Targeted sequencing of the *PIGA* gene can confirm different mutational profiles in the two cell populations [118], implying they must have arisen due to independently occurring mutations.

5.2. Applying the Moran model

To assess the likelihood of neutral drift expansion, we are interested in obtaining probabilities related to the appearance and growth of sizeable PNH clones in the hematopoietic stem cell pool of an individual. The Markov chain formulation of the Moran dynamics of the HSC population introduced in Section 4.3.2 provides the perfect tool for this: in a two population system – here normal cells and *PIGA* mutants – it allows us to calculate the probabilities $P_m[T]$ of all possible states $m \in 0, 1, \ldots, N_{HSC}$ corresponding to the possible size of the *PIGA* population. To facilitate a close comparison with the results of [31], we evolve the discrete-time Markov chain of (4.2) rather than the continuous-time ODE's (4.5), where (as in the original paper) the real time between division events is taken to be the expected value $1/\lambda$ of its exponential distribution (see Section 4.2.2).

5.2.1. Transition probabilities

As PNH is an acquired disease, we must take into account the fact that a stem cell pool does not begin with a *PIGA* mutant. Only when a normal cell divides is there a probability μ of such a mutation occurring in one of its daughter cells. Thus we require a slightly altered set of transition probabilities $p_{i,j}$ compared to the standard Moran model

introduced in (4.1), which we find by listing all possible occurrences during a division event. Denoting m as the number of *PIGA* mutated cells in the HSC pool, these are:

- *A normal HSC divides without PIGA mutation & a normal HSC differentiates*
  Effect: $m \to m$
  Occurs with probability: $\left(1-\frac{m}{N}\right)(1-\mu)\left(1-\frac{m}{N}\right)$
- *A normal HSC divides with PIGA mutation & a normal HSC differentiates*
  Effect: $m \to m+1$
  Occurs with probability: $\left(1-\frac{m}{N}\right)\mu\left(1-\frac{m}{N}\right)$
- *A normal HSC divides without PIGA mutation & a mutated HSC differentiates*
  Effect: $m \to m-1$
  Occurs with probability: $\left(1-\frac{m}{N}\right)(1-\mu)\frac{m}{N}$
- *A normal HSC divides with PIGA mutation & a mutated HSC differentiates*
  Effect: $m \to m$
  Occurs with probability: $\left(1-\frac{m}{N}\right)\mu\frac{m}{N}$
- *A mutated HSC divides & a normal HSC differentiates*
  Effect: $m \to m+1$
  Occurs with probability: $\frac{m}{N}\left(1-\frac{m}{N}\right)$
- *A mutated HSC divides & a mutated HSC differentiates*
  Effect: $m \to m$
  Occurs with probability: $\frac{m}{N}\frac{m}{N}$

Combining all events which lead to the same effect we obtain the following adapted transition probabilities:

$$\begin{cases} p_{m,m-1} = \dfrac{m}{N_{HSC}}\left(1-\dfrac{m}{N_{HSC}}\right)(1-\mu) \\ p_{m,m+1} = \dfrac{m}{N_{HSC}}\left(1-\dfrac{m}{N_{HSC}}\right) + \left(1-\dfrac{m}{N_{HSC}}\right)^2\mu \\ p_{m,m} = 1-(p_{m,m-1}+p_{m,m+1}) \end{cases} \tag{5.1}$$

5.2.2. Ontogenic growth

Since we are interested in evolving the system from birth, we must also take into account the fact that the HSC pool increases in size during ontogenic growth. The growth curve in an average individual has been documented before [33, 142] – shown in Figure 5.1 – and can be applied to the early time evolution of the system. This is done by – at predetermined times – inserting additional self-renewal events without an accompanying differentiation event, which affect both the clone size probabilities as well as the *total* population size. During such a self-renewal event the following transition probabilities $q_{i,j}$ are found:

$$\begin{cases} q_{m,m-1} = 0 \\ q_{m,m+1} = \dfrac{m}{N_{HSC}} + \left(1 - \dfrac{m}{N_{HSC}}\right)\mu \\ q_{m,m} = 1 - (q_{m,m-1} + q_{m,m+1}) \end{cases} \tag{5.2}$$

During the growth phase of the system the normal division/self-renewal Moran dynamics are periodically interrupted by a number of such self-renewal-only events. Since the growth function $N_{HSC}(t)$ obtained from [33] is given in time intervals Δt of two weeks, after each interval $N_{HSC}(t) - N_{HSC}(t - \Delta t)$ self-renewal events are performed to increase the population size.

5.2.3. Observing multiple clones

While the above described method provides a powerful way to describe the evolution of a single PNH population, it does not facilitate tracking the evolution of multiple clones that arise from separate mutational events. Indeed, the one-dimensional state space described by $m \in 0, 1, \ldots, N$ cannot contain information about a third sub-population, although we might expand the system to account for multiple clones. In particular, the tracking of $M+1$ distinct populations (M mutant clones + the unmutated population) requires an M-dimensional state space. For large M the system quickly becomes computationally unsolvable, as the number of states scales with $\sim N^M$. However, we can take advantage of previous estimates showing that the likelihood of more than two *PIGA* clones in

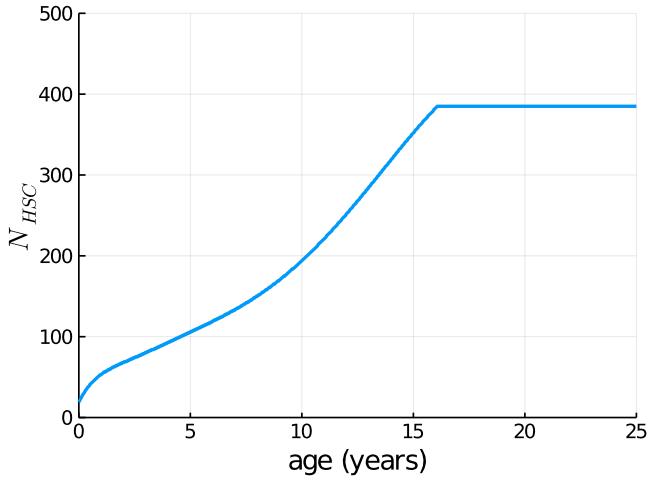


Figure 5.1.: *Size of the HSC pool over time during ontogenic growth, derived by Dingli and Pacheco [33]. The population size remains approximately stable from adulthood.*

the same stem cell pool is vanishingly small [135, 34], to limit our state space to two dimensions and thus account for two different clones. A first thought would be to evolve a $P_{m,n}[T]$ as the probability of the first clone having size m and the second size n. There is, however, some useful information related to the history of a trajectory not contained in such a description. Consider for example the initial state $(m = 0, n = 0)$ at $T = 0$ before any clones exist. If at some point later in time during a trajectory a clone arises and then disappears, the system is once again in $(m = 0, n = 0)$, and there is no way of knowing whether a clone ever existed or not; and the same is true if two clones appear and subsequently vanish. This problem stems from the Markovian property of the time evolution and the fact that these different scenarios lead to the same state. One way to circumvent this is by separating the cases we wish to disentangle into different states. To this end we introduce an extended state space that is divided into three separate "histories" corresponding to states with different pasts, as shown in Figure 5.2. Each history comprises a collection of states where a specific number of mutations occurred in the system's past – zero, one, or two. In this picture, an evolutionary trajectory in which a mutation occurred but the resulting clone eventually died ends in a different state than

one where no mutation occurred in the first place. Note that the master equation (3.22) must be altered to include transitions to different histories:

$$P^{i}_{m_1,m_2}[T+1] = \sum_{m'_1,m'_2,j} p^{i,j}_{m'_1,m'_2,m_1,m_2} \times P^{j}_{m'_1,m'_2}[T] \tag{5.3}$$

where any state is now characterized by the clone sizes m_1 and m_2, and the appropriate history i. The tensor elements $p^{i,j}_{m_1,m_2,m'_1,m'_2}$ represent transitions from state (m'_1, m'_2) to m_1, m_2 and history i to j, and are found in an identical manner as before, though now more transitions are possible, as shown in Figure 5.2.

5.2.4. Parameter values and diagnosis threshold

Taking the Moran replication rate of HSCs at 1/cell/year [119] and the known normal mutation rate of the *PIGA* gene 5×10^{-7} per replication [6], we can iteratively evolve the master equation from state $P_0[0] = 1,\ P_m[0] = 0\ \forall m \neq 0$ to obtain probabilities for all possible size combinations of the mutant clones. However, since the mere existence of a variant *PIGA* clone does not yet constitute a clinical diagnosis of PNH, we require a threshold for the minimum clone size that would lead to diagnosis. In general symptoms of hemolysis only become mildly present from clones of at least 10% [109], with the affect varying among individuals, so that we take this threshold at 20%.

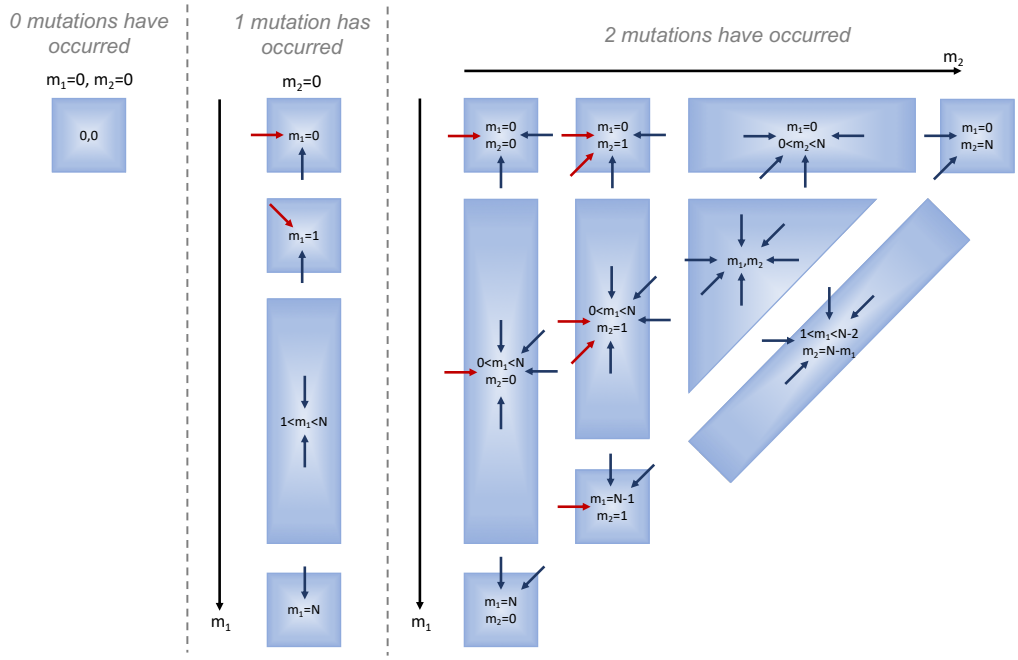


Figure 5.2.: The extended state space and allowed transitions. *Each history describes a possible evolution where either 0, 1 or 2 mutations were acquired by distinct healthy cells. Whenever a mutation occurs the system jumps to the next panel on the right (to the next history), until the final history is reached where mutations are no longer allowed. The dark blue arrows represent incoming transitions from nearest neighbor states in the same history, while the red arrows represent transitions from states in the previous history. Note that every state also has a transition onto itself, which has been left out for readability.*

5.3. Results and predictions

Using the model and parameter values described above we can make predictions for the likelihood of clonal expansion, along with expected clone sizes, expansion rates, the expected age for acquiring the disease, and more. Whenever possible we compare these results to available data, much of which comes from the International PNH Registry [123]. However, for quantities related to incidences within a population – such as the disease prevalence and average clone size – the probabilities for a single individual are not immediately comparable with the data, since in the real world the ages of the individuals that make up any population are non-uniformly distributed. The fact that the probability state space changes over time implies that prevalences in a population will be skewed by the age groups which have the highest representation. In order to perform a meaningful comparison with data from a real world population the projected age-specific probabilities must therefore be weighed with the relevant age distribution. To this end we use the reported distribution from the United States 2010 census [21].

5.3.1. Probability and prevalence of PNH

Evolving the system for a long time period (100 years) we find that the probability for a single individual to develop clinical PNH through neutral drift increases with age according to the curves shown in Figure 5.3, being especially low early in life. Indeed, the limited data available suggests that PNH is quite rare in children [139, 27], generally only occurring in the context of bone marrow failure. While *classical hemolytic* PNH (i.e. in the absence of other disorders such as aplastic anemia) represents about 10% of pediatric patients with a *PIGA* mutant population, data from the International PNH registry suggests that perhaps half of adults with PNH have classical hemolytic disease [123].

We find that the probability of a patient having clinical PNH with two independent clones arising in the HSC pool is approximately 10^3 times smaller than the probability of the same diagnosis with a single clone, while the occurrence of patients with 3 or more distinct PNH clones contributing to hematopoiesis would be another 2 orders of

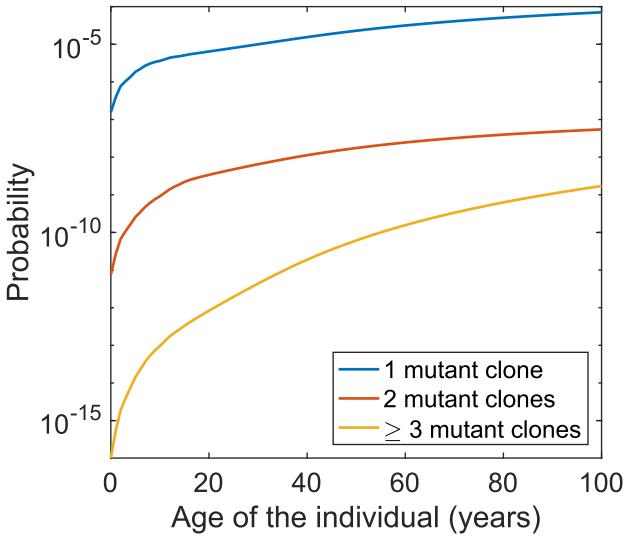


Figure 5.3.: Probability of clinical PNH occurring in a single individual over time. *Absolute probability of clone size ≥20% for 1 or more coexisting clones.*

magnitude lower (Figure 5.3). This implies that approximately only 1 in 1000 cases of clinical PNH would host more than a single mutant clone that arose in the stem cell compartment. Note that these numbers result from a model dealing only with stem cell dynamics, and thus do not preclude the occurrence of mutations farther downstream among progenitor cells, which are present in larger numbers and divide faster than the HSCs [135, 35]. Moreover, *PIGA* mutations occurring in early progenitors will also remain contributing to hematopoiesis for years before any eventual wash-out [143, 136]. Thus this model estimates that clonally distinct *PIGA* mutations found in mature cells are more likely to have originated at later stages of differentiation [135] than in independent mutations occurring in the active stem cell population.

Using the population age distribution data from the 2010 United States Census, we estimate the incidence of clinical PNH for both mono- and multiclonal cases in the USA, by weighing the clonal size probabilities for each age with the respective prevalence of that age in the population, the result of which is shown in Figure 5.4. We obtain an

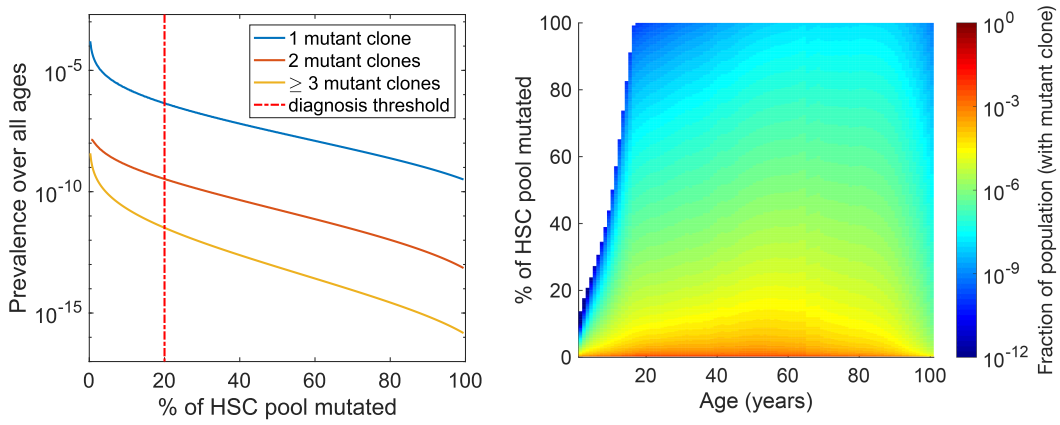


Figure 5.4.: Incidence of PNH in an age varying population. ***(a)*** *Predicted incidence of PNH clone sizes in the US population, found by folding the Markov chain probabilities for ages 1-100 with population data from the 2010 US census. Any clone above the 20% threshold (vertical dashed line) counts as a clinical diagnosis.* ***(b)*** *Predicted incidence of PNH clone sizes in the US population for all ages separately.*

expected total prevalence – calculated by summing over all clones sizes from 20-100% in Figure 5.4a – of 1.76 cases per 10^5 citizens for any diagnosis of clinical PNH (mono- or multiclonal), which is similar to what has been reported in a well-defined population by Hill et al. [56]. To investigate the likelihood of multiple clones, we also calculate the expected number of patients with biclonal disease arising at the level of the HSC, obtaining a prevalence of 1.29 per 10^8 individuals. For the US population, this would amount to approximately 3000 patients with a single clone and 2 patients with biclonal disease, respectively. The number of individuals in the population with a subclinical ($< 20\%$) *PIGA* mutated clone is estimated to be much higher, at 6.0 per 10^4 for monoclonal and 1.9 per 10^7 for biclonal cases, which amounts to respectively 184,495 and 60 individuals in the US.

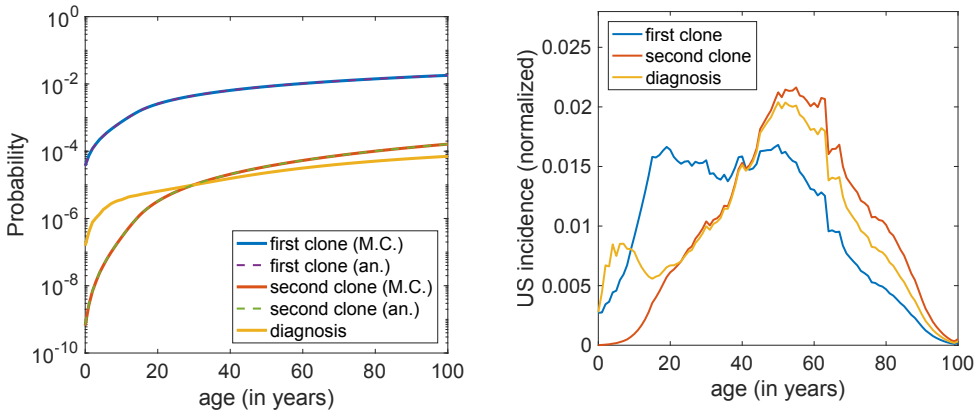
5.3.2. Average clone sizes

Using the census data we can also estimate the average clone size $\langle m \rangle$ in individuals in the US population. If we condition upon the existence of at least one mutant HSC, the average size is estimated in our model to be at 3.4% of the total pool. The more interesting statistic however is the average clone size in those suffering from clinical PNH ($m \geq 20\%$), as it can potentially be compared to clinical data. For diagnosable individuals, it is predicted to be 31.1%, with a very large standard deviation of 32.6%. Comparing this to data from the International PNH registry [123], this appears to fit well with patients with AA-PNH syndrome (categorized as simultaneously suffering from aplastic anemia), which shows a mean clone size of 28.3% and standard deviation of 32.8%, though other categories such as classical PNH present much greater average clone sizes.

5.3.3. Arrival times of mutated clone and clinical PNH

Another useful type of prediction relates to the arrival time of a mutant cell or a clinically diagnosable clone. We find that the first mutated cell in the HSC pool can occur quite early in an individual's life, as shown in Figure 5.5a, and the probability of harboring a mutant clone in the stem cell population grows one order of magnitude from age 20 ($\sim 2 \times 10^{-3}$) to age 100 ($\sim 2 \times 10^{-2}$). Despite these values appearing quite high, in a neutral drift picture the second line of defense against PNH is the significantly low likelihood of clonal expansion, a fact that is illustrated well by comparing the probability of occurrence of a clone (which is quite common in healthy people [7]) with the probability of having clinical PNH. For example, in an individual of age 60, the probability of having acquired a mutant clone is 1×10^{-2}, while the probability of having clinical PNH is 2×10^{-5}, three orders of magnitude smaller.

The average ages of clonal occurrence in an age distributed population are projected at 41 and 54 years for mono- and biclonal (stem cell) cases respectively Figure 5.5b. In general, it appears that on average most clones arrive only after adulthood is reached and the hematopoietic stem cell pool has attained its maximal size.



(a) *(b)*

Figure 5.5.: Clonal existence and arrival times ***(a)*** *Likelihood of existence of clones over time. As a test of accuracy, the probabilities for the existence of the primary and secondary clones were also calculated analytically from a cumulative negative binomial distribution.* ***(b)*** *The probability of obtaining a first or second clone in a given year as well as the probability of reaching the diagnosis threshold (20% of the HSC pool) folded with the 2010 US population distribution to obtain the age incidence in the population. Each curve has been normalized, so that they may be interpreted as the age distribution of the clone and diagnosis arrival times. (M.C.: Markov Chain simulations; an.: Analytical calculations.)*

The average age at diagnosis – taken as the time at which the total number of mutated HSCs reaches 20% for the first time – is found to be 49 years (or 44 years if threshold for diagnosis is taken at 10%) and is quite similar to what has been reported by the International PNH registry, with 43.2% of patients diagnosed between 30 and 59 years of age. [123, 83].

5.3.4. Clonal expansion

A final aspect of the disease we look at is the rate a which a clone changes in size, and how this relates to the potential spontaneous loss of PNH. This rate can be estimated

with the Markov chain method by initiating a clone at size m_0 with probability 1, and subsequently observing the probability distribution of the clone size at later time, as shown in Figure 5.6. This distribution clearly diffuses over time, however it does not occur entirely symmetrically (note the slight skew to the distributions) as one would expect from a typical markovian process. The reason for this can be seen from the Moran transition probabilities (5.1). Ignoring for the moment the small probability of a mutation occurring ($\mu \approx 0$), we obtain once again the basic Moran transitions

$$p[m] = p_{m,m+1} = p_{m,m-1} = m/N_{HSC} - (m/N_{HSC})^2 \tag{5.4}$$

The fact that the probabilities for moving up or down from the current state are identical implies that the expected value for the state cannot change[1], however it does not imply that the diffusion across all states will be symmetrical. Indeed, interpreting $p[m]$ as (half) the amount of probability that leaves the state m in a time step, and noting that $p[m]$ is itself a concave function of m, we see that the rate of change is in fact state dependent, occurring fastest at the maximum of $p[m]$ (at $N_{HSC}/2$) and slowest at the points 0 and N – where it is effectively 0.

We can compare the predicted diffusion over time with measurements from Araten et al. [8], who reported a $\geq 5\%$ size increase per year in 12 out of 36 patients, while most of the other patients experienced either a reduction or no change at all; though the authors did not specify these amounts quantitatively. The study found no significant expansion or reduction ($\approx 0\%$) when calculating the mean over all patients, which is exactly what the neutral model predicts. On the other hand, our model projected the fraction of patients that would experience a $\geq 5\%$ increase after 1 year to be between 5% and 10% depending on the size of the initial clone, which is significantly lower than their observed 33%. While this discrepancy could be attributed to statistical confounding factors such as the relatively small size of study's patient cohort, there may be a more fundamental explanation, which will be discussed in Section 5.5.

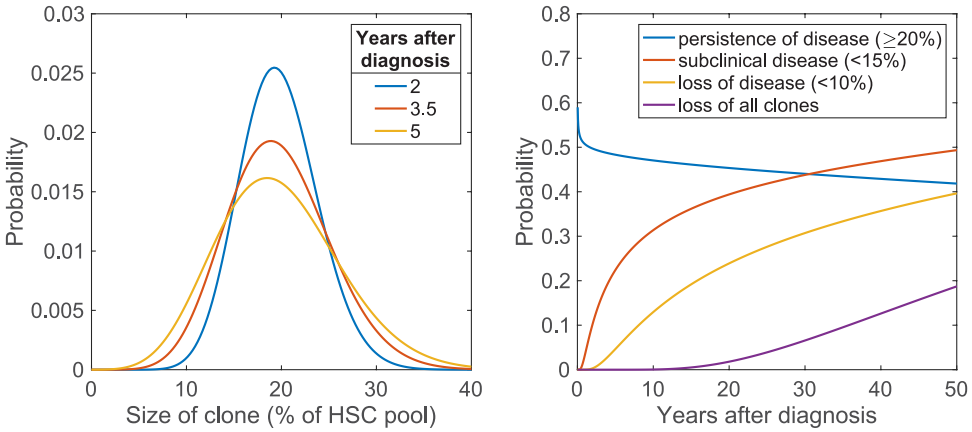
[1] in probability theory this is known as a Martingale

5.3.5. Disease reduction

While there is currently no known cure for PNH other than bone marrow transplantation (though the hitherto successful complement inhibiting therapy *eculizumab* can possibly be administered indefinitely to prevent hemolysis [61, 19]), it is nevertheless possible for the disease to disappear spontaneously without intervention. While this occurrence is difficult to reconcile with a positive selection acting on the variant *PIGA* clone, the neutral hypothesis inherently predicts it, since at any point in time the mutant clone an would have equal chance of shrinking as is would of growing.

We calculate the probability of a recently diagnosed case of clinical PNH becoming subclinical again by evolving a system from the state $m = 0.2N_{HSC}$ and observing the probability of it being in a lower state $m < 0.2N_{HSC}$ at a later time. The result shown in Figure 5.6b demonstrates that starting from a clone of size 20%, over time it is in fact more likely for the disease to recede than to persist. This may seem counterintuitive, since earlier we showed that (ignoring new mutations) the expected value of the size of an established clone does not change. The reason for this apparent contradiction is found in the skewness of the diffused distribution. While the expected clone size $\langle m \rangle$ remains fixed at $m = 0.2N_{HSC}$, the right-sided tail of the distribution (where $m > 0.2N_{HSC}$) stretches out farther than the left-sided tail ($m < 0.2N_{HSC}$) due to the higher diffusion rate determined by $p[m]$; thus the fixed expectation value implies there must be more total probability contained in the $m < 0.2N_{HSC}$ subset of the phase space than in the $m > 0.2N_{HSC}$ subset. On the other hand, since $p[m]$ is symmetric around $0.5N_{HSC}$, clones which consist of more than 50% of the HSC population will be less likely the disappear than to persist.

We find that after diagnosis ($m = 0.2N_{HSC}$) it would take at least 2-10 years for significantly smaller clone sizes ($< 15\%$ or $< 10\%$) to be reached, while the possibility of the clone becoming truly extinct is only realizable after 20-50 years, and in reality clinically detectable extinction will depend on the assay that is used to determine the presence or absence of the clone. Importantly, the model predicts that after 10 years from diagnosis, the probability that the clone is small enough not to be associated with



(a) *(b)*

Figure 5.6.: Clonal expansion and disappearance. ***(a)*** *The size probability distribution for an established clone ($m = 0.2N_{HSC}$) multiple years after diagnosis.* ***(b)*** *Probabilities for an established clone to recede or vanish after diagnosis over time.*

clinical PNH is upwards of 30% (Figure 5.6b), which is comparable to what Hillmen et al [60] reported in their cohort of patients, with 12 out of 35 surviving patients having spontaneous clinical remission within a 10 year followup.

5.4. Discussion

The appearance of mutations in HSCs and their fate over time is an important clinical problem, since many diseases such as myelodysplastic syndromes and several leukemias (e.g. chronic myeloid leukemia, some subtypes of acute myeloid leukemia) arise due to mutations within the HSC. Landmark studies in PNH have shown that it is an acquired clonal HSC disorder [104] with very interesting dynamic properties, including an uncanny probability of spontaneous clonal extinction [60]. The mechanism of clonal expansion in PNH has been a source of great debate and several hypotheses have been proposed to explain it, such as a selective advantage of the mutant cells due to an immune attack on normal HSC (extrinsic advantage), or the presence of a second mutation that grants a

fitness advantage (intrinsic advantage). Some evidence for either hypothesis exists, but both also suffer from deficiencies as described earlier. In particular, in is unclear how an extrinsic advantage would be relegated to the HSC compartment alone, and It is also difficult to argue how a cell could acquire multiple mutations sequentially in the absence of genomic instability, which has not been observed in PNH [6]. Dingli et al. have proposed that the *PIGA* mutant cells generally possess no fitness advantage or disadvantage, and that clonal expansion is simply a consequence of neutral drift within the (small) active HSC pool that maintains hematopoiesis [31]. Neutral drift may come as a surprise for many in the field of hematology and oncology who are accustomed to associate malignant clonal expansions in cancer with some form of selective advantage. Nevertheless, it is not uncommon for mutations in populations to expand in this way, as suggested by Kimura many years ago [72]. This hypothesis leads to the simplest of explanations of PNH, and our stochastic modeling suggests that this could be the case – at least in some patients – since we are able to predict the prevalence of the disease, average age at diagnosis, average clone size and the probability of clonal extinction purely from first principles through a Markov chain evolution of a Moran model of HSCs, with results quite similar to what has been reported in the literature. Furthermore, two important observations – finding a non-varying expected clone size in a large cohort of patients, and the occurrence of spontaneous remission of the clinical disease in a large fraction of patients – would require complicated explanations in the selection picture, whereas they are immediate consequences of applying a neutral model. Thus, although it is difficult to deliver conclusive proof of the neutral hypothesis, the close parallel between these predictions and the clinical reality provides considerable support for it.

While the Moran model's predictions match much of the clinical data surprisingly well, some quantities obtained appear to consistently underestimate what is reported in the literature. In particular, the average clone size in diagnosed patients with classical hemolytic PNH is reported at $69.8 \pm 32.9\%$ in a cohort of 550 patients – much higher than the 28.3% calculated here – and the rate of expansion reported by Araten et al. is a

$\geq 5\%$ increase per year in 1/3 of patients [8] – whereas the current model predicts this to only occur in 1/10 to 1/20 of patients. We might simply gloss over these discrepancies, attributing them to a failure of the model, or even a knock against the neutral drift hypothesis, however it is worth digging a bit deeper and considering whether there may be a more fundamental piece of the true system that we have left out of the model.

5.5. Perspective: HSCs under perturbed hematopoiesis

In constructing the Moran model in this chapter we have, perhaps somewhat unknowingly, made a very important assumption related to the HSC dynamics; one that might not – or is even unlikely – to be true in reality: that the system always remains in dynamical equilibrium. In other words, we have assumed that the Moran dynamics do not change, even as a PNH clone grows or shrinks in size. Considering the effect the disease can have on the body, in particular a chronic and severe loss of red blood cells and other hemocytes, one might however expect the hematopoietic system to react appropriately and attempt to mitigate this loss by increased cell production. This type of reaction from the bone marrow to a perturbation of blood cell counts is well known, having already been documented as much as 50 years ago [58]. Whether or not the HSC compartment would become involved in such a response is up for debate, as to my knowledge the literature provides no direct evidence for or against this idea in PNH. There are however some indirect clues which mark this as a credible possibility, in particular the well-documented response of the body to bone marrow transplantation or sublethal radiation of the bone marrow [55, 22, 16]. It has been shown that under such a severe disruption of the hematopoietic system, the response is an extremely rapid and highly clonal – i.e driven by a small group of HSCs – repopulation of the bone marrow before a return to normal hematopoiesis. If the disruption orchestrated by severe PNH were to cause a similar response, this would result in an increased rate of divisions of the HSC pool, which in turn would speed up the dynamics of the Moran system we have studied here.

While the following short discussion is not part of the research paper in which the

previous results of this chapter are presented, we will briefly look at how the Moran model might be extended to investigate the effects of a feedback loop where the size of the PNH clone directly influences the speed of the dynamics.

5.5.1. Feedback driven division rates

In order to capture a dependence of the stem cell dynamics on the severity of the disease, the main extension to perform is to have the Moran division rate λ no longer constant in time, but rather a function of some underlying process which causes it to speed up. Of course, we don't actually have any tangible knowledge of this underlying mechanism from which to build a model, other than the qualitative observation that a reduced number of functional hemocytes in the blood increases the cell production. Luckily this is enough to build a simple heuristic extension to test our intuition.

Starting from the basic premise that the size of a mutant *PIGA* clone on the level of the HSCs should correlate with the size of the GPI deficient population in the blood, and that the amount of GPI- cells in the blood directly determines the amount of hemolysis, we take the division rate as a function of the mutant clone size: $\lambda \to \lambda(m)$, with the requirement that λ must be an increasing function of m, since a larger PNH clone causes more hemolysis. Because we have no detailed knowledge of this dependence – quantitative or qualitative – it suffices to choose the simplest option, which is a linear function:

$$\lambda(m) = \lambda_0 + \alpha \frac{m}{N_{HSC}} \tag{5.5}$$

where λ_0 is the Moran division rate under normal hematopoiesis, and $\alpha > 0$ determines the strength of the coupling. Since the dependence on m is taken as a fraction of the total population m/N_{HSC}, α can be interpreted as the difference between the highest possible division rate at $m = N_{HSC}$ and λ_0; or in other words, the HSCs divide maximally at rate $\lambda_0 + \alpha$ when $m = N_{HSC}$.

While this extension to the model seems simple enough, it introduces a problem with respect to the method of solution used in this chapter. In the previous sections we evolved the probability of each state $m \in 0, 1, \ldots, N_{HSC}$ at discrete time increments

corresponding to the times at which divisions occurred. But this discrete evolution was only possible thanks to the fact that the division rate was the same for each state, allowing for all states to be evolved simultaneously. This is of course no longer the case here, given that λ is now a function of m. Fortunately there is a simple solution to this, which is to apply the continuous-time Markov chain approach introduced in Section 4.3.3. Using equation (4.5) we can write for the time evolution of any state m:

$$\frac{1}{\lambda(m)}\frac{\mathrm{d}P_m(t)}{\mathrm{d}t} = p_{m-1,m}P_{m-1}(t) - 2p_{m,m}P_m(t) + p_{m+1,m}P_{m+1}(t) \tag{5.6}$$

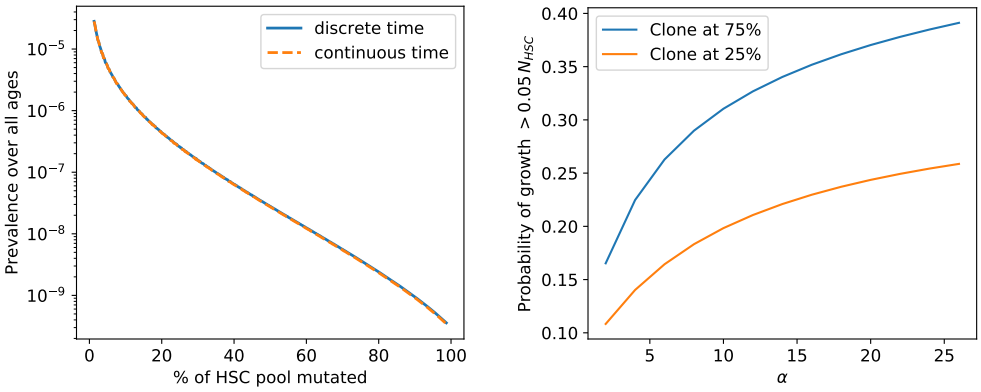
where the $p_{n,m}$ are given by (5.1) or (5.2), depending on the type of division (with or without concomitant differentiation). While this set of differential equations for the P_m may not have an easily obtainable solution, we can still evolve it numerically in a similar manner to the discrete-time system, though now using an appropriate ODE solving algorithm.

5.5.2. Heuristic results

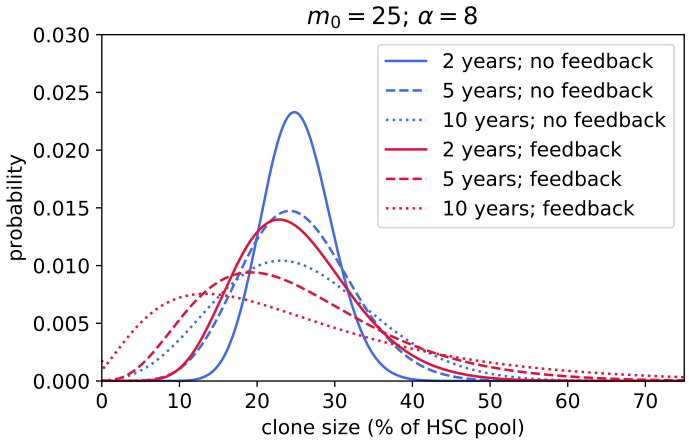
Before investigating how the coupling affects the system, it is worthwhile to first check whether the continuous-time formulation indeed gives the same results as the discrete-time model used before. A comparison of the predicted population prevalence of PNH calculated from both methods is shown in Figure 5.7a, and given their nearly identical result should reassure us that the move to continuous-time is valid.

Introducing the feedback through (5.5), we once again look at the change of an established clone over time, as shown in Figure 5.7c. We observe that compared to using a fixed division rate the feedback driven rate exaggerates both the speed at which the diffusion occurs, as well as the skewness of the distribution, even though the martingale requirement of a constant expectation value remains fulfilled. Thus, the single new parameter α – essentially representing the strength of the coupling between the HSC compartment and the severity of the PNH driven anemia – determines the speed with which expansion occurs. With this in mind we can revisit the data of Araten et al. [8], who reported 1/3 of patients experiencing a growth of $> 5\%$ of the total HSC pool within

one year. Whereas in the feedback free model we predicted this fraction to be between 1/20 and 1/10 of patients, we can now vary α to see how this prediction changes for stronger couplings, as shown in Figure 5.7b. The fact remains that the rate of growth (or decline) depends on the initial size of the clone; firstly because of the size dependent transition rate given in (5.4), which assigns the highest rate of change to the state $m = N_{HSC}/2$, and now secondly because of the size dependent division rate $\lambda(m)$, which is maximal at $m = N_{HSC}$. We could in theory fit α to the value (fraction of patients with growth $> 0.05N_{HSC}$) reported by Araten et al. by calculating the predicted fraction for every clone size, and extracting an average value by weighing these with the expected clone size in the population under the feedback model. However, given the small sample size in the report, the high standard deviations with respect to clone sizes, and the handwaving approach with which we constructed the coupling, there would little interpretive weight to this particular value. On the other hand, from simply eyeballing the results shown for the two initial clone sizes shown in Figure 5.7b, we can already see that it would take an α on the order of ~ 20 to obtain a value similar to what Araten et al. reported. Thus, it would require the HSCs to increase their Moran divisions (each equivalent to two divisions, one with and one without differentiation) from 1 up to a maximum of around 20 times a year, which is not unimaginable, given that they divide much faster during early ontogenic growth or bone marrow reconstitution [16, 22].



(a) *(b)*



(c)

Figure 5.7.: Feedback driven division rates can increase the speed of clonal expansion. ***(a)*** *Comparison of discrete time and continuous time Markov chain evolution for the prevalence within th U.S. population.* ***(b)*** *Probability of a clone expanding for an increase greater than 5% of the HSC pool within 1 year, calculated for initial clone sizes 25% and 75%.* ***(c)*** *Probability of clone size at different time points after initiation at 25%, for constant division rate (no feedback) and for linearly coupled division rate with $\alpha = 8$ according to (5.5).*

5.6. Conclusion

In this chapter we have applied the simple Moran model of HSC dynamics to the investigation of clonal expansion in the rare blood disorder paroxysmal nocturnal hemoglobinuria. While this disease is known to occur due to a debilitating mutation of the *PIGA* gene in the HSC pool, the method of expansion of this clone remains debated, in the past typically assumed to occur through some selective advantage of the variant population. Here, through a simple extension of Moran model introduced in Chapter 4, we showed that the simplest explanation – neutral drift – need not be discounted, as under its assumptions we can accurately predict a number of statistical quantities related to the disease's prevalence in a population. Furthermore, two clinical observations requiring complicated explanations in a selection picture arise naturally from the neutral drift hypothesis: the fact that the average variation of clone size in a cohort of individuals is 0, and the occurrence of spontaneous remission in a significant fraction of patients. While some clinical observations – such as the disease prevalence in an age distributed population – were predicted surprisingly well, others – such as average clone sizes and expansion rates – were underestimated by the model. To understand whether these dissimilarities occur due to a flaw in the neutral drift hypothesis or could be attributed to some component of the true system which was not incorporated in the model, we investigated the possible influence of varying division rates, driven by a bodily response to the anemia occurring alongside a significantly large PNH clone. We show that the existence of such a coupling between HSCs and the blood – to some extent hinted at by the observed behavior after bone marrow transplantation [55, 22, 16] – would indeed likely imply a expediting of the clonal dynamics, leading to larger clone sizes and expansion rates.

6. Subclonal dynamics in hematopoietic stem cells

The shapes are always changing. Changing is their normal state, like us. Even if we're not changing on the outside, we're changing on the inside constantly.
— Jake The Dog, *Adventure Time*

In the previous chapter we focused on the stochastic size fluctuations of a single clone in the HSC pool, defined by a particular mutation which sets its constituent cells apart from the rest. However, it was already hinted at that in a more detailed picture even this collection of mutants would be heterogeneous with respect to the entire genome, given that new somatic mutations are arising constantly [87]. In fact, every time a new mutation occurs somewhere in the HSC pool, one can consider this a new clone with the ability to expand and in turn acquire its own subclones. The result is a complex mosaic of thousands of variants randomly distributed over the stem cells and occurring at random positions in the genome [93], some appearing in large groups of cells, others only in a single cell. Most of these mutations are not particularly adverse, being synonymous or appearing in non-coding or otherwise unimportant regions; however, in situations where they are observable, they may provide us with an opportunity to learn about the stochastic dynamics underlying their existence.

In this chapter we will expand our model of HSC dynamics with the goal of applying it to this more detailed picture of many competing clones. Such a model will allow us to characterize the types of stochastic fluctuations we expect to see and predict

how they might change as an individual ages. To this end we will first formalize what kind of clonal relationships arise within a cell population and how they are observed in sampled data (Section 6.1). From here it will become clear that in order to model this extended picture of clonal competition the asymmetric cell divisions can no longer be neglected, and as such we will extend the Moran model introduced in Chapter 4 to account for these (Section 6.2). This extended formalism will be used to derive analytical expressions for the dynamics of two quantities which can be measured in a sampled dataset: the *single cell mutational burden* (Section 6.4) and the *variant allele frequency* (VAF) (Section 6.5), both of which are effective (but different) reductions of the full state space of a multiclonal cell population. Finally, in Section 6.7 we will compare the resulting predictions of both the mutational burden and the VAF to a recent dataset containing information on the variants in a single human individual [78], which will allow us ascertain difficult to measure values of the fundamental parameters involved in the stochastic dynamics, such as the number of actively proliferating HSCs, their individual self-renewal and differentiation rates, and the balance of symmetric versus asymmetric divisions.

6.1. Clonality

When using the term *clonality* we are referring to the relatedness of cells through their genome. While a strict definition of the term *clone* would mean a group of cells with completely identical genomes, we have so far used it (as is fairly customary in the field) to also refer to cells which share a specific variant in their genome, such as for example the mutated *PIGA* gene discussed in Chapter 5. Given that the picture of clonality we are interested in here is somewhat more complicated than before, it will prove useful to first formalize the various types of relatedness we expect to model.

Consider the simple ancestral tree shown in Figure 6.1a, depicting the mutant accumulation of three generations of cell divisions from a single ancestral cell to its progeny of eight great-granddaughter cells. If we were to observe this population during the last generation we would only see the final eight cells, however their history is to an extent

encoded in the pattern of mutations carried by each them. For example, two cells which share a variant must have a common ancestor which they do not have in common with a cell that does not have this variant. In a similar vein, a variant that is shared by very few cells is likely to have occurred more recently than one that is shared by many cells. There are various existing techniques – based on finding binary conditions for the order of mutational events – for reconstructing such an ancestral tree from a single time point observation [20, 78, 120, 29] such as the (albeit extremely simple) example given in Figure 6.1a. While impressive as accomplishments, such reconstructions do not immediately provide quantitative information on the fundamental processes (i.e. the mutations and cell divisions) driving the system, and while they can in turn be analyzed with various statistical techniques, the small sample size from which they are typically constructed often results in limited inferences. Here, we will instead attempt to derive predictions – based on our assumptions for HSC behavior – for simpler quantities related to the pattern of mutations, such as for example the total number of variants per cell, as these can directly be compared with the data to infer quantitative aspects of the model.

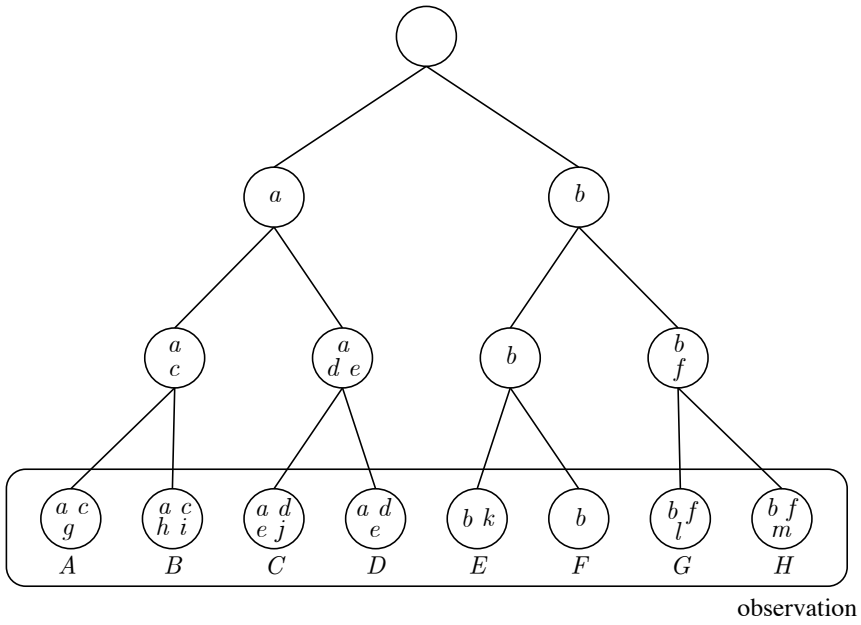
To this end let us first formalize the observation in the final step of Figure 6.1a. We might denote each of the cells in the population as a set $\mathcal{C}_i$ ($i \in 1, 2, \ldots, 8$) with the respective variants they carry as their elements. This picture is a bit messy, since some variants appear in multiple cells, and the various sets are thus partially overlapping. But we can also invert this description, as shown in Figure 6.1b, by taking the cells themselves as elements C_i of the sets $\mathcal{V}_j$ ($j \in 1, 2, \ldots, 14$) given by the variants. An interesting thing to note is that in this view the variants can be subsets of one another $\mathcal{V}_i \subset \mathcal{V}_j$ – which we might call a *subclonal relationship* – but can never be partially overlapping, i.e. $\mathcal{V}_i \cap \mathcal{V}_j$ must be either $\mathcal{V}_i$, $\mathcal{V}_j$, or $\emptyset$. This follows from the fact that newborn cells can only appear within an existing clone, and new clones only occur with new cells.

With the goal of predicting such clonality and comparing it to real world data in mind, we ask ourselves what state spaces (as defined in Section 3.2.1) can be defined on a system of the form in Figure 6.1. Considering we are not really interested in what part of the

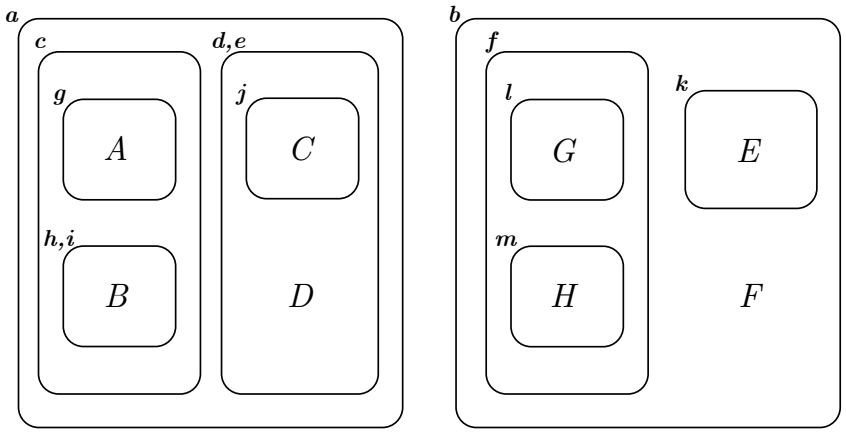
genome is affected by each mutation but only that different variants are distinct, a space which contains all information of the system on observation would cover the possible numbers of clones, the numbers of cells contained by each clone, and the relationship between clones, i.e. whether $\mathcal{V}_i \subset \mathcal{V}_j$, $\mathcal{V}_j \subset \mathcal{V}_i$, or $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$. For a large population this would be a very high dimensional space, given the possible combinations of clonal relationships across all variants. An example of this would be the space of all possible ancestral trees such as the one shown in Figure 6.1a, though as hinted at earlier, its high dimensionality makes it difficult to analyze quantitatively. Instead, we will forego some information contained in the observation in order to construct simpler state spaces, which we will see can already be applied effectively to extract information from real world data. In particular, we will focus on two measurements which can be performed on observations such as that of Figure 6.1a. The first is the number of mutations found in a cell, which we refer to as the *single cell mutational burden*. From the example it should be clear that as a stochastic quantity this can vary between cells, depending on their divisional and mutational past. This variation will end up being quite useful, as we will see that from the model a predicted probability distribution for the burden can be derived (Section 6.4) and effectively compared with even a modestly sized dataset (Section 6.7). The second is the *variant allele frequency* – essentially the histogram of all the variant sizes in the population – for which we will also derive a prediction of its time evolution in Section 6.5. Because these two quantities are different reductions of the more "information-rich" ancestral tree, we can use them both in tandem to extract information from a dataset, as illustrated in Section 6.7.

6.2. Moran model with asymmetric divisions

While in the previous chapters we have neglected any effects of asymmetric HSC divisions, it will at this point become important to include them in our treatment. The reason for ignoring the asymmetric division previously was that it has no effect on the size of the clone which the dividing cell belongs to: while one of the daughter cells is removed from the pool the other simply maintains its HSC characteristics, meaning any



(a)



(b)

Figure 6.1.: Mutation accumulation in a dividing cell population. ***(a)*** *The divisional history of a small population of cells shown as a lineage tree. Cells are depicted by circles, while mother-daughter relationships are shown through connecting lines. Random mutations (letters a-m) can be acquired during cell divisions and are shown within the cells by which they are carried. The observable cells after three generations are denoted by the letters A-H.* ***(b)*** *The observable system shown as a collection of sets associated with the distinct clones a-m to which the cells A-H belong.*

variants which it might carry are unchanged in size. However, since we are now interested in accounting for the arrival of any new variants – no matter their location in the genome – the asymmetric division will play a much greater role: even if existing clones do not change in size due to such an occurrence, the daughter cell that remains in the HSC pool still has a probability of acquiring new mutations, which in turn adds new clones to the system.

One important question that arises then is whether the mutation rate is the same for the asymmetric division as for the symmetric case. Given that during normal cell divisions both the original and copied DNA strands are distributed randomly among the two daughter cells [2], a naive guess would be to take the same rate per daughter cell. However, the immortal strand hypothesis – for which some evidence exists [111, 141] – posits that HSC cell divisions include a mechanism for conserving the original strand within the non-differentiating daughter cell. In our model, this would imply a lower mutation rate for variants arising during asymmetric divisions. Given that this theory currently remains a subject of some debate [134], we will take the simplest approach and assume the (per daughter) mutation rates are equal in symmetric and asymmetric divisions, though it is worth keeping in mind that this can be improved upon in the future, when a clearer picture of asymmetric HSC divisions exists. On the other hand, there is no reason to assume the *rates* at which these divisions occur are in any way related, so that we will need to introduce a new parameter to denote the asymmetric division rate.

To avoid confusion let us briefly restate the mechanics of our now extended Moran model to include asymmetric divisions. In Section 4.2 we introduced the possible events occurring for a single cell in the HSC pool as *self-renewal* (symmetric division where both daughter cells remain HSCs), *symmetric differentiation*, and *asymmetric division* (where we ignored cell death or senescence). Recall that in the Moran model we take every self-renewal to occur simultaneously with a symmetric differentiation, and moving to a real time picture we have opted to take these Moran events – one self-renewal and one symmetric differentiation – as Poisson distributed, so that they occur with fixed a

probability rate ρ per cell in time. Furthermore, adding asymmetric divisions we include another Poisson process occurring with a different rate ϕ per cell. From Section 3.1.2 we know that we can restate the two independent occurrences as a single Poisson process coupled with a probability, so that alternatively we can introduce $\lambda = \rho + \phi$ as the total rate of events, and p the probability of an event being an asymmetric division, i.e.:

$$\begin{cases} \rho = \lambda(1-p) \\ \phi = \lambda p \end{cases} \tag{6.1}$$

Note that λ is not to be confused with a "total division rate" which, as discussed in Section 4.3.3, does not exist for this model in the sense of a Poisson process. Still, we may also wish to think in terms of the average number of *total* divisions – i.e. of any kind – in a unit time. This takes the form $\tilde{\lambda} \equiv 2\rho + \phi$ (since each symmetric Moran event contains two divisions), though one should keep in mind that it does not describe the rate of a Poisson process for divisions.

Depending on the context either the (ρ, ϕ) or the (λ, p) notation might be more convenient, and both will be used interchangeably throughout this chapter. Finally, since all cells act independently, we may state that in the total population simultaneous symmetric self-renewals and differentiations occur at rate $N\rho = N\lambda(1-p)$ and asymmetric divisions at rate $N\phi = N\lambda p$. For reference, the parameters associated with this model are given in Table 6.1.

6.3. Testing with simulations

While the main goal of this chapter is to derive expressions for the dynamics of certain quantities related to the clonality of the HSC pool, it will prove useful to have a mechanism for testing the results obtained here. To this end we will make use of a direct simulation of the Moran based dynamics described above – developed by Marius Möller (Queen Mary University of London, School of Mathematical Sciences) – that performs a Gillespie style algorithm to stochastically evolve a population of individual HSCs and the variants they accumulate. The specifics of the simulation are described in Appendix

parameter name	description
N	Total number of cells in the HSC pool
ρ	Rate of self-renewal/differentiation events per unit time
ϕ	Rate of asymmetric divisions per unit time
λ	Total rate of Poisson events per unit time
p	Probability of event being an asymmetric division
μ	Rate of mutation accumulation per daughter cell per division

Table 6.1.: *Parameter names and description used in extended Moran model.*

A.2. A result of this simulation can thus be considered to be a single possible trajectory of a system which obeys the previously described dynamics exactly. In particular, it allows for observations of the form exemplified in Figure 6.1, so that we may run such ensembles of simulations to test the validity of the results obtained hereafter.

6.4. The single cell mutational burden

Mutations accumulate within the HSC pool as time passes and cell divisions continuously occur. As such we expect the mutational burden (the number of mutations present) to increase, both in the entire population as well as in each cell separately. However, due to the stochasticity in divisional events and mutation occurrence, the number of mutations found in each cell may vary, as can be seen from the example in Figure 6.1a. Still, we can attempt to find the probability distribution for the burden in a single cell using the assumptions made in Section 6.2.

6.4.1. Mutational burden as a compound Poisson process

Consider a single HSC carrying m_c variants in an adult individual. If we were to somehow "rewind the tape" and watch this particular cell's life in reverse, we would witness it

undergoing a great number of cell divisions – during each of which a random number of mutations occurred – all the way back to its ancestral zygote that contained the pristine reference genome. Now if we somehow knew the number of divisions y_c this cell underwent as well as the number of mutations x_i that occurred during each division, we could write the mutational burden as the sum:

$$m_c = \sum_{i=1}^{y_c} x_i \tag{6.2}$$

In our stochastic picture both y_c and x_i are random variables, so that the distribution of m_c will depend on the distributions of y_c and x_i. In Section 4.2 we have already argued for taking both as Poisson distributed, and thus m_c follows a distribution known as the *compound Poisson*, which is a sum over random variables whereby the number of terms in the sum is Poisson distributed. It can be shown (see Appendix A.3) that the mean and variance of this distribution are given by:

$$\mathrm{E}(m_c) = \mathrm{E}(y_c)\mathrm{E}(x_i) \tag{6.3}$$

$$\mathrm{Var}(m_c) = \mathrm{E}(y_c)\mathrm{E}(x_i^2) \tag{6.4}$$

Whereby we have already seen that for a Poisson distribution the mean and variance are both given by the process' rate ((3.17) and (3.18)). In the previous section we argued for taking identical per daughter cell mutation rates for symmetric and asymmetric divisions, and thus all x_i are Poisson distributed with fixed rate μ. To find the number of divisions y_c undergone by the cell we can sum in (6.2) into two parts:

$$m_c = \sum_{i=1}^{s_c} x_i + \sum_{j=1}^{a_c} x_j \tag{6.5}$$

where s_c and a_c are the number of self-renewal and asymmetric divisions which occurred in the cells past. While from Section 6.2 we know a_c is Poisson distributed with rate ϕ, we might naively take $s_c \sim \mathrm{Pois}(\rho t)$ as well. However, since a symmetric division introduces two new HSCs, the total increase in mutational burden in the population comes from two $x_i \sim \mathrm{Pois}(\mu)$, which on average results in an *effective* rate for s_c of 2ρ. Since we may combine the two Poisson processes (see Section 3.1.2) we have $y_c \sim \mathrm{Pois}([2\rho + \phi]t)$,

and can write the mean and variance as

$$\mathrm{E}(m_c) = (2\rho + \phi)t\mu \tag{6.6}$$

$$\mathrm{Var}(m_c) = (2\rho + \phi)t(\mu + \mu^2) \tag{6.7}$$

Obtaining an analytical form of the probability distribution of (6.2) is difficult, however, its moments are easily calculated, and if the rates ρ, ϕ, and μ are known it can also be sampled. On the other hand, given that the manner in which we arrived at the effective division rate was not entirely rigorous, let us for completeness take another approach to obtain the probability distribution by evolving a Markov chain.

6.4.2. Markov chain approach

The state space we are interested in evolving is the histogram of single cell mutational burdens. We thus define the function:

$$\begin{aligned} \tilde{n} : (\mathcal{M} \subset \mathbb{N}) \times (\mathcal{T} \subset \mathbb{R}) &\to \left(\tilde{\mathcal{N}} = 0, 1, \ldots, N \subset \mathbb{N}\right) \\ m, t &\mapsto \tilde{n}_m(t) \end{aligned} \tag{6.8}$$

which maps the number of mutations m at a time t to the number of cells $\tilde{n}$ in the population with m mutations; we can think of it as the histogram of mutational burdens of all individual cells. If we know its shape at a time t_0, for example $\tilde{n}(m) = 0, \ \forall m \in \mathcal{M}$, given a set of transition rates we can evolve it in time. According to our model, the occurrence of self-renewal/differentiation evens at rate ρ and asymmetric divisions at rate ϕ are what change the state of the system. For a single event, each has the following effect:

(i) Symmetric self-renewal and differentiation

- The differentiating cell with m_a mutations is removed and thus decreases $\tilde{n}_{m_a}$ by 1:

  $\tilde{n}_{m_a} \to \tilde{n}_{m_a} - 1$, for one $m_a \in \mathcal{M}$ selected with probability $q_{m_a} = \tilde{n}_{m_a}/N$.
- The self-renewing cell with m_b mutations is removed and thus decrease $\tilde{n}_{m_b}$ by 1:

  $\tilde{n}_{m_b} \to \tilde{n}_{m_b} - 1$, for one $m_b \in \mathcal{M}$ selected with probability $q_{m_b} = \tilde{n}_{m_b}/N$.

- A new daughter cell with u new mutations is added to $\tilde{n}_{m_b+u}$:
  $\tilde{n}_{m_b+u} \to \tilde{n}_{m_b+u} + 1$, for one $u \in \mathbb{N}$ selected with probability $p_u = \text{Pois}(u;\mu)$
- A new daughter cell with v new mutations is added to $\tilde{n}_{m_b+v}$:
  $\tilde{n}_{m_b+v} \to \tilde{n}_{m_b+v} + 1$, for one $v \in \mathbb{N}$ selected with probability $p_v = \text{Pois}(v;\mu)$

(ii) Asymmetric division

- The dividing cell with m_a mutations is removed and thus decreases $\tilde{n}_{m_a}$ by 1:
  $\tilde{n}_{m_a} \to \tilde{n}_{m_a} - 1$, for one $m_a \in \mathcal{M}$ selected with probability $q_{m_a} = \tilde{n}_{m_a}/N$.
- A single daughter cell with u new mutations is added to $\tilde{n}_{m_b+u}$:
  $\tilde{n}_{m_b+u} \to \tilde{n}_{m_b+u} + 1$, for one $u \in \mathbb{N}$ selected with probability $p_u = \text{Pois}(u;\mu)$

While these changes might be used to find the transition probability to any possible state, this problem is high dimensional, since $\tilde{n}_m$ varies probabilistically in $\tilde{\mathcal{N}} = 0, 1, \ldots, N$ for all possible $m \in \mathbb{N}$. Obtaining the probability distribution for m_c is much simpler though, as it should be realized by the normalized histogram $\tilde{n}_m$ in the limit of infinite cells. Thus we introduce the function $n_m(t)$:

$$\begin{aligned} n : (\mathcal{M} \subset \mathbb{N}) \times (\mathcal{T} \subset \mathbb{R}) &\to (\mathcal{N} = [0,1] \subset \mathbb{R}) \\ m, t &\mapsto n_m(t) \end{aligned} \tag{6.9}$$

which is similar to $\tilde{n}$ except now the image is in $\mathbb{R}$, since the normalized histogram is not an integer function. In this limit of infinite cells $n_m(t)$ can be interpreted as the average over all cells in $\tilde{n}_m(t)$ for burden m, and the probabilities q_m and p_k as the fractions of cells undergoing their respective events. Thus evolving $n_m(t)$ we have for each division event:

(i) Symmetric self-renewal and differentiation

- differentiation: $n_m \to n_m - q_m : \quad \forall m \in \mathcal{M}$
- self-renewal: $n_m \to n_m - q_m : \quad \forall m \in \mathcal{M}$
- daughter cells: $n_m \to n_m + 2q_k\, p_{m-k} : \quad \forall k \in 0, 1, 2, \ldots, m$

(ii) Asymmetric division

- differentiation: $n_m \to n_m - q_m : \quad \forall m \in \mathcal{M}$
- daughter cell: $n_m \to n_m + q_k\, p_{m-k} : \quad \forall k \in 0, 1, 2, \ldots, m$

These changes can be summarized in a master equation, which using the rates ρ and ϕ, and $q_m = n_m/N$ and $p_k = \text{Pois}(k;\mu)$, becomes

$$\frac{N}{2\rho+\phi}\frac{\mathrm{d}n_m(t)}{\mathrm{d}t} = -n_m(t) + \sum_{k=0}^{m} n_k(t)\text{Pois}(m-k;\mu) \tag{6.10}$$

While this differential equation can be solved numerically to find the distribution at a future time, the sum on the right hand side can become extremely long. However, if the mutation rate μ is low, the Poisson distribution $\text{Pois}(m-k;\mu)$ goes to zero quickly, so that in practice only a small number of terms need to be taken into account.

6.4.3. Discussion: single cell mutatational burden

To test the results of Sections 6.4.1 and 6.4.2 we can compare them with direct single cell simulations of the Moran model, as discussed in Section 6.3. Since we do not have the exact probabilities for the compound Poisson, we instead sample a large number of values from it – by sampling from the Poisson distributions for y_c and x_i in (6.2) – as its histogram converges to the true probability distribution. A comparison of the three methods is shown in Figure 6.2, which confirms they all produce the same result. Given the symmetry of the resulting distribution, at first glance it may appear to have a shape similar to that of a Poisson distribution, however we can see from (6.6) and (6.7) that this is not the case: the Poisson distribution has equal mean and variance, whereas the variance we have derived for the burden is a factor $1 + \mu^2$ higher. Furthermore, it is worth noting that burden distribution does not depend on the size of the population N. In fact, for a specified time t, it is entirely determined by the mutation rate per cell division μ and the quantity $2\rho + \phi$, which happens to be the total division rate $\tilde{\lambda}$ introduced earlier.

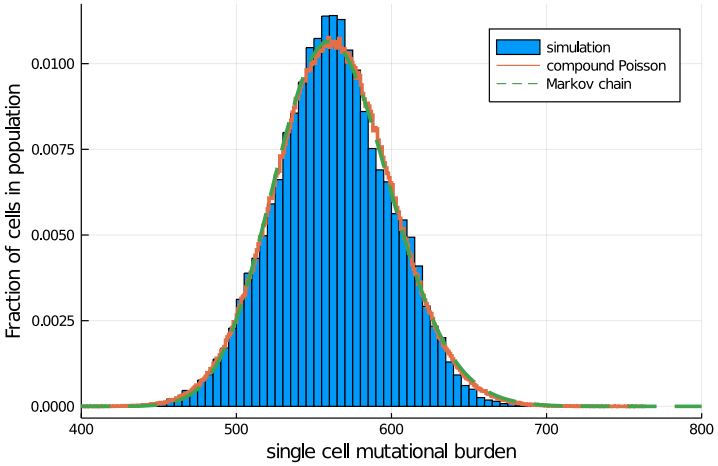


Figure 6.2.: *Distribution of the single cell mutational burden for a HSC population with parameters* $N = 500$, $\rho = 2.5$, $\phi = 2.5$, $\mu = 1.5$, *and* $t = 50$. *The simulation histogram is the average of 100 simulations, the compound Poisson distribution is found by sampling equation (6.2)* 10^6 *times, and the Markov chain solution is found from evolving (6.10).*

6.5. The variant allele frequency spectrum (VAF)

We have seen that the single cell mutational burden contains useful information about the system dynamics such as the mutation and division rates. However, as previously discussed, it is a reduction of the complete state of the system and therefore does not use all of the information encoded in an observation. We now turn to another measurable quantity that is often used to characterize a population's clonality: the *variant allele frequency* (abbreviated as VAF), which describes the number of variants having a particular frequency in the population:

$$\begin{aligned} \tilde{v}: \quad & \mathcal{F} = 1/N, 2/N, \ldots, 1 \subset \mathbb{Q} \quad \rightarrow \quad \tilde{\mathcal{V}} = 0, 1, 2, \ldots, V \subset \mathbb{N} \\ & f \mapsto \tilde{v}_f \end{aligned} \tag{6.11}$$

Where V is the total number of variants in the population. Take as an example the system in Figure 6.1, from which the VAF can be obtained by counting the number of cells found in each of the 13 clones, and then distributing the clones into a histogram where each bin conforms to a fraction of cells in the population: $\tilde{v}_{1/8} = 7$, the number

of variants (g, h, i, j, k, l, m) containing only one cell, $\tilde{v}_{2/8} = 4$, the number of variants (c, d, e, f) containing two cells, and so forth.

6.5.1. Dynamics of the VAF expected value

The state space of the VAF is again of high dimensionality: for each frequency $f \in \mathcal{F}$ there are V available values, meaning that a single state in the complete space conforms to one possible combination of values for every f. Because of this a probability distribution for the entire space is both difficult to construct and unwieldy to analyze. However, we can attempt to find the expected value of $\tilde{v}_f$ for each frequency at a specified time t – which takes the form of a single "average" VAF – by constructing a (continuous-time) Markov chain in a similar manner as we did for the single cell mutational burden in Section 6.4.2). To this end, we introduce $v_f(t) = \langle \tilde{v}_f \rangle$ at time t:

$$\begin{aligned} v: \quad & (\mathcal{F} = 1/N, 2/N, \ldots, 1 \subset \mathbb{Q}) \times (\mathcal{T} \subset \mathbb{R}) \to \mathcal{V} \subset \mathbb{R}; \\ & f, t \mapsto v_f(t) \end{aligned} \tag{6.12}$$

as the expected VAF at time t. To obtain the dynamics of $v_f(t)$, we note that we have already derived the dynamics of a single clone in Section 4.3.2. Given that only the symmetric self-renewal/differentiation events influence the sizes of existing clones, the probability $P_k(t)$ of a single clone having size k varies according to (4.5). Now if we take a population in which a large number of clones exist, if the clones evolve independently (we will discuss this supposition more in detail later) we can interpret the transition probability of a single clone going from size k to l as the expected fraction of clones at k going to l. In other words, the infinitesimal-time dynamics of $v_{k/N}(t)$ are given by the transitions $p_k v_{k/N}(t)$. However, we must still account for the arrival of new clones in the system. Specifically, each self-renewal event introduces on average 2μ new clones at state $f = 1/N$, while each asymmetric division introduces an expected μ variants, so that in time we expect the state $f = 1/N$ to increase at rate $N(2\rho + \phi)\mu$ on top of the

existing clonal dynamics. Thus, finally we obtain:

$$\begin{cases} \dfrac{\mathrm{d}v_{m/N}(t)}{\mathrm{d}t} = N\rho\Big[p_{m-1}v_{m-1/N}(t) - 2p_m v_{m/N}(t) + p_{m+1}v_{m+1/N}(t)\Big] \\ \dfrac{\mathrm{d}v_{1/N}(t)}{\mathrm{d}t} = N\rho\Big[p_0 v_0(t) - 2p_{1/N}v_{1/N}(t) + p_{2/N}v_{2/N}(t)\Big] + N(2\rho+\phi)\mu \end{cases} \tag{6.13}$$

where

$$p_m = \frac{m}{N}\Big(1 - \frac{m}{N}\Big) \tag{6.14}$$

Diffusion approximation

As discussed at the end of Chapter 4, such a set of differential equations can become unwieldy to solve, even numerically, if the population size N is large. Fortunately, we may resort to the diffusion approximation to recast the system as a partial differential equation in a continuous state space $\mathcal{X} = [0,1] \in \mathbb{R}$. We introduce the function

$$v : \mathcal{X} \times \mathcal{T} \mapsto \mathcal{V}; \quad x, t \mapsto v(x,t) \tag{6.15}$$

which represents the VAF in the continuous frequency space picture. We have seen that the dynamics of a single clone are given by (4.18), though we must now also include the incoming flux of clones. To this end the source term in 6.13 can be replaced by a Dirac delta function, so that we obtain:

$$\frac{\partial v(x,t)}{\partial t} = \frac{\rho}{N}\frac{\partial^2[x(1-x)v(x,t)]}{\partial x^2} + \delta(x - N^{-1})N(2\rho+\phi)\mu \tag{6.16}$$

While we have no analytical solution for this expression, numerical approximation techniques can be applied, which prove to be much faster than solving (6.13).

6.5.2. Dynamics of the VAF variance

While we have obtained an expression for the time evolution of the expected value of the VAF, it would be useful to have information on how a single experiment might deviate from $v_f(t)$. In particular, we would like to know the variance of the VAF for each state $f \in \mathcal{F}$ at a future time, which we will denote as $r_f(t)$. This is not easily done from the probabilities dynamics of Section 4.3.3 which we used to derive $n_m(t)$ and $v_f(t)$, and so

we will take a different approach here. Unfortunately this attempt will ultimately fail, however, the reason for its failure will in itself provide an interesting perspective, which is why it is included here.

Denote the exact number of variants in the system with frequency f (or size $k = Nf$) at time t by $V_f(t)$ (the VAF at t is thus given by the $V_f(t)$ for all $f \in \mathcal{F}$). We might write this as an integral over time:

$$V_f(t) = \int_{\tau=0}^{t} dY_f(t-\tau) \tag{6.17}$$

where $dY_f(t-\tau)$ is the number of variants which appeared in the time span $[\tau, \tau + d\tau]$ and have size f at t. This quantity is in itself another sum which we can write as

$$dY_f(t-\tau) = \sum_{i=1}^{M(d\tau)} F_i(t-\tau) \tag{6.18}$$

where the sum goes over all variants i that arose in $[\tau, \tau + d\tau]$ – the total number of which is in itself a random variable given by $M(d\tau)$ – and where $F_i(t-\tau) = 1$ if that variant has size f at t and $F_i(t-\tau) = 0$ otherwise. Or more concretely, we have:

$$F(t-\tau) = \begin{cases} 1 & \text{probability} \quad P_f(t-\tau) \\ 0 & \text{probability} \quad 1 - P_f(t-\tau) \end{cases} \tag{6.19}$$

which is a Bernoulli distribution (with $P_f(t)$ the probability found from evolving the continuous time Moran Markov chain 4.5), and also

$$M(d\tau) = \sum_{i=1}^{Q} m_i \tag{6.20}$$

with the number of divisions occurring in $d\tau$ given by $Q \sim \text{Pois}(N(2\rho + \phi)d\tau)$ and the number of mutations per division by $m_i \sim \text{Pois}(\mu)$. This is a compound Poisson distribution, which we already know from (6.3) can be written as

$$\begin{cases} \text{E}(M(d\tau)) = N(2\rho + \phi)\mu \, d\tau \\ \text{Var}(M(d\tau)) = N(2\rho + \phi)(\mu + \mu^2) \, d\tau \end{cases} \tag{6.21}$$

As a test of this derivation we might first look at $v_f(t)$, the expectation value of $V_f(t)$, since we have already derived this in another manner earlier. For the expectation value we may write

$$v_f(t) = \mathrm{E}(V_f(t)) = \int_{\tau=0}^{t} \mathrm{E}(dY_f(t-\tau)) \tag{6.22}$$

Using the law of total expectation we have for the sum of expected values:

$$\mathrm{E}(dY_f(t-\tau)) = \mathrm{E}(M(d\tau))\mathrm{E}(F(t-\tau)) \tag{6.23}$$

with the same holding for the compound Poisson:

$$\mathrm{E}(M(d\tau)) = N(2\rho+\phi)\mu d\tau \tag{6.24}$$

so that

$$\mathrm{E}(dY_f(t-\tau)) = N(2\rho+\phi)\mu P_f(t-\tau)\, d\tau \tag{6.25}$$

and thus finally

$$v_f(t) = \mathrm{E}(V_f(t)) = N(2\rho+\phi)\mu \int_0^t P_f(t-\tau)\, d\tau \tag{6.26}$$

$$= N(2\rho+\phi)\mu \int_0^t P_f(\tau)\, d\tau \tag{6.27}$$

In order to calculate this expected value we take its derivative to obtain an expression which we may evolve numerically:

$$\frac{\mathrm{d}v_f(t)}{\mathrm{d}t} = N(2\rho+\phi)\mu\, P_f(t) \tag{6.28}$$

We may test that this gives the same result as (6.13), which is shown in Figure 6.3. However it is clear this expression is computationally more expensive, as it requires the simultaneous evolution of both $v_f(t)$ as well as $P_f(t)$.

We may now apply the same approach for the variance, however, taking the variance inside the integral requires the additional assumption of independence of the $dY_f(t)$:

$$v_f(t) = \mathrm{Var}(V_f(t)) = \int_{\tau=0}^{t} \mathrm{Var}(dY_f(t-\tau)) \tag{6.29}$$

To find the integrand we use the law of total variance to write:

$$\begin{aligned}
\mathrm{Var}(dY_f(t-\tau)) &= \mathrm{Var}\left(\sum_{i=1}^{M} F_i(t-\tau)\right) && (6.30)\\
&= \mathrm{E}(\mathrm{Var}[dY_f|M]) + \mathrm{Var}(\mathrm{E}[dY_f|M]) && (6.31)\\
&= \mathrm{E}(M\,\mathrm{Var}[F]) + \mathrm{Var}(M\,\mathrm{E}[F]) && (6.32)\\
&= \mathrm{E}(M)\mathrm{Var}(F) + (\mathrm{E}[F])^2\mathrm{Var}(M) && (6.33)
\end{aligned}$$

Plugging in the variance and expectation values of M (compound Poisson distribution) and F (Bernoulli distribution) this becomes:

$$\begin{aligned}
\mathrm{Var}(dY_f(t-\tau)) &= N(2\rho+\phi)\mu\big[P_f(t-\tau) - P_f^2(t-\tau)\big]\,d\tau\\
&\quad + [N(2\rho+\phi)](\mu+\mu^2)P_f^2(t-\tau)\,d\tau\\
&= N(2\rho+\phi)\big[\mu P_f(t-\tau) + \mu^2 P_f^2(t-\tau)\big]\,d\tau
\end{aligned} \qquad (6.34)$$

We can now once again write a differential form to evolve $r_f(t)$:

$$\begin{aligned}
\frac{\mathrm{d}r_f(t)}{\mathrm{d}t} &= N(2\rho+\phi)\mu\,\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^t P_f(t-\tau)\,d\tau + \mu\int_0^t P_f^2(t-\tau)\,d\tau\right] && (6.35)\\
&= N(2\rho+\phi)\mu\,\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^t P_f(\tau)\,d\tau + \mu\int_0^t P_f^2(\tau)\,d\tau\right] && (6.36)
\end{aligned}$$

So that finally

$$\frac{\mathrm{d}r_f(t)}{\mathrm{d}t} = N(2\rho+\phi)\mu\left[P_f(t) + \mu P_f^2(t)\right] \qquad (6.37)$$

Comparing this prediction with an ensemble of simulations, we can see from Figures 6.3a and 6.3c that it greatly underestimates the true variance across many realizations. However, we saw earlier that the same method did succeed in predicting the expected value. So what went wrong? The difference lies in the very first step of this derivation, (6.22) for the expected value versus (6.29) for the variance. Taking the variance operator into the integral required the assumption of independent evolution of clones, whereas the expected value did not. Thus our approach succeeded for the expected value because it is a linear operator irrespective of the dependence of its operands (while the variance only acts linearly if its operands are independent).

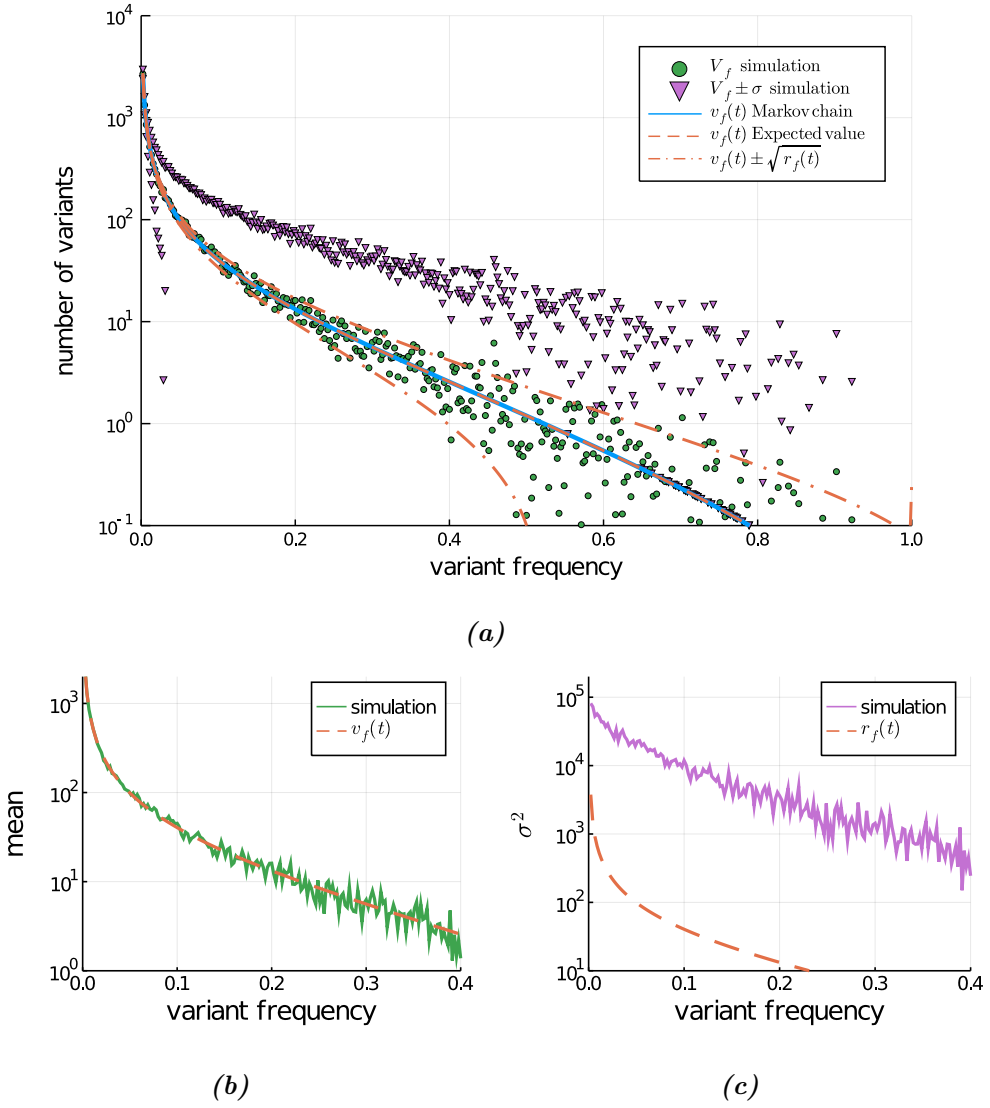


Figure 6.3.: *Comparison of the mean V_f and standard deviation σ for the single cell simulation (averaged of 500 simulations), the Markov chain evolved expected value $v_f(t)$ from (6.16) and (6.28), and the evolved variance estimate $r_f(t)$ from (6.37).* ***(a)*** *The standard deviations are shown as the positive and negative displacement from the expected value, i.e. $V_f \pm \sigma$ and $v_f(t) \pm \sqrt{r_f(t)}$.* ***(b-c)*** *Direct comparisons of the mean and variance between the simulations and the evolved predictions.*

6.5.3. Equilibrium distributions

While we now have a method for obtaining the expected VAF of the system at any point in time, it is worth wondering whether equilibrium solutions exists for this quantity. Since the partial differential equation representation (6.16) takes the form of a diffusion with absorbing boundaries coupled with a source term, our intuition might tell us that a state could exist where the incoming flux of variants exactly balances the amount being lost to the boundaries (variants which fixate or go extinct), though the boundary states $f = 0$ and $f = 1$ would still continue to increase. In fact, this problem has been studied before [REF], and it can be shown that for a fixed population

$$v_{eq}(x,t) = (2 + \phi/\rho)\mu \frac{1}{x} \tag{6.38}$$

if the system is in equilibrium. We can check this result with our approach by evolving a system for a long time $t \to \infty$, as shown in 6.4a. We note that the system indeed converges to the $\sim 1/f$ distribution over time, with the property that the low frequency variant number stabilize sooner than those at high frequency, as shown in Figure 6.4b.

6.5.4. Discussion: VAF

From the diffusion approximation (6.16) we have seen that the dynamics take the character of a diffusion process with an added source term. Furthermore, if a system is in a state with few variants, over time it will evolve towards an equilibrium of the VAF (if the boundaries are excluded), where the incoming flux of clones completely balances they clones lost to extinction and fixation. Let us briefly consider the implications of these findings. The diffusion term in (6.16) drives the system's evolution towards equilibrium, meaning that – similar to the single clone picture – increasing the rate of self-renewals ρ speeds up the overall rate of expansion, whereas increasing the population size N slows this down. The equilibrium distribution (6.38) on the other hand is mostly determined by the source term, being independent of N, and ρ only appearing to limit the contribution of ϕ. Clearly higher mutation or asymmetric division rates result in equilibrium distributions with more variants, as shown in Figure 6.4c.

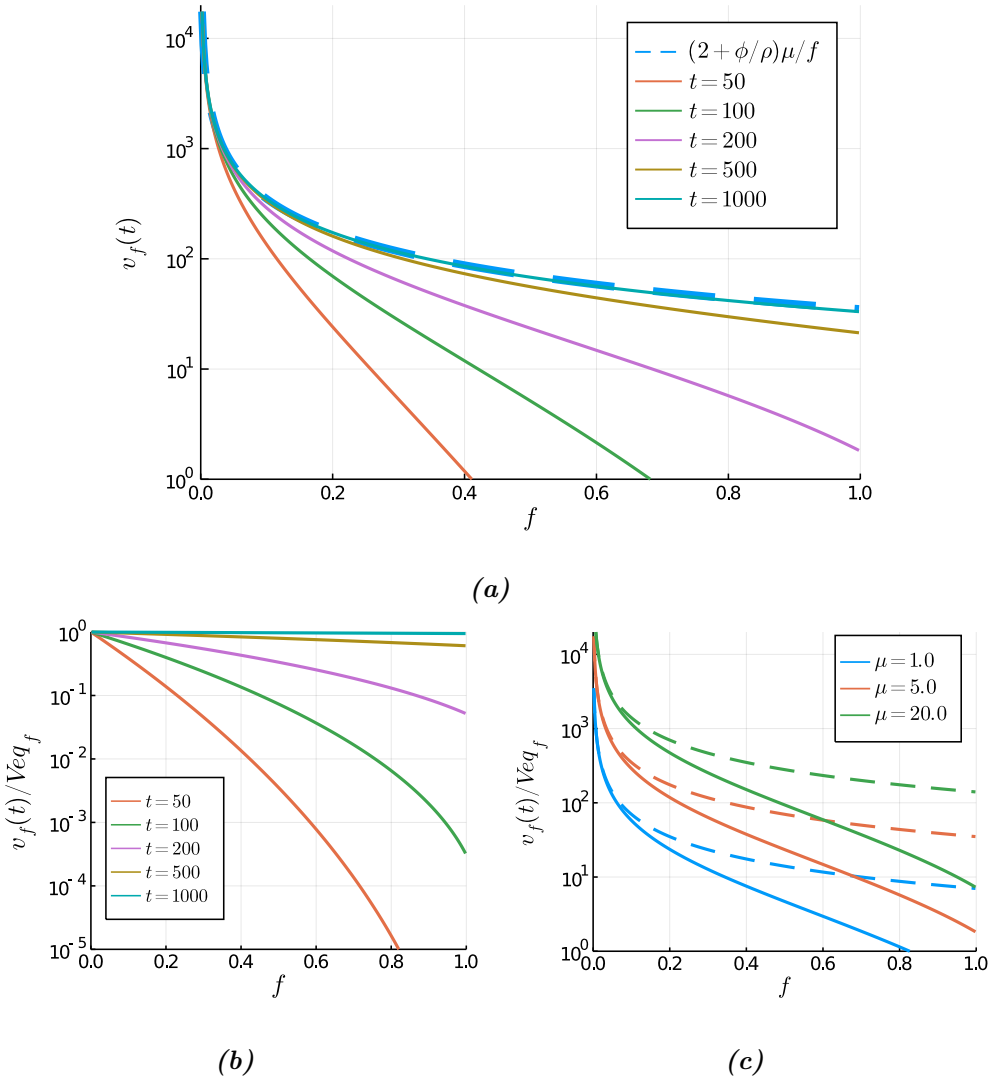


Figure 6.4.: Sparsely populated VAFs evolve to an equilibrium over time. (a) *For a population with parameters values $N = 500$, $\rho = 1.0$, $\phi = 5.0$, and $\mu = 5.0$, as the time t increases $v_f(t)$ approaches its predicted equilibrium (6.38).* ***(b)*** *Lower frequency states of the VAF converge relatively faster to the equilibrium distribution.* ***(c)*** *Expected VAF $v_f(t)$ for three populations with $N = 500$, $\rho = 1.0$, $\phi = 5.0$, and differing mutation rate μ. The full lines the denote $v_f(t)$ after a time $t = 200$, while the dotted lines denote the expected equilibria.*

It may seem a bit paradoxical that we obtain an equilibrium distribution for the spectrum of mutations while we showed earlier that the single cell burden increases indefinitely (6.2). The subtle distinction to be made here is that we have ignored the boundaries of $\mathcal{F} = [0, 1]$ when defining the equilibrium VAF, but did no such thing for the single cell burden. While variants that have gone extinct are of course no longer present, any fixated mutations still are, which is something that we did not take into account in the derivation of (6.2). It is straightforward to see how we might account for these, by taking the fixation probability of a single mutant over time as the fraction of mutants that are lost. However, we will leave such a treatment for possible future development.

6.6. The sampling problem

If we wish to compare our predictions to a dataset, complications arise if the sampled data does not cover the entire population. A naive guess might be that sampling simply reduces the accuracy with which we can determine the stochastic distributions, an intuition which is true for some quantities. Taking for example the single cell mutational burden, we have found that the probability of a single cell carrying m variants follows a compound Poisson distribution; so that whether our sample contains one tenth, one hundredth, or one thousandth of the total population, the distribution from which we are sampling remains the same and thus the sample size only influences the resolution with which we observe it. This is not the case for the VAF, however, as we will see shortly, making it necessary to quantify in what manner sampling causes the observed VAF to differ from that of the complete population.

First let us revisit the set oriented perspective of the system introduced in Section 6.1. With a population of N cells $C_i \in \mathcal{N}$, we can denote the variants as subsets $\mathcal{V}_j \subseteq \mathcal{N}$ with sizes n_j corresponding to the number of cells they contain. Now, taking a random sample of $S(< N)$ cells from this population is akin to taking a subset $\mathcal{S} \subset \mathcal{N}$. Denote $\mathcal{U}_j$ as the set of cells belonging to the variant $\mathcal{V}_j$ that are sampled into $\mathcal{S}$, i.e.

$\mathcal{U}_j = \mathcal{S} \cap \mathcal{V}_j$. Depending on the realization of the sampling, the size s_j of $\mathcal{U}_j$ can be anywhere between 0 – if none of its cells are sampled – and n_j – if all of its cells are sampled. The random variable s_j can be phrased as the sampling process of obtaining s_j successes after S draws *with replacement*, for which the probability distribution is the well-known hypergeometric distribution [45]. In particular, the probability $\mathbb{P}\{s \mid n\}$ of a variant of size n in $\mathcal{N}$ being sampled to size s in $\mathcal{S}$ is given by

$$\mathbb{P}\{s \mid n\} = \frac{\binom{n}{s}\binom{N-n}{S-s}}{\binom{N}{S}} \tag{6.39}$$

Denoting v_n as the number of variants with size n (frequency n/N) in $\mathcal{N}$ – i.e. the VAF in the true population – we can write for the number of variants with size s in $\mathcal{S}$

$$u_s = \sum_{n=s}^{N} v_n \cdot \mathbb{P}(s \mid n) \tag{6.40}$$

which is the VAF of the sample. We can see from Figure 6.5 that sampling indeed affects the shape of the expected VAF, as opposed to the single cell mutational burden where it does not.

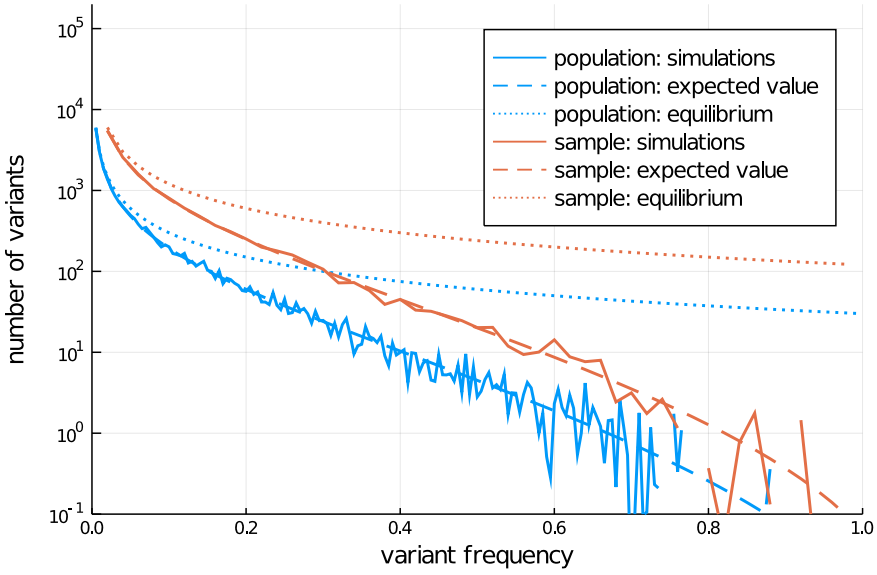


Figure 6.5.: The effect of sampling on the observable VAF. *A population with parameters* $N = 200$, $\rho = 1.0$, $\phi = 4.0$, *and* $\mu = 5.0$ *is evolved for a time* $t = 40$ *and then sampled to an observation of* $S = 50$ *cells. The curves shown for the simulations are the average of 500 trajectories, each evolved stochastically according to the model and then sampled randomly. The expected values* $v_f(t)$ *are calculated with* (6.13) *and sampled according to (6.40), while the equilibrium states are projected by (6.38) and sampled through (6.40). Both the dynamical and equilibrium states are affected by the sampling.*

6.7. Applications to a human HSC dataset

In the previous section of this chapter we have found a number of methods for predicting the clonal evolution of the hematopoietic stem cell pool, based on the assumption of stem cell behavior following Moran-like dynamics. In this section we test our model by comparing its results to a dataset of human HSCs containing information on mutational variants. In particular, we wish to determine on the one hand to what extent the data qualitatively fits with the mathematical predictions, and on the other whether we can infer quantitative properties (such as parameter values) of the system by an appropriate fitting.

6.7.1. Data: somatic mutations in single HSCs

The dataset used here was created by H. Lee-Six et al. [78] and is publicly available. It contains high resolution mutational information on the frequency of some 90'000 mutational variants found in a sample of 89 singly identified HSCs drawn from the bone marrow and peripheral blood of a 59 year old male. The experimental design provided an accurate method for assessing the mutational burden of the individual cells without the risk of false positives inherent in single cell sequencing techniques. First a cohort of hematopoietic stem and progenitor cells (though we are only interested in the former) were identified from samples by sorting based on known cell surface markers [103], with the HSCs characterized by a $CD34^+$ $CD38^-$ $CD90^+$ $CD45RA^-$ profile. All individual cells were then separately cultured to obtain colonies that were bulk sequenced at high depth (around $15\times$), allowing for the accurate detection of somatic mutations that were present in the originally extracted stem cells, as these would be the variants present at frequency 1 in each colony. Thus the authors obtained distinct mutational profiles for 89 individually identified hematopoietic stem cells, an observation equivalent to the example previously discussed in Figure 6.1. These can be visualized as a boolean matrix, which for each cell denotes which of the observed variants it carries.

6.7.2. Single cell mutational burden

We first look at the single cell mutational burden, the distribution of which is shown in Figure 6.6. At this point we have no knowledge of the parameter values which best fit our model, however from (6.6) and (6.7) we know that the compound Poisson distribution is completely determined by its mean and variance, so that we may write

$$(2\rho + \phi)t = \frac{\mathrm{E}(m_i)^2}{\mathrm{Var}(m_i) - \mathrm{E}(m_i)} \tag{6.41}$$

$$\mu = \frac{\mathrm{Var}(m_i) - \mathrm{E}(m_i)}{\mathrm{E}(m_i)} \tag{6.42}$$

with the m_i the single cell burdens in the dataset. Note that both the mutation rate and the total rate of divisions are completely determined by the data, for which we find the values $\mu = 4.3$ *mutations per division per daughter cell* and – using the fact that the donor was 59 years of age at the time of measurement – 4.2 *divisions per year per cell*. The mutation rate is somewhat higher than the 1.2 estimated by Lee-Six et al. using a different method, and the 1.14 estimated in another study using this dataset [145], though the order of magnitude is the same. The expected 18.1 mutations per year fits well with another study performed recently by Osorio et al. where this value was estimated at 14.2 ± 8.1 [105]. Furthermore, from Figure 6.6 we see that the compound Poisson provides an excellent fit for the observed burden distribution, whereas a simple Poisson distribution – what we might consider the naive guess – fails to match the observation.

6.7.3. Variant allele frequency spectrum: fitting parameters with Approximate Bayesion Computation

Having extracted the mutation rate μ, let us now turn to the VAF. In our model the current state of the system is determined by four parameters given in 6.1 – N, λ, p, μ – and the elapsed time t. However, since the data we work with forms only a (small) sample of the total population, we have seen in Section 6.6 that we must also take this sampling into account, meaning that we have a sixth parameter for the sample size S. Of these, the elapsed time and the sample size are known from the experiment, and the

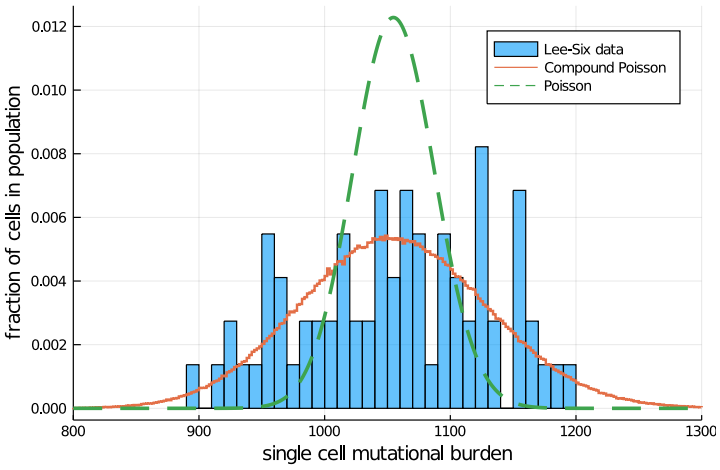


Figure 6.6.: Single cell mutational burden of the Lee-Six dataset. *The distribution of burdens for the 73 bone marrow derived stem cells in the dataset is shown, the mean and variance of which are used to fit the compound Poisson distribution through (6.41) and (6.42). As a contrary example, a Poisson distribution obtained by a least-squares fit to the data is shown as well, which clearly underestimates the variance of the true distribution.*

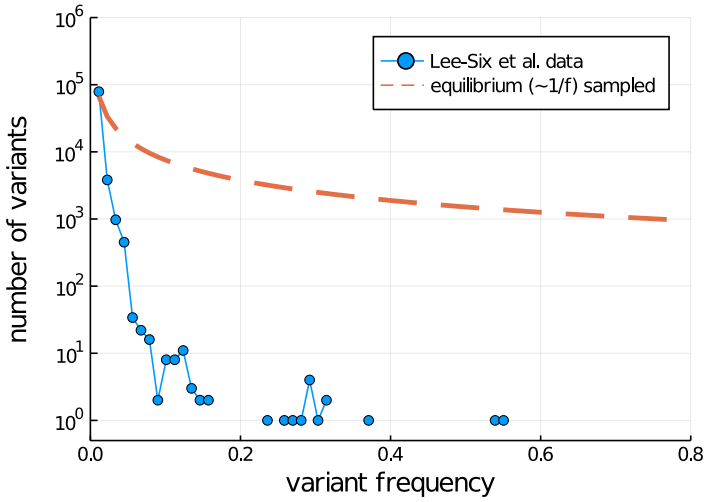


Figure 6.7.: *Variant allele frequency spectrum of the HSCs in the Lee-Six et al. dataset. The shape of the distribution differs strongly from the* $\sim 1/f$ *form associated with an equilibrium VAF, implying that the system has not yet reached the equilibrium state.*

mutation rate and total division rate ($\tilde{\lambda}$) have already been derived from the distribution of single cell mutational burdens, leaving only two unknown parameters: the fraction of "within HSC pool" divisions that are asymmetric p and the population size N. While the latter has been estimated some few times in the past [32, 78](including the original work by Lee-Six et al. using this dataset), the former has proven more elusive to gauge, meaning its value would be of particular interest. Before attempting to fit the model, we might first check whether an equilibrium state has been reached, since if this is the case the prediction for the VAF simplifies greatly – the true VAF of the total population should be proportional to $1/f$ with the slope given by $N\mu$, as described in Section 6.5.3, while the the sampled VAF would be found by applying (6.40) to this. From Figure 6.7 it is clear that this is not the case, meaning that the HSC pool of the dataset is still in a dynamically evolving regime.

Instead we may attempt to fit the free parameters through (6.13) or (6.16). One approach to this is to apply a Monte Carlo type of method, where the master equation

is evolved a large number of times for randomly chosen pairs of values for p and N, and the resulting VAFs which best match the data are recorded. A simple method known as *approximate Bayesian computation* (ABC) provides a convenient scheme for this [116, 130]: given a metric for characterizing the distance between the reference $\tilde{v}(x,t)$ (e.g. the VAF obtained from the measurement) and a predicted $v_i(x,t)$ generated by the model and a parameter pair (N_i, p_i), it is straightforward to visualize which pairs fall within a chosen distance ϵ. The question remains whether a single unique (N,p) pair determines $v(x,t)$, or if other combinations might exist leading to the same result. Rewriting the diffusion picture (6.16) in terms of λ and p

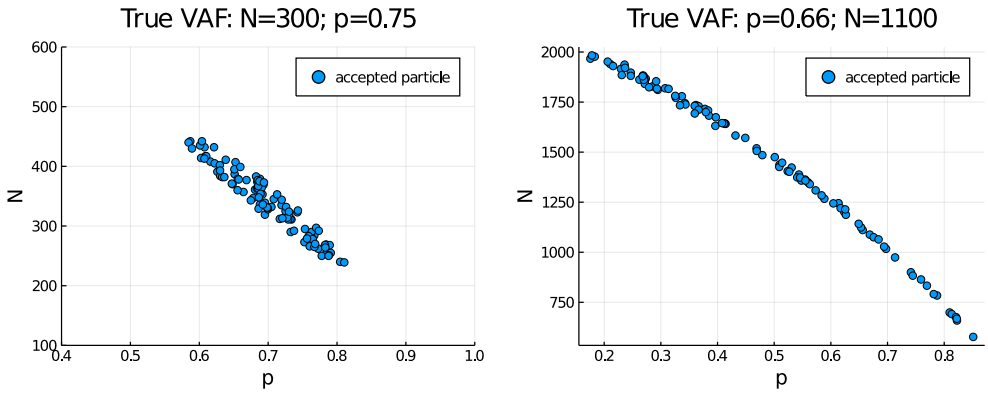
$$\frac{\partial v(x,t)}{\partial t} = \frac{\lambda(1-p)}{N}\frac{\partial^2[x(1-x)v(x,t)]}{\partial x^2} + \delta(x-N^{-1})N\lambda(2-p)\mu \tag{6.43}$$

we can see that for a fixed time t the asymmetric fraction p and the population size N should decouple, since in the diffusion term they appear in the numerator and denominator respectively, whereas in the source term they are both proportionality factors. However, when comparing to data we must also account for sampling, and it is not immediately clear from (6.40) whether this decoupling remains true; or in more mathematical terms: we do not know whether this functional is injective (i.e. no two $v_i(x,t)$ and $v_j(x,t)$ can map to the same $v_s(x,t)$) if we ignore the states $k=0$ (lost variants) and $k=S$ (fixated variants). As a simple test for this we can take an expected VAF evolved with a known parameter set as reference, to see whether the ABC method converges to a unique (N,p) pair. Taking as a distance metric the sum of the squared relative distances of each frequency in the VAF

$$d_i = \sum_f \left(\frac{v_i(f,t) - \tilde{v}(f,t)}{v_i(f,t)}\right)^2 \tag{6.44}$$

the results of this are shown in Figure 6.8a where no sampling is performed and Figure 6.8b for a sampled experiment. The correct parameters can clearly be found before sampling, however after this step we observe the ABC to converge to a line in the parameter space, indicating that some information may indeed be lost during sampling.

Even if a unique set of parameters cannot be found, achieving a reduced parameter space such as found in figure Y is a useful result. However, there is another problem



(a) *(b)*

Figure 6.8.: Inferring parameter values through ABC fitting. ***(a)*** *A reference VAF $v_f(t)$ constructed through (6.13) was used to infer parameters $N = 300$ and $p = 0.75$ using an ABC algorithm.* ***(b)*** *A reference VAF $\tilde{v}_f(t)$ constructed through (6.13) and sampled with (6.40) to size $S = 89$ was used to infer parameters $N = 1100$ and $p = 0.67$ using an ABC algorithm.*

that arises when using a dataset as reference, which related to the chosen distance metric d_i. Where in the previous example the model solutions were compared to an exact result of the expected VAF, in reality the the observed VAF is subject to stochastic fluctuations with respect to its expected value, the strength of which are characterized by the process' variance. Furthermore, we have seen earlier that the variance $r(f,t)$ grows with increasing f, meaning that we expect greater uncertainty at higher frequencies. While the distance d_i in (6.44) is a sum over all points in the frequency space $\mathcal{F}$, is does not take into account the expected variation for each point. For example, a fluctuation of 5% might be highly improbable for some state f_i, but completely reasonable for another f_j. A much better metric therefore would be:

$$d_i = \sum_f \left(\frac{v_i(f,t) - \tilde{v}(f,t)}{r_i(f,t)} \right)^2 \tag{6.45}$$

which takes into account the likelihood of deviation at each point $f \in \mathcal{F}$. However now the problem is clear: we do not yet have a correct method for obtaining the variance of

a solution other than running a great number of simulations, which is too inefficient for the purpose of sampling numerous parameter sets through the ABC scheme.

6.7.4. Discussion: applications to a dataset

It is clear that the application of this simple model to datasets with high resolution information on the somatic mutational burden of HSCs can provide useful insights into the stem cell behavior. In particular, using the model's projected distribution of single cell mutational burdens we have estimated the rate of mutations per division at 4.3 per daughter cell – which is on the order of magnitude as other recent estimates [145, 78] – and the average number of divisions (including both symmetric and asymmetric divisions) per HSC at 4.2 per year. Furthermore, we have shown that the variant allele frequency spectrum of the HSC population is in a dynamically evolving state before equilibrium. In principle this state could be fit to the model's prediction for the expected value of the VAF through any Monte Carlo style procedure, however we have shown that this would require an exact prediction of the variance of the VAF, which we do not have at this time.

6.8. Conclusions and perspective

In this chapter we have extended our inspection of mutations in the hematopoietic stem cell pool from considering a single variant to approaching the entire network of clonal relationships. This more detailed picture of clonality has proven useful for testing our assumptions of the system and estimating the fundamental quantities related to its behavior. From the few basic assumptions of our model established in Chapter 4 we obtained predictions for the distribution of single cell mutational burdens in the population, as well as the expected form of the variant allele frequency spectrum, however, using a similar method to predict the variance of the VAF failed. This might be an indication that separate clones do not evolve independently – an assumption which we showed was implicit in the variance calculation, but not in the expected value – so that

a different approach is required.

It is worth noting that up to this point we have still ignored a few key aspects of the hematopoeitic stem cell pool which are worth investigating in future work. First of all, we have not taken into account the ontogenic growth of the HSC pool, i.e. the fact that from birth to adulthood the population must increase in size. This can have a significant influence on the expansion of mutations arising early on, as they occur in a smaller population in which we have shown it is easier to reach higher frequencies. Such *mosiac mutations* are in fact well studied phenomena, given their importance in genetically acquired diseases [66]. Thus investigating the effect of an initial growth phase in which the population increases may be worth the effort.

The second process we have ignored is the possibility of cell death or senescence in the population, which is known to occur as the HSC compartment ages [51, 65]. In particular it has been shown that the genetic diversity of blood cells decreases with age, with senescence hypothesized as a major contributor to this phenomenon [121].

Finally, throughout our treatment we have put little consideration into the possibility of selective advantages occurring in mutations. Given that we were interested in normal hematopoiesis it has made sense to ignore these, however, investigating their effect in the model could find useful applications in understanding the occurrence and dynamics of cancer [146].

7. Feedback-driven compartmental dynamics of hematopoiesis

Arthur Dent: What happens if I press this button?
Ford Prefect: I wouldn't-
Arthur Dent: Oh.
Ford Prefect: What happened?
Arthur Dent: A sign lit up, saying "Please do not press this button again."
— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

In the previous chapters our mathematical treatment of the hematopoietic system has been restricted to the study of hematopoietic stem cells. At this point we will broaden the scope of our interest to consider the cellular dynamics of the entire process, as cells transition from HSCs through various progenitor stages to the mature types released into the bloodstream.

Because many blood related disorders – often hereditary in origin – are related to improper development or problematic behavior in the bone marrow [70], it is imperative to understand how they influence the cellular dynamics of the system which feeds our transient population of blood cells. Indeed, we have seen that even a disease which can be entirely attributed to issues occurring in the HSC pool can in its progression be influenced by the dynamics of cells outside the stem cell niche, as discussed in Section 5.5. And while current understanding of the hematopoietic architecture is qualitatively detailed, from specific knowledge of maturation lineages [102, 63] to the identification

of various signaling pathways [113, 75, 50, 71, 4] – its quantitative dynamical nature remains for the most part unknown, in no small part due to the fact that (as was the case for hematopoietic stem cells) *in vivo* studies of the bone marrow cell dynamics present numerous challenges. Thus a model which projects the developmental process' dynamics based on the established architecture, both during normal hematopoiesis and if the system is perturbed, could be very useful in understanding and projecting the progression of related disorders. At face value this may appear overly ambitious given the complexity of (and time spent on) the stem cell pool alone, however the goal is once again to find general principles of the system which may be understood without specific knowledge of the underlying processes. Indeed, the pyramidal architecture discussed in Chapter 2 suggests that we may in simplistic terms think of the system as describing a flow (albeit one with peculiar properties) of cells through various sequentially ordered states of maturation, originating in the stem cell pool and arriving after a certain number of steps in a familiar blood cell type. Considering the complex differentiation landscape the cells pass through (Section 2.2) it seems unlikely that different maturation pathways (the particular sequence of differentiation "choices" made by a cell developing towards a particular cell type) wouldn't differ in various quantitative aspects, such as their number of divisions between the stem and mature states, or the flux of cells passing through particular progenitor types. However, the structure of the system in itself presents the possibility of highly non-linear dynamics, which we can study even while remaining agnostic as to the specific values underlying different lineages.

In the past decades a handful of mathematical models of hematopoiesis have been developed to this end [12, 13, 1, 35, 91, 82, 40, 107, 36, 39, 79], however these have typically been constructed to either describe a particular differentiation pathway in equilibrium [35, 40] or a specific deregulation caused by disease [26, 38, 39], but not to consider how the system responds to perturbations in general. Still, one such model developed by Dingli et al. [35] introduces a simple framework which provides a useful starting point for examining the cell dynamics. It describes the hematopoietic architecture as a sequence of *compartments* (corresponding to increasing "levels" of differentiation and

loss of pluripotency) which cells pass through during maturation. Unfortunately, the dynamics as proposed in [35] can only describe the system as it behaves close to equilibrium, as it contains no interactions with the blood compartment which are known to exist in reality [57, 96]. To this end we will introduce a conceptually simple extension to the model, by introducing regulatory feedback mechanisms that allow the system to react to perturbations, for example a loss of cells due to bleeding or hemolysis (destructing of red blood cells). The existence of such feedback loops is not in question, as there are many cytokines and hormones known to be involved in the regulation of cellular proliferation and differentiation [75, 50, 71, 4]. However, such regulatory processes can often in themselves require complex circuit descriptions for the purpose of modeling, and introducing such parameter heavy components quickly reduces both the interpretability as well as the applicability of a model. For this reason we will take an agnostic approach to their functioning, introducing them as black box components which require no knowledge of the underlying signaling circuitry. Through this extended formalism we will study the types of behavior which may occur following perturbations of both transient (for example bleeding) and chronic (for example PNH) nature, as well as validate the model using data from a quantitative study on erythrocyte dynamics [58].

Finally, note that in this chapter – as opposed to much of the previous work – we will for the most part ignore stochastic effects and instead take a deterministic approach. While the processes of division and differentiation are still assumed to occur with some level of stochasticity, especially on the level of single cells, for the quantities of interest here we will simply take their rates of occurrence as deterministic on the population level.

7.1. A compartmental model of hematopoiesis

7.1.1. Dingli model

The model of Dingli et al. [35] constitutes our starting point. It describes the maturation process of hematopoietic cells through a fixed number M of discrete compartments associated with progressive "levels" of differentiation that all cells traverse before leaving the bone marrow, as shown in Figure 7.1. As differentiation is assumed to occur in only one direction (with the cell moving *farther* from its original multipotent stem cell state, and *closer* to a final mature state), higher numbered compartments are sometimes referred to as being *downstream* with respect to lower numbered compartments which are considered *upstream*. Within each compartment j a cell divides at a predefined rate r_j, where each division is considered symmetric (for simplicity), that is, it gives rise to two identical daughter cells. These are either exact replicas of the parent – with probability $1-\epsilon_j$ – and thus remain in the current compartment j, or have differentiated – with probability ϵ_j – and thus move to the subsequent compartment $j+1$. From these assumptions the dynamics of the size (or *number of cells within*) N_j of a compartment j are given by:

$$\partial_t N_j = 2\epsilon r_{j-1} N_{j-1} - (2\epsilon - 1) r_j N_j \tag{7.1}$$

Under homeostatic conditions the number of cells in each compartment should remain approximately constant in time, while compartment sizes increase toward maturity at a fixed ratio $N_{j+1}/N_j = \eta$ to accommodate the expansion of a small number of stem cells (N_0, of the order of several hundred for humans) to the daily output of the bone marrow ($N_M \approx 10^{11}$). This exponential increase is mirrored by the division rates: $r_{j+1}/r_j = \rho$, while the differentiation probability is taken the same for all compartments: $\epsilon_j = \epsilon$. Values for these parameters can be derived by fixing the initial and final compartment sizes and division rates, and using the equilibrium requirement $\partial_t N_j = 0$. (see Appendix A.4 for how this is done specifically).

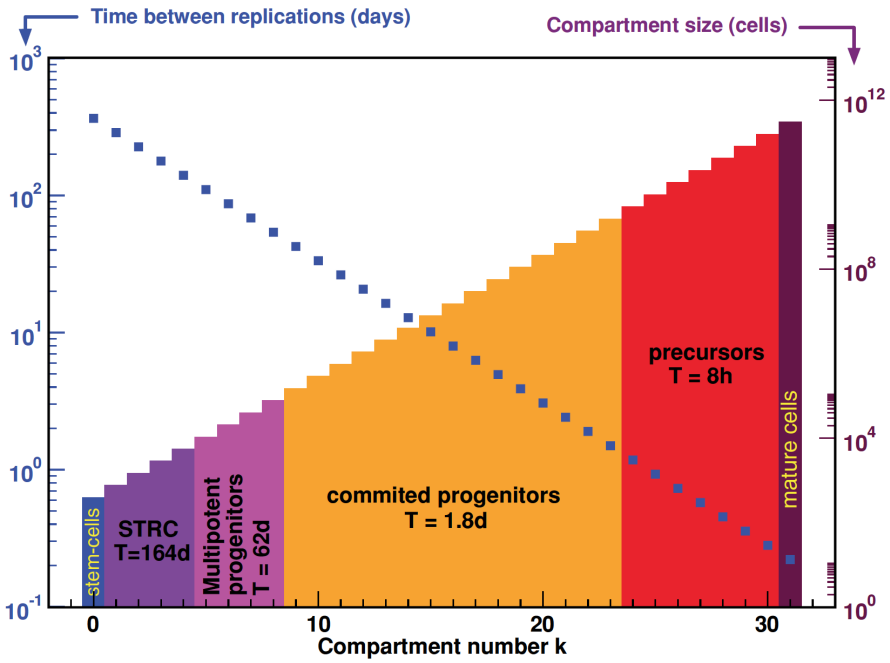


Figure 7.1.: Compartmental hematopoiesis model of Dingli et al. *Successive compartments k represent discrete stages of differentiation which cells pass through during maturation. The number of cells in a compartment N_k at any point in time is denoted as its compartment size and increases geometrically with k. Similarly the division rate r_k (= 1/time between replications) in each compartment also grows geometrically with k. This figure was reproduced from [35].*

7.1.2. Introducing feedback

In order to address the coupling between compartments through feedback, we now alter the existing formalism. First, we formally describe both types of division – self-renewal ($j \to j$) and differentiation ($j \to j+1$) – as independent Poisson processes occurring with rates v_j and s_j respectively. From Section 3.1.2 we know that this is equivalent to the description of Dingli et al. through the relations $r_j = s_j + v_j$ and $\epsilon_j = s_j(s_j + v_j)^{-1}$ (see Appendix A.1 for the detailed proof), so that rewriting the dynamics for the number of cells in each compartment j (7.1) results in

$$\partial_t N_j = 2s_{j-1}N_{j-1} - (s_j - v_j)N_j \tag{7.2}$$

Under homeostatic conditions the system is stable with $N_j(t) = N_j^*$ and $\partial_t N_j^* = 0$, and the division rates are given by their homeostatic values v_j^* and s_j^*. We introduce feedback through sequential coupling between successive compartments, by allowing the rate parameters of each compartment to vary depending on the number of cells in a neighboring downstream compartment. Given a perturbation $n_i = (N_i - N_i^*)/N_i^*$ on the cell number in compartment i, we are thus looking for non-negative functions $v_j(n_i)$ and $s_j(n_i)$ that produce a negative feedback response – i.e. opposing the sign of n_i. Furthermore, we assume there is an upper limit to how many divisions a cell can undergo, thus determining upper bounds on $v_j(n_i)$ and $s_j(n_i)$. Naturally, the fact that homeostasis is maintained in the absence of any perturbation implies that $v_j(0) = v_j^*$ and $s_j(0) = s_j^*$.

From the outset, the functions $v_j(n_i)$ and $s_j(n_i)$ are expected to be the solution of a highly non-linear ecological network of various cell types, nutrients, and signaling factors [106]. Here, instead, we look for the simplest functional form that fulfills the requirements above; this leads us to the linear form

$$u(n) = u^*(1 - \alpha n) \tag{7.3}$$

(where u is either v or s) with $\alpha > 0$ to ensure negative feedback, such that $0 \leq u(n) \leq u_{max} = k_u u^*$ (Figure 7.2). A smoother version is easy to define by drawing inspiration from classical ecological systems, which mirror the competition for promoting

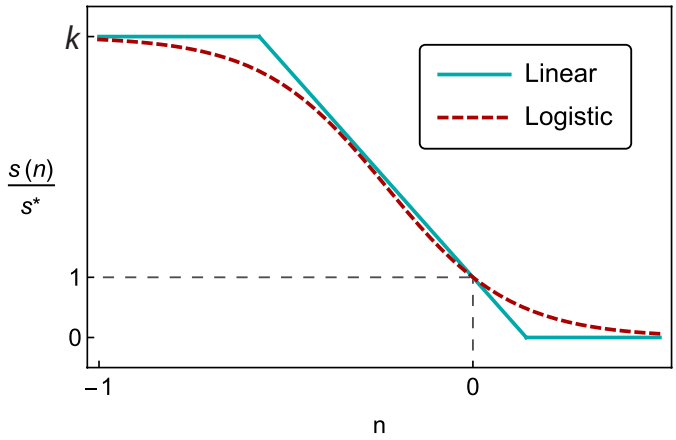


Figure 7.2.: Illustration of linear and logistic differentiation rate functions. *Both are bounded between* 0 *and* ks^**, and have* $s(0) = s^*$*.*

or inhibiting factors among different cell groups [106, 126] where the logistic function arises:

$$\frac{u(n)}{u^*} = \frac{k}{1 + (k-1)e^{\alpha n}} \tag{7.4}$$

where the parameters k and α play analogous roles in determining respectively the maximum and the slope (Figure 7.2). While the rate functions defined above provide a useful method for coupling any pair of compartments, modeling the full hematopoietic system requires an interaction network that defines the pairwise connections between compartments. Many complex circuits are possible, and the number of potential interaction combinations (through pairs or higher orders) increases dramatically with the number of compartments. Here we explore a simple case, in which all compartments are coupled sequentially to their downstream neighbors, so that the rate functions have the form $s_j(n_{j+1})$ and $v_j(n_{j+1})$ for all j. Given this interaction network, as well as the rate functions and their parameters $\alpha_s, k_s, \alpha_v, k_v \in \mathbb{R}^+$, the solution to (7.2) for M compartments can be obtained numerically through any finite difference method.

7.2. Analysis

7.2.1. Sequential coupling elicits three types of behavior

We start by examining the case in which hematopoiesis proceeds under homeostasis when a perturbation occurs in a single compartment. The response in the absence of feedback mechanisms has been studied previously in [143] and can be recovered in this model by fixing the division rates to their homeostatic values: $v_j(t) = v_j^*$ and $s_j(t) = s_j^*$. Equation (7.2) shows that without feedback the compartmental coupling is entirely one-directional and upstream: the dynamics of N_j depends on N_{j-1} but not on N_{j+1}, meaning that compartments will not respond to disturbances taking place in downstream compartments. Still, when a transient perturbation from equilibrium occurs in a given compartment j, the homeostatic equilibrium is eventually restored (Figure 7.3a), though in the absence of downstream coupling the relaxation time is too long to match real recovery times (see discussion). While all upstream neighbors $j-k$ remain in homeostatic conditions ($n_{j-k} = 0$) all downstream $j+k$ are affected as the perturbation moves successively through these compartments.

This behavior will change when feedback – as described above – is introduced: A dependence of N_j on N_{j+1} is now included and a similar wavelike propagation upstream is now expected. The key components of our model that determine the dynamics following a perturbation are the ratio of the coupling strengths of the differentiation/self-renewal rates $s_j(n_{j+1})/v_j(n_{j+1})$, and the total number of feedback stages in the system. The latter will be discussed in the next section. To understand the former – the effect of the relative coupling strengths – we define the simplest possible network with just a single coupled "pair", and turn our attention to the state of the system at time t_0 immediately after a perturbation n_{j+1} is introduced, so that n_j is still 0. Then the dynamics (7.2) of the reacting compartment j can be rewritten as

$$\partial_t n_j|_{t_0} = \frac{2s_{j-1}^*}{\eta} - (s_j(n_{j+1}) - v_j(n_{j+1})) \tag{7.5}$$

If the outgoing flux is unchanged from the homeostatic case, i.e.

$$s_j(n_{j+1}) - v_j(n_{j+1}) = s_j^* - v_j^* \tag{7.6}$$

the equilibrium condition is achieved and we have $\partial_t n_j = 0$. If this remains true for any value of n_{j+1} then under sequential coupling this condition prevents the perturbation from moving upstream with respect to the first responding compartment, and thus protects all upstream compartments from deviating from homeostasis while significantly reducing the time required to return to homeostasis (Figure 7.3a). For the linear rate functions (7.3), equation (7.2) leads to the following condition to ensure that (7.6) is fulfilled:

$$\alpha_v = \frac{s_j^*}{v_j^*}\alpha_s \tag{7.7}$$

While in general no such solution exists for the logistic functions, we use this relation as a first order approximation and denote α_s^* and α_v^* to indicate parameter values which fulfill this requirement. Whenever (7.6) is not fulfilled, then $\partial_t n_j \neq 0$ and one can expand the rate functions about $n_{j+1} = 0$ which, after cancellation of the zeroth order terms (due to the homeostatic condition) gives:

$$\partial_t n_j|_{t_0} = -\left(\frac{\partial s_j}{\partial n_{j+1}}\bigg|_0 - \frac{\partial v_j}{\partial n_{j+1}}\bigg|_0 \right) n_{j+1} + \vartheta\big(n_{j+1}^2\big) \tag{7.8}$$

Ignoring higher order terms and recalling that we have required $\partial s_j/\partial n_{j+1} < 0$ and $\partial v_j/\partial n_{j+1} < 0$ to ensure negative feedback, we see that the sign of $\partial_t n_j|_{t_0}$ can either oppose or match the sign of n_{j+1}, depending on the difference in the brackets: $\partial_n v < \partial_n s$ or $\partial_n s < \partial_n v$ respectively. In the latter case the matching sign means the feedback can actually amplify rather than dampen the perturbation, provided the difference is large enough, since a loss (or excess) of cells would induce further losses (or excesses) in upstream compartments that are required to provide the incoming flux of cells; this in fact corresponds to a positive feedback, as shown in Figure 7.3b. Conversely, if $\partial_n v < \partial_n s$ we obtain the desired negative feedback regime. Nonetheless, damped oscillations may emerge (Figure 7.3c) which, if severe, can prolong the time necessary for the system to return to homeostasis. Recalling that s and v are strictly decreasing functions, the condition $\partial_n v < \partial_n s$ implies that the rate of self-renewal changes faster with n than the rate of differentiation, while $\partial_n s < \partial_n v$ implies the contrary. Thus intuitively we can interpret these conditions as determining whether the response is driven more by

increased self-renewal ($\partial_n v < \partial_n s$), increased differentiation ($\partial_n s < \partial_n v$), or if the increase is balanced across both processes ($\partial_n v = \partial_n s$).

7.2.2. Increasing cell amplification between compartments reduces stability

The number of interacting compartments M also influences the overall dynamics. Note that M need not necessarily be the same as the number of differentiation stages found through traditional methods such as surface marker identification or transcriptional profiling, as our treatment is flexible enough to loosely describe stages of development which interact through feedback, and thus these compartments may encompass multiple maturation stages found in other models. To ensure a meaningful comparison, we change M assuming the same number of cells at the root of the hematopoietic tree and under circulation. Thus, smaller M implies larger cell amplification rates between consecutive compartments. Varying M is found to influence the stability of the hematopoietic system with respect to the rate parameters. Indeed, when deviations from the conditions in (7.6) and (7.7) take place, one obtains an increase in amplitude of oscillations with decreasing M (Figures 7.3c and d). The origin for this can be seen even when employing the linear coupling function in (7.5) (which is equivalent to keeping only the linear terms in the logistic function): there, the inequality in the first derivative becomes $\alpha_v/\alpha_s \neq s^*/v^*$, meaning that the amplitude of the oscillations is determined by how much α_v/α_s deviates from s^*/v^*, the ratio of homeostatic division rates. In particular, perturbations on α_v^* or α_s^* will have a larger impact the smaller this ratio is. Furthermore, it can be shown that s^*/v^* decreases monotonically with increasing cell number amplification η between compartments, which in our model is akin to decreasing M. In this sense hematopoietic models with lower M are less stable under perturbations on the parameters α_s and α_v. It is worth noting here that stability under variation of these parameters forms an important requirement for the system itself and will be discussed in detail later.

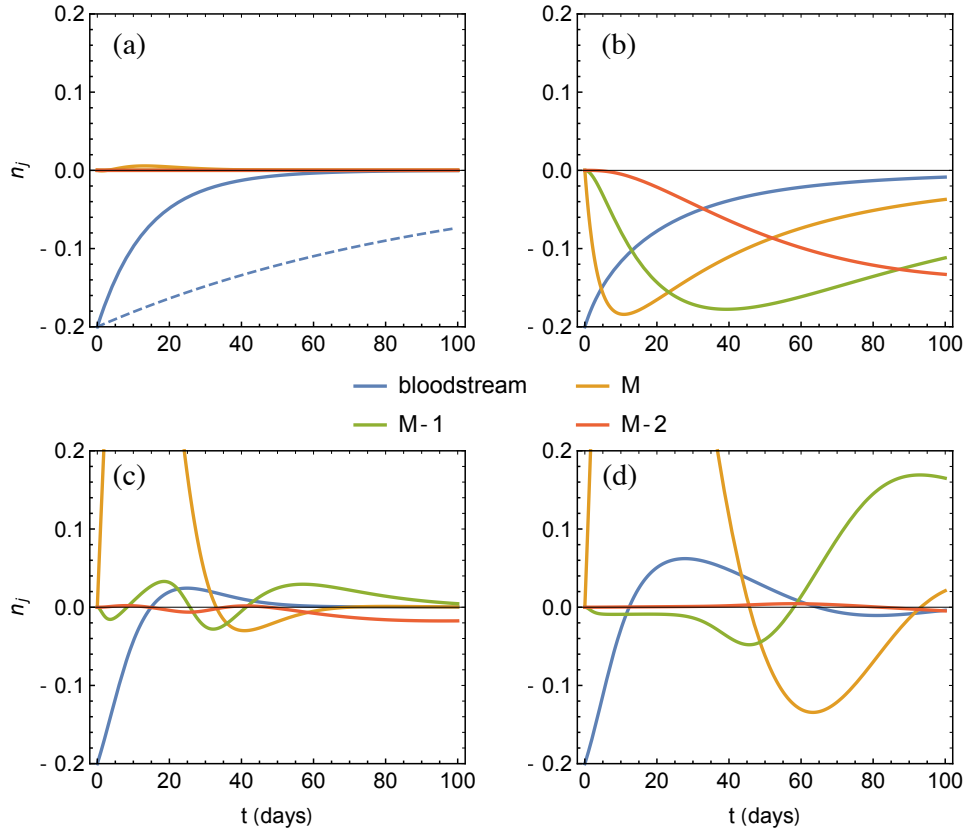


Figure 7.3.: Compartment number dynamics of logistically coupled feedback systems following a sudden loss of cells in the bloodstream. *The cell number is expressed in relative perturbation* $n_i = (N_i(t) - N_i^*)/N_i^*$*, and the bloodstream and final three compartments* $(M-2,\ M-1,\ M)$ *are shown. Parameters* k_s*,* k_v*, and* α_s *are obtained from parameterization to Hillman et al. [58] (see main text and Figure 7.4 for details). (a) An* $M = 5$ *compartment model without feedback (dotted line) and with balanced response* $(\partial_n v \approx \partial_n s)$ *feedback (full lines). The response is not entirely without upstream propagation due to the logistic character of the rate functions, for which no perfectly balanced solution (see (7.6)) exists. (b) differentiation-driven response* $(\partial_n s < \partial_n v)$ *with resulting positive feedback* $(M = 5)$*. (c) self-renewal-driven response* $(\partial_n v < \partial_n s)$ *with oscillatory behavior* $(M = 5)$*. (d) Same as (c) except that* $M = 3$*.*

7.2.3. Recovery time as a measure of efficiency

The time for a compartment to recover from a perturbation is an important measure of the efficiency of hematopoiesis, as an expedited recovery can be considered more advantageous for the host. This recovery time is directly determined by the strength of the response to a loss of cells, which the model itself sets little restriction on: The k and α parameters – respectively determining the maximal increase in divisions and the severity of the perturbation at which this maximal value is reached – can technically (i.e. as long as (7.6) is fulfilled) be taken arbitrarily high without inducing oscillations or positive feedback. However, in real hematopoiesis one would expect physical limitations to apply to these, such as for example the time and/or resources required for cells to undergo additional divisions.

In addressing the recovery time, we should take all possible recovery types into consideration. Indeed, we should keep in mind that hematopoietic cell numbers fluctuate in time even under homeostatic conditions [70]. Consequently, it is reasonable to assign some range around the model's equilibrium value within which a compartment can be considered "recovered". For example, while Figure 7.3a shows a greatly improved response compared to the feedback-free model, the oscillatory behavior in Figure 7.3c presents a qualitatively superior result with respect to the recovery time – effectively halved in this scenario – if we consider a compartment to be recovered once it has returned to within approximately 2% of its homeostatic value (well within the range of normal hematocrit measurements [58]). Thus a slight emphasis on self-renewal rather than differentiation in the response can be beneficial if the resulting oscillations are small in amplitude. Conversely, while the regime depicted in Figure 7.3b (emphasis on differentiation) also improves upon the feedback-free model, it is less efficient than that of Figure 7.3a, as the resulting positive feedback always reduces efficiency.

7.2.4. Inclusion of feedback allows prediction of erythrocyte dynamics

To evaluate the predictive power of the model we use data from Hillman et al. [58], who study the human bone marrow response to a severe loss of erythrocytes. The authors

mark the increase in erythrocyte production as a function of the normal output for different levels of depletion of the *hematocrit* (the volume percentage of erythrocytes in the blood), noting that the efficiency of the response depends strongly on the amount of iron available to the patient. We can translate the hematocrit measurements to perturbations in our model by taking the ratio of the depleted to the normal value; for example if the patient's normal hematocrit is 50%, a reduction to 40% would equate to a 20% loss, which is a perturbation in the bloodstream compartment (B) of $n_B = -0.2$. A summary of their findings is shown in Figure 7.4. We estimate our parameter values by assuming $\alpha_v = (s_j^*/(v_j^*)\alpha_s$ and taking $k_s = k_v \equiv k$. For this coupling the dynamics of the perturbed bloodstream compartment can be written as $\partial_t n_B = 2s_M - \beta_B(1 + n_B)$, with β_B the constant loss rate of circulating cells; which is independent of the replication rate function v_M of the preceding compartment. Thus α_v is fixed by the response requirement and only k and α_s are free. A least-squares fit of the logistic coupling (7.4) results in parameter pairs for the three patient cohorts defined by the authors (based on the patients' body iron stores). The values for the normal patient cohort ($k = 3.5$, $\alpha_s = 7.5$) are used in Figure 7.3. Different parameter pairs are found for the other cohorts, with a clear effect being an increase in maximal production factor κ for increasing iron availability. This implies that the response relies not only on the severity of the perturbation but on the availability of essential resources as well, so that the parameters α_s, α_v, κ_s and κ_v should in fact depend on other parameters reflecting a dynamic environment. The values for $k = k_v = k_s$ found here to range between 3.5 (normal cohort) and 7.6 (hemochromatosis) fit with current knowledge of production rates of mature red blood blood cells, where the highest reported rate increases are 8- to 10-fold the normal rate [58]. For this range of k we thus estimate the slope parameter α_s to be in the range of 7.2–11.3, while α_v is then determined by the compartment number through $\alpha_v = \alpha_s\, s^*/v^*$.

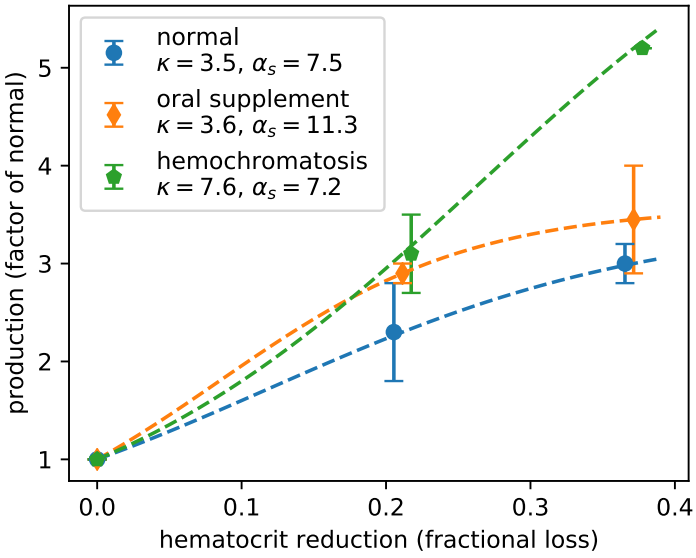


Figure 7.4.: Parameter estimates based on Hillman et al. [58]. *Three patient cohorts are defined by the authors based on the size of their available iron stores: a 'normal' control group, a group which was administered supplementary iron intakes, and a number of individuals suffering from hemochromatosis, a disorder characterized by an increased amount of total body iron stores. Each production factor shown (symbols) is the center of the range (error bars) measured within a patient cohort, as no individual measurements or averages are specified. Dashed lines result from a least-squares fit to the data employing the logistic coupling model (7.4).*

7.2.5. Chronic perturbations lead to new equilibrium states

As a final exploration of the model, we turn our attention to perturbations with a long-lasting character. These are of particular interest in medicine, as many genetic disorders such as inherited red cell membrane defects (hereditary spherocytosis, elliptocytosis, ovalocytosis), thalassemia syndromes and hemoglobinopathies (sickle cell disease, hemoglobin SC disease) all result in a chronic reduction of red cell survival times and anemia. Autoimmune hemolytic anemia due to autoantibodies against red blood cell antigens can also cause chronic destruction of red blood cells and anemia. We can take paroxysmal nocturnal hemoglobinuria (see Chapter 5) as a model example, as it is characterized by severe hemolysis of the PNH afflicted red blood cell population in circulation. If the *PIGA* mutant clone is large enough, a significant portion of circulating erythrocytes will have a severely reduced lifespan. In our model we can take this into account by splitting the bloodstream compartment into a healthy (H) and a PNH afflicted (PNH) population, $N_B = N_H + N_{PNH}$, where the death rate of the PNH group is significantly higher than that of the healthy cells ($\beta_{PNH} > \beta_H$). For a clone which comprises a fraction p of bone marrow cells, we obtain the dynamics

$$\begin{cases} \partial_t N_H = 2s_M(n_B)\,(1-p)\,N_M(t) - \beta_B^* N_H(t) \\ \partial_t N_{PNH} = 2s_M(n_B)\,p\,N_M(t) - \beta_{PNH} N_{PNH}(t) \end{cases} \tag{7.9}$$

To determine which values of p might occur in humans, we note that PNH clones can comprise up to 100 percent of the blood cell population [123], while clones smaller than 10-20% could be considered subclinical. To obtain a realistic value for the rate at which these cells are hemolysed we use a study on the in-vivo survival rate of transfused erythrocytes from a PNH afflicted individual [28]. While no such death rate is derived in the paper itself, the authors describe a fast initial decay of the transfused population to 50% after only 5 days, followed by a slower decay down to 30% at the 10th day. We can describe this behavior by means of two exponentially decaying populations to estimate the donor's PNH fraction at $p \approx 0.8$ and a death rate of $\beta_{PNH} \approx 0.2$, which means that a PNH erythrocyte will be destroyed after on average 5 days in the bloodstream, 20 times faster than its normal counter-part.

Using the same parameter set derived in the previous section, we observe that, in the long term, new steady states emerge for all reacting compartments in any response regime (Figure 7.5). Using Figure 7.5a as a reference, we observe a marginal improvement in mitigating the loss in the self-renewal-driven regime ($\partial_n v < \partial_n s$) (Figures 7.5b and e) whereas, in the differentiation-driven regime ($\partial_n s < \partial_n v$) (Figures 7.5c and f) a reduced efficiency is observed. In contrast with the normal recovery that is realized under transient perturbations (Figure 7.3), the model also predicts a new stationary state for the bloodstream hemoglobin content, which in general remains below the normal homeostatic value. Furthermore, the model captures scenarios where the enduring reduced hemoglobin and red cell mass in circulation is accompanied by a persistent expansion of the upstream compartments (Figures 7.5b and e), as often seen in classic hemolytic PNH as well as other chronic hemolytic disorders [70]. As this expansion does not occur in the differentiation-driven regime we conclude that the adaptive response in chronic hemolytic states must (at least) at times take place in a self-renewal-driven regime.

7.3. Discussion and conclusions

The formalism described here provides a simple method for understanding the type of dynamics that populations of maturing hematopoietic cell precursors undergo in the bone marrow after being subject to different types of perturbations (from mild to severe), such as sudden or chronic blood loss. While the starting model of Dingli et al. [35] provides a useful framework for describing the hematopoietic system under homeostatic conditions, it does not account for the dynamics under perturbations such as those discussed here, as the time for a compartment to return to equilibrium is too long to fit clinically observed timescales [70], and it does not take into account interactions with cell dynamics in the blood stream. The addition of sequential feedback to the model not only produces swifter recoveries, but also reproduces observed dynamic behaviors such as the response to a transient loss of erythrocytes, and the persistence of anemic states following chronic hemolysis with an associated chronic expansion of precursor cells in the bone marrow. The increased complexity, on the other hand, calls for a careful analysis of the properties

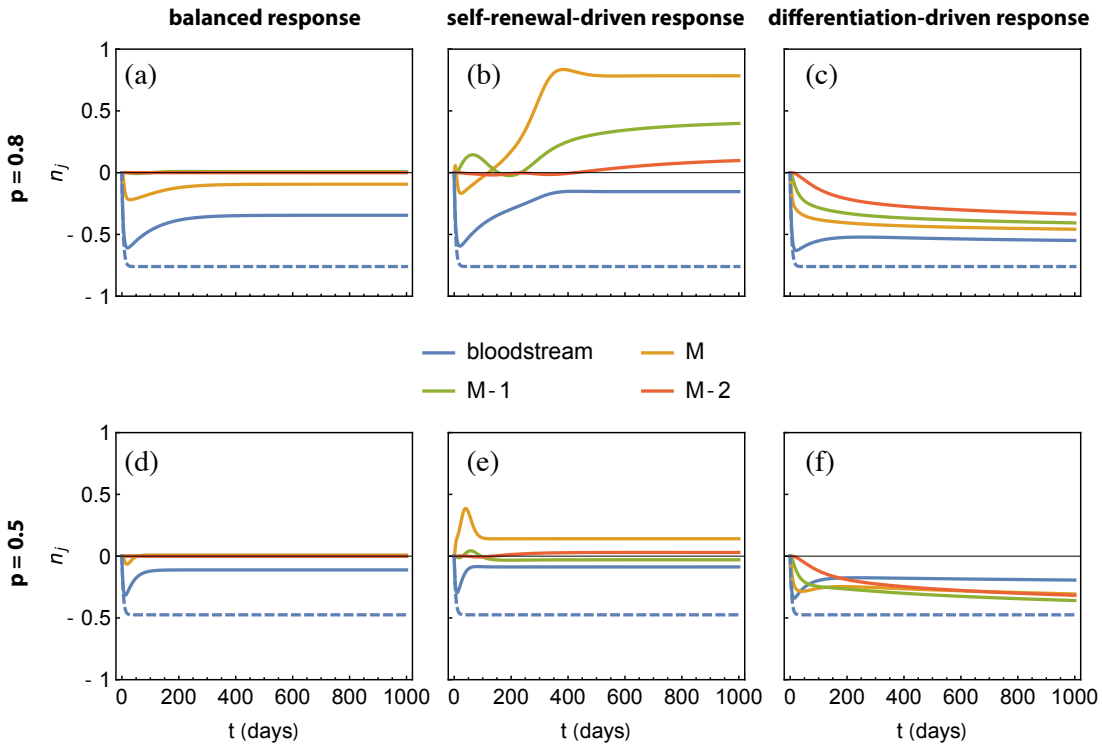


Figure 7.5.: Dynamics following a chronic loss of cells in the bloodstream. *Responses of an $M = 5$ compartmental model employing logistic coupling with normal parameter values taken from Figure 7.4 (full lines) alongside the feedback-free response (dashed line). The rate of hemolysis of PNH afflicted erythrocytes is taken at $\beta_{PNH} = 0.2$. Two different clone sizes are shown: $p = 0.8$ (panels (a)-(c)) and $p = 0.5$ (panels (d)-(f)). Balanced response to clone of size p is shown in panels (a) and (d), self-renewal-driven response ($\partial_n v < \partial_n s$) to clone of size p is shown in panels (b) and (e), and differentiation-driven response ($\partial_n s < \partial_n v$) to clone of size p is shown in panels (c) and (f).*

of the feedback coupling introduced.

We identify three response types for any coupled pair of compartments, determined by the relative strengths of the differentiation and self-renewal coupling, $s(n)$ and $v(n)$ respectively. A perfectly balanced response prevents the perturbation from moving further upstream, thus providing the simplest reaction profile for hematopoiesis as a whole; it occurs whenever the equality $s(n) - v(n) = s^* - v^*$ is fulfilled, and can intuitively be associated with a response where both differentiation and self-renewal increase (or decrease) in a balanced manner such that the compartment's own cell number remains constant. This is however a very strict condition which is difficult to meet, even on average, in hematopoiesis, given its stochastic nature. Thus one expects that, in general, this detailed balance does not occur, and the dynamic behavior depends on which of the rates comes to dominate. When the differentiation rate dominates, the cell number in the compartment will change in the same direction as the perturbation – decreasing if the perturbation is a loss of cells, increasing if it is an excess – effectively introducing a positive feedback. When the self-renewal rate dominates, the compartment's cell number varies in opposition with the perturbation – increasing with a loss, decreasing with an excess – which can lead to an overcompensation of the loss/excess followed by damped oscillations in the cell number. Therefore, if the feedback strengths of self-renewal and differentiation are not tuned to each other, nearly undamped oscillating cell counts in the blood can occur, associated with extreme cases found in certain hematologic disorders such as cyclic neutropenia [107, 39]. It is, however, important to take into consideration that in real hematopoiesis cell numbers in circulation are subject to stochastic noise, even under homeostatic conditions [70]. Thus it is appropriate to introduce a range of values for the cell numbers within which hematopoiesis can be considered to be in (dynamic) equilibrium. In this sense small oscillations within this range predicted by our model can be presumed to be undetectable (and even if detectable, irrelevant) in a clinical setting. This in turn implies the rate parameters have some leeway to be out of sync without disturbing the bloodstream compartment in a detectable way, adding to the overall robustness of hematopoiesis.

The possibility of long lasting disruptions caused by poorly synced feedback loops highlights the importance of the stability of hematopoiesis with respect to the division rate parameters. Here, the coupling functions $s(n)$ and $v(n)$ posit a deterministic dependency of the division rates on downstream cell counts. In reality, these dependencies will be subject to noise from the underlying stochastic biological circuits and – as already pointed out – are unlikely to have perfectly balanced response solutions in the first place. Furthermore, since the response also depends on the availability of resources [58] which may vary or become depleted over time, the balance between s and v adaptation required for stability may itself change in time. However, an important observation is that this stability increases with increasing compartment number, or more specifically decreasing amplification between coupled compartments. The result furthermore adds an interesting angle to the currently favored view that normal hematopoiesis is mostly driven by 'short-term' stem cells which would be found further downstream then the small pool of long term HSCs [129, 23], as such a larger pool of feedback coupled 'drivers' would increase stability.

An important quantifiable characteristic of the feedback driven system is the strength of the coupling between two compartments (determined by the values of the α and k parameters), as it governs the speed with which a return to equilibrium is attained. We find that while balanced responses (Figure 7.3a) allow for arbitrarily strong coupling, the physical limit of how fast a single cell can divide of course cannot be exceeded. Furthermore, the coupling strength may also depend on the availability of essential resources, as can be seen from a human erythropoiesis study where individuals with increased access to iron present amplified responses [58]. This observation raises the question of how long a particular response can be maintained, especially in the case of persistent losses.

Finally, it is worth remarking upon the differences between the compartmental dynamics under transient and chronic perturbations. In the former case, a short-lived perturbation such as bleeding can be swiftly remedied by increased cell divisions in the higher compartments, without propagating to earlier progenitor stages if the homeo-

static balance between self-renewal and differentiation is maintained. In this sense the earliest compartments may not even be requested to respond to an acute loss of blood. On the other hand, chronic perturbations to the system – found in various hematopoietic disorders such as paroxysmal nocturnal hemoglobinuria and other hereditary or acquired hemolytic anemias – lead to the emergence of new equilibrium states that do not correspond to normal homeostasis. For example while the altered dynamics might mitigate a persistent loss of erythrocytes due to hemolysis by increasing the bone marrow output, the resulting steady-state number of erythrocytes in circulation may still be significantly lower than in the unperturbed system – a scenario which fits the observation of anemia occurring in severe cases of PNH as well as other hereditary or acquired hemolytic states. Furthermore, experimental data from telomere length analysis in both PNH and sickle cell disease show that circulating mononuclear cells have shorter telomeres compared to age matched controls [68, 94]. Given that telomere attrition is generally associated with cell divisions, this could be explained by our results here, which posit that under chronic hemolysis progenitor and downstream cells can undergo more replication events than aged matched cells from healthy individuals. In fact, in one study [68] it was found that the shorter telomere length occurred in both PNH afflicted and unafflicted cells, suggesting that the cause indeed lies within the hematopoietic process itself, and that the feedback intrinsic to hematopoiesis does not discriminate between the *PIGA* mutant and normal cells that co-exist in the bone marrow of patients with PNH.

Bibliography

[1] M. Adimy, F. Crauste, and S. Ruan. "A Mathematical Study of the Hematopoiesis Process with Applications to Chronic Myelogenous Leukemia". In: *SIAM Journal on Applied Mathematics* 65.4 (Jan. 1, 2005), pp. 1328–1352. ISSN: 0036-1399. DOI: 10.1137/040604698. URL: https://epubs.siam.org/doi/abs/10.1137/040604698 (visited on 04/25/2019).

[2] Bruce Alberts. *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, 2018.

[3] Matthew Allen. "Compelled by the diagram: thinking through CH Waddington's epigenetic landscape". In: *Contemporaneity* 4 (2015), p. 119.

[4] David Allman, Jon C Aster, and Warren S Pear. "Notch signaling in hematopoiesis and early lymphocyte development". In: *Immunological reviews* 187.1 (2002), pp. 75–86.

[5] Bruce N Ames, Mark K Shigenaga, and Tory M Hagen. "Oxidants, antioxidants, and the degenerative diseases of aging". In: *Proceedings of the National Academy of Sciences* 90.17 (1993), pp. 7915–7922.

[6] David J Araten and Lucio Luzzatto. "The mutation rate in PIG-A is normal in patients with paroxysmal nocturnal hemoglobinuria (PNH)". In: *Blood* 108.2 (2006), pp. 734–736.

[7] David J Araten et al. "Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal

individuals". In: *Proceedings of the National Academy of Sciences* 96.9 (1999), pp. 5209–5214.

[8] DJ Araten et al. "Dynamics of hematopoiesis in paroxysmal nocturnal hemoglobinuria (PNH): no evidence for intrinsic growth advantage of PNH clones". In: *Leukemia* 16.11 (2002), pp. 2243–2248.

[9] Armin Attar. "Changes in the cell surface markers during normal hematopoiesis: a guide to cell isolation". In: *Global Journal of Hematology and Blood Transfusion* 1.1 (2014), pp. 20–28.

[10] Nick Barker et al. "Identification of stem cells in small intestine and colon by marker gene Lgr5". In: *Nature* 449.7165 (2007), pp. 1003–1007.

[11] Andrew J Becker, Ernest A McCulloch, and James E Till. "Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells". In: *Nature* 197.4866 (1963), pp. 452–454.

[12] Jacques Bélair, Michael C. Mackey, and Joseph M. Mahaffy. "Age-structured and two-delay models for erythropoiesis". In: *Mathematical Biosciences* 128.1 (July 1, 1995), pp. 317–346. ISSN: 0025-5564. DOI: 10.1016/0025-5564(94)00078-E. URL: http://www.sciencedirect.com/science/article/pii/002555649400078E (visited on 03/07/2019).

[13] Samuel Bernard, Jacques Bélair, and Michael C. Mackey. "Oscillations in cyclical neutropenia: new evidence based on mathematical modeling". In: *Journal of Theoretical Biology* 223.3 (Aug. 7, 2003), pp. 283–298. ISSN: 0022-5193. DOI: 10.1016/S0022-5193(03)00090-0. URL: http://www.sciencedirect.com/science/article/pii/S0022519303000900 (visited on 03/07/2019).

[14] Monica Bessler et al. "Mutations in the PIG-A gene causing partial deficiency of GPI-linked surface proteins (PNH II) in patients with paroxysmal nocturnal haemoglobinuria". In: *British journal of haematology* 87.4 (1994), pp. 863–866.

[15] M Bessler et al. "Paroxysmal nocturnal haemoglobinuria (PNH) is caused by somatic mutations in the PIG-A gene." In: *The EMBO journal* 13.1 (1994), pp. 110–117.

[16] Luca Biasco et al. "In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases". In: *Cell stem cell* 19.1 (2016), pp. 107–119.

[17] Cédric Blanpain and Elaine Fuchs. "Epidermal homeostasis: a balancing act of stem cells in the skin". In: *Nature reviews Molecular cell biology* 10.3 (2009), pp. 207–217.

[18] Ben Boward, Tianming Wu, and Stephen Dalton. "Concise review: control of cell fate through cell cycle and pluripotency networks". In: *Stem cells* 34.6 (2016), pp. 1427–1436.

[19] Robert A Brodsky. "Paroxysmal nocturnal hemoglobinuria". In: *Blood* 124.18 (2014), pp. 2804–2811.

[20] Yehuda Brody et al. "Quantification of somatic mutation flow across individual cell division events by lineage sequencing". In: *Genome research* 28.12 (2018), pp. 1901–1918.

[21] U.S. Census Bureau. *Annual Estimates of the Resident Population by Single Year of Age and Sex for the United States*. https://data.census.gov/cedsci/. Accessed 13 June 2017. Apr. 2010.

[22] Katrin Busch and Hans-Reimer Rodewald. "Unperturbed vs. post-transplantation hematopoiesis: both in vivo but different". In: *Current opinion in hematology* 23.4 (2016), p. 295.

[23] Katrin Busch et al. "Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*". In: *Nature* 518.7540 (Feb. 2015), pp. 542–546. ISSN: 1476-4687. DOI: 10.1038/nature14242. URL: https://www.nature.com/articles/nature14242 (visited on 04/30/2019).

[24] Rui Chen et al. "Impaired growth and elevated Fas receptor expression in PIGA+ stem cells in primary paroxysmal nocturnal hemoglobinuria". In: *The Journal of Clinical Investigation* 106.5 (2000), pp. 689–696.

[25] Tom H Cheung and Thomas A Rando. "Molecular regulation of stem cell quiescence". In: *Nature reviews Molecular cell biology* 14.6 (2013), pp. 329–340.

[26] Caroline Colijn and Michael C. Mackey. "A mathematical model of hematopoiesis—I. Periodic chronic myelogenous leukemia". In: *Journal of Theoretical Biology* 237.2 (Nov. 21, 2005), pp. 117–132. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2005.03.033. URL: http://www.sciencedirect.com/science/article/pii/S0022519305001542 (visited on 03/07/2019).

[27] Kevin J Curran et al. "Paroxysmal nocturnal hemoglobinuria in pediatric patients". In: *Pediatric blood & cancer* 59.3 (2012), pp. 525–529.

[28] J. V. Dacie and P. L. Mollison. "SURVIVAL OF TRANSFUSED ERYTHROCYTES FROM A DONOR WITH NOCTURNAL HÆMOGLOBINURIA". In: *The Lancet*. Originally published as Volume 1, Issue 6549 253.6549 (Mar. 5, 1949), pp. 390–392. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(49)90704-7. URL: http://www.sciencedirect.com/science/article/pii/S0140673649907047 (visited on 05/19/2020).

[29] Alexander Davis and Nicholas E Navin. "Computing tumor trees from single cells". In: *Genome biology* 17.1 (2016), p. 113.

[30] David Dingli, Lucio Luzzatto, and Jorge M. Pacheco. "Neutral evolution in paroxysmal nocturnal hemoglobinuria". In: *Proceedings of the National Academy of Sciences* 105.47 (Nov. 25, 2008), pp. 18496–18500. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0802749105. URL: http://www.pnas.org/content/105/47/18496 (visited on 11/05/2018).

[31] David Dingli, Lucio Luzzatto, and Jorge M Pacheco. "Neutral evolution in paroxysmal nocturnal hemoglobinuria". In: *Proceedings of the National Academy of Sciences* 105.47 (2008), pp. 18496–18500.

[32] David Dingli and Jorge M Pacheco. "Allometric scaling of the active hematopoietic stem cell pool across mammals". In: *PLoS One* 1.1 (2006), e2.

[33] David Dingli and Jorge M Pacheco. "Ontogenic growth of the haemopoietic stem cell pool in humans". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1624 (2007), pp. 2497–2501.

[34] David Dingli, Jorge M Pacheco, and Arne Traulsen. "Multiple mutant clones in blood rarely coexist". In: *Physical Review E* 77.2 (2008), p. 021915.

[35] David Dingli, Arne Traulsen, and Jorge M Pacheco. "Compartmental architecture and dynamics of hematopoiesis". In: *PloS one* 2.4 (2007), e345.

[36] David Dingli, Arne Traulsen, and Jorge M. Pacheco. "Dynamics of haemopoiesis across mammals". In: *Proceedings of the Royal Society of London B: Biological Sciences* 275.1649 (Oct. 22, 2008), pp. 2389–2392. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2008.0506. URL: http://rspb.royalsocietypublishing.org/content/275/1649/2389 (visited on 11/05/2018).

[37] David Dingli, Arne Traulsen, and Jorge M Pacheco. "Stochastic dynamics of hematopoietic tumor stem cells". In: *Cell cycle* 6.4 (2007), pp. 461–466.

[38] David Dingli et al. "Evolutionary Dynamics of Chronic Myeloid Leukemia". In: *Genes & Cancer* 1.4 (Apr. 1, 2010), pp. 309–315. ISSN: 1947-6019. DOI: 10.1177/1947601910371122. URL: http://journals.sagepub.com/doi/abs/10.1177/1947601910371122 (visited on 11/05/2018).

[39] D. Dingli et al. "Progenitor cell self-renewal and cyclic neutropenia". In: *Cell Proliferation* 42.3 (June 1, 2009), pp. 330–338. ISSN: 1365-2184. DOI: 10.1111/j.1365-2184.2009.00598.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2184.2009.00598.x (visited on 10/24/2018).

[40] M. Doumic et al. "A Structured Population Model of Cell Differentiation". In: *SIAM Journal on Applied Mathematics* 71.6 (Jan. 1, 2011), pp. 1918–1940. ISSN: 0036-1399. DOI: 10.1137/100816584. URL: https://epubs.siam.org/doi/abs/10.1137/100816584 (visited on 10/10/2018).

[41] Ariane Dröscher. "Images of cell trees, cell lines, and cell fates: the legacy of Ernst Haeckel and August Weismann in stem cell research". In: *History and philosophy of the life sciences* 36.2 (2014), pp. 157–186.

[42] Connie J. Eaves. "Hematopoietic stem cells: concepts, definitions, and the new reality". In: *Blood* 125.17 (Apr. 23, 2015), pp. 2605–2613. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2014-12-570200. URL: http://www.bloodjournal.org/content/125/17/2605 (visited on 12/20/2018).

[43] Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*. Vol. 27. Springer Science & Business Media, 2012. Chap. 4.

[44] Willliam Feller. *An introduction to probability theory and its applications, vol 1*. John Wiley & Sons, 1968.

[45] Willliam Feller. *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

[46] CE Ford et al. "Cytological identification of radiation-chimaeras". In: *Nature* 177.4506 (1956), pp. 452–454.

[47] Umberto Galderisi, Francesco Paolo Jori, and Antonio Giordano. "Cell cycle regulation and neural differentiation". In: *Oncogene* 22.33 (2003), pp. 5208–5219.

[48] Lucia Gargiulo et al. "Glycosylphosphatidylinositol-specific, CD1d-restricted T cells in paroxysmal nocturnal hemoglobinuria". In: *Blood, The Journal of the American Society of Hematology* 121.14 (2013), pp. 2753–2761.

[49] Lucia Gargiulo et al. "Highly homologous T-cell receptor beta sequences support a common target for autoreactive T cells in most patients with paroxysmal nocturnal hemoglobinuria". In: *Blood, The Journal of the American Society of Hematology* 109.11 (2007), pp. 5036–5042.

[50] Christian R Geest and Paul J Coffer. "MAPK signaling pathways in the regulation of hematopoiesis". In: *Journal of leukocyte biology* 86.2 (2009), pp. 237–250.

[51] Hartmut Geiger, Gerald De Haan, and M Carolina Florian. "The ageing haematopoietic stem cell compartment". In: *Nature Reviews Immunology* 13.5 (2013), pp. 376–389.

[52] Tatyana Grinenko et al. "Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice". In: *Nature communications* 9.1 (2018), pp. 1–10.

[53] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[54] Penelope Hayward, Tibor Kalmar, and Alfonso Martinez Arias. "Wnt/Notch signalling and information processing during development". In: *Development* 135.3 (2008), pp. 411–424.

[55] Jonathan Henninger et al. "Clonal fate mapping quantifies the number of haematopoietic stem cells that arise during development". In: *Nature cell biology* 19.1 (2017), pp. 17–27.

[56] Anita Hill et al. *The incidence and prevalence of paroxysmal nocturnal hemoglobinuria (PNH) and survival of patients in Yorkshire.* 2006.

[57] Robert S. Hillman. "Characteristics of marrow production and reticulocyte maturation in normal man in response to anemia". In: *The Journal of Clinical Investigation* 48.3 (Mar. 1, 1969), pp. 443–453. ISSN: 0021-9738. DOI: 10.1172/JCI106001. URL: https://www.jci.org/articles/view/106001 (visited on 11/07/2019).

[58] Robert S Hillman, Perry A Henderson, et al. "Control of marrow production by the level of iron supply". In: *The Journal of clinical investigation* 48.3 (1969), pp. 454–460.

[59] Peter Hillmen, Jill M Hows, and Lucio Luzzatto. "Two distinct patterns of glycosylphosphatidylinositol (GPI) linked protein deficiency in the red cells of patients with paroxysmal nocturnal haemoglobinuria". In: *British journal of haematology* 80.3 (1992), pp. 399–405.

[60] Peter Hillmen et al. "Natural history of paroxysmal nocturnal hemoglobinuria". In: *New England Journal of Medicine* 333.19 (1995), pp. 1253–1258.

[61] Peter Hillmen et al. "The complement inhibitor eculizumab in paroxysmal nocturnal hemoglobinuria". In: *New England Journal of Medicine* 355.12 (2006), pp. 1233–1243.

[62] Anthony D. Ho and Wolfgang Wagner. "The beauty of asymmetry: asymmetric divisions and self-renewal in the haematopoietic system". In: *Current Opinion in Hematology* 14.4 (July 2007), p. 330. ISSN: 1065-6251. DOI: 10.1097/MOH.0b013e3281900f12. URL: https://journals.lww.com/co-hematology/Abstract/2007/07000/The_beauty_of_asymmetry__asymmetric_divisions_and.5.aspx (visited on 03/07/2019).

[63] Thomas Höfer and Hans-Reimer Rodewald. "Differentiation-based model of hematopoietic stem cell functions and lineage pathways". In: *Blood* 132.11 (Sept. 13, 2018), pp. 1106–1113. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2018-03-791517. URL: http://www.bloodjournal.org/content/132/11/1106 (visited on 04/23/2019).

[64] Norimitsu Inoue et al. "Molecular basis of clonal expansion of hematopoiesis in 2 patients with paroxysmal nocturnal hemoglobinuria (PNH)". In: *Blood* 108.13 (2006), pp. 4232–4236.

[65] D Leanne Jones and Thomas A Rando. "Emerging models and paradigms for stem cell ageing". In: *Nature cell biology* 13.5 (2011), pp. 506–512.

[66] Young Seok Ju et al. "Somatic mutations reveal asymmetric cellular dynamics in the early human embryo". In: *Nature* 543.7647 (2017), pp. 714–718.

[67] Anastasios Karadimitris et al. "PNH cells are as sensitive to T-cell-mediated lysis as their normal counterparts: implications for the pathogenesis of paroxysmal nocturnal haemoglobinuria". In: *British journal of haematology* 111.4 (2000), pp. 1158–1163.

[68] Anastasios Karadimitris et al. "Severe telomere shortening in patients with paroxysmal nocturnal hemoglobinuria affects both GPI– and GPI+ hematopoiesis". In: *Blood* 102.2 (July 15, 2003). Publisher: American Society of Hematology, pp. 514–516. ISSN: 0006-4971. DOI: 10.1182/blood-2003-01-0128. URL: https://ashpublications.org/blood/article/102/2/514/17382/Severe-telomere-shortening-in-patients-with (visited on 06/16/2020).

[69] Takamasa Katagiri et al. "A cure for paroxysmal nocturnal hemoglobinuria using molecular targeted therapy specific to a driver mutation". In: *Blood, The Journal of the American Society of Hematology* 126.23 (2015), pp. 1215–1215.

[70] Kenneth Kaushansky, ed. *Williams hematology*. Ninth edition. New York: McGraw-Hill, 2016. 2 pp. ISBN: 978-0-07-183300-4.

[71] SJ Kim and J Letterio. "Transforming growth factor-β signaling in normal and malignant hematopoiesis". In: *Leukemia* 17.9 (2003), pp. 1731–1737.

[72] Motoo Kimura. "DNA and the neutral theory". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 312.1154 (1986), pp. 343–354.

[73] Juergen A Knoblich. "Mechanisms of asymmetric stem cell division". In: *Cell* 132.4 (2008), pp. 583–597.

[74] Augustine Kong et al. "Rate of de novo mutations and the importance of father's age to disease risk". In: *Nature* 488.7412 (2012), pp. 471–475.

[75] Jonas Larsson and Stefan Karlsson. "The role of Smad signaling in hematopoiesis". In: *Oncogene* 24.37 (2005), pp. 5676–5692.

[76] Elisa Laurenti and Berthold Göttgens. "From haematopoietic stem cells to complex differentiation landscapes". In: *Nature* 553.7689 (Jan. 24, 2018), pp. 418–426. ISSN: 1476-4687. DOI: 10.1038/nature25022. URL: https://www.nature.com/articles/nature25022 (visited on 04/23/2019).

[77] Henry Lee-Six et al. "Population dynamics of normal human blood inferred from somatic mutations". In: *Nature* 561.7724 (Sept. 2018). Number: 7724 Publisher: Nature Publishing Group, pp. 473–478. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0497-0. URL: https://www.nature.com/articles/s41586-018-0497-0 (visited on 03/02/2020).

[78] Henry Lee-Six et al. "Population dynamics of normal human blood inferred from somatic mutations". In: *Nature* 561.7724 (2018), pp. 473–478.

[79] Tom Lenaerts et al. "Tyrosine kinase inhibitor therapy can cure chronic myeloid leukemia without hitting leukemic stem cells". In: *Haematologica* 95.6 (June 1, 2010). Publisher: Haematologica, pp. 900–907. ISSN: 0390-6078, 1592-8721. DOI: 10.3324/haematol.2009.015271. URL: http://www.haematologica.org/content/95/6/900 (visited on 05/25/2020).

[80] Smadar Ben-Tabou de-Leon and Eric H Davidson. "Gene regulation: gene control network in development". In: *Annu. Rev. Biophys. Biomol. Struct.* 36 (2007), pp. 191–212.

[81] Linheng Li and Hans Clevers. "Coexistence of quiescent and active adult stem cells in mammals". In: *science* 327.5965 (2010), pp. 542–545.

[82] Wing-Cheong Lo et al. "FEEDBACK REGULATION IN MULTISTAGE CELL LINEAGES". In: *Mathematical biosciences and engineering : MBE* 6.1 (Jan. 2009), pp. 59–82. ISSN: 1547-1063. DOI: 10.3934/mbe.2009.6.59. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2756546/ (visited on 04/23/2019).

[83] Michael Loschi et al. "Impact of eculizumab treatment on paroxysmal nocturnal hemoglobinuria: a treatment versus no-treatment study". In: *American Journal of Hematology* 91.4 (2016), pp. 366–370.

[84] Richard Losick and Claude Desplan. "Stochasticity and Cell Fate". In: *Science* 320.5872 (Apr. 4, 2008), pp. 65–68. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1147888. URL: http://science.sciencemag.org/content/320/5872/65 (visited on 11/13/2018).

[85] Lucio Luzzatto. "Paroxysmal nocturnal hemoglobinuria: an acquired X-linked genetic disease with somatic-cell mosaicism". In: *Current Opinion in Genetics & Development*. Genetics of disease 16.3 (June 1, 2006), pp. 317–322. ISSN: 0959-437X. DOI: 10.1016/j.gde.2006.04.015. URL: http://www.sciencedirect.com/science/article/pii/S0959437X06000761 (visited on 08/17/2020).

[86] Lucio Luzzatto and Monica Bessler. "The dual pathogenesis of paroxysmal nocturnal hemoglobinuria". In: *Current Opinion in Hematology* 3.2 (1996), pp. 101–110.

[87] Michael Lynch. "Rate, molecular spectrum, and consequences of human mutation". In: *Proceedings of the National Academy of Sciences* 107.3 (2010), pp. 961–968.

[88] Jaroslaw P Maciejewski et al. "Impaired hematopoiesis in paroxysmal nocturnal hemoglobinuria/aplastic anemia is not associated with a selective proliferative defect in the glycosylphosphatidylinositol-anchored protein-deficient clone". In: *Blood, The Journal of the American Society of Hematology* 89.4 (1997), pp. 1173–1181.

[89] Andreas-Holger Maehle. "Ambiguous cells: the emergence of the stem cell concept in the nineteenth and twentieth centuries". In: *Notes and Records of the Royal Society* 65.4 (2011), pp. 359–378.

[90] Anna Marciniak-Czochra, Thomas Stiehl, and Wolfgang Wagner. "Modeling of replicative senescence in hematopoietic development". In: *Aging (Albany NY)* 1.8 (July 23, 2009), pp. 723–732. ISSN: 1945-4589. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2830082/ (visited on 04/25/2019).

[91] Anna Marciniak-Czochra et al. "Modeling of Asymmetric Cell Division in Hematopoietic Stem Cells—Regulation of Self-Renewal Is Essential for Efficient Repopulation". In: *Stem Cells and Development* 18.3 (Aug. 27, 2008), pp. 377–386. ISSN: 1547-3287. DOI: 10.1089/scd.2008.0143. URL: https://www.liebertpub.com/doi/abs/10.1089/scd.2008.0143 (visited on 03/07/2019).

[92] Iñigo Martincorena et al. "High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237 (2015), pp. 880–886.

[93] Iñigo Martincorena et al. "Somatic mutant clones colonize the human esophagus with age". In: *Science* 362.6417 (2018), pp. 911–917.

[94] Armand Mekontso Dessap et al. "Telomere attrition in sickle cell anemia". In: *American Journal of Hematology* 92.6 (June 1, 2017). Publisher: John Wiley & Sons, Ltd, E112–E114. ISSN: 0361-8609. DOI: 10.1002/ajh.24721. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/ajh.24721 (visited on 06/16/2020).

[95] Donald Metcalf. "On hematopoietic stem cell fate". In: *Immunity* 26.6 (2007), pp. 669–673.

[96] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. "Transition states and cell fate decisions in epigenetic landscapes". In: *Nature Reviews Genetics* 17.11 (Nov. 2016), pp. 693–703. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.98. URL: https://www.nature.com/articles/nrg.2016.98 (visited on 11/12/2018).

[97] Sean J Morrison, Nirao M Shah, and David J Anderson. "Regulatory mechanisms in stem cell biology". In: *Cell* 88.3 (1997), pp. 287–298.

[98] Y Mortazavi et al. "N-RAS gene mutation in patients with aplastic anemia and aplastic anemia/paroxysmal nocturnal hemoglobinuria during evolution to clonal disease". In: *Blood, The Journal of the American Society of Hematology* 95.2 (2000), pp. 646–650.

[99] Khedoudja Nafa et al. "Mutations in the PIG-A gene causing paroxysmal nocturnal hemoglobinuria are mainly of the frameshift type". In: (1995).

[100] Sonia Nestorowa et al. "A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation". In: *Blood* 128.8 (Aug. 25, 2016). Publisher: American Society of Hematology, e20–e31. ISSN: 0006-4971. DOI: 10.1182/blood-2016-05-716480. URL: https://ashpublications.org/blood/article/

128/8/e20/35749/A-single-cell-resolution-map-of-mouse (visited on 06/18/2020).

[101] Ashley P. Ng and Warren S. Alexander. "Haematopoietic stem cells: past, present and future". In: *Cell Death Discovery* 3.1 (Feb. 6, 2017). Number: 1 Publisher: Nature Publishing Group, pp. 1–4. ISSN: 2058-7716. DOI: 10.1038/cddiscovery.2017.2. URL: https://www.nature.com/articles/cddiscovery20172 (visited on 06/18/2020).

[102] Faiyaz Notta et al. "Distinct routes of lineage development reshape the human blood hierarchy across ontogeny". In: *Science* 351.6269 (Jan. 8, 2016), aab2116. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab2116. URL: http://science.sciencemag.org/content/351/6269/aab2116 (visited on 08/07/2018).

[103] Faiyaz Notta et al. "Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment". In: *Science* 333.6039 (2011), pp. 218–221.

[104] SB Oni, BO Osunkoya, and L Luzzatto. "Paroxysmal nocturnal hemoglobinuria: evidence for monoclonal origin of abnormal red cells". In: *Blood* 36.2 (1970), pp. 145–152.

[105] Fernando G Osorio et al. "Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis". In: *Cell reports* 25.9 (2018), pp. 2308–2316.

[106] Sarah P Otto and Troy Day. *A biologist's guide to mathematical modeling in ecology and evolution*. Princeton University Press, 2011.

[107] Jorge M. Pacheco et al. "Cyclic neutropenia in mammals". In: *American Journal of Hematology* 83.12 (2008). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.21295, pp. 920–921. ISSN: 1096-8652. DOI: 10.1002/ajh.21295. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajh.21295 (visited on 05/25/2020).

[108] Artur Pappenheim. *Atlas der menschlichen Blutzellen*. Vol. 1. Fischer, 1905.

[109] Charles Parker et al. "Diagnosis and management of paroxysmal nocturnal hemoglobinuria". In: *Blood* 106.12 (2005), pp. 3699–3709.

[110] John W Pepper and Matthew D Herron. "Does biology need an organism concept?" In: *Biological Reviews* 83.4 (2008), pp. 621–627.

[111] Thomas A Rando. "The immortal strand hypothesis: segregation and reconstruction". In: *Cell* 129.7 (2007), pp. 1239–1243.

[112] Hannes Risken. "Fokker-planck equation". In: *The Fokker-Planck Equation*. Springer, 1996, pp. 63–95.

[113] L. Robb. "Cytokine receptors and hematopoietic differentiation". In: *Oncogene* 26.47 (Oct. 2007). Number: 47 Publisher: Nature Publishing Group, pp. 6715–6723. ISSN: 1476-5594. DOI: 10.1038/sj.onc.1210756. URL: https://www.nature.com/articles/1210756 (visited on 02/27/2020).

[114] Ingo Roeder and Markus Loeffler. "A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity". In: *Experimental Hematology* 30.8 (Aug. 1, 2002), pp. 853–861. ISSN: 0301-472X. DOI: 10.1016/S0301-472X(02)00832-9. URL: http://www.sciencedirect.com/science/article/pii/S0301472X02008329 (visited on 04/23/2019).

[115] WF Rosse, JV Dacie, et al. "Immune lysis of normal human and paroxysmal nocturnal hemoglobinuria (PNH) red blood cells. I. The sensitivity of PNH red cells to lysis by complement and specific antibody." In: *The Journal of clinical investigation* 45.5 (1966), pp. 736–748.

[116] Donald B Rubin. "Bayesianly justifiable and relevant frequency calculations for the applies statistician". In: *The Annals of Statistics* (1984), pp. 1151–1172.

[117] Stephan C Schuster. "Next-generation sequencing transforms today's biology". In: *Nature methods* 5.1 (2008), pp. 16–18.

[118] Wenyi Shen et al. "Deep sequencing reveals stepwise mutation acquisition in paroxysmal nocturnal hemoglobinuria". In: *The Journal of clinical investigation* 124.10 (2014), pp. 4529–4538.

[119] Bryan E Shepherd et al. "Estimating human hematopoietic stem cell kinetics using granulocyte telomere lengths". In: *Experimental hematology* 32.11 (2004), pp. 1040–1050.

[120] Hidetoshi Shimodaira. "An approximately unbiased test of phylogenetic tree selection". In: *Systematic biology* 51.3 (2002), pp. 492–508.

[121] Liran I Shlush. "Age-related clonal hematopoiesis". In: *Blood* 131.5 (2018), pp. 496–504.

[122] Louis Siminovitch, Ernest A McCulloch, and James E Till. "The distribution of colony-forming cells among spleen colonies". In: (1963).

[123] Gérard Socié et al. "Changing prognosis in paroxysmal nocturnal haemoglobinuria disease subcategories: an analysis of the International PNH Registry". In: *Internal medicine journal* 46.9 (2016), pp. 1044–1053.

[124] Gerald J Spangrude, Shelly Heimfeld, and Irving L Weissman. "Purification and characterization of mouse hematopoietic stem cells". In: *Science* 241.4861 (1988), pp. 58–62.

[125] Thomas Stiehl and Anna Marciniak-Czochra. "Characterization of stem cells using mathematical models of multistage cell lineages". In: *Mathematical and Computer Modelling*. Mathematical Methods and Modelling of Biophysical Phenomena 53.7 (Apr. 1, 2011), pp. 1505–1517. ISSN: 0895-7177. DOI: 10.1016/j.mcm.2010.03.057. URL: http://www.sciencedirect.com/science/article/pii/S0895717710001755 (visited on 04/25/2019).

[126] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[127] C Sugimori et al. "Paroxysmal nocturnal hemoglobinuria and concurrent JAK2 V617F mutation". In: *Blood cancer journal* 2.3 (2012), e63–e63.

[128] Kaoru Sugimoto, Sean P Gordon, and Elliot M Meyerowitz. "Regeneration in plants and animals: dedifferentiation, transdifferentiation, or just differentiation?" In: *Trends in cell biology* 21.4 (2011), pp. 212–218.

[129] Jianlong Sun et al. "Clonal dynamics of native haematopoiesis". In: *Nature* 514.7522 (Oct. 2014), pp. 322–327. ISSN: 1476-4687. DOI: 10.1038/nature13824. URL: https://www.nature.com/articles/nature13824 (visited on 12/19/2018).

[130] Evgeny Tankhilevich et al. "GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation". In: *Bioinformatics* 36.10 (2020), pp. 3286–3287.

[131] James E Till and Ernest A McCulloch. "A direct measurement of the radiation sensitivity of normal mouse bone marrow cells". In: *Radiation research* 14.2 (1961), pp. 213–222.

[132] James E Till and Ernest A McCulloch. "Hemopoietic stem cell differentiation". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 605.4 (1980), pp. 431–459.

[133] James E Till, Ernest A McCulloch, and Louis Siminovitch. "A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells". In: *Proceedings of the National Academy of Sciences of the United States of America* 51.1 (1964), p. 29.

[134] Cristian Tomasetti and Ivana Bozic. "The (not so) immortal strand hypothesis". In: *Stem cell research* 14.2 (2015), pp. 238–241.

[135] Arne Traulsen, Jorge M Pacheco, and David Dingli. "On the origin of multiple mutant clones in paroxysmal nocturnal hemoglobinuria". In: *Stem Cells* 25.12 (2007), pp. 3081–3084.

[136] Arne Traulsen et al. "Somatic mutations and the hierarchy of hematopoiesis". In: *Bioessays* 32.11 (2010), pp. 1003–1008.

[137] Lars Velten et al. "Human haematopoietic stem cell lineage commitment is a continuous process". In: *Nature cell biology* 19.4 (2017), pp. 271–281.

[138] Bert Vogelstein and Kenneth W Kinzler. "The multistep nature of cancer". In: *Trends in genetics* 9.4 (1993), pp. 138–141.

[139] Russell E Ware, Sharon E Hall, and Wendell F Rosse. "Paroxysmal nocturnal hemoglobinuria with onset in childhood and adolescence". In: *New England Journal of Medicine* 325.14 (1991), pp. 991–996.

[140] Meltem Weger et al. "Stem cells and the circadian clock". In: *Developmental biology* 431.2 (2017), pp. 111–123.

[141] Benjamin Werner and Andrea Sottoriva. "Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates". In: *PLoS computational biology* 14.6 (2018), e1006233.

[142] Benjamin Werner, Arne Traulsen, and David Dingli. "Ontogenic growth as the root of fundamental differences between childhood and adult cancer". In: *Stem Cells* 34.3 (2016), pp. 543–550.

[143] Benjamin Werner et al. "Dynamics of mutant cells in hierarchical organized tissues". In: *PLoS Comput Biol* 7.12 (2011), e1002290.

[144] Benjamin Werner et al. "Measuring single cell divisions in human tissues from multi-region sequencing data". In: *Nature Communications* 11.1 (Feb. 25, 2020). Number: 1 Publisher: Nature Publishing Group, pp. 1–9. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14844-6. URL: https://www.nature.com/articles/s41467-020-14844-6 (visited on 03/06/2020).

[145] Benjamin Werner et al. "Measuring single cell divisions in human tissues from multi-region sequencing data". In: *Nature communications* 11.1 (2020), pp. 1–9.

[146] Marc J Williams et al. "Quantification of subclonal selection in cancer from bulk sequencing data". In: *Nature genetics* 50.6 (2018), pp. 895–903.

Part II.

Statistical mechanics of proliferating cells

8. Cell movement as a stochastic process

You have brains in your head.
You have feet in your shoes.
You can steer yourself any direction you choose.
You're on your own. And you know what you know.
And YOU are the guy who'll decide where to go.
— Dr. Seuss, *Oh! The Places You'll Go!*

Many tasks in the human body require cells to move around, often as part of a coordinated effort to achieve objectives on the tissue level. Obvious examples of this are found in the earliest stages of human life, during embryonic development and tissue growth [65]. However even in adults such migratory processes can be identified, for example in wound healing [19] and vascularization (the creation of blood vessels) [45, 48]. Understanding and characterizing the properties of the underlying cell migration, both in the context of the individual cells as well as the collective population and its environment, can thus lead to important insights in the fields of medicine and human biology. One particular application in which the relevance of cell motility has only recently surfaced is cancer, specifically in the late stage of the disease when additional tumors begin to arise at different locations in the body [27], where the successful migration of cancerous cells from the primary location appears to be aided by a phenotypic transformation to a more motile morphology [40]. As the relevance of this process currently remains a topic of some debate [57], quantitative models for the cells' movement can greatly aid in designing and interpreting experiments [43, 36]. However, the cells which make up the tissues, organs, malignant tumors, or other collective processes we may be interested in are in themselves highly complex organisms, so that from the modeling perspective

we once again run into the now familiar problem of scale: observed phenomena on the population level may arise due to many interactions on the single cell level, whose behavior is in itself similarly emergent from a multitude of processes on a subcellular scale. Thus, similar to how the divisional behavior of hematopoietic cells was modeled in the previous chapters, single cells must be envisioned as black box particles endowed with only a select set of properties that take on a stochastic character.

With possible applications to cell motion in mind, in this part we will make a brief foray into the field statistical mechanics and its application to biological components, which in a sense leans closely to the field of *active matter* [44] – the study of population mechanics in which the particles of interest transform energy from their environment into movement. While the statistical study of biological agents such as cells has advanced rapidly [3], applications in cancer have been somewhat lacking until recently [68]. For this reason it may be of use to obtain a better understanding of the influence of certain fundamental differences found in the context of malignant tumors compared to other cell systems. In particular, inspired by recent experimental developments [35, 68] we will investigate basic statistical properties of motility in a growing population that is spatially confined. While this has been previously done for a specific model of active matter – the *run-and-tumble* particle [7] – there is little known about the influence of growth in a more generic context. To expand and generalize our understanding of this phenomenon we will investigate proliferation effects in what is considered the quintessential model of stochastic movement: Brownian motion. In this chapter we will motivate and discuss the application of statistical mechanics in the context of cancer, as well as introduce the basic mathematical formalism of the *Langevin equation* by which we will develop our treatment in Chapter 9. The model we will introduce and discuss there entails a general method for including growth and its resulting crowding effects in a statistical mechanics context, whereby specifics of the described moving "particles" are left open. This agnostic approach to the nature of the system's constituents means that it can be applied in many different contexts besides cellular populations, and as such, the

cancer cell system described in the following section should be considered more as a motivating example of a possible application, rather the ultimate goal of the forthcoming mathematical formulation.

8.1. Motility in cancer: a motivating example

Cancer is perhaps the most widely studied disease of the past century. While from a clinical perspective it is more appropriate to consider it a collection of illnesses – the specific type depending on its source tissue and the particular properties it has acquired during its development – the general nature and origin of cancers allows for a rather simple definition: it is the unchecked proliferation of a population of corrupted cells within the body, caused by the accumulation of specific somatic mutations which facilitate the escape from the cell's preprogrammed purpose within the collective tissue [66, 21]. The fact that it arises from the host's own tissue forms an important part of its success at avoiding destruction, as a cancerous cell population can unlock many of the mechanisms and processes encoded in the human genome to serve its own survival. Nevertheless, the body has many built-in defense mechanisms against cancer which must be overcome before a true tumor is formed, from a low mutation rate and robust DNA repair mechanisms to apoptosis programs (self-induced cell death), and even a well-prepared immune system [1, 50, 66]. In fact, a single mutation – even in a large group of cells – is in itself not sufficient to provoke the invasive and disruptive behavior seen in cancer. Instead a handful of mutations in key genes – so called *oncogenes* – are typically required for a population of cells to be considered a tumor [63]. Unfortunately, this generally means that by the time a cancer is detected, it is already a highly evolved malignant system, with many properties and abilities which are difficult to combat. In two seminal papers [22, 21] Douglas Hanahan and Robert A. Weinberg classified the capabilities of cancers in a number of proposed *hallmark* traits acquired during oncogenesis (the process of transformation from a population of normal cells to a cancer), each representing a distinct barrier the developing malignancy must overcome in order to continue is expansion. One of these is *metastasis* – the spread of the cancer from its original location (in solid

tumors) to colonies at distant locations in the body [27, 20]. Despite being the only non-essential hallmark for the formation of a tumor – as some tumors never progress to this stage – it may be considered the most deadly, given the bleak prognosis of metastatic disease [53]. Unfortunately it has proven to be incredibly difficult to combat, with the metastatic state associated with increased likeliness of treatment failures and acquired drug resistances [31]. While the general process appears superficially the same for most cancer types – a sequence of events involving malignant cells which detach from the primary tumor and enter the circulatory system, exit though the capillaries into distant tissues and there successfully establish new cancerous microenvironments [42] – few specific pathogenic principles have been identified which are shared across different tumor types, and the biological underpinnings of the various steps in the process are still poorly comprehended [53]. Some progress towards understanding the pathogenicity has been made though, with one particular cellular program linked to the metastatic capability of many cancer types [12] – called the *epithelial-to-mesenchymal-transition* – which occurs naturally in the body during embryonic development as well as wound healing [29, 58]. It constitutes a phenotypic transformation of epithelial cells from their normal state to a more mobile build (a *mesenchymal* cell) that facilitates migratory behavior [40]. Cancers can co-opt this process, whereby it appears the mesenchymal phenotype aids in the invasion of distant tissues. In this context the cell's mobility becomes a relevant area of interest. Even if the success of EMT in metastatic cancers is unrelated to the associated increased motility – the EMT has been shown to grant other situational benefits to the cancer cell such as an increased resistance to chemotherapy [52] and evasion of the adaptive immune system [57, 13] – quantitative models and a heuristic understanding of the movement are useful for extracting valuable information from various experimental setups.

The inspiration for this particular project – and an example of a possible application of the model we will develop in Chapter 9 – is an experimental design by Lin et al. [35], in which the behavior is studied of individual cells from a prostate cancer metastasis derived cell-line PC3 in a complex drug landscape, using a state of the art

microfluidic cell culture device. The cell population was divisible into two phenotypes, the first – denoted as PC3-EPI – is an E-cadherin/CDH1 positive/vimentin negative PC3 clone, while the second – denoted here as PC3-EMT – has an E-cadherin negative/vimentin positive phenotype that can be induced by culturing in the presence of human macrophages. While the former is a cancer cell of epithelial origin, the latter – being a highly mobile and less proliferative phenotype – is characterized as having gone through the epithelial-to-mesenchymal transition resulting in its enhanced movement capabilities. It is hypothesized that this phenotype may prove advantageous for the cell in initiating the first steps of metastasis – such as detachment from the primary lesion and intravasation in the bloodstream – as well as assist it in escaping immune surveillance. A mixed population of both cell types was placed in a heterogeneous surface environment in the form of a gradient of docetaxyl (Figure 8.1b) – a chemotherapeutic drug – and monitored over the course of a few weeks. Both phenotypes suffered effects of the drug induced stress, especially in areas containing high densities of docetaxyl, however the PC3-EMT cells were able to sustain their presence under higher doses than the PC3-EPI phenotype. The use of a state of the art monitoring system allowed for the precise visualization of the system at short time intervals, including the ability to track individual cells over long periods of time (see Figure 8.1a). The resulting single cell path measurements illustrate a highly stochastic picture of motion, where the motility of cells correlates not only with the subtype but also spatially with the drug concentration. In particular, the cells' *mean squared displacement* (see Section 8.3.1) presents a sublinear character as shown in Figure 8.1c, presumably due to growth of the cell populations. The authors furthermore document interesting phenomena such as varying motility distributions and the formation of protective niches for non-drug resistant cells. However, distinguishing the contributions of the different experimental factors – the phenotypic motility, the local drug concentration, and the local population density – to the individual cell behaviors is difficult to realize from a purely qualitative analysis. A model of stochastic motion which includes the proliferative behavior would provide a first step in separating the causal relationships between the observed phenomena.

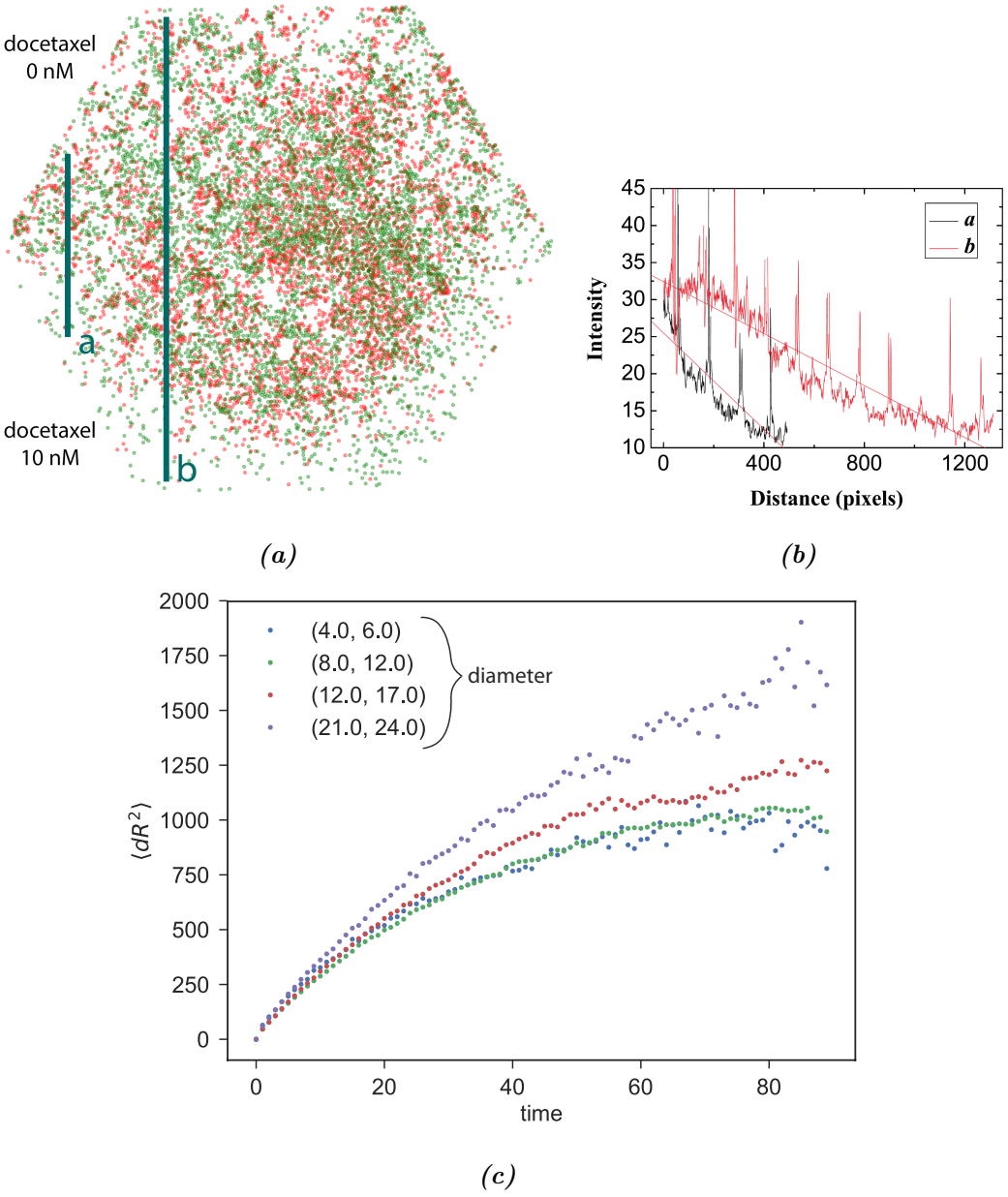


Figure 8.1.: Experimental setup of Lin et al. [35]. *(a) The positions of cells on the hexagonal surface of the culturing device. Untransformed PC3 cells (epithelial phenotype) are depicted as red circles, while EMT transitioned cells (mesenchymal phenotype) are shown in green. The applied drug increases linearly from top to bottom. (b) Intensity of the drug according to the lines* a *and* b *shown in (a) (reproduced from [35]). Measurement of the mean squared displacement over time of cells grouped per diameter. All groupings clearly diffuse in a sublinear manner.*

8.2. Cells as motile particles

Cells are incredibly complex organisms. Across an enormous multitude of phenotypically differing types they can perform a wide variety of tasks, from sending and detecting signals to and from their surroundings, to initiating metabolic processes with reactants from their environment, with some types even capable of certain forms of self-propelled motion [1]. As such, their behavior is driven by an immense number of internal processes. A population level model which attempts to take into account the interactions and internal decision making of all of its constituent cells would clearly devolve into an intractable mess long before any useful information can be gleaned from it. Instead we turn to statistical modeling, where only a handful of key properties are taken into account and the driving processes are chosen to follow some (sometimes empirically determined) probability distributions. Arguably the earliest of such approaches is the now widely known model for Brownian motion developed by Albert Einstein at the start of the 20th century [16], which describes the apparent random movement of a pollen suspended in a liquid due to stochastically varying force interactions with the (then still hypothetical) atoms within the liquid. This type of approach later found its way into many applications in physics, from classical statistical mechanics – Newtonian systems with too many individual parts to resolve all interactions deterministically – to quantum field theory and astronomy [32], resulting in a wealth of useful methods and techniques. Their application in biological systems is a more recent topic of interest. While biochemical systems with more fundamental building blocks – such as for example the diffusion of macromolecules within cellular compartments [2, 10, 11] – have proven quite amenable to this approach, systems on larger scales (e.g. bacteria, animals, etc.) have shown to require some rethinking of the processes applied. In particular, certain fundamental differences are apparent between the previously studied physical systems of "simple" interacting particles and those typically found in the microscopic or even macroscopic biological world. The foremost, which has already been discussed to an extent, is the complexity. While the properties of the fundamental particles and interactions in physical systems can often be summarized exactly in a handful of equations, with the number

of components under consideration leading to the system's intractability; the biological "particles" we consider are already highly complex machines, which we only approximate as black boxes with simple interactions. This adds a layer of additional questions and problems that must be addressed, involving which approximations to make and the extent of their validity under particular circumstances. In the case of describing their movement there is another perhaps more subtle difference, in the fact that the particles of interest in biological systems are typically self-propelled, i.e. their motion is driven by an internal process, whereas the particles in physical systems are generally moved by external forces. This is an important distinction, as it means that the biological particles must constantly consume and expend energy. It can result in more complicated patterns of motion [47] and furthermore implies that unless some external energy source is modeled – which complicates the construction of tractable models and reduces our ability to obtain general principles – the very useful property of energy conservation is effectively violated [5]. In a system of stochastic motion, elementary units with such an internal driver are typically referred to as *active* particles, as opposed to the *passive* particles whose movement is entirely attributed to external forces [44]. It is clear that for some applications certain cells may fall under the active moniker, given their ability to obtain energy from nutrients and, depending on their morphology and function, their capabilities of self-propulsion.

8.3. Basics of stochastic motion

As discussed previously, the stochastic motion of a particle of interest is principally an emergent phenomenon, caused by an underlying process which – due to its complexity or some lack of information – cannot be resolved exactly. In this sense the stochastic model is an approximation of the true system, one that importantly does not specify a deterministic prediction of the particle's future state. Instead, its purpose is to provide the likelihood of possible futures, which can be used to predict quantities related to the collective behavior of many particles, or the average behavior of a single particle over a longer period of time. An example of this is diffusion: the well-known expression for the

spatial density $n(x)$ of a collection of particles with diffusion coefficient $\mathcal{D}$

$$\frac{\partial n(x,t)}{\partial t} = \mathcal{D}\nabla^2 n(x,t) \tag{8.1}$$

can be obtained from certain models of stochastic motion of single particles by considering the time evolution of their position probability distribution. Similarly one might examine predicted distributions of the particle velocities, energies, or any other quantities which may be of interest for the system in question.

8.3.1. Brownian motion

One of the most ubiquitously applied models of random movement is that of Brownian motion. Its prevalence, while in part due to historical reasons, may also be attributed to the fact that its associated stochastic fluctuations have turned out to carry a fundamental significance in stochastic processes in general, being the only *Lévy process* with continuous paths [30]. Here we will introduce the concepts of Brownian motion which will be used in the following chapter.

The original purpose of the model was the quantitative description of the seemingly random motion of a pollen immersed in water, described in 1827 by the botanist Robert Brown to which the process owes its name. Einstein's original description [16] relied on the supposition that the motion was caused by interactions of the pollen with the water molecules surrounding it. In it he considered the time-dependent displacement of a large density of Brownian particles $n(x,t)$, which in a contemporary context describes a Fokker-Planck formulation of the stochastic process [46]. A more modern definition of Brownian motion is that it is a Gaussian Markov process with stationary independent increments [15]. Specifically, this means that if $B(t)$ is a Brownian motion (i.e. the position of the particle at time $t \geq 0$), it has the properties:

(i) $B(t_0)$, $B(t_1) - B(t_0)$, $\ldots$, $B(t_n) - B(t_{n-1})$ are independent for $t_0 < t_1 < \cdots < t_n$.

(ii) $B(t) - B(s)$ is normally distributed with mean 0 and variance $t - s$ for $0 < s < t$.

(iii) $t \mapsto B(t)$ is continuous with probability 1.

As done in Chapter 3, for our purposes we will eschew further levels of mathematical rigor in favor of a more physical approach, which will facilitate the introduction of a growing population in Chapter 9. In particular, we will study the process through the method first presented by Langevin [33] merely a few years after Einstein's treatment, which introduces the use of a *stochastic differential equation* (SDE). Rather than model the displacement for a statistical ensemble of particles, Langevin formulated the force acting on a single particle in order to obtain the dynamics of a possible trajectory. To this end he introduced the differential equation of the form:

$$m\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\gamma v(t) + \mathcal{F}(t) \tag{8.2}$$

which contains a driving stochastically fluctuating force $\mathcal{F}(t)$ and a viscous frictional force $\gamma v(t)$ with linear velocity dependence. This was the first of a class of equations known as *Langevin equations* (LEs). It may be interpreted in one or more spatial dimensions, in which case we will note the components as $v_i(t)$. Since (8.2) contains a random variable its integral is in itself a random value. The mathematical framework for rigorously constructing such an SDE is somewhat technical and involves selecting one of two possible definitions – referred to as the *Itô* and *Stratonovich* formulations – for the integral of a random variable [61]. We will not go into detail on this subject, and instead simply note that where relevant we take the Itô interpretation, and refer the reader to a more comprehensive reference such as [41] for more information.

As a phenomenological model it is perhaps not immediately clear how (8.2) exactly represents the system of a microscopic pollen being bombarded by millions of nanoscopic molecules, however we will later show how the friction and the driving force together represent this single phenomenon.

The stochastic random force $\mathcal{F}(t)$ is specified to be Gaussian distributed. While this requirement was shown by Ornstein and Uhlenbeck some years later [60], it may somewhat superficially be argued from the central limit theorem: the large number of collisions occurring in a short timespan guarantees that the total intensity and direction of the resulting force converges to a Gaussian. The mean must be zero, so that the expected position of the particle to be fixed in time (as specified in *(ii)* of the formal

definition given above), and the variance is determined by the force's autocorrelation function, which is taken to be a delta-function to ensure independent increments in every direction, (from *(i)* of the formal definition). Introducing $\xi(t)$ as the noise term with intensity 1, we write $\mathcal{F}(t) = \sqrt{2D}\,\xi(t)$, where the parameter D thus characterizes the strength of the fluctuating force, which is the same in every direction. We then have

$$\begin{aligned} \langle \xi(t) \rangle &= 0 \\ \left\langle \xi_i(t)\xi_j(t') \right\rangle &= \delta(t-t')\delta_{i,j} \end{aligned} \tag{8.3}$$

with the $\xi_i(t)$ the components in each spatial dimension. For simplicity we will set the particle mass m to 1, so that we may restate 8.2 as

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\gamma v(t) + \sqrt{2D}\,\xi(t) \tag{8.4}$$

Velocity autocorrelation

We can integrate (8.4) to obtain [46]

$$v(t) = v_{t_0} e^{-\gamma(t-t_0)} + \sqrt{2D} \int_{t'=t_0}^{t} e^{-\gamma(t-t')} \xi(t') dt' \tag{8.5}$$

The first term shows how the initial velocity is lost over time, meaning that if we are far enough away from this initial time point, i.e. $t - t_0 \gg 1/\gamma$ it goes to zero and we may simplify the expression

$$v(t) = \sqrt{2D} \int_{t'=t_0}^{t} e^{-\gamma(t-t')} \xi(t') dt' \tag{8.6}$$

With (8.3) this can be used to calculate the the velocity's autocorrelation function

$$\langle v(t)v(t+\tau) \rangle = \int_{t'=t_0}^{t} \int_{t''=t_0}^{t} e^{-\gamma(2t+\tau-t'-t'')} \left\langle \xi(t')\xi(t'') \right\rangle \, dt' dt'' \tag{8.7}$$

which by (8.3) results in

$$\langle v(t)v(t+\tau) \rangle = \frac{dD}{\gamma} e^{-\gamma\tau} \tag{8.8}$$

where d is the number of spacial dimensions. Thus, while the stochastic force pushes were taken to be uncorrelated, the particle's velocity does contain memory, which decreases at an exponential rate (not coincidentally the rate at which the initial velocity is lost in (8.5)) with correlation time $\tau = 1/\gamma$.

Fluctuation-dissipation relation

From (8.6) we see that the average velocity disappears, while the variance σ_v^2 results in

$$\sigma_v^2(t) = \left\langle v^2(t) \right\rangle - \langle v(t) \rangle^2 = \left\langle v^2(t) \right\rangle \tag{8.9}$$

Thus from (8.8) we have that the first moment of the velocity can be written as $\left\langle v^2(t) \right\rangle = dD/\gamma$. Given that we may write the average energy of the particle as $E = \left\langle v^2(t) \right\rangle /2$ (with $m = 1$) we obtain the relation

$$E = \frac{dD}{2\gamma} \tag{8.10}$$

which relates the strength of the random force pushes D to the friction coefficient γ through the particle's average energy. This is known as a *fluctuation-dissipation relation*, and it is the mathematical formulation of the fact that, as hinted earlier, the stochastic and friction terms in (8.2) act as two sides of the same coin: the energy given to the particle through $\mathcal{F}(t)$ is compensated by the energy lost due to friction.

Speed distribution

It can be shown (most easily by transforming to a Fokker-Planck description [46, 47]) that as the memory of any initial velocities are lost (8.8) the full probability density describing the Brownian particle's velocity components converges to the Maxwell-Boltzmann distribution [49]. While the average of the velocity $v_i(t)$ is then zero (as shown previously from (8.6)), the average speed of the particle – defined as $s(t) = |v(t)| = \sqrt{\sum_i v_i^2}$ – does not. For example, in two dimensions the speed is given by the Rayleigh distribution [47]

$$\mathbb{P}\{s\} = \frac{\gamma}{D} s \exp\left(-\frac{\gamma s^2}{2D}\right) \tag{8.11}$$

which has moments

$$\langle s \rangle = \sqrt{\frac{D}{\gamma}} \sqrt{\frac{\pi}{2}} \tag{8.12}$$

and

$$\left\langle s^2 \right\rangle = 2\frac{D}{\gamma} \tag{8.13}$$

Mean squared displacement

An important metric for characterizing types of stochastic motion is the displacement of a particle over time. While the expected value of a particle's position is clearly fixed in time (from the fact that $\langle v(t)\rangle = 0$), the variance is not, meaning that the second moment of the particle's position $x^2(t)$ encodes some information about its probabilistic displacement. The famously linear character of the mean squared displacement [16] can be found by twice integrating (8.6) to obtain [46]

$$\left\langle x^2(t)\right\rangle = 2d\frac{D}{\gamma^2}\left[t - \frac{1}{\gamma}\left(1 - e^{-\gamma t}\right)\right] \tag{8.14}$$

which reduces to $\left\langle x^2(t)\right\rangle = (2dDt/\gamma^2)t$ in the long time limit $t \gg 1/\gamma$. This exactly matches Einstein's initial result for the diffusion of a Brownian particle $\left\langle x^2(t)\right\rangle = 2d\mathcal{D}t$ [16], which allows us to identify the diffusion coefficient of (8.1) as

$$\mathcal{D} = D/\gamma^2 \tag{8.15}$$

The linear form of the mean squared displacement is an important result, as it encodes the classical type of diffusion we are familiar with. Its ubiquity has led to systems which do not present this behavior to be referred to as *sublinear* – if the diffusion occurs on a slower than linear scale – or *superlinear* – if the mean squared displacement grows faster. Similarly, particles whose displacement is proportional to time (instead of the square root) are referred to as *ballistic*.

Alternative formulations

As discussed at the start of this section, other equivalent formulations of Brownian motion can be constructed, allowing the researcher to pick whichever is most convenient for the problem they wish to tackle. Perhaps the most common alternate formulation is through the Fokker-Planck equation, which has already been referred to before. While the Langevin equation describes the stochastic evolution of a single particle, the Fokker-Planck equation gives the deterministic evolution of the associated probability distribution. In fact, it can be shown that for every Langevin equation there is an

equivalent Fokker-Planck equation [46]. In particular, for a general LE of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = h(x,t) + g(x,t)\xi(t) \tag{8.16}$$

the following Fokker-Planck equation can be constructed (in the Itô convention) [46]

$$\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}[h(x,t)P(x,t)] + \frac{\partial^2}{\partial x^2}\left[\frac{g^2(x,t)}{2}P(x,t)\right] \tag{8.17}$$

where we have used the shorthand $P(x,t) = \mathbb{P}\{x,t \mid x_0,t_0\}$ for the probability density. Note that if $h(x,t) = 0$ and $g(x,t)$ is constant, the Fokker-Planck equation reduces to diffusion equation (8.1), where we identify $\mathcal{D} = g^2(x,t)/2$. For the case of the Langevin equation (8.4), taking the so-called *overdamped approximation* [47] – which corresponds to assuming large friction so that inertial effects may be neglected ($\mathrm{d}v(t)/\mathrm{d}t = 0$) – the resulting Langevin equation for the particle position becomes

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \sqrt{\frac{2D}{\gamma^2}}\xi(t) \tag{8.18}$$

which after transformation to the Fokker-Planck equation becomes exactly the diffusion equation (8.1), which again provides the relation (8.15).

Finally, another useful formulation of Brownian motion is that it can also be obtained formally from the limit of a *random walk* if the number paths in a fixed time interval goes to infinity [14, 15], a result known as *Donsker's theorem*. Without going into too much details, the heart of the formulation is the following: For a random walk S_n defined as

$$S_n = \sum_{i=1}^{n} X_i \tag{8.19}$$

with the X_i independent and identically distributed random values, it can be show that the rescaled random walk

$$W_n(t) = \frac{S_{nt}}{\sqrt{n}}, \quad t \in [0,1] \tag{8.20}$$

converges to Brownian motion as $n \to \infty$. This formulation will be illustrative in motivating our use of the motion in Section 9.2.

8.3.2. Generalizations and other models

While the formalization of Brownian motion by Einstein [16] and the conception of the random noise term by Langevin [33] (along with important contributions from contemporaries such as Smolukowski [64] and Wiener [67]) kickstarted the study of statistical motion, the century that followed saw the introduction of many generalizations and related models in fields concerning fluctuations and noise [23]. Some notable examples are the introduction of the *Lévy process* (of which both Brownian motion and the Poisson process are well known examples) which formally generalizes the concept of non-differentiable trajectories [30], the so-called *generalized Langevin equation* [39] which introduces memory into the LE through finitely correlated fluctuations, and the formal development of stochastic differential equations [26, 25, 54], which are ubiquitously used in mathematical finance.

As mentioned in Section 8.2, the most relevant extension of this formalism for cancer cells is perhaps found in *active particle* models, which attempt to include the self-propelled character of certain agents such as cells. While our treatment will not fall entirely in this active category, it is worth briefly describing these extensions. Perhaps the most noteworthy is the class of particles referred to as *active Brownian particles*, whose motion is described by some modified form of (8.4). One such alteration is obtained by allowing a dynamically varying friction coefficient $\gamma(x, v, t)$ which can reach negative values [47]; this effective "negative dissipation" of energy can be interpreted as the result of the active internal motor and the assumption of energy being obtained from the environment, and thus renders the fluctuation-dissipation relation (8.10) invalid. Alternatively, active Brownian models have been proposed where the energy obtained from the environment is stored internally (the *depot model*), which is modeled by a separate balance equation $e(t)$ so that the frictional force becomes $[-\gamma + e(t)]v(t)$ [51]. Besides active Brownian motion, another popular model for cellular locomotion is the *run-and-tumble* model – phenomenologically inspired by the observed erratic movement of Escherichia coli (a famously studied type of bacteria) [8] – which characterizes the particle's movement as short straight lines at constant speed ("runs") that are occa-

sionally interrupted at stochastic intervals by random changes of direction ("tumbles"). Its Langevin equation does not derive from the Brownian form discussed previously, but rather constitutes a two- (or three-) dimensional form of the *telegrapher's equation* [28, 56, 6].

9. Stochastic motion under population growth

9.1. The problem of growth

From the experimental example [35] discussed in Section 8.1, we might conclude that the increased cell density in the confined space reduces the mobility of the individual cells, resulting in the observed sublinear diffusion. The qualitative argument for this is that cells cannot move through or over one another and thus act as obstacles to each other's motion, over time effectively reducing the available space at random positions. Some parallel's can be made to the crowding effects studied in macromolecular solutions – for example the diffusion of macromolecules inside the cellular cytoplasm, which have situationally been show to present anomalous (non-linear) diffusion [2, 17, 37] – however here we are interested in the phenomenon of a dynamically changing density of cells caused by a varying population size. One approach to model crowding would involve defining an interaction potential for particles that run into each other, which can be superimposed on some stochastic motion. Such a system can be probed by evolving a local density $n(x,t)$ in time [9], as done for the run-and-tumble model by Cates et al. in [7]. While this can offer highly useful spatial information of the system, it requires the explicit choice of the single particle dynamics and interaction potential, which can be cumbersome to work with if these take a complicated form.

In this chapter we show that a simple form of dynamic crowding can instead be obtained readily from the Langevin equation (8.4). From its interpretation as describing the dynamics of an ideal gas, we first identify a key physical quantity that is implicitly

encoded and relates to the particle density – the particle mean free time – which we then use to generalize the LE to allow for a time-varying density. The predictions of this LE are compared to simulations of two-dimensional hard disk colliding particles simultaneously undergoing population growth (the specifics of which are described in Section B.1), with highly convergent results.

9.2. Brownian motion in an ideal gas

While in Section 8.3.1 we introduced Brownian motion as a model for the movement of a large particle (a pollen) undergoing a multitude of collisions with the various atoms and molecules surrounding it, it is worth wondering whether it can also serve as a model for the movement of these smaller particles themselves, as in for example an ideal gas. Indeed, the erratic trajectory of a single gas molecule is also caused by collisions with other molecules, and the Maxwell-Boltzmann distribution (8.11) to which the velocities of an ensemble of Brownian particles were shown to converge was in fact originally derived to describe the velocities within an ideal gas [38]. However, there is a notable difference in timescales: The microscopically visible movement of the pollen is caused by the net force of thousands of collisions with the particles in its suspension – since a push from a single molecule would not transfer enough momentum to effect a displacement on the scale of the pollen – which results in the apparent fractal structure of the trajectory. Conversely, a single collision *does* cause noticeable displacement of the ideal gas molecule on its own scale, meaning that its trajectory is perhaps better described by a *random walk*, i.e. a succession of short straight displacements in random directions. Recall however that we saw in Section 8.3.1 that Brownian motion can be constructed from a random walk by taking "infinitely short" straight lines. Thus, it might accurately depict the long-time behavior of this short-time random walk. To see where the two descriptions overlap, we will first derive the velocity autocorrelation for the random walk of the gas particle. To facilitate a comparison with the particle simulations (see Section B.1) we will from this point onward constrain our treatment to movement in two spatial dimensions.

9.2.1. Velocity correlation of the random walk

To obtain the autocorrelation $\langle v(t)v(t+\tau)\rangle$ of our test particle's velocity, we first write out the scalar product explicity:

$$v(t)v(t+\tau) = \cos\theta(\tau)\, s(t)s(t+\tau) \tag{9.1}$$

with $s(t) \equiv \sqrt{v_x(t) + v_y(t)}$ the particle's speed, and $\theta(\tau)$ the angle between the velocities at t and $t+\tau$. In our simple random walk model the particle moves in a straight line at constant speed until it collides, at which point it obtains a new direction and possibly a new speed. If the particle does not undergo a collision in the short time τ, then $\theta(\tau) = 0$, $s(t+\tau) = s(t)$ and thus $v(t)v(t+\tau) = s^2(t)$. If on the other hand the particle has undergone a collision, $\theta(\tau)$ is a random (uniformly distributed) angle such that the many particle average disappears $\langle\cos\theta(\tau)\rangle = 0$. For a large ensemble of particles we can separate the average into those that have and have not collided in τ:

$$\begin{aligned}
\langle v(t)v(t+\tau)\rangle &= \frac{\sum_{\text{no coll}} \cos\theta_i(\tau)\, s_i(t)s_i(t+\tau) + \sum_{\text{coll}} \cos\theta_i(\tau)\, s_i(t)s_i(t+\tau))}{N_{\text{coll}} + N_{\text{no coll}}} \\
&= \frac{\sum_{\text{no coll}} \cos\theta_i(\tau)\, s_i(t)s_i(t+\tau)}{N_{\text{no coll}}} \frac{N_{\text{no coll}}}{N_{\text{coll}} + N_{\text{no coll}}} \\
&\qquad + \frac{\sum_{\text{coll}} \cos\theta_i(\tau)\, s_i(t)s_i(t+\tau)}{N_{\text{coll}}} \frac{N_{\text{coll}}}{N_{\text{coll}} + N_{\text{no coll}}} \\
&= \langle\cos\theta(\tau)\, s(t)s(t+\tau)\rangle_{\text{no coll}}\, \mathbb{P}\{\text{no collision in } \tau\} \\
&\qquad + \langle\cos\theta(\tau)\, s(t)s(t+\tau)\rangle_{\text{coll}}\, \mathbb{P}\{\text{collision in } \tau\} \\
&= \left\langle s^2(t)\right\rangle \mathbb{P}\{\text{no collision in } \tau\}
\end{aligned} \tag{9.2}$$

Where the final equality comes from the assumption that the new angle after a collision is independent of the speed. So what is the probability of no collision occurring in this finite time τ? A first guess would be that – assuming the gas is in thermal equilibrium – the occurrence of collisions with a selected particle is a Poisson process (see Section 3.1.2), which would make the time between collisions exponentially distributed. From

(3.10) $\mathbb{P}\{\text{no collision in } \tau\}$ is then $e^{-\lambda\tau}$, where λ is the probability rate of collisions with the particle in time. Plugging this into (9.2) gives a familiar result, as we previously saw that the velocity autocorrelation of the Brownian particle (8.8) contains a similar exponential decay. Furthermore, having seen the distribution of speeds in the Brownian model to be Rayleigh distributed, the prefactor in (8.8) can be identified as the second moment of the speed (8.13), so that the velocity correlation of the Brownian particle can be rewritten as

$$\langle v(t)v(t+\tau)\rangle = \left\langle s^2(t)\right\rangle e^{-\gamma\tau} \tag{9.3}$$

This is exactly the autocorrelation function of the random walk velocity upon identification of $\lambda \equiv \gamma$.

It is worth taking a moment to summarize what this tells us. We have seen that despite its "jaggedness" on all scales (i.e. its fractal character), the Brownian path of the LE still contains memory in the velocity vector (Section 8.3.1), in other words the direction of motion of the particle is briefly correlated with its past. Here we have shown that the exponential character of this correlation function is *exactly the same* as the average autocorrelation of a particle which moves ballistically but undergoes sudden changes in velocity at stochastic intervals, where the friction coefficient γ in the Brownian model is equivalent to the rate of collisions in the random walk. From this we can conclude that while a single Brownian trajectory differs from that of a random walk on the scale of the moving particle, an ensemble of random walks presents on average the same speed and direction changes as the Brownian motion if the timescale on which the measurements are performed is larger than $1/\gamma$.

9.3. Coupling the Brownian Langevin equation to the particle density

9.3.1. Fixed density populations

If γ is the Poisson rate of collisions in time undergone by a single particle in the gas, then the average time between collisions is given by $\langle\tau\rangle = 1/\gamma$ (the expected value (3.12) of the exponential distribution). For a particle in a gas, this quantity is sometimes referred to as the *mean free time* [18]. It is closely related to another quantity known as the *mean free path* – the average distance a particle travels in between collisions, which we will denote as l – through the relation

$$l = \langle s\rangle \langle\tau\rangle \tag{9.4}$$

with $\langle s\rangle$ the particle's average speed. The fact that the mean free path is thus encoded in the Langevin equation (through the friction coefficient and the average speed) will prove very useful, as it has a simple dependence on the number density n of particles [18]:

$$l = \frac{1}{\sqrt{2}\sigma n} \tag{9.5}$$

where σ is the collisional *cross-section* – the area perpendicular to the direction of motion covered by two colliding particles – which in two dimensions is simply the sum of their diameters. This means the Brownian LE for our gas particles can be made to depend explicitly on the particle density. With $\gamma = \langle s\rangle/l$ from (9.4), the average speed determined by the Rayleigh distribution (8.12), and using the fluctuation-dissipation relation $E = D/\gamma$ (8.10) we obtain:

$$\begin{cases} \gamma = \dfrac{\sqrt{E}}{l}\sqrt{\dfrac{\pi}{2}} = \sqrt{\pi E}\,\sigma n \\ D = \dfrac{E^{3/2}}{l}\sqrt{\dfrac{\pi}{2}} = \sqrt{\pi E^3}\,\sigma n \end{cases} \tag{9.6}$$

which completely determines the LE (8.4).

9.3.2. Growing populations

At this point it should be clear how we intend to model a varying population size. If we swap out the gas particles for cells, the density which arises in the LE derived in the previous section can be made to vary according to some growth rate, which would model the dynamic crowding of a growing population. This of course implicitly fixes the motion type of the cells: ballistic movement in between circular disk elastic collisions. In Section 9.3.3 we will discuss how this can be extended to allow for other forms of movement in between collisions.

Equilibrium in slowly varying populations

Before moving from the fixed to a varying density $n \to n(t)$ let us briefly recall an important presumption we have made in deriving this density dependence. We have at multiple points – when taking the speeds as Rayleigh distributed and when invoking the fluctuation-dissipation relation – made use of the assumption that the population of interacting particles is in equilibrium. This state of the system is of course questionable if particles are being added: wherever a new particle appears the density becomes (at that point in time) locally higher than its surroundings, meaning that the new total density $(N+1)/V$ does not describe spatial homogeneity until this local "bump" has diffused. On the other hand, a return to equilibrium might occur very fast depending on the particle energies, making such a brief variation differ little from the local density fluctuations one would observe in any case for fixed a particle number. Thus we can presume a separation of timescales, meaning that we may add (or remove) particles to the system and still maintain equilibrium as long as this is done *slow enough*. Under these circumstances the statistical quantities described in Section 8.3.1 should hold, even for slowly varying $\gamma(t)$ an $D(t)$.

Logistic growth

The time dependent particle density $n(t)$ is determined by the growth of the population. While such growth is inherently an exponentially varying process, biological populations

are often found to increase with a sigmoid character – i.e. an "S" shape – over long periods of time, principally because unrestricted exponential growth cannot be sustained by their environment. In our case the limited space is clearly such a restricting factor, further supported by evidence that cell proliferation in mammalian tissues can be slowed in response spatial constraints [55] . Many functional forms are used for modeling sigmoid growth curves in different biological applications, and are typically specific variations of the logistic model [59], which has an exponential increase whereby the growth rate decreases over time. For our purpose it is sufficient to consider the simplest form proposed by Verhulst almost two centuries ago [62], characterized by a growth rate λ and a carrying capacity K which represents the limit to which the population size converges. For a population of size $N(t)$ it is given by

$$N(t) = \frac{KN_0e^{\lambda t}}{K + N_0(e^{\lambda t} - 1)} \tag{9.7}$$

so that we have for the density $n(t) = N(t)/V$.

9.3.3. Alternative types of motion

As previously mentioned, the Langevin equation (8.4) with γ and D determined by (9.6) – while density dependent – only describes a particular form of motion, i.e. that of a random walk where direction changes are caused by elastic collisions. Let us now briefly discuss how this may be extended to describe other types of movement as well as possibly different forms of interactions.

Generalizing the type of trajectory in between collisions fairly simple. It suffices to add additional terms to account for a different type of motion. For example, to couple the collisional LE of (9.6) to the LE of an active Brownian particle, the equation of motion would have some form of

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = [\gamma_c(n) + \gamma_a(x, v, t)]v(t) + \sqrt{2D_c(n)}\xi_c(t) + \sqrt{2D_a(x, v, t)}\xi_a(t) \tag{9.8}$$

where the collisional parameters $\gamma_c(n)$ and $D_c(n)$ are given by (9.6) and the active parameters $\gamma_a(x, v, t)$ and $D_a(x, v, t)$ depend on the chosen model of motion [47].

Altering the type of interactions is more complicated, as γ_c and D_c clearly encode elastic collisions even though the interaction potential never explicitly appeared in our derivation in Section 9.3.1. Still, since both depend on the average particle energy E, allowing it to vary can introduce inelastic effects, while lifting the fluctuation-dissipation requirement can involve "active" interactions.

9.4. Comparison of the Langevin equation with direct particle simulations

Having developed a formalism for density dependent motion, let us now compare the Langevin equation derived here with the exact simulations of individual particles (the specifics of which are described in Appendix B.1) shown in Figure 9.1. As discussed in Section 9.3.1, while the single particles in the simulation move in straight lines in between collisions, we expect the statistical properties of the particle ensemble to be well-described by the Langevin equation (8.4) if the timescale on which we measure is much larger than the particle mean free time $1/\gamma$. To assess the accuracy of this model we first consider the system at fixed density. The quantities described in Section 8.3.1 can be obtained from the particle simulations as direction measurements of the individual particle, and can be compared to both their theoretical predictions as well as numerically simulated trajectories of the Langevin equation [24]. (see Addendum B.2 for specifics).

Note that given the comparative nature of this section, we will refrain from specifying dimensions explicitly when reporting chosen parameter values, as they are usually implied by the quantities themselves, and their implicit existence removes the need for defining units of measurement. For example, if a reported particle energy is denoted as $E = 0.5$, it is tacitly implied the dimensions are of the form $\left[mass \times distance^2/time^2\right]$ whereby the distinct units for *mass*, *distance*, and *time* are universal to any relevant comparisons.

9.4.1. Fixed density results

For a population of fixed density, we saw (in Section 8.3.1) that the speed of the Brownian particles is expected to be Rayleigh distributed, according to

$$\mathbb{P}\{s\} = \frac{s}{E} e^{\frac{-s^2}{2E}} \tag{9.9}$$

where E is the average particle energy from the fluctuation-dissipation relation (8.10). A comparison of the equilibrium Langevin and ballistic particle simulations with (9.9) is shown in Figure 9.2. Note that this distribution is independent of the particle density and depends only on the average energy. Conversely, the velocity autocorrelation function

$$\langle v(t)v(t+\tau)\rangle = 2Ee^{-\gamma\tau} \tag{9.10}$$

and the mean squared displacement

$$\left\langle x^2(t)\right\rangle = 4\frac{E}{\gamma}\left[t - \frac{1}{\gamma}\left(1 - e^{-\gamma t}\right)\right] \approx \frac{4E}{\gamma}t \tag{9.11}$$

(obtained from plugging in the fluctuation-dissipation relation $E = D/\gamma$ in (8.8) and (8.14)) of the particles do depend on the population density through (9.6), as shown in Figures 9.3 and 9.4 respectively, shows both models to be in good agreement for different population densities. We note how the ballistic regime – when $e^{-\gamma t}$ is not yet ≈ 1 – is clearly visible in the early time of the mean squared displacement, and is shortened for higher densities where γ is larger.

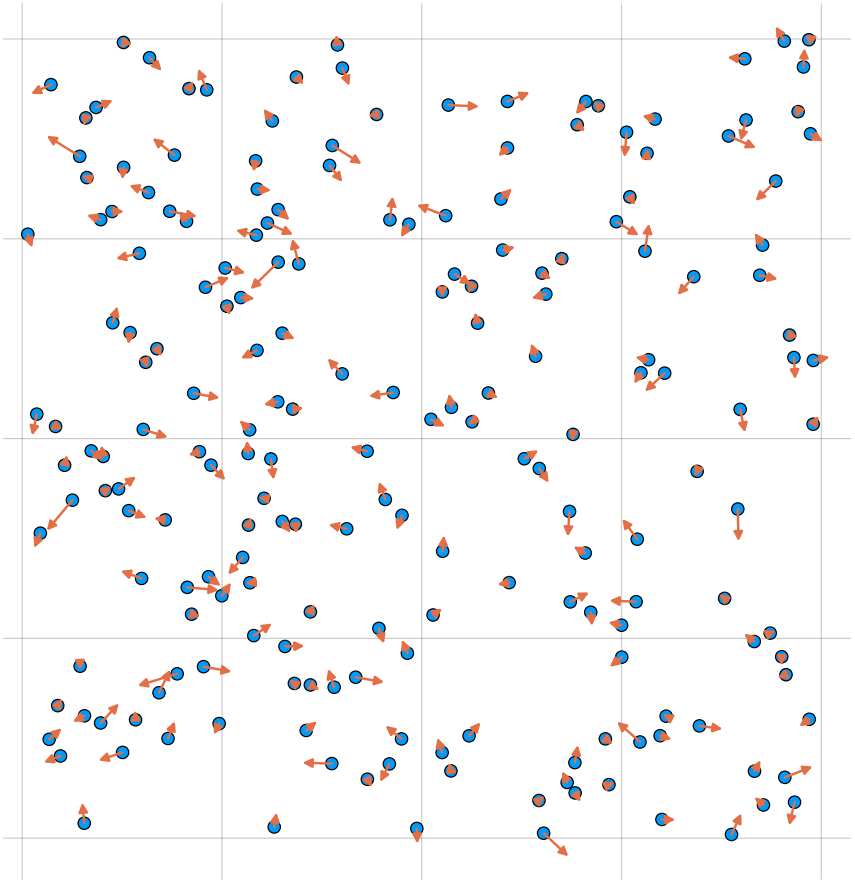


Figure 9.1.: Visualization of the particle simulation. *An example of a possible state the system can be in at any given time point. The individual particles are represented as blue-filled circles at their respective positions in the plane, with their velocities depicted by the orange arrows.*

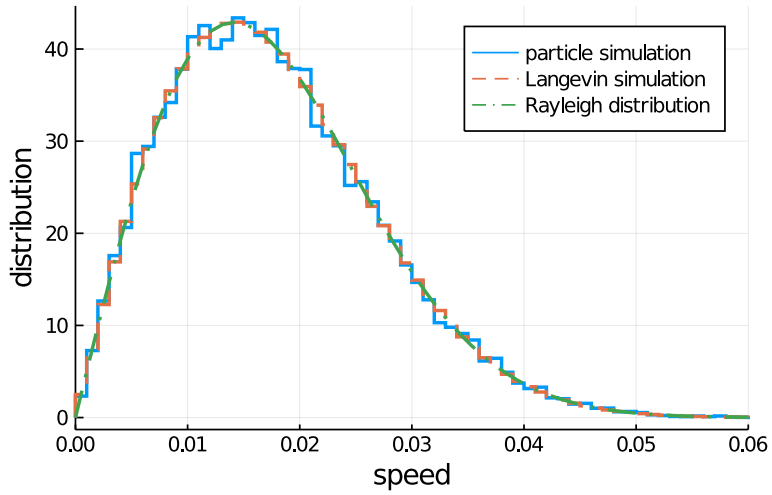
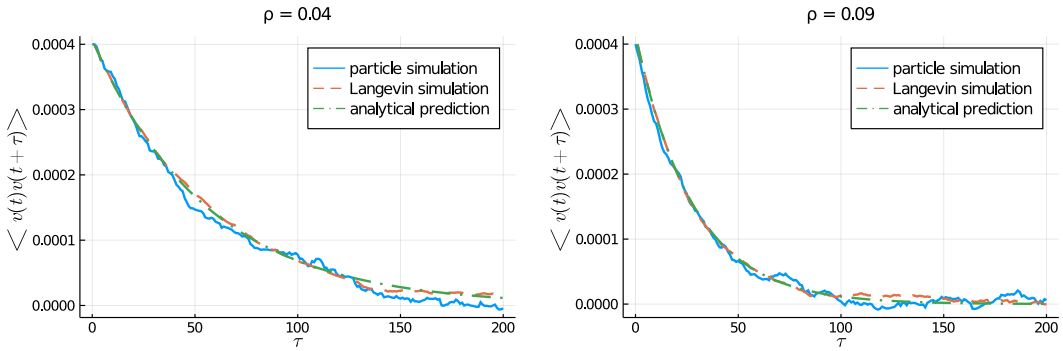


Figure 9.2.: Distribution of single particle speeds $s(t) = \sqrt{v^2(t)}$***.*** *The measured distributions are shown for the particle simulation and the Langevin equation, alongside the predicted Rayleigh distribution. The particle simulation consisted of 1000 particles of radius* $r = 0.08$ *in a square volume with dimensions* 15×15*, resulting in a particle number-density of* $n = 4.44$ *and a volume-density of* $\rho = 0.09$*. The average particle energy was taken at* $E = 0.002$ *by initiating the simulation with fixed particle speeds* $s(t = 0) = 0.02$*, and subsequently allowing the system to relax to its equilibrium distribution. The distribution for the Langevin particles was obtained by numerical simulation of 5000 trajectories of the Langevin equation (8.4) with parameters (9.6) with* E*,* n*, and* r *as denoted above. Both distributions were obtained from all particle speeds at multiple time points.*



(a) ***(b)***

Figure 9.3.: Velocity autocorrelation of the single particles in a fixed density population. *The particle simulation was run for a* 15×15 *square volume with 500 (a) and 1000 (b) particles of radius* $r = 0.08$ *with initial speed* $s(t = 0) = 0.02$ *and thus energy* $E = 2 \times 10^{-4}$*. The Langevin simulations were performed for 5000 trajectories with the same parameter values. The analytical prediction is the function (8.8) obtained from analytically solving the LE.*

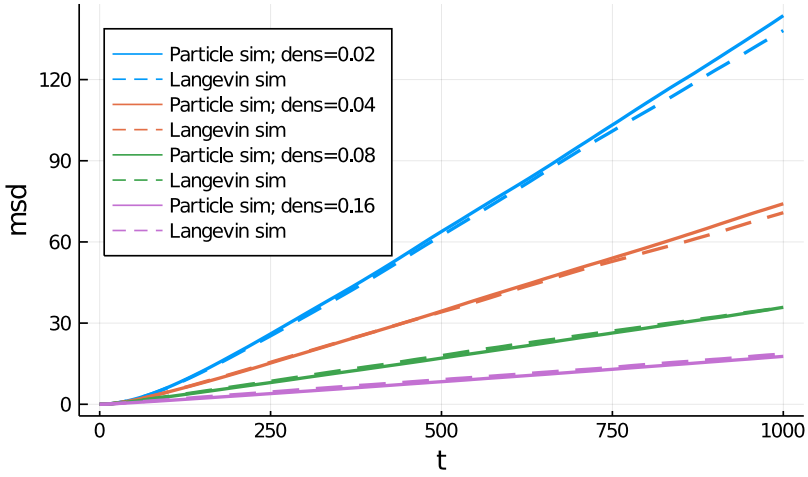


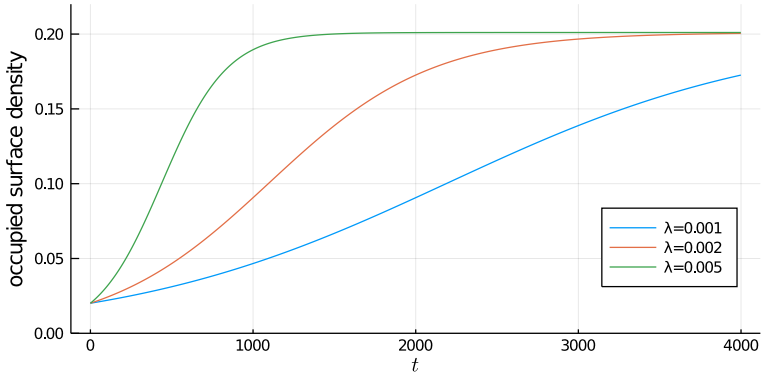
Figure 9.4.: Mean squared displacement for a fixed density population. *The measured quantities are shown for the particle simulation and the simulated Langevin equation (the analytical prediction (8.14) obtained from the LE matched the numerical LE simulations almost exactly, and is therefore left out for clarity), each performed for four different population densities. All simulations were performed for particles with radius* $r = 0.08$ *and initial speed* 0.03 *(average energy* 4.5×10^{-4}*).*

9.4.2. Growing population results

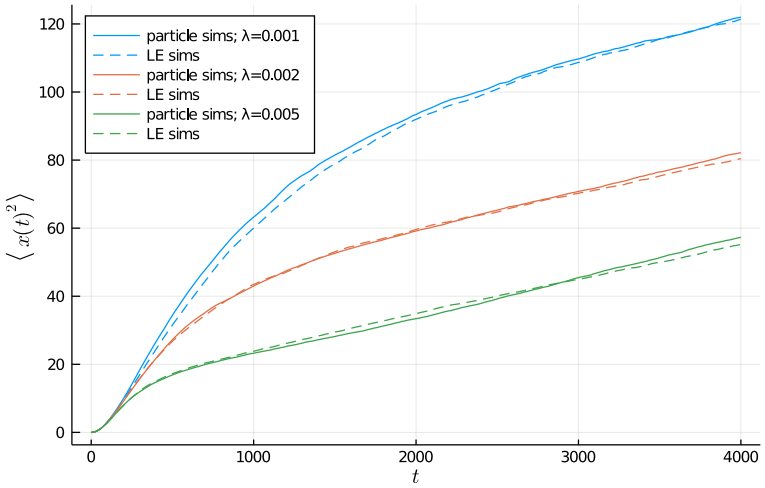
Let us now turn to the case of a varying population. Taking the growth function (9.7) for the particle density n in (9.5) we obtain the density dependent Langevin equation. As before, this can be compared with the particle simulation, where growth is implemented as described in B.1.5, the result of which is shown in Figure 9.5. We observe sublinear diffusion as the population increases, followed by a return to a linear slope as the carrying capacity K is reached, whereby the Langevin model generally agrees well with the particle simulation.

9.5. Perspective: localizing the LE for interacting particles

Up until this point we have only considered a spatially homogeneous system – made possible by the assumption of a slowly varying population size – where density variation is therefore assumed to occur simultaneously at all points in space. It would be useful to extend this formalism to a spatially localized model, as this could be effectively used in modeling experimental conditions with spatial inhomogeneity. One possible route to achieve this would be by following the method of [9] and [56], where the single particle Langevin equation is used to obtain a stochastic time evolution of the spatial density $\rho(x,t)$ by its constructing as a sum of the *single particle densities* $\rho(x) = \sum_i \rho_i(x) = \sum_i \delta(x - x_i)$. This avenue of investigation provides an interesting perspective for future work.



(a)



(b)

Figure 9.5.: Mean squared displacement in a growing population. *For the logistic growth function (9.7) with maximum occupied volume density $K/V = 0.2$ different values of the growth parameter λ were simulated: $\lambda = 0.001$, $\lambda = 0.002$, and $\lambda = 0.005$. (a) Growth curve for different values of the growth rate λ. (b) Mean squared displacement measured from particle simulations (full lines) and Langevin simulated solutions (dashed lines)*

9.6. Discussion

In this chapter we have seen how the friction and force intensity parameters γ and D of the Brownian Langevin equation for a ballistic particle undergoing a random walk due to collisions can be made to depend explicitly on the density n of the system, by identification of the *mean free path* arising in the velocity autocorrelation. Through numerical simulations of the 2D Langevin equation we tested this model for a growing population confined to a plane by coupling the density to an appropriate growth curve – such as the logistic function (9.7) – and comparing the predictions with a 2D simulation of individually colliding particles in which the particle number increases according to the projected growth. The Brownian model performed well as an estimate for the averaged effect of increasing crowding. We propose this Brownian Langevin equation can be used to model dynamical crowding effects on cell motion in situations where the population size variation plays a role, whereby possible extensions of the LE can be made to include active types of motion in between interactions. An interesting perspective is the possibility of constructing a similar Langevin equation for the density field $\rho(x,t)$, in which spatially localized density variations can be modeled, as such a tool would be useful in analyzing more complex environments in which there is no spatial homogeneity.

The model introduced here can be useful in interpreting experiments in the field of cancer such as that by Lin et al. [35] discussed in Section 8.1, where cell proliferation plays a key role in the systems' dynamics. Furthermore, because this approach remains agnostic as to both the characteristics of the particles' motion in between interactions (as discussed in Section 9.3.3) as well as the mechanism by which the particle density varies (Section 9.5), it is potentially applicable in any situation where there is an interest in studying properties of statistical motility in a population presenting density dynamics. Some examples include spatial ecological systems [4], for example the exploratory habits of animals of interest could exhibit effects of dynamic crowding as a result of population cycles; epidemiological systems [34], since the quantity of information-spreading interactions depends on the density of information-carrying individuals; and even human societal contexts, for example the characterization of mobility in urban locations which

present crowding dynamics.

Bibliography

[1] Bruce Alberts. *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, 2018.

[2] Tadashi Ando and Jeffrey Skolnick. "Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion". In: *Proceedings of the National Academy of Sciences* 107.43 (2010), pp. 18457–18462.

[3] Clemens Bechinger et al. "Active particles in complex and crowded environments". In: *Reviews of Modern Physics* 88.4 (2016), p. 045006.

[4] Robert Stephen Cantrell and Chris Cosner. *Spatial ecology via reaction-diffusion equations*. John Wiley & Sons, 2004.

[5] Michael E Cates. "Diffusive transport without detailed balance in motile bacteria: does microbiology need statistical physics?" In: *Reports on Progress in Physics* 75.4 (2012), p. 042601.

[6] Michael E Cates and Julien Tailleur. "When are active Brownian particles and run-and-tumble particles equivalent? Consequences for motility-induced phase separation". In: *EPL (Europhysics Letters)* 101.2 (2013), p. 20010.

[7] Michael E Cates et al. "Arrested phase separation in reproducing bacteria creates a generic route to pattern formation". In: *Proceedings of the National Academy of Sciences* 107.26 (2010), pp. 11715–11720.

[8] Nicholas C Darnton et al. "On torque and tumbling in swimming Escherichia coli". In: *Journal of bacteriology* 189.5 (2007), pp. 1756–1764.

[9] David S Dean. "Langevin equation for the density of a system of interacting Langevin processes". In: *Journal of Physics A: Mathematical and General* 29.24 (1996), p. L613.

[10] Carmine Di Rienzo et al. "Probing short-range protein Brownian motion in the cytoplasm of living cells". In: *Nature communications* 5.1 (2014), pp. 1–8.

[11] James A Dix and AS Verkman. "Crowding effects on diffusion in solutions and cells". In: *Annu. Rev. Biophys.* 37 (2008), pp. 247–263.

[12] Anushka Dongre and Robert A. Weinberg. "New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer". In: *Nature Reviews Molecular Cell Biology* 20.2 (Feb. 2019). Number: 2 Publisher: Nature Publishing Group, pp. 69–84. ISSN: 1471-0080. DOI: 10.1038/s41580-018-0080-4. URL: https://www.nature.com/articles/s41580-018-0080-4/briefing/signup/ (visited on 09/29/2020).

[13] Anushka Dongre et al. "Epithelial-to-mesenchymal transition contributes to immunosuppression in breast carcinomas". In: *Cancer research* 77.15 (2017), pp. 3982–3989.

[14] Monroe D Donsker. "Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems". In: *The Annals of mathematical statistics* (1952), pp. 277–281.

[15] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[16] Albert Einstein. "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen". In: *Annalen der physik* 4 (1905).

[17] Dominique Ernst et al. "Fractional Brownian motion in crowded fluids". In: *Soft Matter* 8.18 (2012), pp. 4886–4889.

[18] Richard P Feynman, Robert B Leighton, and Matthew Sands. "The feynman lectures on physics; vol. i". In: *American Journal of Physics* 33.9 (1965), pp. 750–752.

[19] Peter Friedl and Darren Gilmour. "Collective cell migration in morphogenesis, regeneration and cancer". In: *Nature reviews Molecular cell biology* 10.7 (2009), pp. 445–457.

[20] Gaorav P Gupta and Joan Massagué. "Cancer metastasis: building a framework". In: *Cell* 127.4 (2006), pp. 679–695.

[21] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[22] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer". In: *cell* 100.1 (2000), pp. 57–70.

[23] Peter Hänggi and Fabio Marchesoni. *Introduction: 100 years of Brownian motion.* 2005.

[24] Desmond J Higham. "An algorithmic introduction to numerical simulation of stochastic differential equations". In: *SIAM review* 43.3 (2001), pp. 525–546.

[25] Kiyosi Itô. *On stochastic differential equations.* 4. American Mathematical Soc., 1951.

[26] Kiyosi Itô. "On stochastic processes (I)". In: *Japanese journal of mathematics: transactions and abstracts.* Vol. 18. The Mathematical Society of Japan. 1941, pp. 261–301.

[27] Johanna A. Joyce and Jeffrey W. Pollard. "Microenvironmental regulation of metastasis". In: *Nature Reviews Cancer* 9.4 (Apr. 2009), pp. 239–252. ISSN: 1474-1768. DOI: 10.1038/nrc2618. URL: https://www.nature.com/articles/nrc2618 (visited on 02/12/2018).

[28] Mark Kac. "A stochastic model related to the telegrapher's equation". In: *The Rocky Mountain Journal of Mathematics* 4.3 (1974), pp. 497–509.

[29] Raghu Kalluri, Robert A Weinberg, et al. "The basics of epithelial-mesenchymal transition". In: *The Journal of clinical investigation* 119.6 (2009), pp. 1420–1428.

[30] Sato Ken-Iti. *Lévy processes and infinitely divisible distributions*. Cambridge university press, 1999.

[31] Florian Klemm and Johanna A. Joyce. "Microenvironmental regulation of therapeutic response in cancer". In: *Trends in Cell Biology* 25.4 (Apr. 1, 2015), pp. 198–213. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2014.11.006. URL: http://www.sciencedirect.com/science/article/pii/S0962892414001998 (visited on 03/12/2018).

[32] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*. Vol. 31. Springer Science & Business Media, 2012.

[33] Paul Langevin. "Sur la théorie du mouvement brownien". In: *Compt. Rendus* 146 (1908), pp. 530–533.

[34] Andrew B Lawson. *Statistical methods in spatial epidemiology*. John Wiley & Sons, 2013.

[35] Ke-Chih Lin et al. "Epithelial and mesenchymal prostate cancer cell population dynamics on a complex drug landscape". In: *Convergent Science Physical Oncology* 3.4 (2017), p. 045001. ISSN: 2057-1739. DOI: 10.1088/2057-1739/aa83bf. URL: http://stacks.iop.org/2057-1739/3/i=4/a=045001 (visited on 02/12/2018).

[36] Marianne Lintz, Adam Muñoz, and Cynthia A Reinhart-King. "The mechanics of single cell and collective migration of tumor cells". In: *Journal of biomechanical engineering* 139.2 (2017).

[37] TT Marquez-Lago, Andre Leier, and Kevin Burrage. "Anomalous diffusion and multifractional Brownian motion: simulating molecular crowding and physical obstacles in systems biology". In: *IET systems biology* 6.4 (2012), pp. 134–142.

[38] James Clerk Maxwell. "V. Illustrations of the dynamical theory of gases.—Part I. On the motions and collisions of perfectly elastic spheres". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19.124 (1860), pp. 19–32.

[39] Hazime Mori. "Transport, collective motion, and Brownian motion". In: *Progress of theoretical physics* 33.3 (1965), pp. 423–455.

[40] M Angela Nieto et al. "EMT: 2016". In: *Cell* 166.1 (2016), pp. 21–45.

[41] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[42] Klaus Pantel and Ruud H Brakenhoff. "Dissecting the metastatic cascade". In: *Nature reviews cancer* 4.6 (2004), pp. 448–456.

[43] Colin D Paul, Panagiotis Mistriotis, and Konstantinos Konstantopoulos. "Cancer cell motility: lessons from migration in confined spaces". In: *Nature Reviews Cancer* 17.2 (2017), pp. 131–140.

[44] Sriram Ramaswamy. "The mechanics and statistics of active matter". In: (2010).

[45] Werner Risau. "Mechanisms of angiogenesis". In: *Nature* 386.6626 (1997), pp. 671–674.

[46] Hannes Risken. "Fokker-planck equation". In: *The Fokker-Planck Equation*. Springer, 1996, pp. 63–95.

[47] Pawel Romanczuk et al. "Active brownian particles". In: *The European Physical Journal Special Topics* 202.1 (2012), pp. 1–162.

[48] Jeroen Rouwkema, Nicolas C Rivron, and Clemens A van Blitterswijk. "Vascularization in tissue engineering". In: *Trends in biotechnology* 26.8 (2008), pp. 434–441.

[49] JS Rowlinson*. "The Maxwell–Boltzmann distribution". In: *Molecular Physics* 103.21-23 (2005), pp. 2821–2828.

[50] Ton N Schumacher and Robert D Schreiber. "Neoantigens in cancer immunotherapy". In: *Science* 348.6230 (2015), pp. 69–74.

[51] Frank Schweitzer, Werner Ebeling, and Benno Tilch. "Complex motion of Brownian particles with energy depots". In: *Physical Review Letters* 80.23 (1998), p. 5044.

[52] Tsukasa Shibue and Robert A Weinberg. "EMT, CSCs, and drug resistance: the mechanistic link and clinical implications". In: *Nature reviews Clinical oncology* 14.10 (2017), p. 611.

[53] Patricia S Steeg. "Targeting metastasis". In: *Nature reviews cancer* 16.4 (2016), pp. 201–218.

[54] RL Stratonovich. "A new representation for stochastic integrals and equations". In: *SIAM Journal on Control* 4.2 (1966), pp. 362–371.

[55] Sebastian J Streichan et al. "Spatial constraints control cell proliferation in tissues". In: *Proceedings of the National Academy of Sciences* 111.15 (2014), pp. 5586–5591.

[56] J Tailleur and ME Cates. "Statistical mechanics of interacting run-and-tumble bacteria". In: *Physical review letters* 100.21 (2008), p. 218103.

[57] Stéphane Terry et al. "New insights into the role of EMT in tumor immune escape". In: *Molecular oncology* 11.7 (2017), pp. 824–846.

[58] Jean Paul Thiery et al. "Epithelial-mesenchymal transitions in development and disease". In: *cell* 139.5 (2009), pp. 871–890.

[59] Anastasios Tsoularis and James Wallace. "Analysis of logistic growth models". In: *Mathematical biosciences* 179.1 (2002), pp. 21–55.

[60] George E Uhlenbeck and Leonard S Ornstein. "On the theory of the Brownian motion". In: *Physical review* 36.5 (1930), p. 823.

[61] Nicolaas G Van Kampen. "Itô versus stratonovich". In: *Journal of Statistical Physics* 24.1 (1981), pp. 175–187.

[62] Pierre-François Verhulst. "Notice sur la loi que la population suit dans son accroissement". In: *Corresp. Math. Phys.* 10 (1838), pp. 113–126.

[63] Bert Vogelstein and Kenneth W Kinzler. "The multistep nature of cancer". In: *Trends in genetics* 9.4 (1993), pp. 138–141.

[64] Marian Von Smoluchowski. "Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen". In: *Annalen der physik* 326.14 (1906), pp. 756–780.

[65] Cornelis J Weijer. "Collective cell migration in development". In: *Journal of cell science* 122.18 (2009), pp. 3215–3223.

[66] Robert A Weinberg. *The biology of cancer*. Garland science, 2013.

[67] Norbert Wiener. "The mean of a functional of arbitrary elements". In: *Annals of Mathematics* (1920), pp. 66–72.

[68] Pei-Hsun Wu, Daniele M. Gilkes, and Denis Wirtz. "The Biophysics of 3D Cell Migration". In: *Annual Review of Biophysics* 47.1 (2018), pp. 549–567. DOI: 10.1146/annurev-biophys-070816-033854. URL: https://doi.org/10.1146/annurev-biophys-070816-033854 (visited on 06/22/2018).

Conclusions

10. Conclusions

In this thesis we have investigated a number of cell-based biological systems through the application of stochastic mathematical models. In each of these applications the goal has been to either achieve novel insights into the system's behavior, test existing assumptions related to its functioning, or quantify qualitatively understood behavior in a manner which may contribute to accurate predictions.

The results described in this work were roughly divided into two parts: the first part concerned three separate investigations of the human *hematopoietic system*, the process by which precursor cells of the blood are matured in the bone marrow. The second – somewhat shorter – part described a study of the effect of proliferation in populations of motile particles, with the intent of its possible application in growing cell populations which may be relevant to the occurrence of metastasis in cancer. Here we will recapitulate the conclusions and perspectives of projects.

10.1. Population dynamics of hematopoiesis

The hematopoietic system has for many decades been an important subject of study, both due to its importance in the normal functioning of the body, as well as its relevance in the study of various blood-related diseases. Its hierarchical architecture has been meticulously mapped over time, with current understanding describing a complex picture of differentiation lineages which facilitate transitions from the stem cell compartment to the mature cell types found in the blood. The quantitative underpinnings of this picture are however to an extent still lacking, and numerous question marks remain with respect

to the underlying mechanisms that drive the system.

Over the course of three overarching research questions we have studied the behavior of hematopoietic stem cells and their mutation accumulation through an adapted Moran model approach, as well as the plasticity of the hematopoietic system as a whole through the development of a heuristic feedback driven model of its encapsulated cell dynamics.

The first question (Chapter 5) related to the blood disorder *paroxysmal nocturnal hemoglobinuria* (PNH) caused by the expansion of a mutant *PIGA* clone in the hematopoietic stem cell (HSC) pool. While one hypothesis attributes this clonal growth to some infrequently occurring selective advantage, we have shown the possibility of a simpler alternative: expansion due to neutral drift. While this may be a highly unlikely event, we investigated the exact probability of its occurrence by constructing a Markov chain for the probabilistic arrival and evolution of such a selectively neutral mutant in the HSC pool. Requiring only a handful of parameters from the literature, among which a small size $N \approx 400$ of the stem cell population, a slow rate of symmetric self-renewal $\lambda = 1/\text{year}$ and a mutation rate of the gene $\mu = 5 \times 10^{-7}$ per cell division, we have shown that the expected number of individuals in which this neutral expansion would occur is not negligible, and in fact, after taking into account the distribution of ages using data from the 2010 United States Census – fits with the incidence of the disease in a population, where we predict an expected prevalence of 1.76 cases per 10^5 citizens. Under the neutral hypothesis, the model furthermore rightly predicts the possibility of a spontaneous loss of the disease, a phenomenon which is observed in many patients and is difficult to reconcile with the selection-based hypothesis. On the other hand, we found that the same model of HSC dynamics alone fails to capture the rate at which clones are seen to expand in patients, even though the observed distribution of clonal expansion and reduction qualitatively fits the selection-free hypothesis. We suggest that this can be explained by the fact that the model does not take into account the cellular dynamics outside of the stem cell pool, in particular the extreme loss of blood caused by the PNH phenotype in mature cells, which would through feedback cause a compensating reaction

in the HSC pool, increasing the speed of its divisional dynamics thus the rate at which clones vary in size.

This model, while simple, still required the knowledge of a select few parameter values. Because of the difficulty of performing *in vivo* measurements and the complexity of the hematopoietic system such values are often debated in the literature. Because the strength of such predictive models may rely heavily on accurate parameterization, devising methods for obtaining these quantities can be highly beneficial for future research in the field. For this reason, the second research project (Chapter 6) dealt with the question of extracting this quantitative information from what is ultimately often noisy experimental data. Focusing our attention on the stem cell pool, we posited that valuable information on the behavioral processes of HSCs – such as the per-daughter-cell mutation rate, the symmetric and asymmetric division rates, and the size of the stem cell pool – is encoded in the stochastic occurrence of somatic mutations, which thanks to recent advances in genome sequencing techniques and experimental design has become increasingly observable. Thus, building on the model of a neutrally expanding mutant, we next examined the clonal landscape of all possible occurring mutations in HSCs. Given that the space of all outcomes of such a stochastic evolution is too complex to analyze directly, we considered two reductions of this state space, both containing different information on the underlying quantitative processes. The first was the *single cell mutational burden* (the total number of mutations per cell), which encoded both the mutation rate and the total division rate. The second was the variant allele frequency (VAF, the distribution of clone sizes within the population), for which we derived an expression to numerically evolve its expected value, which relies on all underlying parameter values at once. We were, however, unable to obtain a prediction for the evolution of the variance of the VAF. These results were then applied to a recent dataset containing high resolution mutational information on 89 human HSCs derived from a single patient, from which we extracted a Poisson mutation rate of 4.3 per daughter cell per division, as well as a total division rate of 4.2 per year. Fitting the VAF measured from the data

proved on the other hand difficult, as it was shown that information on the size of the HSC pool is lost during the process of sampling. These results highlighted the usefulness of this mutational information for robustly quantifying the processes driving stem cell dynamics, likely even in other tissues. It is clear that further research into quantifying this clonality – such as, for example, obtaining a prediction for the time evolution of the variance on the VAF – can greatly benefit the interpretation of the growing wealth of experimental data.

Finally, our treatment of the hematopoietic system concluded with an investigation of its vivo cell dynamics (Chapter 7), with a possible generalization to any differentiation based systems in the body. While our modern map of hematopoiesis – describing the various existing lineage pathways – is considerably intricate, there is still little known of the dynamic character of the system as a whole. It is known that various disorders can cause great variations in cell numbers in the blood, however the hematopoietic system presents an uncanny ability to react to these and resume some form of normal functioning. To better understand the cell dynamics underlying this plasticity we formulated a model of hematopoiesis whereby successive stages of differentiation are modeled as distinct compartments which interact through feedback. This system acts as a flow of cells from the HSC stage to the mature compartment, whereby the density of cells in each compartment grows exponentially due to proliferation. Applying a perturbation to this system, such as for example a loss of cells in a compartment, the preceding compartments must alter their behavior in order to mitigate the duration of the disruption to the system. We found that a precise balance of altered differentiation – providing the downstream flow – and self-renewal – providing the exponential growth – is required for stability, as disproportionate responses can lead to oscillations or sudden losses in the compartment numbers. The former behavior is in fact observed in certain hematopoietic diseases such as *cyclic neutropenia* in the form of cycling cell numbers in the blood. Applying long lasting perturbations, as are expected to occur in various chronic diseases such as PNH or *chronic myeloid leukemia*, we found the the system evolves to new steady

states, with moderately to extremely altered cell numbers in the blood, which is in line with observations in patients.

While this model should not be considered a predictive tool for a specific hematopoietic lineage pathway, it serves as an important illustration for the expected types of dynamics one would observe under stresses of the hematopoietic system. Through its investigation we have shown the importance of the effects of feedback when interpreting observations of disease progression in patients. Indeed, this model illustrates clearly how the final inclusion of feedback in the previously discussed model of PNH (Chapter 5) constitutes a far more accurate depiction of the mutational dynamics. A similar observation can be made for the mutational diversity discussed in Chapter 6: an individual who has at some point in their life suffered from an extreme perturbation to the blood may have a more aged stem cell compartment due to its increased activity in the past.

The three research questions addressed here have highlighted both the importance of the quantitative approach to the complex system of hematopoiesis, as well the many open questions still remaining. The accumulating wealth of increasingly qualitative data provides ample opportunity to answer these, and simultaneously necessitates the further investigation into the mathematical properties of the system's underlying behaviors.

10.2. Statistical mechanics of proliferating cells

The possibility of self-driven motion in living matter can lead to complex dynamics. However, while the statistical characteristics of such *active motion* are the topic of great scrutiny, little attention has been payed to the influence of proliferation in active populations. Such system are abundant though, with a prime example found in motile cell populations which can occur in regenerating tissues, cell culturing experiments, and cancer. In Part II of this thesis (Chapters 8 and 9) we have investigated the influence of such proliferation on the stochastic motion of interacting particles by returning to the simplest model from first principles: Brownian motion. While this model is traditionally applied to describe a particle's stochastic movement caused by an external

medium such as a fluid, we instead used it as a course-grained model for the ballistic motion of particles in between collisions in a crowded environment. In this interpretation, its formulation as a Langevin equation (LE) allowed for the explicit introduction of a dependence on the density through the friction and force parameters, by identifying their relation to the particle's mean free path. In this manner a dynamically varying density could be introduced, which can be effectively used to model crowding effects in proliferating populations. For a growing population in a confined environment this led to the observation of sublinear diffusion, caused by the dynamically reducing availability of space. We furthermore tested the validity of this approach by comparing numerically simulated trajectories of the Langevin equation to a direct simulation of elastically colliding ballistic particles, for which we found good agreement. A potentially interesting extension to this formalism would be to construct a spatially localized form of the equation, allowing for the characterization of non-heterogeneous systems whereby stochastic fluctuations may play an increased role.

While the Langevin equation investigated here is not active in the sense of the single particle motion – as we made no inclusion of any auto-locomotive properties – it can easily be extended to cover such systems by including additional *active* terms in the LE. For this reason it can be a useful tool for introducing a proliferative effect in models of active systems such as, for example, motile cells in culture. Furthermore, thanks to the generality of the approach the model can be applicable in other fields as well, wherever there are questions pertaining to stochastic movement in space and a dynamically varying mobility caused by crowding.

Appendix

A. Population dynamics of hematopoiesis

A.1. Combining Poisson processes

Consider two independent Poisson processes with respective events V and S, each with (independent) exponentially distributed waiting times with rates v and s. We have

$$\begin{cases} \mathbb{P}\{V_t\} = 1 - e^{-vt} \\ \mathbb{P}\{S_t\} = 1 - e^{-st} \end{cases} \qquad \begin{cases} f_{t_V}(\tau) = v\, e^{-v\tau} \\ f_{t_S}(\tau) = s\, e^{-s\tau} \end{cases} \tag{A.1}$$

with $\mathbb{P}\{X_t\}$ the probability of at least one event occurring in a time interval t and $f_{t_X}(\tau)$ the density distribution of the waiting times. We now introduce the new event $\tilde{V}_t$ as the occurrence of at least one V, occurring before any S in t; and the complementary event $\tilde{S}_t$ with the converse definition. Note that $\tilde{V}_t$ and $\tilde{S}_t$ are mutually exclusive, and cover all possible outcomes except for those where no S or V occur. We may write them equivalently as the following sets:

$$\begin{aligned} \tilde{V}_t &= \{V_t \cap S_t^c, V_t \cap S_t \cap (t_v < t_s)\} \\ \tilde{S}_t &= \{S_t \cap T_t^c, S_t \cap V_t \cap (t_s < t_v)\} \end{aligned} \tag{A.2}$$

Since both events in each set are mutually exclusive we may write the probabilities of $\tilde{V}_t$ and $\tilde{S}_t$ as the sum of the probabilities of their respective elements. The first term is

$$\mathbb{P}\{V_t \cap S_t^c\} = \mathbb{P}\{V_t\}\mathbb{P}\{S_t^c\} = e^{-st}\big(1 - e^{-vt}\big) \tag{A.3}$$

since V_t and S_t are independent (with the analogous argument for $\mathbb{P}\{S_t \cap V_t^c\}$). For the second we obtain

$$\begin{aligned}\mathbb{P}\{t_v < t_s \cap S_t \cap V_t\} &= \mathbb{P}\{t_v < t_s \cap S_t\} \\ &= \int_0^t \mathbb{P}\{t_v < \tau\} f_{t_s}(\tau)\, d\tau \\ &= \int_0^t \left(1 - e^{-v\tau}\right) s\, e^{-s\tau}\, d\tau \\ &= 1 - e^{-st} + \frac{se^{-t(s+v)} - s}{s+v}\end{aligned}$$

and summing the two gives

$$\mathbb{P}\left\{\tilde{V}_t\right\} = \frac{v}{s+v}\left(1 - e^{-(s+v)t}\right) \tag{A.4}$$

From this we identify $\left(1 - e^{-(s+v)t}\right) = \mathbb{P}\{V_t \cup S_t\}$, which is the probability of any event occurring in t. Thus $\mathbb{P}\{\tilde{V}_t\}$ and $\mathbb{P}\{\tilde{S}_t\}$ can readily be interpreted as the probabilities of an event occurring in t, multiplied by a probability which determines whether that event is V or S. The above expression can furthermore be expanded for an infinitesimal timestep dt to obtain a rate:

$$\mathbb{P}\left\{\tilde{V}_{dt}\right\} = \frac{v}{s+v}(s+v)dt + \vartheta(dt^2) \tag{A.5}$$

which shows that S and V occur at rates s and v respectively, and allows us to identify

$$\begin{aligned}\epsilon &= \frac{s}{s+v} \\ r &= s+v\end{aligned} \tag{A.6}$$

The argument can be extended for more than two processes analogously.

A.2. Simulations of the Moran model with mutant accumulation

As a check for the correctness of the results on clonality derived in Chapter 6, we make use of direct simulations of the underlying Moran model, the primary code for which was developed by Marius Möller (Queen Mary University of London, School of Mathematical

Sciences). The simulation involves a Gillespie algorithm [2] which stochastically performs the divisional events described in Section 4.2 in a population of cells, whereby distinct mutations which occur in the population are tracked.

A.2.1. The cell population

The cell population is modeled as a collection of cells, each of which may carry an unlimited amount of mutations. Thus the state of the system at any point in time can be represented by a Boolean matrix A whereby the rows represent distinct cells and the columns distinct mutations, so that a 1 represents the existence of a particular mutant (column) in a cell (row) and a 0 depicts its absence. Because the population is constant the number of rows is fixed, however the number of columns increases as new mutations are added to the system.

A.2.2. Events which alter the population

Two distinct events can occur which alter the state (A) of the population. Specifically these are (see Chapter 4.3.2) a simultaneously occurring self-renewal and symmetric differentiation, and an asymmetric differentiation. Their effect on the population state is as follows:

self-renewal + symmetric differentiation

- One cell is randomly (with uniform probability) selected for self-renewal:
  - This cell is copied (i.e. the copy contains the same set of mutations), and both the original and the copy undergo mutation.
- One cell is randomly (with uniform probability) selected for symmetric differentiation:
  - This cell is removed.

asymmetric differentiation

- One cell is randomly selected (with uniform probability) for asymmetric division:
  - This cell undergoes mutation.

A.2.3. Mutations

Whenever a division is performed in the population a cell undergoes mutation, meaning it can acquire new mutations (though it cannot lose any existing mutations it carries). Each new mutation arising in a cell is distinct from the existing mutations in the population, thus introducing a new column in A. The number of mutations added to a cell during a division event is drawn from a Poisson distribution with parameter μ the mutation rate.

A.2.4. Time evolution

During time evolution, events occur probabilistically according to their Poisson rates ρ (self-renewal + symmetric differentiation) and ϕ (symmetric differentiation). Following the Gillespie algorithm, at initiation and after each occurrence of an event, the time until the next event is drawn from the exponential waiting time of all possible events – $\mathrm{Exp}(-[\rho+\phi]t)$ (see Section 3.1.2) – and the specific event is chosen randomly according to their respective likelihoods $\rho/(\rho+\phi)$ and $\phi/(\rho+\phi)$ (see Section 3.1.2).

A.3. Obtaining the mean and variance of the compound Poisson distribution

The *compound Poisson distribution* is defined as the distribution of the random variable

$$m = \sum_{i=1}^{y} x_i \tag{A.7}$$

whereby the x_i are independent and identically distributed random variables, and y is a Poisson distributed random variable. The expected value of m can be found using the law of total expectation [5], which states that

$$\mathrm{E}(m) = \mathrm{E}\left[\mathrm{E}\left(m \mid y\right)\right] \tag{A.8}$$

With y fixed, $\mathrm{E}\left(m \mid y\right)$ is expectation of a sum of random variables, which is given by the sum of their respective expectation values, so that we obtain

$$\mathrm{E}(m) = \mathrm{E}\left[y\,\mathrm{E}(x)\right] = \mathrm{E}(y)\,\mathrm{E}(x) \tag{A.9}$$

To obtain the variance, we similarly use the law of total variance [5], which states that

$$\mathrm{Var}(m) = \mathrm{E}[\mathrm{Var}(m \mid y)] + \mathrm{Var}[\mathrm{E}(m \mid y)] \tag{A.10}$$

The first term can be found from the fact that all x_i are independently distributed – so that the variance of the sum can be written as the sum of the variances – which leads to

$$\mathrm{Var}(m) = \mathrm{E}[y\,\mathrm{Var}(x)] + \mathrm{Var}[y\,\mathrm{E}(x)] \tag{A.11}$$

$$= \mathrm{E}(y)\,\mathrm{Var}(x) + (\mathrm{E}(x))^2\,\mathrm{Var}(y) \tag{A.12}$$

Now using the fact that since y is Poisson distributed $\mathrm{E}(y) = \mathrm{Var}(y)$, we obtain

$$\mathrm{Var}(m) = \mathrm{E}(y)\left[\mathrm{Var}(x) + (\mathrm{E}(x))^2\right] \tag{A.13}$$

$$= \mathrm{E}(y)\,\mathrm{E}(x^2) \tag{A.14}$$

A.4. Compartment model of hematopoiesis: fixing parameter values

In the model of Dingli et al. [1] the dynamics of a compartment j are given by

$$\partial_t N_j = 2\epsilon r_{j-1} N_{j-1} - (2\epsilon - 1) r_j N_j \tag{A.15}$$

The first term on the right hand side of the equation is the flux of cells coming in from the nearest upstream compartment $j-1$ (where the factor 2 comes from the fact that two daughter cells are created per division), while the second term is the sum of the fluxes of cells being removed due to differentiation (at rate $\epsilon r_j N_j$) and added due to self-renewal (at rate $(1-\epsilon) r_j N_j$). Given the number of HSCs N_0, the number of (non-HSC) compartments M, and the daily bone marrow output β_M; the homeostatic values N_j^*, r_j^*, and ϵ^* can be found by simultaneously solving the equilibrium condition (found by taking $\partial N_j = 0$) $\eta\rho = 2\epsilon/(2\epsilon - 1)$, the geometric growth equations $N_j = N_0 \eta^j$ and $r_j = r_0 \rho^j$, and the bone marrow output rate $2\epsilon r_M N_M = \beta_M$ for ϵ, η, and ρ. For example, given a system with $M = 28$, $N_0 = 400$, and $\beta_M = 3.5 \times 10^1 1$; the values $\epsilon = 0.82$, $\eta = 1.97$, and $\rho = 1.31$ are obtained.

B. Statistical mechanics of cell motion

B.1. Particle simulation

To test the validity of the Langevin equation with parameters (9.6), a simulation of colliding particles was constructed to perform exact realizations of the system modeled by the LE. In order to parallel typical cell culturing experimental conditions we consider a system in two spatial dimensions. From the discussion in Section 9.3.1 the particles we wish to simulate should exhibit the following behavior:

- They inhabit a finite volume within a confined space.
- Their free trajectories (in between collisions) are straight lines traversed at constant speed.
- They cannot inhabit the same space, and instead collide elastically.
- They can proliferate resulting in increased number density over time.

With this in mind, a simulation was constructed which evolves the positions of all particles in the system over time.

B.1.1. Particle properties

We take a set of particles $c_i \in \mathcal{C}$, each equipped with a velocity $\mathbf{v}_i$ and a position $\mathbf{x}_i$ and inhabiting a circular area of volume πr_i^2 centered around $\mathbf{x}_i$. For simplicity all particles are taken to have equal size $r_i = r$ and mass $m_i = 1$. The $\mathbf{v}_i$ of a particle is constant in time until it collides, at which point the new velocity is calculated as an elastic moment transfer.

B.1.2. Particle collisions

For a collisions between two particles c_i and c_j, the new velocity $\mathbf{v}_i'$ of c_i is given by

$$\mathbf{v}_i' = \mathbf{v}_i - \frac{(\mathbf{v}_i - \mathbf{v}_j)(\mathbf{x}_i - \mathbf{x}_j)}{(\mathbf{x}_i - \mathbf{x}_j)^2}(\mathbf{x}_i - \mathbf{x}_j) \tag{B.1}$$

Collisions themselves are detected as overlapping cells. With time evolved in fixed discrete increments Δt, the time t' at which a collision occurred during the increment – i.e. the point at which the distance between their centers of mass was $2r$ – can be found from the individual particle velocity vectors and their Euclidean distance. Thus the overlap which occurring during the increment can be reversed and the remainder of their movement can be performed in the new velocities. If a mutual overlap of more than two cells is detected – for example c_1 overlapping both c_2 and c_3 – the earliest collision is performed, as the collisions detection algorithm must be run again in order to correctly identify the any subsequent collisions that occurred in the timestep.

B.1.3. Confined space: minimum image periodic boundaries

The particles are confined to a two-dimensional rectangular plane $\mathcal{S} \subset \mathbb{R}^2$ with minimum image periodic boundary conditions. Concretely, this means that a particle is allowed to leave the plane $\mathcal{S}$ – i.e. $\mathbf{x}_i$ may reach any value in $\mathbb{R}^2$ – however its interactions are calculated with its *image* in S, which is the position it would have if the bounds were periodic. This convention allows for obtaining displacement measurements which are much larger than the dimensions of $\mathcal{S}$.

B.1.4. Accounting for center of mass drift

The lack of true boundaries in the simulation implies that any nonzero momentum associated with the center of mass at initiation will remain present throughout the entire time evolution. In the ensemble of particles this translates to an average drift in the particles' velocity. As the Langevin equation (8.4) assumes there is no drift present, it is necessary to ensure it is zero at initiation. This is done by measuring the center of mass velocity $\mathbf{V}$, subtracting it from each individual particle's velocity, and rescaling all

velocities to ensure the average energy remain unchanged. This can be summarized as performing the following operation on the particle velocities $\mathbf{v}_i$:

$$\mathbf{v}_i \to \sqrt{\frac{E}{E - \mathbf{V}^2/2}}(\mathbf{v}_i - \mathbf{V}) \tag{B.2}$$

B.1.5. Population growth

The number of particles $N(t)$ in the system follows a predetermined growth curve such as (9.7). With the time-evolution performed in increments Δt, particles are added to the system according to

$$\text{floor}\,[N(t + \Delta t)] - \text{floor}\,[N(t)] \tag{B.3}$$

at random non-occupied spaces in $\mathcal{S}$. The particles are added at velocity $\mathbf{v} = 0$ and for each addition all particle speeds are rescaled according to (B.2).

B.1.6. Sketch of the simulation algorithm

The system is initialized with the number of particles determined by $N(t = 0)$ at randomly distributed unoverlapping positions in $\mathcal{S}$ and given initial velocities at fixed speeds $|\mathbf{v}_i| = s_0$ and random directions, after which the center of mass distribution is removed as described in Section B.1.6. The time evolution is then performed in fixed increments Δt. Each timestep contains the following events:

(i) New particles are added to the system as described in Section B.1.5.

(ii) Each particle in $c_i \in \mathcal{C}$ is moved according to its current velocity $\mathbf{v}_i$.

(iii) Collisions detection and resolution is performed successively until no more are found:

a) Overlapping cells are detected as collisions. If a collision involves more than 2 cells, only the first that occurred is taken into account.

b) Any recorded collisions are resolved as described in Section B.1.2.

(iv) The current state of the system ($\mathbf{x}_i(t)$ and $\mathbf{v}_i(t)$) is recorded.

B.2. Numerically simulating the Langevin equation

Stochastic differential equations (SDEs) of the form

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + g(t, X_t)\,\mathrm{d}W_t \tag{B.4}$$

whereby $\mathrm{d}W_t$ is Wiener noise – i.e. Gaussian distributed with mean 0 and variance $\mathrm{d}t$ – can be simulated numerically in a similar manner as for standard ODEs, however the inclusion of a stochastic term requires adapted algorithms. Any numerical solutions to a Langevin equation described in this thesis were obtained using the Julia package DifferentialEquations.jl [3], and the SOSRI algorithm developed by C. Rackauckas and Q. Nie [4].

Bibliography

[1] David Dingli, Arne Traulsen, and Jorge M Pacheco. "Compartmental architecture and dynamics of hematopoiesis". In: *PloS one* 2.4 (2007), e345.

[2] Daniel T Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of computational physics* 22.4 (1976), pp. 403–434.

[3] Christopher Rackauckas and Qing Nie. "Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia". In: *Journal of Open Research Software* 5.1 (2017).

[4] Christopher Rackauckas and Qing Nie. "Stability-Optimized High Order Methods and Stiffness Detection for Pathwise Stiff Stochastic Differential Equations". In: *arXiv:1804.04344 [math]* (2018). URL: `http://arxiv.org/abs/1804.04344`.

[5] Neil A Weiss. *A course in probability*. Addison-Wesley, 2006.