Max-Planck-Institut für
Evolutionsbiologie, Plön

From the Department of Evolutionary Theory,
Max Planck Institute of Evolutionary Biology, Plön,
of the University of Lübeck
Director: Prof. Dr. Arne Traulsen

Population Dynamics with Epistatic Interactions and its Applications to Mathematical Models of Cancer

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences

Submitted by
Benedikt Bauer
from Lünen
Lübeck, 2015

First referee: Prof. Dr. Arne Traulsen

Second referee: Prof. Dr. Karsten Keller

Date of oral examination: September 11, 2015

Approved for printing. Lübeck, September 14, 2015

Abstract

Cells of every organism undergo somatic mutations. Many mutations do not significantly affect the gene's function, while other mutations impair the gene's function. Often, this impairment leads to decreased survival chance of the cell, such that the cell dies. If necessary, this cell can be unproblematically be replaced. Sometimes, however, that impairment disturbs the cell's life cycle in a way that decreases its chance of cell death, or apoptosis, or increases its rate of cell division. Uncontrolled cell proliferation can then lead to the formation of cancer, and ultimately to a tumor. Usually not only one, but a handful of mutations are necessary to affect different safety mechanisms of the cell.

Often, genes influence each other, which is called *epistasis*. Also some oncogenes underlie epistatic interactions. In the Burkitt Lymphoma the hallmark mutation – an IG/MYC translocation – is believed to actually lower the cell's chance of survival. In concert with other mutations, however, it forms a lymphoma. Basic knowledge about the initiation of cancers where the genes of the cancerous mutations underlie epistatic interactions is rare and difficult to acquire in experimental system.

The work described in this thesis is a theoretical analysis of systems with epistatic interactions in cancer initiation. The population dynamics of an abstract system with two different types of mutations between which epistatic interactions exist is analyzed. One type is deleterious by itself, the other one is (nearly) neutral. If the deleterious mutation is accompanied by enough mutations of the other types, the cell has a fitness advantage. We find, amongst others, that the cancer deploying cell lineage has most likely acquired that specific mutation only subsequently to the other, non-deleterious mutations, which inhibit the negative effect of that particular mutation. It is conceivable that epistatic effects could change the order of mutations not only in the survival and proliferation rates of the cell, but also in the mutation rates. Hence, we further pursue the question: "If the deleterious mutation increases the mutation rate for acquiring the necessary, additional mutations, how does this change the probability for the order of mutations?". We develop a recursive algorithm for the computation of the probability density functions of the different mutational pathways over time. Finally, we develop a model aiming at describing the initiation of Burkitt Lymphoma. Lastly, an outlook is given explaining future research directions based on epistasis in cancer initiation.

Kurzfassung

In den Zellen eines jeden Organismus häufen sich Mutationen an. Viele dieser Mutationen üben keinen signifikanten Effekt auf die Funktionalität des entsprechenden Genes aus, andere wiederum können die Funktionalität beeinträchtigen. Häufig führt diese Beeinträchtigung zu einer verringerten Lebensdauer der Zelle, sodass diese stirbt. Falls nötig wird die Zelle wird daraufhin ersetzt. In seltenen Fällen kommt es jedoch vor, dass die Beeinträchtigung den Lebenszyklus der Zelle so ändert, dass die Sterberate verringert wird oder die Teilungsrate erhöht. Unkontrollierte Zellvermehrung kann so zu Krebs und letztendlich zu einem Tumor führen. In der Regel braucht es dafür nicht nur eine, sondern mehrere Mutationen, die verschiedene Sicherheitsmechanismen der Zelle beeinflussen.

Häufig beeinflussen sich die Gene gegenseitig; dies wird *Epistase* genannt. Auch einige Onkogene unterliegen epistatischen Effekten. Beim Burkitt Lymphoma ist die gängige Meinung, dass die Hauptmutation – eine *IG/MYC* translocation – die Lebensdauer der Zelle herabsetzt. Zusammen mit weiteren Mutationen bilden sich jedoch Lymphomzellen. Ein grundlegendes Verständnis über die Initiierung von Krebs, bei welchem die Gene der krebserregenden Mutationen epistatischen Effekten unterliegen, ist unvollständig und nur schwierig über Experimente zu erlangen.

Diese Arbeit beschäftigt sich mit der theoretischen Untersuchung eines Systems mit epistatischen Interaktionen in der Initiierung von Krebs. Wir analysieren die Populationsdynamik eines abstrakten Systems mit zwei verschiedenen Mutationstypen, welche epistatisch miteinander interagieren. Der eine Mutationstyp ist alleine nachteilig, der andere (annähernd) neutral. Wenn die nachteilige Mutation von genügend Mutationen des anderen Typs zusammen trifft in einer Zelle, erhält diese einen Fitnessvorteil. Laut unseren Ergebnissen muss die krebserzeugende Zelllinie sehr wahrscheinlich die neutralen Mutationen, die den Fitnessnachteil der anderen Mutation aufheben, erlangen. Erst anschließend kann die eigentlich nachteilige Mutation sich in der Zelle halten, da diese in dem jetzigen Mutationshintergrund vorteilig ist. Es ist jedoch auch möglich, dass epistatische Effekte nicht nur auf die Zellteilung und ihre Lebensdauer wirken, sondern auch auf die Mutationsrate. Daher gehen wir weiter der Frage nach "Falls die nachteilige Mutation die Mutationsrate für die zusätzlichen, notwendigen Mutationen erhöht, inwiefern ändert dies die Reihenfolge der Mutationen?". Wir entwickeln einen rekursiven Algorithmus, der die Wahrscheinlichkeitsdichte der verschiedenen Mutationspfade über die Zeit numerisch berechnet. Anschließend erarbeiten wir ein Model, welches die Initiierung von Burkitt Lymphoma beschreibt. Schlussendlich wird ein Ausblick gegeben für zukünftige Forschungsrichtungen basierend auf Epistase in der Krebsentstehung.

Contents

Chapter 1

Introduction

1.1 Motivation

Cancer is older than mankind. The earliest written recordings of cancer in humans dates back to approximately 3000 BC [Hajdu, 2011]. Despite being known for such a long time, cancer is still the leading cause of death in the industrialized world and the second leading cause of death in developing countries [Jemal et al., 2011]. Cancer research in the past decades has led to a much better understanding of this disease. Especially improvements in early diagnostic techniques have greatly improved the chances of getting cured [Hochberg et al., 2013]. At the same time, the aging of the world population and adoption of cancer-causing behaviors, above all smoking and poor dietary choices, increases the global burden of cancer [Jemal et al., 2011].

Furthermore, knowledge about the dynamics of cancer tissue is still young. It was not until Charles Darwin has formulated the theory of evolution [Darwin, 1859] that one could think of cancer as an evolving disease, which originates from the organism's own body cells.

Nowadays, it is somewhat more common to think of cancer as an evolutionary process, where different alterations have to be acquired in one cell that affect the cell's ability to produce daughter cells or die [Armitage and Doll, 1954; Attolini and Michor, 2009; Gerstung and Beerenwinkel, 2010; Greenman et al., 2007; Hanahan and Weinberg, 2000; Jones et al., 2008; Lengauer et al., 1998; Michor et al., 2004; Parmigiani et al., 2009; Sjöblom et al., 2006; Traulsen et al., 2010; Wodarz and Komarova, 2005; Wood et al., 2007]. If the cells divide too often, the population grows very fast. Throughout this thesis, this can be thought of as the formation of a tumor.

Cancer is undoubtedly a major health risk and consequently an important field of research. The attentive reader, who might not be familiar with mathematical biology, might ask himself at this point: How can mathematics help

in cancer research? It is true that the investigation of biological questions traditionally requires experiments in biology. To a large extent, this is still the case. The questions, however, become more and more complex and doing experiments often takes a long time and is very expensive. In vivo experiments, especially in humans, are often even impossible to perform. Moreover, instead of analyzing the actual system using biological experiments, one can simplify it into a more abstract, theoretical model. Using this theoretical model allows one to perform investigations that are not possible with a with a biological system. Figure 1.1 shows how biological investigations, which might be too expensive or not possible at all to perform, can be supplemented by using mathematical modeling.

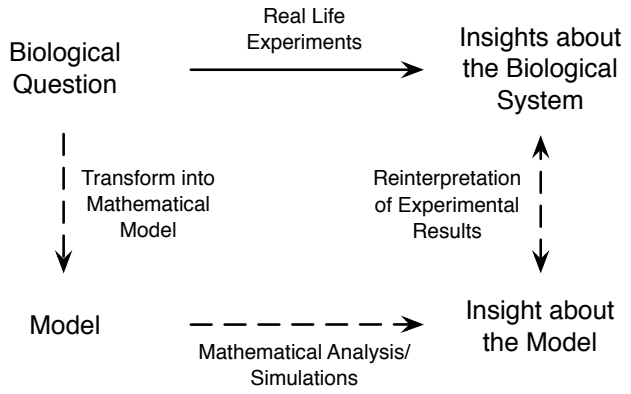

Figure 1.1: A biological question is traditionally investigated using experiments in biology (solid arrow). If performing those experiments is not possible, developing an abstract, theoretical model representing the biological system can introduce further tools to analyze the system (dashed arrows).

While in modeling there is always a system that is supposed to be represented by the model, pure theory deals with abstract theoretical systems. The difference is that there is not necessarily a specific real life system that is linked to the theory. Evolution, for instance, is in itself a theoretical concept. It holds true for any conceivable species or population. Often, pure theoretical results help with the modeling process which is why the two subfields usually go hand in hand and are not easily disentangled.

In physics, for example, mathematical modeling is extremely common. We only need to think of models for the description of gravity [Newton, 1872]. Also

in chemistry, theoretical modeling takes up an inherent part. Amongst others, the research field of *Molecular Dynamics* helps to understand the movement and nature of proteins and other molecules, cf., e.g., [Levitt, 1976]. Modeling is not as prevalent in biology as in the other life sciences, partly due to the enormous complexity of biological systems. The explanation of the commonly observed sex ratio of 1:1 in most species was one of the earliest stepping stones in biological modeling [Fisher, 1930]. Nowadays, theoretical cancer models help to understand the initiation and progression of cancerous cells [Antal and Krapivsky, 2011; Beerenwinkel et al., 2007; Bozic et al., 2010; Gerstung and Beerenwinkel, 2010; Michor et al., 2004; Reiter et al., 2013] as well as improving treatment strategies [Leder et al., 2014].

Of course, theoretical models cannot capture reality at its fullest, but they can reflect specific aspects of the underlying biology reasonably well. The art of modeling is to create a model that captures all the biology necessary and that is, at the same time, simple enough to be understood and handled. One can compare modeling to the creation of a map. If we want to gain knowledge about the subway system of a city, a map showing all the streets might be more complete, but at the same time, it is more confusing to analyze. At the same time, street information might be of interest to see how to get to a subway system. It is barely ever clear what to include in a model, and depending on the question we ask about a specific system, we might even need to use different models for the same system. Having a complete subway system without street connection and one where we only see the stations but therefore also streets, is an example where it makes sense to have two models for one system. Combined, we know how to get to a subway station and where we can go by subway. Merging both maps in one would possibly make the map too confusing, though. Therefore, additionally one needs to make sure that models, which seem similar on a microscopic level, also gives similar results on the macroscopic level. In [Wu et al., 2015], the macroscopic feature of the probability of fixation of a subpopulation is investigated for two processes, which seem very similar on a microscopic level. Interestingly, we find that the fixation probability is only the same for very strict constraints. To avoid these issues, a collaboration of people with expertise in different fields is therefore often inevitable. Biologists and physicians know the biological processes, what

is important, what can be left out, or what can be regarded as constant, while people with a background in mathematics or physics know how to phrase a biological process in a theoretical model. Moreover, modeling is a dynamical process; often some macroscopic properties, which result from the model, can be checked against the corresponding properties of the biological system. Depending on the comparison, the microscopic assumptions of the model can be refined. This way, a compatible model can be developed and at the same time the refinement process already contributes to the understanding of the system.

Cancer is a complex disease, and while many aspects of cancer can be examined by experiments in mice or similar model organisms, there are still a lot of things that cannot be tested experimentally. Experiments in humans are not possible at all, and only a few things can be tested. Theoretical biology, therefore, adds a powerful tool for the investigation and understanding of cancer. The work described in this thesis aims at helping understand cancer initiation especially in cancers where the effect of mutations depends on other mutations in that particular cell.

In the following sections of the introduction, a brief overview is given over the biological background and the mathematical tools used throughout this thesis.

1.2 Biological Background

For the modeling of cancer, we first need to introduce the necessary biological terminology as well as the basic biological background knowledge used throughout this thesis. The explanations here are supposed to give only a short overview. For a more detailed description, see for example [Watson et al., 2014; Wodarz and Komarova, 2014].

First, we need to know that the cell's information is stored in deoxyribonucleic acid (DNA). The DNA's building blocks are *nucleic acids*. The sequence of these building blocks represents (most of the) information a cell needs to function. For the purpose of this thesis, we do not need other mechanisms of information storage. Whenever the cell needs to produce or repair something, it is reading the information from the DNA to build the necessary proteins.

We call a consecutive segment of the DNA, which is read off for a specific protein, a *gene*. Note, that this is only a simplified definition for gene. Upon cell division, some errors might happen. Errors in the DNA are called *mutation*.

The DNA is not just stored as a single strand, but in different packages, so called *chromosomes*. In humans we have 23 different chromosomes. Upon cell division, parts of the chromosomes might break and are transferred onto a different chromosome. These kinds of mutations are called *translocations*. When a part of chromosome a is transferred onto chromosome b, this is written as *t(a;b)*. In Chapters 2 and 4, the translocation of the *MYC* gene from chromosome 8 onto an *IG* gene, e.g., on chromosome 14, is of importance. We will write t(8;14) for this translocation. Also, just single nucleic acids can be deleted, inserted, or substituted. Those mutations are called *point mutations*.

Healthy cells divide in a controlled manner. Immune cells, for example, divide upon an infection to increase the number of immune cells and strengthen the immune response. If cells die, they are over the long term replaced by descendants of other cells. Cells might die because of mutations that disturb the cell's functioning. An abnormally behaving cell often realizes itself that it does not function regularly and hence induces a mechanism that ultimately leads to cell death. This process is called *intrinsic apoptosis*. Cells that behave abnormally and do not die by themselves are artificially taken out of the body's system by immune cells. This is called *extrinsic apoptosis*. Lost cells are being replaced by descendants of neighboring cells. Some mutations do not affect the cell's ability to survive or proliferate. The cell then simply continues to live. A few mutations, however, disturb the cell's reproductive cycle, so that it either cannot be killed in a regulated way or it divides uncontrollably. This can lead to the formation of a tumor.

The ratio between a cell's chance of undergoing apoptosis and its rate of proliferation is called *fitness* throughout this thesis, following e.g., [Bozic et al., 2010]. We view fitness always from the cell's point of view. A deleterious mutation therefore is a mutation that lowers the cell's fitness and thus increases its chance of undergoing apoptosis compared to its proliferation rate. Analogously, an advantageous mutation increases the cell's fitness, and the ratio between proliferation and apoptosis rate is increased. This increased rate

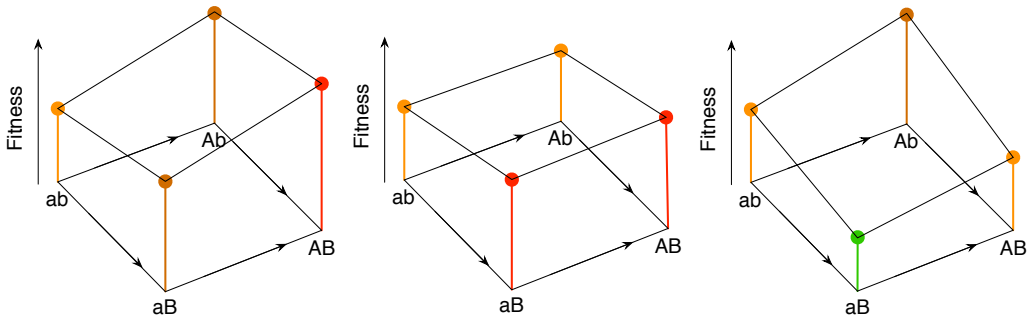

Figure 1.2: Different examples for a minimal system with two genes and without any epistatic interactions are given. Left: The positive fitness effects of the single A and single B mutations are simply added for an individual with both the A and B mutation. Middle: The gene type A can be seen as a neutral mutation. Right: For an AB individual the fitness advantage of the A mutation is reversed by the deleterious effects of the B mutation.

in proliferation compared to apoptosis can lead to the formation of a tumor. The higher fitness of cells is consequently unfavorable for the organism.

Usually, more than one mutation is necessary for the development of cancer. Mutations that seems to be deleterious might lead to an advantage in combination with other mutations. The dependence between genes is generally referred to as *epistasis*, which is explained in greater detail in the following section.

1.3 Epistasis

Epistasis means that the effect of one gene depends on the genetic background. For example, the gene for the hair color and the gene causing albinism are subject to epistatic effects, since the gene for the hair color does not play a role when the person is albinotic anyway.

In a non-epistatic fitness landscape, the change in fitness induced by a mutation is independent on other mutations possibly present in the cell. Figure 1.2 shows different fitness landscapes for systems without epistasis. In an *epistatic fitness landscape* however, the change in fitness induced by a mutation does depend on the genetic background. Hereby, so called *sign epistasis* refers to a system in which one mutation alone is deleterious, but in concert

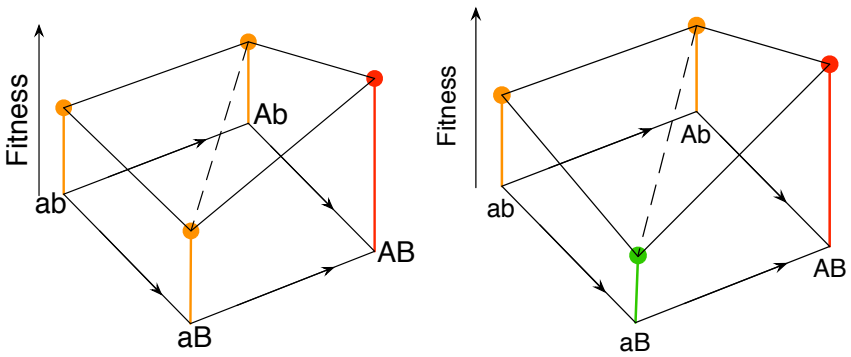

Figure 1.3: Two examples for a system with two genes and epistatic interactions are shown. Left: The single A and B mutations can be seen as recessive mutations in a diploid organism, where the mutation of only one copy does not have an effect on the individual's phenotype. Only when both genes are mutated the phenotype is affected. Right: A presumably neutral mutation and a deleterious one ultimately increase the individual's fitness when both genes are mutated.

with another mutation it is advantageous or the other way around, i.e., the sign of the effect on fitness changes. Figure 1.3 shows two examples for systems with epistatic effects, where the right example refers to sign epistasis. Biologically, the left example could denote a gene in a diploid individual. A mutation of one of the alleles (Ab or aB) has no effect at all, and only a mutation of both alleles (AB) affects the individual's fitness.

Epistatic effects in cancer initiation seem to be relevant for a wide range of cancers. For example, we can think of the inactivation of a tumor suppressor gene discussed by Knudson in the context of retinoblastoma [Knudson, 1971]. This inactivation is neutral for the first mutation but highly advantageous for the second mutation, and can hence be viewed as an interaction of genes [Iwasa et al., 2005; Nowak et al., 2002, 2004; Vogelstein and Kinzler, 2004]. Another case is found in lung carcinomas, where activation of each of two oncogenes (*SOX2* and *PRKCI*) alone is insufficient, but in concert they initiate cancer [Justilien et al., 2014]. In other cases, there is clear evidence for sign epistasis: The *ras* family of proto-oncogenes is also discussed to underlie epistatic effects. Amplification of *ras* leads to senescence in the cell. Nevertheless, *ras* is a well known oncogenic driver gene. Hence, the *ras* mutation needs to be accompanied by other mutations [Elgendy et al., 2011; Serrano

et al., 1997].

Often, both fitness and mutation rates underlie epistatic interactions. This is of particular interest when a deleterious mutation increases the mutation rate for other advantageous mutations. For the system depicted in Figure 1.3 on the right side, this would mean that the mutation rate from aB to AB is increased. Such a system is analyzed in more detail in Chapter 3.

1.4 Branching Process and Probability Generating Function

A branching process is an individual based process. In a classical branching process (also known as Galton-Watson Process), each individual lives for one time unit and then reproduces independently according to some probability density and dies. There is, in general, no restriction on the number of offspring. In biological processes it is common to assume two offspring or no offspring, representing cell division and apoptosis, respectively. Also one offspring is possible, because not all cells divide at the same time and one offspring would represent simply living on. For other organisms, a distribution for a larger number of a large number of offspring makes sense. Trees, for example, are able to produce many offsprings at once. The individuals can be of different types. Mutation or adaptation can be represented as the production of offspring of a different type.

The independence between individuals makes it easy to calculate extinction probabilities, more specifically the probability for different types to be present in the system or not to be present. In a branching process, this probability can be recursively obtained using *probability generating functions* (PGFs). Since the relation between PGFs and the probability for a type to be present is the main tool we are using, we devote this subsection to giving a short overview about this connection, although it is rather technical and well studied [e.g., Haccou et al., 2005; Kimmel and Axelrod, 2002].

The PGF for a time discrete, one-type process is in general defined as

$$f(z) = \sum_{k=0}^{\infty} p_k z^k, \tag{1.1}$$

where k denotes the number of offspring and p_k represents the probability of having k offspring (the focal individual dies in this context) [Haccou et al., 2005]. For many biological processes, for example cell multiplication, it makes sense to only consider offspring numbers of 0 (death), 1 (nothing happens), and 2 (cell division). But in other biological systems, it makes sense to consider many offspring at once. Our analysis is not restricted to any particular offspring distribution. However, for the sake of simplicity, we restrict our example to the so called *binary splitting*, i.e., either two descendants (p_2) or death of the focal individual (p_0). The use of the argument z is not obvious at this point. If we set z equal to 0, the probability generating function reduces to $f(0) = p_0$, which is the extinction probability for a population of one individual in one time step. Since all individuals behave independently, $f(0)^N = p_0^N$ is the extinction probability for a population of size N in one time step. Now looking at the extinction probability within two time steps, we note that with probability p_2, we would have two individuals in the next time step originating from one individual. Hence, the extinction probability for a single individual within two time steps is

$$p_0 + p_2 p_0^2 = f(f(0)) = f^{\circ(2)}(0), \tag{1.2}$$

and that of population with N individuals is

$$\left(p_0 + p_2 p_0^2\right)^N = \left(f(f(0))\right)^N = \left(f^{\circ(2)}(0)\right)^N. \tag{1.3}$$

Continuing for further time steps, we see that $f^{\circ(t)}(0)$ is the extinction probability for a single individual and $\left(f^{\circ(t)}(0)\right)^N$ is the extinction probability for the system within t time steps.

As of now we assumed that individuals reproduce clonally, i.e., giving rise to the same type. Now we continue investigating the extinction probability for a two-type process. Let us think of the two types A and B, where an A individual can produce any number of A or B individuals, and respectively for B. Then the general PGFs, if the process starts with one type A or one

type B individual, are defined as

$$f_A(z_A, z_B) = \sum_{k_A=0}^{\infty} \sum_{k_B=0}^{\infty} p^A_{k_A,k_B} z_A^{k_A} z_B^{k_B}, \tag{1.4}$$

$$f_B(z_A, z_B) = \sum_{k_A=0}^{\infty} \sum_{k_B=0}^{\infty} p^B_{k_A,k_B} z_A^{k_A} z_B^{k_B}, \tag{1.5}$$

where $p^A_{k_A,k_B}$ ($p^B_{k_A,k_B}$) denotes the probability of one A (B) individual producing k_A A and k_B B individuals in the next time step. Let us recover the extinction probability as for the one-type process. If we set both z_A and z_B equal to zero and assume that we start with one A individual, we obtain a similar result as above for the total extinction probability

$$f_A(0,0) = p^A_{0,0}. \tag{1.6}$$

Oftentimes, one is rather interested in the extinction or non-presence of just one particular type. Let us assume we are only interested in the presence of B individuals. The probability of having no B individuals in time step 1 is the sum over all probabilities where no B offspring is being produced $\sum_{k_A=0}^{\infty} p^{A(B)}_{k_A,0} = f_{A(B)}(1,0)$, starting with one A (B) individual. Now looking at the probability of having no B individuals in time step 2, we need to account for the probability of having k_A A and k_B B individuals being produced in the first time step. This leads to

$$\sum_{k_A=0}^{\infty} \sum_{k_B=0}^{\infty} p^A_{k_A,k_B} f_A(1,0)^{k_A} f_B(1,0)^{k_B} = f_A(f_A(1,0), f_B(1,0)) =: f_A^{\circ(2)}(1,0). \tag{1.7}$$

Continuing this procedure and analogous to the one-type process, the probability of having no B individual in time t is $f_A^{\circ(t)}(1,0)$.

This procedure can be extended to a multi-type process with an arbitrary number of types in a similar fashion. For further information and detailed insights into extinction of branching processes, we refer to [Kimmel and Axelrod, 2002] and [Haccou et al., 2005].

We have seen above that the independence between individuals is beneficial to formulate the PGF of the process, and that we can compute the extinction

probability from the PGF. The independence has hence an advantage when it comes to analyzing the system. At the same time, it can be a huge drawback on the realism of the system. It is therefore important to address this concern. This thesis mostly focuses on modeling the initiation of cancer, i.e., until the first successful cancer lineage has been established. Up to this point, the number of cancer cells is still very small, allowing for all cells – cancer and healthy cells – to harbor enough nutritions as well as having enough space. Since resource limitation does not play a great role in this scenario, we neglect competition among cells. Therefore, we can safely assume all cells to act independently.

Another concern is the lack in consideration of spatial structure. For the population dynamics of solid tumors, spatial structure might be an important factor to be taken into account. Non-solid tumors, however, are usually structureless, particularly in the initiation process. This thesis is motivated by Burkitt Lymphoma which forms a structure in the later stage of the disease. The presumed cell of origin for Burkitt Lymphoma is a germinal center dark zone cell [Basso and Dalla-Favera, 2015; Klein and Dalla-Favera, 2008]. Cells in the germinal center are not structurally organized. Hence, there is no need to assume spatial structure for the initation of Burkitt Lymphoma.

At the same time, we want to model a growing population size to capture a more biologically realistic environment. For these reasons, a classical branching process captures enough realism whereas at the same time it allows for some analytical exploration due to its simplicity. One such analytical investigation possible is the calculation of the extinction probability for a branching process [Athreya and Ney, 1972]. This allows for developing interesting properties of a multi-type system, such as the time distribution and mutational path probabilities as explained in Chapter 3 in detail.

Branching processes are a common choice for cancer modeling [Jagers, 1970]. For example, Bozic et al. [2010] use a time-discrete branching process to model the progression of glioblastoma multiforme and pancreatic cancer. The authors provide a correlation between the number of driver mutations and the total number of mutations (driver and neutral passenger mutations) in the tumor. Moreover, they are able to calculate the selective advantage for the cell's fitness provided by the driver mutations. They find a surprisingly

small fitness advantage of only 0.4% (±0.04%).

Further, in [Durrett and Moseley, 2010], a branching process is used modeling the "evolution of resistance and progression to disease during clonal expansion of cancer". The authors consider a process with an arbitrary number of mutations. Those mutations are being acquired in a sequential order. The different types can have arbitrary cell division and apoptosis rates, where the division rate has to be greater than the death rate. In the paper an approximative closed form solution for the waiting time of obtaining a cell lineage with an arbitrary number of mutations is developed.

These are just two well-known examples of the application of branching processes in cancer modeling; cf., [Antal and Krapivsky, 2011; Bozic et al., 2013; Durrett et al., 2010; Kimmel and Axelrod, 1991; Kimmel et al., 1992; Reiter et al., 2013] for further references.

While for the purpose of this thesis a classical branching process is a good choice, a more general branching process extends the range of realism. Allowing arbitrary life spans, dependence on resources, and population size or density [Haccou et al., 2005], would make it possible to analyze the population dynamics also after the initiation of the cancer. The calculation of the extinction probability is, however, considerably more complicated when individuals do not behave independently. In Chapter 5 we derive the probability generating function for a frequency dependent branching process. In particular, we use a time-continuous two-type branching process where the division rates depend on the number of individuals of both subpopulations.

1.5 Structure of the Thesis

The following Chapter 2 describes a theoretical (cancer) model motivated by Burkitt Lymphoma. We investigate the dynamics of a system with epistatic interactions between one driver and several passenger mutations. We also introduce the term *secondary driver mutation* for a mutation that has no direct effect on the cell's fitness, but an indirect one due to interactions with the (primary) driver mutation.

Knowing the order of mutations, i.e., which mutation happens first, which one second and so on, can be of particular interest in biology since it gives

indication on possible subpopulations. Those subpopulations present a risk for potential relapses in cancer. Chapter 3 deals with this question. First, the simpler issue of calculating the time distribution until a certain mutant type is reached is being studied. Modifying the resulting approach, we are able to describe a framework with which it is possible to directly investigate the order of mutations needed.

In Chapter 4 we complete the picture of epistatic interactions in cancer initiation by developing a simple, yet meaningful model about Burkitt Lymphoma. To ensure biological correctness, we worked with Prof. Dr. Siebert and Dr. Aukema from the Institute for Human Genetics of the University of Kiel. Despite the apparent simplicity of the model, it is already very hard to handle and henceforth an analytic investigation is not feasible. We therefore analyze this system numerically by running simulations.

While this thesis gives a complete picture about epistatic interactions in cancer initiation, it can only serve as a starting point. In research one can always dig deeper, extend models, and understand a scientific question in more detail. In Chapter 5 we collect ideas on questions arising from this work, which could have a strong impact on our understanding of cancer initiation.

Chapter 2

Cancer Initiation with Epistatic Interactions Between Driver and Passenger Mutations

---

In Chapter 1.3 we have briefly discussed epistasis. Apart from modeling the inactivation of two copies of one tumor suppressor gene in a diploid organism [Iwasa et al., 2005; Komarova et al., 2003], epistasis in modeling cancer has to our knowledge not been discussed in detail in the literature so far. However, epistatic interactions are for many cancers evidentially part of the initiation process of cancer. In this chapter, we therefore analyze the effect of epistasis in a theoretical framework. This chapter is based on the publication [Bauer et al., 2014], coauthored by Reiner Siebert and Arne Traulsen. A detailed summary of the authors contributions to this publication can be found at the end of this thesis. While the model presented here is not designed as a specific model for a particular cancer, the underlying idea is motivated by clinical and experimental observations in Burkitt Lymphoma. Our analysis is based on a multi type branching process. Using simulations allows us to investigate single realizations. We further give analytical results for the average number of cells with different mutations. Lastly, we discuss the time distribution for cancer cells to occur, i.e., the incidence curve. We find that this model shows a very interesting dynamical behavior, which is distinct from the dynamics of cancer initiation in the absence of epistasis.

2.1 Introduction

We were motivated by genetic studies in Burkitt Lymphoma, a highly aggressive tumor, where a single genetic alteration has an impact on a wide range

of other genes, some of them affect cell growth while others induce apoptosis. More specifically, a chromosomal translocation between the *MYC* protooncogene on chromosome 8 and one of three immunoglobulin (*IG*) genes is found in almost every case of Burkitt Lymphoma [Allday, 2009; Hummel et al., 2006; Richter et al., 2012; Sander et al., 2012]. This leads to deregulated expression of the *MYC* RNA and in consequence, to deregulated *MYC* protein expression. The *MYC* protein acts as a transcription factor and has recently been shown to be a general amplifier of gene expression [Lin et al., 2012; Nie et al., 2012], targeting a wide range of different genes. Most importantly, *MYC* expression induces cell proliferation. In Burkitt Lymphoma, the *IG-MYC* fusion is evidently the key mutation for tumorigenesis [Campo, 2012; Salaverria and Siebert, 2011; Schmitz et al., 2014; Zech et al., 1976]. However, *MYC* plays also a key role in inducing apoptosis [Hoffman and Liebermann, 2008; Pelengaris et al., 2002; Wang et al., 2011]. Thus, the *IG-MYC* translocation alone would lead to cell death. Therefore, the *IG-MYC* translocation has to be accompanied by additional mutations, which deregulate the apoptosis pathways, such as mutations affecting for example the tumor suppressor gene *TP53* or *ARF* [Allday, 2009; Richter et al., 2012; Sander et al., 2012]. These additional mutations have probably only little direct impact on the cell's fitness, since apoptosis is rare. Hence, these mutations cannot be considered as primary driver mutations in the context of Burkitt Lymphoma. However, in combination with the *MYC* mutation these additional mutations decrease the apoptosis rate. Consequently, the cells proliferate fast and the population grows accordingly, leading to tumorigenesis. Because all cells carry the *MYC* mutation in Burkitt Lymphoma, but fast growth does not start immediately with that mutation, it seems to confer its large fitness advantage only in a certain genetical context. Thus, interactions between different mutations may crucially affect the dynamics of cancer progression. Due to the fact that those additional mutations do not confer a direct fitness advantage, they cannot be considered as driver mutations. Nevertheless, at least some of them are necessary in order for the *MYC* mutation to become advantageous for the cell. Therefore, they cannot be regarded as true passenger mutations, either. Throughout this chapter, we therefore call these additional mutations "secondary driver mutations". Note that a system with neutral tumor suppressor

Here, we are interested in the dynamics of such an epistatic model, which we illustrate by stochastic, individual based simulations. In addition, we derive analytical results for the average number of cells with different combinations of mutations and find a good agreement with the average dynamics in individual based computer simulations. Furthermore, we discuss the computation of the waiting time until cancer initiation. Our results show that the dynamics in such systems of epistatic interactions are distinct from previous models of cancer initiation [Antal and Krapivsky, 2011; Beerenwinkel et al., 2007; Bozic et al., 2010; Gerstung and Beerenwinkel, 2010; Michor et al., 2004; Reiter et al., 2013], which may have important consequences for the treatment of such cancers. While in previous models there is a steady increase in growth with every new mutation, in our model there is a period of stasis followed by a rapid tumor growth.

Of course, the biology of Burkitt Lymphoma is much more complex than modeled herein. To make the model more realistic one would have to distinguish between the different secondary driver mutations, since different genes contribute differently to the cells fitness, especially in a cell where the *IG-MYC* fusion is present. Our model is not aimed to realistically describe such a situation in detail. Instead, we focus on the extreme case of the so called *all-or-nothing* epistasis [see, e.g., Barrick and Lenski, 2013; Meyer et al., 2012, from experimental evolution] to illustrate its effect on the dynamics of cancer initiation. As there is no theoretical analysis of epistatic effects in cancer initiation so far, a well understood minimalistic model seems to be necessary in order to illustrate the potential impact of epistasis on cancer progression. Our minimalistic model clearly shows that epistasis can lead to a qualitatively different dynamics of cancer initiation.

2.2 Mathematical Model

We analyze cancer initiation in a homogenous population of initially N cells with discrete generations. In every generation, each of the N cells can either die or divide. If a cell divides, its two daughter cells can mutate with mutation probabilities μ_{D} for the driver mutation and μ_{P} for secondary driver mutations (where the P indicates that these would be called passenger muta-

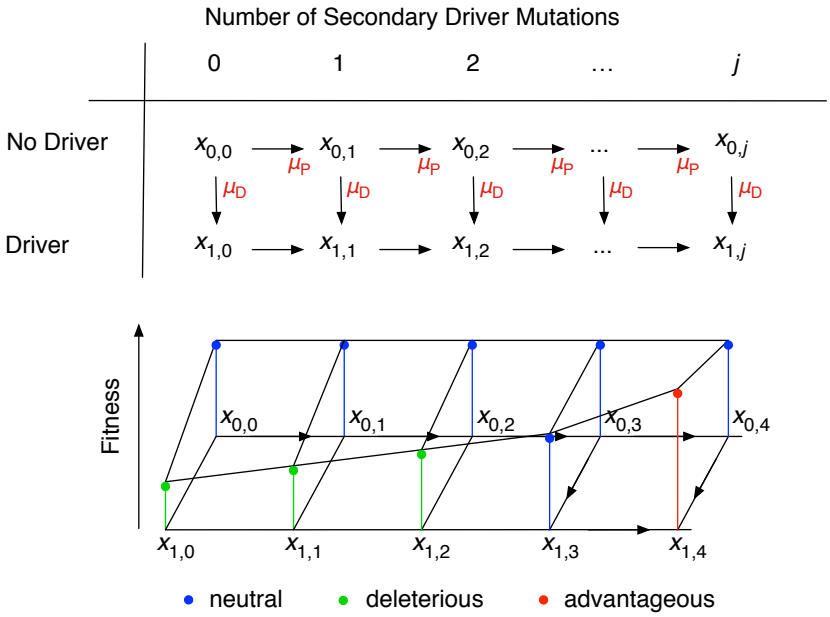

Figure 2.1: Mutational pathways of the model. Top: The entries $x_{i,j}$ denote the number of cells with or without the primary driver mutation ($i = 1$, or $i = 0$ respectively), and j secondary driver mutations. Bottom: Cells with only secondary driver mutations have neutral or nearly neutral fitness. The fitness of cells with the primary driver mutation depends on the number of secondary driver mutations within the cell, leading to an epistatic fitness landscape.

tions in closely related models). In principle, we could drop the assumption that these two mutation probabilities are independent on the cell of origin, but this would lead to inconvenient notation. We neglect back mutations and multiple mutations within one time step, because their probabilities are typically very small. Figure 2.1 summarizes the possible mutational pathways of the model. The variables $x_{i,j}$ denote the number of cells with or without the primary driver mutation ($i = 1$ or $i = 0$ respectively), and j secondary driver mutations.

A cell's probability for apoptosis and proliferation depends on the presence of the primary driver mutation and on the number of secondary driver mutations it has accumulated. For cells with no mutations, the division and apoptosis probabilities are both equal to $\frac{1}{2}$. This implies that the number of

cells is constant on average as long as no further mutations occur. We assume that the initial number of cells is high and thus we can neglect that the population would go extinct [Haccou et al., 2005]. For our parameter values, the expected extinction time of our critical branching process exceeds the average life time of the organism by far. The average time until extinction is for a critical branching process infinity [Haccou et al., 2005; Kimmel and Axelrod, 2002]. From Figure 2.5 we see that the simulations did not need longer than 70000 time steps. The probability to have gone extinct until 70000 for a critical, binary splitting branching process is of the order of 10^{-6}.

For cells without the primary driver mutation, each secondary driver mutation leads to a change in the cell's fitness by s_{P}, where fitness is modeled in terms of division probability. For cells with the primary driver mutation, the fitness advantage obtained with each secondary driver mutation is s_{DP}. The driver mutation increases both the apoptosis rate and the proliferation rate. The increase in the apoptosis rate is s_{D^-} and the increase in the division rate is s_{D^+}. With these parameters, the proliferation rate $b_{0,j}$ for cells with j secondary driver mutations but without the primary driver mutation is

$$b_{0,j} = \frac{1}{2}(1 + s_{\mathrm{P}})^j, \tag{2.1}$$

whereas the proliferation rate $b_{1,j}$ for such cells with the primary driver mutation is

$$b_{1,j} = \frac{1}{2} \cdot \frac{1 + s_{\mathrm{D}}^+}{1 + s_{\mathrm{D}}^-}(1 + s_{\mathrm{DP}})^j, \tag{2.2}$$

where $s_{\mathrm{D}}^+ < s_{\mathrm{D}}^-$. The apoptosis rates, denoted as $d_{0,j}$ and $d_{1,j}$ are simply one minus the proliferation rate

$$\begin{aligned} d_{0,j} &= 1 - \frac{1}{2}(1 + s_{\mathrm{P}})^j, \\ d_{1,j} &= 1 - \frac{1}{2} \cdot \frac{1 + s_{\mathrm{D}}^+}{1 + s_{\mathrm{D}}^-}(1 + s_{\mathrm{DP}})^j. \end{aligned} \tag{2.3}$$

Note, that we could also incorporate the driver fitness effect in terms of the product $(1+s_{\mathrm{D}}^+)(1-s_{\mathrm{D}}^-)$. However, we then need to take care that $s_{\mathrm{D}}^- < 1$. Using the fraction as in (2.2) and (2.3), we can freely choose s_{D}^+ and s_{D}^-.

For small values of s_D^+ and s_D^- the term for the driver fitness effect can be approximated by the product

$$\frac{1+s_{\mathrm{D}}^+}{1+s_{\mathrm{D}}^-} \approx \left(1+s_{\mathrm{D}}^+\right)\left(1-s_{\mathrm{D}}^-\right). \tag{2.4}$$

2.3 Results

2.3.1 Simulations

Mutations occur at fixed rates μ_{D} and μ_{P} for primary and secondary drivers, respectively. For a long time, the overall fitness does not increase noticeably. For $s_{\mathrm{P}} = 0$, it stays on average constant. Hence, the total number of cells stays approximately constant. Only when a cell with enough secondary driver mutations and also the primary driver mutation arises, the cell's fitness is increased substantially beyond the fitness of other cells and its chance of proliferation is significantly increased. At that point, the total number of cells starts to increase rapidly, see Figure 2.2. In models presented in literature so far, the cell's fitness is increased independently with every (driver) mutation (see e.g., [Beerenwinkel et al., 2007; Bozic et al., 2010]). Although the total number of cells increases exponentially, these models do not find a sudden burst in the number of cells. Instead, the number of cells starts growing slowly with the first (driver) mutation, where the average growth of population increases with every (driver) mutation.

In Figure 2.3, the total number of cells is subdivided into the number of cells with different numbers of mutations. The left panel presents the cells that have not acquired the primary driver mutation, the right one shows cells with the primary driver mutation. Cells with the primary driver mutation, but not enough secondary driver mutations, arise occasionally, but those cells die out quickly again – thus, their average abundance is small. Cells without the primary driver mutation do not die out, they also do not induce fast growth, cf. Figure 2.3. Only cells that have obtained enough secondary driver mutations and in addition acquire the primary driver mutation, divide so quickly that the population size increases rapidly.

The parameters in Figure 2.2 and 2.3 have been chosen such that a cells acquires a substantial growth advantage once the primary driver mutation

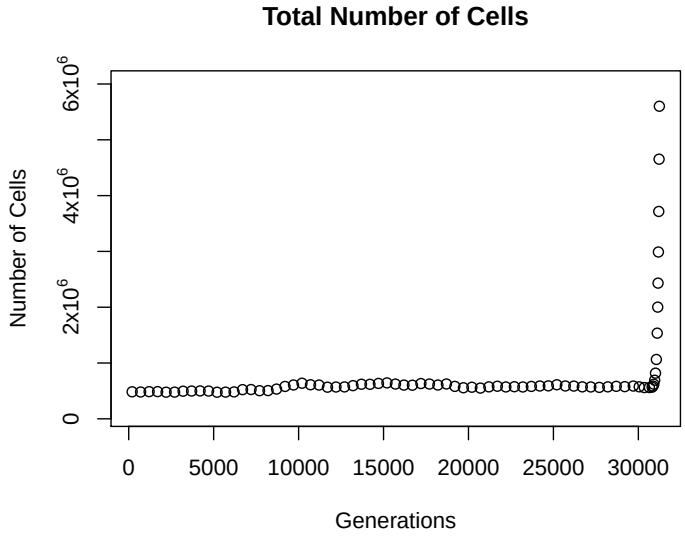

Figure 2.2: The dynamics of the total number of cells. Initially, the total cell count increases only marginally but at some point, a combination of primary and secondary driver mutations within one cell with a large fitness benefit arises and leads to rapid exponential proliferation (parameters: Initial number of cells $N = 500000$, secondary driver fitness advantage $s_{\mathrm{P}} = 10^{-5}$, the primary driver fitness advantage $s_{\mathrm{D}^+} = 0.05$, primary driver disadvantage $s_{\mathrm{D}^-} = 0.1$, advantage of a secondary driver mutation in the presence of the primary driver mutation $s_{\mathrm{DP}} = 0.015$, mutation rates for secondary driver mutations $\mu_{\mathrm{P}} = 2 \cdot 10^{-5}$, mutation rate for the primary driver mutation $\mu_{\mathrm{D}} = 5 \cdot 10^{-6}$).

co-occurs with 4 secondary driver mutations. This event can occur at any time and hence, in some simulation the number of cells can increase very early, whereas in other simulations the number of cells does not undergo fast proliferation for many generations. Consequently, the rate of progression has an enormous variation. For the parameters from our figures, the time at which rapid proliferation occurs varied between ≈ 9300 and ≈ 63000 generations in 500 simulations. The average number is therefore not particularly meaningful. The distribution of these times is discussed in more detail below.

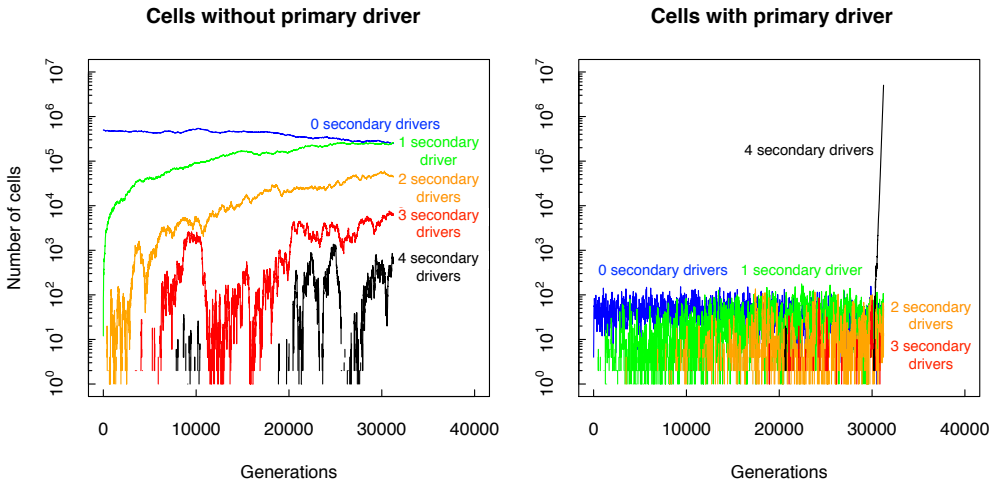

Figure 2.3: The dynamics of the number of cells with different numbers of mutations in a single simulation. Top: The number of cells without any mutation decreases slightly, whereas the number of cells with secondary driver mutations, but no primary driver mutation, slowly increases. Bottom: While a small number cells with the primary driver mutation is present from the beginning, at first these primary driver mutations are not accompanied by sufficiently many secondary driver mutations to compensate the disadvantage arising from the primary driver. Only when a primary driver mutation is co-occurring with enough secondary driver mutations (in this case four), the number of cells with the primary driver starts to increase rapidly. (parameters: $N = 500000$, $s_{\mathrm{P}} = 10^{-5}$, $s_{\mathrm{D}^+} = 0.05$, $s_{\mathrm{D}^-} = 0.1$, $s_{\mathrm{DP}} = 0.015$, $\mu_{\mathrm{P}} = 2 \cdot 10^{-5}$, $\mu_{\mathrm{D}} = 5 \cdot 10^{-6}$)

2.3.2 Analytical Results

2.3.2.1 Average Number of Cells

We can calculate the average number of cells with a certain number of mutations at a given generation t. The number of cells which do not have the primary driver mutation and k secondary driver mutations (i.e., $x_{0,k}(t)$) changes on average by means of the cell's fitness and it decreases by the mutation rate

$$x_{0,k}(t) = (1 - (\mu_{\mathrm{P}} + \mu_{\mathrm{D}}))(1 + s_{\mathrm{P}})^k x_{0,k}(t-1) + \mu_{\mathrm{P}}(1 + s_{\mathrm{P}})^{k-1} x_{0,k-1}(t-1), \tag{2.5}$$

where $x_{0,-1}(t) \equiv 0$. The solution of Equation (2.5) for $s_P \neq 0$, i.e., if the secondary driver mutations are not neutral, is

$$x_{0,k}(t) = N\mu_P^k(1-(\mu_D+\mu_P))^{t-k}(1+s_P)^{k(k-1)/2}\prod_{i=0}^{k-1}\frac{1-(1+s_P)^{t-i}}{1-(1+s_P)^{i+1}}, \quad (2.6)$$

where N denotes the initial number of cells. The mathematical proof of Equation (2.6) is given in 7.1. Note, that the product can be written in terms of a q-binomial coefficient [Koekoek et al., 2010],

$$\prod_{i=0}^{k-1}\frac{1-(1+s_P)^{t-i}}{1-(1+s_P)^{i+1}} = \begin{bmatrix} t \\ k \end{bmatrix}_{1+s_P}. \quad (2.7)$$

For the case $s_P = 0$, we take the limit of the q-binomial coefficient [e.g., Kac and Cheung, 2002]

$$\lim_{s_P\to 0}\begin{bmatrix} t \\ k \end{bmatrix}_{1+s_P} = \binom{t}{k} \quad (2.8)$$

and obtain

$$x_{0,k}(t) = N\mu_P^k(1-(\mu_D+\mu_P))^{t-k}\binom{t}{k}, \quad (2.9)$$

which is the result that is also expected if the secondary driver mutations are neutral and accumulated independently of each other.

Intuitively, the term $\mu_P^k(1-(\mu_D+\mu_P))^{t-k}$ describes the probability of obtaining exactly k mutations in t generations. There are different possibilities when the mutations happen, these possibilities are captured by the binomial coefficient $\binom{t}{k}$. Thus, we have a growing polynomial term in t and a declining exponential term in t, since $(1-(\mu_D+\mu_P)) < 1$.

In the case of $s_P \neq 0$, the interpretation is similar. Here, additionally the fitness advantage for secondary driver mutations has to be taken into account. Since the number of cells with j secondary driver mutations grows with $(1+s_P)^j$, also the number of cells that can mutate grows. Hence, the factor $(1+s_P)^{k(k-1)/2}$ is multiplied to the expression and the binomial coefficient turns into the q-binomial coefficient.

For cells that have obtained the primary driver mutation and k secondary

μ_{P}	Mutation rate for secondary driver mutations
μ_{D}	Mutation rate for the primary driver mutation
$\nu_{\mathrm{P}} = 1 - \mu_{\mathrm{D}} - \mu_{\mathrm{P}}$	Probability for a cell without the primary driver mutation to not mutate
$\nu_{\mathrm{D}} = 1 - \mu_{\mathrm{P}}$	Probability for a cell with the primary driver mutation to not mutate
s_{P}	Fitness change of a secondary driver mutation (see (2.1), (2.2))
s_{D^+}	Fitness advantage of the primary driver mutation (see (2.1), (2.2))
s_{D^-}	Fitness disadvantage of the primary driver mutation (see (2.1), (2.2))
s_{DP}	Fitness advantage of combination of a secondary driver and the primary driver mutation
$\varsigma_{\mathrm{P}} = 1 + s_{\mathrm{P}}$	Fitness according to a secondary driver without the primary driver mutation
$\varsigma_{\mathrm{DP}} = 1 + s_{\mathrm{DP}}$	Fitness according to the combination of primary and secondary driver mutation
$\varsigma_{\mathrm{D}} = \frac{1+s_{\mathrm{D}}^+}{1+s_{\mathrm{D}}^-} \approx 1 + s_{\mathrm{D}}^+ - s_{\mathrm{D}}^-$	Fitness according to a primary driver mutation with no secondary driver

Table 2.1: Summary of our abbreviations

driver mutations, the situation is slightly more complex. There are $k+1$ different possibilities on how to obtain k secondary driver and the primary driver mutation, since some of the secondary driver mutations may have occurred before the primary driver mutation has been acquired, whereas others may have occurred afterwards. Let $x_{1,k}^{(p)}(t)$ denote the number of cells with the primary driver mutation and k secondary driver mutations, when the primary driver mutation has happened in a cell with p secondary driver mutations. Note that $0 \leq p \leq k$. The change in the number of cells now depends on p. Using the abbreviations from Table 2.3.2.1 to simplify our notation, we have

$$x_{1,k}^{(p)}(t) = \begin{cases} \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k} x_{1,k}^{(p)}(t-1) + \mu_{\mathrm{P}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k-1} x_{1,k-1}^{(p)}(t-1), & \text{if } p < k \\ \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k} x_{1,k}^{(p)}(t-1) + \mu_{\mathrm{D}}\varsigma_{\mathrm{P}}^{k} x_{0,k}(t-1), & \text{if } p = k. \end{cases} \tag{2.10}$$

To express the average number of cells in total we need to sum over all

possible pathways,

$$x_{1,k}(t) = \sum_{p=0}^{k} x_{1,k}^{(p)}(t). \tag{2.11}$$

In 7.2 in the Appendix, we show that the analytical solution of Equation (2.11) is

$$\begin{aligned} x_{1,k}(t) =& N \sum_{p=0}^{k} \mu_D \mu_P^k \varsigma_{\mathrm{D}}^{k-p} \varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2} \frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{n+1})} \\ &\times \left[\nu_{\mathrm{P}}^{t-p} \left(\sum_{r=0}^{p} \frac{\left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^r \varsigma_{\mathrm{P}}^{\frac{r(r-1)}{2}}}{\prod_{n=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{n})} \begin{bmatrix} p \\ r \end{bmatrix}_{\varsigma_{\mathrm{P}}} \right) \right. \\ &- \sum_{j=p}^{k} \nu_{\mathrm{P}}^{j-p} (\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})^{t-k} \prod_{m=j}^{k-1} \frac{1-\varsigma_{\mathrm{DP}}^{t-m-1}}{1-\varsigma_{\mathrm{DP}}^{k-m}} \\ &\left. \times \left(\sum_{r=0}^{p} \frac{\left(-\varsigma_{\mathrm{P}}^{j-p+1}\right)^r \varsigma_{\mathrm{P}}^{\frac{r(r-1)}{2}}}{\prod_{n=p}^{j}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{n})} \begin{bmatrix} p \\ r \end{bmatrix}_{\varsigma_{\mathrm{P}}} \right) \right], \end{aligned} \tag{2.12}$$

if $s_{\mathrm{P}} \neq 0$. The summation over p indicates the different mutational pathways. An intuitive explanation of this somewhat lengthy equation is given in 7.3 in the Appendix.

Interestingly, the case for $s_{\mathrm{P}} = 0$ is much more challenging. The underlying problem is that the normal binomial coefficient cannot be expressed in a sum in the way the q-binomial coefficient can be expressed [Koekoek et al., 2010],

$$\begin{bmatrix} t \\ p \end{bmatrix}_{\varsigma_{\mathrm{P}}} = \prod_{j=0}^{p-1} \frac{1-\varsigma_{\mathrm{P}}^{t-j}}{1-\varsigma_{\mathrm{P}}^{j+1}} = \frac{\sum_{r=0}^{p} \left(-\varsigma_{\mathrm{P}}^{t}\right)^r \left(1/\varsigma_{\mathrm{P}}\right)^{\frac{r(r-1)}{2}} \begin{bmatrix} p \\ r \end{bmatrix}_{1/\varsigma_{\mathrm{P}}}}{\prod_{j=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{j+1})}. \tag{2.13}$$

When summing over all generations of the population with p secondary driver mutations to derive the expression for the population of cells with $p+1$ secondary driver mutations, we have to calculate the sum

$$\sum_{i=0}^{t-p-1} (\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{p})^{i} \binom{t-i-1}{p}. \tag{2.14}$$

When we go further and calculate the expression for the population with k

secondary driver mutations, we need to apply this sum $(k-p)$-times and hence we obtain a multi sum,

$$\sum_{i_0=0}^{t-k-1} \left(\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\right)^{i_0} \sum_{i_1=0}^{t-k-i_0-2} \left(\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k-1}\right)^{i_1} \cdots \sum_{i_{k-p}=0}^{t-2k+p-i_0-i_1-\cdots-i_{k-p-1}-1} \left(\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{p}\right)^{i_{k-p}} \times \binom{t-2k+p-i_0-i_1-\cdots-i_{k-p-1}-1}{p}. \tag{2.15}$$

Only an analytical expression for this multi sum would allow a closed solution of the problem with $s_{\mathrm{P}} = 0$. Also, taking the limit $s_{\mathrm{P}} \to 0$ of our expression for $s_{\mathrm{P}} \neq 0$ is a substantial mathematical challenge. However, we can use our solution for $s_{\mathrm{P}} \neq 0$ for arbitrarily small values of s_{P}. Moreover, numerical considerations show that the result for $s_{\mathrm{P}} = 0$ is very close to the case of $s_{\mathrm{P}} \ll 1$.

In Figure 2.4, the dynamics of the average number of cells with a certain number of mutations, is shown, both without and with the primary driver mutation. Simulation results for $s_{\mathrm{P}} = 0$ agree very well with the analytical result obtained for $s_{\mathrm{P}} \neq 0$.

2.3.2.2 Distribution of time until cancer initiation

Next, let us take a look at the distribution of the time it takes until rapid proliferation occurs. We are able to give a recursive algorithm by using the probability generating functions for the calculation of the time distribution. The derivation of the algorithm is given in following Chapter 3. Figure 2.5 shows the good agreement between simulations and the recursive calculation. The time distribution for low t can be approximated by a power law, as shown in the inset of Figure 2.5. The exponent of the power law is approximately 3.4. If all mutations were neutral, one would expect a lead coefficient of approximately 4 to accumulate five mutations, as derived by [Armitage and Doll, 1954]. In our case, the curve increases slower. Numerical considerations show that the main reason for this is that, in contrast to [Armitage and Doll, 1954], we allow extinction: Many lineages that have accumulated mutations go extinct before the final, cancer causing mutation arises.

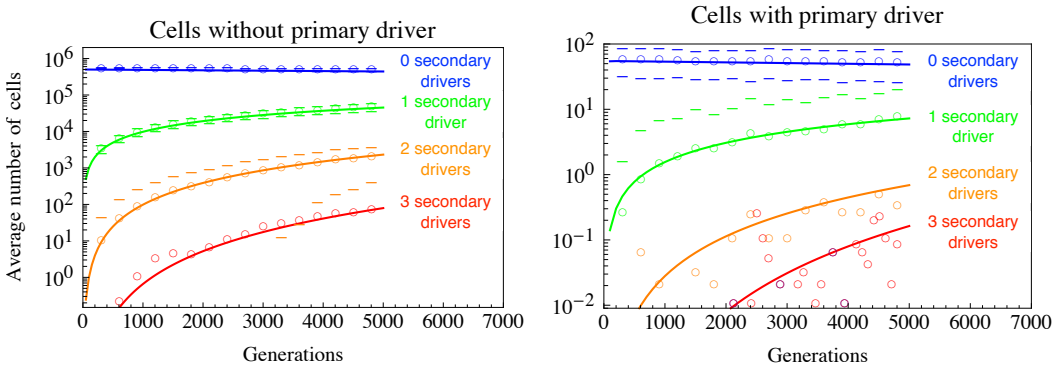

Figure 2.4: Dynamics of the number of cells with different number of secondary driver mutations, without (left) and with (right) the primary driver mutation. Simulation results averaged over 500 independent realizations for $s_P = 0$ (circles) agree almost perfectly with the analytical result obtained for $s_P = 10^{-5}$. The bars represent the standard deviation. Cells with no mutation have a very small relative standard deviation and cells with one mutation (i.e., one passenger only or the driver only) have a relatively small standard deviation. In contrast, cells with two passenger mutations for instance have a very broad standard deviation in the beginning that is approximately four times the average number. Only in few realizations, a primary driver mutations co-occurs with several secondary drivers, hence the simulations for these cases shows a large spread (parameters: $N = 500000$, $s_P = 10^{-5}$, $s_{D^+} = 0.05$, $s_{D^-} = 0.1$, $s_{DP} = 0.015$, $\mu_P = 2 \cdot 10^{-5}$, $\mu_D = 5 \cdot 10^{-6}$).

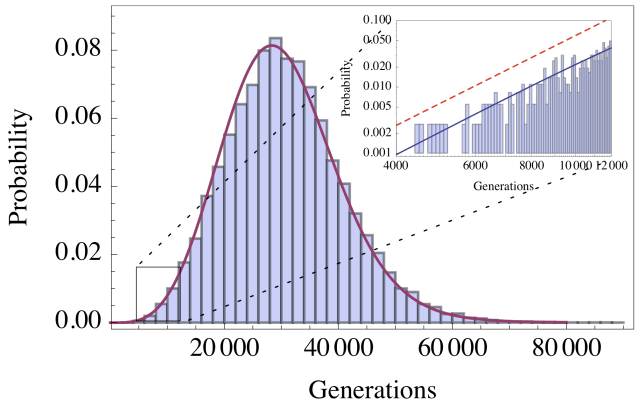

Figure 2.5: Comparison between the analytical calculation and simulations of the distribution until cancer initiation (main panel). Solid lines represents analytic solution. The analytical calculation and simulations of the model agree very well. This inset illustrates that the time distribution initially follows a power law with an exponent of ≈ 3.4 shown as a dashed line (parameters: $N = 500000$, $s_{\mathrm{P}} = 10^{-5}$, $s_{\mathrm{D}^+} = 0.05$, $s_{\mathrm{D}^-} = 0.1$, $s_{\mathrm{DP}} = 0.015$, $\mu_{\mathrm{P}} = 2 \cdot 10^{-5}$, $\mu_{\mathrm{D}} = 5 \cdot 10^{-6}$, distribution over 20000 independent realizations).

2.4 Discussion

Most models in literature assume that each mutation leads to an independent and steady increase in the cells' fitness [Beerenwinkel et al., 2007; Bozic et al., 2010; Gerstung and Beerenwinkel, 2010; Michor et al., 2004; Reiter et al., 2013]. In this context, neutral passenger mutations have no causal impact on cancer progression. Only recently, some authors have considered passenger mutations not only as neutral byproducts of the clonal expansion of mutagenic cells, but as having a deleterious impact on the cells' fitness [McFarland et al., 2013].

Here, we have described a model in which the fitness of the driver mutation strongly depends on the number of passenger mutations the cell has acquired. These passenger mutations, which we have termed secondary driver mutations, lead only to a small change in fitness or no change in fitness at all. As illustrated in Figures 2.2, 2.3, and 2.4, the number of cells stays roughly constant for a long time before it rapidly increases, despite the fact that mutations occur in the process permanently. This effect of the population dynamics of cancer initiation is very different from models in which mutations do not interact with each other. We speculate that this kind of dynamics can have

important implications for diagnosis and treatment. In principle, the dynamics presented in Figure 2.2 can also be the result of one highly advantageous, but very unlikely driver mutation. But in such a case, cells with the driver mutation should not be present in the population before tumorigenesis. This contradicts with current knowledge about the *MYC* translocation which has also been detected in humans without lymphoma [Müller et al., 1995]. This effect is well captured by our model, as shown in Figure 2.3.

In some tumors, such as Burkitt Lymphoma, the neoplasms is only diagnosed after fast tumor growth has started. In this case, sequencing studies have shown that several mutations are present at the time of examination [Alexandrov et al., 2013; Love et al., 2012; Richter et al., 2012; Schmitz et al., 2012]. Since the patients typically do not have any symptoms before diagnosis of the cancer, it is possible that some mutations have virtually no direct impact on the cells fitness. Nevertheless, they are necessary for the initiation of the cancer, as they indirectly allow the driver mutation to initiate rapid cell growth. This agrees well with our epistatic model, where (nearly) neutral secondary driver mutations occur at a fixed rate before the cancer can be diagnosed.

Of course, not all mutations have such an epistatic effect on primary driver mutations, some might even be considered deleterious [McFarland et al., 2013]. Nevertheless, our work shows that mutations that appear to be neutral in one context should not only be regarded as a neutral byproduct of the clonal expansion of mutagenic cells. Instead, in some cases passenger mutations can have a serious impact in cancer initiation, in particular when there are non-trivial interactions between different mutations. In this case the term "passenger" may not be the most appropriate one. To understand the impact of those interactions can be essential for a deeper understanding of the initiation of cancer.

Chapter 3

Calculation of Time Distribution and Path Probabilities

In many biological systems it is interesting to know how long it takes for a mutational process to happen. For example, to be able to predict how long it takes for the bird flu to accumulate the necessary mutations to cross the interspecies barrier. In the previous chapter we were interested in the time it takes for self cells to harvest the mutations necessary to turn into cancer cells.

This chapter is based on the publication [Bauer and Gokhale, 2015], coauthored by Chaitanya Gokhale. We present an algorithm on how to recursively compute the time distribution for processes in which the individuals proliferate independently of other individuals. By having the time distribution, it is possible to say up to which time the mutational process happens with a probability above a certain threshold. We go even a step further in our manuscript and develop a procedure computing the time distribution for the single mutational pathways. This allows us to derive probabilities for the order in which the mutations are accumulated. Ultimately, this gives information about the subpopulations present in the system. In cancer, for example, subpopulations pose a high threat for a relapse after treatment: A subpopulation might not be targeted by normal chemotherapy and in the worst case only one additional mutation for cells of that subpopulation is needed to become cancerous. Therefore, by knowing which subpopulations are present, it might be possible to develop therapies targeted specifically to those subpopulations.

While there is no closed form solution for the time distributions, a direct computation of the mutational path probabilities would make it possible to interactively analyze the parameter space, even when not all required input parameters, such as mutation, birth, and death rates, are known. Our recursive approach allows to avoid time consuming simulations.

This the issue itself is not only cancer related, but also appears in exper-

imental evolution [Cooper et al., 2003; de Visser et al., 1997; Lenski et al., 1991]. For this reason, this chapter is mostly free of cancer specific terminology, and we rather approach the more general question "How repeatable is evolution?" [Beatty, 2006].

3.1 Introduction

As the metaphor by Stephen J Gould goes 'if we run the tape of life back from the start how likely is it that we will get the same outcome that we see around us today?' [Beatty, 2006]. The pioneering work of Lenski et al. tackled this question experimentally with E. coli. In their system, it is now possible to literally play back evolution from a certain starting point and see where it leads [Blount et al., 2012; Cooper et al., 2003; Lenski et al., 1991; Meyer et al., 2012].

Such empirical explorations made the until then theoretical concept of fitness landscapes tangible. The concept of a fitness landscape is a mapping between the genotype and the phenotype of an organism. Since selection acts on the phenotype, the genotype of each phenotype can be attributed a certain fitness. Connecting the genotypes which are one mutational step away from each other leads to the concept of fitness landscapes [Fisher, 1930; Haldane, 1927]. Such empirical studies do make it clear that predictions will not be based on simple rules but complicated phenomena such as epistasis and epigenetics which play a major role in the process of evolution [Travisano et al., 1995; Travisano and Shaw, 2013; Weinreich et al., 2005].

Epistasis is any deviation from the additive effects of alleles at different loci [Fisher, 1918]. Epistasis gives rise to rugged fitness landscapes which have been found to be quite common in experimental observations in a variety of model systems [de Visser et al., 1997; Jain and Krug, 2007; Szendro et al., 2013; Weinreich et al., 2006]. In particular, reciprocal *sign epistasis* is a necessary condition for having a rugged fitness landscape [Poelwijk et al., 2007]. While in *magnitude epistasis* the fitness always increases (or decreases) with every additional mutation in a non-additive manner, in *sign epistasis*, however, valleys appear in the fitness landscape. A certain mutation might have a lower fitness than the previous state although it eventually leads to

higher fitness. In such a case not all paths in the fitness landscape might be accessible by the population [Weinreich et al., 2006]. Comparing experimental systems to theoretical predictions made on the basis of the underlying fitness landscape helps elucidate the role of microscopic properties of the system in determining the macroscopic evolutionary trajectory. The details of the process such as the mutation rate, fitnesses of individual states and the global population size act as constraints on the accessibility of paths [Szendro et al., 2013]. Using the assumption of strong selection and weak mutation rates (SSWM), the system advances on the fitness landscape in a stepwise fashion. This automatically limits the possible number of adaptive paths [Weinreich et al., 2005].

Evolutionary predictability and the speed of the dynamics is not only determined by the molecular constraints of fitness and mutation rate, but also by population dynamics [Poelwijk et al., 2007]. Theoretical explorations often assume a fixed population size starting at one node of the fitness landscape and its movement is tracked over the course of time [Gokhale et al., 2009]. Increasing the population size, or the mutation rate, we observe the phenomenon of clonal interference [Park and Krug, 2007; Weinreich et al., 2006]. This occurs when a second step mutant arises in a population even when the first step mutation is not fixed. In other words, the SSWM assumption is no longer valid. Clonal interference has been extensively explored experimentally [Elena and Lenski, 2003; Hegreness et al., 2006; Imhof and Schlotterer, 2001] as well as theoretically [Desai et al., 2007; Gerrish and Lenski, 1998; Gokhale et al., 2009; Iwasa et al., 2004; Park and Krug, 2007; Weinreich and Chao, 2005; Weissman et al., 2009]. This phenomenon removes the limit on the accessibility of non-adaptive trajectories. If the fitnesses and mutation rates follow particular conditions, i.e. the mutation rates also underlie epistatic interactions, then such valley crossings might be faster than adaptive trajectories [Gokhale et al., 2009; Lynch and Abegg, 2010].

Populations in real systems are finite and their size can undergo fluctuations which can lead to possible extinction events. Together with the phenomena of clonal interference and epistatic interactions between mutations (correlated rugged fitness landscapes), predicting evolution through a given fitness landscape seems like an impossible task. Herein, we develop a gen-

eral methodology for predicting the most probable path in a fitness landscape with epistatic interactions in a multi-dimensional fitness landscape. To reflect a realistic scenario we use a multi-type branching process (e.g., [Haccou et al., 2005]) to drop the assumption of a constant population size. For presentation purposes we limit ourselves to systems without back mutations. The model in its full generality is free of this assumption, although it is unclear how to define pathways when back mutations are allowed (see Section 7.4 in the Appendix for a detailed explanation). To introduce the framework we begin with a simple model in which the wild type can have two independent mutations leading to the fittest type. Then we increase the number of mutational events it takes to get to the corresponding type leading to a generalization of the methodology. We briefly mention an application of this approach by linking it to a cancer initiation model [Bauer et al., 2014] showing how mutational epistasis changes the path probabilities. Finally we provide an outline on how to extend the model to a general system where different mutations need to be acquired to reach the final mutant.

3.2 Model and Results

3.2.1 Two Dimensional Fitness Landscape

We begin with a minimal fitness landscape. Envision a wildtype ab which can mutate at the two loci to A and B, respectively. With both mutations, the system is in the final state of AB. In such a system there are two different paths as illustrated in Figure 3.1. Traditionally, epistatic models are discussed in terms of different fitness values, whereas the mutation rates stay the same [Poelwijk et al., 2007; Szendro et al., 2013]. A fitness landscape for a system with *sign epistasis* is shown in Figure 3.1. In such a system where the mutation rates stay the same, i.e. $\mu_A = \mu_A^B$ and $\mu_B = \mu_B^A$, it is clear that the path via Ab is the most probable one. However, if the mutation rates change, e.g., $\mu_A^B \gg \mu_A$, also the path via aB can become accessible. Changing mutation rates amounts to including epistasis in the mutational landscape in addition to epistasis in the fitness landscape [Sasaki and Nowak, 2003].

For the four types of the above model, we need to consider four different

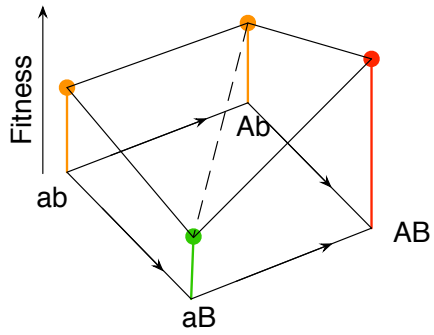

Figure 3.1: **Mutational pathways for a system with two loci.** There are two different pathways to reach the final mutant. Fitness is represented by the size of the circles denoting the types. Thus the wildtype ab and Ab have a similar fitness whereas AB has a significantly greater fitness compared to the wildtype while aB is much less fit than the wildtype. When all mutation rates are the same, the pathway via aB would be not adaptive, since this type has a low fitness. If the mutation rate μ_A^B is large enough, especially if $\mu_A^B \gg \mu_A$ (indicated by the thick arrow), this pathway becomes accessible.

PGFs, one for each type

$$\begin{aligned} f_{ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{ab} + b_{ab}((1-\mu_A-\mu_B)z_{ab} + \mu_A z_{Ab} + \mu_B z_{aB})^2, \\ f_{Ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{Ab} + b_{Ab}((1-\mu_B^A)z_{Ab} + \mu_B^A z_{AB})^2, \\ f_{aB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{aB} + b_{aB}((1-\mu_A^B)z_{aB} + \mu_A^B z_{AB})^2, \\ f_{AB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{AB} + b_{AB} z_{AB}^2, \end{aligned} \tag{3.1}$$

where b_i and d_i are the birth and death probabilities of type i. The exponent of 2 arises from a branching process with binary splitting. The arguments $z_{ab}, \ldots, z_{AB}$ correspond to extinction probabilities of the respective type as discussed above. The functions f_i correspond to the extinction probability of the whole process given that the process starts with a single individual of type i. The PGF f_i at time t is recursively calculated as [Haccou et al., 2005; Kimmel and Axelrod, 2002]

$$f_i^{(t)}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) = f_i(f_{ab}^{(t-1)}, f_{Ab}^{(t-1)}, f_{aB}^{(t-1)}, f_{AB}^{(t-1)}). \tag{3.2}$$

3.2.2 Time Distribution

Using the generating functions we now approach the extinction time distribution of the binary branching process. Particularly starting with 1 wild type individual, the probability of having no AB-individual at time t is $f_{ab}^{(t)}(1,1,1,0) =: f(t)$. Thus the probability of having at least 1 AB-individual at time t is $1 - f(t)$. The probability, that at least 1 AB-individual appears *exactly* at time t is the probability, that there is an AB-individual at t minus the probability that there was already one at time $t-1$:

$$\tau(t) = (1 - f(t)) - (1 - f(t-1)) = f(t-1) - f(t) \tag{3.3}$$

Starting with N wild type individuals the probability that there are no AB-individual at time t is then $f(t)^N$. This leads to the time distribution,

$$\tau(t) = f^N(t-1) - f^N(t). \tag{3.4}$$

However, the arising AB should start a lineage that does not die out. Hence we are interested in the probability of having a *successful* AB-individual. To calculate this we use the known extinction probability of an AB-individual in place of z_{AB}. The probability of an AB-individual going extinct is its death probability divided by its birth probability $e_{AB} := d_{AB}/b_{AB}$ [Athreya and Ney, 1972]. The modified PGFs for this purpose then read as

$$\begin{aligned}
f_{ab}(z_{ab}, z_{Ab}, z_{aB}) &= d_{ab} + b_{ab}((1 - \mu_A - \mu_B)z_{ab} + \mu_A z_{Ab} + \mu_B z_{aB})^2, \\
f_{Ab}(z_{ab}, z_{Ab}, z_{aB}) &= d_{Ab} + b_{Ab}((1 - \mu_B^A)z_{Ab} + \mu_B^A e_{AB})^2, \\
f_{aB}(z_{ab}, z_{Ab}, z_{aB}) &= d_{aB} + b_{aB}((1 - \mu_A^B)z_{aB} + \mu_A^B e_{AB})^2.
\end{aligned} \tag{3.5}$$

Note, that the PGF for the final mutant type is not necessary anymore. We can now calculate the time distribution until the first *successful* mutant appears the same way as described above. Figure 3.2 shows the perfect agreement between the recursive solution and 5000 simulations. The parameters, specified in the Figure 3.2's caption, are entirely arbitrarily chosen to reflect an epistatic fitness landscape as sketched in Figure 3.1. The reason we chose a very slightly advantageous fitness for the type Ab-individuals is solely to stress the fact, that this method holds for any fitness values, not only if some

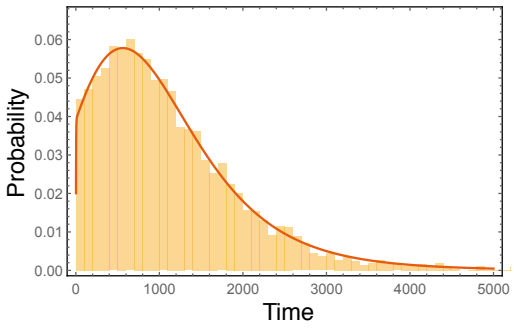

Figure 3.2: **Time distribution of reaching the final mutant for a four type fitness landscape as in Fig. 3.1.** Solid line represents the recursive solution and the bars represent 5000 simulations. The parameters are: Death probabilities: $d_{ab} = 0.5, d_{Ab} = 0.49995, d_{aB} = 2/3, d_{AB} = 0.25$. Birth probabilities are 1 minus the corresponding death probability. Mutation probabilities are $\mu_B = \mu_B^A = 2 \cdot 10^{-6}$, $\mu_A = 2 \cdot 10^{-5}$, $\mu_A^B = 0.005$. Initial population size: $N = 30000$.

are restricted, for example to being neutral.

For a three-type continuous time branching process, as in $A \xrightarrow{\mu_B} B \xrightarrow{\mu_C} C$, the time distribution was computed in [Bozic et al., 2013]. This was done using the analytical solution of the probability generating function for the two-type process $A \xrightarrow{\mu_B} B$ [Antal and Krapivsky, 2011] and the fact that in continuous time mutations follow a Poisson distribution. Adding a second intermediate type, e.g., B_2, would also give such a process but immediately results in unwieldy analytical calculations.

3.2.3 Path Probabilities

In the current example there are two possible paths by which the wildtype can reach the final mutant AB, either $ab \to Ab \to AB$ or $ab \to aB \to AB$. Experimental evidence shows that not all paths are equally probable [Lee et al., 1997; Weinreich et al., 2006]. Beginning with ab then what is the probability of the first AB mutant arising via either path and how long does it take for the different pathways?

The probability, that the first mutant arises exactly at time t via pathway

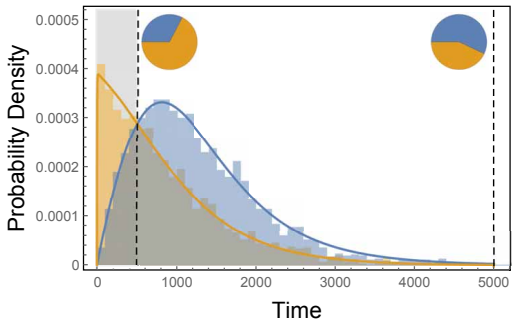

Figure 3.3: **Probability distribution for the different pathways.** Orange represents the pathway via aB and blue the pathway via Ab. The bars are the results of simulations, the solid lines depict the computed results. In the pie charts the distribution of the pathways are illustrated up to 500 time steps (shaded area, left pie chart) and up to 5000 time steps (right pie chart). Stopping after a few lineages have reached the final mutant might lead to a false distribution: The other pathway might just need longer, but have a smaller extinction probability. The parameters are: Death probabilities: $d_{ab} = 0.5, d_{aB} = 2/3, d_{Ab} = 0.49995, d_{AB} = 0.25$. Birth probabilities are 1 minus the corresponding death probability. Mutation probabilities are $\mu_B = \mu_B^A = 2 \cdot 10^{-6}$, $\mu_A = 2 \cdot 10^{-5}$, $\mu_A^B = 0.005$. Initial Population size is $N = 30000$.

Ab is (derived in the Appendix),

$$\rho_{Ab}(t) = f^N(t-1) - (\bar{f}^{(Ab)}(t))^N, \tag{3.6}$$

where $\bar{f}^{(Ab)}(t)$ is defined in Section 7.6 in the Appendix and is being computed in a similar fashion as $f(t)$. The total probability for this path ϱ_{Ab} is then the summation of $\rho_{Ab}(t)$

$$\varrho_{Ab} = \sum_{t=1}^{\infty} \rho_{Ab}(t). \tag{3.7}$$

Computationally the sum would go up to a t_{max}, where $f^{(Ab)}(t_{max}-1) - f^{(Ab)}(t_{max}) < \varepsilon$ (where usually *machine epsilon* is chosen as ε). The total extinction probability of a multi-type branching process is determined by the smallest fixed point $\mathbf{z}^* = (z^*_{ab}, z^*_{Ab}, z^*_{aB}, z^*_{AB})$ of the probability generating functions $\mathbf{f}(\mathbf{z}^*) = \mathbf{z}^*$, where z^*_{ab} is the extinction probability, if the process starts with one ab-individual [Haccou et al., 2005]. Nevertheless those total extinc-

tion probabilities are not suitable for the question, via which path the first successful AB-mutant arises. The problem lies in the time; the pathway via Ab for example could have a very low extinction probability whereas the pathway via aB might have an extinction probability of $1/2$. Intuitively one would expect the path via Ab to be more frequent. However, if the path via aB is much faster (e.g., due to $\mu_A^B \gg \mu_B^A$), one would actually find that each path happens with a probability that approaches $1/2$. Therefore, it is important to do the recursive analysis to include the probability that a successful mutant did not arise through any other path beforehand.

Figure 3.3 shows the probability densities for the different pathways of the minimal model. Interestingly, the pathway via aB is predominantly prominent in the beginning but overall less likely. Hence if experiments are stopped after a short time interval then they might provide conclusions which can be upended by looking at the experiments at a later time point.

3.2.4 Multiple Mutations in two Dimensions

In the earlier model the wildtype had two possible mutations $a \to A$ and $b \to B$. It is possible, that a to A and b to B are a multi-step process. Hence we can assume that it takes m mutations to go from a to A and n to go from b to B. Hence for $m = n = 1$ we recover the simple model as discussed above. The calculation of the time distribution can be directly transferred from the simple model by including all necessary probability generating functions for all available types. Increasing the length of the dimensions has a direct impact on the number of paths leading from the wildtype to the final mutant. In particular there are $N = \binom{m+n}{m}$ possible paths. Assuming in general m mutations in the A dimension and n in the B dimension we enumerate the paths as follows. Path 1 is the path where at first all A mutations and subsequently all B mutations happen. Path 2 is the path where all but one A mutations happen first, then one B, then the last A, and finally all other B mutations. Figure 3.4 shows the different paths for a system with three type A and three type B mutations (left), and for a system with four mutations for type A and one mutation for type B (right). Thus calculating the path probability for

any particular path p now takes the form,

$$\rho_p(t) = f^N(t-1) - \left(\bar{f}^{(p)}(t)\right)^N, \tag{3.8}$$

where $f(t)$ is the probability generating function as in Eq. A.2 and $\bar{f}^{(p)}$ is defined analogously to Eq. A.9 in the Appendix

$$\begin{aligned}\bar{f}^{(p)}(t) &:= \bar{f}_{p_0}^{\circ(t)}\left(\underbrace{1,1,\ldots,1}_{m+n},\frac{d_{m,n}}{b_{m,n}},1,\ldots,1\right)\\ &= \bar{f}_{p_0}\left(\bar{f}_{p_0}^{\circ(t-1)},\bar{f}_{p_1}^{\circ(t-1)},\ldots,\frac{d_{m,n}}{b_{m,n}},\bar{f}_{q_1}^{\circ(t-2)},\ldots,\bar{f}_{q_{mn}}^{\circ(t-2)}\right).\end{aligned} \tag{3.9}$$

Here, the probability generating functions with a p index belong to types along the regarded path (which in total are $m+n+1$ without back mutations, beginning at 0, with which we always label the subindex for the wild type). Accordingly, probability generating functions with a q index are associated with types, that do not belong to the respective path (which are in total $m \times n$). The probability generating function for the final mutant type is again replaced by the extinction probability of this type. We use our framework with this extension on the cancer initiation model proposed in [Bauer et al., 2014]. Therein a model with several mutational steps to reach state A and one mutational step for state B is analyzed (cf. Fig. 3.4). The direct change in fitness for the A mutations is (nearly) zero, and the B mutation alone is even deleterious. However, if an individual obtains all A mutations and the B mutation, the fitness is enhanced which in the model leads to rapid proliferation. Here, we provide an example on how the path probabilities change, when epistasis is not just in the fitness landscape but in the mutational landscape as well. Figure 3.5 compares the path probability distributions with and without epistasis in the mutational landscape. The fitness values, the birth and death probabilities respectively, as well as the "nonepistatic" mutation probabilities, are the same as in chapter 2.

3.2.5 Multi Dimensional Fitness Landscapes

The cancer landscape discussed above is a two dimensional system. In principle it is possible to extend this approach to higher dimensions. For fitness

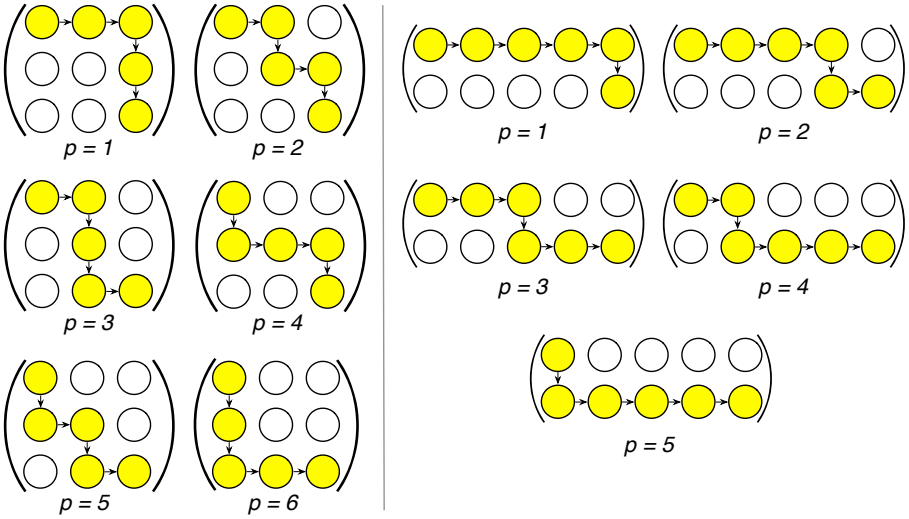

Figure 3.4: **Exemplary numbering of the different mutational pathways for a system with $m = n = 3$ mutations for type A and type B mutations (left), and for a system with $m = 4$ mutations for type A and $n = 1$ mutation for B (right).**

landscapes of higher orders [Khan et al., 2011; Weinreich et al., 2006] it is still possible to write down the system of probability generating functions and apply the approach explained here. The concept remains the same. For each type the probability generating functions are needed except for the final mutant type, here only the extinction probability is necessary (Appendix). Finally the probability generating function for the wild type needs to be recursively calculated for the time distribution. For the path probabilities the probability generating functions related to types not along the considered path again are one time step behind, similar as in Eq. 3.9. However, while we can get accurate experimental data elucidating the fitness landscape, the mutational landscape is usually hard to determine.

3.3 Discussion

A theoretical framework to study mutational pathways in epistatic systems has been presented in this chapter. The crucial part is that in our analysis epistasis affects not only fitness (i.e. proliferation and death rates) but also

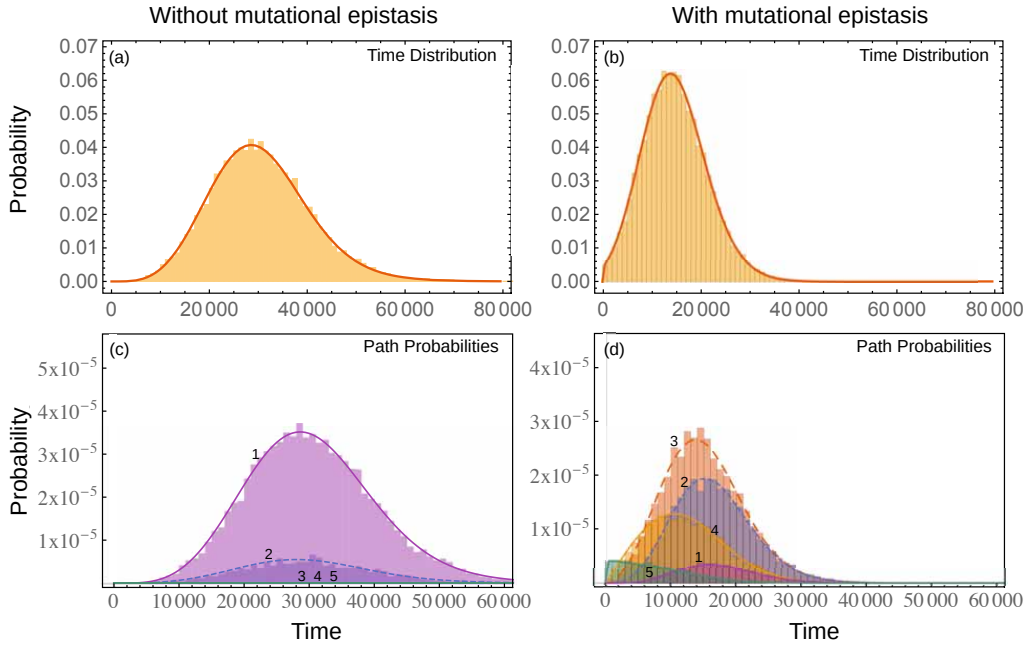

Figure 3.5: **Comparison between the path probability distributions of a minimal Burkitt Lymphoma model.** Top: Time distributions for the model (a) without epistatic effects on mutation probabilities and (b) with mutational epistasis. The probability to obtain an A mutation is 100 times higher, if the B mutation is present in that individual. Bottom: In (c) the path probabilities for the model without epistatic effects on mutations are illustrated, whereas in (d) the mutation probability is again increased by 100 for an A mutation if the B mutation is present. Pathway 1 corresponds the the mutational pathway, where first all necessary extra mutations have to be acquired, and the B mutates last. Pathway 2 denotes the pathway, where 3 of 4 extra mutations have been obtained, then the B mutation happens, and at last the final extra mutation is acquired. Respectively for the other pathways (cf. Figure 3.4). The parameters are the same as in the previous chapter: The birth probability for an individual with j passenger mutations and without the B mutation is $b_{0,j} = 0.5(1+10^{-5})^j$, and with the B mutation $b_{1,j} = \frac{1.05}{2.2} \cdot 1.015^j$. The mutation probability for the B mutation is $\mu_D = 5 \cdot 10^{-6}$, for an A mutation without the B mutation being present $\mu_P = 2 \cdot 10^{-5}$, and with the B mutation being present (only necessary for (b) and (d)) $\mu_D^P = 2 \cdot 10^{-5}$. The population size in the beginning is $N = 500000$.

mutation rates. Hereby we could show, that pathways become accessible, that without mutational epistatic effects are mostly unlikely to happen (cf. e.g., Figure 3.5). Our analysis is based on multi-type branching processes and hence it does not rely on the assumption of a constant population size.

While we have focused on a fairly simple system with a fitness landscape with a single peak, the approach can be extended to a rugged fitness landscape. Moreover, if back mutations are involved, one can still calculate the time distribution, although pathways are not clearly defined in a system with back mutations anymore (see Appendix). Furthermore in the current scenario in each time step the individuals could replicate or die. In addition we could have a resting probability where the individuals remain in the same state with a certain probability. Such complicated scenarios can be incorporated in our framework as well (Appendix). The computations can be precisely represented in analytic terms and need to be solved recursively.

We apply our framework to a cancer model including mutational epistasis [Bauer et al., 2014] and show how the path probabilities are altered by it. Mutational epistasis can thus lead to heterogeneity in the density of different mutant types between different age groups, as reaching the final mutant early is only possible by certain mutational pathways.

As shown here the mutational landscape can undermine the current predictions based solely on fitness landscapes. Just like in long term evolution, experimental as well as theoretical approaches ought to be balanced between studying effects of selection *and* the strengths of mutations. The theoretical analysis based on the approach explained here helps in understanding the importance of mutational epistasis, even though the computations have to be solved recursively. In particular, it makes analyzing the fitness and mutational landscapes more interactive, since long-lasting simulations are not necessary any more.

Chapter 4

Model for the Initiation of Burkitt Lymphoma

In Chapters 2 and 3 we have investigated abstract, general models and methods. In this chapter we approach a model particularly designed for the initiation of Burkitt Lymphoma. Of particular interest is the sequence of initiation mutations, as well as the nature of relapses that occur after successful treatment.

4.1 A Model for the Sequence of Cancer Initiating Events in Burkitt Lymphoma

Burkitt Lymphoma is a highly aggressive B-cell lymphoma. It is, with a doubling time of 24-48 hours, presumably the most rapidly dividing tumor in men [Abe et al., 1992; Burmeister et al., 2006; Matsuo et al., 1997]. In childhood it is the most frequent lymphoma, especially in equatorial Africa. Studies using conventional cytogenetics/karyotyping, array-CGH and gene expression profiling (GEP) have not shown (gross) differences between adult and pediatric Burkitt lymphoma suggesting a similar lymphoma genesis [Boerma et al., 2009; Klapper et al., 2008; Onciu et al., 2006; Salaverria et al., 2008]. However, a recent study using high-resolution SNP-arrays found a higher number of single nucleotide variants (SNVs) in adult compared to pediatric tumors [Lundin et al., 2013], and in addition the mutational landscape might reveal differences. The *MYC* translocation, characterized by a chromosomal translocation between the *MYC* protooncogene on chromosome 8 and one of the three Immunoglobulin genes, is evidently the hallmark mutation of Burkitt Lymphoma. It is likely found in every known case of this malignancy [Salaverria et al., 2014]. This mutation leads to an increased level of *MYC* protein pro-

duction [Johnson et al., 2012; Kluk et al., 2012; Tapia et al., 2011]. However, *MYC* affects a wide range of genes [Dang et al., 2006; Dave et al., 2006]. Hence, higher levels of *MYC* protein not only increases the rate of proliferation of a cell, it also increases its chance of undergoing apoptosis. Other mutations inhibiting the effect on apoptosis need to also be present in the same cell in order for the *MYC* mutation to gain a proliferative advantage. Indeed, the sole presence of the *MYC* translocation is insufficient to drive lymphoma genesis [Müller et al., 1995], suggesting an important etiologic role for additional genetic events, including mutations [Lundin et al., 2013]. Nevertheless, this seems to contradict the assumption that the *MYC* translocation is a primary event in cancer initiation, since cells with this mutation have a higher chance of dying than dividing. However, an increased level of *MYC* protein presumably leads to higher mutation rate of the cell, such that the chance of getting the necessary additional mutations rises drastically in a cell with the *MYC* mutation. The goal of our study is to infer how the prominent role of the *IG/MYC* translocation in cancer initiation is compatible with the idea that this mutation alone is insufficient to drive the cancer.

Presumably, there is a set of core mutations of which two mutations (additional to the *IG/MYC* translocation) are sufficient to drive the cancer in a patient [Drost et al., 2015; Tomasetti et al., 2015]. Our hypothesis is that the *MYC* mutation increases the probability of at least this set of core mutations. Based on these assumptions we develop a theoretical model. By means of simulations, we show that we can recapitulate a qualitatively similar modeled incidence curve compared to an empirically observed incidence curve. Afterwards, we investigate the conditions necessary in order for the *IG/MYC* translocation to be the initiating event.

4.1.1 Materials and Methods

We assume a time discrete model, where time is measured in generations. In each generation, a cell either divides, undergoes apoptosis, or nothing changes (cf. Figure 4.1). For cells with no mutation, the division and apoptosis probabilities (b_W and d_W) are both the same. The probability that the cell neither divides nor dies is consequently $1-b_W-d_W$. Upon an infection, B-cells divide in the dark zone of a germinal center with a rate of approximately six hours

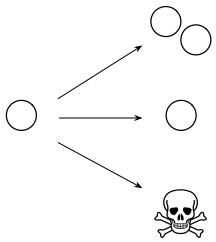

Figure 4.1: In each time step a cell either divides, dies, or there is no change.

[Liu et al., 1991; Radmacher et al., 1998]. Otherwise, B-cells divide rarely. We assume an average division rate of 2 days, corresponding to approximately 15 infection days per year, assuming that B-cells divide only every 3 days during non infection periods. During the early time steps of each realization, which correspond to the age of children, the number of B-cells grows to a higher level, stays at this level for a while, and decreases again to the baseline level (cf. Figure 4.2). The implementation is described in detail in 7.7 in the Appendix. We assumed this temporary increase to reflect the typical, exponential growth of B-cells in children by which new antigens are being produced [Kliegman, 2012].

When a cell divides, its two daughter cells can mutate. In our model, we consider the *MYC* translocation as the hallmark mutation of Burkitt Lymphoma, additional core mutations, and finally minor mutations without phenotypic effects. Of the core mutations we assume that at least two need to be present.

Minor mutations (see Table 4.1) are much less specific and, in the context of Burkitt Lymphoma, are most likely rather neutral mutations. Figure 4.4 shows the possible mutational pathways of the model. The front of the hypercube illustrates the types with (right) or without (left) the *IG/MYC* translocation and 0, 1, 2, ... core mutations and no minor mutation. The types represented in the next layer have one minor mutation and so on.

The different mutations have different effects on the cell's fitness. Core mutations are assumed to have a small direct impact on the cell's fitness. With each core mutation the probability for apoptosis is reduced by the fitness parameter $s_C \approx 0$. The probability for apoptosis thus changes from d_W for a cell with no mutations, to $d_W(1 - s_C)^{j_C}$, where j_C denotes the number of core mutations present. The minor mutations are assumed to be completely

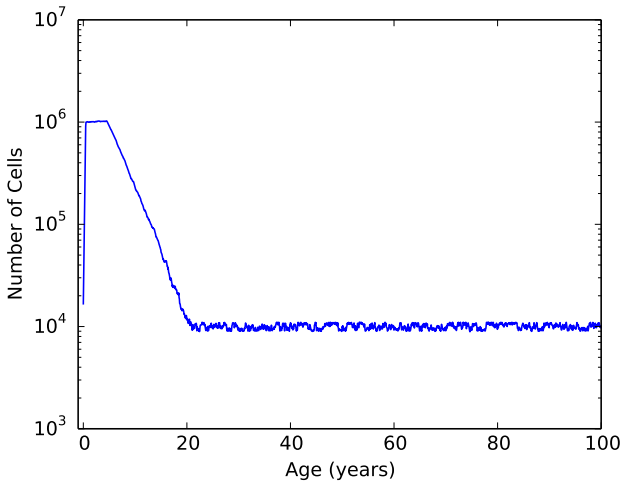

Figure 4.2: The number of B-cells rises quickly at a young age and drops again to a base line level (here 10^4).

Core mutations	Passenger mutations
CCND3	*ABCC5*
TP53	*ADAMTS5*
ID3	*CHD4*
TCF3	*E2F2*
FBXO11	*PHF6*
SMARCA4	*TBL1XR1*

Table 4.1: Exemplary list of possible core and passenger mutations.

neutral (without the *MYC* translocation being present). Consequently, the probability that nothing happens in a particular time step increases slightly, from $1 - d_W - b_W$ to $1 - d_W(1 - s_C)^{j_C} - b_W$.

We assume the *IG/MYC* translocation to have both an advantageous and a disadvantageous effect on the cell's fitness. However, as potential cancer cells, we assume cells with the *MYC* translocation to have an individual division and apoptosis rate, irrespective of potential infection time. For a full-blown Burkitt cell, i.e., a cell with the *MYC* translocation and four core mutations, to have a doubling rate between 30 and 35 hours [Woo et al., 1980]. The average number of Burkitt lymphoma cells $N(t)$ with four core mutations in our model can be computed as

$$N(t) = (1 + b_M - d_M(1 - s_{MC})^4)^t N_0, \tag{4.1}$$

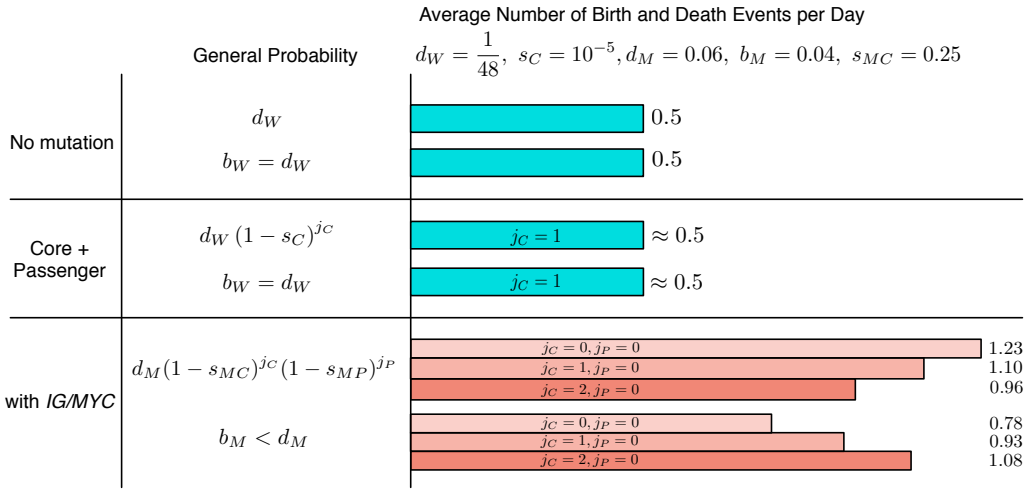

Figure 4.3: The rates for cells with different mutations. Core mutations decrease the apoptosis probability by a factor of s_C. Minor mutations alone are assumed to be neutral. The *IG/MYC* translocation increases both proliferation and apoptosis probability, where the proliferation probability in comparison with the apoptosis probability is smaller by a factor of s_M. With the *IG/MYC* translocation present, core and minor mutations affect the cell's division and apoptosis probability differently. The proliferation probability in this case is increased by factors s_{MC} for core and s_{MP} for minor mutations. The apoptosis probability is decreased accordingly.

where N_0 is the initial number of cells. The doubling time can be approximated by setting the left side to $2N_0$ and solving for t

$$t_{doubling} \approx \frac{\log(2)}{\log\left(1 + b_M - d_M(1-s_{MC})^4\right)}. \tag{4.2}$$

Further, from [Woo et al., 1980] we assume a ratio between the death and division rates of a full-blown Burkitt lymphoma cell of approximately 0.5, i.e., $\frac{d_M(1-s_{MC})^4)}{b_M} = 0.5$. Combining this condition with (4.2) and the fact that $t_{doubling}$ is between 30 and 35 hours, we obtain a death rate of $d_M = 0.06$, a proliferation rate of $b_M = 0.04$, and a fitness advantage for a core mutation of $s_{MC} = 0.25$. In the genetic background with the *MYC* translocation, passenger mutations have a small fitness advantage of $s_{MP} = 0.01$. The apoptosis and proliferation rates as well as the mutational fitness effects are listed in Figure 4.3.

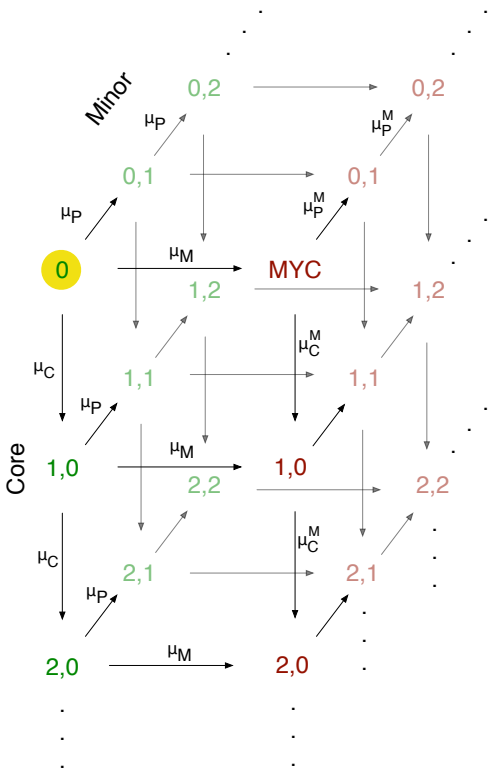

Figure 4.4: Mutational Pathways of the Model. Initially, the cells start with no mutation at all (shaded in yellow). Upon cell division, the daughter cells can acquire a core mutation (arrow pointing down) a minor mutation (arrow pointing to the back) or the *IG/MYC* translocation (arrow pointing right). The different types are indicated by a tuple X,Y, where X denotes the number of core mutation and Y the number of minor mutation. Further, a green colored tuple indicates no *IG/MYC* translocation in contrast to red colored tuples.

Mutations cannot only affect the cell's fitness, they can also influence certain mutation rates. In order for the *IG/MYC* translocation to be the first, initiating event, our simulations show that it needs to increase mutation rates, i.e., $\mu_C^M > \mu_C$ and $\mu_P^M > \mu_P$. Otherwise, it is far more likely that non-disadvantageous core and minor mutations happen first to pave the way for the advantageous effect of the *MYC* translocation, whereas a disadvantageous *IG/MYC* translocation first would die out too quickly, if mutation rates were not increased.

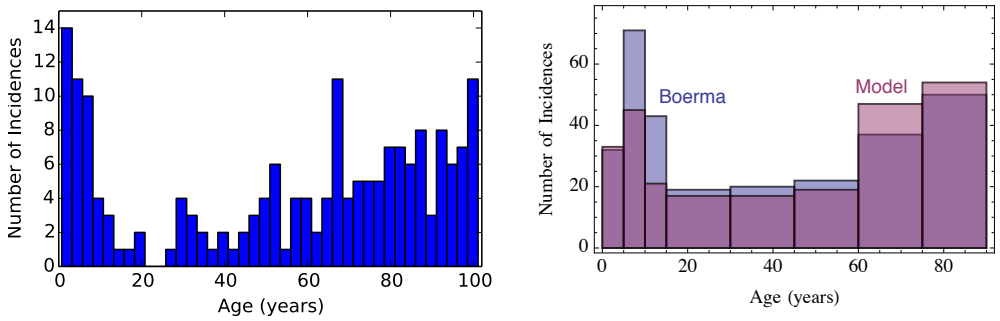

Figure 4.5: Age incidence curves from simulations (a) and in comparison to [Boerma et al., 2004] (b). The simulations recapitulate a peak for individuals at a young age, corresponding to the peak seen in the age of children. The subsequent raise in incidences occurs at a higher rate. Since we do not account for any other death possibilities in our model, this makes sense to some extend. (Parameters: Mutation probabilities: MYC translocation: $\mu_M = 2 \times 10^{-7}$, Core mutations: $\mu_C = 2 \times 10^{-6}$, Minor mutations: $\mu_P = 10^{-7}$, Core mutation with MYC present: $\mu_C^M = 5 \times 10^{-4}$, Minor mutation with MYC present: $\mu_P^M = 10^{-3}$.
Change in fitness for a cell with: Core mutation: $s_C = 10^{-5}$, Core mutation with MYC present: $s_{MC} = 0.25$, Minor mutation with MYC present: $s_{MP} = 0.01$. Baseline level for number of cells: 10^4, High level: 10^6, increase by 0.1%, decrease by 0.003%.)

4.1.2 Results and Discussion

4.1.2.1 Time Distribution

Simulations of the model described above show a time distribution of the disease incidence qualitatively similar to the age incidence curve of Burkitt Lymphoma as seen in real life [Boerma et al., 2004]. The similarity is illustrated in Figure 4.5. Due to the temporary rise in B-cells at early age, there is a small hump in the beginning and then a very subtle increase. This latter increase eventually vanishes again to a decrease in incidents, but this occurs beyond the normal life expectancy, such that what we see in the age incidence is only the subtle increase. Figure 4.5 shows the time distribution of 26000 realizations of the model with a specific set of parameters, specified in the figure caption, which lead to a disease case at a certain age.

Note, that the overall incidence is much higher than expected in real life (here 250 out of 4500). This is due to the fact, that computing enough simula-

tions of the model with quantitatively realistic parameters is computationally extremely expensive. The incidence of Burkitt Lymphoma is about 1 in 1 Million per year. Thus, we would need to do on average 1 Million simulations to get one case within the life expectancy. To be able to get a time distribution fine enough to be able to recover the first hump of the incidence in children, one would need billions of simulations. However, despite this lack in quantitative realism, our qualitative arguments still hold.

If we would assume a roughly constant population size of cells at risk, such an early peak in the incidence would not be possible. Therefore, we postulate that the increased number of B centroblasts in children [Kliegman, 2012] is driving the higher incidence in children.

4.1.2.2 Sequence of Mutational Events

Of particular interest is the order of mutations during the initiation of the disease. Not only is knowledge over the order of mutations crucial for the basic understanding of the initiation of this cancer, it also holds important information for therapeutic applications. The cell lineage that initiates the cancer originates from a cell lineage that lacks (at least) one mutation for the formation of a tumor. We call cells from this cell lineage precursor cells. These precursor cells pose a high risk for a potential relapse, in case the original cancer is being successfully treated. Therefore, knowing which mutations are commonly present in possible Burkitt precursor cells, one might be able to specifically target those cells in order to reduce the risk for a potential relapse. It is commonly assumed, that the MYC translocation is the initial mutation in the cell lineage initiating Burkitt Lymphoma. Empirical evidence is however hard to obtain. Further, since the MYC translocation itself presumably increases the apoptosis rate more than the proliferation rate, the accumulation of subsequent core mutations is only possible if the mutation rate for those core mutations is increased substantially by the MYC translocation. By further analyzing our presented theoretical model, we aim to acquire an idea on the change in mutation rate.

We have kept track of the order of mutations in 20000 realizations. To see how often the MYC translocation has happened first, or second, or even later, Figure 4.6 shows the frequencies of the initiating cancer cell lineage

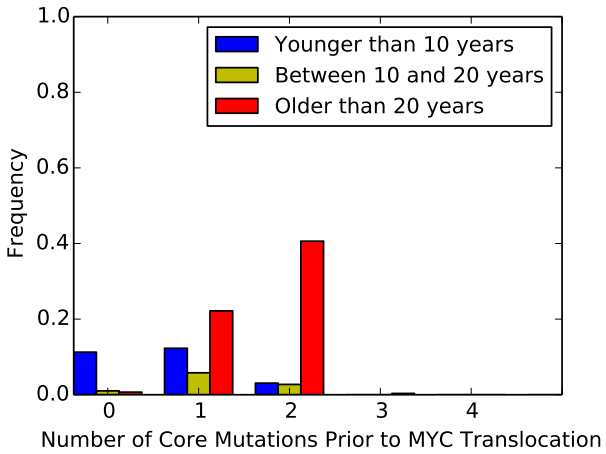

Figure 4.6: In our model for the initiation of the cancer an MYC translocation and two core mutations are necessary. Shown here are the Frequencies for the initiating cancer cell lineages to have accumulated different numbers of core mutations prior to the MYC translocation. The frequencies are distinguished between three different age groups (see legend). The parameters are the same as in Figure 4.5.

to have accumulated no, one, or up to 4 core mutations prior to the MYC translocation.

Interestingly, although the mutation probability for core mutations is increased by 250 it seems rather unlikely for the MYC translocation to be the initiating event. Further, it is important to note that the number of core mutations prior to the MYC translocation increases in probability with age. The simulations suggest that for very young patients the MYC translocation was very likely the initiating event, or only one core mutation has happened beforehand. In contrast, for older patients the probability that more than one core mutation is acquired prior to the MYC translocation increases according to our model.

Let us investigate now which parameters can be altered in order to get more cases where the MYC translocation really is the first event.

First, we can increase the mutation probability of core mutations for $MYC+$ cells even further. This leads to even more incidences in the young age group and respectively less incidences in the older age group. In Figure 4.7, left subfigure, we have used similar parameters as before, but the MYC

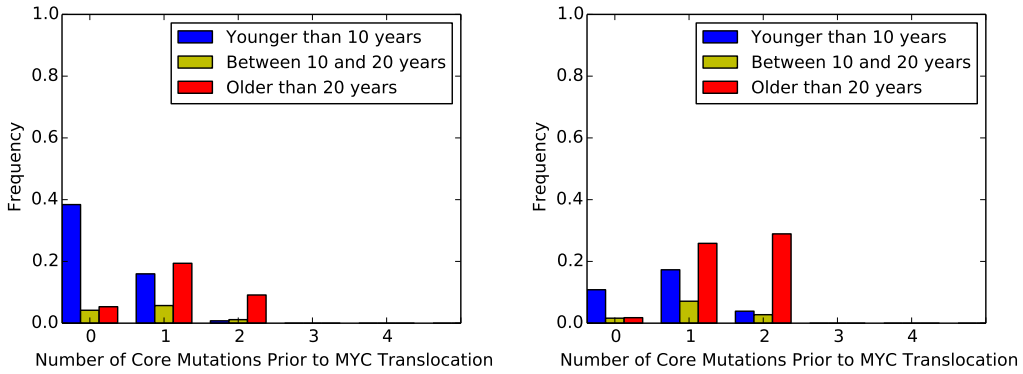

Figure 4.7: The probability for the cancer initiating cell lineage to have acquired different numbers of core mutations prior to the MYC translocation (compare Figure 4.6) with different parameters is shown here. Again the IG/MYC translocation and two core mutations are assumed to be necessary for the initiation of the cancer. Most of the parameters are the same as in Figure 4.5. In the left figure the effect of the mutation rate on the core mutation is analyzed. The increase in mutation rate that comes along with the IG/MYC translocation is set to be even greater. Instead of $\mu_C = 2 \times 10^{-6}$ for the mutation rate of core mutations and $\mu_C^M = 5 \times 10^{-4}$ for the mutation rate of core mutations in presence of the IG/MYC translocation, we have used here $\mu_C = 10^{-6}$ and $\mu_C^M = 10^{-3}$. In the right figure the mutation rate of the IG/MYC translocation is $\mu_M = 5 \times 10^{-7}$ instead of $\mu_M = 2 \times 10^{-7}$.

translocation enhances the mutation rate for core mutations by 1,000 instead of 250. We find that this increase in core mutation rate has a great effect on the number of mutation events prior to the MYC translocation. It also shifts the simulated incidence curve towards early incidences.

Second, the mutation probability μ_M affects the population dynamics. A higher mutation probability means once again a higher probability for the IG/MYC translocation to be the initiating event, but only up until a certain point. The reason is that also for cells with core mutations, the mutation probability for the IG/MYC translocation, and hence the probability for this mutational pathway, is enhanced. In Figure 4.7, right subfigure, the order of mutations is shown for similar parameters as in Figure 4.5, but the mutation rate for the MYC translocation is 5×10^{-7} here.

Further, we can change the ratio of birth and death for $MYC+$ cells, which is defined here as the cell's fitness. Increasing this ratio would result in more

division events compared to apoptosis events, and thus increasing the chance of getting (additional) core mutations. However, one has to make sure to use values in the range found by [Woo et al., 1980]. With the parameters used for b_M, d_M, and s_{MC} we cannot increase the ratio substantially.

Increasing the fitness advantage for the core mutations s_C gives a further advantage for cells with this mutation. Therefore, the chance for core mutations to have happened in the cancer cell lineage first increases. Consequently, decreasing that parameter would make it less likely for core mutations to have been acquired prior to the IG/MYC translocation. The parameter we are using is $s_C = 2 \times 10^{-6}$, which is already very small. A negative value for this parameter is biologically not reasonable.

We conclude that the fitness of a $MYC+$ cell cannot be too disadvantageous, and the mutation probability has to be increased a lot in order for the MYC translocation to be the initiation event. The observation that older patients have an increased probability for the MYC translocation not to be the initiating event in our simulations questions the hypothesis that this particular event is always the initiating one. This would have direct implications particularly for preventive screening. Patients, for whom the MYC translocation was not the initiating event, must have cell lineages with only core mutations. Those lineages grow only slowly and could in principle be detected by screening. Contrary, a cell with the MYC translocation alone (no core mutations) has to acquire further core mutations very quickly in order to survive, and then initiate the cancer (cf. Figure 4.8). This might be too fast to allow detection of Burkitt precursor cells in preventive screening.

4.2 Timing and Nature of Relapses

We assess the timing and nature of relapses of Burkitt Lymphoma in this section. The analysis presented here is based on the publication [Aukema et al., 2015]. Our model is based on a time-continuous branching process [Antal and Krapivsky, 2011; Bozic et al., 2010, 2013; Durrett et al., 2010; Haccou et al., 2005; Kimmel and Axelrod, 2002]. We assume two different types of cells, MYC+ precursor cells, which are not affected by therapy, and Burkitt lymphoma cells, which survived cancer therapy. Both types of cells

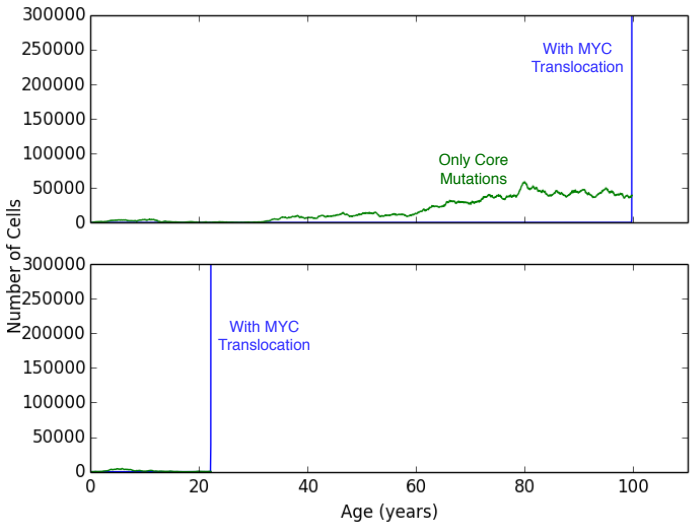

Figure 4.8: Number of cells with the *MYC* translocation (blue) and only core and minor mutations (green). Top: Cancer is developed by a cell lineage with two core mutations prior to the *MYC* translocation. Cells with a core mutation are present in the organism for a long time before explosive proliferation happens. Bottom: Cancer happens early and starting from a lineage with the *IG/MYC* translocation first. There is a low number of cells with core mutations (due to exponential growth after birth). Otherwise, there is no increased level in the number of cells with core mutations.

can lead to a relapse, but precursor cells need an additional mutation. For each of the two types, we assume different proliferation and apoptosis rates. As the presumed cell of origin for Burkitt lymphoma is a germinal center dark zone cell [Basso and Dalla-Favera, 2015; Klein and Dalla-Favera, 2008; Victora et al., 2012] and cell cycle rates of 6 – 12 hours have been reported for centroblasts [Klein and Dalla-Favera, 2008; Meyer-Hermann et al., 2012; Radmacher et al., 1998; Victora and Nussenzweig, 2012] we assume that the MYC+-precursor cells divide at most every 6 hours, i.e., 4 times per day. In contrast to the t(14;18)/IGH-BCL2, (virtually) no benign neoplasms or lymph nodes with the t(8;14)/MYC-IGH have been described [Kluin, 2014; Limpens et al., 1991; Mamessier et al., 2014; Nagy et al., 2009; Tellier et al., 2014]. Moreover, precursor cells do not lead to a tumor. Therefore, the apoptosis rate for precursor cells has to be (at least) equal to the division rate, i.e. 4 per day.

The division and apoptosis rates for full-blown Burkitt Lymphoma cells are

obtained using the empirically deduced doubling time of Burkitt Lymphoma and fraction between proliferation and apoptosis of Burkitt Lymphoma cells. For a time-continuous one-type branching process, the average number of cells $N(t)$ grows over time as $N(t) = N_0 e^{(b-d)t}$, where b and d denote the proliferation and death rates, and N_0 states the initial number of cells. To acquire the difference of proliferation and apoptosis rates $b - d$ for a known doubling time, we set the left hand side equal to $2N_0$ and solve for $b - d$

$$2N_0 = N_0 e^{(b-d)t_d} \Leftrightarrow b - d = \ln(2)/t_d, \tag{4.3}$$

where t_d, denotes the time it takes for a population to double its size. For our model we have considered three different doubling times, 24 hours, $\approx$ 33.3 hours, and 48 hours [Woo et al., 1980]. This leads to differences of approximately $b - d = 0.7$, 0.5, and 0.35. Additionally, from Woo et al., [Woo et al., 1980] we have used three different values for the fractions $d/b = 0.25$, 0.5, and 0.667. The ratio and difference of the proliferation and apoptosis rate allows us to determine both the division rate and the apoptosis rate. Initially, there is presumably enough space and nutrition for all cells, which allows us to assume constant proliferation and apoptosis rates. If a precursor cell divides, the daughter cells can mutate, leading to a new lineage of full-blown Burkitt lymphoma cells. The various possible events are illustrated in Figure 4.9.

The value for the mutation rate is varied across simulations. As the upper limit we consider a very high mutation rate, 0.1. As the lower limit, we use the biological estimate of 10^{-10} [Tomasetti et al., 2013].

Finally, we need to assume an initial number of cells. Before therapy, there are more Burkitt Lymphoma cells than precursor cells. Chemotherapy is most harmful for fast proliferating cells. Since the Burkitt lymphoma cells divide on average more often than precursor cells (unless they are in a centroblast stage), we assume that the Burkitt lymphoma cells are targeted more effectively than the precursor cells. As a conservative estimate we use an initial number of cells for both populations of 10. For illustration purposes, Figure 4.10 shows a single realization of a simulation of this process. The simulation in continuous-time is based on a Gillespie Algorithm [Gillespie, 1977]. The remaining lymphoma cell quickly produces a lineage that grows exponentially fast. After less than one month, the population of Burkitt lymphoma cells

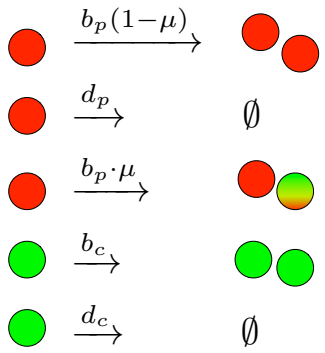

Figure 4.9: The rates of division of precursor cells (red) and Burkitt lymphoma cells (green) are depicted here. A precursor cell divides with rate b_p and undergoes apoptosis with rate $d_p = b_p$. If it divides, one of the daughter cells mutates with probability μ and becomes a full-blown lymphoma cell. This Burkitt lymphoma cell originates from a new lineage and is therefore illustrated in a different coloring. A Burkitt lymphoma cell divides with rate b_c, and it undergoes apoptosis with rate d_c, where $d_c < b_c$.

already reaches about 1 Million cells.

Figure 4.11 shows the estimated probability for a relapse to occur via a Burkitt lymphoma cell that has survived therapy compared to a relapse via a mutated precursor cell. We notice that the mutation rate has to be enormously high (far exceeding those reported in literature [Tomasetti et al., 2013]) in order for a relapse to happen via a mutated precursor cell. Overall the probability for a relapse to occur via a left-over Burkitt lymphoma cell is greater than about 80%. Even for a rather slow doubling time of 48 hours, a high cell loss factor of 0.667 (red line in (c)), and an enormously high mutation rate of 0.1 the probability for a Burkitt lymphoma relapse is still approximately 75%. Already for a mutation rate of 10^{-5} the probability for a relapse via a precursor cell is virtually 0%. Therefore, testing for even lower mutation rates is not necessary here.

To stress the robustness of the probability of a relapse via a left-over Burkitt lymphoma cell, we have run additional simulations with parameters that reflect an even greater advantage for precursor cells compared to the already conservative set of parameters described above. In particular we have used a proliferation and apoptosis rate of 2 hours for precursor cells and have assumed 100 precursor cells initially. The initial number of left-over Burkitt lymphoma cells has been kept at 10. Figure 4.11 (d)-(f) show the resulting

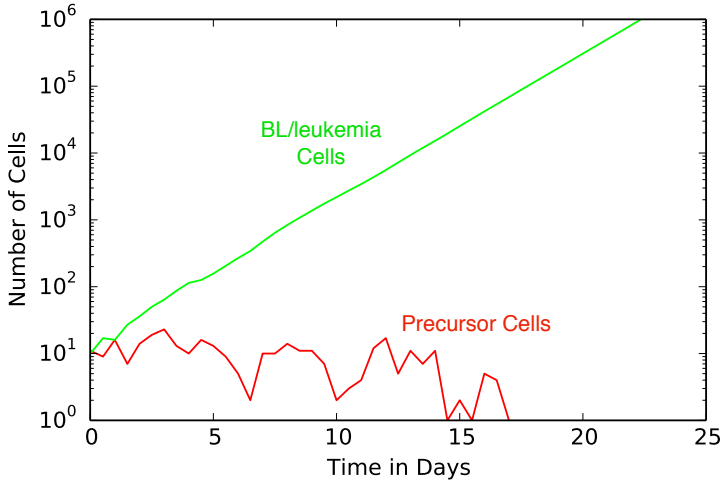

Figure 4.10: A single realization of the model. The population of precursor cells fluctuates around the initial number of cells until it goes extinct. The Burkitt lymphoma cell population grows exponentially. After less than one month, the number of cells reaches 10^6.

estimated probabilities for a relapse via a left-over Burkitt lymphoma cell. These simulations show that a relapse via precursor cells remains extremely unlikely for biologically meaningful mutation rates.

Additionally, in Figure 4.12 the time distributions of relapses after the termination of therapy are shown for two parameter sets. We here define a relapse in our model when 10^8 Burkitt lymphoma cells are present again (a number of Burkitt lymphoma cells which could, at least by sensitive breakpoint-specific PCR, be detected in the bone marrow of, as reference, a seven year old boy of approximately 25 kg) [Bianconi et al., 2013; Boerma et al., 2004; Busch et al., 2004; Harrison, 1962; Kuczmarski et al., 2002].

Figure 4.12 shows, that according to our model relapses from left-over Burkitt lymphoma cells need a little over a month to occur with the parameters stated in the caption. Also relapses from a precursor cell seem to occur very quickly. However, the mutation rate of 10^{-3} is chosen to be very high. With a lower mutation rate we would see even less relapses occurring via a precursor cell, because relapses via a left-over Burkitt lymphoma cell are much faster. To check the timing of relapses occurring via a mutated precursor cell, we

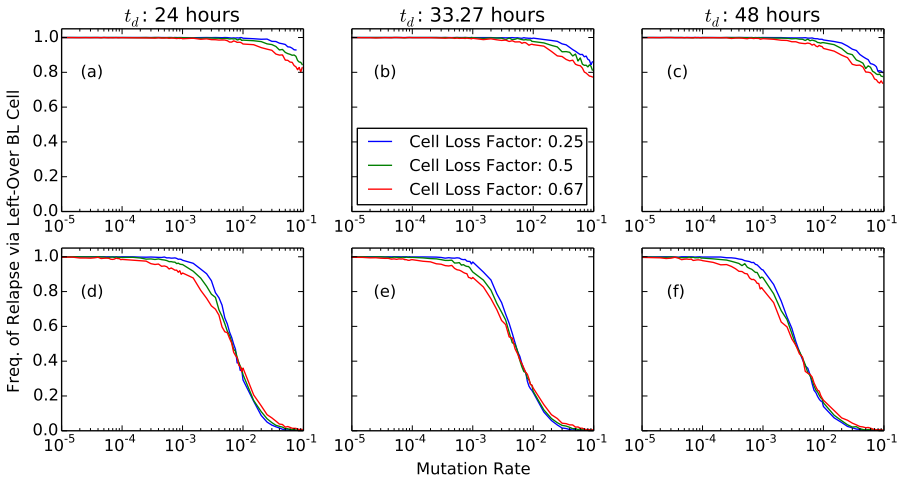

Figure 4.11: The estimated probability that a relapse happens via a Burkitt lymphoma cell that has survived therapy and not via precursor cells.
In (a)-(c) the initial numbers of cells for precursor and left-over Burkitt lymphoma cells are 10. For a relapse to happen via a precursor cell, the mutation rate needs to be very high. Even for a mutation rate of 0.1 the probability is approximately 80% for a relapse to happen via a left-over Burkitt lymphoma cell. With a high doubling time t_d and a high cell loss factor (d/b) (red line in (c)) that probability drops to approximately 75%.
In (d)-(f) the initial number of cells for precursor cells is increased to 100. Further, the rate for proliferation and apoptosis is increased to 2 hours. Also for these extreme parameters for Burkitt lymphoma cell, our model suggests a probability of a relapse via a Burkitt lymphoma cell of 100% for a mutation rate of 10^{-5} or slower.

therefore set the number of left-over BL/Burkitt lymphoma cells to one and use a more realistic mutation rate. Setting the number of left-over Burkitt lymphoma cells to one leads to a higher of probability for this lineage to die out by stochastic effects, which leaves time for the precursor lineage to produce a relapse. The timing for these parameters is shown in Figure 4.13.

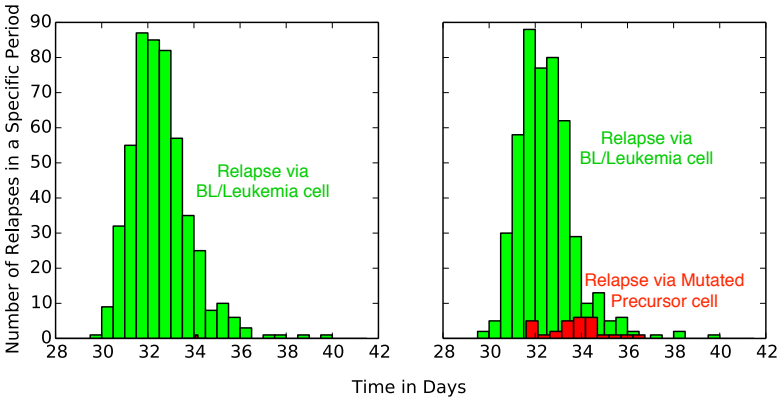

Figure 4.12: The distributions of relapse times after therapy. In both histograms we have a doubling time of 33.27 (i.e., $b_c - d_c = 0.5$) hours and a cell loss factor d_c/b_c of 0.5. The mutation rate is 10^{-3}.
Left: The precursor cells divide and die with rate 6 hours and initially there are 10 left-over Burkitt lymphoma cells and 10 precursor cells. There are no relapses happening via a mutated precursor cell.
Right: The precursor cells proliferate and die with rate 2 hours and there are 100 precursor cells present initially. There are some relapses occurring via a newly founded Burkitt lymphoma lineage (red bars). Relapses via mutated precursor cells tend to take slightly more time on average to happen compared relapses via a left-over Burkitt lymphoma cell.

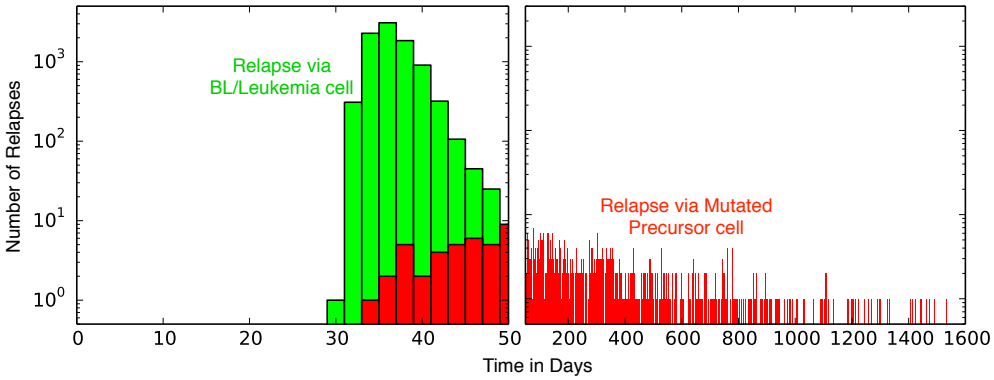

Figure 4.13: The distributions of relapse times after therapy for a different set of parameters. As in Figure 4.12, a doubling time of 33.27 (i.e., $b_c - d_c = 0.5$) hours and a cell loss factor d_c/b_c of 0.5 are assumed. The initial number of cells for left over Burkitt lymphoma cells is only one and for precursor cells it is 1000. The mutation rate is set to 10^{-7}. Precursor cells divide and die every 6 hours. While the relapses occurring from a left over Burkitt lymphoma cell need on average approximately 38 days, the relapses occurring via a mutated precursor cell need much longer, approximately 200 days on average. A slower turnover rate and a smaller mutation rate would lead to an even later timing of the relapses via precursor cell.

Chapter 5

Further Research

5.1 Branching Process with Frequency Dependent Fitness

As already mentioned in Section 5.1, the independence of individuals in a Branching Process is on the one hand convenient for the calculation of extinction probabilities. On the other hand, however, it reduces the applicability. Even though for cancer initiation the assumption of independence is largely valid, taking into account population size dependent interactions is important for certain investigations. One very prominent example is immunosurveillance. At some point of cancer progression the cancer cells are usually being recognized by immune cells as being harmful. At the same time, the cancer cells try to escape immunosurveillance and acquire further mutations [Schreiber et al., 2011]. This interplay between immune cells and cancer cells leads to interactions between cells, such that proliferation and apoptosis of cancer cells cannot be assumed independent anymore. For this reason, a short overview on how to introduce such dependencies in a Branching Process and how one could "in principle" calculate extinction probabilities is given in the following.

For the sake of simplicity we first start by examining the principles in a one-type process. We are using a time-continuous process here [Haccou et al., 2005]. The individuals live for a while and then produce a certain number of offspring (zero offspring is also valid) and die. The time they live is exponentially distributed, where the parameter depends on the birth and death rates. Let b_m and d_m be the birth and death rates in a population of size m. The process is completely characterized by the probabilities of having m individuals at a certain point of time $p_m(t)$. The change of those probabilities obeys the Kolmogorov forward or backward equations. For density independent processes, it is easier to calculate the probability generating function of

the process using the Kolmogorov backward equations. However, since we are ultimately interested in density dependent processes, we need to use Kolmogorov forward equations. Let us therefore derive the Kolmogorov forward equations for the one-type process we are discussing here. The probability $p_m(t)$ changes forward in time with the birth and death rates, resulting in $m+1$ and $m-1$ individuals respectively

$$\dot{p}_m(t) = (m-1)b_{m-1}p_{m-1}(t) + (m+1)d_{m+1}p_{m+1}(t) - m(b_m + d_m)p_m(t). \tag{5.1}$$

The probability generating function (over time) is defined as

$$f(z,t) = \sum_{m=0}^{\infty} p_m(t) z^m. \tag{5.2}$$

Hence, the derivate with respect to t of the PGF is calculated as

$$\dot{f}(z,t) = \sum_{m=0}^{\infty} \dot{p}_m(t) z^m. \tag{5.3}$$

Inserting Equation 5.1 and rearranging the terms leads to the following

$$\dot{f}(z,t) = \sum_{m}^{\infty} m p_m z^{m-1} (z-1)\left(z b_m - d_m\right). \tag{5.4}$$

If the birth and death rates would be density independent ($b_m = b$, $d_m = d$), we can use the fact that $m z^{m-1} = \frac{\partial}{\partial z} z^m$ and thus obtain

$$\dot{f}(z,t) = (z-1)\left(zb - d\right) \frac{\partial}{\partial z} \sum_{m}^{\infty} p_m z^m = (s-1)\left(zb - d\right) \partial_z f(z,t). \tag{5.5}$$

The solution of this partial differential equation (PDE) is [Antal and Krapivsky, 2011; Athreya and Ney, 1972]

$$f(z,t) = 1 - \frac{b-d}{b\left(1 - \left(1 - \frac{b-d}{b(1-z)}\right)\exp\left(-(b-d)t\right)\right)}. \tag{5.6}$$

For density dependent processes, the resulting PDE is only seldom analytically

solvable. Let us for now assume that both the birth and death rate for one individual depend linearly on the number of other individuals $b_m = (m-1)b$, $d_m = (m-1)d$. The resulting PDE for the PGF is

$$\dot{f}(z,t) = \sum_{m}^{\infty} m(m-1)p_m z^{m-2} z(z-1)\,(zb-d) = z(z-1)\,(zb-d)\,\partial_{zz} f(z,t), \tag{5.7}$$

which resembles a generalized heat equation.

The solution of (5.7) allows us to analytically calculate the extinction probability and numerically also the extinction time. The total extinction probability would be given by $p_{extinction} = \lim_{t\to\infty} f(0,t)$. The probability for extinction until a certain time point τ is $f(0,\tau)$. Extending 5.7 to multi-type processes would make it possible to include interactions between cell types. The integration of competition between cells would make it potentially possible to not only analyze the population dynamics during cancer initiation, but also at a later stage.

5.2 Epistasis in Spatially Structured Populations

Generalizing the analysis of epistatic interaction with respect to density dependent populations is a very interesting subject for future research. As stated above, integrating competition between cells allows to analyze the population dynamics of cancer also at a later stage. Cancer cells are not always well-mixed. In a solid tumor, the cancer cells might be structurally organized. Another exciting research direction is therefore the investigation of the effect of population structure. The idea is to place individuals on the nodes of a graph and let them interact only with their neighbors. The neighbor is represented by the links between nodes. In particular, we are using the *Moran Process* on a graph. In the following section, the Moran Process is described in general. Afterwards an algorithm is presented, with which the average time of fixation for a certain type in a population with fixed size, can be computed. One is also able to compute the average path probabilities by modifying this

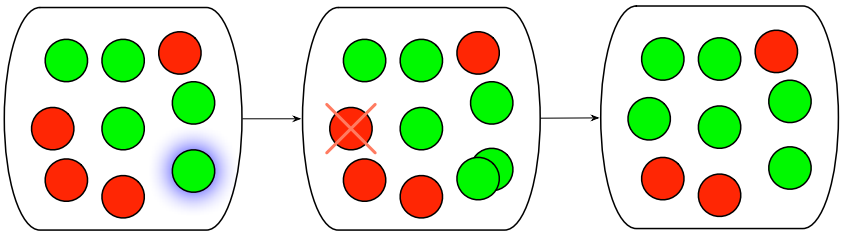

Figure 5.1: Illustration of the Moran Process. First one individual is chosen according to its fitness for reproduction (highlighted by a blue shadow). Afterwards an individual is chosen for death, which is replaced by the offspring the reproducing individual.

algorithm. In Chapter 2 we have seen that the secondary driver mutations have a much higher probability to happen first, because the driver mutation is deleterious. Komarova et al. [2014] have shown that in a structured population deleterious intermediate types reach the beneficial type faster than advantageous intermediate ones. We therefore envision, that the population dynamics and the path probabilities look differently in a structured population compared to the well-mixed case as in 2.

Moran Process

Let us start with a brief overview of the time-discrete Moran Process. The Moran Process involves a population of constant size N, where the individuals can be of different types. For a type k individual the interaction with other individuals in the population is described by its payoff function π_k. In a well-mixed system all individuals of a specific type have the same payoff, because all individuals interact with each other. If the individuals are placed on a network, they interact only with its neighbors. In each time step one individual is chosen for reproduction proportional to its fitness, where the fitness is a function $f(\beta\pi_k)$ of the payoff and the selection intensity β. Afterwards another individual is chosen randomly for death. Thus, the offspring of the first chosen individual replaces the second chosen individual. Figure 5.1 illustrates the procedure for one time step. This is the basic set up, upon which different properties have been analyzed in literature. Fitness can for instance depend on the frequencies of the different types. Further, one could introduce mutations between a fixed number of types, or give even the possibility to

create new types with a stochastic fitness, cf. [Huang et al., 2012].

Let us first consider a system with only two types. We call the wild type individuals A, and the mutant type individuals B. Let i denote the number of B individuals, the number of A individuals is hence $N-i$. The interaction between A and B individuals is often described by a *payoff matrix*, where the four single payoffs for the interactions for an A with another A or B individual, and respectively for B individuals. With the payoff matrix

	A	B
A	a	b
B	c	d

an A individual would receive a payoff of a from every other A individual and b from every B individual. The total payoff is then $\pi_A^i = (N-i-1)\cdot a + i\cdot b$. Analogously we can calculate the total payoff for a B individual $\pi_B^i = (N-i)\cdot c + (i-1)\cdot d$. Since in each time step only one individual reproduces and one dies, the number of mutants can increase or decrease only by one in each time step. The probability for an increase (T_i^+) and decrease (T_i^-) is calculated as

$$\begin{aligned} T_i^+ &= \frac{if(\beta\pi_B)}{N\langle f\rangle}\frac{N-i}{N-1}, \\ T_i^- &= \frac{(N-i)f(\beta\pi_A)}{N\langle f\rangle}\frac{i}{N-1}, \end{aligned} \tag{5.8}$$

where $\langle f\rangle = if(\beta\pi_B^i) + (N-i)f(\beta\pi_A^i)$ is the average fitness. The probability that the number of mutants does not change is consequently $T_i^0 = 1-T_i^+-T_i^-$. We impose that the derivative of $f(\beta\pi_B^i)$ be always greater than zero, such that the fitness increases with payoff.

Here, we are interested in the fixation probability of a single mutant (B) in an otherwise homogenous population of wild type individuals (A). We will analyze the fixation probabilities for the different states of the system, $\rho_i, i \in \{0,\dots,N\}$. These fixation probabilities can be derived by recursive calculations of the so called *Master Equation*. In the Master Equation we represent the fixation probability in terms of transition probabilities and the corresponding fixation probabilities for the resulting states of the system

$$\rho_i = T_i^+\rho_{i+1} + T_i^-\rho_{i-1} + (1-T_i^+-T_i^-)\rho_i, \tag{5.9}$$

where the transition probabilities T_i^+ and T_i^- are defined in Equation (5.8) and $\rho_0 = 0$, $\rho_N = 1$. Let us collect all transition probabilities in a matrix of dimension $\mathbb{R}^{(N+1)\times(N+1)}$

$$\mathbf{T} = \begin{pmatrix} T_0^0 & T_0^+ & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ T_1^- & T_1^0 & T_1^+ & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & T_2^- & T_2^0 & T_2^+ & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & T_{N-1}^- & T_{N-1}^0 & T_{N-1}^+ \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & T_N^- & T_N^0 \end{pmatrix}. \tag{5.10}$$

The master equations can now be written in terms of an Eigenvector problem

$$\mathbf{T}\underline{\rho} = \underline{\rho}, \tag{5.11}$$

where $\underline{\rho} = (\rho_0, \rho_1, \ldots, \rho_N)^T$.

We now transform this Eigenvector problem into a linear system of equations. The states 0 and N are absorbing states and thus $T_0^0 = T_N^0 = 1$ and $T_0^+ = T_N^- = 0$ irrespective of fitness and selection values. Let us formulate the transition matrix as a block matrix, distinguishing between absorbing and transient states

$$\mathbf{T} = \left(\begin{array}{c|c|c} 1 & \underline{0} & 0 \\ \hline \underline{t_1} & \mathcal{T} & \underline{t_2} \\ \hline 0 & \underline{0} & 1 \end{array}\right), \tag{5.12}$$

where $\underline{t_1} = (T_1^-, 0, \ldots, 0)^T$, $\underline{t_2} = (0, \ldots, 0, T_{N-1}^+)$, and $\mathcal{T}$ represent the transition probabilities between transient states. Further, $\rho_0 = 0$ and $\rho_N = 1$ are also independent of fitness and selection values. Due to the fact that $\rho_0 = 0$, the transition value into this state T_1^- does not hold information which is directly necessary for the computation of the fixation probability, because it enters in the Master Equation as $T_1^-\rho_0 = 0$. Using this block notation, we can

write Equation (5.11) as

$$\begin{aligned} 1\rho_0 + \underline{0}^T \underline{\varrho} + 0\rho_N &= \rho_0 = 0, \\ 0\rho_0 + \underline{0}^T \underline{\varrho} + 1\rho_N &= \rho_N = 1, \\ \mathcal{T}\underline{\varrho} + \underline{t_2} &= \underline{\varrho}, \end{aligned} \tag{5.13}$$

where $\underline{\varrho} = (\rho_1, \ldots, \rho_{N-1})^T$.

As discussed above, the first two rows are always true irrespective of fitness and selection values. Hence, we only need to consider the last row. By bringing $\underline{\varrho}$ onto the left side and subtracting both sides by $\underline{t_2}$, we obtain the system of linear equations

$$\mathcal{T}\underline{\varrho} = \left(0, \ldots, 0, -T^+_{N-1}\right)^T. \tag{5.14}$$

The fixation probabilities can now be computed as the solution of this system of linear equations.

For this simple system, the fixation probabilities have an analytically closed form solution [Karlin and Taylor, 1975; Nowak, 2006; Traulsen and Hauert, 2009]

$$\rho_i = \frac{1}{\sum_{k=0}^{N-1} \prod_{i=1}^{k} \frac{T_i^-}{T_i^+}}. \tag{5.15}$$

For a structured population, however, a closed form solution is not know. Approximative solutions have been obtained for certain network structure, cf. eg., [Hindersin and Traulsen, 2014; Kaveh et al., 2015]. The approach using the master equations and rewriting them into the system of linear equations as in Equation 5.14 is always possible. Note, that in a structured population the number of states is much larger, because not only the number of B individuals matters, but also the position on the structure. While the ultimate transition matrix $\mathcal{T}$ in Equation 5.14 is of size $(N-1) \times (N-1)$, the matrix for the same system but on a network would scale with $(2^N - 2) \times (2^N - 2)$. To analyze epistatic interactions we need at least a system with four different types, wild type, final type and two intermediate types, where the fitness difference for the final type compared to the wild type is not just the sum of

the fitness differences of the two intermediate types (cf. Figure 3.1). This leads to a transition matrix $\mathcal{T}_{epistasis}$ that scales with $(4^N - 2) \times (4^N - 2)$. Note, that the transition probabilities in such a system consists of two parts now, because mutations are introduced. One part is similar to the transition probabilities used beforehand in case of no mutation. Additionally, the offspring of the proliferating individual can mutate, which adds an additional term to the transition probabilities.

Most entries of the ultimate transition matrices are zero, which is very convenient for memory purposes. And for more than two types, some theoretical states are not possible, see for example the state as depicted in Figure 5.2 Nevertheless, the size of the matrix increases exponentially. As a starting

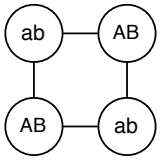

Figure 5.2: In a regular grid with four individuals and four types, this state is not accessible. Since we neglect double mutations, one of the neighbors of an AB individuals would have to have already one mutation.

point, we have therefore looked at a minimal grid of size 2x2 with four types, similar to the system depicted in Figure 3.1, where initially all cells are of type *ab*.

Comparing the fixation times between the well-mixed and spatially structured populations for different mutation rates, we find the interesting effect that the grid can accelerate or slow down the time of fixation, depending on the mutation rates. In Figure 5.3 the ratio between the time necessary for fixation of the final AB mutant with a population on a grid (t_{grid}) and a well-mixed one ($t_{well-mixed}$) is shown. The parameters are stated in the caption.

Another possibility to compute the fixation times would be averaging over simulations. However, for the parameters we are using here, the differences in the mean fixation times between the distinct populations (well-mixed and on a grid) are very close. Additionally, the variance in fixation time is high. To be able to correctly classify differences in mean fixation times, a very large number of realizations is necessary.

Pursuing research in this field will allow us to further understand the role

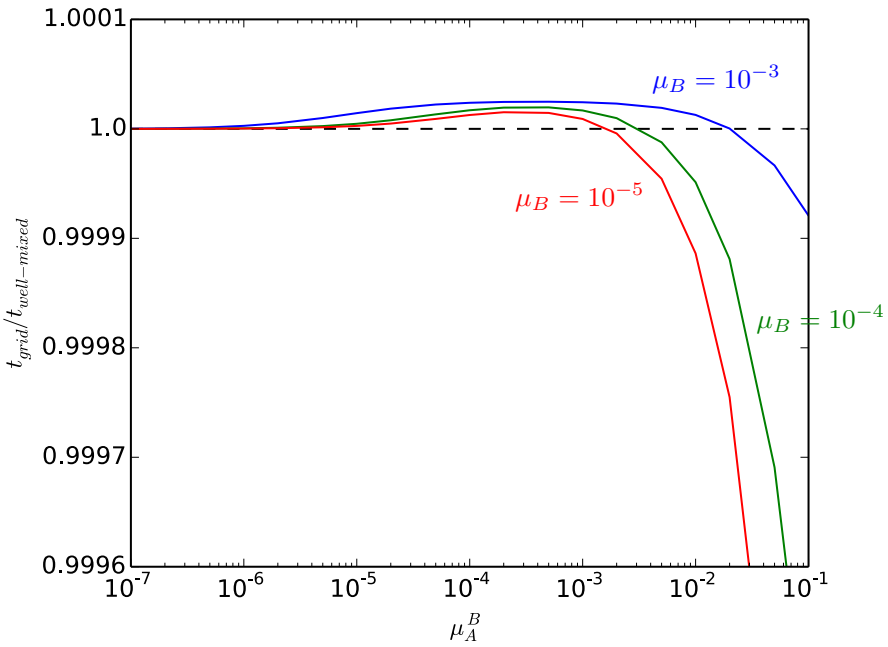

Figure 5.3: Ratio between the time necessary for fixation of the final AB mutant with a population on a grid (t_{grid}) and a well-mixed one ($t_{well-mixed}$). Interestingly, the grid structure accelerates or slows down fixation depending on the mutation rates. The mutation rate for the A mutation is $\mu_A = 10^{-4}$. The mutation rates for the B mutation (irrespective of the mutational background, i.e., $\mu_B = \mu_B^A$) are stated in the figure. The fitness for ab type individuals is 1, for aB individuals it is 1.00004, Ab individuals have a fitness of 0.9545, and AB individuals 1.0614.

of spatial structures for epistatic systems in cancer initiation.

Chapter 6

Summary

Throughout this thesis a framework has been presented to theoretically and mathematically study epistatic effects in cancer initiation.

In Chapter 2 it has been shown that the common distinction between driver and passenger mutations is often not enough. In some cancers, apparent neutral mutations have a fitness effect in a different genetic background. Hence, especially in systems in which epistatic interactions play a key role (cf. Chapter 2, 4) one might need to introduce terms such as secondary driver mutations. We have seen that in epistatic systems the order of mutations is not straight forward to reconstruct, even more so when there are also epistatic effects regarding mutation rates.

In Chapter 3 an algorithm for the likelihood of different mutational pathways has therefore been derived. Even though a closed form solution cannot be acquired, the procedure introduced makes exploring the parameter space much more interactive. Long lasting simulations are not necessary anymore to compute the probability density over time for the different mutational pathways. Moreover, our approach presented here does not use the usual assumption of strong selection and weak mutation. Rather any parameter set can be used. This helps investigating and understanding systems with epistatic fitness and mutation landscapes.

In the following chapter, we have looked at a realistic model for the initiation of Burkitt Lymphoma. The model used there exceeds the necessary conditions for the previously developed algorithm. The qualitative dynamics in dependence of mutation rates and fitness values has therefore been analyzed by means of simulations. We have investigated the sequence in which the different mutations occur. Interestingly, in order for the *IG/MYC* translocation to be the initiating event, our model suggests that the mutation rate for additional (core) mutations needs to be increased by the *IG/MYC* translocation greatly.

Further in that chapter, a model describing the formation of a relapse after therapy in Burkitt Lymphoma is analyzed. The cell lineage that originates the relapse can either be the original Burkitt Lymphoma lineage, which has not been completely eradicated by the therapy, or a precursor lineage. Cells from that precursor lineage lack at least one mutation. It has been shown that a relapse most likely originates from a Burkitt lymphoma cell, which survived therapy, unless the mutation rate for the lacking mutation for the precursor cell lineage is enormously high.

In January 1971, Richard Nixon has declared the "War on Cancer" and has increased the efforts to find a cure for cancer. More than four decades later, the war is still far from being won. Theoretical biology provides a promising tool and new hope to advance in this war. The theoretical models worked out in this thesis present a small part in this huge venture and help understanding fundamental questions in cancer initiation.

Chapter 7

Appendix

7.1 Analytic Expression for the Average Number of Cells without the Primary Driver Mutation at Generation t

7.1.1 Secondary Driver Fitness Advantage is unequal to Zero - k Secondary Driver Mutations

We first consider the case without the primary driver mutation. We assume $s_{\mathrm{P}} \neq 0$ (and consequently $(1 + s_{\mathrm{P}}) = \varsigma_{\mathrm{P}} \neq 1$), as discussed in the main text. The rate change of cells with k secondary driver mutations is

$$x_{0,k}(t) = \nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k} x_{0,k}(t-1) + \mu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k-1} x_{0,k-1}(t-1), \tag{7.1}$$

where $x_{0,-1}(t) \equiv 0$. The solution of (7.1) is formulated in the following theorem:

Theorem 1 *For any integer $k \geq 0$, the number of cells with k secondary driver mutations and no primary driver mutation is*

$$x_{0,k}(t) = N\mu_{\mathrm{P}}^{k}\nu_{\mathrm{P}}^{t-k}\varsigma_{\mathrm{P}}^{k(k-1)/2}\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n}}{1-\varsigma_{\mathrm{P}}^{n+1}}. \tag{7.2}$$

Proof 1 *Since solutions for recursive functions are unique, (7.2) would be the only solution if it fulfills (7.1). Hence, we proof (7.2) by inserting the equation on the right hand side of (7.1).*

$$
\begin{aligned}
&\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k}x_{0,k}(t-1)+\mu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k-1}x_{0,k-1}(t-1)\\
&\quad=\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k}N\mu_{\mathrm{P}}^{k}\nu_{\mathrm{P}}^{t-k-1}\varsigma_{\mathrm{P}}^{k(k-1)/2}\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n-1}}{1-\varsigma_{\mathrm{P}}^{n+1}}\\
&\quad+\mu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k-1}N\mu_{\mathrm{P}}^{k-1}\nu_{\mathrm{P}}^{t-k}\varsigma_{\mathrm{P}}^{(k-1)(k-2)/2}\prod_{n=0}^{k-2}\frac{1-\varsigma_{\mathrm{P}}^{t-n-1}}{1-\varsigma_{\mathrm{P}}^{n+1}}\\
&\quad=N\mu_{\mathrm{P}}^{k}\nu_{\mathrm{P}}^{t-k}\varsigma_{\mathrm{P}}^{k(k-1)/2}\left(\varsigma_{\mathrm{P}}^{k}\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n-1}}{1-\varsigma_{\mathrm{P}}^{n+1}}+\prod_{n=0}^{k-2}\frac{1-\varsigma_{\mathrm{P}}^{t-n-1}}{1-\varsigma_{\mathrm{P}}^{n+1}}\right).
\end{aligned}
\tag{7.3}
$$

We can write each of the two products as a q-binomial coefficient, $\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n}}{1-\varsigma_{\mathrm{P}}^{n+1}}=\begin{bmatrix}t\\k\end{bmatrix}_{\varsigma_{\mathrm{P}}}$. *Thus, with the q-Pascal rule [Kac and Cheung, 2002]*

$$
\varsigma_{\mathrm{P}}^{k}\begin{bmatrix}t-1\\k\end{bmatrix}_{\varsigma_{\mathrm{P}}}+\begin{bmatrix}t-1\\k-1\end{bmatrix}_{\varsigma_{\mathrm{P}}}=\begin{bmatrix}t\\k\end{bmatrix}_{\varsigma_{\mathrm{P}}}
\tag{7.4}
$$

Equation (7.3) *simplifies to*

$$
\begin{aligned}
&\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k}x_{0,k}(t-1)+\mu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{k-1}x_{0,k-1}(t-1)=N\mu_{\mathrm{P}}^{k}\nu_{\mathrm{P}}^{t-k}\varsigma_{\mathrm{P}}^{k(k-1)/2}\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n}}{1-\varsigma_{\mathrm{P}}^{n+1}}\\
&=x_{0,k}(t),
\end{aligned}
\tag{7.5}
$$

which concludes the proof.

7.2 Analytic Expression for the Average Number of Cells with the Primary Driver Mutation at Generation t

We now turn to the cells which have obtained the primary driver mutation. As discussed in the main text, we only look at the case where the fitness change of the secondary driver mutation is not equal to zero, $s_{\mathrm{P}}\neq 0$. Cells without the primary driver mutation can only arise through cells that lack one secondary driver mutation. Hence, there is only one mutational pathway to

cells without the primary driver mutation. Conversely, cells with the primary driver mutation can be reached via different mutational pathways, because cells that get the primary driver mutation might have different amounts of secondary driver mutations. Hence, we need to sum over all those possible pathways. Let p be the number of secondary drivers that are present in the cell which acquires the primary driver mutation. Then $x_{1,k}^{(p)}(t)$ denotes the number of cells with the primary driver mutation and k secondary driver mutations, when the primary driver mutation has happened in a cell with p secondary driver mutations ($0 \leq p \leq k$). With this, the total number of cells with the primary driver mutation is

$$x_{1,k}(t) = \sum_{p=0}^{k} x_{1,k}^{(p)}(t). \tag{7.6}$$

The change in the number of cells now depends on p. We have

$$x_{1,k}^{(p)}(t) = \begin{cases} \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k} x_{1,k}^{(p)}(t-1) + \mu_{\mathrm{P}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k-1} x_{1,k-1}^{(p)}(t-1), & \text{if } p < k \\ \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k} x_{1,k}^{(p)}(t-1) + \mu_{\mathrm{D}}\varsigma_{\mathrm{P}}^{k} x_{0,k}(t-1), & \text{if } p = k. \end{cases} \tag{7.7}$$

The solution of (7.7) is given by the following theorem:

Theorem 2 *The average number of cells with the primary driver mutation and k secondary driver mutations, given that the primary driver mutation happens in a cell with p secondary driver mutations, is given by*

$$\begin{aligned} x_{1,k}^{(p)}(t) =& N\mu_D\mu_P^k\varsigma_{\mathrm{D}}^{k-p}\varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2}\frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}\left(1-\varsigma_{\mathrm{P}}^{n+1}\right)} \\ &\times\left[\nu_{\mathrm{P}}^{t-p}\Psi_{p,k}(t) - \sum_{j=p}^{k}\nu_{\mathrm{P}}^{j-p}\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\varsigma_{\mathrm{D}}\right)^{t-k}\Psi_{p,j}(j)\prod_{n=j}^{k-1}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-1}}{1-\varsigma_{\mathrm{DP}}^{k-n}}\right], \end{aligned} \tag{7.8}$$

where the function Ψ is defined as

$$\Psi_{p,k}(t) = \sum_{r=0}^{p}\frac{\left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^{r}\varsigma_{\mathrm{P}}^{\frac{r(r-1)}{2}}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}}{\prod_{j=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r}-\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})}. \tag{7.9}$$

Proof 2 *Again we proof the theorem by inserting (7.8) in (7.7) and showing that the equality holds true. We need to distinguish between the two cases as in (7.7). First, we proof the theorem for the case $p < k$.*

$$
\begin{aligned}
x_{1,k}^{(p)}(t) &= \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}x_{1,k}^{(p)}(t-1) + \mu_{\mathrm{P}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k-1}x_{1,k-1}^{(p)}(t-1) \qquad (7.10)\\
&= \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}N\mu_{D}\mu_{P}^{k}\varsigma_{\mathrm{D}}^{k-p}\varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2}\frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}\left(1-\varsigma_{\mathrm{P}}^{n+1}\right)}\\
&\times\left[\nu_{\mathrm{P}}^{t-p-1}\Psi_{p,k}(t-1) - \sum_{j=p}^{k}\nu_{\mathrm{P}}^{j-p}\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\varsigma_{\mathrm{D}}\right)^{t-k-1}\Psi_{p,j}(j)\prod_{n=j}^{k-1}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n}}\right]\\
&+ \mu_{\mathrm{P}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k-1}N\mu_{D}\mu_{P}^{k-1}\varsigma_{\mathrm{D}}^{k-p-1}\varsigma_{\mathrm{DP}}^{((k-1)(k-2)-p(p-1))/2}\frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}\left(1-\varsigma_{\mathrm{P}}^{n+1}\right)}\\
&\times\left[\nu_{\mathrm{P}}^{t-p-1}\Psi_{p,k-1}(t-1) - \sum_{j=p}^{k-1}\nu_{\mathrm{P}}^{j-p}\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\varsigma_{\mathrm{D}}\right)^{t-k}\Psi_{p,j}(j)\prod_{n=j}^{k-2}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n-1}}\right]
\end{aligned}
$$

By factoring out $N\mu_{D}\mu_{P}^{k}\varsigma_{\mathrm{D}}^{k-p}\varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2}\frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}\left(1-\varsigma_{\mathrm{P}}^{n+1}\right)}$ and simplifying the second factor, we obtain

$$
\begin{aligned}
x_{1,k}^{(p)}(t) &= N\mu_{D}\mu_{P}^{k}\varsigma_{\mathrm{D}}^{k-p}\varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2}\frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{n=0}^{p-1}\left(1-\varsigma_{\mathrm{P}}^{n+1}\right)} \qquad (7.11)\\
&\times\Bigg[\nu_{\mathrm{P}}^{t-p-1}\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{p,k}(t-1)+\Psi_{p,k-1}(t-1)\right) - \nu_{\mathrm{P}}^{k-p}(\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k})^{t-k}\Psi_{p,k}(k)\\
&- \sum_{j=p}^{k-1}\nu_{\mathrm{P}}^{j-p}\Psi_{p,j}(j)\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\right)^{t-k}\left(\varsigma_{\mathrm{DP}}^{k-j}\prod_{n=j}^{k-1}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n}} + \prod_{n=j}^{k-2}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n-1}}\right)\Bigg].
\end{aligned}
$$

When we compare (7.8) and (7.10), we see that the two equations are equal if

$$\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{p,k}(t-1) + \Psi_{p,k-1}(t-1) = \nu_{\mathrm{P}}\Psi_{p,k}(t) \qquad (7.12)$$

and

$$\varsigma_{\mathrm{DP}}^{k-j}\prod_{n=j}^{k-1}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n}} + \prod_{n=j}^{k-2}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-2}}{1-\varsigma_{\mathrm{DP}}^{k-n-1}} = \prod_{n=j}^{k-1}\frac{1-\varsigma_{\mathrm{DP}}^{t-n-1}}{1-\varsigma_{\mathrm{DP}}^{k-n}}. \qquad (7.13)$$

Multiplying both sides of Equation (7.13) by $\prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{k-n}\right)$ and factoring

out $\left(1-\varsigma_{\mathrm{DP}}^{t-n-2}\right)$*, we obtain*

$$\begin{aligned}
&\Leftrightarrow \quad \varsigma_{\mathrm{DP}}^{k-j}\prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{t-n-2}\right)+\left(1-\varsigma_{\mathrm{DP}}^{k-j}\right)\prod_{n=j}^{k-2}\left(1-\varsigma_{\mathrm{DP}}^{t-n-2}\right)=\prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{t-n-1}\right)\\
&\Leftrightarrow \quad \left(1-\varsigma_{\mathrm{DP}}^{k-j}+\varsigma_{\mathrm{DP}}^{k-j}\left(1-\varsigma_{\mathrm{DP}}^{t-k-1}\right)\right)\prod_{n=j}^{k-2}\left(1-\varsigma_{\mathrm{DP}}^{t-n-2}\right)=\prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{t-n-1}\right)\\
&\Leftrightarrow \quad \prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{t-n-1}\right)=\prod_{n=j}^{k-1}\left(1-\varsigma_{\mathrm{DP}}^{t-n-1}\right).
\end{aligned} \tag{7.14}$$

For Equation (7.12)*, we need to insert the definition of* Ψ

$$\begin{aligned}
&\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{p,k}(t-1)+\Psi_{p,k-1}(t-1)\\
&\quad=\sum_{r=0}^{p}\left(-\varsigma_{\mathrm{P}}^{t-p}\right)^{r}\varsigma_{\mathrm{P}}^{r(r-1)/2}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}\\
&\quad\quad\times\left(\frac{\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}}{\prod_{j=p}^{k}\left(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r}-\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\right)}+\frac{1}{\prod_{j=p}^{k-1}\left(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r}-\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\right)}\right).
\end{aligned} \tag{7.15}$$

Using the common denominator for the summands in the parentheses leads to

$$\begin{aligned}
&\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{p,k}(t-1)+\Psi_{p,k-1}(t-1)\\
&=\sum_{r=0}^{p}\left(-\varsigma_{\mathrm{P}}^{t-p}\right)^{r}\varsigma_{\mathrm{P}}^{r(r-1)/2}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}\left(\frac{\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}+\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r}-\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}}{\prod_{j=p}^{k}\left(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r}-\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j}\right)}\right).
\end{aligned} \tag{7.16}$$

Using $\left(-\varsigma_{\mathrm{P}}^{t-p}\right)^{r}\varsigma_{\mathrm{P}}^{r}=\left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^{r}$ *we finally obtain*

$$\begin{aligned}
&\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{p,k}(t-1)+\Psi_{p,k-1}(t-1)\\
&=\nu_{\mathrm{P}}\sum_{r=0}^{p}\left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^{r}\varsigma_{\mathrm{P}}^{r(r-1)/2}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}\\
&=\nu_{\mathrm{P}}\Psi_{p,k}(t).
\end{aligned} \tag{7.17}$$

This concludes the proof for the case $p<k$*. Now we look at the case* $p=k$*.*

We have

$$\begin{aligned}
&\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}x_{1,k}^{(k)}(t-1)+\mu_{\mathrm{D}}\varsigma_{\mathrm{P}}^{k}x_{0,k}(t-1)\\
=&\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}N\mu_{\mathrm{D}}\mu_{\mathrm{P}}^{k}\frac{\varsigma_{\mathrm{P}}^{k(k+1)/2}}{\prod_{n=0}^{k-1}(1-\varsigma_{\mathrm{P}}^{n+1})}\left[\nu_{\mathrm{P}}^{t-k-1}\Psi_{k,k}(t-1)-\left(\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\right)^{t-k-1}\Psi_{k,k}(k)\right]\\
+&\mu_{\mathrm{D}}\varsigma_{\mathrm{P}}^{k}N\mu_{\mathrm{P}}^{k}\nu_{\mathrm{P}}^{t-k-1}\varsigma_{\mathrm{P}}^{k(k-1)/2}\prod_{n=0}^{k-1}\frac{1-\varsigma_{\mathrm{P}}^{t-n-1}}{1-\varsigma_{\mathrm{P}}^{n+1}} \qquad (7.18)\\
=&N\mu_{\mathrm{D}}\mu_{\mathrm{P}}^{k}\frac{\varsigma_{\mathrm{P}}^{k(k+1)/2}}{\prod_{n=0}^{k-1}(1-\varsigma_{\mathrm{P}}^{n+1})}\\
\times&\left[\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\nu_{\mathrm{P}}^{t-k-1}\Psi_{k,k}(t-1)-(\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k})^{t-k}\Psi_{k,k}(k)+\nu_{\mathrm{P}}^{t-k-1}\prod_{n=0}^{k-1}\left(1-\varsigma_{\mathrm{P}}^{t-n-1}\right)\right].
\end{aligned}$$

In order for this to be equal to $x_{1,k}^{(k)}$*, we need*

$$\nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{k}\Psi_{k,k}(t-1)+\prod_{n=0}^{k-1}\left(1-\varsigma_{\mathrm{P}}^{t-n-1}\right)=\nu_{\mathrm{P}}\Psi_{k,k}(t). \qquad (7.19)$$

Analogue to (7.12) *this equation holds true if*

$$\prod_{n=0}^{k-1}\left(1-\varsigma_{\mathrm{P}}^{t-n-1}\right)=\Psi_{k-1,k}(t-1)=\sum_{r=0}^{k}\left(-\varsigma_{\mathrm{P}}^{t-k}\right)^{r}\varsigma_{\mathrm{P}}^{r(r-1)/2}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}. \qquad (7.20)$$

By writing the summation as a q-Pochhammer symbol, we have

$$\sum_{r=0}^{k}\left(-\varsigma_{\mathrm{P}}^{t-k}\right)^{r}\varsigma_{\mathrm{P}}^{r(r-1)/2}\begin{bmatrix}p\\r\end{bmatrix}_{\varsigma_{\mathrm{P}}}=\left(\varsigma_{\mathrm{P}}^{t-k};\varsigma_{\mathrm{P}}\right)_{k}=\prod_{n=0}^{k-1}\left(1-\varsigma_{\mathrm{P}}^{t-k+n}\right)=\prod_{n=0}^{k-1}\left(1-\varsigma_{\mathrm{P}}^{t-n-1}\right). \qquad (7.21)$$

This concludes the proof also for $p=k$*.*

7.3 Intuitive Description of Equation (11)

Here, we describe this equation in a more intuitive way. For each generation t, the number of possibilities to distribute the p secondary driver mutations over t time steps is given by the q-binomial coefficient $\begin{bmatrix}t\\p\end{bmatrix}_{\varsigma_{\mathrm{P}}}$. But the growth of the cells depends on the time when the secondary driver mutations are

first acquired. Due to fitness advantage, the earlier the mutations have been acquired, the faster the population grows, and also the sooner the primary driver mutation can be obtained. As in Equation (5), the effect of the fitness advantage on the cells without the primary driver mutation itself is captured by multiplying $\varsigma_{\mathrm{P}}^{p(p+1)/2}$. The effect on the primary driver mutation is more intricate. To capture this effect, we start from a q-binomial coefficient and rewrite the q-Pochhammer symbol in the numerator $\prod_{j=0}^{p-1} 1-\varsigma_{\mathrm{P}}^{t-j}$ in terms of a sum [Koekoek et al., 2010],

$$\begin{bmatrix} t \\ p \end{bmatrix}_{\varsigma_{\mathrm{P}}} = \prod_{j=0}^{p-1} \frac{1-\varsigma_{\mathrm{P}}^{t-j}}{1-\varsigma_{\mathrm{P}}^{j+1}} = \frac{\sum_{r=0}^{p} \left(-\varsigma_{\mathrm{P}}^{t}\right)^{r} \left(1/\varsigma_{\mathrm{P}}\right)^{\frac{r(r-1)}{2}} \begin{bmatrix} p \\ r \end{bmatrix}_{1/\varsigma_{\mathrm{P}}}}{\prod_{j=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{j+1})}. \tag{7.22}$$

To make this resemble the term in the parentheses in the second line of Equation (11), we divide the numerator by $\prod_{j=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})$ and we obtain

$$\frac{\sum_{r=0}^{p} \frac{1}{\prod_{j=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})} \left(-\varsigma_{\mathrm{P}}^{t}\right)^{r} \left(1/\varsigma_{\mathrm{P}}\right)^{\frac{r(r-1)}{2}} \begin{bmatrix} p \\ r \end{bmatrix}_{1/\varsigma_{\mathrm{P}}}}{\prod_{j=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{j+1})}. \tag{7.23}$$

With

$$\begin{bmatrix} p \\ r \end{bmatrix}_{\varsigma_{\mathrm{P}}} = \prod_{j=0}^{r-1} \frac{1-\varsigma_{\mathrm{P}}^{p-j}}{1-\varsigma_{\mathrm{P}}^{j+1}} = \prod_{j=0}^{r-1} \frac{\varsigma_{\mathrm{P}}^{p-2j-1}(1-1/\varsigma_{\mathrm{P}}^{p-j})}{1-1/\varsigma_{\mathrm{P}}^{j+1}} \tag{7.24}$$

$$= \varsigma_{\mathrm{P}}^{r(p-r)} \prod_{j=0}^{r-1} \frac{1-1/\varsigma_{\mathrm{P}}^{p-j}}{1-1/\varsigma_{\mathrm{P}}^{j+1}} = \varsigma_{\mathrm{P}}^{r(p-r)} \begin{bmatrix} p \\ r \end{bmatrix}_{1/\varsigma_{\mathrm{P}}} \tag{7.25}$$

Equation (7.23) can be written as

$$\frac{\sum_{r=0}^{p} \frac{1}{\prod_{j=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})} \left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^{r} \varsigma_{\mathrm{P}}^{\frac{r(r-1)}{2}} \begin{bmatrix} p \\ r \end{bmatrix}_{\varsigma_{\mathrm{P}}}}{\prod_{j=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{j+1})}. \tag{7.26}$$

For the numerator of this modified q-binomial coefficient, we introduce the abbreviation

$$\Psi_{p,k}(t) = \sum_{r=0}^{p} \frac{\left(-\varsigma_{\mathrm{P}}^{t-p+1}\right)^{r} \varsigma_{\mathrm{P}}^{\frac{r(r-1)}{2}} \begin{bmatrix} p \\ r \end{bmatrix}_{\varsigma_{\mathrm{P}}}}{\prod_{j=p}^{k}(\nu_{\mathrm{P}}\varsigma_{\mathrm{P}}^{r} - \nu_{\mathrm{D}}\varsigma_{\mathrm{D}}\varsigma_{\mathrm{DP}}^{j})}. \tag{7.27}$$

In terms of this Ψ-function Equation (11) can be written in a more compact form as

$$x_{1,k}(t) = N \sum_{p=0}^{k} \mu_D \mu_P^k \varsigma_{\mathrm{D}}^{k-p} \varsigma_{\mathrm{DP}}^{(k(k-1)-p(p-1))/2} \frac{\varsigma_{\mathrm{P}}^{p(p+1)/2}}{\prod_{j=0}^{p-1}(1-\varsigma_{\mathrm{P}}^{j+1})} \times \left[\nu_{\mathrm{P}}^{t-p} \Psi_{p,k}(t) - \sum_{j=p}^{k} \nu_{\mathrm{P}}^{j-p} (\nu_{\mathrm{D}} \varsigma_{\mathrm{D}} \varsigma_{\mathrm{DP}}^{j})^{t-k} \Psi_{p,j}(j) \prod_{m=j}^{k-1} \frac{1-\varsigma_{\mathrm{DP}}^{t-m-1}}{1-\varsigma_{\mathrm{DP}}^{k-m}} \right]. \tag{7.28}$$

7.4 General Probability Generating Functions

In Chapter 3, we considered only the case where each individual has to die or divide in every time step. Here we relax this assumption and consider a more realistic scenario where only some individuals proliferate or die, whereas others do not take any action at all (Fig. 7.1). Then, the probability generating functions for the four types: wild type, individuals with mutation A, individuals with mutation B, and individuals with both mutations are defined as

$$\begin{aligned}
f_{ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{ab} + (1 - b_{ab} - d_{ab}) z_{ab} \\
&\quad + b_{ab}((1-\mu_A-\mu_B) z_{ab} + \mu_A z_{Ab} + \mu_B z_{aB})^2, \\
f_{Ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{Ab} + (1 - b_{Ab} - d_{Ab}) z_{Ab} \\
&\quad + b_{Ab}((1-\mu_B^A) z_{Ab} + \mu_B^A z_{AB})^2, \\
f_{aB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{aB} + (1 - b_{aB} - d_{aB}) z_{aB} \\
&\quad + b_{aB}((1-\mu_A) z_{aB} + \mu_A z_{AB})^2, \\
f_{AB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{AB} + (1 - b_{AB} - d_{AB}) z_{AB} + b_{AB} z_{AB}^2.
\end{aligned} \tag{7.29}$$

The functions are similar to the scenario of binary splitting (cf. Eq. 1 in the main text). There is only one term added: $(1-b_i-d_i)z_i, i \in \{ab, Ab, aB, AB\}$ which denotes the case of the individual neither dividing nor dying. To make

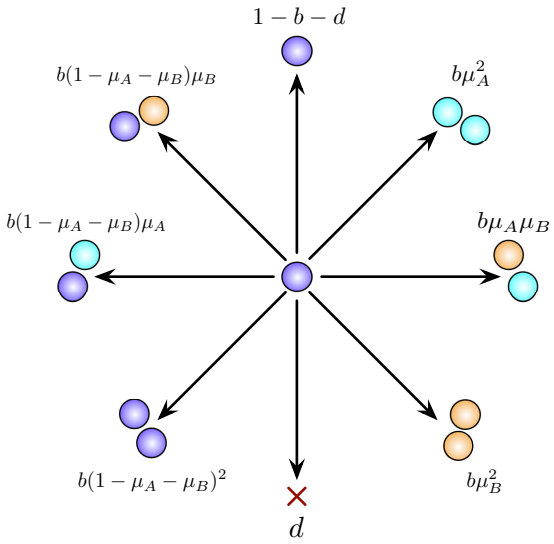

Figure 7.1: **Process described by the general pgf.** An individual can either die, proliferate, or neither and just live. If it proliferates the offspring can mutate. In case of including back mutations additional mutation terms appear leading as in Eq. (7.31).

increase the applicability of the model, one could also include back mutations,

$$\begin{aligned}
f_{ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{ab} + (1 - b_{ab} - d_{ab})z_{ab} \\
&\quad + b_{ab}((1 - \mu_A - \mu_B)z_{ab} + \mu_A z_{Ab} + \mu_B z_{aB})^2 \\
f_{Ab}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{Ab} + (1 - b_{Ab} - d_{Ab})z_{Ab} \\
&\quad + b_{Ab}((1 - \mu_B^A)z_{Ab} + \mu_{ab}^A z_{ab} + \mu_B^A z_{AB})^2 \\
f_{aB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{aB} + (1 - b_{aB} - d_{aB})z_{aB} \\
&\quad + b_{aB}((1 - \mu_A^B)z_{aB} + \mu_{ab}^B z_{ab} + \mu_A^B z_{AB})^2 \\
f_{AB}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) &= d_{AB} + (1 - b_{AB} - d_{AB})z_{AB} \\
&\quad + b_{AB}\left((1 - \mu_A^{AB} - \mu_B^{AB})z_{AB} + \mu_A^{AB} z_{Ab} + \mu_B^{AB} z_{aB}\right)^2
\end{aligned} \tag{7.30}$$

If the fitness landscape is rugged, i.e., if it has multiple local optima, they would be inaccessible from certain “downstream" directions if back mutations are not allowed. Hence allowing back mutations, allows to have a rugged fitness landscape with local optima accessible from multiple directions. The probability generating functions seem more complex, but the principle of the

computation as discussed in the main text does not change at all.

7.5 Time Distribution

Here, we give a more detailed description on how to calculate the time distribution for the minimal model with four types, and two paths, but with back mutations.

1. Calculate the extinction probability of the final mutant type AB as in [Athreya and Ney, 1972]

$$e_{AB} = \frac{d_{AB} + b_{AB}\left(\mu_A^{AB} + \mu_B^{AB}\right)^2}{b_{AB}(1 - \mu_A^{AB} - \mu_B^{AB})^2}. \tag{7.31}$$

   Note, that without back mutations, i.e., $\mu_A^{AB} = \mu_B^{AB} = 0$, the extinction probability reduces to $e_{AB} = \frac{d_{AB}}{b_{AB}}$ as in the main text.

2. Until some t_{max} calculate recursively

$$\begin{aligned}
f_{AB}^{\circ(t)} &= d_{AB} + (1 - b_{AB} - d_{AB}) f_{AB}^{\circ(t-1)} \\
&\quad + b_{AB}\left((1 - \mu_A^{AB} - \mu_B^{AB}) f_{AB}^{\circ(t-1)} + \mu_A^{AB} f_{Ab}^{\circ(t-1)} + \mu_B^{AB} f_{aB}^{\circ(t-1)}\right)^2, \\
f_{aB}^{\circ(t)} &= d_{aB} + (1 - b_{aB} - d_{aB}) f_{aB}^{\circ(t-1)} \\
&\quad + b_{aB}\left((1 - \mu_A - \mu_{aB}^{B}) f_{aB}^{\circ(t-1)} + \mu_{aB}^{B} f_{ab}^{\circ(t-1)} + \mu_A f_{AB}^{\circ(t-1)}\right)^2, \\
f_{Ab}^{\circ(t)} &= d_{Ab} + (1 - b_{Ab} - d_{Ab}) f_{Ab}^{\circ(t-1)} + \\
&\quad b_{Ab}\left((1 - \mu_B - \mu_{aB}^{A}) f_{Ab}^{\circ(t-1)} + \mu_{aB}^{A} f_{ab}^{\circ(t-1)} + \mu_B f_{AB}^{\circ(t-1)}\right)^2, \\
f(t) := f_{ab}^{\circ(t)} &= d_{ab} + (1 - b_{ab} - d_{ab}) f_{ab}^{\circ(t-1)} \\
&\quad + b_{ab}\left((1 - \mu_A - \mu_B) f_{ab}^{\circ(t-1)} + \mu_A f_{Ab}^{\circ(t-1)} + \mu_B f_{aB}^{\circ(t-1)}\right)^2
\end{aligned} \tag{7.32}$$

   where $f_{aB}^{\circ(0)} = f_{Ab}^{\circ(0)} = f_{ab}^{\circ(0)} = 1$ and $f_{AB}^{\circ(0)} = e_{AB}$. Note, that without back mutations these functions would not be coupled anymore and one can first calculate f_{Ab}^t and f_{aB}^t for all t, since those functions would not depend on f_{ab}. Moreover, $f_{AB}^{\circ(t)}$ would be equal to e_{AB} $\forall$ t. Hence, one would not need to recursively calculate $f_{AB}^{\circ(t)}$. However, the complexity does not change.

3. The probability to get the final, successful AB mutant, i.e., an individual that produces a lineage that does not die out again, exactly at time t is

$$\tau(t) = f^N(t-1) - f^N(t). \tag{7.33}$$

where N is the number of individuals in the beginning. Calculating this for all $t \in \{0, \ldots, t_{max}\}$ we obtain the time distribution.

7.6 Single-Path Time Distribution

Here, we explain the computation of the probability distribution of the pathway via type Ab exemplarily. Allowing back mutations it is unclear how to specify different mutational pathways. For instance, for the pathway $ab \to aB \to ab \to Ab \to AB$, it is unclear via which type the final mutant has been reached. Obviously the final mutant has been reached via type Ab, but it might be necessary for the population to first reach type aB. Hence, aB might play a vital role for reaching AB, too. For this reason, we neglect back mutations in the computation of the path probabilities, obtaining clear distinguishable pathways.

Let $\mathrm{Ab}(t)$ ($\mathrm{aB}(t)$) denote the random variable that there is an AB mutant until time t via pathway Ab (aB). Thus, $\neg\mathrm{Ab}(t)$ corresponds to the random variable, that there is no AB mutant until time t vial pathway Ab. Then the probability, that the first mutant arises exactly at time t via pathway Ab (i.e., not via pathway aB beforehand) is

$$\begin{aligned}\rho_{Ab}(t) =& P(Ab(t) \cap \neg Ab(t-1) \cap \neg aB(t-1)) \\ =& P(\neg Ab(t-1) \cap \neg aB(t-1)) - P(\neg Ab(t) \cap \neg aB(t-1)).\end{aligned} \tag{7.34}$$

The first term is calculated by the pgf as in Eq. (7.29). For the second term however, the time points for the different pathways are different. Let us derive a recursive function for this second term at this point. To do so, let us first consider the extinction probability for the subprocess of $Ab \to AB$, where the process starts with one Ab individual. As discussed previously, this extinction probability within $t-1$ time steps can be recursively calculated by

its probability generating function

$$f_{Ab}^{\circ(t-1)} = d_{Ab} + (1 - b_{Ab} - d_{Ab}) f_{Ab}^{\circ(t-2)} + b_{Ab} \left((1 - \mu_B) f_{Ab}^{\circ(t-2)} + \mu_B e_{AB} \right)^2, \tag{7.35}$$

with $f_{Ab}^{\circ(0)} = 1$. Similarly, the extinction probability for the subprocess $aB \to AB$ within $t-2$ time steps can be calculated recursively using the probability generating function for aB

$$f_{aB}^{\circ(t-2)} = d_{aB} + (1 - b_{aB} - d_{aB}) f_{aB}^{\circ(t-3)} + b_{aB} \left((1 - \mu_A) f_{aB}^{\circ(t-3)} + \mu_A e_{AB} \right)^2, \tag{7.36}$$

with $f_{aB}^{\circ(0)} = 1$. When we now consider the extinction probability of the whole process starting with an individual of type ab, it can either go extinct right away, or if it divides we can refer to the individual extinction probabilities for the different types (in case of mutation), i.e., their probability generating functions

$$\begin{aligned} \bar{f}_{ab}^{\circ(t)} &:= d_{ab} + (1 - b_{ab} - d_{ab}) \bar{f}_{ab}^{\circ(t-1)} \\ &\quad + b_{ab} \left((1 - \mu_A - \mu_B) f_{ab}^{\circ(t-1)} + \mu_A f_{Ab}^{\circ(t-1)} + \mu_B f_{aB}^{\circ(t-2)} \right)^2 \\ &= \bar{f}_{ab}(\bar{f}_{ab}^{\circ(t-1)}, f_{Ab}^{\circ(t-1)}, f_{aB}^{\circ(t-2)}), \end{aligned} \tag{7.37}$$

with $\bar{f}_{ab}^{\circ(0)} = 1$, $f_{Ab}^{\circ(0)} = 1$, and $f_{aB}^{\circ(0)} = 1$. Note, that in contrast to the normal probability generating function, here the probability generating function for type aB has one time step less, which agrees with the second term in (7.34). To not confuse this modified probability generating function with the common one, we use function names with a bar. Again, no probability generating function for the AB-type is necessary, since the actual extinction probability for this type is used.

We define this recursive function as

$$\bar{f}_{ab}^{\circ(t)}(z_{ab}, z_{Ab}, z_{aB}, z_{AB}) := \bar{f}^{(Ab)}(t). \tag{7.38}$$

The index Ab denotes, that this is the modified probability generating function

for the pathway via Ab.

With this we now describe the algorithm for the path probability.

1. Calculate the extinction probability of the final mutant type AB as above.

2. Until some t_{max} calculate recursively $f(t)$ as explained above in Eq. (7.32).

3. Until some t_{max} calculate recursively

$$\begin{aligned} f_{aB}^{\circ(t)} =& d_{aB} + (1 - b_{aB} - d_{aB}) f_{aB}^{\circ(t-1)} + b_{aB} \left((1-\mu_A) f_{aB}^{\circ(t-1)} + \mu_A e_{AB} \right)^2, \\ f_{Ab}^{\circ(t)} =& d_{Ab} + (1 - b_{Ab} - d_{Ab}) f_{Ab}^{\circ(t-1)} + b_{Ab} \left((1-\mu_B) f_{Ab}^{\circ(t-1)} + \mu_B e_{AB} \right)^2, \end{aligned} \tag{7.39}$$

$$\begin{aligned} \bar{f}^{(Ab)}(t) :=& \bar{f}_{ab}^{\circ(t)} = d_{ab} + (1 - b_{ab} - d_{ab}) \bar{f}_{ab}^{\circ(t-1)} \\ & + b_{ab} \left((1-\mu_A-\mu_B) \bar{f}_{ab}^{\circ(t-1)} + \mu_A f_{Ab}^{\circ(t-1)} + \mu_B f_{aB}^{\circ(t-2)} \right)^2, \end{aligned}$$

where $f_{aB}^0 = f_{aB}^{-1} = f_{Ab}^0 = f_{ab}^0 = 1$. Note, that the only difference is that the probability generating function of types not along the pathway considered is one time step behind (marked in red). This is also the reason, why there are two initial conditions needed for type aB.

4. The probability to get the final, successful AB mutant exactly at time t via path Ab and not getting a successful AB mutant beforehand is then computed by

$$\rho_{Ab} = f^N(t-1) - \left(\bar{f}^{(Ab)}(t) \right)^N. \tag{7.40}$$

Analogously, one can calculate the path probability for reaching the final mutant via aB. Note, that while this computation gives the correct path probabilities, the sum over all paths can be slightly greater than the overall time distribution. This is due to the fact that in time discrete systems the final mutant can be reached by different pathways at the same time. In the description here, such cases count for all pathways that succeed at the time.

7.7 Implementation of Burkitt Lymphoma Model

In this section the details for the implementation of the Burkitt Lymphoma Model as developed Section 4.1 are given.

At first the initial numbers of different cell types are set. The number of wild type cells is set to $N_0 = 10^4$, while for all other types the number of cells is initially zero. Now, for each cell a random number is drawn. This random number is then compared to the division and death probability of the respective cell type as defined in Section 4.1.1. As long as the upper limit of $N_{max} = 10^6$ has not been reached, the division probability for cells without the MYC translocation is increased by a parameter c_1, here $c_1 = 0.001$. As soon as the number of wild type cells reaches N_{max}, the division probabilities are not being increased anymore.

If a cell divides, its daughter cells can mutate. Hence, another random number is drawn and is compared to the mutation probabilities.

When all cells have been looked at, the number of cells is updated according to proliferation, apoptosis, and mutation events.

From a certain age on, e.g., 4000 hours $\approx$ 4.5 years, the division probability of cells without the MYC translocation is decreased by a parameter c_2, here $c_2 = 0.00003$, until the number of wild type cells reaches the baseline level again, $N_0 = 10^4$.

After the baseline level has been reached, the probabilities for division and apoptosis stay in general as defined in Section 4.1. If the number of wild type cells is larger (smaller) than $N_0 + (-)0.1N_0$, the proliferation probability is decreased (increased) by C_1. This way, the number of wild type cells stays within a 10% area of the baseline level. Note, that cells with mutations are not affected by this 10% constraint. The population of cells with a core mutation can therefore grow beyond $1.1N_0$.

Bibliography

Abe, M., K. Tasaki, K. Tominaga, S. Fukuhara, S. Imai, T. Osato, and H. Wakasa

1992. Characterization and comparison of two newly established epstein-barr virus (ebv)-negative and ebv-positive burkitt's lymphoma cell lines. ebv-negative cell line with a low level of expression of icam-1 molecule and ebv-positive cell line with a high level of expression of icam-1 molecule. *Cancer*, 69(3):763–771. (Cited on page 45.)

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton

2013. Signatures of mutational processes in human cancer. *Nature*, 500:415–421. (Cited on page 29.)

Allday, M. J.

2009. How does epstein-barr virus (ebv) complement the activation of myc in the pathogenesis of burkitt's lymphoma? *Semin Cancer Biol*, 19:366–376. (Cited on page 16.)

Antal, T. and P. Krapivsky

2011. Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, 2011:P08018. (Cited on pages 3, 12, 17, 37, 55 and 64.)

Armitage, P. and R. Doll

1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8:1–12. (Cited on pages 1 and 26.)

Athreya, K. B. and P. E. Ney

1972. *Branching Processes*. Berlin: Springer. (Cited on pages 11, 36, 64 and 84.)

Attolini, C. S.-O. and F. Michor

2009. Evolutionary theory of cancer. *Annals of the New York Academy of Sciences*, 1168:23–51. (Cited on page 1.)

Aukema, S. M., L. Theil, M. Rohde, B. Bauer, J. Bradtke, B. Burkhardt, B. R. Bonn, A. Claviez, S. Gattenlöhner, O. Makarova, I. Nagel, I. Oschlies, C. Pott, M. Szczepanowski, A. Traulsen, P. M. Kluin, W. Klapper, R. Siebert, and E. M. Murga Penas

2015. Sequential karyotyping in burkitt lymphoma reveals a linear clonal evolution with increase in karyotype complexity and a high frequency of recurrent secondary aberrations. *Br J Haematol*. (Cited on pages 55 and 112.)

Barrick, J. E. and R. E. Lenski

2013. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827–39. (Cited on page 17.)

Basso, K. and R. Dalla-Favera

2015. Germinal centres and b cell lymphomagenesis. *Nat Rev Immunol*, 15(3):172–84. (Cited on pages 11 and 56.)

Bauer, B. and C. S. Gokhale

2015. Repeatability of evolution on epistatic landscapes. *Sci Rep*, 5:9607. (Cited on pages 31 and 112.)

Bauer, B., R. Siebert, and A. Traulsen

2014. Cancer initiation with epistatic interactions between driver and pas-

senger mutations. *Journal of Theoretical Biology*, 358:52–60. (Cited on pages 15, 34, 40, 43 and 112.)

Beatty, J.
2006. Replaying life's tape. *The Journal of philosophy*, 103(7):336–362. (Cited on page 32.)

Beerenwinkel, N., T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and M. A. Nowak
2007. Genetic progression and the waiting time to cancer. *PLoS Computational Biology*, 3:e225. (Cited on pages 3, 17, 20 and 28.)

Bianconi, E., A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider
2013. An estimation of the number of cells in the human body. *Ann Hum Biol*, 40(6):463–71. (Cited on page 59.)

Blount, Z. D., J. E. Barrick, C. J. Davidson, and R. E. Lenski
2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417):513–518. (Cited on page 32.)

Boerma, E. G., R. Siebert, P. M. Kluin, and M. Baudis
2009. Translocations involving 8q24 in burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge. *Leukemia*, 23(2):225–234. (Cited on page 45.)

Boerma, E. G., G. W. van Imhoff, I. M. Appel, N. J. G. M. Veeger, P. M. Kluin, and J. C. Kluin-Nelemans
2004. Gender and age-related differences in burkitt lymphoma–epidemiological and clinical data from the netherlands. *Eur J Cancer*, 40(18):2781–2787. (Cited on pages 51 and 59.)

Bozic, I., T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak
2010. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences USA*, 107:18545–18550. (Cited on pages 3, 5, 11, 17, 20, 28 and 55.)

Bozic, I., J. G. Reiter, B. Allen, T. Antal, K. Chatterjee, P. Shah, Y. S. Moon, A. Yaqubie, N. Kelly, D. T. Le, E. J. Lipson, P. B. Chapman, L. A. Diaz, Jr, B. Vogelstein, and M. A. Nowak
2013. Evolutionary dynamics of cancer in response to targeted combination therapy. *Elife*, 2. (Cited on pages 12, 37 and 55.)

Burmeister, T., R. A. F. Macleod, R. Reinhardt, V. Mansmann, C. Loddenkemper, O. Marinets, H. G. Drexler, E. Thiel, and I. W. Blau
2006. A novel sporadic burkitt lymphoma cell line (blue-1) with a unique t(6;20)(q15;q11.2) rearrangement. *Leuk Res*, 30(11):1417–1423. (Cited on page 45.)

Busch, K., A. Borkhardt, W. Wössmann, A. Reiter, and J. Harbott
2004. Combined polymerase chain reaction methods to detect c-myc/igh rearrangement in childhood burkitt's lymphoma for minimal residual disease analysis. *Haematologica*, 89(7):818–25. (Cited on page 59.)

Campo, E.
2012. New pathogenic mechanisms in burkitt lymphoma. *Nat Genet*, 44(12):1288–1289. (Cited on page 16.)

Cooper, T. F., D. E. Rozen, and R. E. Lenski
2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. *Proceedings of the National Academy of Sciences USA*, 100(3):1072–1077. (Cited on page 32.)

Dang, C. V., K. A. O'Donnell, K. I. Zeller, T. Nguyen, R. C. Osthus, and F. Li
2006. The c-myc target gene network. *Semin Cancer Biol*, 16(4):253–264. (Cited on page 46.)

Darwin, C.
1859. *On the origin of species by means of natural selection.* Cambridge-London. Reprinted in Harvard University Press (1964). (Cited on page 1.)

Dave, S. S., K. Fu, G. W. Wright, L. T. Lam, P. Kluin, E.-J. Boerma, T. C. Greiner, D. D. Weisenburger, A. Rosenwald, G. Ott, H.-K. Müller-Hermelink, R. D. Gascoyne, J. Delabie, L. M. Rimsza, R. M. Braziel,

T. M. Grogan, E. Campo, E. S. Jaffe, B. J. Dave, W. Sanger, M. Bast, J. M. Vose, J. O. Armitage, J. M. Connors, E. B. Smeland, S. Kvaloy, H. Holte, R. I. Fisher, T. P. Miller, E. Montserrat, W. H. Wilson, M. Bahl, H. Zhao, L. Yang, J. Powell, R. Simon, W. C. Chan, L. M. Staudt, and Lymphoma/Leukemia Molecular Profiling Project
2006. Molecular diagnosis of burkitt's lymphoma. *N Engl J Med*, 354(23):2431–2442. (Cited on page 46.)

de Visser, J., R. F. Hoekstra, and H. van den Ende
1997. Test of interaction between genetic markers that affect fitness in aspergillus niger. *Evolution*, Pp. 1499–1505. (Cited on page 32.)

Desai, M. M., D. S. Fisher, and A. W. Murray
2007. The Speed of Evolution and Maintenance of Variation in Asexual Populations. *Current Biology*, 17(5):385–394. (Cited on page 33.)

Drost, J., R. H. van Jaarsveld, B. Ponsioen, C. Zimberlin, R. van Boxtel, A. Buijs, N. Sachs, R. M. Overmeer, G. J. Offerhaus, H. Begthel, J. Korving, M. van de Wetering, G. Schwank, M. Logtenberg, E. Cuppen, H. J. Snippert, J. P. Medema, G. J. P. L. Kops, and H. Clevers
2015. Sequential cancer mutations in cultured human intestinal stem cells. *Nature*, 521(7550):43–47. (Cited on page 46.)

Durrett, R., J. Foo, K. Leder, J. Mayberry, and F. Michor
2010. Evolutionary dynamics of tumor progression with random fitness values. *Theoretical Population Biology*, 78(1):54–66. (Cited on pages 12 and 55.)

Durrett, R. and S. Moseley
2010. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1):42–8. (Cited on page 12.)

Elena, S. F. and R. E. Lenski
2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, 4:457–469. (Cited on page 33.)

Elgendy, M., C. Sheridan, G. Brumatti, and S. J. Martin
2011. Oncogenic ras-induced expression of noxa and beclin-1 promotes autophagic cell death and limits clonogenic survival. *Mol Cell*, (1):23–35. (Cited on page 7.)

Fisher, R. A.
1918. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433. (Cited on page 32.)

Fisher, R. A.
1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford. (Cited on pages 3 and 32.)

Gerrish, P. J. and R. E. Lenski
1998. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103:127–144. (Cited on page 33.)

Gerstung, M. and N. Beerenwinkel
2010. Waiting time models of cancer progression. *Mathematical Population Studies*, 17:115–135. (Cited on pages 1, 3, 17 and 28.)

Gillespie, D. T.
1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361. (Cited on page 57.)

Gokhale, C. S., Y. Iwasa, M. A. Nowak, and A. Traulsen
2009. The pace of evolution across fitness valleys. *Journal of Theoretical Biology*, 259:613–620. (Cited on page 33.)

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis,

P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton
2007. Patterns of somatic mutation in human cancer genomes. *Nature*, 446:153–8. (Cited on page 1.)

Haccou, P., P. Jagers, and V. A. Vatutin
2005. *Branching processes: variation, growth, and extinction of populations*, volume 5. Cambridge: Cambridge University Press. (Cited on pages 8, 9, 10, 12, 19, 34, 35, 38, 55 and 63.)

Hajdu, S. I.
2011. A note from history: landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102. (Cited on page 1.)

Haldane, J. B. S.
1927. A mathematical theory of natural and artificial selection. v. selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23:838–844. (Cited on page 32.)

Hanahan, D. and R. Weinberg
2000. The hallmarks of cancer. *Cell*, 100:57–70. (Cited on page 1.)

Harrison, W. J.
1962. The total cellularity of the bone marrow in man. *J Clin Pathol*, 15:254–259. (Cited on page 59.)

Hegreness, M., N. Shoresh, D. Hartl, and R. Kishony
2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617. (Cited on page 33.)

Hindersin, L. and A. Traulsen
2014. Counterintuitive properties of the fixation time in network-structured populations. *Journal of The Royal Society Interface*, 11:20140606. (Cited on page 69.)

Hochberg, M. E., F. Thomas, E. Assenat, and U. Hibner
2013. Preventive evolutionary medicine of cancers. *Evol Appl*, 6(1):134–43. (Cited on page 1.)

Hoffman, B. and D. A. Liebermann
2008. Apoptotic signaling by c-myc. *Oncogene*, 27:6462–6472. (Cited on page 16.)

Huang, W., B. Haubold, C. Hauert, and A. Traulsen
2012. Emergence of stable polymorphism driven by evolutionary games between mutants. *Nature Communications*, 3:919. (Cited on page 67.)

Hummel, M., S. Bentink, H. Berger, W. Klapper, S. Wessendorf, T. F. E. Barth, H.-W. Bernd, S. B. Cogliatti, J. Dierlamm, A. C. Feller, M.-L. Hansmann, E. Haralambieva, L. Harder, D. Hasenclever, M. Kühn, D. Lenze, P. Lichter, J. I. Martin-Subero, P. Möller, H.-K. Müller-Hermelink, G. Ott, R. M. Parwaresch, C. Pott, A. Rosenwald, M. Rosolowski, C. Schwaenen, B. Stürzenhofecker, M. Szczepanowski, H. Trautmann, H.-H. Wacker, R. Spang, M. Loeffler, L. Trümper, H. Stein, R. Siebert, and Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe
2006. A biologic definition of burkitt's lymphoma from transcriptional and genomic profiling. *New England Journal of Medicine*, 354(23):2419–30. (Cited on page 16.)

Imhof, M. and C. Schlotterer
2001. Fitness effects of advantageous mutations in evolving Escherichia coli populations. *Proceedings of the National Academy of Sciences USA*, 98(3):1113–1117. (Cited on page 33.)

Iwasa, Y., F. Michor, N. L. Komarova, and M. A. Nowak
2005. Population genetics of tumor supressor genes. *Journal of Theoretical Biology*, 233:15–23. (Cited on pages 7 and 15.)

Iwasa, Y., F. Michor, and M. A. Nowak
2004. Stochastic tunnels in evolutionary dynamics. *Genetics*, 166:1571–1579. (Cited on page 33.)

Jagers, P.

1970. The composition of branching populations: a mathematical result and its application to determine the incidence of death in cell proliferation. *Mathematical Biosciences*, 8(3):227–238. (Cited on page 11.)

Jain, K. and J. Krug

2007. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics*, 175:1275–1288. (Cited on page 32.)

Jemal, A., F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman

2011. Global cancer statistics. *CA Cancer J Clin*, 61(2):69–90. (Cited on page 1.)

Johnson, N. A., G. W. Slack, K. J. Savage, J. M. Connors, S. Ben-Neriah, S. Rogic, D. W. Scott, K. L. Tan, C. Steidl, L. H. Sehn, W. C. Chan, J. Iqbal, P. N. Meyer, G. Lenz, G. Wright, L. M. Rimsza, C. Valentino, P. Brunhoeber, T. M. Grogan, R. M. Braziel, J. R. Cook, R. R. Tubbs, D. D. Weisenburger, E. Campo, A. Rosenwald, G. Ott, J. Delabie, C. Holcroft, E. S. Jaffe, L. M. Staudt, and R. D. Gascoyne

2012. Concurrent expression of myc and bcl2 in diffuse large b-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. *J Clin Oncol*, 30(28):3452–9. (Cited on page 46.)

Jones, S., W.-D. Chen, G. Parmigiani, F. Diehl, N. Beerenwinkel, T. Antal, A. Traulsen, M. A. Nowak, C. Siegel, V. Velculescu, K. W. Kinzler, B. Vogelstein, J. Willis, and S. Markowitz

2008. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences USA*, 105:4283–4288. (Cited on page 1.)

Justilien, V., M. P. Walsh, S. A. Ali, E. A. Thompson, N. R. Murray, and A. P. Fields

2014. The prkci and sox2 oncogenes are coamplified and cooperate to activate hedgehog signaling in lung squamous cell carcinoma. *Cancer Cell*, 25:139–151. (Cited on page 7.)

Kac, V. G. and P. Cheung

2002. *Quantum calculus*, Universitext. New York: Springer. (Cited on pages 23 and 76.)

Karlin, S. and H. M. A. Taylor

1975. *A First Course in Stochastic Processes*, 2nd edition edition. London: Academic. (Cited on page 69.)

Kaveh, K., N. L. Komarova, and M. Kohandel

2015. The duality of spatial death–birth and birth–death processes and limitations of the isothermal theorem. *Royal Society Open Science*, 2(4). (Cited on page 69.)

Khan, A. I., D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper

2011. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196. (Cited on page 41.)

Kimmel, M. and D. E. Axelrod

1991. Unequal cell division, growth regulation and colony size of mammalian cells: a mathematical model and analysis of experimental data. *Journal of theoretical biology*, 153(2):157–180. (Cited on page 12.)

Kimmel, M. and D. E. Axelrod

2002. *Branching Processes in Biology*. Springer NY. (Cited on pages 8, 10, 19, 35 and 55.)

Kimmel, M., D. E. Axelrod, and G. M. Wahl

1992. A branching process model of gene amplification following chromosome breakage. *Mutation Research/Reviews in Genetic Toxicology*, 276(3):225–239. (Cited on page 12.)

Klapper, W., M. Szczepanowski, B. Burkhardt, H. Berger, M. Rosolowski, S. Bentink, C. Schwaenen, S. Wessendorf, R. Spang, P. Möller, M. L. Hansmann, H.-W. Bernd, G. Ott, M. Hummel, H. Stein, M. Loeffler, L. Trümper, M. Zimmermann, A. Reiter, R. Siebert, and Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe

2008. Molecular profiling of pediatric mature b-cell lymphoma treated in population-based prospective clinical trials. *Blood*, 112(4):1374–1381. (Cited on page 45.)

Klein, U. and R. Dalla-Favera
2008. Germinal centres: role in b-cell physiology and malignancy. *Nat Rev Immunol*, 8(1):22–33. (Cited on pages 11 and 56.)

Kliegman, R. M.
2012. *Nelson textbook of pediatrics*. Saunders Elsevier. (Cited on pages 47 and 52.)

Kluin, P. M.
2014. The missing link in early follicular lymphoma development. *Blood*, 123(22):3371–3372. (Cited on page 56.)

Kluk, M. J., B. Chapuy, P. Sinha, A. Roy, P. Dal Cin, D. S. Neuberg, S. Monti, G. S. Pinkus, M. A. Shipp, and S. J. Rodig
2012. Immunohistochemical detection of myc-driven diffuse large b-cell lymphomas. *PLoS One*, 7(4):e33813. (Cited on page 46.)

Knudson, A. G.
1971. Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences USA*, 68:820–823. (Cited on page 7.)

Koekoek, R., P. Lesky, and R. F. Swarttouw
2010. *Hypergeometric orthogonal polynomials and their q-analogues*, Springer monographs in mathematics. Heidelberg: Springer. (Cited on pages 23, 25 and 81.)

Komarova, N. L., A. Sengupta, and M. A. Nowak
2003. Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. *J Theor Biol*, 223(4):433–450. (Cited on page 15.)

Komarova, N. L., L. Shahriyari, and D. Wodarz
2014. Complex role of space in the crossing of fitness valleys by asexual populations. *Journal of The Royal Society Interface*, 11(95):20140014. (Cited on page 66.)

Kuczmarski, R. J., C. L. Ogden, S. S. Guo, L. M. Grummer-Strawn, K. M.

Flegal, Z. Mei, R. Wei, L. R. Curtin, A. F. Roche, and C. L. Johnson
2002. 2000 cdc growth charts for the united states: methods and development. *Vital Health Stat 11*, (246):1–190. (Cited on page 59.)

Leder, K., K. Pitter, Q. Laplant, D. Hambardzumyan, B. D. Ross, T. A. Chan, E. C. Holland, and F. Michor
2014. Mathematical modeling of pdgf-driven glioblastoma reveals optimized radiation dosing schedules. *Cell*, 156(3):603–16. (Cited on page 3.)

Lee, T. H., L. M. DSouza, and G. E. Fox
1997. Equally parsimonious pathways through an rna sequence space are not equally likely. *Journal of Molecular Evolution*, 45:278–284. (Cited on page 37.)

Lengauer, C., K. W. Kinzler, and B. Vogelstein
1998. Genetic instabilities in human cancers. *Nature*, 396:643–9. (Cited on page 1.)

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler
1991. Long-term experimental evolution in escherichia coli. I. adaptation and divergence during 2,000 generations. *The American Naturalist*, 138:1315–1341. (Cited on page 32.)

Levitt, M.
1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1):59–107. (Cited on page 3.)

Limpens, J., D. de Jong, J. H. van Krieken, C. G. Price, B. D. Young, G. J. van Ommen, and P. M. Kluin
1991. Bcl-2/jh rearrangements in benign lymphoid tissues with follicular hyperplasia. *Oncogene*, 6(12):2271–2276. (Cited on page 56.)

Lin, C. Y., J. Lovén, P. B. Rahl, R. M. Paranal, C. B. Burge, J. E. Bradner, T. I. Lee, and R. A. Young
2012. Transcriptional amplification in tumor cells with elevated c-myc. *Cell*, 151:56–67. (Cited on page 16.)

Liu, Y. J., J. Zhang, P. J. Lane, E. Y. Chan, and I. C. MacLennan
1991. Sites of specific b cell activation in primary and secondary responses to t cell-dependent and t cell-independent antigens. *Eur J Immunol*, 21(12):2951–62. (Cited on page 47.)

Love, C., Z. Sun, D. Jima, G. Li, J. Zhang, R. Miles, K. L. Richards, C. H. Dunphy, W. W. L. Choi, G. Srivastava, P. L. Lugar, D. A. Rizzieri, A. S. Lagoo, L. Bernal-Mizrachi, K. P. Mann, C. R. Flowers, K. N. Naresh, A. M. Evens, A. Chadburn, L. I. Gordon, M. B. Czader, J. I. Gill, E. D. Hsi, A. Greenough, A. B. Moffitt, M. McKinney, A. Banerjee, V. Grubor, S. Levy, D. B. Dunson, and S. S. Dave
2012. The genetic landscape of mutations in burkitt lymphoma. *Nat Genet*, 44:1321–1325. (Cited on page 29.)

Lundin, C., L. Hjorth, M. Behrendtz, M. Ehinger, A. Biloglav, and B. Johansson
2013. Submicroscopic genomic imbalances in burkitt lymphomas/leukemias: association with age and further evidence that 8q24/myc translocations are not sufficient for leukemogenesis. *Genes Chromosomes Cancer*, 52(4):370–7. (Cited on pages 45 and 46.)

Lynch, M. and A. Abegg
2010. The Rate of Establishment of Complex Adaptations. *Molecular Biology and Evolution*, 27(6):1404–1414. (Cited on page 33.)

Mamessier, E., F. Broussais-Guillaumot, B. Chetaille, R. Bouabdallah, L. Xerri, E. S. Jaffe, and B. Nadel
2014. Nature and importance of follicular lymphoma precursors. *Haematologica*, 99(5):802–810. (Cited on page 56.)

Matsuo, Y., R. A. MacLeod, K. Kojima, K. Kuwahara, A. Sakata, H. G. Drexler, C. Nishizaki, S. Fukuda, Y. Inoue, T. Sezaki, N. Sakaguchi, and K. Orita
1997. A novel all-l3 cell line, balm-16, lacking expression of immunoglobulin chains derived from a patient with hypercalcemia. *Leukemia*, 11(12):2168–2174. (Cited on page 45.)

McFarland, C. D., K. S. Korolev, G. V. Kryukov, S. R. Sunyaev, and L. A. Mirny

2013. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences USA*, 110:2910–5. (Cited on pages 28 and 29.)

Meyer, J. R., D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski

2012. Repeatability and contingency in he evolution of a key innovation in phage lambda. *Science*, 335:428–432. (Cited on pages 17 and 32.)

Meyer-Hermann, M., E. Mohr, N. Pelletier, Y. Zhang, G. D. Victora, and K.-M. Toellner

2012. A theory of germinal center b cell selection, division, and exit. *Cell Rep*, 2(1):162–174. (Cited on page 56.)

Michor, F., Y. Iwasa, and M. A. Nowak

2004. Dynamics of cancer progression. *Nature Reviews Cancer*, 4:197–205. (Cited on pages 1, 3, 17 and 28.)

Müller, J. R., S. Janz, J. J. Goedert, M. Potter, and C. S. Rabkin

1995. Persistence of immunoglobulin heavy chain/c-myc recombination-positive lymphocyte clones in the blood of human immunodeficiency virus-infected homosexual men. *Proceedings of the National Academy of Sciences USA*, 92(14):6577–81. (Cited on pages 29 and 46.)

Nagy, N., G. Klein, and E. Klein

2009. To the genesis of burkitt lymphoma: regulation of apoptosis by ebna-1 and sap may determine the fate of ig-myc translocation carrying b lymphocytes. *Semin Cancer Biol*, 19(6):407–410. (Cited on page 56.)

Newton, I.

1872. Philosophiae naturalis principia mathematica, london 1687; deutsch von. *J. Ph. Wolfers*. (Cited on page 2.)

Nie, Z., G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D. R. Green,

L. Tessarollo, R. Casellas, K. Zhao, and D. Levens
2012. c-myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151:68–79. (Cited on page 16.)

Nowak, M., N. L. Komarova, A. Sengupta, P. Jallepalli, I. Shih, B. Vogelstein, and C. Lengauer
2002. The role of chromosomal instability in tumour initiation. *Proceedings of the National Academy of Sciences USA*, 99(25):16226–16231. (Cited on page 7.)

Nowak, M. A.
2006. *Evolutionary dynamics*. Cambridge MA: Harvard University Press. (Cited on page 69.)

Nowak, M. A., F. Michor, N. L. Komarova, and Y. Iwasa
2004. Evolutionary dynamics of tumor suppressor gene inactivation. *Proceedings of the National Academy of Sciences USA*, 101:10635–10638. (Cited on page 7.)

Onciu, M., E. Schlette, Y. Zhou, S. C. Raimondi, F. J. Giles, H. M. Kantarjian, L. J. Medeiros, R. C. Ribeiro, C.-H. Pui, and J. T. Sandlund
2006. Secondary chromosomal abnormalities predict outcome in pediatric and adult high-stage burkitt lymphoma. *Cancer*, 107(5):1084–1092. (Cited on page 45.)

Park, S.-C. and J. Krug
2007. Clonal interference in large populations. *Proceedings of the National Academy of Sciences USA*, 104(46):18135–18140. (Cited on page 33.)

Parmigiani, G., S. Boca, J. Lin, K. W. Kinzler, V. Velculescu, and B. Vogelstein
2009. Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics*, 93:17–21. (Cited on page 1.)

Pelengaris, S., M. Khan, and G. I. Evan
2002. Suppression of myc-induced apoptosis in beta cells exposes multiple oncogenic properties of myc and triggers carcinogenic progression. *Cell*, 109(3):321–34. (Cited on page 16.)

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans
2007. Empirical fitness landscales reveal accessible evolutionary paths. *Nature*, 445:383–386. (Cited on pages 32, 33 and 34.)

Radmacher, M. D., G. Kelsoe, and T. B. Kepler
1998. Predicted and inferred waiting times for key mutations in the germinal centre reaction: evidence for stochasticity in selection. *Immunol Cell Biol*, 76(4):373–381. (Cited on pages 47 and 56.)

Reiter, J. G., I. Bozic, B. Allen, K. Chatterjee, and M. A. Nowak
2013. The effect of one additional driver mutation on tumor progression. *Evolutionary Applications*, 6:34–45. (Cited on pages 3, 12, 17 and 28.)

Richter, J., M. Schlesner, S. Hoffmann, M. Kreuz, E. Leich, B. Burkhardt, M. Rosolowski, O. Ammerpohl, R. Wagener, S. H. Bernhart, D. Lenze, M. Szczepanowski, M. Paulsen, S. Lipinski, R. B. Russell, S. Adam-Klages, G. Apic, A. Claviez, D. Hasenclever, V. Hovestadt, N. Hornig, J. O. Korbel, D. Kube, D. Langenberger, C. Lawerenz, J. Lisfeld, K. Meyer, S. Picelli, J. Pischimarov, B. Radlwimmer, T. Rausch, M. Rohde, M. Schilhabel, R. Scholtysik, R. Spang, H. Trautmann, T. Zenz, A. Borkhardt, H. G. Drexler, P. Möller, R. A. F. MacLeod, C. Pott, S. Schreiber, L. Trümper, M. Loeffler, P. F. Stadler, P. Lichter, R. Eils, R. Küppers, M. Hummel, W. Klapper, P. Rosenstiel, A. Rosenwald, B. Brors, R. Siebert, and ICGC MMML-Seq Project
2012. Recurrent mutation of the id3 gene in burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nature Genetics*, 44(12):1316–20. (Cited on pages 16 and 29.)

Salaverria, I., I. Martin-Guerrero, R. Wagener, M. Kreuz, C. W. Kohler, J. Richter, B. Pienkowska-Grela, P. Adam, B. Burkhardt, A. Claviez, C. Damm-Welk, H. G. Drexler, M. Hummel, E. S. Jaffe, R. Küppers, C. Lefebvre, J. Lisfeld, M. Löffler, R. A. F. Macleod, I. Nagel, I. Oschlies, M. Rosolowski, R. B. Russell, G. Rymkiewicz, D. Schindler, M. Schlesner, R. Scholtysik, C. Schwaenen, R. Spang, M. Szczepanowski, L. Trümper, I. Vater, S. Wessendorf, W. Klapper, R. Siebert, Molecular Mechanisms in Malignant Lymphoma Network Project, and Berlin-Frankfurt-Münster

Non-Hodgkin Lymphoma Group
2014. A recurrent 11q aberration pattern characterizes a subset of myc-negative high-grade b-cell lymphomas resembling burkitt lymphoma. *Blood*, 123(8):1187–1198. (Cited on page 45.)

Salaverria, I. and R. Siebert
2011. The gray zone between burkitt's lymphoma and diffuse large b-cell lymphoma from a genetics perspective. *Journal of Clinical Oncology*, 29:1835–1843. (Cited on page 16.)

Salaverria, I., A. Zettl, S. Beà, E. M. Hartmann, S. S. Dave, G. W. Wright, E.-J. Boerma, P. M. Kluin, G. Ott, W. C. Chan, D. D. Weisenburger, A. Lopez-Guillermo, R. D. Gascoyne, J. Delabie, L. M. Rimsza, R. M. Braziel, E. S. Jaffe, L. M. Staudt, H. K. Müller-Hermelink, E. Campo, A. Rosenwald, and Leukemia and Lymphoma Molecular Profiling Project (LLMPP)
2008. Chromosomal alterations detected by comparative genomic hybridization in subgroups of gene expression-defined burkitt's lymphoma. *Haematologica*, 93(9):1327–1334. (Cited on page 45.)

Sander, S., D. P. Calado, L. Srinivasan, K. Köchert, B. Zhang, M. Rosolowski, S. J. Rodig, K. Holzmann, S. Stilgenbauer, R. Siebert, L. Bullinger, and K. Rajewsky
2012. Synergy between pi3k signaling and myc in burkitt lymphomagenesis. *Cancer Cell*, 22(2):167–79. (Cited on page 16.)

Sasaki, A. and M. A. Nowak
2003. Mutation landscapes. *Journal of Theoretical Biology*, 224(2):241–7. (Cited on page 34.)

Schmitz, R., M. Ceribelli, S. Pittaluga, G. Wright, and L. M. Staudt
2014. Oncogenic mechanisms in burkitt lymphoma. *Cold Spring Harb Perspect Med*, 4(2). (Cited on page 16.)

Schmitz, R., R. M. Young, M. Ceribelli, S. Jhavar, W. Xiao, M. Zhang, G. Wright, A. L. Shaffer, D. J. Hodson, E. Buras, X. Liu, J. Powell, Y. Yang, W. Xu, H. Zhao, H. Kohlhammer, A. Rosenwald, P. Kluin, H. K.

Müller-Hermelink, G. Ott, R. D. Gascoyne, J. M. Connors, L. M. Rimsza, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, M. D. Ogwang, S. J. Reynolds, R. I. Fisher, R. M. Braziel, R. R. Tubbs, J. R. Cook, D. D. Weisenburger, W. C. Chan, S. Pittaluga, W. Wilson, T. A. Waldmann, M. Rowe, S. M. Mbulaiteye, A. B. Rickinson, and L. M. Staudt
2012. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418):116–20. (Cited on page 29.)

Schreiber, R. D., L. J. Old, and M. J. Smyth
2011. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*, 331(6024):1565–70. (Cited on page 63.)

Serrano, M., A. W. Lin, M. E. McCurrach, D. Beach, and S. W. Lowe
1997. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16ink4a. *Cell*, 88(5):593–602. (Cited on page 7.)

Sjöblom, T., S. Jones, L. Wood, D. Parsons, J. Lin, T. Barber, D. Mandelker, R. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. Markowitz, J. Willis, D. Dawson, J. Willson, A. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. Park, K. Bachman, N. Papadopoulos, B. Vogelstein, K. Kinzler, and V. Velculescu
2006. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314:268–274. (Cited on page 1.)

Szendro, I. G., J. Franke, J. A. G. M. de Visser, and J. Krug
2013. Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences USA*, 110:571–576. (Cited on pages 32, 33 and 34.)

Tapia, G., R. Lopez, A. M. Muñoz-Mármol, J. L. Mate, C. Sanz, R. Marginet, J.-T. Navarro, J.-M. Ribera, and A. Ariza
2011. Immunohistochemical detection of myc protein correlates with myc gene status in aggressive b cell lymphomas. *Histopathology*, 59(4):672–678. (Cited on page 46.)

Tellier, J., C. Menard, S. Roulland, N. Martin, C. Monvoisin, L. Chasson,

B. Nadel, P. Gaulard, C. Schiff, and K. Tarte
2014. Human t(14;18)positive germinal center b cells: a new step in follicular lymphoma pathogenesis? *Blood*, 123(22):3462–3465. (Cited on page 56.)

Tomasetti, C., L. Marchionni, M. A. Nowak, G. Parmigiani, and B. Vogelstein
2015. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A*, 112(1):118–123. (Cited on page 46.)

Tomasetti, C., B. Vogelstein, and G. Parmigiani
2013. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci U S A*, 110(6):1999–2004. (Cited on pages 57 and 58.)

Traulsen, A. and C. Hauert
2009. Stochastic evolutionary game dynamics. In *Reviews of Nonlinear Dynamics and Complexity*, H. G. Schuster, ed., volume II, Pp. 25–61. Weinheim: Wiley-VCH. (Cited on page 69.)

Traulsen, A., J. M. Pacheco, and D. Dingli
2010. Reproductive fitness advantage of bcr-abl expressing leukemia cells. *Cancer Letters*, 294:43–48. (Cited on page 1.)

Travisano, M., J. A. Mongold, A. F. Bennett, and R. E. Lenski
1995. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194):87–90. (Cited on page 32.)

Travisano, M. and R. G. Shaw
2013. Lost in the map. *Evolution*, 67(2):305–314. (Cited on page 32.)

Victora, G. D., D. Dominguez-Sola, A. B. Holmes, S. Deroubaix, R. Dalla-Favera, and M. C. Nussenzweig
2012. Identification of human germinal center light and dark zone cells and their relationship to human b-cell lymphomas. *Blood*, 120(11):2240–2248. (Cited on page 56.)

Victora, G. D. and M. C. Nussenzweig
2012. Germinal centers. *Annu Rev Immunol*, 30:429–57. (Cited on page 56.)

Vogelstein, B. and K. Kinzler

2004. Cancer genes and the pathways they control. *Nature Medicine*, 10:789–799. (Cited on page 7.)

Wang, C., Y. Tai, M. P. Lisanti, and D. J. Liao

2011. c-myc induction of programmed cell death may contribute to carcinogenesis: a perspective inspired by several concepts of chemical carcinogenesis. *Cancer Biology and Therapy*, 11:615–626. (Cited on page 16.)

Watson, J. D., T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick

2014. Molecular biology of the gene. *Molecular biology of the gene.*, (7th edition). (Cited on page 4.)

Weinreich, D., N. Delaney, M. DePristo, and D. Hartl

2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312:111–114. (Cited on pages 32, 33, 37 and 41.)

Weinreich, D. M. and L. Chao

2005. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution*, 59:1175–1182. (Cited on page 33.)

Weinreich, D. M., R. Watson, and L. Chao

2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, 56(6):1165–1174. (Cited on pages 32 and 33.)

Weissman, D. B., M. M. Desai, D. S. Fisher, and M. W. Feldman

2009. The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology*, 75(4):286–300. (Cited on page 33.)

Wodarz, D. and N. Komarova

2005. *Computational biology of cancer: Lecture notes and mathematical modeling*. World Scientific Publishing. (Cited on page 1.)

Wodarz, D. and N. L. Komarova

2014. *Dynamics of cancer: mathematical foundations of oncology*. World Scientific Publishing Company. (Cited on page 4.)

Woo, K. B., W. K. Funkhouser, C. Sullivan, and O. Alabaster
1980. Analysis of the proliferation kinetics of burkitt's lymphoma cells. *Cell Tissue Kinet*, 13(6):591–604. (Cited on pages 48, 49, 55 and 57.)

Wood, L. D., D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein
2007. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113. (Cited on page 1.)

Wu, B., B. Bauer, T. Galla, and A. Traulsen
2015. Fitness-based models and pairwise comparison models of evolutionary games are typically different—even in unstructured populations. *New Journal of Physics*, 17:023043. (Cited on page 3.)

Zech, L., U. Haglund, K. Nilsson, and G. Klein
1976. Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with burkitt and non-burkitt lymphomas. *Int J Cancer*, 17:47–56. (Cited on page 16.)

Acknowledgments

I thank especially my supervisor Arne Traulsen. His patience and guidance has shaped my scientific interest and has helped me enormously in finding the strength and will to accomplish my Ph.D. Trusting in me where I did not trust myself has further supported both my scientific and personal development. I am deeply grateful for all his help.

I also thank the whole group of the Department of Theory! All the support and fruitful discussions have helped a lot. You guys have provided a pleasant working atmosphere, which has made it enjoyable to work in.

Many thanks also to my coauthors that have participated in my work. Special thanks to Prof. Dr. Reiner Siebert and Dr. Sietse Aukema for fruitful discussions about the Burkitt Lymphoma and a great collaboration! Furthermore, I express my sincere gratitude to Chaitanya Gokhale for his support and guidance.

And lastly I thank my family and friends. Having a friendly ear in difficult times was just as important as scientific advice.

Curriculum Vitæ

	Born December 27th 1986 in Lünen
2006	Abitur at the Freiherr-vom-Stein Gymnasium, Lünen, Germany
2006 – 2009	Bachelor of Science: Computational Life Sciences – Grade 1.6 University of Lübeck, Germany
2009 – 2012	Master of Science: Computational Life Sciences – Grade 1.3 Master's Thesis: Structural and Functional Protein Alignment Supervisor: Prof. Dr. Jürgen Prestin University of Lübeck, Germany
2012 – 2015	Ph.D. student: Mathematical Biology Max Planck Institute for Evolutionary Biology in Plön Supervisor: Prof. Dr. Arne Traulsen

Declaration

This thesis is a presentation of my original work, apart from my supervisor, Arne Traulsen's guidance. This thesis has not been submitted partly or wholly as a part of a doctoral degree to any other examining body. This thesis has been prepared according to the rules of Good Scientific Practice of the German Research Foundation.

Three papers published separately in the Journals *Journal of Theoretical Biology*, *Scientific Reports*, and *British Journal of Haematology*, as well as two chapters based on work in progress, are included in this thesis.

- Published paper [Bauer et al., 2014]
  I (B.B.) have developed the model, did the mathematical analysis, and performed simulations. B.B., Reiner Siebert (R.S.), and Arne Traulsen (A.T.) wrote the manuscript.

- Published paper [Bauer and Gokhale, 2015]
  B.B. did the mathematical analysis and performed simulations. B.B. and Chaitanya S. Gokhale (C.S.G.) developed the recursive algorithm and wrote the manuscript.

- Published paper [Aukema et al., 2015]
  B.B. and Sietse M. Aukema (S.M.A.) developed the theoretical model. B.B. performed simulations. B.B., S.M.A., and A.T. wrote the manuscript.

- Work in progress, Chapter 4.1
  B.B., R.S., S.M.A., and A.T. developed the theoretical model. B.B. performed simulations. B.B. and A.T. wrote the manuscript.

- Work in progress, Chapter 5.1
  B.B. and Bin Wu (B.W.) have performed the mathematical analysis.

- Work in progress, Chapter 5.2
  B.B. and A.T. have developed the model. B.B. and Miriam Otto (M.O.) have written the computational program and analyzed the model.