Computational modelling of hematopoietic stem cell division and regulation dynamics

Richard C. van der Wath
St John’s College
University of Cambridge

A thesis submitted for the degree of
Doctor of Philosophy
September 2009
Declaration

I hereby declare that my thesis entitled *Computational modelling of hematopoietic stem cell division and regulation dynamics* is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University, and does not exceed 60,000 words.

*Richard C. van der Wath*
Abstract

Computational modelling of hematopoietic stem cell division and regulation dynamics

Richard C. van der Wath

Abstract

In healthy adult multicellular organisms, daily production and loss of millions of blood cells due to normal cellular turnover are in equilibrium, a condition called homeostasis. During hematopoietic stress or after injury, homeostasis is rapidly restored by a small number of tightly regulated hematopoietic stem cells (HSCs) in the bone marrow. Until recently it was unclear how HSCs could be so effective in these very different tasks, since they seldom replicate to avoid accumulating DNA mutations, replicative senescence or exhaustion. The work presented in this thesis is at the frontier of a new paradigm in hematopoietic regulation that posit the existence of two related mouse HSC populations. First, a dormant HSC (d-HSC) population that actively divides only after hematopoietic injury, returning to a resting state once homeostasis has been restored. And second, an active HSC subset that by and large produces the progenitors and mature cells required for maintenance of day-to-day hematopoiesis.

We evaluate experimental and conceptual support for the d-HSC hypotheses from a systems biology perspective. Through extensive modelling of experimental DNA label-retaining cell data, we demonstrate how quantifying HSC division kinetics can assist in testing hypotheses about the population biology of HSCs. Our comprehensive analysis introduces novel discrete (agent-based), continuous (ordinary differential equation), and stochastic (Markov process) computational models, based on two independent datasets. All the results consistently show that under our assumptions only models accounting for a slowly dividing subpopulation of cells (putative d-HSCs) can describe the observed data satisfactorily. We subsequently derive a new conceptual model on dynamic hematopoietic regulation. In particular we investigate what the impact of the d-HSC population on dynamic maintenance and restoration of homeostasis is. Assuming the models we define are sound representations of the real biological system, bifurcation analysis reveals that malfunctioning d-HSCs are likely to play a central role in hematological diseases characterised by oscillatory peripheral cell numbers. Using elasticities as a measure, we provide well-grounded arguments that d-HSC regulation of hematopoiesis is evolutionarily more likely than conventional (d-HSC free) regulation of hematopoiesis.

The discovery of the d-HSC population has important implications for cancer treatment since recent evidence suggests that some cancer stem cells are also dormant, providing an explanation for their remarkable resistance to chemotherapy.
Acknowledgements

For he who desires to build a future dare not neglect the past. Seek therefore all that is good and beautiful in the past. Build on it your ideal and strive to realise that ideal for the future.

- S.J.P. Kruger. Clarens (Vaud), Switzerland. 29 June 1904.

I am indebted to the biologists Anne Wilson, Andreas Trumpp, and Elisa Laurenti for their hospitality in Lausanne and for eagerly sharing their wealth of knowledge (and of course their data) on hematopoietic stem cells. I thank my supervisor Pietro Liò for setting the waypoints but still allowing me the freedom to satisfy my curiosity. I acknowledge the Bradlow Foundation who through their prestigious scholarship provided generous financial support. My appreciation also goes to St John’s College for promoting a family-friendly atmosphere thereby offering me the opportunity to experience Cambridge traditions at its best.

A special thanks goes to my family for their prayers, encouragement, and financial support during the course of this work. Also to my friends for making my time at Cambridge more enjoyable. I am deeply grateful to my lovely wife Elizabeth for her endless love, support and enthusiasm. I thank God to be blessed with such a wonderful wife and three beautiful children, Willem, Sybrand, and Alet to whom I dedicate this thesis.

Richard C. van der Wath, Cambridge, September 2009
# Contents

Declaration ......................................................... i
Abstract ............................................................ ii
Acknowledgements .................................................. iii

1 Introduction ................................. 1

   1.1 Biological background ......................... 2
      1.1.1 Central dogma of molecular biology ...... 2
      1.1.2 Stem cells .................................. 4
      1.1.3 Hematopoietic stem cells .................. 5
      1.1.4 Cell division (mitosis) ...................... 7
   1.2 Measuring cell division kinetics ............. 8
      1.2.1 DNA labelling ................................ 8
      1.2.2 Non-DNA labelling ......................... 10
   1.3 Objective ........................................ 11
      1.3.1 Estimating HSC division rates and dynamics 11
      1.3.2 Modelling dynamic hematopoietic regulation 12
      1.3.3 Assumptions ................................ 13
   1.4 Publications ......................................... 14
   1.5 Outline ........................................... 16

2 Agent-based model ....................... 18

   2.1 Introduction ....................................... 18
   2.2 Ordinary differential equation model .......... 19
      2.2.1 Experimental data .......................... 20
      2.2.2 Model assumptions .......................... 20
      2.2.3 Asymmetric chromosome segregation equations 21
      2.2.4 Random chromosome segregation equations .... 22
      2.2.5 Analytic solutions .......................... 24
   2.3 Agent-based model ............................... 25
      2.3.1 Model assumptions .......................... 26
4 Label-retaining cell dynamics as a stochastic kinetic process 78
  4.1 Introduction .................................................. 78
  4.1.1 Markov process ................................................. 80
  4.2 Master equations ............................................... 81
    4.2.1 LRC assay reactions .......................................... 81
    4.2.2 One-population master equation .......................... 83
    4.2.3 Two-population master equation ......................... 83
    4.2.4 Stochastic Petri net representation .................... 85
    4.2.5 Relationship to ODE model ............................... 86
  4.3 Markov process simulation .................................... 87
    4.3.1 Event-time distribution ...................................... 87
    4.3.2 Event distribution ........................................... 89
  4.4 Results ......................................................... 90
    4.4.1 Stochastic variation cannot explain long-term label retention .... 90
    4.4.2 Two-sample Kolmogorov-Smirnov test .................... 93
  4.5 Conclusion ..................................................... 94

5 Parallel parameter exploration on the Grid 97
  5.1 Introduction ................................................... 97
  5.2 High dimensional parameter space ........................... 98
    5.2.1 Exploration boundaries ....................................... 99
    5.2.2 Parallel strategy ............................................ 99
    5.2.3 Grid deployment ............................................. 101
  5.3 Results ......................................................... 103
    5.3.1 Grid performance ............................................ 103
    5.3.2 Best parameter sets .......................................... 106
  5.4 Conclusion ..................................................... 109

6 Modelling dynamic hematopoietic regulation 110
  6.1 Introduction ................................................... 110
  6.2 Conventional model of hematopoietic regulation ............. 111
    6.2.1 Linear feedback ............................................ 113
    6.2.2 Saturating feedback ......................................... 113
  6.3 Dormant HSC model of hematopoietic regulation ............. 114
    6.3.1 Linear feedback ............................................ 115
    6.3.2 Saturating feedback ......................................... 116
  6.4 Dynamical analysis ............................................. 117
    6.4.1 Bifurcation analysis ......................................... 117
6.4.2 Elasticities: a measure of hematopoietic regulation effectiveness . . 122
6.4.3 Derivation of elasticities . . . . . . . . . . . . . . . . . . . . . . . 122
6.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
6.5.1 Simulations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
6.5.2 Bifurcation diagrams . . . . . . . . . . . . . . . . . . . . . . . . . . 126
6.5.3 Evaluating model elasticities . . . . . . . . . . . . . . . . . . . . . . 129
6.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133

7 Summary and outlook . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135
7.1 Reflections . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139

A Acronyms and definitions . . . . . . . . . . . . . . . . . . . . . . . . . . 141
A.1 List of acronyms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 141
A.2 Biological nomenclature . . . . . . . . . . . . . . . . . . . . . . . . . . 142

B Grid deployment and example files . . . . . . . . . . . . . . . . . . . . . 145
B.1 Grid deployment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 145
B.1.1 EGEE Grid architecture . . . . . . . . . . . . . . . . . . . . . . . . 145
B.1.2 JDL design and job types . . . . . . . . . . . . . . . . . . . . . . . . 147
B.1.3 gLite data management . . . . . . . . . . . . . . . . . . . . . . . . 149
B.2 JDL file . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
B.3 Repast parameter file . . . . . . . . . . . . . . . . . . . . . . . . . . . . 154

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155
List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 162
Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 165
Chapter 1

Introduction

In this thesis we describe the methodology and results of various computational models on hematopoietic stem cell (HSC) division kinetics and dynamic hematopoietic regulation. The therapeutic potential of stem cells due to their ability to build and maintain tissues and organs is widely recognised. Fuelled by intense public and media interest, stem cell research has experienced unprecedented growth in recent years and in response research groups in the field are becoming more and more multidisciplinary. As such, the work here investigates HSC biology from a systems biology perspective. Whilst classical bioinformatics is primarily concerned with the *structure* of biological components (like DNA sequences), systems biology studies the *function* and *interaction* of biological components (Weston & Hood, 2004) and therefore allows us to model and simulate dynamic and emergent behaviour of biological systems.

Despite the enormous potential of stem cells to revolutionise 21st century healthcare, biologist still understand very little of how stem cells function and clinical stem cell therapy has not progressed much further than bone marrow transplantation which was pioneered in the 1950s through to the 1970s. Two of the reasons why progress has been slower than expected is the sheer volume of data being generated by new high throughput experimental technologies like microarrays and mass spectrometry, and the inherent complexity of the processes that control within-cell and between-cell interactions in stem cell systems. In fact, this is a general phenomenon in most molecular biology fields and one of the main reasons that led to the emergence of systems biology. Whilst the bulk of the work in systems biology studies biochemical and molecular dynamics, the focus here is mostly on cellular level dynamics. Before we motivate the importance in having a good understanding of cellular division kinetics, we will first give an overview of the necessary biological background.
1.1 Biological background

This section briefly describes the biological background that is needed to have an appreciation for the problems encountered and benefits gained by the computational models in this thesis. Alberts et al. (2002) gives an in depth treatment of most of the concepts that are related here, and relevant biological nomenclature is defined in Appendix A.

1.1.1 Central dogma of molecular biology

Molecular biology mainly concentrates on the dynamics among the several entities and pathways of a cell, namely the interacting and regulatory relationships among Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and proteins. The flow of information in a cell, known as the central dogma of molecular biology (Crick, 1970), can in its most basic form be depicted as

\[
\text{DNA [transcription]} \rightarrow \text{RNA [translation]} \rightarrow \text{Protein [folding]} \rightarrow \text{Structure}
\]

Here actions that transfer sequence information from one biopolymer to the next is depicted in square brackets. Below we summarise the main components:

**DNA**

DNA is the main information carrier molecule in a cell and is formed by a chain of four different small molecules called nucleotides: adenine, guanine, cytosine and thymine. These nucleotides are called bases and a DNA chain is represented by a sequence of the initial letters of the nucleotides that builds the chain: A, C, G and T. The double helix structure that DNA assumes was discovered by James Watson and Francis Crick at the University of Cambridge (Watson & Crick, 1953) and provides a perfect medium for information storage. The two strands of the double helix are complementary since A can only bind to T, and C only to G and hence forms a stable structure. Consequently only one strand is needed for storing all the genetic information since each strand fully determines the other. One strand ends in a terminal phosphate group and is referred to as the 5’ (5-prime) strand whilst its complementary strands ends in a terminal hydroxyl group and is referred to as the 3’ (3-prime) strand.

**Genes.** A small number of sub-sequences in a DNA sequence contain the information to build specific proteins. These sub-sequences are called genes and they correspond to units of inheritance. The regions of DNA which regulates the behaviour of genes (when in time and space and to what amplitude they are expressed) are called transcription
factor binding sites. The sections of DNA that contain the actual gene information are also referred to as coding regions and it is these regions that are transcribed into RNA.

**Chromosomes.** At certain stages of the eukaryotic cell cycle DNA is packaged by DNA-bound proteins (chromatin) into organised structures called chromosomes. During the first stages of cell division the chromatin strands gradually become more compact to form chromatids causing the separate chromosomes to become visible. These structures are highly condensed, four armed, and consist of paired sister chromatids attached to each other at a point known as the centromere. Ploidy is a term that refers to the number of complete sets of chromosomes in an organism. Human cells are diploid as they contain two sets of chromosomes, with the exception of sex cells which are haploid. After cell division, chromatids uncoil themselves to function as chromatin and hence transcription can take place again.

**RNA**

Transcription is the process where DNA sequence information is transcribed into RNA. This event is also known as RNA synthesis and is facilitated by the enzyme RNA polymerase. RNA is single stranded as opposed to the double helix structure of DNA and is formed by the same nucleotides, except for thymine which is replaced by uracil. RNA has many roles but is mostly important for protein synthesis. Here it carries the blueprint information from DNA in the form of messenger RNA (mRNA) to structural entities called ribosomes. These ribosomes are a type of reader that relates the information carried by mRNA into the production of important proteins.

**Proteins**

Proteins are the workhorses of the cell, they are macromolecules that serve as building blocks for cellular structures and carry out all the important functions of a cell: gene activation, transcription, translation, regulating the cell cycle, etc. Amino acids are the building blocks of proteins and there are twenty-two standard types of them. mRNA is coded in such a way that every three nucleotides (called a codon) corresponds to a specific amino acid. Synthesised amino acid chains fold onto themselves after being translated and it is believed that the function of a protein is determined by these resulting tertiary three dimensional structures.
1.1.2 Stem cells

There is a constant need for new cells in humans and animals. This is due to normal growth and development (e.g. muscle tissue which can be built up according to the body’s needs) and maintenance of tissues and organs (e.g. blood and skin that need continuous cell replacement or bone that needs repair after disturbance or injury). New cells are generated through cell division, but specialised cells (like blood, muscle and skin cells) are unable to divide and replicate themselves - they have to be replenished from populations of stem cells. Although the exact manner by which stem cells should be defined is still undecided, the functional definition of stem cells as cells that can self-renew (produce copies of themselves) or differentiate into mature specialised progeny (cells that the body needs to grow and maintain itself) is the most widely accepted (Potten & Loeffler, 1990). Currently four types of stem cells are distinguished:

**Embryonic stem cells (ESCs),** found from the inner cell mass of the blastocyst (very early embryo), are pluripotent - i.e. they have the potential to differentiate into most of the cell types of the adult organism. Being pluripotent gives ESCs a huge therapeutic potential but since the extraction of ESCs sacrifices the donor embryo, any use of human ESCs in research is a sensitive issue and raises many ethical questions.

**Induced pluripotent stem cells (iPSCs)** are artificially created pluripotent cells by genetically reprogramming an adult somatic cell to the state of an ESC (Yates et al., 2007; Takahashi et al., 2007). This is done by manipulating the adult cell to express certain genes and factors critical to the defining properties of an ESC, forcing the cell back to the pluripotent state. iPSCs are regarded as a major scientific advance but the true nature of the relation between induced pluripotent and pluripotent stem cells are not yet clear.

**Adult stem cells (ASCs)** are multipotent cells - i.e. they have the potential to make only a limited range of cell types in the body. ASCs are found in various regenerative tissues like the skin, blood, bone marrow, and gut. They have the capacity to provide both life-long self-renewal and to generate all the terminally differentiated cell types of each tissue. Research on ASCs is not plagued by any ethical issues but rather by difficulty in identifying and culturing them in vitro.

**Cancer stem cells (CSCs)** are cancer cells which possess the distinct self-renewal and differentiation characteristics associated with normal stem cells. There is growing evidence that CSCs are at the root of carcinogenesis as well as relapse and metastasis (Alison et al., 2007), however it is still not certain whether CSCs are malignant stem
cells or somatic cells which acquired stem cell properties. CSCs are key drug targets in cancer therapy research with some promising results recently published (Gupta et al., 2009).

### 1.1.3 Hematopoietic stem cells

In healthy adult multicellular organisms daily production and loss of millions of blood cells due to normal cellular turnover are in equilibrium, a condition called homeostasis. Even under conditions of hematopoietic stress, like a sudden loss of blood or a toxic insult, homeostasis is rapidly restored. A small number of tightly regulated HSCs in the bone marrow are responsible for both maintenance of homeostasis and restoration after injury. HSCs are thus crucial for maintaining lifelong production of all blood cells. Functional proof of the existence of adult HSCs is shown through long-term reconstitution of the hematopoietic system of a lethally irradiated mouse by transplantation of a single purified HSC (Domen & Weissman, 1999). Due to the technological advances provided by flow cytometry, the existence of multiple monoclonal antibodies directed to stem cell specific cell surface antigens, and *in vitro* and *in vivo* assays that can quantitate their functional capacity, mouse bone marrow HSCs are amongst the most well characterised (both phenotypically and functionally) adult stem cells.

#### The HSC niche

Unfortunately in spite of being so well characterised, the future therapeutic potential of HSCs are impeded by the current failure to successfully expand and culture HSCs *in vitro*. The HSC niche refers to the bone marrow microenvironment of these cells that is thought to regulate HSC self-renewal and differentiation (Jones & Wagers, 2008). It has been postulated that *in vitro* expansion of HSCs will only be possible when a three-dimensional niche can be reconstructed *ex vivo* (Wilson & Trumpp, 2006). Because HSCs exhibit such different behaviour *in vitro*, much emphasis is placed on *in vivo* techniques and data.

#### The hematopoietic hierarchy

The hematopoietic hierarchy is a conceptualisation that views HSCs as losing their self-renewal potential as they progress down the hierarchy by becoming more and more specialised and developed. The resulting tree-like structure (Figure 1.1) determines the possible properties and developmental stages of daughter cells after division. The leaves represent terminally differentiated cells which have lost their ability to self-renew or differentiate.
1.1. Biological background

Figure 1.1: The hematopoietic hierarchy (adapted from Bryder et al. (2006)). The hierarchical organisation of the hematopoietic system has long been recognised, with rarely-dividing multipotential HSCs producing rapidly dividing lineage-restricted transit-amplifying and committed progenitors which in turn will give rise to all differentiated cell types of the blood. HSCs with the strongest self-renewal potential is at the top and each successive node towards the bottom of the tree represents stages of differentiation. The proliferative index indicates division rate and assumes a water droplet shape with the leave nodes (mature cells) not proliferating at all.

In practice each node in the hierarchy can be identified by a set of cell surface markers. Although there is mostly consensus about what the exact markers for each cell population (the nodes) should be, there is sometimes conflict of opinion in one or two molecules - especially at the higher end of the tree. In this thesis (apart from Chapter 2 which uses a different dataset) we regard the Lin⁻Sca1⁺cKit⁺CD150⁺CD48⁻CD34⁻CD135⁻ set of markers as referring to phenotypic HSCs (the topmost node in the hierarchy) that comprise around 0.001% of mouse bone marrow (Wilson et al., 2007)). This marker set was also used when HSCs were isolated for the main dataset introduced in Chapter 3.
1.1.4 Cell division (mitosis)

The main organelles of the eukaryotic cell are its membrane, cytoplasm, mitochondria, and nucleus which contains the hereditary information (DNA). Mitotic cell division (Figure 1.2) is the process by which an eukaryotic cell duplicates its DNA and transfers it to two daughter cells. Mitotic cell division is performed by going through a series of phases called the cell cycle. These phases are known as G\(_1\) (growth 1), S (synthesis), G\(_2\) (growth 2) and M (mitosis). If this cycle is temporarily or permanently halted the cell would have entered a state of quiescence referred to as the G\(_0\) phase.

**Figure 1.2: Mitotic cell division.** This is the process through which an eukaryotic cell duplicates itself. Firstly the genetic material is copied and doubled. Mitosis follows, during which the cell’s duplicated chromosomal material is separated and moved to opposites sides. Cytokinesis completes the process by dividing the cytoplasm to construct the two individual daughter cells.

After completion of the previous M phase, the cell cycle enters the G\(_1\) phase during which the various enzymes that are needed for DNA replication are produced. G\(_1\) is followed by the S-phase which is marked by the start of DNA synthesis and ends with the successful replication of all the chromosomes. Each chromosome now comprises two sister chromatids and even though the amount of DNA in the cell has doubled, the ploidy of the cell remains unchanged. The G\(_2\) phase is reached at the end of the S-phase and ends itself when the cell enters mitosis. As in the G\(_1\) phase, protein synthesis commences on a large scale and mainly results in the production of microtubules which are essential for mitosis. The M phase consist of two inter-dependable processes which are known as mitosis and cytokinesis. Mitosis refers to the formation of two daughter nuclei through the separation of the duplicated chromosomal material in the original cell nucleus. This separation is achieved by firstly aligning the new sister chromatids constructed during the S-phase along the centre of the cell. Microtubules produced during the G\(_2\) phase then attach themselves
to the centres of these sister chromatids (known as the centromeres) and begin to shorten, pulling the individual sister chromatids in opposite directions. The duplicated genetic material is thus halved resulting in two genetic identical cells. Cytokinesis commences directly after the completion of mitosis, in which the nuclei, cytoplasm, organelles and finally the cell membrane are divided into two distinct daughter cells. Therefore the M phase as a whole is the process in which the cell’s genome is physically divided into two genetically identical daughter cells.

**Symmetric and asymmetric stem cell division**

Conceptually a distinction is made between two possible division types based on the fate of the two daughter cells after division. Asymmetric division happens when one daughter cell retains the stem cell properties and the other one differentiates to a more mature cell lineage. When both daughter cells remain stem cells or differentiate it is considered to be a symmetric division. In principle homeostasis can ideally be maintained by asymmetric division exclusively, however symmetric division is required for expansion (and hence transplantation assays) to be possible. Experimental evidence for both types of division has been found (Takano et al., 2004; Lechler & Fuchs, 2005) and some studies suggest that the division type plays an important role in development and cancer (Morrison & Kimble, 2006; Dingli et al., 2007).

**1.2 Measuring cell division kinetics**

Experimental techniques for measuring cell division rates and dynamics take an important role in this thesis. We therefore present some background on the most common approaches. A marker found naturally in cells that can be used to estimate division rates is telomere length (De Boer & Noest, 1998). A telomere is a short repeated DNA sequence at the end of a chromosome and its length is reduced with each cell division. The state of art biotechnology, however, for measuring *in vitro* or *in vivo* turnover kinetics of cells is to use some form of biochemical marker that labels cells once they divide. The biochemical marker is typically modified in some way by any of the phases of the cell cycle and hence acts as a counter (in the true sense of the word) of mitotic events. These techniques can be broadly classified as DNA labelling or non-DNA labelling.

**1.2.1 DNA labelling**

As the name suggests, a DNA labelling biochemical marker is incorporated into strands of newly synthesised DNA during cell division. The most commonly used techniques in this
category uses bromodeoxyuridine (BrdU), deuterated glucose ([\(^2\)H]glucose), or tritiated thymidine ([\(^3\)H]dT).

**BrdU**, introduced by Gratzner (1982), is a synthetic nucleotide and an analogue of thymidine (a combination of thymine and deoxyribose) and can thus be incorporated in synthesised DNA by substituting for thymidine (Figure 1.3). Fluorescently marked antibodies that attaches to BrdU are used to detect cells that are BrdU\(^+\). BrdU is usually applied by adding it to the drinking water of animals and/or by intraperitoneal injection. When BrdU\(^+\) cells divide without BrdU present, their label is diluted. Tracking the proportion of BrdU\(^+\) cells during both the uptake and loss period provides a mechanism to measure the turnover kinetics of a given population of cells.

\[^2\)H]glucose\ is the only division tracking compound that is safe for human use since it is a stable isotope that labels the pentose sugar (as opposed to the actual base like BrdU) of nucleotides synthesised during the labelling period (Hellerstein & Neese, 1992). Labelled nucleotides have a different mass than unlabelled ones and hence the ratio of labelled to unlabelled nucleotides can be measured using mass spectrometry. The fraction of labelled DNA in a population of cells is thus measured instead of measuring the fraction of labelled cells as is done for BrdU. Cellular death is the only means by which label can be reduced since the added mass of each originally labelled DNA strand remains constant whether scattered among any number of different daughter cells in post labelling divisions (Hellerstein, 1999).
[^3H]dT is a radioactive analogue of thymidine (Lebowitz & Rubinow, 1969; Cotsarelis et al., 1990) and thus works similar than BrdU by also substituting for thymidine.[^3H]dT label is however detected by using a liquid scintillation counter.

### 1.2.2 Non-DNA labelling

Non-DNA labelling techniques label cellular components other than DNA molecules like the cytoplasm or nuclear proteins of a cell. Whilst the DNA labelling techniques described above only label cells that have divided during the labelling period (this can be viewed either as an advantage or disadvantage, depending on your perspective and purpose of study), cells are randomly labelled (not only those that divide) by non-DNA labelling techniques. Quantification of division rates is only done during the post labelling period by tracking the change in label intensity as the label is segregated to daughter cells during division. Techniques in this category include chromosome damage, carboxyfluorescein diacetate succinimidyl ester (CFSE), and histone H2B-green fluorescent protein (H2B-GFP).

**Chromosome damage** by radiation treatment can be used to track cell division history since damaged chromosomes are either inherited by one of the daughter cells upon division, or cause the cell to die (Michie et al., 1992; Mclean & Michie, 1995). It goes without saying that this technique has not gained widespread use for measuring cell division kinetics.

**CFSE** is a fluorescent dye that divides equally between daughter cells upon division of a stained cell (Lyons & Parish, 1994). When cells are initially stained in culture, they can be adoptively transferred back to the host where cells that have undergone the same number of divisions can later be observed in cohorts of approximately log-normally distributed intensities. From this intensity profile the division history of populations of cells which have undergone multiple rounds of divisions can be studied and division rates estimated (Hasbold et al., 1999; Gett & Hodgkin, 2000; Pilyugin et al., 2003; Ganusov et al., 2005; De Boer et al., 2006; Yates et al., 2007). The standard labelling protocol (*ex vivo* labelled cells transferred back to the host) place enormous strain on cells which could alter their kinetics and thus place doubts on how useful CFSE results are in practice. When initial labelling is done *in vivo*, staining is heterogeneous so that the distinctly observable cohorts of intensities are subsumed into a single unimodal distribution (Asquith et al., 2006). CFSE are thus not very useful as an *in vivo* labelling technique.

**H2B-GFP** is a nuclear protein label and also the most recent technique for measuring cell division rates (Tumbar et al., 2004; Wilson et al., 2008; Foudi et al., 2009).
1.3. Objective

The use of H2B-GFP however requires transgenic organisms that firstly express the GFP protein, and secondly have some regulatory system that can turn GFP reporter gene expression on or off. During “labelling” GFP expression is turned on and cells randomly produce fluorescent proteins. When GFP expression is turned off, no new fluorescent proteins are produced and the existing ones are diluted equally between daughter cells upon division. Similar to CFSE, H2B-GFP permits the isolation and functional assessment of live (labelled) cells (not possible with the other labelling techniques), and in addition H2B-GFP allows for in vivo labelling of cells.

1.3 Objective

The general purpose of modelling in systems biology as stated by Wilkinson (2006) can be summarised as aiming to:

1. Test whether our understanding of a system is correct.
2. Validate whether the implications of the model is consistent with observed experimental data.
3. Use the model predictively to do ”virtual experiments“.
4. Guide the design and analysis of new biological experiments.

With these aims in mind, the main objective of this thesis is to elucidate key questions surrounding the kinetics of HSC division in particular, and hematopoietic regulation in general, through the use of computational models and sound reasoning. The largest and foremost part of the work presented here focusses on modelling of experimentally observed cell labelling data in order to ascertain HSC division rates and dynamics. This is followed by a purely conceptual model (i.e. not designed to explain observed data) to shed further light on our new understanding of dynamic hematopoietic regulation.

1.3.1 Estimating HSC division rates and dynamics

A solid understanding of cell division rates and dynamics is fundamental for effective treatment of the most serious diseases of our time like cancer, diabetes, and human immunodeficiency virus (HIV). Most cell kinetic studies to date have focused on the immune system (Asquith et al., 2009), for example cell division tracking has been used to compare the proliferation and death rates of CD4$^+$ cells$^1$ in normal and HIV infected subjects (Mohri et al., 1998; Ribeiro et al., 2002). It was found that HIV infection induces

$^1$HIV causes loss of CD4$^+$ T lymphocytes on the long-term.
an increase in CD4$^+$ cell proliferation rates and hence eventual exhaustion. However when the effects of antiretroviral therapy were studied, contradicting results have been reported. One study found that antiretroviral therapy increases T lymphocyte division rates (Hellerstein et al., 1999) whilst the other reported a decrease (Mohri et al., 2001), highlighting the need for careful design of mathematical models (Asquith et al., 2002). Another example is the work by Michie et al. (1992) and Mclean & Michie (1995) on the lifespan of naive and memory T-cells which (surprisingly) revealed that memory T-cells divide much more rapidly than naive T-cells. This study dramatically altered the view of immune memory maintenance which in turn influenced vaccine design.

The experimental data on which modelling in this thesis is based are all results of in vivo BrdU label tracking in HSCs\(^2\). We will demonstrate how quantifying HSC division kinetics can help test hypotheses about the population biology of HSCs. There are many different approaches to modelling complex systems like that found in biology, and often individuals tend to advocate only one type of model for specific systems. Instead we follow a comprehensive modelling approach. Based on two related but independent datasets we define mathematical models that are discrete, continuous, stochastic, or deterministic. Each model has its own strengths and weaknesses and with each we learn something different about the data. All support the same hypothesis though and all are defined by static parameters. Our comprehensive approach thus concludes with a dynamic model that evaluates the findings of the static models on the concept of hematopoietic regulation - independent of experimental data.

### 1.3.2 Modelling dynamic hematopoietic regulation

As mentioned in Section 1.1.3 bone marrow HSCs are responsible for both lifelong daily maintenance of all blood cells and for repair after cell loss caused by infection or toxic insults. Until recently the cellular mechanisms by which HSCs accomplish these two very different tasks remained an open question since they must seldom replicate to avoid accumulating DNA mutations, replicative senescence or exhaustion (Fuchs, 2009). Although HSCs have been shown to be predominantly in a transient resting state of cell cycle and are therefore thought to divide infrequently, it has always been assumed that this is a stochastic process with the entire HSC pool turning over every few weeks. Indeed, the earliest studies estimated the doubling time of individual HSCs to be between 17.8 and 30 days with the entire HSC pool turning over every 57 days (Bradford et al., 1997; Cheshier et al., 1999). Thus the conventional view was that despite their relative transient quiescence, all HSCs nevertheless regularly entered and exited the cell cycle.

\(^2\)Incidentally HSCs are precursors to the lymphocytes mentioned above, thus also part of the immune system.
The dormant hematopoietic stem cell hypothesis

Alternatively, biological evidence has recently been found for the dormant HSC hypothesis (Wilson et al., 2008; Raaijmakers & Scadden, 2008; Foudi et al., 2009) which postulates the existence of two related mouse HSC populations. First, a dormant HSC (d-HSC) population which harbours the highest self-renewal potential of all blood cells but is only induced into active self-renewal in response to hematopoietic stress, returning to a resting state once homeostasis has been restored. And second, an active HSC (a-HSC) subset that by and large produces the progenitors and mature cells required for maintenance of day-to-day hematopoiesis. The implications of this hypothesis is far reaching and seriously challenge the conventional view of HSCs and the hematopoietic hierarchy (Figure 1.1). Of particular interest is the implications for cancer treatment (Essers et al., 2009; Viale & Pelicci, 2009) since recent evidence suggests that some cancer stem cells are also dormant, providing an explanation for their remarkable resistance to chemotherapy.

In this thesis we investigate the d-HSC concept from a computational point of view. The d-HSC hypothesis thus stands central in the work presented here. Firstly, our analysis on the experimental BrdU data predicts that at least a third of the HSC population are d-HSCs that divide about once every 149-193 days (about five times in the lifespan of a mouse) with a-HSCs dividing once every 28-36 days. Secondly, a d-HSC model of dynamic hematopoietic regulation shows that malfunctioning d-HSCs are likely to play a central role in hematological diseases. Based on a carefully defined mathematical measure, we can provide well-grounded arguments that d-HSC regulation of hematopoiesis is evolutionarily more likely than conventional (d-HSC free) regulation of hematopoiesis.

1.3.3 Assumptions

As mentioned above our comprehensive modelling approach includes a number of different model types. As a result the assumptions on which each model is based might be slightly different. These model-specific assumptions like division type (symmetric or asymmetric) and rate of label uptake or loss will be clearly indicated when each model is described. All models are however based on the set of universal assumptions below and all our conclusions in this thesis are made in light of these.

Firstly, the underlying structure of each model is designed around the hematopoietic hierarchical view (Figure 1.1) and thus implicitly assumes the hierarchy is a valid conceptualisation of the real biological system. There have been alternative proposals and models that criticise the hierarchical (or compartmental) approach though. Kirkland (2004) proposed a two-dimensional phase space model of hematopoiesis. The $x$ and $y$ coordinates of
the plane represent the differentiative state of the cell (as opposed to the discrete nodes of a hierarchical tree), and movement of cells on the plane is described by a probability density function. Another example is the single cell-based Monte-Carlo model which Röder (2003) developed around the concept of within-tissue plasticity. In this model cell fate rely on dynamically controlled switches between two growth environments, one of which represents the stem cell niche. The HSC hierarchical view remains, however, by far the most practical and widely accepted conceptualisation by both biologists and model builders.

Lastly it is important to note that the computational work in this thesis is strongly tied to the current biological understanding and constraints of the hematopoietic system. Accordingly we only consider two classes of models (the conventional d-HSC free models and the d-HSC models) and compare which of these seems more consistent with empirical data and produces the most realistic output. From a purely mathematical perspective many other modelling classes and architectures that explain the observed phenomena as effective could be envisaged. However, to our knowledge the two classes of models evaluated here represent the most plausible biological scenarios as dictated by current biological knowledge. If new experimental evidence comes to light which dramatically changes the current view and understanding of HSC population biology and their plausible dynamics, alternative models might proof more suitable and hence possible different conclusions might be derived. However, currently the d-HSC concept is gaining widespread acceptance in the biological community (Raaijmakers & Scadden, 2008; Sarrazin et al., 2009; McKay et al., 2009; Benveniste et al., 2010) and similar conclusions to what is reported here were made by a another group using different experimental models and protocols (Foudi et al., 2009).

1.4 Publications

The following published articles and posters have a direct relevance to the material presented in this thesis:


The following articles and posters are indirectly related to the work presented in this thesis:


\(^3\)At the time of writing this paper received a Faculty of 1000 Biology rating of 9.8 (Exceptional), see http://f1000biology.com/article/id/1144922.
1.5 Outline

This thesis is organised as follows:

Chapter 2 presents a stochastic agent-based model of BrdU kinetics. This model is applied to a dataset that was used to study the chromosome segregation mechanism of HSCs when they divide. We demonstrate how its discrete properties provides desirable flexibility to the agent-based model and enable it to describe the observed data much better than a previously published continuous model. Although the focus is not primarily on the d-HSC hypothesis, related concepts like heterogeneity in the HSC population do emerge here.

Chapter 3 introduces a new much more comprehensive BrdU dataset in which long-term label retention is apparent. Here the focus is primarily in gauging the support of the data for the d-HSC hypothesis and since accurate parameter estimation is of utmost importance, an ordinary differential equation (ODE) model is developed. The model equations are analytically tractable since they are all linear and hence we are able to optimise parameters very effectively. However, compared to the agent-based model, much more work are required to model different HSC population structures with the continuous deterministic model described in this chapter.

Chapter 4 models label-retaining cell dynamics as a stochastic kinetic process. The aim is to estimate the variation for the parameters optimised by the ODE model in Chapter 3 in order to explore whether stochastic variation in HSC division kinetics can explain the observed label retention, rather than a slowly cycling d-HSC subset. The bulk of this chapter is focused on the theoretical derivation of a stochastic model whose predictions on average matches that of the previously defined ODE model.

Chapter 5 remodels the data presented in Chapter 3 with the agent-based model defined in Chapter 2 to compare the parameter estimates to that of the ODE model. But, the only option to optimise parameters for the agent-based model is a brute-force search approach. We therefore define a parallel search strategy to explore the parameter space more effectively. The technical details of deploying this strategy on a Grid supercomputer is explained in this chapter.

Chapter 6 introduces models of dynamic hematopoietic regulation. The purpose here is to assess the qualitative nature of d-HSC hematopoiesis as opposed to the conventional understanding of hematopoiesis, and to quantify how effective each model is in hematopoietic regulation. Elementary bifurcation theory is presented to study
conditions that results in oscillatory behaviour of the d-HSC model. In addition we define a mathematical measure to quantify the effectiveness of each model at re-establishing homeostasis after injury.

Chapter 7 summarises the work presented in this thesis and likewise underlines how the general aims of modelling in systems biology have been achieved. In addition potential future directions are pointed out.
Chapter 2

Agent-based model

2.1 Introduction

In Chapter 1 we presented a number of biochemical compounds that can be used to label dividing cells, one of which was BrdU. Tracking the proportion of BrdU+ cells during both the labelling (monotonic increase) and de-labelling (monotonic decrease) period provides a mechanism to measure the in vivo division rates of a given population of cells. BrdU data can however be misleading when interpreted directly and a number of pitfalls awaits the unprepared (Hellerstein, 1999; Perelson, 2002). Much more certainty about turnover rates can be gained by using computational and mathematical models to fit observed BrdU dynamics (Mohri et al., 1998; Cheshier et al., 1999; Bonhoeffer et al., 2000; MacKey, 2001).

Because it is a DNA labelling technique and hence labels the chromosomes of a cell, BrdU could provide information about chromosome segregation in addition to cell division rates. Chromosome segregation refers to the way in which chromosomes segregate between the two daughter cells during mitosis. Two opposing theories exist: random segregation or asymmetric segregation. The asymmetrical segregation or “immortal strand hypothesis” was first proposed by Cairns (1975) and suggests that adult stem cells retain older DNA strands during mitotic cell division and newly synthesised DNA strands is asymmetrically segregated to differentiating daughter cells (Figure 2.1). The older DNA strands act as templates for all divisions thereby providing a mechanism to limit the accumulation of DNA mutations.

A recent study by Kiel et al. (2007) uses an ODE-based model to fit BrdU data on murine HSCs. The authors use the result of their model to refute the asymmetrical segregation of chromosomes hypothesis in favour of the random segregation hypothesis. In this chapter we remodel the Kiel et al. BrdU dataset as a first introduction to BrdU and HSC turnover kinetics. We develop an agent-based model of BrdU dynamics and evaluate its performance in relation to the ODE model. Our approach is to model single cells as
Figure 2.1: **Chromosome segregation.** Observed BrdU dynamics in stem cells (encircled with black) would be different for asymmetric and random chromosome segregation as depicted here (BrdU$^+$ DNA strands indicated with light green, immortal strands in red). In the asymmetric case BrdU intensity remains constant after the first division during uptake and cells loose all labelling after the first chase division. In contrast, many different patterns are possible for both uptake and chase in the random segregation case.

well as their chromosomes as agents which make probabilistic decisions over fixed intervals of time. Our results demonstrate the advantages of the agent-based approach in terms of being stochastic, discrete and flexible to easily model heterogeneous cell populations.

This chapter proceed as follows. We first describe the equations of the Kiel et al. model in Section 2.2 where we propose alternative uptake equations for the random chromosome segregation case (Section 2.2.4). All the equations are solvable in closed form and their solutions are given in Section 2.2.5. In Section 2.3 we present a description of a stochastic agent-based BrdU model. We use this model to repeat the study of Kiel et al. in Section 2.4 as a case study to compare the results of the two approaches. Most of the material in this chapter have been published in van der Wath & Liò (2008).

## 2.2 Ordinary differential equation model

We have mentioned above that information can be learned by studying the rate of BrdU incorporation or dilution. Hence it logically follows that rate equations, otherwise known as ODEs should be a suitable approach to model BrdU dynamics. It’s no coincidence that most published models of BrdU data to date are ODE-based (Mohri et al., 1998; Bonhoeffer et al., 2000; De Boer et al., 2003). The model of Kiel et al. too consists of
2.2. Ordinary differential equation model

sets of coupled ODEs that describe the rate of change of the fraction of BrdU labelled cells over time. Different sets of equations describe the dynamics during BrdU uptake and loss for the random and asymmetric segregation case separately. The experimental data of Kiel et al. are however markedly different than that of Mohri et al., Bonhoeffer et al., and De Boer et al.. We will thus first describe the Kiel et al. dataset before we proceed to describing the actual model.

### 2.2.1 Experimental data

Adult mice were given BrdU through injection as well as in their drinking water for 10 days, followed by a 120 day chase period (BrdU free). The percentage of BrdU$^+$ lineage$^{-}$Sca-1$^+$c-kit$^+$CD150$^+$CD48$^-$CD41$^-$ HSCs were tracked using immunofluorescence microscopy during uptake at day 1, day 4, and day 10, and during chase at day 40, day 70, and day 120 (Table 2.1).

Note that cells have to be fixed\(^1\) in order to measure their BrdU intensity, so it is not possible to track the HSC BrdU uptake and dilution of a single mouse. Each mouse can be evaluated at only one time point. In contrast Mohri et al., Bonhoeffer et al., and De Boer et al. measured BrdU labelling in T-cells from peripheral blood samples and could track label intensity in individual animals.

### 2.2.2 Model assumptions

It is assumed that HSCs proliferate through asymmetric division only at a rate $\alpha$ which was estimated by Kiel et al. to be 6% per day or about once in every 17 days. BrdU is removed at day $T$ when the chase period starts. During chase, for random chromosome

---

\(^1\)Frozen and hence killed.
segregation, it is assumed that BrdU intensity is more or less halved with each round of division. Since T-cells are relatively short lived, Mohri et al., Bonhoeffer et al., and De Boer et al. assumed that once a cell has been labelled it remains so until it dies. HSCs are however much longer lived and the rate of apoptosis is assumed to be negligible. The total stem cell population thus remains constant and since cellular death (apoptosis) is not accounted for in this model, the only way for an HSC to go from BrdU\(^+\) to BrdU\(^-\) is to dilute its label below a detectable intensity through multiple divisions. The number of divisions \(N\) to dilute incorporated BrdU to below the level of detection needs to be estimated by the model.

**Labelled DNA strands**

Albeit on a less dramatic scale, different label intensities are also possible during BrdU uptake for random chromosome segregation. Cells that divide more in the presence of BrdU will have a higher label intensity and will obviously dilute the label over more chase divisions. Kiel et al. modelled this phenomenon in terms of the number of labelled DNA strands and distinguished between cells with one strand BrdU\(^+\) (after one division) and cells with both strands BrdU\(^+\) (after two or more divisions) in their equations. Their approach effectively regard cells as having 100% of its DNA labelled after only two divisions in the presence of BrdU - a somewhat naive assumption. We therefore introduce alternative uptake equations in Section 2.2.4 which (biologically) seems to give more realistic results. It turns out that the relationship between the BrdU detection threshold (BDT) and number of uptake divisions is much more complex than assumed by either version of uptake equations used here, so we will revisit the relationship in Chapter 3 where we analyse it much more extensively.

### 2.2.3 Asymmetric chromosome segregation equations

When chromosomes segregate asymmetrically the possible distribution of label intensities during both uptake and chase are few and well defined (Figure 2.1). The differential equations are given below.

**Uptake equations**

When an unlabelled cell divides one of its DNA strands (the newly synthesised one) incorporates BrdU and hence is labelled. If a labelled cell divides a second time its label intensity will remain the same because the old unlabelled strand is asymmetrically retained and the old labelled strand is simply replaced by a newly synthesised labelled strand. The rate of change of the proportion of BrdU\(^+\) cells at time \(t\) during BrdU application, \(y_1(t)\),
2.2. Ordinary differential equation model

is thus equivalent to the proportion of BrdU\(^{-}\) cells that has divided at time \(t\), \(y_0(t)\):

\[
\frac{dy_1}{dt} = \alpha y_0 .
\]  

(2.2.1)

There is an implicit assumption here (discussed in more detail in Chapter 3) namely that one division in the presence of BrdU labels a cell sufficiently for it to be detected as BrdU\(^{+}\). Because cells leave the \(y_0\) population when they divide, the equation for \(y_0\) in turn is

\[
\frac{dy_0}{dt} = -\alpha y_0 .
\]  

(2.2.2)

The initial conditions are \(y_0(0) = 1.0\) and \(y_1(0) = 0\) since all cells are initially unlabelled.

Chase equations

Modelling loss of labelling when BrdU is removed at day \(T\) is rather straightforward as well since only one DNA strand can take up and consequently lose BrdU for the asymmetric segregation case. All BrdU\(^{+}\) cells will thus lose their labelling after a single division, so the rate equation is:

\[
\frac{dy_{10}}{dt} = -\alpha y_{10} ,
\]  

(2.2.3)

where \(y_{10}\) represent the fraction of cells with one labelled strand that hasn’t divided yet after day \(T\). The initial condition is \(y_{10}(0) = y_1(T)\).

2.2.4 Random chromosome segregation equations

Our discussion so far already subtly hinted that more care has to be taken when modelling the random chromosome segregation case. Firstly, BrdU uptake modelling now has to account for more than one division in the presence of BrdU since both DNA strands of a chromosome can take up BrdU. In addition the uptake predictions at time \(T\) determines the initial conditions for chase which implies that we have to define more than one set of chase equations (see below). Moreover, BrdU dilution is a more complex process since the model also needs to take the BDT into account.

Uptake equations

As discussed the uptake equations need to predict both the proportion of cells with 1 labelled strand and 2 labelled strands during BrdU uptake. For this purpose Kiel et al. introduced \(y_1\) to represent the fraction of cells with one DNA strand BrdU\(^{+}\) after only
one division and $y_2$ to represent the fraction of cells with both DNA strands BrdU$^+$ after two or more divisions. The uptake equations now define a two step process:

$$\frac{dy_0}{dt} = -\alpha y_0$$

(2.2.4)

$$\frac{dy_1}{dt} = -\alpha y_1 + \alpha y_0$$

(2.2.5)

$$\frac{dy_2}{dt} = \alpha y_1.$$  

(2.2.6)

Note that the total labelled proportion $\frac{dy_0}{dt} + \frac{dy_2}{dt} = \alpha y_0$ simplify to Equation 2.2.1, the rate of uptake during asymmetric chromosome segregation. The initial conditions are $y_0(0) = 1.0$, $y_1(0) = 0$ and $y_2(0) = 0$.

**Chase equations**

Two sets of coupled equations describe the kinetics during chase. With each division, as label intensity is halved, cells$^2$ “move” from equation $k$ to equation $k+1$. A cell with 2 labelled strands would need one extra division to be on the same label intensity as a cell with only 1 labelled strand. The equations for modelling this process are:

$$\frac{dy_{10}}{dt} = -\alpha y_{10}$$

$$\frac{dy_{11}}{dt} = -\alpha y_{11} + \alpha y_{10}$$

$$\vdots$$

$$\frac{dy_{1N}}{dt} = \alpha y_{1(N-1)}$$

(2.2.7)

$$\frac{dy_{20}}{dt} = -\alpha y_{20}$$

$$\frac{dy_{21}}{dt} = -\alpha y_{21} + \alpha y_{20}$$

$$\vdots$$

$$\frac{dy_{2N}}{dt} = \alpha y_{2(N-1)},$$

where $y_{ij}: i \in \{1, 2\}, j \in \{0, 1, ..., N\}$ is the fraction of cells with $i$ labelled DNA strand(s) that has divided $j$ times after day $T$. The initial conditions follows from the uptake

$^2$Strictly this should read ‘the fraction of labelled cells’. Interestingly all differential equations in this chapter, although describing continuous entities, can be directly applied to cell numbers instead of labelled cell proportions. Predictions for cell numbers can simply be divided by the (constant) total population size to get the proportions currently being modelled.
predictions at the time when BrdU is removed: \( y_{10}(0) = y_1(T) \) and \( y_{20}(0) = y_2(T) \). Labelled cells that have diluted their BrdU below detection are accumulated in \( y_{iN} \).

**Alternative uptake equations**

Careful study of Equation 2.2.6 reveals that it is based on an assumption that a cell has 100\% of its DNA labelled after only 2 uptake divisions. Even if a cell has only one chromosome this would happen only in one out of every two cases (depending on which daughter cell remains an HSC and which one differentiates). The data we model are from mouse cells which have 40 (\( 2 \times 20 \)) chromosomes each. When a cell cycles, its DNA (and hence chromosomes) is doubled during the synthesis phase. This means that during mitosis of a mouse cell 80 chromosomes are segregated between the two new daughter cells. If sister chromatids are individually unique by some measure (like the BrdU label status of each of its DNA strands for example), then there are \( 2^{40} \) different ways in which the chromosomes can segregate. For only one of these it is the case that all 40 chromosomes of the stem cell daughter will have both DNA strands labelled after two uptake divisions. Equation 2.2.6 thus quite likely over-predicts \( y_{20}(0) \), the proportion of cells that has both strands labelled when BrdU is removed (see Figure 2.5 in Section 2.4). An alternative approach is to rather let \( y_1 \) represent the fraction of cells that has divided once or twice and then letting \( y_2 \) represent the fraction of cells that has divided three or more times, making provision for an intermediate level of labelling between \( y_1 \) and \( y_2 \). Instead of modelling the number of labelled strands explicitly like Kiel et al., we define equations for two classes of labelling, allowing cells to “stay” in the first class for one round of divisions:

\[
\begin{align*}
\frac{dy_{11}}{dt} &= -\alpha y_{11} + \alpha y_0 \\
\frac{dy_{12}}{dt} &= -\alpha y_{12} + \alpha y_1 \\
\frac{dy_{2}}{dt} &= \alpha y_{12},
\end{align*}
\]  

(2.2.8, 2.2.9, 2.2.10)

where \( y_1 = y_{11} + y_{12} \). There is no need to define a third class in this study, since we will use \( T = 10 \) and \( \alpha = 0.06 \) (see Section 2.4) and hence we can expect few cells to divide four or more times during the 10 days of BrdU uptake. Cells that do divide four or more times are not lost but are retained in \( y_2 \) as is evident in Equation 2.2.10.

**2.2.5 Analytic solutions**

Fortunately all the equations above are linear and hence have analytic solutions. The asymmetric segregation equations in particular can be solved directly with

\[
y_0(t) = y_0(0)e^{-\alpha t} = e^{-\alpha t}
\]  

(2.2.11)
2.3 Agent-based model

\[ y_{10}(t) = y_{10}(0)e^{-\alpha t} = y_1(T)e^{-\alpha t}, \quad (2.2.12) \]

since both are describing standard exponential decay. The solution of \( y_1(t) \) is

\[ y_1(t) = y_1(0) + y_0(0) - y_0(0)e^{-\alpha t} = 1 - e^{-\alpha t} = 1 - y_0(t), \quad (2.2.13) \]

which turns out to be the Cumulative Distribution Function (CDF) of an exponential distribution with parameter \( \alpha \). This makes sense because \( y_1 \) effectively acts as a counter of the number of times cells in \( y_0 \) divide and as we will show in Chapter 4, the time between divisions follows an exponential distribution.

The uptake equations for random segregation follows in a similar fashion

\[ \begin{align*}
    y_0(t) &= e^{-\alpha t} \\
    y_1(t) &= \alpha te^{-\alpha t} \\
    y_2(t) &= 1 - e^{-\alpha t} - \alpha te^{-\alpha t} = 1 - y_0 - y_1,
\end{align*} \quad (2.2.14-15) \]

The alternative uptake equation solutions are

\[ \begin{align*}
    y_{11}(t) &= \alpha te^{-\alpha t} \\
    y_{12}(t) &= \frac{1}{2} \alpha^2 t^2 e^{-\alpha t} \\
    y_2(t) &= 1 - e^{-\alpha t} - \alpha te^{-\alpha t} - \frac{1}{2} \alpha^2 t^2 e^{-\alpha t} \\
    &= 1 - y_0 - y_{11} - y_{12}.
\end{align*} \quad (2.2.16-18) \]

Solutions for the sets of chase equations in 2.2.7 have the general form

\[ y_{ij}(t) = \left[ \sum_{k=0}^{j} \frac{1}{j!}(\alpha t)^{j-k}y_{ik}(0) \right] e^{-\alpha t}, \quad (2.2.20) \]

with \( i \in \{1, 2\} \), and \( j \in \{0, 1, \ldots, N-1\} \). Since \( y_{ij}(0) = 0 \ \forall j \in \{1, 2, \ldots, N-1\} \) it follows that

\[ y_{ij}(t) = \frac{1}{j!}(\alpha t)^j y_{i0}(0)e^{-\alpha t}. \quad (2.2.21) \]

2.3 Agent-based model

In the previous section we have described an ODE-based model of BrdU dynamics in HSCs. The equations in this model are based on the average behaviour of a population of cells. They provide deterministic descriptions of how the (continuous) proportion of BrdU\(^+\) cells changes over time. The change in BrdU label status of a single cell is however determined by the dynamics of a discrete entity, namely strands of DNA that segregate
in units of chromosomes after each mitotic cell division. It is quite possible for some cells
to have low numbers of labelled DNA strands during any particular stage of tracking,
indicating that a deterministic description might be inaccurate. In such cases stochastic
models usually perform better since they are able to capture the non-deterministic effects
associated with noisy low numbers (Goss & Peccoud, 1998; Wilkinson, 2006).

From this point of view there is thus motivation for a discrete stochastic approach
to modelling BrdU labelling dynamics. In the rest of this chapter we develop an agent-
based model that exhibits these properties. We test the model on the BrdU labelling data
reported in Table 2.1 and compare results to that of the ODE model defined in Section
2.2. The model we introduce here is based on a one-to-one correspondence between a real
cell (including its DNA strands organised in chromosomes) and a software implementation
of a cell (which we will refer to as an agent). In practice it is often found that biologists
can more easily relate to such single cell based models as opposed to more mathematically
complex ones. In addition agent-based modelling of complex biological systems is lately
gaining more widespread use, with applications to stem cells (d’Inverno & Saunders,
2005), diseases (Abbott et al., 2006), intracellular pathways (Pogson et al., 2006) and
epithelial morphology (Walker et al., 2004).

Since our aim is to fit observed BrdU uptake and dilution data rather than modelling
dynamic hematopoietic regulation (a detailed explanation of the difference can be found
in Chapter 6), a lightweight agent-based model sufficed for achieving our goal. We were
fortunate to be in a position where we could steer clear of the computational burden
imposed when modelling size, shape, inter-communication or movement of individual
cells. Rather our focus was to develop a method which is comparable to the ODE model
of Section 2.2 (in terms of proliferation parameters) but is instead stochastic (so that
variation can be simulated) and provides a discrete estimate of the BDT. In addition we
investigated heterogeneity in the HSC population which required our agent-based model
to be easily adaptable in terms of cell population structure (i.e. being able to model
different nodes of the hematopoietic hierarchy).

We will first describe the agent-based model, its treatment of time and its division
algorithms, before we proceed to modelling the experimental data in Table 2.1 with both
the ODE and agent-based model. The adaptability of the agent-based model is then
demonstrated in Section 2.4.4 where we investigate heterogeneity in the HSC population
taking into account the different ways in which chromosomes can segregate.

2.3.1 Model assumptions

The assumptions for the agent-based model are similar to that of the ODE model reported
in Section 2.2.2, namely that HSCs proliferate through asymmetric division only and the
rate of cell death is negligible. The differences are that the time $T$ when BrdU is removed and the number of divisions $N$ to go below the BDT do not have any influence on the actual model definition in this case. $T$ can be dynamically set during a simulation by adjusting a boolean value, similar to how the real life experiment is performed. The BDT is actually not modelled in terms of $N$ anymore, but as the minimum number of chromosomes that needs to be labelled for a cell to be detectable as BrdU$^+$. This parameter too can be dynamically set during a simulation so that the model definition is independent of the actual BDT simulated. In fact, multiple BDT values can be simulated in parallel by simply calculating the label status of every cell for each of the thresholds being simulated.

### 2.3.2 Model description

Since stem cells are rare entities and relatively low in numbers compared to other cells (e.g. the data in this chapter is based on a maximum of 400 cells - see (Kiel et al., 2007, Figure 2 and 3), it is quite feasible to simulate large stem cell systems with single cell based methods in a reasonable amount of time. There are however additional properties of the adult stem cell systems that further appeals to an agent-based approach (Wooldridge, 2002). Firstly, strong evidence has been found that adult stem cells reside and interact with a specialised microenvironment or niche (Wilson & Trumpp, 2006; Scadden, 2006). Additionally adult stem cells directly and indirectly interact with each other and with other molecules in their environment in such a way that their functions (e.g. tissue maintenance or repair) are emergent rather than individually performed. Lastly, each cell has a size and shape as well as a control mechanism (genome and proteome) determining when and which individual actions they perform (grow, divide, migrate, etc.). Implementing many of these properties are, however, not required for an agent-based model of BrdU dynamics. Our agent-based model is thus surprisingly simple since all that it relies on is representation of cells, chromosomes, time, and cell division. We will describe our model under each of these components.

**Cells**

Cells are implemented as agents. Agents can be of different types to represent distinct cell populations and to facilitate heterogeneity. Each agent is member of a masterlist which is used to keep track of the total cell population and its relevant properties, in our case status of BrdU labelling. A hierarchical relationship that resembles the hematopoietic hierarchy depicted in Figure 1.1 is imposed between different cell populations by having sublists of the masterlist for each cell population. Agents move between sublists, when
they differentiate, according to the precursor-progeny relationship defined by the imposed hierarchy. In addition each different type of agent can have their own set of kinetic parameters. The ability of the agent-based model to be effortlessly adapted to model heterogeneous cell populations gives it a very powerful property as we will show later in this chapter. Apart from the cell type, the state of each agent is determined by the label status of its DNA strands, organised in discrete packages of chromosomes.

**Chromosomes**

Chromosomes are implemented as collections of boolean pairs. Each unit of the boolean pair represents one of the DNA strands of a cell’s chromosome (3-prime or 5-prime strand) and is set to `true` if the strand has taken up BrdU and `false` otherwise (Figure 2.2). Depending on the organism we are simulating, the number of chromosomes as well as the ploidy of cells can be set. A global boolean variable `brdu_present` indicates whether BrdU is currently applied or not, so that both BrdU uptake and loss can be dynamically simulated. With our agent-based approach we can count the explicit number of strands that are BrdU positive at each time point for each cell and this provides a very accurate mechanism to model the sensitivity of BrdU detection. For each cell it is thus possible to concurrently calculate whether it would be classified as BrdU$^+$ or BrdU$^-$ for any number of BDTs.

**Time**

The way in which time is treated has a major influence on the structure of a single cell based model. In general there are two approaches to model time - either as continuous or discrete. Continuous time models simulate the time between consecutive events and are thus usually exact since all events are accounted for. Continuous time models are
2.3. Agent-based model

Figure 2.3: **Agent-based approach to time modelling.** The grey box at the top shows a toy example of BrdU dynamics for 4 cells in real time. Cell division events are shown as circles along the time axis with the amount of green fill indicating BrdU intensity. The period up to the red dotted line represents BrdU uptake (intensity increases) and the period thereafter represents BrdU chase (intensity decreasing). The white box at the bottom shows how this toy example would be modelled with our agent-based model. Each individual cell is tracked and at fixed discrete intervals of time (ticks) the probability of dividing is calculated for each cell agent. Some divisions might be missed (black arrows - tick 7) if more than one occurs during a tick. Mostly special cases of stochastic processes like the Poisson or Markov process and we will present such a model of BrdU dynamics in Chapter 4.

Here we follow a discrete time approach by evaluating each individual cell at fixed intervals of time. Temporal space is divided into fixed discrete intervals (ticks) and at the end of each tick one possible set of events that could have occurred in the time interval are predicted and executed. This approach can be thought of as an analogue of taking snapshots of the system at certain time intervals as illustrated in Figure 2.3. A tick can be set to represent any level of granularity (seconds, hours, days, etc.) with smaller values producing more exact results but requiring more computational power. The time interval of a tick needs to be sufficiently small since each agent is allowed only one action per tick. Some actions will be missed (black arrows in Figure 2.3) - and hence inaccurate predictions occur - if the probability of an action per tick is too high. Since cells are traversed sequentially at each tick, concurrency is simulated by randomising the order of traversal.
2.3. Agent-based model

<table>
<thead>
<tr>
<th>Algorithm 1: Asymmetric segregation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> mother</td>
</tr>
<tr>
<td>new(daughter1)</td>
</tr>
<tr>
<td>new(daughter2)</td>
</tr>
<tr>
<td><strong>for</strong> i = 1 to (mother.num_chromosomes) <strong>do</strong></td>
</tr>
<tr>
<td>daughter1.chr(i).5prime $\leftarrow$ mother.chr(i).5prime</td>
</tr>
<tr>
<td>daughter1.chr(i).3prime $\leftarrow$ brdu_present</td>
</tr>
<tr>
<td>daughter2.chr(i).5prime $\leftarrow$ brdu_present</td>
</tr>
<tr>
<td>daughter2.chr(i).3prime $\leftarrow$ mother.chr(i).3prime</td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td>remove(mother)</td>
</tr>
<tr>
<td>add(daughter1)</td>
</tr>
<tr>
<td>add(daughter2)</td>
</tr>
</tbody>
</table>

Cell division

The proliferation rate $\alpha$ of the ODE model is based on the average proportion of HSCs in a population of cells that enter cell cycle each day. The same parameter is implemented slightly differently for the agent-based model since cells are modelled individually rather than as a population aggregate. Here $\alpha$ is defined as the probability that a single HSC will divide in a time period of one day (assuming a tick represents one day). In addition it is assumed that cell division events are independent, so $\alpha$ for a particular cell remains constant irrespective of whether or not that cell has divided the previous day. Ultimately this scheme will also result in $\alpha\%$ of the HSCs dividing per day on average but the use of probabilities adds uncertainty to the behaviour of agents and is at the core of the non-determinism of the model.

2.3.3 Chromosome segregation

When a stem cell agent divides it duplicates its hereditary information (i.e. chromosomes) and spawns two new daughter agents after which itself is deleted from the masterlist and the population sublist it belonged to. The way in which the duplicated hereditary information gets transferred to the two daughter cells depends on whether we simulate asymmetric or random chromosome segregation. For asymmetric segregation (Algorithm 1) only the 3-prime DNA strand of each chromosome can be labelled or unlabelled (5-prime for daughter 2). Random segregation (Algorithm 2) allows either the 3-prime or 5-prime strand to have their labelling state changed in each chromosome independently, for any of the two daughter cells. Unlike the BDT, the two different chromosome segregation
2.3. Agent-based model

<table>
<thead>
<tr>
<th>Algorithm 2: Random segregation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> mother</td>
</tr>
<tr>
<td>new(daughter1)</td>
</tr>
<tr>
<td>new(daughter2)</td>
</tr>
<tr>
<td>for i = 1 to (mother.num_chromosomes) do</td>
</tr>
<tr>
<td>rnd ⇐ uniform(0, 1.0)</td>
</tr>
<tr>
<td>if rnd &gt; 0.5 then</td>
</tr>
<tr>
<td>daughter1.chr(i).5prime ⇐ mother.chr(i).5prime</td>
</tr>
<tr>
<td>daughter1.chr(i).3prime ⇐ brdu_present</td>
</tr>
<tr>
<td>daughter2.chr(i).5prime ⇐ brdu_present</td>
</tr>
<tr>
<td>daughter2.chr(i).3prime ⇐ mother.chr(i).3prime</td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td>daughter1.chr(i).5prime ⇐ brdu_present</td>
</tr>
<tr>
<td>daughter1.chr(i).3prime ⇐ mother.chr(i).3prime</td>
</tr>
<tr>
<td>daughter2.chr(i).5prime ⇐ mother.chr(i).5prime</td>
</tr>
<tr>
<td>daughter2.chr(i).3prime ⇐ brdu_present</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>remove(mother)</td>
</tr>
<tr>
<td>add(daughter1)</td>
</tr>
<tr>
<td>add(daughter2)</td>
</tr>
</tbody>
</table>

schemes cannot be modelled concurrently and each has to be simulated separately.

2.3.4 BrdU detection threshold

The discrete nature of the agent-based model, not only by accounting for individual cells but also by describing the dynamics of individual DNA strands of cells packaged and segregated in units of chromosomes, makes estimation of the BDT for this model much easier and more intuitive compared to what was the case for the continuous ODE model. Here we can, during runtime, simply traverse any number of BDTs we wish to simulate and count for each BDT the number of cells that will be detected as BrdU+. For example only cells with 10 or more chromosomes set to true (either the 5-prime or 3-prime strand can be true) will be counted as BrdU+ at a BDT of 10 chromosomes. The best we could do to implement the BDT with the ODE model was to keep track of the number of times a cell has divided from a reference point. Disadvantages of this approach have already being apparent in the fact that we had to define two versions of the random chromosome segregation uptake equations (Equations 2.2.5, 2.2.8, and 2.2.9). Moreover, it turns out
that much more work is needed when mapping the BrdU intensity to the number of times a cell has divided in the presence and absence of BrdU, as we will show in Chapter 3.

2.4 Simulation results

We have implemented the agent-based model described in Section 2.3 as an Object Oriented application in Java using the Repast Agent Simulation Toolkit (North et al., 2006). Here we evaluate the usefulness of this model as a stochastic BrdU dynamics simulator over the deterministic models described in Section 2.2. Our measure of usefulness is based on how well experimental data is explained by each respective model.

2.4.1 Model evaluation

Two key criteria when evaluating a model’s suitability are its complexity and how well it describes empirical data (goodness-of-fit). The quest is thus to find the simplest model that fits experimental data at a satisfactory level, known as the Ockham’s razor principle. A model of BrdU uptake and dilution dynamics can often be quickly and accurately evaluated simply through visual inspection of the plot comparing model prediction with empirical data. The reason is because the output is a simple one dimensional time-series and the expected shape of the time-series curve is well understood. It is namely monotonically increasing during uptake and monotonically decreasing during chase. Throughout this thesis we quantify the visual goodness-of-fit measure using the Residual Sum of Squares (RSS) statistic. The RSS statistic we use is defined by

\[ RSS = \sum_{i=1}^{n}(e(t_i) - y(t_i))^2, \]  

where \( e(t_i) \) is the experimental value observed at time \( t_i \) and \( y(t_i) \) is the predicted value of the model at time \( t_i \). Smaller RSS values indicate a better fit. The RSS statistic is particularly suitable for this study since more complex model evaluation statistics (likelihood based criteria for example) for the agent-based model are difficult or even impossible to define. In addition, the RSS statistic enables us to compare two very different modelling approaches.

2.4.2 ODE model predictions

When Kiel et al. present their ODE model predictions (Kiel et al., 2007, Figure 3a) they explicitly indicate the prediction values for day 70 only. They conclude that their data are most consistent with random chromosome segregation and that BrdU detection is lost
2.4. Simulation results

Figure 2.4: The deterministic model by Kiel et al. (2007), showing the model prediction for: BrdU uptake, asymmetric segregation BrdU chase, and random segregation BrdU chase on different detection thresholds. A: The prediction when using the original random segregation uptake equations (Equations 2.2.5 & 2.2.6). B: The prediction when using the alternative random segregation uptake equations (Equations 2.2.8 - 2.2.10).

After approximately three divisions. It seems that this conclusion is based solely on the prediction of day 70 since a three division BDT is highly unlikely if the observed data on day 40 and day 120 are also taken into account (see below). To evaluate the model based on all the experimental data, we implemented the equations of Kiel et al. as given in Section 2.2 and plotted two versions of the predictions (together with the observed values and their standard deviation bars). Figure 2.4A shows the prediction using the original random segregation uptake Equations 2.2.5 and 2.2.6 and is the same as the predictions of (Kiel et al., 2007, Figure 3a). Figure 2.4B shows the model prediction when the alternative random segregation uptake Equations 2.2.8, 2.2.9 and 2.2.10 are used instead. We also show the predicted distribution of $y_1$ and $y_2$ during BrdU uptake for the two versions of uptake equations in Figure 2.5. From these figures the following observations can be made:

- The predictions for BrdU uptake are very accurate. The BrdU administration phase is a very robust system to model, independent of chromosome segregation mechanism and the BDT. The only real free parameter is thus the proliferation rate $\alpha$ and it appears that the estimate of 6% per day is very good. In contrast, both models perform poorly in the chase period (days 0-130).

- For the model in Figure 2.4A, if all experimental data are taken into account, the 1-2 division BDT has the best fit (see Table 2.2) which is different to what Kiel et al. concluded. This model therefore predicts that BrdU will not be detected
2.4. Simulation results

Figure 2.5: The proportion of $y_2$ and $y_1$ during BrdU uptake. A: The prediction of the original random segregation uptake equations (Equations 2.2.5 & 2.2.6). B: The prediction of the alternative random segregation uptake equations (Equations 2.2.8 - 2.2.10).

unless each and every chromosome of a cell has at least one BrdU$^+$ strand.

- For the model in Figure 2.4B, the 2-3 division BDT has the best fit (Table 2.2) which seems to be more realistic than in the case of Figure 2.4A. For this model the 1-2 division BDT prediction is very similar to the asymmetric segregation curve. This is due to the smaller predicted proportion of $y_2$ as can be seen in Figure 2.5B, which we postulate is a more realistic prediction than Figure 2.5A over a period of 10 days for cells with a daily turnover rate of 6%. It makes sense that cells with a 1-2 division BDT should have a trajectory close to the asymmetric segregation trajectory, which effectively has a 1 division BDT.

- From this perspective the model of Figure 2.4B can consequently be regarded as an improved version, but nevertheless, prediction in general remains poor. At any threshold, either the early (day 40), middle (day 70) or late stage (day 120) chase data can be fitted, but no model succeeds in satisfactorily describing all stages simultaneously. One possible explanation is that the poor fit is an artifact of a continuous deterministic model describing a discrete stochastic process. It can also imply the existence of a small more quiescent cell population that retains BrdU for a longer period. Using our agent-based model, we can take both of these two possibilities into account. Our suspicions will be confirmed in the much more detailed study we perform in Chapter 3 where a higher resolution and longer term BrdU dataset is presented. In this dataset the long-term label retention is even more apparent and here too a homogeneous model fails to describe all the data simultaneously.
2.4. Simulation results

Figure 2.6: Homogeneous agent-based model simulation results. A: Showing each of the 50 unique simulations along the z-axis. B: The average (of 50 simulations) predicted percentage of 400 cells that are BrdU positive after an uptake period of 10 days. A BDT of 20 chr means at least 20 chromosomes needs to be labelled before the agent is regarded as BrdU +.

2.4.3 Agent-based model predictions

Since it is stochastic, each simulation of the agent-based model for the same parameter settings is slightly different (Figure 2.6A). In order to calculate the RSS we therefore run many stochastic simulations and computed the average trajectory. In addition we could calculate the variance and thereby estimate confidence intervals for predictions. The agent-based model thus allows for a much more statistically rich analysis than the ODE model.

The predictions of our stochastic agent-based model (assuming one homogeneous cell population akin to the ODE model) are shown in Figure 2.6 and Figure 2.7. To improve comparison, parameter values were chosen to match the values reported in Kiel et al.:

- Every tick represents one day. The probability of division for each agent per tick is 0.06. We used a homogeneous population of 400 agents each with 40 chromosomes. Mus musculus (mice) have 20 chromosomes, but since they are diploid the actual number of DNA strands that can take up BrdU is 80.

- BrdU uptake is simulated for 10 days, followed by a chase period of 130 days. Since each simulation produces slightly different results, 50 repetitions were done and the location and spread of the predictions calculated. Using more than 50 runs did not improve the statistical consistency of location and spread estimates, whilst using fewer runs resulted in irregular and inconsistent (between different simulations) sample averages and variances.

Like the model of Kiel et al., this version of our model also supports the random
2.4. Simulation results

Figure 2.7: Homogeneous agent-based model predictions. Showing individual simulations with their average and 95% confidence interval. **A:** Asymmetric segregation model predictions. **B:** Random segregation model predictions on a 20 chromosome BDT.

<table>
<thead>
<tr>
<th>Kiel et al. model</th>
<th>Agent-based model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>original</td>
</tr>
<tr>
<td>immortal</td>
<td>55.9</td>
</tr>
<tr>
<td>1-2 div</td>
<td>29.7</td>
</tr>
<tr>
<td>2-3 div</td>
<td>73.7</td>
</tr>
<tr>
<td>3-4 div</td>
<td>421.4</td>
</tr>
<tr>
<td>4-5 div</td>
<td>1095</td>
</tr>
</tbody>
</table>

Table 2.2: **RSS values** for the two versions of the Kiel et al. model and our agent-based model (homogeneous population). The best RSS for each model is indicated in bold.

segregation hypothesis (comparing Figure 2.7A to Figure 2.7B). The best fit suggests BrdU is detectable at a 20 chromosomes threshold (Table 2.2). However from Figure 2.7B we see that even with our stochastic simulator, the 2.0 ± 1.0% observed value at day 120 falls outside the 95% confidence interval of prediction. In the following section we will investigate the possibility that the observed data was generated by a heterogeneous cell population. We will show that if this is the case the data’s support for the random chromosome segregation hypothesis is much weaker.

Note that our model is much more intuitive in terms of simulating the BDT, since the model of Kiel et al. simulates the number of divisions $N$ before BrdU becomes undetectable rather than the labelled chromosomes itself. In addition, as pointed out in Section 2.2, their model assumes that both DNA strands are BrdU$^+$ after just two divisions during BrdU application under random segregation (Equation 3 in Kiel et al. (2007)). In
2.4. Simulation results

2.4.4 Heterogeneous population

For this version of our model we assumed the existence of a smaller quiescent population of cells which will retain BrdU for longer periods of time. We will refer to the other group of cells as the active population. Quiescence and smaller population size are associated with stronger stemness properties (Wilson et al., 2007), so we further assume that the quiescent cells are precursors of the active cells. Differentiated daughter cells of the quiescent population are thus accrued in the active population. Implementing the additional population only required a rather trivial extension of the agent-based model but complexity is increased in that the size and division rate of the quiescent population are now free parameters which have to be estimated. Fortunately they are constrained since the quiescent population size must be less than 50% of the total population size and the combined daily turnover rate must still be 6% to accurately simulate BrdU uptake. We found that a 30% (120 out of 400) quiescent population with a 0.0156 probability of dividing per day (once in 64.1 days) and 0.079 for the active population (once in 12.7 days) result in better RSS values than the homogeneous population model (Table 2.2 and 2.3). The predictions based on random segregation are shown in Figure 2.8A. The most likely threshold is again 20 chromosomes and the average with 95% confidence interval of this threshold is shown in Figure 2.9A.

The assumption of heterogeneity has a dramatic effect on the predictions under the reality, at the single cell level, this is only one possible outcome in the $2^{\#\text{chromosomes}}$ ways in which the chromosomes can segregate during the second division. Our discrete model successfully captures this process.

**Figure 2.8:** Heterogeneous agent-based model predictions assuming random chromosome segregation. A: Single-phase turnover. B: Bi-phase turnover.
immortal strand hypothesis, with two possible assumptions:

**Immortal-immortal segregation.** The first option is to assume that both populations segregate their chromosomes asymmetrically. In this case a scenario will be created where a small subset of cells (differentiated progeny of the quiescent population during BrdU application) in the active population will have immortal BrdU$^+$ DNA strands. This model is thus extremely efficient in describing long-term BrdU retaining data, with a lower RSS than the random segregation model (Figure 2.10A and Table 2.3).

**Immortal-random segregation.** The second option is for the quiescent population to have asymmetrical segregation and the active population to have random segregation. This might be a biological more realistic assumption since one can expect only the most potent stem cells to have asymmetric chromosome segregation. Interestingly this model is not as effective as the previous option in describing the observed data (Figure 2.11A). There is nevertheless a significant reduction in RSS compared to the immortal strand option of the homogeneous model (Table 2.2 and 2.3).

### 2.4.5 Heterogeneity over time

In addition to a heterogeneous (in terms of proliferation rate) cell population, we also investigated the possibility of heterogeneity over time. As we will point out in Chapter 3, such a scenario can be observed if HSCs and their progeny are sensitive to the presence of BrdU. To take time heterogeneity into account we simulated a bi-phase heterogeneous
2.4. Simulation results

Figure 2.10: Heterogeneous agent-based model predictions assuming immortal-immortal chromosome segregation. A: Single-phase turnover. B: Bi-phase turnover.

Figure 2.11: Heterogeneous agent-based model predictions assuming immortal-random chromosome segregation and a 20 chromosome BDT. A: Single-phase turnover. B: Bi-phase turnover.

population whose division probability during BrdU uptake and the first 10 days of chase are higher than the division probability during chase days 10 - 130. This is a trivial implementation for the agent-based model since a cell populations’ proliferation rate can be dynamically adjusted during runtime. The combined probability of uptake is however still constrained to the optimised estimate of 0.06 per day. The big disadvantage is that the proliferation rates for chase days 10 - 130 had to be found through a systematic process of trial-and-error. Parameters that produced good results are (in the order of population fraction, first phase division probability, second phase division probability): for the random and immortal-random segregation case: quiescent population (30%,0.0156,0.0096),
active population (70%,0.079,0.072); for the immortal-immortal segregation case: quiescent population (10%,0.0156,0.008), active population (90%,0.0649,0.04)\(^3\). It makes sense that the immortal-immortal model would support a smaller quiescent population since the long-term label retention in this case is due to the differentiated progeny of the 10% quiescent population (with immortal BrdU\(^+\) strands) rather than the actual slow turnover rate of the quiescent population itself.

The results are shown in Figures 2.8B and Figure 2.9B for the random segregation case and in Figures 2.10B and Figure 2.11B for the asymmetric segregation case with the corresponding RSS values in Table 2.3. All three cases (random segregation with 20 chromosomes detection threshold, immortal-random segregation with 20 chromosomes threshold, and immortal-immortal segregation) provide near-perfect goodness of fit statistics. Although the random segregation model has the lowest RSS at 4.3, no specific preference can be given to any of the three hypotheses based on the RSS alone. For instance, the immortal-immortal model seems to be the most effective in capturing the large variance observed at day 40 (Figure 2.10B). Both versions of the heterogeneous population model thus have equal support for both random and asymmetric segregation.

<table>
<thead>
<tr>
<th></th>
<th>1-phase</th>
<th>2-phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>immortal-immortal</td>
<td>7.5</td>
<td>5.9</td>
</tr>
<tr>
<td>immortal-random</td>
<td>13.7</td>
<td>5.7</td>
</tr>
<tr>
<td>20 chr</td>
<td>9.3</td>
<td><strong>4.3</strong></td>
</tr>
<tr>
<td>10 chr</td>
<td>47.3</td>
<td>87.9</td>
</tr>
<tr>
<td>5 chr</td>
<td>295.1</td>
<td>401.2</td>
</tr>
<tr>
<td>3 chr</td>
<td>598</td>
<td>746.9</td>
</tr>
</tbody>
</table>

Table 2.3: **RSS values** for the agent-based model (heterogeneous population). Comparing the single phase to the bi-phase version. The best RSS values for each model is indicated in bold.

### 2.5 Conclusion

The model of Kiel et al. and that of other well known BrdU models in literature (Grossman et al., 1999; Cheshier et al., 1999; Bonhoeffer et al., 2000; MacKey, 2001; De Boer et al., 2003) are all ODE-based and thus deterministic and continuous in nature. There are several disadvantages when modelling biological systems as continuous deterministic

\(^3\)These estimates in number of days are (30%,64.1,104.2), (70%,12.7,13.9), (10%,64.1,125), (90%,15.4,25) respectively.
processes. The most obvious being the fact that a deterministic model needs to assume complete knowledge of the biological system under consideration (Wilkinson, 2006). This is not possible for most biological systems that researchers are interested in (due to the mere complexity of the spatial position, size, velocity, etc. of billions of molecules). Hence deterministic models invariably have to adopt a higher level view, representing actual biomolecular reactions as some form of aggregate. For systems where all interacting components are present in high numbers individual fluctuations get subsumed in the population average and hence deterministic models are very effective. However, if some entities are present only in low numbers, the system dynamics behaves in a stochastic manner and needs to be modelled as such (Goss & Peccoud, 1998; Elowitz et al., 2002; Paulsson, 2005). A second advantage of stochastic models over deterministic models is that it provides for much more detailed statistical analysis as we have shown in this study. The disadvantage is that stochastic models are usually much more computationally intensive and parameter estimates and inference are much harder for these models and not as well established as in the case of deterministic models (Wilkinson, 2006; Boys et al., 2008; Wilkinson, 2009).

We have demonstrated how an agent-based approach, with probabilistic decision making, resulted in a stochastic simulator of BrdU uptake and dilution dynamics which can generate Monte Carlo samples from the underlying biological process. Analysis of these samples in turn provides statistical information on the process, whilst being able to account for each individual cell and its chromosomes, provide useful advantages over conventional approaches as we have shown in this study.

We evaluated three versions of our model, systematically increasing its complexity and also model prediction accuracy. Taking the results of all three versions of our model together, we can conclude that:

1. The data of Kiel et al. are unsuitable for making conclusions about the “immortal strand hypothesis”. More experimental data for the chase period 0-40 days and post 120 days might rectify this situation.

2. It is very likely that the data were generated by a heterogeneous population, supporting the possibility of a smaller quiescent subset of cells that are the precursors of the larger active cell population.

3. There is further support for heterogeneous proliferation rates over time. In particular changes occur during the period that BrdU is added to or removed from the system, which raises the possibility that application of BrdU induces an injury signal whereby turnover rates are increased. We will discuss the implications of this philosophy in greater detail in Chapter 3.
4. The method used by Kiel et al. to detect BrdU+ cells (immunofluorescence microscopy) has a maximum BDT of 20 chromosomes. Stem cells with less than 20 labelled chromosomes are thus classified as BrdU− cells.

One of the central topics of this chapter is the ease with which our agent-based approach can be extended to model different cell population structures, time heterogeneity and intuitively account for the BDT. Our agent-based model has however one major drawback - it is very inefficient for parameter estimation. To our knowledge the only way is a brute-force search or trial-and-error approach. The fact that the model is stochastic implies that many simulations need to be performed to evaluate prediction accuracy and this renders these brute-force strategies even more time consuming.

We were fortunate here since a minimum number of parameter estimation was involved and in each case the combined uptake rate was constrained at 6% per day. The question now arises whether the ODE model, since it is much more suitable for parameter estimation, can also be extended to model more than one cell population. In the next chapter we show that this is indeed possible, although considerable more effort and care is needed when mapping the BDT to the number of times a cell has divided.
Chapter 3

Ordinary differential equation model

3.1 Introduction

In Chapter 2 we have compared an agent-based model to a continuous ODE model of BrdU label-retaining cell (LRC) data. We have shown the flexibility of the agent-based model and demonstrated the ease in which it could be adapted to various HSC architectures (modelling cell population heterogeneity) as well as its intuitive way of describing the BDT. In this chapter we introduce a new dataset of experimental BrdU tracking data in bone marrow HSCs. This is a much higher resolution and longer term dataset and was obtained with the specific aim to elucidate issues surrounding the dormant hematopoietic stem cell hypothesis (introduced in Chapter 1). Many of the questions raised could precisely be answered by making use of the flexibility demonstrated by the agent-based approach in Chapter 2. However, since the dormant hematopoietic stem cell hypothesis is somewhat controversial, accurate parameter estimation making optimal use of all the information conveyed by the new dataset was crucial. Modelling the new dataset with an approach for which high performance optimisation techniques exists thus became of critical importance. Since, as we have pointed out at the end of Chapter 2, the best parameter searching strategy to our knowledge for the agent-based model is a brute-force search approach, the agent-based model were deemed unsuitable here.

Whilst parameter estimation for continuous deterministic models is well established, it is not clear to what extend an ODE model can successfully capture heterogeneity in both HSC populations and time, or exhibit similar flexibility (whilst clamping the number of unknown parameters) as has been demonstrated by our agent-based model. In order to achieve this goal the ODE approach needs to model both BrdU uptake and dilution of a possibly heterogeneous cell population and carefully implement the way in which the model accounts for the BDT, since it turns out that there exist a multiple nonlinear dependency between the number of divisions during BrdU administration, the number of
3.1. Introduction

divisions during BrdU dilution, and the BDT (see Section 3.2.3). By carefully elucidating this relationship and incorporating the results into our model we will show that it is indeed possible to model more than one cell population with an ODE approach, albeit with much more work than what was required for the agent-based model. Fortunately we can restrict ourselves to modelling on a cellular level and hence successfully describe the system using only linear differential equations. We are thus able to find analytic solutions for the equations which in turn allows the use of well established optimisation techniques to search for optimal parameters. Moreover, by optimising parameters over carefully motivated subregions of the data, it became possible to implicitly model time heterogeneity.

The ultimate purpose of mathematical modelling in this chapter is to ascertain which one of two hypotheses best describes the observed BrdU kinetics (Figure 3.2) of the bone marrow Lin-Scal+Kit+CD150+CD48-CD34-CD135- population, which is highly enriched for functional HSCs. The first hypothesis states that this population is a homogeneous group of cells with a homogeneous rate of division across the whole population. The second hypothesis states that the HSC population comprises two subpopulations: a (smaller) dormant subpopulation (which we shall refer to as the d-HSC population) with a very slow turnover rate and an activated subpopulation (which we shall refer to as the a-HSC population) with a faster turnover rate. Together with time heterogeneity, confirmation of this last hypothesis would provide support for the dormant hematopoietic stem cell hypothesis (d-HSC hypothesis hereafter). Recall from Chapter 1 that the d-HSC hypothesis provides an explanation of how a small number of tightly regulated HSCs in the bone marrow can be so effective in both maintenance of homeostasis and restoration after injury. The first compelling biological evidence of the d-HSC concept is given by Wilson et al. (2008) where the conventional view of HSCs and the hematopoietic hierarchy (Bradford et al., 1997; Cheshier et al., 1999) is seriously challenged. A comparison between the conventional and d-HSC hematopoietic hierarchy is shown in Figure 3.1, where all the different progenitor populations of Figure 1.1 are represented as a single collection of progenitors and transit amplifying cell populations.

3.1.1 Experimental data

The empirical BrdU LRC data on which modelling in this chapter is based have been reported in Wilson et al. (2008) and are given in Tables 3.1 and 3.2. This is the most rigorous in vivo BrdU dataset to date and track dilution up to 306 days - more than half the lifetime of a typical mouse. Each time point represents between 5 and 11 mice. The dose of BrdU administered is 180mg intraperitoneal injection per mouse at the start followed by water containing 800 micrograms per ml BrdU continuously for 10-13 days. In
3.1. Introduction

Figure 3.1: Conventional vs. dormant population HSC hierarchy. Within the HSC population, two possible HSC hierarchy models can be envisaged. In the conventional model (left panel), the HSC population is homogeneous with respect to cell cycle entry with the entire HSC pool turning over every few weeks. In contrast, in the dormant HSC model (right panel), the hierarchical organisation of the hematopoietic system includes the phenotypic HSC pool, in which two subpopulations can be defined based on their relative turn-over frequencies. An active HSC (a-HSC) population is responsible for the day-to-day maintenance of the hematopoietic system, while a second population, the dormant HSC pool (d-HSC), cycles only a few times over the life span of the mouse in a homeostatic situation (dashed arrow) but is activated and participates in replenishment of the hematopoietic system after injury (solid arrow). LT: long-term.

Total 137 mice were examined over a period of 319 days. BrdU staining was quantitated by flow cytometry as described in Wilson et al. (2004). The novelty of this dataset is the exceptionally long chase period, as previous BrdU datasets only included uptake (Bradford et al., 1997; Cheshier et al., 1999) or tracked label dilution no longer than 70 days (Mohri et al., 1998; Bonhoeffer et al., 2000; De Boer et al., 2003) or 120 days (Kiel et al., 2007) at most. The data is plotted in Figure 3.2 with the day 0 time point set to the time BrdU was removed. Strikingly, the rate of label dilution in Figure 3.2 decreases after chase day 70 flattening out as label is retained on the long-term. It is precisely this
3.1. Introduction

Figure 3.2: Plot of experimental data in Tables 3.1 and 3.2. Green line: BrdU uptake; Blue line: BrdU chase starting at day 0; red x: mean observed data. The long-term label retention is clearly visible and was found most interesting by biologists.

long-term label retention which we aim to describe with computational modelling in this chapter. Other peculiarities that strongly influence model design can also be identified and is discussed in Section 3.5.1. This high resolution\(^1\) dataset together with an ODE approach enabled us to support the d-HSC hypotheses and also elucidate properties of the dormant and active HSCs in much more detail and with more certainty than would have been possible with the agent-based model described in Chapter 2.

The rest of this chapter is organised as follows. In Section 3.2 we do an in depth study of the nature of our BrdU dataset and issues that have to be addressed by a continuous model. The actual model equations and their solutions are presented in Sections 3.3 and 3.4 respectively. Section 3.5 discusses our parameter estimation strategy and simulation results before we end the chapter with a few concluding remarks in Section 3.6. The work presented here has been published in Wilson et al. (2008) and van der Wath et al. (2009).

\(^1\)In terms of typical in vivo chromosomal LRC data.
Table 3.1: % BrdU+ HSCs (uptake). Uptake data from (Wilson et al., 2008, Figure 7C). H: Hour; SD: Standard deviation.
<table>
<thead>
<tr>
<th>Num</th>
<th>D 0</th>
<th>D 3</th>
<th>D 10</th>
<th>D 14</th>
<th>D 35</th>
<th>D 42</th>
<th>D 56</th>
<th>D 70</th>
<th>D 112</th>
<th>D 141</th>
<th>D 170</th>
<th>D 177</th>
<th>D 213</th>
<th>D 240</th>
<th>D 306</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85.81</td>
<td>86.4</td>
<td>83.3</td>
<td>78.0</td>
<td>61.35</td>
<td>46.5</td>
<td>40.23</td>
<td>30.25</td>
<td>24.1</td>
<td>21.81</td>
<td>15.4</td>
<td>17.9</td>
<td>15.1</td>
<td>7.1</td>
<td>5.0</td>
</tr>
<tr>
<td>2</td>
<td>87.58</td>
<td>94.7</td>
<td>57.7</td>
<td>70.97</td>
<td>59.1</td>
<td>46.9</td>
<td>40.44</td>
<td>30.85</td>
<td>24.06</td>
<td>21.32</td>
<td>12.1</td>
<td>12.0</td>
<td>6.6</td>
<td>16.25</td>
<td>5.2</td>
</tr>
<tr>
<td>3</td>
<td>86.0</td>
<td>91.3</td>
<td>69.5</td>
<td>64.3</td>
<td>50.0</td>
<td>48.09</td>
<td>40.52</td>
<td>33.48</td>
<td>24.6</td>
<td>21.6</td>
<td>12.8</td>
<td>17.2</td>
<td>19</td>
<td>7.8</td>
<td>5.1</td>
</tr>
<tr>
<td>4</td>
<td>95.0</td>
<td>90.7</td>
<td>68.5</td>
<td>60.1</td>
<td>54.8</td>
<td>52.27</td>
<td>29.0</td>
<td>28.86</td>
<td>21.0</td>
<td>25.97</td>
<td>14.7</td>
<td>12.0</td>
<td>7.8</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>98.9</td>
<td>90.8</td>
<td>75.41</td>
<td>76.0</td>
<td>43.3</td>
<td>41.24</td>
<td>30.4</td>
<td>26.4</td>
<td>20.5</td>
<td>10.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>94.94</td>
<td>76.29</td>
<td>68.5</td>
<td>46.9</td>
<td>35.0</td>
<td>34.0</td>
<td>16.9</td>
<td>21.5</td>
<td>14.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>90.27</td>
<td>42.1</td>
<td></td>
<td>28.74</td>
<td></td>
<td>24.6</td>
<td>22.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>91.39</td>
<td>48.9</td>
<td></td>
<td>26.91</td>
<td></td>
<td>11.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>87.49</td>
<td></td>
<td></td>
<td>26.83</td>
<td></td>
<td>10.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>100.0</td>
<td></td>
<td></td>
<td>26.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>28.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>91.74</td>
<td>90.78</td>
<td>71.78</td>
<td>69.65</td>
<td>56.31</td>
<td>46.87</td>
<td>37.74</td>
<td>29.52</td>
<td>23.44</td>
<td>21.07</td>
<td>17.06</td>
<td>13.97</td>
<td>13.57</td>
<td>9.74</td>
<td>4.63</td>
</tr>
<tr>
<td>SD</td>
<td>5.23</td>
<td>2.95</td>
<td>8.72</td>
<td>6.83</td>
<td>5.01</td>
<td>3.17</td>
<td>4.84</td>
<td>2.6</td>
<td>1.65</td>
<td>5.88</td>
<td>4.3</td>
<td>3.08</td>
<td>6.34</td>
<td>4.35</td>
<td>0.95</td>
</tr>
</tbody>
</table>

**Table 3.2:** % BrdU⁺ HSCs (chase). Chase data from (Wilson et al., 2008, Figure 2D). D: Day; SD: Standard deviation. Day 0 corresponds to Hour 312 of Table 3.1.
3.2 BrdU modelling for continuous models

Although we have already presented a continuous ODE model of BrdU labelling dynamics in Chapter 2, we also highlighted the problems of a continuous approach in accounting for the BDT which is inherently discrete. In particular difficulties arise when grouping cells in terms of the number of times they have divided instead of modelling label intensity per se. In general any system with some temporal dependant state can be described by differential equations. In theory it should thus be possible to define an ODE model with a more direct implementation of label intensity. But in reality there exist an infinite number of possible intensities and it is therefore not feasible to define a separate equation for every possible label intensity. Also, here it is important to limit the number of unknown parameters since inference on the available data must be strong (too many free parameters create a substantial risk of overfitting). A good understanding of the process of BrdU labelling and de-labelling dynamics are thus prerequisites before an ODE model in particular can be properly defined on BrdU data. Here we describe some of the difficulties and typical characteristics of BrdU data that can be exploited to reduce the number of equations required without sacrificing faithfulness in terms of the BDT. But first let us briefly transcend reality and consider in vivo cell division tracking in an ideal world.

**In vivo cell division tracking in an ideal world**

Imagine a method in which an exact cell cycle counter can be embedded in all cells in vivo, either from some known point in time or, even better, from the embryonic stem cell stage (so that all cells in a particular cell lineage have the counter). After any desirable period the cell cycle counters of each cell can be examined and the distribution of cells that have divided once, twice, thrice, etc. be determined. Using mathematical models similar to that of in vitro CFSE labelling (Gett & Hodgkin, 2000; De Boer et al., 2006; Ganusov et al., 2005; Yates et al., 2007) very reliable turnover rates could be estimated from such a distribution and the time period it represents. An even more useful method would be some form of wireless transmitter that uniquely identifies individual cells and signals mitotic events. Real-time reconstruction of a cell population’s ancestral tree would then be possible and calculating any cell’s turnover rate would be a simple division of branch depth by time interval.

3.2.1 Difficulties in interpreting BrdU data

BrdU labelling is one of the oldest and most widely used in vivo cell division labelling compounds. Unfortunately the labelling and detection process of BrdU is far from the
ideal world example described above - there are many confounding factors that have to be taken into account when interpreting BrdU data.

**Number of divisions**

The most obvious confounding factor is the fact that labelled cells are relabelled when they divide two or more times during uptake. But in terms of the standard output there is no distinction between the number of times a cell has taken up BrdU - cells are simply classified as BrdU+ (labelled) or BrdU− (not labelled). It is thus fundamentally wrong to conclude that a group of cells divide at a rate of \( \frac{x}{10} \) % per day if it is observed that \( x\% \) of these cells are BrdU+ after a labelling period of 10 days. Label incorporation thus needs to be tracked over a number of time points so that the rate of uptake can be estimated with mathematical modelling.

The situation is even more dire during BrdU dilution - many different patterns in staining arise if chromosomes are assumed to segregate randomly (Figure 3.3). It is in fact possible for a cell to receive only unlabelled chromosomes in just one division, or it might take several divisions to dilute the same label. Fortunately, the population-wide average intensity is more or less halved with each division, so the combination of a high quality dataset and good continuous models can produce accurate results.

![Figure 3.3: Mitosis schematic of BrdU uptake and chase.](image)
Change in intensity in a population of cells during uptake is uniform, but in practice each intensity is transformed to a boolean classification of labelled or not labelled. It is thus not possible to directly know the number of times a cell has divided. During BrdU chase many different intensity patterns are possible, depending how chromosomes segregate.
3.2. BrdU modelling for continuous models

Chromosome segregation

From the discussions (Sections 2.2 and 2.3) and model predictions (Section 2.4) in Chapter 2 it was clear that interpretation of BrdU data is dependant on the chromosome segregation mechanism (asymmetric or random) assumed. We also showed that BrdU experiments need to be carefully designed otherwise it will not provide much information about chromosome segregation if the cell population under scrutiny is heterogeneous. In particular we showed that the BrdU data of Kiel et al. (2007) cannot (as they have done) be used to refute the asymmetrical chromosome hypothesis. Nevertheless most recent evidence does support random chromosome segregation in HSCs (see our discussion on biological explanations for long-term BrdU retention in Section 3.2.4).

Toxicity of BrdU

Since HSCs are responsible for maintenance and repair of all mature blood cells, any peripheral injury caused by BrdU is likely to induce a proliferative signal in HSCs (thereby affecting interpretation of observed BrdU data). Evidence can be found in literature for (Stockdale et al., 1964; Yen & Forbes, 1990; Caldwell et al., 2005) and against (Schneider et al., 1977; Mohri et al., 1998; Cheshier et al., 1999) claims that BrdU is toxic for cells. It spite of the evidence that it is not toxic BrdU has never been deemed safe enough for human use. We will show that the data presented here strongly supports effects that can be best explained by toxicity of BrdU. In fact, it turns out to be a key assumption without which the data cannot be described by computational models at all.

3.2.2 Two phases of BrdU kinetics

From our discussion so far it clearly emerged that BrdU labelling assays are typically characterised by two distinct phases, the shorter uptake or pulse phase (BrdU present, labelling) followed by a longer chase phase (BrdU absent, de-labelling). Although the data observed during both phases are “generated” by exactly the same underlying cellular dynamics (i.e. proliferation and death), different information about the cellular processes are conveyed during each phase. When modelling BrdU kinetics it is thus of utmost importance to have a precise understanding of the nature of the information that can be learned from the data during each phase.

Uptake phase. During BrdU uptake all newly synthesised DNA strands take up BrdU and hence cells that are observed as labelled must be progeny of cells that have divided at least once since BrdU was added to the system\(^2\). The proportion of

\(^2\)It is also possible for cells to take up BrdU during DNA repair but \textit{in vitro} studies on the
BrdU+ cells in a given population thus increases monotonically over time and the rate at which this happens tells us something about cell proliferation and death rates. But as we have mentioned above, some difficulties in interpretation arise during uptake due to the fact that we cannot tell how many times an observed BrdU+ cell has divided. Fortunately, mathematical modelling makes it possible to better estimate turnover rates if BrdU uptake are tracked over a number of time points. Short uptake phases also help to limit the number of cells that divide multiple times during this time period. A major advantage when modelling BrdU uptake data is its independence of the particular chromosomal segregation mechanism and the BDT (see Section 2.4.2 of Chapter 2), it is therefore quite robust and limits the number of unknown parameters when modelled. Moreover, in terms of the BDT it is also evident from the uptake data that 50% labelled DNA strands (thus after one or more uptake divisions) is an upper bound for our detection threshold estimate. If BrdU detection by flow cytometry was less sensitive than 50%, cells would need to divide at least twice before we could detect them as BrdU+, making BrdU labelling very inefficient.

**Chase phase.** The chase phase starts at the end of the uptake phase when BrdU is removed. From here on newly synthesised DNA do not take up BrdU anymore with the result that the amount of BrdU in a cell is reduced each time it divides. In most cases each daughter cell will have less BrdU molecules than their mother cell with the exact number of labelled DNA strands depending on how the chromosomes segregated during mitosis. At some point the intensity of BrdU in a cell will fall below the level of detection and consequently the cell will be classified as BrdU−. The total proportion of BrdU+ cells during chase is therefore monotonically decreasing. Unlike during uptake this proportion is not only determined by the cell proliferation rate but also by the chromosome segregation mechanism as well as the BDT. There are thus two additional parameters that need to be estimated for chase phase data.

When modelling BrdU data there is the option to focus on one phase only, depending on the purpose of the study. For example, it is possible to estimate cellular turnover kinetics by only studying BrdU uptake as in Cheshier et al. (1999). Alternatively, only the chase data are of interest when label retention or heterogeneity is studied (Wilson et al., 2008; Glauche et al., 2009). Since LRCs can, by definition, only be observed during the chase phase when labelling is diluted, we can expect the discriminative power for heterogeneity in HSC proliferation parameters to lie only in the chase phase of the density of BrdU labelled DNA suggest label uptake is most likely due to proliferation (Pang et al., 2003).
data. There are, however, also other possible causes for the observation of long-term label retention in HSCs (Section 3.2.4) which our models have to address. Although BrdU uptake data has little discriminative power when investigating HSC heterogeneity, actual parameter estimates are in fact strongly influenced when they are based on both BrdU uptake and chase data as opposed to uptake or chase data only. This observation is based on the concept that cells that have divided more in the presence of BrdU need more divisions to dilute the BrdU they have taken up (see the discussion below). We therefore regard the two BrdU phases as an advantage: two different datasets that carries the same information implies it is easier to learn that information. We already used this to our benefit in Chapter 2 where the proliferation rate $\alpha$ was estimated at 6% per day using only the uptake data and without having to take the chromosome segregation mechanism or the BDT into account. Knowing the population turnover rate we could more easily study possible heterogeneous subpopulations using the chase data. Unfortunately some of the gains for modelling in having two phases are lost due to the toxic effects of BrdU that contaminates the homeostatic proliferation rate during uptake. However, as we will discuss later the advantages of BrdU toxicity outweighs the disadvantages: the injury signal induced by BrdU is precisely what makes BrdU labelling of slowly cycling populations (typical stem cells) possible.

### 3.2.3 BrdU detection threshold

In addition to HSC proliferation kinetics, the BDT needs to be estimated since it can have a confounding effect on the observed LRC data. Although the BDT can only be estimated from chase phase data, we at least know the upper bound from the uptake data. This BDT upper bound is nevertheless also apparent in the chase phase data where a detection threshold of more than 50% would imply rapid loss of labelling, similar to what would be the case if chromosomes segregated asymmetrically (Kiel et al., 2007; van der Wath & Liò, 2008).

At first glance it appears as if an ODE model that accounts for the BDT needs to track the infinite number of possible label intensities in cells - clearly not a viable solution. The reason being that an ODE model has to define an equation for each “species” in the system it models and since each cell can have a real valued BrdU intensity, the system we model has an infinite number of possible “species”. One solution is to discretise the intensity by measuring it as the number of labelled DNA strands. This approach is still troublesome because of the exponential number of ways in which chromosomes can segregate randomly during chase (Section 2.4.3 of Chapter 2). ODE models of signalling pathways are confronted with a similar problem because of the combinatorial explosion in the number of possible protein complexes. A number of elegant solutions in the form
of rule-based process calculi have been put forth by the Computer Science community in recent years, namely \(\pi\)-calculus (Regev et al., 2001), Bio-PEPA (Calder et al., 2006), \(\kappa\)-calculus (Danos et al., 2007), and BlenX (Dematte et al., 2008). Fortunately in our case, cells that have divided an equal number of times in the presence or absence of BrdU will tend to have a similar label intensity. The BDT can thus also be defined in terms of the number of times a BrdU\(^+\) cell can divide before it is not detected as BrdU\(^+\) anymore (defined as \(N\) in Section 2.2.2) as opposed to the minimum number of BrdU molecules that need to be present before a cell is detected as BrdU\(^+\). Therefore we don’t need to define thousands of equations but can map the BrdU intensity to the number of times a cell has divided during uptake or chase. As we will show in the next section doing this properly is not straightforward.

### Mapping BrdU intensity to number of divisions

Accurate modelling of the BDT by grouping cells by the number of times they have divided during uptake or chase is at the heart of the continuous approach described in this chapter and requires careful analysis. The reason why the BDT complicates matters is because some cells will take \(n\) chase divisions to transcend the detection threshold whilst others that have divided more during uptake will take \(n+1\) divisions. However, for some other BDT these same cells might take an equal number of divisions to dilute their label. This multiple nonlinear relationship is depicted graphically in Figure 3.4 where we have calculated the average real-valued labelling percentage of a single cell based on up to 4 uptake divisions as well as the decrease in label corresponding to each of the 1 to 4 uptake divisions. Also shown is a slice in more detail where 5 different detection thresholds are indicated with red lines. Here we can now clearly see that at a threshold of 5\%, 1 and 2 uptake division cells will lose label after 4 divisions and 3 and more uptake division cells will lose label after 5 divisions. At a 7.5\% threshold 1 uptake division cells will lose label after 3 divisions whilst all the others (2 or more uptake divisions) will take 4 divisions. Interestingly all cells will lose label after 3 divisions at a 12.5\% threshold. It is thus clear that there is a complex interplay (summarised in Table 3.3) between the number of divisions during uptake, the BDT, and the number of divisions during chase. Any continuous BrdU model used to infer cell division kinetics should carefully address these three factors.

The proportions in Figure 3.4 can be calculated when realising that half of a cell’s DNA is newly synthesised after each mitotic cell division. All newly synthesised DNA during BrdU uptake will thus be BrdU\(^+\) so that \(\text{str}_u(i)\), the average number of labelled DNA strands in a single cell after uptake division \(i\), is given by the recursive expression

\[
\text{str}_u(i) = 0.5 \times (\text{str}_u(i-1) + \text{str}_t) \quad \text{for} \quad i \in \{1, 2, 3, \cdots \},
\]
Figure 3.4: Average theoretical BrdU percentage of a single cell for a given number of divisions during uptake and chase. Yellow: cell has divided once during uptake; green: cell has divided twice during uptake; red: cell has divided thrice during uptake; violet: cell has divided four or more times during uptake; red lines: detection thresholds. Slices of each uptake pie are cumulative and BrdU percentage is calculated in terms of the number of DNA strands that are labelled and not in terms of the proportions of nucleotides that are replaced by BrdU. Uptake divisions follow in a clockwise direction and chase divisions follow in an anticlockwise direction. Slices intersected by red lines indicate the number of divisions for the cell to fall below the detection threshold.

where \( s\text{tr}_u(0) = 0 \) and \( s\text{tr}_t \) is the total number of DNA strands (twice the number of chromosomes). Here we assume that chromosomes segregate randomly and hence it is possible for chromosomes to have both their DNA strands labelled after two or more uptake divisions. Conversely all newly synthesised DNA during chase is BrdU\(^+\) (unlabelled) so that \( s\text{tr}_c(j) \), the average number of labelled DNA strands in a single cell after chase division \( j \), is given by

\[
 s\text{tr}_c(j) = 0.5 \times s\text{tr}_c(j - 1) \quad \text{for} \quad j \in \{1, 2, 3, \cdots \},
\]

where \( s\text{tr}_c(0) = s\text{tr}_u(n) \) after \( n \) uptake divisions. Note that we are calculating the average number of labelled DNA strands for a single cell and non-integer numbers of strands are therefore possible. The biological interpretation must however be made in a cell population context to make sense. For example \( s\text{tr}_c(j) = 2.5 \) means that 50% of cells in a particular population have 2 labelled DNA strands and the other 50% have 3 labelled DNA strands.
The total number of uptake and chase divisions we have to consider are dependant on the assumed BDT. In theory the smallest change in BrdU intensity is by one nucleotide which poses a potential dilemma in respect of the number of possible detection thresholds to be modelled. Fortunately BrdU intensity is diluted in units of chromosomes upon cell division so that intensity differences on a nucleotide level can be safely ignored. Moreover, no chromosome can have both its DNA strands labelled after one chase division. We thus know that label intensity will be reduced by multiples of single DNA strands of chromosomes during chase (after the first division) and that the least possible label intensity in a cell is $1.25\% = 1 \div 80 \times 100$ in the case of mice. Note that we have assumed that differences in the sizes of chromosomes are negligible.

<table>
<thead>
<tr>
<th>BDT</th>
<th>% DNA strands labelled</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$n_{c_1}$</th>
<th>$n_{c_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.25%</td>
<td>1,2</td>
<td>3,4,..</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2.5%</td>
<td>1,2</td>
<td>3,4,..</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3.75%</td>
<td>1</td>
<td>2,3,..</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5%</td>
<td>1,2</td>
<td>3,4,..</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6.25%</td>
<td></td>
<td>1,2,..</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>7.5%</td>
<td>1</td>
<td>2,3,..</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>8.75%</td>
<td>1</td>
<td>2,3,..</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>10%</td>
<td>1,2</td>
<td>3,4,..</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>11.25%</td>
<td>1,2,3</td>
<td>4,5,..</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>12.5%</td>
<td></td>
<td>1,2,..</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3.3: **BrdU detection threshold.** BDT: BrdU detection threshold (minimum number of labelled DNA strands that can be detected). For each BDT we divide cells into two groups during the chase phase ($c_1$ and $c_2$) depending on how many times a cell has divided during BrdU uptake. We then give $n_{c_1}$ and $n_{c_2}$, the number of times cells in each group has to divide to go below the detection threshold. Consider a BDT of 4 strands for example: cells that have divided once or twice during uptake will dilute their BrdU in 4 divisions, whilst it will take 5 divisions for cells that have divided 3 or more times during uptake. This table was compiled using Figure 3.4.

### 3.2.4 Long-term retention of BrdU

Several factors can cause long-term retention of experimentally labelled DNA in cells as is observed in Figure 3.2. In general these factors can be either classified as artefacts of the experimental procedure used (i.e. experimental error) or can they be explained by some biological property of the cells under scrutiny.
Experimental error. We have identified two possible experimental confounding factors that might create an illusion of label retention. Good mathematical models need to consider both of these factors. The first is controlling for the BDT which, as highlighted above, is an absolute necessity for more accurate parameter estimation. Apart from its influence on parameter estimation the BDT also provides a possible explanation for label retention, since if BrdU detection by flow cytometry was extremely sensitive, cells would still be detected as BrdU\(^+\) even after many chase divisions. Hence it will appear as if cells have long-term BrdU retention. Alternatively it might be possible that the natural stochastic variation between mice at the HSC level is much larger than expected. In this case the observed long tail in Figure 3.2 may be due to an inadequate or unintentional skew sample of mice examined at each time point. Since the model we describe here is deterministic, we will investigate this possibility for the dataset reported here in Chapter 4 where we predict confidence intervals using variance estimates of a stochastic version of our model.

Biological explanations. In terms of biological properties of stem cells explaining label retention two independent hypotheses have been postulated. Firstly it has been proposed in the “immortal strand hypothesis” (first introduced in Chapter 2) that stem cells divide frequently, but in doing so asymmetrically recognise and retain the “old” labelled mother DNA strand, while the newly synthesised unlabelled chromosomes are selectively distributed to the non-stem cell daughter (Cairns, 1975; Potten et al., 1978). Alternatively stem cells with a very slow division rate during chase will retain DNA label much longer than those that divide more frequently. It has recently been shown that label retention, which cannot be explained convincingly by an asymmetric segregation mechanism, has also been observed in nuclear protein labelling assays, such as that using an H2B-GFP fusion protein under control of specific promoters (Wilson et al., 2008; Foudi et al., 2009). Moreover the “immortal strand hypothesis” has recently been seriously challenged for HSCs (Kiel et al., 2007; Lansdorp, 2007; Shinin et al., 2006), although as we have shown in Chapter 2 the LRC data of Kiel et al. provides insufficient information about the segregation mechanism of chromosomes, but nevertheless supports HSC heterogeneity including a dormant population. The models we present here have therefore assumed random chromosome segregation so that the presence or absence of dormancy is the only biological hypothesis we have to address.
3.3 Model equations

We model the LRC data of Figure 3.2 by a system of coupled ODEs that describe the rate of change of the proportion of BrdU$^+$ cells over time. In order to test for label retention due to slow cellular proliferation, two versions of the LRC model were implemented. First, a one-population version that models the data as being observed from a single population of cells with homogeneous turnover rates was defined. Secondly, this model was extended to a two-population version that assumes heterogeneous turnover rates between the two subpopulations. These two models are defined below.

3.3.1 One-population model parameters and assumptions

The one-population model assumes a single active cell population represented by $A$. As before a separate set of equations is defined for BrdU uptake (indicated by $A_u$) and two sets of equations for the chase phase, with $A_{c_1}$ indicating the group of cells diluting labelling after $n$ divisions and $A_{c_2}$ indicating the group of cells diluting labelling after $n+1$ divisions. Dynamics are determined by the rates at which cells self-renew $\sigma$, differentiate $\delta$, or die $\gamma$. A cell that self-renews has undergone cell cycling and hence is replaced by two daughter cells with changed label intensities but retaining stem cell properties (symmetric division). Label intensity is mapped to the number of chase divisions a cell has undergone from a reference point (see Figure 3.4 and Table 3.3) so that an equation can be defined for cells at each division level. For example, during uptake active cells that have divided $i$ times in the presence of BrdU will be represented by $A_{ui}$, and chase group $A_{c_1}$ active cells that have divided $i$ times without BrdU present will be represented by $A_{c_1i}$. Notation for $A_{c_2}$ follows in a similar fashion. Differentiation happens when a cell changes its phenotype to that of its direct progeny - independent of cell division and hence label intensity is unaffected. Cells that die are simply removed from the system.

3.3.2 One-population model equations

The equations of the one-population model are related to the random chromosome ODE model described in Chapter 2. It differs in that the equations here assume symmetric self-renewal and include a parameter for cellular death (for reasons that will be clear when we describe the two-population model). In addition this model has a more elaborate treatment of the BDT (Section 3.2.3) and hence equations are defined for levels of divisions (akin to the classes of labelling of the alternative uptake equations in Section 2.2.4 but using the scheme described in Section 3.2.3) rather than the number of labelled DNA strands.
3.3. Model equations

Uptake equations

\[
\frac{dA_{u_0}}{dt} = - (\sigma_a + \delta_a + \gamma_a) A_{u_0}
\]
\[
\frac{dA_{u_i}}{dt} = 2\sigma_a A_{u_{i-1}} - (\sigma_a + \delta_a + \gamma_a) A_{u_i} \tag{3.3.1}
\]
\[
\frac{dA_{u_n}}{dt} = 2\sigma_a A_{u_{n-1}} + (\sigma_a - \delta_a - \gamma_a) A_{u_n},
\]

where \( i \in \{1, 2, \ldots n - 1\} \). For the BDT values reported in Table 3.3, we can see that the largest \( n \) we have to model is 4. Initial conditions are \( A_{u_0}(0) = 1 \) and \( A_{u_i} = 0 \) \( \forall i \geq 1 \) since all cells are unlabelled when BrdU application starts.

Chase equations

\[
\frac{dA_{c_i0}}{dt} = - (\sigma_a + \delta_a + \gamma_a) A_{c_{i0}}
\]
\[
\frac{dA_{c_{ij}}}{dt} = 2\sigma_a A_{c_{i(j-1)}} - (\sigma_a + \delta_a + \gamma_a) A_{c_{ij}} \tag{3.3.2}
\]
\[
\frac{dA_{c_{1n}}}{dt} = 2\sigma_a A_{c_{1(n-1)}} + (\sigma_a - \delta_a - \gamma_a) A_{c_{1n}}
\]
\[
\frac{dA_{c_{2(n+1)}}}{dt} = 2\sigma_a A_{c_{2n}} + (\sigma_a - \delta_a - \gamma_a) A_{c_{2(n+1)}},
\]

where \( i \in \{1, 2\} \) and \( j \in \{1, 2, \ldots n - 1\} \) if \( i = 1 \), or \( j \in \{1, 2, \ldots n\} \) if \( i = 2 \). Depending on the assumed BDT, the initial conditions for chase are determined by Table 3.3. It is interesting to note that the equations for uptake and chase are almost identical. The major difference is that the equations for unlabelled cells are mirrored, \( A_{u_0} \) in the case of uptake; \( A_{c_{1n}} \) and \( A_{c_{2(n+1)}} \) in the case of chase. Under homeostasis the level of \( A \) will remain constant on average so that we’ll have \( \sigma_a = \delta_a + \gamma_a \). In this case the equations simplify to an asymmetric division case like that in Chapter 2 with \( 2\sigma_a = \alpha \).

3.3.3 Two-population model parameters and assumptions

The two-population model defines HSC population heterogeneity by assuming and additional dormant cell population (\( D \)) whose differentiated daughter cells enter the \( A \) population (right panel of Figure 3.1). The main purpose of the two-population model is to create a platform on which to test the d-HSC hypothesis. In addition to heterogeneity in the HSC population, the d-HSC hypothesis also postulates that d-HSCs can reversibly switch between a dormant and active (a-HSC) state (Wilson et al., 2008). Even the most simple mathematical description of dynamical activation and deactivation of d-HSCs (i.e. feedback) would require at least one nonlinear term (a Hill function for example) and
hence introduce at least two additional unknown parameters that need to be estimated. In order to limit the number of free parameters (so that estimation for the amount of data available remain accurate and reliable) we did not explicitly implement reversible $D$ to $A$ activation. Our three stage estimation strategy (Section 3.5.1) did however enable implicit modelling of the activation and deactivation of $d$-HSCs so that the model remained useful in this context. Nevertheless, compared to the one-population model, this model does introduce a considerable amount of new parameters. First of all the active and dormant proportions (AP and DP respectively) need to be estimated. Secondly, both cell populations have their own kinetic rates. We thus have (in order of self-renewal, differentiation, and apoptosis) $\sigma_d, \delta_d, \gamma_d$ for the $D$ population and $\sigma_a, \delta_a, \gamma_a$ for the $A$ population. All notation for uptake and chase equations follows the same convention as described for the one-population model above.

### 3.3.4 Two-population model equations

The two-population model is therefore simply an extended one-population model where we allow for heterogeneity in cellular kinetics. In addition the two subpopulations are coupled by letting differentiated progeny of one subpopulation enter the other. Although a simple extension, the two-population model has more than double the number of free parameters in comparison to the one-population model. Fortunately, by avoiding an activation term, all ODEs are still linear and parameters first order so that closed form solutions can be found for both models. Moreover, stochastic modelling using parameter values optimised here are also simpler as we will explain in Chapter 4.

#### Uptake equations

\[
\begin{align*}
\frac{dD_{u0}}{dt} &= -(\sigma_d + \delta_d + \gamma_d)D_{u0} \\
\frac{dD_{ui}}{dt} &= 2\sigma_d D_{ui-1} - (\sigma_d + \delta_d + \gamma_d)D_{ui} \\
\frac{dD_{un}}{dt} &= 2\sigma_d D_{un-1} + (\sigma_d - \delta_d - \gamma_d)D_{un} \\
\frac{dA_{u0}}{dt} &= \delta_d D_{u0} - (\sigma_a + \delta_a + \gamma_a)A_{u0} \\
\frac{dA_{ui}}{dt} &= \delta_d D_{ui} + 2\sigma_a A_{ui-1} - (\sigma_a + \delta_a + \gamma_a)A_{ui} \\
\frac{dA_{un}}{dt} &= \delta_d D_{un} + 2\sigma_a A_{un-1} + (\sigma_a - \delta_a - \gamma_a)A_{un},
\end{align*}
\]
for \( i \in \{1, 2, \ldots n - 1\} \). We see that the equations for \( D \) are similar to those of \( A \) in the one-population model. Here the \( A \) equations differ from the one-population version in the differentiated progeny of \( D \) that enters \( A \) at a rate of \( \delta_d \).

**Chase equations**

\[
\begin{align*}
\frac{dD_{ci0}}{dt} &= -(\sigma_d + \delta_d + \gamma_d)D_{ci0} \\
\frac{dD_{cij}}{dt} &= 2\sigma_d D_{c(i-1)} - (\sigma_d + \delta_d + \gamma_d)D_{cij} \\
\frac{dD_{c1n}}{dt} &= 2\sigma_d D_{c(n-1)} + (\sigma_d - \delta_d - \gamma_d)D_{c1n} \\
\frac{dD_{c2(n+1)}}{dt} &= 2\sigma_d D_{c2n} + (\sigma_d - \delta_d - \gamma_d)D_{c2(n+1)} \\
\frac{dA_{ci0}}{dt} &= \delta_d D_{ci0} - (\sigma_a + \delta_a + \gamma_a)A_{ci0} \\
\frac{dA_{cij}}{dt} &= \delta_d D_{cij} + 2\sigma_a A_{c(i-1)} - (\sigma_a + \delta_a + \gamma_a)A_{cij} \\
\frac{dA_{c1n}}{dt} &= \delta_d D_{c1n} + 2\sigma_a A_{c(n-1)} + (\sigma_a - \delta_a - \gamma_a)A_{c1n} \\
\frac{dA_{c2(n+1)}}{dt} &= \delta_d D_{c(n+1)} + 2\sigma_a A_{c2n} + (\sigma_a - \delta_a - \gamma_a)A_{c2(n+1)},
\end{align*}
\]

(3.3.4)

where \( i \in \{1, 2\} \) and \( j \in \{1, 2, \ldots n - 1\} \) if \( i = 1 \), or \( j \in \{1, 2, \ldots n\} \) if \( i = 2 \). From these equations it is clear why an assumption of asymmetric division would have caused difficulties. All the differentiated daughters of \( D \) enter the \( A \) population which will thus not be able to maintain constant numbers under an asymmetric division setting. Parameter constraints, if homeostasis is assumed, are \( \sigma_d = \delta_d + \gamma_d \) and \( \sigma_a = \delta_a + \gamma_a - \pi\delta_d \) where \( \pi = \frac{DP}{AP} = \frac{DP}{1-PP} \).

It is worthwhile to note that there are other ways to implement differentiation to what we have shown here. Here we have assumed that differentiation is independent of cell division which means cells can differentiate without having to divide. Alternatively we can assume (like asymmetric cell division) that a differentiation event is linked to a cell division event, i.e. immediately after a cell has divided one or both daughter cells differentiate. In the case of modelling BrdU kinetics like we have, this means differentiation would change a cell’s label intensity - just like self-renewal. Our choice of differentiation being independent from cell division is motivated by the advantages of having only one possible kinetic event (self-renewal) changing a cell’s label intensity.
3.4 Analytic solutions

All the above ODEs are linear and thus have analytic solutions in the form of constrained nonlinear multivariate functions. The sets of chase equations are coupled to the uptake equations by partitioning of the uptake cells at chase day 0 into chase group $c_1$ or $c_2$ depending on the BDT we model. Analytical expression of Equations 3.3.1, 3.3.2, 3.3.3 and 3.3.4 can be derived quite conveniently by writing each system of ODEs in matrix form as

$$\dot{y} = K \cdot y,$$

where $y$ is the vector of cell species for which equations are defined in each of the four systems, $\dot{y}$ contains the derivatives of these variables and $K$ is the coefficient matrix. The solution can be written directly in analogy to that of first order differential equations

$$y(t) = \exp(K \cdot t) \cdot y(0),$$

where $\exp$ is the matrix exponential function. The analytical expression of Equations 3.3.1, 3.3.2, 3.3.3 and 3.3.4 only differ in their version of $y$ and $K$ and these are given below.

3.4.1 One-population model solutions

The expression for $y$ and $K$ during uptake is

$$y = \begin{bmatrix} A_{u0} \\ A_{u1} \\ A_{u2} \\ \vdots \\ A_{un} \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} -\Phi & 0 & 0 & \cdots & 0 \\ 2\sigma_a & -\Phi & 0 & \cdots & 0 \\ 0 & 2\sigma_a & -\Phi & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\sigma_a & (2\sigma_a - \Phi) \end{bmatrix}.$$

During chase, for cells loosing labelling after $n$ divisions, they are

$$y = \begin{bmatrix} A_{c10} \\ A_{c11} \\ A_{c12} \\ \vdots \\ A_{c1(n-1)} \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} -\Phi & 0 & 0 & \cdots & 0 \\ 2\sigma_a & -\Phi & 0 & \cdots & 0 \\ 0 & 2\sigma_a & -\Phi & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\sigma_a & -\Phi \end{bmatrix},$$

where $\Phi = \sigma_a + \delta_a + \gamma_a$ in both cases above. The expression for $y$ and $K$ during chase, for cells loosing labelling after $n + 1$ divisions, follow in a similar fashion.
3.4.2 Two-population model solutions

Uptake expressions:

\[ y = \begin{bmatrix} D_{u_0} & D_{u_1} & \cdots & D_{u_n} & A_{u_0} & A_{u_1} & \cdots & A_{u_n} \end{bmatrix} \]'^

\[
\begin{bmatrix}
-\Phi_d & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
2\sigma_d & -\Phi_d & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 2\sigma_d & -\Phi_d & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 2\sigma_d & (2\sigma_d - \Phi_d) & 0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]

Chase expressions, for cells losing labelling after \( n \) divisions:

\[
y = \begin{bmatrix} D_{c_{10}} & D_{c_{11}} & \cdots & D_{c_{1(n-1)}} & A_{c_{10}} & A_{c_{11}} & \cdots & A_{c_{1(n-1)}} \end{bmatrix} '\
\]

\[
\begin{bmatrix}
-\Phi_d & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
2\sigma_d & -\Phi_d & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 2\sigma_d & -\Phi_d & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 2\sigma_d & -\Phi_d & 0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]

\[ \Phi_d = \sigma_d + \delta_d + \gamma_d \text{ and } \Phi_a = \sigma_a + \delta_a + \gamma_a \text{ in both cases above.} \]

The expressions for cells losing labelling after \( n + 1 \) divisions follow in a similar fashion. Note that for both the one-population and two-population chase phase solutions we do not include the expressions for \( D_{c_{1n}} \) or \( A_{c_{1n}} \) \( (D_{c_{2(n+1)}} \text{ or } A_{c_{2(n+1)}} \text{ similarly}) \) since we only need to track the labelled proportion of the cell population during chase.
3.5 Simulation results

Two ODE models were defined in Section 3.3.2 and 3.3.4 respectively to ascertain which one of two hypotheses best describes the observed BrdU kinetics reported in Tables 3.1 and 3.2. Each model is a system of coupled ODEs that describe the rate of change of the proportion of BrdU labelled cells over time. In this section we relate the process and results of finding the unknown parameters so that each model describes the time evolution of the observed data optimally. Since all ODEs have analytic solutions in the form of constrained nonlinear multivariate functions the problem can be formulated as finding which one of two constrained nonlinear multivariate functions (sources) are most likely to have generated the empirical label-retaining data: is it best described by a single homogeneous population source, or by a source that is a combination of a dormant and a more rapidly dividing cell type? Non-linear optimisation algorithms to solve problems such as these are by now well established.

3.5.1 Parameter estimation

![Figure 3.5: Observed experimental data of BrdU uptake and chase.](image)

In Figure 3.5 we replot the average observed data as a time series for visual analysis with the uptake data on a timescale of hours for improved readability. As mentioned earlier parameters were assumed to be constant and constrained to maintain a steady state. By this assumption we thus effectively estimated the average turnover rates over
particular periods in time. If HSC dynamics are constant, as during homeostasis, this approach poses no problem and serves as a convenient description of the net effect of state-dependant (i.e. non-constant) parameters. However, careful study of the observed LRC data in Figure 3.5 reveals 5 points in time where a peculiar change in kinetic slope is apparent. These are indicated by black arrows. Clearly visible is an initial lag phase after BrdU is first added to the system. After 24 hours a sudden increase in the BrdU uptake rate can be observed (first arrow), which extends until 312 hours (13 days) when BrdU is removed (second arrow). At this time point the cells immediately start to dilute their label as they continue to divide. Hence the percentage of BrdU\(^+\) cells rapidly decreases until day 10 of chase (third arrow) when it seems that the BrdU dilution rate decreases as the curve flattens out. Each one of these three time points just noted has a biological interpretation. The most obvious is the second arrow which indicates the time when BrdU is removed at 312 hours and the cells immediately start to dilute out the label. The initial lag phase (0-24h) during uptake followed by the sudden increased incorporation of BrdU after 24 hours of pulse (the first arrow) is however much less obvious. As discussed in Section 3.2.1 experimental evidence exists that BrdU have toxic effects on cycling cells, hence a proliferative signal is most likely induced in HSCs (commencing around 24h) in response to peripheral injury caused by BrdU. Indeed, it has biologically been confirmed that this is the most likely reason for the changes in proliferation kinetics at this point (Wilson et al., 2008). Importantly, this proliferative burst is also the most likely explanation as to why dormant HSCs can be efficiently labelled with BrdU (around 90% after 13 days) in the first place.

Similarly, when BrdU is removed at 312 hours we first observe a rapid loss of BrdU label over about 10 days before HSC proliferation returns to “normal” rates. This initial rapid loss of label is also a consequence of the toxic affects of BrdU on the periphery, as HSCs are still cycling in response to injury signals mediated by the presence of BrdU. Once the BrdU is removed, it may take several days for the injury status to resolve and for homeostasis to be re-established.

Unlike the first 3 time points discussed above, there is no known biological event that can be attributed to the change occurring at day 70. However, the change of kinetic slope at this time point can be regarded as the starting point of the label retaining tail we observe. Since we have implemented measures to control for the observed tail (Section 3.2.4), we are satisfied that the alteration in kinetics of BrdU loss at day 70 can readily be accounted for by the model. At day 177 an unexpected drop in BrdU retention can be observed. This is most likely due to non-specific external influences on the mice (such as a mild infection) during the long chase period which activated the d-HSCs. The change of slope at chase day 177 is different from the previous four in that it is a solitary change
with the subsequent time points returning to the kinetic slope as observed prior to day 177. Thus the observed data at chase day 177 could be regarded as an outlier particularly as omitting it had negligible effects on parameter estimation.

**Three stage parameter estimation strategy**

From the discussion above we thus motivate parameter estimation in three regions, rather than one set of parameter estimates for all time points. Our three-stage parameter estimation strategy proceeded as follows:

1. Starting with 0% of the cells labelled, we fitted parameters to uptake data observed in the first 24 hours. These parameters are an estimate of the homeostatic proliferation rates, but it has to be kept in mind that little information can be conveyed by the only three observed data points (excluding day 0), hence these estimates carry less weight than the homeostatic estimates of stage 3.

2. Continuing with the distribution predicted by stage 1 as the initial condition, we estimated a new set of parameters for the rest of the uptake data and the first 10 days of chase (shaded in grey in Figure 3.5). Since we hypothesised d-HSC to be activated in this stage we expected an increase in parameter estimates ("hematopoietic stress" proliferation rates). In this stage there is a switch between the BrdU uptake and chase phases. Since the BDT has a profound effect during chase the cells have to be partitioned into two groups at the onset of chase - those that lose labelling after \( n \) divisions and those that lose labelling after \( n + 1 \) divisions. See the discussion in Section 3.2.3 for more details.

3. Finally we estimated parameters for the rest of the chase data, from day 10 onwards. The initial labelling proportions amongst the two partitions are obtained from the day 10 predictions in stage 2 so that we maintain parameter integrity between the three different estimate regions. In this region we assumed that HSC turnover has settled back to homeostatic rates, hence activated dormant cells, if present, would have switched back to a dormant state.

**3.5.2 Model evaluation**

As in Chapter 2 we will here also use the RSS goodness-of-fit statistic (Equation 2.4.1) to evaluate consistency of the model with empirical data. Better models have smaller RSS values and hence are more consistent with experimental data. In order to make use of nonlinear optimisation algorithms to search for optimal parameters we needed to
3.5. Simulation results

reformulate our problem as finding the parameter vector $\mathbf{p}$ that locally minimises a scalar objective function $f(\mathbf{p})$ such that the following three constraints hold:

$$
\mathbf{A} \cdot \mathbf{p} \leq \mathbf{b}, \quad \mathbf{A}_{eq} \cdot \mathbf{p} = \mathbf{beq}, \quad 1 \leq \mathbf{p} \leq \mathbf{u}.
$$

The constrained nonlinear multivariate functions representing each model (Equation 3.4.2 with the corresponding expressions for $\mathbf{y}$ and $\mathbf{K}$) could easily be transformed to such an objective function that calculates the RSS for a given set $\mathbf{p}$ of parameters, as follows:

$$
f(\mathbf{p}) = (\mathbf{e} - \hat{\mathbf{y}}) \cdot (\mathbf{e} - \hat{\mathbf{y}})',
$$

where $\mathbf{e}$ is the row vector of observed values (mean values reported in Tables 3.1 and 3.2) and $[\hat{\mathbf{y}}]_i = [1 \cdot \mathbf{y}(t_i)]$ for $\mathbf{1}$ a $1 \times n$ row vector of ones ($n$ the number of elements in $\mathbf{y}(t_i)$) and $t_i \in \{ \text{time points of data in } \mathbf{e} \}$. Here $\mathbf{y}(t_i)$ is the vector of different levels of labelled cells as predicted by the model. For example, for the one-population model

$$
\mathbf{y} = \begin{bmatrix}
A_{u1} & A_{u2} & \cdots & A_{un}
\end{bmatrix}',
$$

when uptake is modelled and

$$
\mathbf{y} = \begin{bmatrix}
A_{c10} & A_{c11} & \cdots & A_{c1(n-1)} & A_{c20} & A_{c21} & \cdots & A_{c2n}
\end{bmatrix}',
$$

when chase is modelled. To perform the optimisation we used the Active-Set algorithm as implemented by Matlab’s \texttt{fmincon()} function. This algorithm uses a sequential quadratic programming method (Gill et al., 1984; Fletcher, 1987), the details of which are beyond the scope of this thesis.

Using heuristic search to find local minima over the parameter hyperspace holds considerable advantages over the brute-force search best option available for the agent-based model of Chapter 2. Not only could much larger areas of the search space be covered in a short time, but results are also guaranteed to be at least a local best. With the agent-based model there is no such guarantee. It is however quite possible that many (different) parameter sets exist for which $f(\mathbf{p})$ has a local minimum. To improve our chances of finding the global minimum, we iterated through the possible values of DP and BDT, each time optimising over $\sigma_a, \delta_a$, and $\gamma_a$ (for the one-population model) or $\sigma_d, \delta_d, \gamma_d, \sigma_a, \delta_a$, and $\gamma_a$ (for the two-population model). Each optimisation iteration was repeated 100 times, starting at a new random position in the parameter landscape. When many local minima were found the lowest was recorded for the relevant DP and BDT.

In order to compare the goodness-of-fit during the three stages of parameter optimisation, we additionally computed a normalised measure of the RSS, the Mean Squared Error ($MSE^k = \frac{RSS}{n_k}$), where $n_k$ is the number of data points of stage $k$. In addition to a low RSS, we were seeking parameter sets that would give comparable MSE values across each three optimisation regions, indicating consistency.
3.5.3 Model predictions

Our search for optimal parameters were subject to the following constraints

\[ l = \begin{bmatrix} 0.0001 & 0.0001 & 0.0001 \end{bmatrix}, \quad u = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \]

\[ A_{eq} = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}, \quad b_{eq} = \begin{bmatrix} 0 \end{bmatrix}, \]

for the one-population model \((p = [\sigma_a \ \delta_a \ \gamma_a])\) and

\[ l = \begin{bmatrix} 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0001 \end{bmatrix}, \quad u = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \]

\[ A = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \]

\[ A_{eq} = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & 0 \end{bmatrix}, \quad b_{eq} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \]

for the two-population model \((p = [\sigma_d \ \delta_d \ \gamma_d \ \sigma_a \ \delta_a \ \gamma_a])\). The equality constraints \(A_{eq}\) and \(b_{eq}\) are just the matrix representations of the homeostatic constraints already mentioned in Section 3.3.2 and 3.3.4. The inequality constraints \(A\) and \(b\) for the two-population model specify some biological properties, namely that d-HSCs self-renewal must be slower or equal to a-HSC self-renewal and differentiation faster or equal to apoptosis. Having these constraints considerably improved optimisation since the number of free parameters (apart from DP and BDT) are reduced to one \((\sigma_a)\) for the one-population model and three \((\sigma_d, \delta_d, \sigma_a)\) for the two-population model. We will discuss prediction results of stages 1 & 2 independently from stage 3 since the different optimisation regions could be modelled with various degrees of success by each model. When plotting simulation results we take a slightly different approach to that of Chapter 2 by not plotting predictions for various BDTs under the same cell division parameter set. The reason is two-fold. Firstly because we have evaluated at least 10 BDTs and plotting predictions for each of them would severely obscure plot readability. Secondly, our optimisation strategy has been different since we optimised cell division parameters for fixed BDT-DP combinations and hence only report the best performing sets of these.

For the fixed BDT-DP sets, we optimised cell division parameters for BDTs ranging from 1 to 10 and DP ranging from 5% to 50%. The reason why the BDT couldn’t be optimised with the other parameters is because it is discrete (the optimisation algorithms require that \(f(p)\) be a continuous function) and also since model equations change for different BDTs. Similarly parameter constraints depend on DP \((\pi)\) in \(A_{eq}\) above and it must therefore be fixed before optimisation can be performed.
3.5. Simulation results

Table 3.4 summarises the best performing parameter sets for each of the one-population and two-population models based on the RSS measure. From these results it is clear that the one-population model does not fit the observed data nearly as well as the two-population model whose best RSS value of 71.7 is more than five times lower than that of the best one-population RSS value of 391.9. The best two-population model (highlighted in bold in Table 3.4) suggests a BDT of 6 or 7 DNA strands which both map to the same 3-4 division partitioning (Figure 3.4 and Table 3.3). Not only does this parameter set result in the lowest RSS, but its MSE for stage 2 and 3 is also consistent. Moreover, the stage 2 estimate for the d-HSC self-renewal rate $\sigma^d_{ii}$ and a-HSC self-renewal rate $\sigma^q_{ii}$ are similar - supporting the idea that d-HSCs are activated to self-renew at the a-HSC rate during stage 2. Indeed all d-HSC self-renewal estimates in Table 3.4 clearly predict a resting-activated-resting scheme for stage 1, stage 2 and stage 3 respectively indicating that activation of d-HSCs is reversible.

First and second stage predictions

The best stage 1 & 2 prediction for both the one-population and two-population model is plotted in Figure 3.6 on a timescale of hours. For the two-population model we also plotted the d-HSC (red) and a-HSC (green) BrdU$^+$ percentages. It is quite apparent

**Figure 3.6: Model predictions of BrdU content during stage 1 & 2.** Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. Predictions are shown on a timescale of hours and the BrdU chase period starts at 312 hours (13 days). Clearly both models can satisfactorily describe stage 1 & 2 data (uptake and first 10 days of chase). **A:** One-population model prediction (stage 1 MSE = 1.54, stage 2 MSE = 6.56). **B:** Red line: d-HSCs; green line: a-HSCs. Two-population model prediction with a 6 or 7 strand BDT and 40% DP (stage 1 MSE = 1.5, stage 2 MSE = 3.89).
that both models are very effective in modelling this part of the data. In fact, there seems to be very little difference between the predictions of the two models, especially for the BrdU uptake phase (dashed line). In spite of the fact that the one-population model is able to satisfactorily describe stage 1 & 2 data, the d-HSC hypothesis is not contradicted here since we can indeed expect to observe a homogeneous rate of BrdU uptake and dilution in both cases. To understand the reason why, it is useful to carefully dissect what constitutes what we are observing at each point in time. The crux is that we are only observing labelled cells and hence dynamics are that of labelled cells only - even if there are heterogeneous cell populations present. During stage 1 (0-24h) the d-HSCs have not been activated yet so we are mostly observing the a-HSCs, allowing the one-population model to present a good fit. Conversely we also start observing the d-HSCs after 24 hours (stage 2) since they have now been activated and hence started taking up BrdU. But even though we are now observing two heterogeneous cell populations, the rate of BrdU uptake and dilution still appear homogeneous since the activated d-HSCs are dividing at an increased rate, most likely close to that of the a-HSCs.

It is most interesting to note that the one-population model predicts a lower value for the last time point (day 10 of chase). The two-population model in contrast predicts this value almost exactly. This is confirmation that we are starting to observe heterogeneity around day 10. It would be insightful to do more experiments between chase day 3 and 10 (i.e. between 384h and 552h in Figure 3.6) to get a better estimate of where exactly (or at what rate) activated d-HSCs return to dormancy.

**Third stage predictions**

The predictions for all the data are shown in Figure 3.7, with the focus on stage 3 data since the timescale is in days. In contrast to stage 1 & 2 data (Figure 3.6), the two-population model can describe the stage 3 data much more satisfactorily than the one-population model. In stage 3 the d-HSCs have switched back to a resting state, but unlike in stage 1 we do observe them now since most of the d-HSCs have been labelled during stage 2. In stage 3 there is therefore heterogeneity in the BrdU dilution rate which cause the one-population model to fail completely in describing the data. In contrast the two-population model describes the stage 3 data extremely well. The fact that the one-population model can successfully describe the uptake data but not the chase data is in agreement with our analysis earlier that the discriminative power for heterogeneity only lies in the chase data.
Table 3.4: Goodness-of-fit results and parameter estimates. BDT: BrdU detection threshold, DP: Dormant proportion, RSS: Residual Sum of Squares, MSE: Mean Squared Error, superscripts indicate modelling stage 1, 2 or 3. The DP of 0% indicates the one-population model. Parameter values $\sigma_d$, $\delta_d$ and $\sigma_a$ are inverted to units of days. The best overall parameter set is highlighted in bold. Estimates for $\frac{1}{\sigma_d^i}$ are always equal to the lower bound $\frac{1}{24 \times 0.0002}$ which simply indicate that d-HSCs are practically not cycling during the first 24h.

<table>
<thead>
<tr>
<th>BDT</th>
<th>DP</th>
<th>RSS</th>
<th>MSE$^i$</th>
<th>MSE$^ii$</th>
<th>MSE$^iii$</th>
<th>$\frac{1}{\sigma_d}$</th>
<th>$\frac{1}{\delta_d}$</th>
<th>$\frac{1}{\sigma_a}$</th>
<th>$\frac{1}{\sigma_d}$</th>
<th>$\frac{1}{\delta_d}$</th>
<th>$\frac{1}{\sigma_a}$</th>
<th>$\frac{1}{\sigma_d}$</th>
<th>$\frac{1}{\delta_d}$</th>
<th>$\frac{1}{\sigma_a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0%</td>
<td>391.9</td>
<td>1.54</td>
<td>6.56</td>
<td>27.54</td>
<td>26.1</td>
<td>10.4</td>
<td>95.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30%</td>
<td>131.8</td>
<td>1.51</td>
<td>11.7</td>
<td>4.31</td>
<td>208.3</td>
<td>16.2</td>
<td>16.9</td>
<td>7.2</td>
<td>122.2</td>
<td>146.2</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30%</td>
<td>113.8</td>
<td>1.51</td>
<td>9.42</td>
<td>3.94</td>
<td>208.3</td>
<td>15.6</td>
<td>15.6</td>
<td>7.6</td>
<td>134.4</td>
<td>136.3</td>
<td>26.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30%</td>
<td>87</td>
<td>1.51</td>
<td>5.79</td>
<td>3.22</td>
<td>208.3</td>
<td>13.8</td>
<td>27.6</td>
<td>8.2</td>
<td>150.2</td>
<td>156.4</td>
<td>32.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 &amp; 7</td>
<td>30%</td>
<td>80.6</td>
<td>1.51</td>
<td>4.07</td>
<td>3.6</td>
<td>208.3</td>
<td>11.3</td>
<td>11.3</td>
<td>9.3</td>
<td>189.9</td>
<td>193.6</td>
<td>35.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 &amp; 7</td>
<td>35%</td>
<td>73.6</td>
<td>1.5</td>
<td>3.97</td>
<td>3.25</td>
<td>208.3</td>
<td>10.5</td>
<td>10.5</td>
<td>9.6</td>
<td>170.3</td>
<td>173.3</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 &amp; 7</td>
<td>40%</td>
<td>71.7</td>
<td>1.5</td>
<td>3.89</td>
<td>3.22</td>
<td><strong>208.3</strong></td>
<td><strong>16.1</strong></td>
<td><strong>10.1</strong></td>
<td><strong>9.8</strong></td>
<td><strong>165.1</strong></td>
<td><strong>326.6</strong></td>
<td><strong>30.5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 &amp; 7</td>
<td>45%</td>
<td>73.7</td>
<td>1.5</td>
<td>3.8</td>
<td>3.41</td>
<td>208.3</td>
<td>9.8</td>
<td>10.1</td>
<td>9.8</td>
<td>149.1</td>
<td>298.1</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>35%</td>
<td>95.2</td>
<td>1.5</td>
<td>3.85</td>
<td>4.93</td>
<td>208.3</td>
<td>10.2</td>
<td>20.3</td>
<td>10.2</td>
<td>193.9</td>
<td>197.8</td>
<td>40.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.5. Simulation results

Figure 3.7: Model predictions of BrdU content during stage 3. Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. A: One-population model predictions with a 9 strand BDT. (RSS = 391.9, stage 3 MSE = 27.54). B: Red line: d-HSCs; green line: a-HSCs. Two-population model predictions with a BDT of 6 or 7 strands and 40% DP. This model gave the best overall goodness-of-fit (RSS = 71.7, Stage 3 MSE = 3.22). Activation of the d-HSCs can be clearly seen in Figure 3.6B where both d-HSC and a-HSC are predicted to take BrdU up at the same rate (cycling about once every 10 days). During chase d-HSCs return to a dormant state and are predicted to divide about once every 165 days, whilst a-HSCs divide once every 31 days, diluting label much faster than the d-HSCs.

The green and red line show how the BrdU kinetics would look like if we were able to trace or isolate the a-HSC (green) and d-HSC (red) populations independently. Interestingly the d-HSC and a-HSC predictions for stage 2 are quite similar (Figure 3.6). In fact, by the end of uptake the a-HSC and d-HSC populations have equal proportions of cells labelled. This is most likely because d-HSCs have been activated to self-renew at the same rate as a-HSCs (about once every 10 days) and are also almost half (40%) of all the cells we observe. For comparison we show the best prediction for a smaller d-HSC population (30% DP, 4 strand BDT) in Figure 3.8. Here the lower BrdU uptake rate for d-HSCs compared to a-HSCs is clearly visible, probably due to the smaller d-HSC population but also because the activated d-HSC self-renew rate is almost half that of the a-HSC rate. Visually it seems as if the stage 2 predictions are quite good, however, the relatively high MSE\textsuperscript{ii} of 9.42 clearly show that this is not actually the case - even the one-population model is better here (recall the one-population MSE\textsuperscript{ii} of 6.56). Other than a slightly higher label retention for d-HSCs in the 4 BDT - 30% DP prediction, the stage 3 predictions of Figures 3.7B and 3.8B seems quite similar, indeed the MSE\textsuperscript{iii} for the 4 BDT - 30% DP model (3.94) and the 6 & 7 BDT - 40% DP model (3.22) are both below 4. An interesting and clear pattern that can be observed is that uptake data (in particular stage
3.5. Simulation results

Figure 3.8: Two-population model predictions with a 4 strand BDT and 30% DP. Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. A: Stage 1 & 2 predictions (uptake and first 10 days of chase) on a timescale of hours. The effect of a smaller d-HSC population is clearly visible (stage 1 MSE = 1.51, stage 2 MSE = 9.42). B: Red line: d-HSC; green line: a-HSC. Uptake and chase predictions on a timescale of days (RSS = 113.8, stage 3 MSE = 3.94).

2) favour larger d-HSC proportions with homogeneous turnover rates between d-HSCs and a-HSCs. In the study of Wilson et al. (2008) where only chase data were modelled for example, the d-HSC proportion were estimated to be 15% which is much lower than our estimate here using both uptake and chase data. This also gives one explanation why previous studies that only considered BrdU uptake excluded the existence of a dormant HSC population (Bradford et al., 1997; Cheshier et al., 1999).

Plotting the proportion of chase group \( c_1 \) and chase group \( c_2 \) amongst BrdU\(^+\) cells during uptake for the two models we have been focussing on in the last paragraph provides interesting insights (Figure 3.9). Figure 3.9A gives the \( c_1 \) and \( c_2 \) prediction during uptake for our 6 & 7 BDT - 40% DP best model. Here \( c_1 \) cells (yellow) have divided once during uptake and \( c_2 \) (green) cells divided 2 or more times during uptake (the largest proportion of cells). Cells in \( c_1 \) will dilute their label in 3 divisions and those in \( c_2 \) in 4 divisions. The majority of cells (according to our parameter set of choice) are thus predicted to loose BrdU labelling in 4 divisions. Figure 3.9B in turn gives the \( c_1 : c_2 \) ratio for the best 4 BDT - 30% DP model. Here \( c_1 \) represent the cells that have divided once or twice in the presence of BrdU with \( c_2 \) representing those that divided 3 or more times. In this case \( c_1 \) cells will dilute their BrdU in 4 divisions whilst it will take 5 divisions for \( c_2 \) cells (Table 3.3). It is interesting to note that the relative BrdU\(^+\) d-HSC proportion for this model stays almost stationary during the first 10 days of chase (Figure 3.8A). One possible explanation is that at a self-renew rate of once in 15 days, d-HSCs cycle twice
3.6 Conclusion

Although the biological evidence per se in Wilson et al. (2008) was strongly in favour of the d-HSC hypothesis, human interpretation of especially the BrdU label-retaining cell data was in need of more formal justification. Moreover, BrdU data are typically obscured by many confounding factors (discussed in Section 3.2.1) and can thus be misleading when interpreted directly (i.e. without computational modelling).

Our mathematical models developed here provide additional support for the d-HSC hypothesis. We showed that amongst the two hypotheses considered in this thesis (Section 1.3.3), those models that assume the existence of a slowly cycling subpopulation were able to describe the observed experimental data the most satisfactorily. We then used these two-population models to elucidate the d-HSC population size and rate of division. Many of the concepts developed here have already emerged in Chapter 2 where a different

**Figure 3.9: Proportion of c1 and c2 in BrdU\(^+\) cells during uptake.** A: 6 & 7 BDT - 40% DP. Here \(c_2\) represents cells that have divided 2 or more times. B: 4 BDT - 30% DP. In this case \(c_2\) represents cells that have divided 3 or more times

as slow as a-HSCs (once in 7.6 days). More importantly, it takes half of the BrdU\(^+\) cells \((c_2)\) 5 divisions to dilute BrdU below detection, so it can be expected that few of them will divide so many times in only 10 days.

In summary our parameter estimates (based on a maximum RSS of 80.6) are:

- **BDT**: 6-7 strands (3-4 divisions)
- **d-HSC proportion**: 30-45%
- **d-HSC self-renew during stage 3**: once every 149-193 days (putative homeostatic rate)
- **a-HSC self-renew during stage 3**: once every 28-36 days
model and dataset have shown support for heterogeneity in both time as well as the HSC population. It is important to keep in mind that the d-HSC hypothesis does not only postulate heterogeneity in primitive HSCs, a phenomenon that has been observed in vitro before (Bernard et al., 2003). It additionally states the d-HSCs are quiescent during normal homeostasis but can be reversibly activated to proliferation during injury. Interestingly the inferred optimal rates of division here also support this notion that the d-HSC can be reversibly activated.

Parameter estimation indicated that at least a third of the HSC population are d-HSCs that divide about once every 149-193 days with a-HSCs dividing once every 28-36 days. We further estimated that mouse HSCs must have at least 6-7 DNA strands labelled before they will have a strong enough intensity to be detected as BrdU+. Our corresponding BDT estimate for the dataset in Chapter 2 was 20 strands. The discrepancy makes sense because BrdU detection in Kiel et al. (2007) was done by hand (immunofluorescence microscopy), whilst here the latest FACS (flow cytometry) technology was used. Detection is hence much more sensitive.

### 3.6.1 Homogeneous parameters

When homogeneous parameter values are assumed for the entire dataset, optimisation simply collapses - it is impossible to fit both the uptake and chase profiles with a homogeneous turnover rate, even with a heterogeneous cell population (Figure 3.10). We thus adopted a three-stage parameter estimation strategy making sure each stage is based on careful biological motivations. Since the LRC dataset at our disposal is the most detailed

![Figure 3.10: Model predictions using homogeneous parameters. Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. A: One-population model. B: Two-population model.](image)

...
to date, and due to our thorough and novel treatment of the correlation between the BDT and number of uptake divisions (Figure 3.4), we were able to estimate parameters very accurately using state of the art nonlinear optimisation algorithms in combination with our three-stage strategy. This is evident in the good fit (Figures 3.6B and 3.7B) and low RSS values (Table 3.4) we were able to achieve. Our estimates for the d-HSC and a-HSC division rates are nevertheless more or less in agreement with the chase only model estimates in Wilson et al. (2008), although we found a larger range of values to fit the data.

3.6.2 More than two populations

An obvious question that arises is, given the results in this chapter, what the possibility of having more than two HSC subpopulations is? Put in another way, is it possible that the kinetic distribution between a dormant and active state is continuous? Whilst we cannot totally exclude these possibilities our results indicate it to be highly unlikely. First of all, just increasing the assumption from one population to two has resulted in a dramatic improvement in the goodness-of-fit. The simplest three-population model will introduce a further 3 free parameters bringing the total number of parameters that need to be estimated to 8. There is simply not enough support in the data for this number of free parameters and the end result would most probably be overfitting. Secondly, the results here strongly support a bimodal distribution between a dormant and an active state. This is evident in the large difference between the homeostatic turnover rate estimates of the d-HSCs vs. the a-HSCs. Since we used an ODE model to estimate these rates, it implies that we have estimated the average turnover rate of a population. Our only input to the optimisation procedure was that there is the possibility of having two distinct rates. We did not specify how far the estimates must be or even that they have to be different (stage 2 estimates for d-HSC and a-HSC are often equal, see Table 3.4). An elaborate search in parameter space, which most probably will have included consideration of closer d-HSC and a-HSC turnover estimates, found values that are significantly distinct from each other (d-HSCs typically around 5 times slower than a-HSCs). This clearly indicates a bimodal switch-like behaviour between the dormant and active state.

3.6.3 Further work

As discussed in Section 3.2.4 there is another possible explanation for the observed long-term label retention which we were unable to address with the ODE models in this chapter: stochastic variability. We investigate this possibility in Chapter 4 where we show how master equations, that describe the label retaining cell dynamics as a stochastic kinetic
3.6. Conclusion

process, can be derived from the equations in the ODE model. We also give formal proof that the average prediction of a stochastic model defined in this way is exactly described by the ODE model from which the master equations were derived (subject to some conditions). Whilst we already defined a stochastic (agent-based) model in Chapter 2, no similar proof for this model is known, so choosing parameter sets from which to estimate variation is uncertain. To gain empirical insights into the relationship between the ODE model and agent-based model predictions and parameter sets, a parallel search of the agent-based model’s parameter hyperspace using a Grid supercomputer is performed in Chapter 5.

An interesting question that arises is whether d-HSC hematopoiesis (Figure 3.1, right panel) is evolutionarily superior to the previously widely accepted dormant-free hematopoiesis (Figure 3.1, left panel). This question is extremely difficult to answer with human reasoning alone but analysis of dynamical models of hematopoietic regulation could provide some useful insights. However, in this chapter we have focused on the modelling of LRC-based HSC kinetic data rather than the modelling of dynamic maintenance and restoration of homeostasis in the hematopoietic system. Any worthwhile dynamical analysis of the equations here to address questions such as these were thus not possible. We therefore define a related model of dynamic hematopoiesis in Chapter 6 to evaluate the impact of the d-HSC population on hematopoietic regulation.

By now we have presented all the necessary HSC and BrdU biology and the following chapters are more technically oriented.
Chapter 4

Label-retaining cell dynamics as a stochastic kinetic process

4.1 Introduction

The ODE-based LRC models described in Chapter 3 result in deterministic solutions, and hence describe average population dynamics. The main purpose was to investigate whether there is support in the experimental data for the d-HSC hypothesis by evaluating possible causes of the long-term BrdU retention observed in Figure 3.2. We concluded that the most likely cause was the presence of a slowly dividing subpopulation of HSCs that take longer to dilute BrdU during chase. Whilst we did highlight the need - under the d-HSC hypothesis - to rule out stochastic variability as a cause of the observed long-term label retention, we were unable to test this possibly because of the deterministic nature of the ODE models defined in Chapter 3. The putative role stochastic variability plays in the context of explaining label retention is illustrated in Figure 4.1. By the central limit theorem we can assume that the distribution of labelled cells, from a sufficiently large population of mice, will be approximately normally distributed at any particular time point of an LRC experiment. At each time point we examine a number of mice, hoping that our sample is representative of the population so that average BrdU retention is as close as possible to the population mean of the real distribution at this point. If it was the case that the mice we examined at a certain time point (in particular at the later stages) were all sampled from the higher end of the distribution tail (Figure 4.1), we might be falsely led to believe that there is significant long-term retention of BrdU. Certainty about the label distribution will increase by increasing the number of mice being examined. However, in practice time and cost constraints often limit sample sizes.

In this chapter we investigate the scenario depicted in Figure 4.1 by defining a stochastic model (different to the agent-based model described in Chapter 2 which is also stochas-
Figure 4.1: **Example of a possible skew sample.** The distribution of the % BrdU+ HSCs of a large mouse population would be approximately normal (blue line). If the mice that we examine (red circles) happen to fall in the tail of this normal distribution we might estimate actual BrdU retention higher (or lower) than what it really is.

It is widely accepted that stochasticity is an inherent property of biological systems and modelling them as such is an area of great interest (Goss & Peccoud, 1998; Elowitz et al., 2002; Allen, 2003; Paulsson, 2005; Wilkinson, 2006, 2009). Describing individual cells as agents and keeping track of the labelling status of their chromosomes proved to be a simple and useful stochastic model of BrdU data (Chapter 2). Its discrete properties made this approach especially effective in implementing the BDT by alleviating the need for continuous approximation and grouping of the number of labelled DNA strands in a cell. However, the relationship between the predictions of our deterministic ODE model and a stochastic agent-based model for the same set of parameters is unclear. Any inference on stochastic variance estimated by an agent-based model using parameters optimised by our ODE model is thus somewhat troublesome. Ideally we would desire a stochastic model defined over the same parameter space as the ODE-based LRC model and which would give exactly the same average predictions as the ODE model for the same set of parameter values. In this chapter we show it is indeed possible to develop such a model by defining the label-retaining cell dynamics as a Markov process. The work we present here has been published in van der Wath et al. (2009).
4.1.1 Markov process

The LRC model described in Chapter 3 can be viewed as a random walk over time moving through a multi-dimensional hyperspace of cellular species (i.e. all the $A_u^i$, $A_{c1}^i$, $A_{c2}^i$, $D_u^i$, $D_{c1}^i$, $D_{c2}^i$ for which we have defined an equation). When the next future state is fully determined by the current state of the system, independent of all previous states, the random walk adheres to the Markov property and is known as a Markov process (MP).

The Markov property is a reasonable assumption for the LRC system we are modelling since, if the current number of cells with their labelling intensities are known, no additional knowledge about the future number of the cells and their labelling states can be gained from earlier cell numbers.

Mathematically such an MP can be formulated as

$$X(t) = \mathcal{S},$$

where $\mathcal{S}$ is a vector representing the current number of each cellular species (i.e. the current state). The continuous time formulation for the probability of being in a particular state is then given by

$$P (X(t + dt) = \mathcal{S} | \{X(t') = \mathcal{S}' | t' \in [0, t]\})$$

$$= P (X(t + dt) = \mathcal{S} | X(t) = \mathcal{S}(t)) \forall t \in [0, \infty), \mathcal{S} \in \mathcal{M}, \quad (4.1.1)$$

with $\mathcal{M}$ the countable state space of the MP.

Figure 4.2 gives a cartoon of how an MP model implementation of BrdU uptake and dilution would proceed, analogous to the agent-based model cartoon of the same real time BrdU dynamics depicted in Figure 2.3. We see that each occurrence of an event (change in label intensity of any of the four cells) induces the MP to jump to a new state. Moreover, the MP model is exact since each event is accounted for in contrast to the agent-based model that evaluated event probabilities at fixed intervals of time and, as was shown in Figure 2.3, miss some events as a result.

An MP model of LRC dynamics such as depicted in Figure 4.2 is fully determined by its transition kernel $P(\mathcal{S}, t)$ and we will show in Section 4.2 how the partial derivative of the transition kernel (the so called master equation) can be easily defined. These partial differential equations are however very difficult if not impossible to solve. Therefore we give in Section 4.3 an algorithm to simulate the MP. The principles on which the algorithm are based can be deduced directly from the cartoon in Figure 4.2 namely: iteratively compute the time to the next event and then compute which event actually occurred. The results of model simulations are given in Section 4.4 where we show that both variance estimates and comparison of the empirical distribution functions of observed data and simulated data do not support the idea that the observed label retention could be due to stochastic effects.
4.2 Master equations

For discrete-state systems, the master equation, also known as the Chapman-Kolmogorov equation (van Kampen, 2007), defines a set of partial differential equations that describe the rate of change in the probability of being in a certain state. The solution of the master equation gives the full transition kernel for the MP system dynamics and can be indirectly derived from the ODEs that describe the deterministic behaviour of the system.

First we will derive a master equation for each of the one-population and two-population ODE models. We will then give the general form of a master equation using stochastic Petri net notation. The mathematical properties of the master equation in this form makes it useful to explore the relationship between the MP formulation and ODEs from which the master equation was derived.

4.2.1 LRC assay reactions

For each cell kinetic event (self-renewal, differentiation, apoptosis) a reaction in analogy to that of chemical reactions is implied by the ODE-based LRC model of Chapter 3. The master equations can be more easily derived if these reactions are first written down. For
example, the one-population model reactions for unlabelled a-HSCs during BrdU uptake would be

\[
\text{self-renewal: } \{A_{u0}, A_{u1}\} \xrightarrow{\sigma_a A_{u0}} \{A_{u0} - 1, A_{u1} + 2\}
\]

\[
\text{differentiation: } A_{u0} \xrightarrow{\delta_a A_{u0}} A_{u0} - 1 \quad \text{(4.2.1)}
\]

\[
\text{apoptosis: } A_{u0} \xrightarrow{\gamma_a A_{u0}} A_{u0} - 1.
\]

Instead of giving the event reactions for every single species defined by the ODEs, a general reaction scheme that applies to all the species can be derived.

**General reaction scheme**

All BrdU event reactions for a cellular species \(S\) are defined by the general reaction scheme \(\mathcal{R}\),

\[
\mathcal{R}(S) \overset{\text{def}}{=} \begin{cases} 
\{S_i, S_{i+1}\} & \xrightarrow{\sigma_s S_i} \{S_i - 1, S_{i+1} + 2\} \\
\{S_i, \tilde{S}_i\} & \xrightarrow{\delta_s S_i} \{S_i - 1, \tilde{S}_i + 1\} \\
S_i & \xrightarrow{\gamma_s S_i} S_i - 1 \\
S_N & \xrightarrow{\sigma_s S_N} S_N + 1 \\
\{S_N, \tilde{S}_N\} & \xrightarrow{\delta_s S_N} \{S_N - 1, \tilde{S}_N + 1\} \\
S_N & \xrightarrow{\gamma_s S_N} S_N - 1
\end{cases}
\]

\[
\text{for } i \in \{0, 1, \ldots, N - 1\}, \text{ and}
\]

Here \(\tilde{S}\) refers to species on the next level of the differentiation hierarchy, thus direct progeny of \(S\). \(S_i\) indicate species at division level \(i\) and its interpretation is dependant on the context of \(\mathcal{R}\), whether it describes BrdU uptake or dilution. For example, in the context of BrdU uptake reactions as described by \(\mathcal{R}(S_u)\), \(S_u\) denote cells that have divided \(i\) times in the presence of BrdU. Alternatively, for reactions of \(\mathcal{R}(S_c)\) (BrdU chase), \(S_c\) would denote cells that have divided \(i\) times without BrdU present in their microenvironment - after they have been labelled. Hence there is an inverse context dependant interpretation of \(S_i\), with unlabelled cells during BrdU uptake and chase indicated by \(S_{u0}\) and \(S_{cN}\) respectively. Note that differentiation happens independent of cell division and hence label intensity is unaffected as it was defined in the ODE model. We can now proceed to derive the actual master equations for the one-population and two-population models respectively.
4.2. One-population master equation

The general form of the master equation for the one-population model equations is given by

\[
\frac{\partial P(S)}{\partial t} = \sum_{i=0}^{N-1} \left[ \sigma_s(S_i + 1)P(S_i + 1, S_{i+1} - 2) - \sigma_sS_iP(S) 
+ \delta_s(S_i + 1)P(S_i + 1) - \delta_sS_iP(S) 
+ \gamma_s(S_i + 1)P(S_i + 1) - \gamma_sS_iP(S) \right]
\]

(4.2.3)

\[
+ \sigma_s(S_N - 1)P(S_N - 1) - \sigma_sS_NP(S) 
+ \delta_s(S_N + 1)P(S_N + 1) - \delta_sS_NP(S) 
+ \gamma_s(S_N + 1)P(S_N + 1) - \gamma_sS_NP(S).
\]

Where the following equivalences are defined for notational convenience:

\[S \equiv [S_0, S_1, \ldots , S_N]\]

\[P(S) \equiv P(S, t)\]

\[P(S_i + k) \equiv P(S_0, \ldots , S_i + k, \ldots , S_N, t)\]

similarly

\[P(S_i + k, S_{i+1} + l) \equiv P(S_0, \ldots , S_i + k, S_{i+1} + l, \ldots , S_N, t).\]

Note that differentiated progeny \(\tilde{S}\) do not appear in the master equation since we only model a single population of cells.

We can now describe the BrdU uptake reactions of a homogeneous a-HSC population \(S = A_u\) with \(R(A_u)\), keeping track of \(N = n\) labelling intensities. The master equation \(\frac{\partial P(A_u)}{\partial t}\) describes the transition kernel of a continuous time MP where we regard \(X(t) = A_u\) as the state space. For the BrdU chase reactions, \(A_u\) is partitioned into two disjunct groups based on label intensity and BDT. Let \(R(A_{c_1})\) be the reactions for cells that take \(N = n\) divisions to lose BrdU labelling, and \(R(A_{c_2})\) be the reactions for cells that take \(N = n + 1\) divisions to lose BrdU labelling. We have two state spaces \(A_{c_1}\) and \(A_{c_2}\), each modelled by the respective master equation \(\frac{\partial P(A_{c_1})}{\partial t}\) and \(\frac{\partial P(A_{c_2})}{\partial t}\).

4.2.3 Two-population master equation

The general form of the master equation for the two-population model equations is given by

\[
\frac{\partial P(S, \tilde{S})}{\partial t} = \sum_{i=0}^{N-1} \left[ \sigma_s(S_i + 1)P(S_i + 1, S_{i+1} - 2) - \sigma_sS_iP(S, \tilde{S}) \right]
\]
Again we introduced some equivalences to simplify notation:

\[ + \delta_s(S_i + 1)P(S_i + 1, \tilde{S}_i - 1) - \delta_s S_i P(S, \tilde{S}) \]

\[ + \gamma_s(S_i + 1)P(S_i + 1) - \gamma_s S_i P(S, \tilde{S}) \]

\[ + \sigma_s(S_N - 1)P(S_N - 1) - \sigma_s S_N P(S, \tilde{S}) \]

\[ + \delta_s(S_N + 1)P(S_N + 1, \tilde{S}_N - 1) - \delta_s S_N P(S, \tilde{S}) \]

\[ + \gamma_s(S_N + 1)P(S_N + 1) - \gamma_s S_N P(S, \tilde{S}) \]

\[ + \sum_{i=0}^{N-1} \left[ \sigma_s(\tilde{S}_i + 1)P(\tilde{S}_i + 1, \tilde{S}_{i+1} - 2) - \sigma_s \tilde{S}_i P(S, \tilde{S}) \right] \]

\[ + \delta_s(\tilde{S}_i + 1)P(\tilde{S}_i + 1) - \delta_s \tilde{S}_i P(S, \tilde{S}) \]

\[ + \gamma_s(\tilde{S}_i + 1)P(\tilde{S}_i + 1) - \gamma_s \tilde{S}_i P(S, \tilde{S}) \left( 4.2.4 \right) \]

\[ + \sigma_s(\tilde{S}_N - 1)P(\tilde{S}_N - 1) - \sigma_s \tilde{S}_N P(S, \tilde{S}) \]

\[ + \delta_s(\tilde{S}_N + 1)P(\tilde{S}_N + 1) - \delta_s \tilde{S}_N P(S, \tilde{S}) \]

\[ + \gamma_s(\tilde{S}_N + 1)P(\tilde{S}_N + 1) - \gamma_s \tilde{S}_N P(S, \tilde{S}) \]

Again we introduced some equivalences to simplify notation:

\[ S \equiv [S_0, S_1, \ldots, S_N] \]

\[ \tilde{S} \equiv [\tilde{S}_0, \tilde{S}_1, \ldots, \tilde{S}_N] \]

\[ P(S, \tilde{S}) \equiv P(S, \tilde{S}, t) \]

\[ P(S_i + k) \equiv P(S_0, \ldots, S_i + k, \ldots, S_N, \tilde{S}, t) \]

\[ P(S_i + k; S_{i+1} + l) \equiv P(S_0, \ldots, S_i + k, S_{i+1} + l, \ldots, S_N, \tilde{S}, t) \]

similarly

\[ P(\tilde{S}_i + k) \equiv P(S, \tilde{S}_0, \ldots, \tilde{S}_i + k, \ldots, \tilde{S}_N, t) \]

\[ P(\tilde{S}_i + k; \tilde{S}_{i+1} + l) \equiv P(S_0, \ldots, \tilde{S}_i + k, \ldots, S_N, \tilde{S}_0, \ldots, \tilde{S}_i + l, \ldots, \tilde{S}_N, t) \]

Let \( D_u \) be the d-HSC species during BrdU uptake and let \( A_u \) be the a-HSC species during uptake. Also let \( \tilde{D}_u \equiv A_u \). The reactions during BrdU uptake are then given by \( \Re(D_u) \) and \( \Re(A_u) \), and the master equation for the MP by \( \frac{\partial P(D_u, A_u)}{\partial t} \).

As before, we have two state spaces when considering BrdU chase, with the partitioning dependant on BrdU detection sensitivity and level of labelling during BrdU uptake. Let \( D_{c_1} \) and \( A_{c_1} \) represent the different cell species that falls below the BDT after \( N = n \) divisions, and let \( D_{c_2} \) and \( A_{c_2} \) be the ones that loose labelling after \( N = n + 1 \) divisions. Reactions for chase in a heterogeneous cell population case is then given by \( \Re(D_{c_1}) \),
4.2. Master equations

\[ \Re(A_{c1}), \Re(D_{c2}) \text{ and } \Re(A_{c2}) \]. Two master equations \( \frac{\partial P(D_{c1},A_{c1})}{\partial t} \) and \( \frac{\partial P(D_{c2},A_{c2})}{\partial t} \) now govern the Markov processes with state spaces \([D_{c1},A_{c1}]\) and \([D_{c2},A_{c2}]\) respectively.

### 4.2.4 Stochastic Petri net representation

Both the ODE and master equation versions of a model can be written in a compact general form by making use of stochastic Petri net notation. A Petri net is a mathematical language that describes a directed bipartite graph representation of a system to be modelled (Murata, 1989) and has found widespread use in computational and systems biology (Pinney et al., 2003; Hardy & Robillard, 2004). A stochastic Petri net is a Petri net where rate variables are parameters of an exponentially distributed time delay (Goss & Peccoud, 1998). Formally, a stochastic Petri net is defined by the 6-tuple \((P, T, Pre, Post, S_0, r(S))\), with \(P\) and \(T\) the bipartite graph’s disjunct sets of vertices, where \(P\) represents \(m\) places (species) and \(T\) represents \(n\) transitions (events or reactions). \(Pre\) and \(Post\) are integer matrices containing the weights of arcs from places to transitions and transitions to places respectively. \(S_0\) gives the initial marking (state) of the system and \(r(S) = \{r_1(S), r_2(S), \ldots, r_n(S)\}\) is the set that contains the stochastic rate law of each transition. The difference between \(Post\) and \(Pre\) gives the stoichiometry matrix \(N\) whose rows \(N_i\) describe the effect of individual reactions on each species in \(S\).

All the preceding concepts are best explained with a simple example. Consider the reactions for unlabelled a-HSCs during BrdU uptake, written in a slightly different format than Equation 4.2.1:

- **self-renewal:** \(A_{u0} \xrightarrow{\sigma_a} 2A_{u1}\)
- **differentiation:** \(A_{u0} \xrightarrow{\delta_a} \emptyset\)
- **apoptosis:** \(A_{u0} \xrightarrow{\gamma_a} \emptyset\).

Assuming we start with 3750 a-HSCs that are all unlabelled the stochastic Petri net representation of this model would be

\[
P = \begin{pmatrix} A_{u0} \\ A_{u1} \end{pmatrix}, \quad T = \begin{pmatrix} \text{self-renewal} \\ \text{differentiation} \\ \text{apoptosis} \end{pmatrix}, \quad Pre = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad Post = \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},
\]

\[
S_0 = [3750 \ 0], \quad r(S) = \{\sigma_a A_{u0}, \delta_a A_{u0}, \gamma_a A_{u0}\}.
\]

And the stoichiometry matrix then equates to

\[
N = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ -1 & 0 \\ -1 & 0 \end{bmatrix}.
\]
Using the stochastic Petri net notation just introduced, the general form of a master equation can be compactly written as

\[
\frac{\partial P(S,t)}{\partial t} = \sum_{i=1}^{n} \left[ r_i(S - N_i)P(S - N_i, t) - r_i(S)P(S, t) \right].
\] (4.2.5)

Similarly the ODE representation of the model can be written as (compare to the representation of Equation 3.4.1)

\[
\dot{y}(t) = N \cdot r(y(t)).
\] (4.2.6)

### 4.2.5 Relationship to ODE model

The master equation and ODE forms given in Equations 4.2.5 and 4.2.6 respectively are useful to deduce the relationship between a stochastic MP model and the corresponding ODE model. It turns out that when all reactions are zero- or first-order (i.e. linear functions of \( S \)), the average trajectory of the MP model exactly equals the deterministic prediction (Proposition 4.1).

**Proposition 4.1.** When \( r_i(S) \) are linear for all \( i \) (i.e. all reactions are zero- or first-order), the deterministic model \( y(t) \) exactly describes the expected value of the master equation based Markov process model \( X(t) \). That is

\[
y(t) = E(X(t)).
\]

**Proof (Wilkinson, 2006).**

\[
\frac{\partial}{\partial t} E(X(t)) = \frac{\partial}{\partial t} \sum_{S \in \mathcal{M}} S P(S, t)
\]

\[
= \sum_{S \in \mathcal{M}} S \frac{\partial}{\partial t} P(S, t)
\]

\[
= \sum_{S \in \mathcal{M}} S \sum_{i=1}^{n} \left[ r_i(S - N_i)P(S - N_i, t) - r_i(S)P(S, t) \right]
\]

\[
= \sum_{i=1}^{n} \left[ \sum_{S \in \mathcal{M}} S r_i(S - N_i)P(S - N_i, t) - \sum_{S \in \mathcal{M}} S r_i(S)P(S, t) \right]
\]

\[
= \sum_{i=1}^{n} \left[ \sum_{S \in \mathcal{M}} (S + N_i) r_i(S)P(S, t) - \sum_{S \in \mathcal{M}} S r_i(S)P(S, t) \right]
\]

\[
\left( \sum_{S \in \mathcal{M}} S r_i(S - N_i)P(S - N_i, t) = \sum_{S \in \mathcal{M}} (S + N_i) r_i(S)P(S, t) \text{ for } \mathcal{M} \text{ countable} \right)
\]

\[
= \sum_{i=1}^{n} \left[ E((X(t) + N_i) r_i(X(t))) - E(X(t) r_i(X(t))) \right]
\]
4.3 Markov process simulation

\[
\begin{align*}
\sum_{i=1}^{n} E(N_ir_i(X(t))) &= \sum_{i=1}^{n} N_ir_i(X(t)) \\
&= \sum_{i=1}^{n} N_ir_i(E(X(t))) \\
&= \sum_{i=1}^{n} N_ir_i(E(X(t))) \quad \text{(linearity of expectation)}
\end{align*}
\]

Putting \(y(t) = E(X(t))\) we get

\[
\dot{y}(t) = \sum_{i=1}^{n} N_ir_i(y(t)) = N \cdot r(y(t)),
\]

which is just the general form of the ODE model given in Equation 4.2.6.

According to Wilkinson (2006), inference for stochastic models is at least an order of magnitude more complex than inference for deterministic models. The result of Proposition 4.1 is therefore rather encouraging, since it tells us that if we would optimise parameters for our stochastic LRC model to minimise the RSS (calculated using the mean predictions), we should get exactly the same results as when using the deterministic model in Chapter 3. In other words, we can simply use the optimal parameter sets we have already found with a model that was highly suitable for parameter inference.

4.3 Markov process simulation

As is often the case for Chapman-Kolmogorov equations, the master equations defined in Sections 4.2.2 and 4.2.3 are analytically intractable. Fortunately one can resort to various Monte Carlo type algorithms available to numerically simulate such processes. From the cartoon in Figure 4.2 it clearly emerges that in order to simulate the MP as depicted, two probability distributions need to be determined, given the current state. The first one being the event-time distribution to sample the time to the next event. The second one is the distribution over all the possible events to determine what event actually occurred.

4.3.1 Event-time distribution

Consider a system whose elements can be in one of two possible states (say \(A\) or \(B\)) at any moment in time. Suppose a particular event causes elements to jump from state \(A\) to state \(B\) and that the probability of this event occurring in the time interval \([t_1, t_2]\) is \(P_e(t_1, t_2)\):

\[
A \xrightarrow{P_e} B.
\]
4.3. Markov process simulation

If it is given that a particular element is in state \( A \) at time \( t \), we can define the probability of the element still being in state \( A \) at a time \( t + dt \) as

\[
P(A, t + dt) = P(A, t) [1 - P_e(t, t + dt|A, t)],
\]

since we know the jump event has not occurred. Simplifying and rearranging gives

\[
P(A, t + dt) - P(A, t) = -P_e(t, t + dt|A, t)P(A, t).
\]

Dividing this expression by \( dt \) and taking the limit as \( dt \) tends to 0 gives the differential equation for the probability of being in state \( A \) at time \( t \)

\[
\lim_{dt \to 0} \frac{P(A, t + dt) - P(A, t)}{dt} = \frac{dP(A, t)}{dt} = -\lim_{dt \to 0} \frac{P_e(t, t + dt|A, t)}{dt} P(A, t).
\]

Here \( \lim_{dt \to 0} \frac{P_e(t, t + dt|A, t)}{dt} \) gives the instantaneous event probability per time unit, otherwise known as the rate of the transition from \( A \) to \( B \). If this rate is constant, the differential equation becomes

\[
\frac{dP(A, t)}{dt} = -\lambda P(A, t),
\]

which describes exponential decay and hence the solution of \( P(A, t) \) is given by

\[
P(A, t) = P(A, 0)e^{-\lambda t} = e^{-\lambda t},
\]

since we start in \( A \) and thus have \( P(A, 0) = 1 \). The event-time distribution of an event is given by the probability that the element is in state \( B \) at time \( t \), since it implies the reaction event must have occurred before time \( t \). The probability of being in state \( B \) at time \( t \) is thus equivalent to the probability that the event-time was \( t \) time-units or less

\[
P(B, t) = 1 - P(A, t) = 1 - e^{-\lambda t} = F(t),
\]

where \( F(t) \) is the cumulative distribution function of an exponential distribution. Since \( r(S_t) \) is constant if we are in a known state \( S_t \) at time \( t \), this result tells us the time to the next reaction event (i.e. the smallest of a number of possible event times) follows an exponential distribution. The parameter of this exponential distribution can be deduced from Proposition 4.2.

Proposition 4.2. If \( X_i \sim \text{Exp}(\lambda_i), i = 1, 2, \ldots, n \) are independent random variables then

\[
X_m \equiv \min_i \{X_i\} \sim \text{Exp}(\lambda_T), \text{ where } \lambda_T = \sum_{i=1}^{n} \lambda_i.
\]
4.3. Markov process simulation

Proof (Ross, 2006).

\[ P(X_m > x) = P\left( \min_{i} \{X_i\} > x \right) \]
\[ = P([X_1 > x] \cap [X_2 > x] \cap \cdots \cap [X_n > x]) \]
\[ = \prod_{i=1}^{n} P(X_i > x) \quad (X_i \text{ are independent}) \]
\[ = \prod_{i=1}^{n} (1 - P(X_i \leq x)) \]
\[ = \prod_{i=1}^{n} (1 - (1 - e^{-\lambda_i x})) \]
\[ = e^{-x \sum_{i=1}^{n} \lambda_i} \]
\[ = e^{-\lambda_T x}. \]

It follows that \( P(X_m \leq x) = 1 - e^{-\lambda_T x} \) and so \( X_m \sim \text{Exp}(\lambda_T) \). \( \square \)

In other words, at each iteration when the MP has jumped to a new state, the time to the next event can be simulated from \( \text{Exp}(r_T(S)) \) where \( r_T(S) = \sum_{i=1}^{n} r_i(S) \).

4.3.2 Event distribution

Knowing the time to the next reaction, we now need to determine which of the possible reaction events actually occurred, i.e. we need to simulate the index of the minimum of the \( n \) independent exponential distributions in Proposition 4.2. The required distribution is deduced in the following two propositions.

Proposition 4.3. If \( X \sim \text{Exp}(\lambda) \) and \( Y \sim \text{Exp}(\mu) \) are independent random variables, then

\[ P(X < Y) = \frac{\lambda}{\lambda + \mu}. \]

Proof (Ross, 2006).

\[ P(X < Y) = \int_{0}^{\infty} P(X < Y | Y = y) f(y) dy \quad \text{(law of total probability)} \]
\[ = \int_{0}^{\infty} P(X < y) f(y) dy \]
\[ = \int_{0}^{\infty} (1 - e^{-\lambda y} \mu e^{-\mu y} dy \]
\[ = 1 - \int_{0}^{\infty} \mu e^{-(\lambda+\mu)y} dy \]
\[4.4. \text{Results} \]

\[= 1 - \left. \frac{\mu e^{-(\lambda + \mu)y}}{-(\lambda + \mu)} \right|_{0}^{\infty} = 1 - \frac{\mu}{\lambda + \mu} = \frac{\lambda}{\lambda + \mu}. \]

**Proposition 4.4.** Suppose \(X_i \sim \text{Exp}(\lambda_i), i = 1, 2, \ldots, n\) are independent random variables and \(j\) is the index of the minimum of all the \(X_i\). Then \(j\) is a discrete random variable with Probability mass function

\[\pi_i = \frac{\lambda_i}{\lambda_T}, i = 1, 2, \ldots, n, \text{ where } \lambda_T = \sum_{i=1}^{n} \lambda_i.\]

**Proof (Ross, 2006).**

\[\pi_j = \Pr(X_j < \min_{i \neq j} \{X_i\}) = \Pr(X_j < Y),\]

where \(Y = \min_{i \neq j} \{X_i\}\), so that \(Y \sim \text{Exp}(\lambda_{-j})\), where \(\lambda_{-j} = \sum_{i \neq j} \lambda_i\) (by Proposition 4.2)

\[= \frac{\lambda_j}{\lambda_j + \lambda_{-j}} = \frac{\lambda_j}{\lambda_T}. \quad \Box\]

After the time to the next reaction is simulated, we thus simply choose the \(i^{th}\) reaction event with probability \(\frac{r_i[S]}{r_T[S]}\). The complete algorithm is given in Algorithm 3, which turns out to be one of the most well known algorithms for simulating biochemical reactions, namely the exact stochastic simulation algorithm (SSA) introduced by Gillespie (1977).

**4.4 Results**

**4.4.1 Stochastic variation cannot explain long-term label retention**

Although performance issues are experienced with the SSA algorithm for some systems, it was not the case here since we easily performed 1000 simulations in a few minutes. Each simulation was started with an initial HSC population of 3750 (Wilson et al., 2007). The one-population MP model predictions with a \(\pm 3\) standard deviations confidence interval is shown in Figure 4.3A. The inset shows a detailed view of the model predictions during
days 200 to 300 which is the most important time period in terms of label retention. It is clear that each of the three observed values (averages) falls well outside the predicted variation (shaded area) in this period. Comparing the individual values observed at day 306 (Figure 4.3B) to the predicted density, there seems to be no correspondence of any sort. In the following section we will apply the two-sample Kolmogorov-Smirnov test to quantify this visual inspection but instead of just day 306 we will take all the observed time points into account. The predicted distribution does seem to resemble that of a normal distribution in which case the reported $\pm 3$ standard deviations interval is a 99% confidence interval.

The corresponding predictions of the two-population MP model is shown in Figure 4.4A. Strikingly in contrast to the one-population model all mean observed values are
4.4. Results

**Figure 4.3:** Variation estimates using the one-population model. A: Dashed line: average uptake; brown line: average chase; shaded area: ± 3 standard deviations; blue x and error bars: observed data. B: Predicted density at chase day 306.

**Figure 4.4:** Variation estimates using the two-population model. A: Dashed line: average uptake; brown line: average chase; shaded area: ± 3 standard deviations; blue x and error bars: observed data. B: Predicted density at chase day 306.

As can be expected from a system with more parameters, a larger confidence interval is predicted. From the day 306 density plot (Figure 4.4B) we can see that 3 of the 4 observed values seem to be similarly distributed as predicted by the model although there are indications that the actual density has a lower mean here than predicted by the model. In general we observe that although the observed averages are described satisfactorily by the two-population model it seems actual variation is larger than predicted at some time points. It has to be kept in mind that the stochastic model defined here only predicts variation in terms of the actual time to either self-renewal, differentiation or apoptosis (thus variation on a cellular scale).
It is very likely that other sources of variation, be it external (experimental error or physiological differences between mice) or on a smaller scale (molecular, genetic) are also at play (Elowitz et al., 2002; Paulsson, 2005). Possible steps that can be taken to improve the consistency of the current model’s variation predictions with the observed variation include:

1. Instead of a constant initial number of cells, define some distribution to take physiological differences between mice into account.

2. Since there is a noticeable difference in observed variation between different time points, evaluating more mice will be advantageous, especially at the late time points. The true label retention distribution will only be known if the sample size at each time point are significantly increased, however, time and cost constraints are the main limiting factors.

3. Expand the model to include smaller scale sources (thus define a multi-scale model). We foresee problems in terms of the number of unknown parameters that will need to be estimated for such a model.

### 4.4.2 Two-sample Kolmogorov-Smirnov test

Our evaluation so far consisted of calculating averages and standard deviations of observed data and stochastic samples. Information that can potentially be useful is lost as a result, as was evident from Figures 4.3B and 4.4B where individual values were evaluated. The two-sample Kolmogorov-Smirnov (K-S) test is a distribution-free method that can be used to test whether two given populations were sampled from the same one-dimensional probability distribution, by quantifying the distance between the empirical distribution functions of the two samples (Massey, 1951). The two-sample K-S test is a generalised version of the one-sample K-S test and for a given reference and test sample evaluates the null hypothesis $H_0$: the two samples come from the same distribution vs. $H_1$: the two samples come from two different distributions.

If we regard the observed data (each individual mouse) as the reference sample and each of the one-population or two-population simulation results as test sample, we can use the two-sample K-S test to evaluate the following two hypotheses:

- $H_0$: The experimental data and one-population simulated data are from the same distribution.

$H_1$: The experimental data and one-population simulated data are from different distributions.
H$_0$: The experimental data and two-population simulated data are from the same distribution.

H$_1$: The experimental data and two-population simulated data are from different distributions.

Each MP realisation simulates the LRC dynamics of one mouse and hence represents a time series. Consecutive values are thus not independent which might reduce the power of the K-S test. To circumvent this problem we performed an *in silico* version of the real life experiment as follows: for the one-population and two-population samples separately do the following

1. Simulate 1000 realisations of the MP model. Each realisation represents an individual simulated mouse.

2. For each observed mouse, randomly select (without repetition) a simulated mouse and record the value predicted at the time point corresponding to the time point for the observed mouse. Discard all the other data of the simulated mouse.

3. Perform the two-sample K-S test comparing the 137 experimental mouse values to the 137 simulated mouse values. Independence is accomplished since the 137 simulated values are each predictions from independent simulations.

4. Repeat steps 2 and 3 100 times, each time storing the p-value (Schervish, 1996).

5. Report the average of 100 p-values.

The average one-population model p-value was 0.0314 and hence the null hypothesis is rejected at a 5% level of significance. In contrast the null hypothesis is not rejected at this level for the two-population model since the average p-value was 0.133. The K-S test therefore confirms our earlier conclusion that the observed data are more likely that of a heterogeneous HSC population, with a smaller slowly cycling subset that is reversibly activated during hematopoietic injury.

### 4.5 Conclusion

In this chapter we defined LRC dynamics as a Markov process to investigate stochastic variability as a possible cause of the long-term BrdU retention observed in Figure 3.2. For each set of equations in our ODE-based LRC model (Chapter 3) we derived a master equation that define the transition kernel of a continuous time state-discrete MP. Fortunately all reaction rates of the ODE model are first-order, which entails the average
MP prediction is exactly described by the ODE model for the same set of parameters (Proposition 4.1). Since the master equations are analytically intractable, we derived, from first principles, the SSA that can be used to simulate the MP.

Thus, by using the best parameter sets deduced with the ODE models, multiple simulations for each of the one and two-population models enabled us to estimate confidence intervals of predictions (shaded area of Figures 4.3 and 4.4). We found that the confidence interval of the one-population model does not encapsulate the late state observed data (Figure 4.3), which renders it quite unlikely to observe long-term label retention without dormant cells. In contrast at least all the mean experimental data falls inside the two-population confidence interval (Figure 4.4). There are, however, a few time points where the observed variability are larger than predicted. This is most likely either due to fewer mice examined at the particular time points or additional variation not accounted for by our model. However, little variation has been observed at chase day 306 which happens to be encapsulated by the two-population confidence interval. Moreover, the two-sample K-S test suggests that the observed data is more likely to be a sample from the two-population model than the one-population model.

Under our assumption thus that the one-population and two-population models we evaluated here and in Chapter 3 are representative of the two most plausible biological understandings of the hematopoietic system (Section 1.3.3) both the deterministic and stochastic models strongly support a dormant subpopulation in HSCs.

**Deterministic parameter values in stochastic models**

In this chapter we have performed stochastic inference using parameter values estimated by deterministic models. In general this is not possible as demonstrated by the following example. Let \( y(t) \) represent a population of cells that self-renew at rate \( \sigma \), differentiate at rate \( \delta \) and die at rate \( \gamma \). The deterministic ODE describing the change in total cell population over time is

\[
\frac{dy(t)}{dt} = (\sigma - \delta - \gamma)y(t),
\]

with solution

\[
y(t) = y(0)e^{(\sigma - \delta - \gamma)t}.
\]

If we fit this model to experimental data we can thus only learn the net growth (or net death) \( \sigma - \delta - \gamma \) and not the individual rates. Models with different \( \sigma \), \( \delta \), or \( \gamma \) rates but the same net growth (or net death) will give the same predictions. However with the type of stochastic modelling presented here the relative size of the individual rates will influence variability, with larger rates causing larger jumps and hence larger variation. Fortunately parameters for our deterministic and stochastic models are equivalent because of the
homeostatic constraints we imposed on parameters in Section 3.5.3. Suppose $\sigma = \delta + \gamma$, then both deterministic and stochastic predictions depend only on $\sigma$. 
Chapter 5

Parallel parameter exploration on the Grid

5.1 Introduction

So far only the limited dataset in Table 2.1 has been modelled with the agent-based approach defined in Chapter 2. We were therefore interested in exploring the predictive power of the agent-based model on the larger more elaborate dataset described in Tables 3.1 and 3.2, especially to compare the BDT estimates of the ODE (continuous BDT implementation) and agent-based (discrete BDT implementation) models. However, unlike what has been the case for the stochastic Markov process model defined in Chapter 4, no formal relationship between the parameter values estimated by the ODE model (Chapter 3) and the agent-based model exists. Informally one might expect at least some correlation, especially since the underlying mechanism of the agent and Markov process based models as depicted in Figures 2.3 and 4.2 seems related. But alas, the agent-based approach is simply not mathematically formulated well enough for any formal comparison to be possible. We were therefore left with no other option but to re-estimate the parameters for the agent-based model and as pointed out already at the end of Chapter 2 this necessarily entails some form of brute-force search approach. In this chapter we explain how we tackled this problem using a parallel search strategy. However, our planned parallel search required vast amounts of computing power to cover a worthwhile region of the parameter hyperspace in a reasonable amount of time. Fortunately it turned out that the required high performance could be achieved by utilising a Grid supercomputer.

The Enabling Grids for E-sciencE (EGEE - http://www.eu-egee.org/) project originated from the necessity for high performance computers due to the increasingly routine use of high throughput techniques and complex computational models in many scientific disciplines. The EGEE infrastructure is based on clustering technologies and high perfor-
mance distributed platforms and creates a seamless international Grid environment that is available to scientists 24 hours per day. Currently the EGEE Grid consists of more than 41,000 CPUs in addition to about 5 Petabytes disk and tape mass-storage systems, and maintains 100,000 concurrent jobs. Having such resources available clearly changes the way scientific research can take place as this infrastructure allows scientists to process complex and time consuming computations in parallel and on demand.

Grid computing should not be confused with the more recent concept of cloud computing (Vaquero et al., 2009). Cloud computing deployments are self managed, often offered as a metered service and usually depend on existing Grid infrastructures or user networks. Some cloud architectures hardly have a centralised infrastructure or billing system such as peer-to-peer networks and volunteer computing.

In this chapter the main focus is on describing our parallel search strategy (Section 5.2) and evaluating the results in terms of both Grid performance and parameter estimation (Section 5.3). The technical details of deploying our strategy on the Grid is given in Appendix B.

5.2 High dimensional parameter space

An agent-based version of the two-population ODE model described in Section 3.3.4 of Chapter 3 has five parameters in total to estimate\(^1\) namely the d-HSC proportion (DP), BrdU detection threshold (BDT), d-HSC self-renew rate \((\sigma_d)\), d-HSC differentiation rate \((\delta_d)\), and the a-HSC self-renew rate \((\sigma_a)\). This implies that a brute-force search have to traverse a five dimensional parameter space in which each point is presented by the parameter set \((\text{DP}, \text{BDT}, \sigma_d, \delta_d, \sigma_a)\). Moreover, since predictions are stochastic, many realisations for each parameter set have to be performed to calculate an unbiased estimate of the mean trajectory. As before we’ve found that 50 realisations give a good trade-off between computational cost and convergence of mean predictions. If we make an idealised assumption that a single simulation of the entire time-frame (13 days uptake and 400 days chase) takes 1 second to complete, an exploration visiting just \(32 = 2^5\) points in parameter space (thus only 2 values for each parameter) will take \(50 \times 32 \div 60 = 26.7\) minutes to complete. In reality a single simulation takes almost 60 seconds to complete (1.86 GHz Pentium with 1 GB RAM) hence a search that only visits two different values for each parameter would take approximately 26 hours to compute. This shows that for any meaningful parameter exploration to be possible, a solution needs to be found that both limit the parameter space boundaries and enable parallel exploration.

\(^1\)Assuming again that under homoeostasis \(\sigma_d = \delta_d + \gamma_d\), and \(\sigma_a = \delta_a + \gamma_a - \frac{\text{DP}}{1-\text{DP}}\delta_d\)
5.2.1 Exploration boundaries

As a matter of necessity we only considered discrete regions (integer values) of the parameter space. For both the BDT and DP this is not a critical limitation since the BDT for the agent-based model is integer by definition and from experience we’ve found that there are little difference in predictions for DP intervals smaller than 5%. We therefore decided to explore BDT values from 1 to 20 (incrementing by 1) and DP values from 5% to 50% (incrementing by 5% giving 10 possible values). The BDT upper bound can be determined if we recall that the estimate for the BDT of the dataset in Chapter 2 was 20 DNA strands, and then taking into account that BrdU detection by flow cytometry (used to obtain the dataset for which we want to estimate the BDT here) is more sensitive than fluorescent microscopy (used for the dataset in Chapter 2). Regarding the DP, a value of 0% would indicate a single population model (which by now we don’t consider) and we consequently estimated the smallest DP value to be 5%. Since d-HSCs are above a-HSCs in the hematopoietic hierarchy (Figure 3.1 - right panel) it is biologically unrealistic to expect the d-HSC population to be larger than the a-HSC population and hence we can assume the DP \( \leq 50\% \).

Because we expected good performing parameter sets around our ODE model estimates, we clamped each kinetic parameter to these regions. Our model was implemented in Java using the Repast Agent Simulation Toolkit (North et al., 2006) which conveniently allows one to specify parameter boundaries and step sizes in a parameter file for batch mode simulations. An example Repast parameter file is shown in Appendix B.

5.2.2 Parallel strategy

When we developed the ODE model in Chapter 3 we carefully motivated parameter optimisation in three disjunct regions in time (Section 3.5.1). Likewise our approach here is to estimate parameter sets over the same three regions, a different set for: 0-24 hours uptake (optimisation region 1); uptake hour 24 to chase day 10 (optimisation region 2); chase day 10 to chase day 400 (optimisation region 3).

The three region parameter estimation strategy can be exploited by so-called directed acyclic graph (DAG) jobs which are specially defined jobs on the Grid that permit the definition of a workflow (see Section B.1.2 of Appendix B). A key question when designing the DAG job structure is the order of iteration when optimising the five parameters. For our strategy to be systematic four parameters need to be fixed at any time and the fifth iterated through its possible values. The end result might be the same but different iteration orders will result in different DAG structures and possibly more difficulty in programming and keeping track of results. It is important to note that both the DP
and BDT are constant over all three regions of optimisation. It is only the cell kinetic parameters that are adjusted between the regions. In addition the region 1 data are independent of the BDT (recall our discussion in Sections 2.4.2 and 3.2.2) and hence the BDT can only be estimated from both region 2 and region 3. An obvious iteration order therefore is to fix the DP to each of its ten possible values and then optimise the other four parameters for each of these DP values. As a result we defined a separate DAG job for each DP value (ten in total) and in each DAG job the other four parameters were optimised. The overall best parameter set is then given by the DAG job with the best goodness-of-fit results.

We defined DAG jobs that consisted of three nodes (or phases) with the first phase precisely comprising region 1 parameter optimisation (i.e. the first 24 hours of uptake). It turns out that in order to limit unnecessary Grid job and worker node file transfers the other two nodes of the DAG job should not exactly correspond to optimisation regions 2 and 3 respectively. The second phase optimise the BDT and the cell kinetic parameters for both optimisation region 2 and 3. Another speciality Grid job (the parametric job, see Section B.1.2 of Appendix B) can be used to iterate through the twenty possible BDT values. At the beginning of phase 2, twenty jobs (one for each BDT value) are spawned. Each of these jobs optimise the cell kinetic parameters for both regions 2 and 3. The third and final DAG node simply collate the results of the twenty parametric jobs and find the best parameter set.

Optimising parameters for the agent-based model in three different regions is non-trivial since we have to make sure that continuity of agent states is maintained throughout. This problem is further exaggerated by the fact that we have to do many realisations for each parameter set. After each optimisation region every agent of the initial agent population in the new region must have the same state as the last day optimised-parameter-set predictions, for each realisation. To accomplish this feat requires an extra round of simulation after each region 1 and region 2 kinetic parameter optimisation, in total this equates to an additional 210 (10 + 200) simulations of 50 runs each. Agent states are stored on file using Java serialisation which is also useful in a Grid setting where subsequent optimisations are often performed on different worker nodes or computing elements. Our carefully designed parallel exploration strategy just described is depicted in Figure 5.1.

The DAG and parametric jobs conveniently allowed us to separate implementation of DP and BDT iteration from cell kinetic parameters iteration which was handled by the underlying Repast engine of our model. When the cell kinetic parameter ranges and step sizes specified in the Repast parameter file were decided on, consideration was given to the applicable DP value and optimisation region. Actual values were chosen based on experiences and trends observed with the ODE model predictions (for example \( \sigma_d \) is
5.2. High dimensional parameter space

Figure 5.1: Parallel brute-force search strategy. Here we show a flow diagram of one of the 10 main DAG jobs (35% DP). Our complete strategy encompasses 10 such DAG jobs.

larger during region 2 - the activated stage). These values are summarised in Table 5.1 which shows that we visit a total of 1080 points in parameter space for region 1, 9420 parameter sets for region 2, and 11880 for region 3. Each of these parameter sets have to be simulated 50 times in order to assess the average predictions for the particular set. The average predictions (or mean trajectory) are then used to calculate the RSS (Equation 2.4.1) and MSE values as a measure of goodness-of-fit.

5.2.3 Grid deployment

The EGEE Grid distributed supercomputer provides easy access to geographical computing and data management resources on an infrastructure that facilitates the execution of numerous parallel jobs and workflows. Several higher level software packages have already been developed which supplies a user with visual methods of workflow construction and presentation of execution success or progress. Most of these have been build into the portals of several research groups and are available to the scientific community without charge. The National Grid Service (http://www.grid-support.ac.uk) currently provides such an online grid access point in addition to storage space, database integration and user support. Most Grid portals and tools were designed for the most common job, task, and workflow formats. Since we had very specific requirements we did not make use of these tools but instead used command line submission procedures. For this reason we describe in Appendix B the basic instructions we used to interact with and deploy jobs.
## 5.2. High dimensional parameter space

<table>
<thead>
<tr>
<th>DP</th>
<th>OR</th>
<th>$\sigma_d$</th>
<th></th>
<th></th>
<th>$\sigma_a$</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>start</td>
<td>stop</td>
<td>$\Delta$</td>
<td>$n$</td>
<td>start</td>
<td>stop</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>start</td>
<td>stop</td>
<td>$\Delta$</td>
<td>$n$</td>
<td>start</td>
<td>stop</td>
<td>$\Delta$</td>
</tr>
<tr>
<td>1</td>
<td>5%</td>
<td>2400</td>
<td>6000</td>
<td>1200</td>
<td>4</td>
<td>528</td>
<td>720</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>765</td>
<td>200</td>
<td>4</td>
<td>25</td>
<td>85</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2400</td>
<td>6000</td>
<td>1200</td>
<td>4</td>
<td>504</td>
<td>696</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10%</td>
<td>240</td>
<td>1200</td>
<td>480</td>
<td>3</td>
<td>192</td>
<td>264</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>150</td>
<td>600</td>
<td>150</td>
<td>4</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>192</td>
<td>264</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>15%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>192</td>
<td>264</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>135</td>
<td>435</td>
<td>150</td>
<td>4</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>165</td>
<td>765</td>
<td>200</td>
<td>4</td>
<td>25</td>
<td>85</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>20%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>168</td>
<td>240</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>115</td>
<td>435</td>
<td>150</td>
<td>4</td>
<td>168</td>
<td>240</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>290</td>
<td>100</td>
<td>3</td>
<td>16</td>
<td>76</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>25%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>168</td>
<td>240</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>290</td>
<td>100</td>
<td>3</td>
<td>16</td>
<td>76</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>230</td>
<td>30</td>
<td>6</td>
<td>15</td>
<td>60</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>30%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>168</td>
<td>240</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>230</td>
<td>30</td>
<td>6</td>
<td>15</td>
<td>60</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>435</td>
<td>150</td>
<td>3</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>40%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>168</td>
<td>240</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>435</td>
<td>150</td>
<td>3</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>215</td>
<td>30</td>
<td>6</td>
<td>15</td>
<td>60</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>45%</td>
<td>240</td>
<td>720</td>
<td>240</td>
<td>3</td>
<td>168</td>
<td>240</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>215</td>
<td>30</td>
<td>6</td>
<td>15</td>
<td>60</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>180</td>
<td>25</td>
<td>6</td>
<td>15</td>
<td>76</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.1: Parameter ranges and step sizes. DP: Dormant proportion. OR: optimisation region. $\Delta$: step size. $n$ is the number of parameters considered. Total gives the number of simulations required for each OR and is calculated as follows: $n_{\sigma_d} \times 3 \times n_{\sigma_a} \times \text{#realisations} \times \text{#BDTs}$. Optimisation regions 1 and 2 are given in a scale of hours while region 3 is given in a scale of days. The values for $\delta_d$ are not shown, since in all cases it iterated through one of three values: either 1.02, 1.33, or 2 times the value of $\sigma_d$. 


5.3 Results

In the previous section we described a parallel strategy to search for good performing parameter sets for the agent-based LRC model defined in Chapter 2. In what follows we summarise the results achieved with regard to Grid execution times and we relate the relationships of our agent-based parameter results to the ODE estimates under each of the optimisation regions.

5.3.1 Grid performance

Grid execution times varies significantly due to the Grid’s dynamic compilation and also because of congestion and resulting queueing times on selected CEs. The results summarised here have been achieved from our submission portal on a selection of CEs and during a certain submission time window. It will vary from any other or even the same submission variables. The configuration of WNs depends on the several clusters’ administrators and hence also varies significantly. Even on the same CE it is sometimes found that some WNs will successfully execute a job whilst others don’t. JVM or memory related job failure could at times be seen through file error output, others were very hard to

<table>
<thead>
<tr>
<th>DP</th>
<th>Submitted</th>
<th>Running</th>
<th>Done</th>
<th>TST</th>
<th>TRT</th>
</tr>
</thead>
</table>

|               | Average | 2:48:45 | 0:55:00 |

Table 5.2: Processing times - phase 1. DP: Dormant proportion, TST: Total submission time, TRT: Total running time. During phase 1 we evaluated parameter sets for the first 24 hours of BrdU administration. This table summarises the execution results with regard to the processing times on the Grid (hh:mm:ss).
5.3. Results

<table>
<thead>
<tr>
<th>BDT</th>
<th>Submitted</th>
<th>Running</th>
<th>Done</th>
<th>TST</th>
<th>TRT</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0:20:36</td>
<td>15:31:16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Processing times - phase 2. BDT: BrdU detection threshold (in number of DNA strands), TST: Total submission time, TRT: Total running time. These are the submission and computation statistics for the twenty BDT values of the 35% DP DAG job. The BDT values were traversed using a parametric job and varying the _PARAM_ variable.

discover as in most cases no files are returned when a job fails and the user consequently not supplied with any meaningful error messages.

Compared to a normal workstation there is thus much more effort involved in job preparation and submission on the Grid. Nevertheless, any lost time is quickly made up if the massive computational power and parallelisation offered by the Grid are properly exploited by resource hungry applications. Such was the case with finding good parameter sets for the agent-based LRC model. The computing time of 50 runs (400 agents) on a single processor (1.86 GHz Pentium with 1 GB RAM) for each optimisation region is as follows: region 1 - 45 seconds, region 2 - 950 seconds, region 3 - 700 seconds. The total computing time if the parallel parameter search described in this chapter was done on a
5.3. Results

Table 5.4: Processing times - phase 3. DP: Dormant proportion, TST: Total submission time, TRT: Total running time. These are the execution statistics of phase 3 which calculated the best BDT parameter sets for all 10 main DAG jobs. Here it is clearly demonstrated why the Grid is more suitable for resource hungry jobs since the submission time far exceeds the actual computation time.

<table>
<thead>
<tr>
<th>DP</th>
<th>Submitted</th>
<th>Running</th>
<th>Done</th>
<th>TST</th>
<th>TRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Jun 21 00:38:39</td>
<td>Jun 21 00:41:57</td>
<td>Jun 21 00:46:22</td>
<td>0:03:18</td>
<td>0:04:25</td>
</tr>
<tr>
<td>10</td>
<td>Jun 22 12:21:52</td>
<td>Jun 22 12:25:49</td>
<td>Jun 22 12:30:03</td>
<td>0:03:57</td>
<td>0:04:14</td>
</tr>
</tbody>
</table>

Average: 0:18:13 0:04:18

single processor sequentially is thus $1080 \times 45 + 9420 \times 950 + 11880 \times 700 = 17313600$ seconds or 200 days.

The submission and execution times for the Grid are summarised in Tables 5.2, 5.3, and 5.4. Table 5.2 lists the job submission and execution statistics for each of the DP parameter estimations during the first 24 hours of BrdU administration. The average submission time, which is the time that passes since we submitted the specific job to the Grid up until actual execution commence on the designated WN, is 2:48:45 (hh:mm:ss). In contrast the actual running time is much quicker - 55 minutes on average. Phase 2 is computationally the most expensive node of the DAG since it spawns twenty parametric jobs and parameters are evaluated for two optimisation regions. The processing times for twenty values of the BDT of a single DAG during phase 2 are given in Table 5.3 where we can see the average execution time was 15:31:16 which was far greater than the average submission time of 00:20:36. When all the top BDT parameter sets from phase 2 were collated to find the ultimate best performing set of the main DAG job (phase 3), quite the opposite can be observed with regard to processing times. Table 5.4 clearly demonstrate why the Grid is not ideally suited for shorter running jobs. Here the actual computation time was on average 00:04:18 minutes as opposed to a submission period of 00:18:13 minutes.

On average the total processing time on the Grid for one of the ten main DAG jobs is
thus more or less $2:48:45 + 0:55:00 + 0:20:36 + 15:31:16 + 0:18:13 + 0:04:18 = 19:58:08$. Since these DAG jobs were executed concurrently in parallel, this is also the average processing time for all ten DAG jobs, significantly lower than the 200 days that would be required on a single processor.

### 5.3.2 Best parameter sets

The best parameter set for each of the ten DAG jobs together with RSS and MSE goodness-of-fit statistics (calculated using the mean trajectory) are given in Table 5.5. Strikingly, although all the RSS values are larger, the results of the best performing sets are not much different than that of the ODE model (Table 3.4). Similar to our ODE model estimates, the agent-based model suggests that the d-HSCs constitutes around 30% - 45% of the HSC population and self-renew once in approximately every 135-165 days and the a-HSCs once in approximately 25-30 days\(^2\). In addition the results suggest that an HSC will be detected as BrdU\(^+\) if it has at least 5-6 labelled DNA strands, a slightly different estimate than previously. We have to keep in mind, however, that we have no guarantee that these estimates are even local best sets. We can merely state that these are good performing parameter sets and that they are the best from the ones we considered (which hopefully have been a good selection). Since the typical BrdU uptake and dilution profile are quite smooth, the fact that we could only afford to explore discrete cell kinetic parameters should not be a major disadvantage in this instance.

We also show the agent-based model prediction using the best parameter set as estimated by the ODE model in Chapter 3 (the starred 40% DP row). Interestingly the MSE for region 2 and 3 are again consistent, probably the most for all the parameter sets. The RSS of 97.8 is only slightly larger than the best RSS of 94.61 (shown in bold) but it is, however, much higher than 71.7, the corresponding RSS of the ODE model. Although the broad conclusions are the same when parameters are explored, it does seem as if the agent-based model gives a different prediction for the same parameter set compared to the ODE model. This result thus confirms our reluctance for using the agent-based model to estimate stochastic variance for a set of parameter estimated by the ODE model. However, it might be that the higher RSS values of the agent-based model are due to it not being exact (Figure 2.3). If this is indeed the case a shorter tick interval might bring the agent-based predictions closer to that of the ODE.

The model prediction using the best parameter set is shown in Figure 5.2. The variation seems to have a similar spread than that of the Markov process model.

\(^2\)These estimates are based on an RSS < 100.
Figure 5.2: Predictions results of best parameter set. Dashed line: average uptake; brown line: average chase; shaded area: ± 3 standard deviations; blue x and error bars: observed data.
Table 5.5: Goodness-of-fit results for the agent-based model. DP: Dormant proportion, BDT: BrdU detection threshold, RSS: Residual Sum of Squares, MSE: Mean Squared Error, superscripts indicate modelling region 1, 2 or 3. Parameter values $\sigma_d$, $\delta_d$ and $\sigma_a$ are inverted to units of days. The best overall parameter set is highlighted in bold. The *40% DP row shows the prediction of the agent-based model using parameters optimised with the ODE model in Chapter 3.

<table>
<thead>
<tr>
<th>DP</th>
<th>BDT</th>
<th>RSS</th>
<th>MSE$^i$</th>
<th>MSE$^ii$</th>
<th>MSE$^iii$</th>
<th>$\frac{1}{\sigma_d}$</th>
<th>$\frac{1}{\delta_d}$</th>
<th>$\frac{1}{\sigma_a}$</th>
<th>$\frac{1}{\sigma_d}$</th>
<th>$\frac{1}{\delta_d}$</th>
<th>$\frac{1}{\sigma_a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>9</td>
<td>343.9</td>
<td>1.62</td>
<td>7.05</td>
<td>23.71</td>
<td>250</td>
<td>333</td>
<td>24</td>
<td>10</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>10%</td>
<td>8</td>
<td>275.4</td>
<td>1.56</td>
<td>3.97</td>
<td>18.69</td>
<td>150</td>
<td>153</td>
<td>24</td>
<td>10</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>15%</td>
<td>8</td>
<td>199.7</td>
<td>1.52</td>
<td>3.79</td>
<td>12.97</td>
<td>100</td>
<td>133</td>
<td>23</td>
<td>10</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>20%</td>
<td>8</td>
<td>155.9</td>
<td>1.64</td>
<td>3.49</td>
<td>9.8</td>
<td>150</td>
<td>300</td>
<td>23</td>
<td>10</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>25%</td>
<td>5</td>
<td>126.0</td>
<td>1.51</td>
<td>6.65</td>
<td>5.53</td>
<td>150</td>
<td>200</td>
<td>21</td>
<td>15</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>30%</td>
<td>5</td>
<td>95.54</td>
<td>1.53</td>
<td>6.53</td>
<td>3.28</td>
<td>100</td>
<td>102</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>35%</td>
<td>6</td>
<td>94.61</td>
<td>1.49</td>
<td>6.07</td>
<td>3.54</td>
<td>200</td>
<td>400</td>
<td>17</td>
<td>11</td>
<td>14.5</td>
<td>8.5</td>
</tr>
<tr>
<td>40%</td>
<td>6</td>
<td>102.0</td>
<td>1.47</td>
<td>5.15</td>
<td>4.57</td>
<td>200</td>
<td>204</td>
<td>16</td>
<td>12.5</td>
<td>25</td>
<td>8.5</td>
</tr>
<tr>
<td>*40%</td>
<td>6</td>
<td>97.8</td>
<td>1.51</td>
<td>5.09</td>
<td>5.03</td>
<td>208.3</td>
<td>416.7</td>
<td>16.1</td>
<td>9.9</td>
<td>10.1</td>
<td>9.8</td>
</tr>
<tr>
<td>45%</td>
<td>6</td>
<td>96.1</td>
<td>1.52</td>
<td>5.1</td>
<td>4.17</td>
<td>100</td>
<td>200</td>
<td>16</td>
<td>11</td>
<td>14.5</td>
<td>8.25</td>
</tr>
<tr>
<td>50%</td>
<td>6</td>
<td>122.8</td>
<td>1.53</td>
<td>4.88</td>
<td>5.62</td>
<td>150</td>
<td>200</td>
<td>14</td>
<td>11.5</td>
<td>23</td>
<td>8.25</td>
</tr>
</tbody>
</table>
5.4 Conclusion

Some applications can perform and scale very well in a Grid environment while others are instead better suited for a dedicated cluster, especially when bound to certain license agreements or when specialised supporting software is required. Current results also suggest that the Grid is better suited for more computationally expensive jobs as job submission times often exceeds execution times of small jobs. One of our parameter space exploration workflows from phase 1 to phase 3 had an average execution time of about 1 day and since we submitted numerous of these in parallel, the Grid proved to be ideally suited for this type of study.

The results of our parallel parameter search indicate that the agent-based model more or less support the same conclusions as the ODE model. However, actual cell division rate estimates differ as well as model prediction for the same set of parameters. It could be that the observed discrepancies are simply due to the agent-based model not being exact, in which case a smaller tick interval would improve consistency - at the expense of higher computational cost. A logical next step is to do further refinement around the good parameter sets we found. For example, the number of BDT values explored can be drastically reduced by only focussing on the range 4-8. Similarly we can decide to only explore DP values in the range 30% - 45%.

The estimates for the DP and BDT are, however, in strong agreement with the ODE estimates. For the BDT in particular this is an encouraging result since the two models have a very different approach in their implementation of the BDT.
Chapter 6

Modelling dynamic hematopoietic regulation

6.1 Introduction

Recall from Chapter 1 that multipotent stem cells are required to regenerate self-renewing tissues such as the skin, gut, and hematopoietic system. They have the capacity to provide both life-long self-renewal and to re-establish homeostasis after injury. In order to protect against oncogenic mutations, most immature adult stem cells are thought to divide infrequently and be predominantly in a quiescent state (reviewed in Wilson et al. (2007)). In addition, quiescence has been postulated to prevent stem cell exhaustion. This conventional view of quiescence in adult stem cells, however, raises questions about how these cells can be so effective in both maintaining homeostasis and restore it after injury.

In the case of bone marrow HSCs, biological evidence has been found in support of a new perspective on hematopoietic homeostasis - the d-HSC hypothesis (Wilson et al., 2008; Raaijmakers & Scadden, 2008; Foudi et al., 2009). It was showed by Wilson et al. (2008) that while d-HSCs possess most of the multi-lineage long-term self-renewal activity, they are efficiently activated in response to bone marrow injury. After re-establishment of homeostasis, activated HSCs return to dormancy, suggesting that HSCs are not stochastically entering the cell cycle but reversibly switch from dormancy to self-renewal under conditions of hematopoietic stress. In the previous chapters we presented deterministic and stochastic models of LRC data, and showed that those models that assume the existence of a d-HSC population and heterogeneity over time (i.e. activation and de-activation of d-HSCs) are able to describe the observed experimental data the most satisfactorily.

Although all experimental evidence and modelling thereof supports the d-HSC hypothesis (under our assumptions detailed in Section 1.3.3), it is not clear from these LRC
models what the impact of the d-HSC population would be on hematopoietic regulation. Here, instead of describing experimental data, our focus is to mathematically model the dynamic process of hematopoietic regulation. In particular we aim to assess the qualitative nature of d-HSC hematopoiesis (Figure 3.1, right panel) as opposed to the conventional understanding of hematopoiesis (Figure 3.1, left panel), as well as to quantify how effective each model is in hematopoietic regulation. We thus first define a “conventional” model of hematopoiesis (Section 6.2) and then extend this model to include a d-HSC population (Section 6.3). The theoretical background on dynamical analysis methods used to analyse and compare the two models is given in Section 6.4. Our results (Section 6.5) show that the d-HSC model not only exhibits dynamics that can be regarded as biologically more realistic but also has mathematical support for being evolutionarily superior in injury repair compared to the conventional model. Furthermore, oscillations are observed under certain conditions which could provide new insights on oscillatory diseases of the blood system. At the same time the d-HSC model retains previously reported properties of the conventional model that demonstrate how unconstrained cell proliferation can occur through accumulation of mutations in cellular control mechanisms.

6.2 Conventional model of hematopoietic regulation

The LRC equations defined in Chapter 3, although modelling proportion of labelled cells rather than actual cell numbers, implicitly define cell population dynamics:

\[
\frac{dD}{dt} = \sigma_d D - \delta_d D - \gamma_d D \\
\frac{dA}{dt} = \delta_d D + \sigma_a A - \delta_a A - \gamma_a A
\]  

(6.2.1)

However, this implicit HSC model cannot explain dynamic homeostasis since all the kinetic parameters (\(\sigma, \delta, \gamma\)) are independent of the population state. Recently Johnston et al. (2007) were confronted with a similar problem for their ODE model of cell populations in the colonic crypt. The colonic crypt is a popular and well understood adult stem cell system for computational modelling (Potten & Loeffler, 1987; van Leeuwen et al., 2006). The very nature of the structure of the crypt lends itself to a precise mathematical description (Potten & Loeffler, 1990). The mathematical model of Johnston et al. defines two feedback mechanisms to model dynamic homeostasis in the colonic crypt. The first is linear feedback where the differentiation rate of a-HSCs (\(A\)) or transit amplifying cells (\(T\)) has a linear dependence on its respective population sizes. The second is saturating feedback which also assumes the differentiation rate depends on population size but in a nonlinear fashion and up to a maximum rate.
6.2. Conventional model of hematopoietic regulation

Figure 6.1: Schematic showing the conventional model of hematopoietic regulation. a-HSCs (A) self-renew and differentiate into transit amplifying cells (T) which in turn also self-renew and differentiate into mature cells (Z). Feedback is indicated with dashed arrows.

The hematopoietic system, although having a less well-defined spatial structure, is commonly regarded as having a similar proliferation hierarchy with cells transforming from stem to semi-differentiated (sometimes called transit amplifying) to fully differentiated (also called terminally differentiated) compartments (Bryder et al., 2006) (Figure 3.1). The hematopoietic system, however, has a much more complex and larger transit amplifying compartment than what is the case with a single colonic crypt (Dingli et al., 2008). We can use a similar model than Johnston et al. for the hematopoietic system but in order to account for the larger transit amplifying compartment we assume that differentiated transit amplifying cells are expanded by a factor $\zeta$ when they enter the mature compartment. Let $A$ denote the active HSC population (compartment), $T$ the transit amplifying population and $Z$ the terminally differentiated cells. We adopt a simplified view of hematopoiesis in which we focus on a single mature blood cell population $Z$ to represent any of the different hematopoietic cell lineages (white blood cells, red blood cells, platelets, etc.). This simplification is based on the assumption that the regulation of a particular lineage is dependant on the mature numbers in that lineage only (Colijn & Mackey, 2007). Like the LRC model defined in Chapter 3 we denote the constant rates of self-renewal, differentiation, and death by $\sigma$, $\delta$, and $\gamma$ with subscript $a$, $t$, or $z$ differentiating between the rates of $A$, $T$, or $Z$ respectively. Note that we don’t need to define equations for different states in cells like in the case of LRC model since we don’t need to keep track of the number of times cells divide. Figure 6.1 shows a schematic of
this conventional model with the feedback indicated by a dashed arrow. The two sets of model equations with their steady state expressions \((A^*, T^*, Z^*)\) follows below.

### 6.2.1 Linear feedback

For linear feedback the differentiation rate of \(A\) and \(T\) grows linearly with the respective population sizes of \(A\) or \(T\), so that \(A\) differentiate at rate \(\delta_a + k_0A\) and \(T\) at a rate of \(\delta_t + k_1T\) where \(k_i > 0\) are constants. The \(Z\) population loses cells due to normal cellular turnover at a rate of \(\gamma_z\) but \(Z\) cells do not self-renew or differentiate.

\[
\begin{align*}
\frac{dA}{dt} &= (\sigma_a - \delta_a - \gamma_a)A - k_0A^2, \\
\frac{dT}{dt} &= (\sigma_t - \delta_t - \gamma_t)T - k_1T^2 + (\delta_a + k_0A)A, \\
\frac{dZ}{dt} &= -\gamma_zZ + \zeta(\delta_t + k_1T)T.
\end{align*}
\]

(6.2.2)

Apart from the differentiation rate, these equations are closely related to the LRC model’s implied cell population equations (Equations 6.2.1). If we let \(\alpha = \sigma_a - \delta_a - \gamma_a\) and \(\beta = \sigma_t - \delta_t - \gamma_t\), the steady state solutions are given by

\[
\begin{align*}
A^* &= \frac{\alpha}{k_0}, \\
T^* &= \frac{1}{2k_1} \left[ \beta + \sqrt{\beta^2 + 4k_1A^*(\delta_a + k_0A^*)} \right], \\
Z^* &= \frac{\zeta}{\gamma_z}(\delta_t + k_1T^*)T^*.
\end{align*}
\]

\(A\) thus simply defines logistic growth with a carrying capacity of \(\frac{\alpha}{k_0}\).

### 6.2.2 Saturating feedback

In the case of saturating feedback the differentiation rate of \(A\) or \(T\) also depends on population size but in a nonlinear fashion and up to a maximum rate. The differentiation rate of \(A\) is \(\delta_a + \frac{k_0A}{1+m_0A}\) and that of \(T\) is \(\delta_t + \frac{k_1T}{1+m_1T}\).

\[
\begin{align*}
\frac{dA}{dt} &= (\sigma_a - \delta_a - \gamma_a)A - \frac{k_0A^2}{1 + m_0A}, \\
\frac{dT}{dt} &= (\sigma_t - \delta_t - \gamma_t)T - \frac{k_1T^2}{1 + m_1T} + \delta_aA + \frac{k_0A^2}{1 + m_0A}, \\
\frac{dZ}{dt} &= -\gamma_zZ + \zeta \left( \delta_tT + \frac{k_1T^2}{1 + m_1T} \right).
\end{align*}
\]

(6.2.3)

Steady state solutions are given by

\[
A^* = \frac{\alpha}{k_0 - m_0\alpha}
\]
6.3 Dormant HSC model of hematopoietic regulation

The Dormant HSC model of hematopoietic regulation is given by the equation:

\[ T^\ast = \frac{1}{2(k_1 - \beta m_1)} \left[ \beta + m_1 C + \sqrt{(\beta - m_1 C)^2 + 4Ck_1} \right] \]

where \( C = \delta_a A^\ast + \frac{k_0 A^\ast^2}{1 + m_0 A^\ast} = (\sigma_a - \gamma_a) A^\ast \)

\[ Z^\ast = \frac{\zeta}{\gamma_z} \left( \delta_i T^\ast + \frac{k_1 T^\ast^2}{1 + m_1 T^\ast} \right). \]

Ultimately, apart from using a different notation, our two models for hematopoiesis differ from those presented by Johnston et al. for the crypt only in the amplification factor \( \zeta \). In fact, Figure 6.2 shows the parameter stability regions of Equations 6.2.2 and 6.2.3 and it exactly matches the regions reported in (Johnston et al., 2007, Figure 3 (ii and iii)). The amplification factor \( \zeta \) we introduced thus only influences the steady state value of the terminally differentiated compartment.

**Figure 6.2**: Stability regions (from Johnston et al. (2007))

- **A**: The linear feedback model (Equation 6.2.2) either predicts a finite cell population \( (\alpha \geq 0) \) or extinction \( (\alpha < 0) \).
- **B**: The saturating feedback model exhibits similar dynamics but can be driven to a state of unbounded growth if \( \alpha \) or \( \beta \) exceeds certain threshold values.

### 6.3 Dormant HSC model of hematopoietic regulation

Our interest is in what the effects of a switching d-HSC population would be on the above (Section 6.2) models of hematopoietic regulation. We subsequently extended each conventional model to a dormant population version. We assume, based on the d-HSC hypothesis, that there exists a constant dormant stem cell pool that reversibly switch
between a resting and active state upon injury. Denote the constant dormant pool by $D$ and let the number of activated dormant stem cells at time $t$ be given by a Hill function $Da(t)$ that depend on the population level of $Z(t)$:

$$Da(t) = \frac{\theta^s}{\theta^s + Z(t)^s} D,$$  \hspace{1cm} (6.3.1)

and hence $Da(t) = 0.5D$ when $Z(t) = \theta$. The parameter $s$ determines how steep the transition between $Da(t) = 0$ and $Da(t) = D$ is and is chosen such that $Da(t) \to D$ when $Z(t)$ is suitably lower than $Z^*(t)$ and $Da(t) \to 0$ during normal homeostasis (Figure 6.3). We further assume that activated dormant cells divide at a rate $\sigma_d$ with one daughter cell remaining a dormant cell and the other differentiating into the active pool.

![Activation function](image)

**Figure 6.3: Activation function.** The steepness of the activation function is determined by $s$, with smaller values for $s$ indicating more gradual activation (dark blue line) and larger values more abrupt activation (green line).

A new model of homeostasis that takes a switching dormant population into account can now be defined. The schematic of this d-HSC model is shown in Figure 6.4 which shows how the activation function (Equation 6.3.1) couples the stem cell population $A$ with the mature cell population $Z$ through a positive feedback loop, allowing us to simply modify the existing equations for $A$ instead of having to define additional differential equations. We thus only give the equations for $A$ since the ones for $T$ and $Z$ remain as in the conventional model.

### 6.3.1 Linear feedback

$$\frac{dA}{dt} = (\sigma_a - \delta_a - \gamma_a)A - k_0A^2 + \sigma_d Da$$
6.3. Dormant HSC model of hematopoietic regulation

Figure 6.4: Schematic of the d-HSC hematopoietic regulation model. Like in the conventional model, a-HSCs (A) self-renew and differentiate into transit amplifying cells (T) which in turn also self-renew and differentiate into mature cells (Z). The d-HSC model has a constant dormant pool D which, depending on the population level of Z, is reversibly activated to Da. When cells in Da divide, one daughter enters the A pool whilst the other remain in Da (and hence can switch back to the D pool). Feedback is indicated with dashed arrows.

\[
\frac{dA}{dt} = (\sigma_a - \delta_a - \gamma_a)A - \frac{k_0A^2}{1 + m_0A} + \sigma_d\frac{\theta^s}{\theta^s + Z^s}D.
\]

The positive steady state for A in this case is given by the expression

\[
A^* = \frac{1}{2k_0} \left[ \alpha + \sqrt{\alpha^2 + 4k_0\sigma_d\theta^s + Z^s D} \right].
\]

6.3.2 Saturating feedback

\[
\frac{dA}{dt} = (\sigma_a - \delta_a - \gamma_a)A - \frac{k_0A^2}{1 + m_0A} + \sigma_dDa(t)
\]

\[
= (\sigma_a - \delta_a - \gamma_a)A - \frac{k_0A^2}{1 + m_0A} + \sigma_d\frac{\theta^s}{\theta^s + Z(t)^s}D.
\]

The positive steady state for A in this case is given by the expression

\[
A^* = \frac{1}{2(k_0 - \alpha m_0)} \left[ \alpha + m_0C + \sqrt{(\alpha - m_0C)^2 + 4k_0C} \right],
\]

where \( C = \sigma_d\theta^s / (\theta^s + Z^s) D \). In both cases \( A^* \) simplifies to \( N_0^* \) of Johnston et al. (2007) if \( Da(t) = 0 \). The activated dormant cells will thus raise the steady state during injury.
but once the terminally differentiated cells have returned to normal levels the active population will return to the steady state as if there were no dormant population. The dormant population thus does not contribute to normal homeostasis.

6.4 Dynamical analysis

The models we’ve just defined describe a dynamical system (i.e. a system that change and evolve over time) and their equations can be classified as systems of nonlinear ordinary differential equations. A proper understanding of how these models behave, both during equilibrium (homeostasis) and under stress conditions (injury), requires dynamical analysis of their respective equations. Fortunately methods and tools to do the required dynamical analysis are well established. In particular we will use numerical methods to integrate the ODE’s (i.e. plot their solution) and also to analyse steady states and qualitative change in steady states dependant on parameter values (bifurcation analysis). As a measure of how the models respond during conditions of stress, we compute normalised sensitivities of model equations to changes in cell population numbers. The precise sensitivity measure we compute is called the elasticity of an equation (see below) and actually quantifies how each equation adjusts back to steady state after a perturbation.

6.4.1 Bifurcation analysis

Bifurcation analysis tells us how the qualitative nature of equilibrium solutions change in response to parameter adjustments. Numerical bifurcation discovery is by now a well established field, have been used for many years and makes especially visualisation in the phase plane and drawing bifurcation diagrams of nonlinear systems much easier. We will follow suit in our analysis of the d-HSC model by using the software package AUTO (Doedel, 1981), whose algorithms are based on numerical continuation methods (Keller, 1977; Doedel et al., 1991a,b), for our bifurcation analysis. Nevertheless for the sake of completeness and since the d-HSC model exhibits Hopf-bifurcations (as we will show later) we find it appropriate to give a brief background here on the basic theory of classifying critical points\(^1\) for linear and nonlinear systems. An excellent source for more information about this topic is Strogatz (2001).

\(^1\)also called fixed or stable points.
Linear systems

We have already showed in Chapter 3 that a system of linear ODEs can be written in matrix form as (Equation 3.4.1)

\[ \dot{y} = K \cdot y, \]  

(6.4.1)

with the solution then given by (Equation 3.4.2)

\[ y(t) = \exp(K \cdot t) \cdot y(0). \]  

(6.4.2)

Since for a given diagonal matrix \( D \) with \( D_{ii} = \lambda_i \) it is the case that

\[ \exp(D) = \sum_{i=0}^{\infty} \frac{D^i}{i!} = \sum_{i=0}^{\infty} \begin{bmatrix} \lambda_1^i & 0 & \ldots & 0 \\ 0 & \lambda_2^i & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n^i \end{bmatrix} = \begin{bmatrix} e^{\lambda_1} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e^{\lambda_n} \end{bmatrix}, \]  

(6.4.3)

and

\[ (V \cdot D \cdot V^{-1})^i \]  

\[ = \sum_{i=0}^{\infty} \frac{D^i}{i!} \cdot V^{-1} = \sum_{i=0}^{\infty} \begin{bmatrix} \sum_{i=0}^{\infty} \frac{\lambda_1^i}{i!} & 0 & \ldots & 0 \\ 0 & \sum_{i=0}^{\infty} \frac{\lambda_2^i}{i!} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sum_{i=0}^{\infty} \frac{\lambda_n^i}{i!} \end{bmatrix} = \begin{bmatrix} e^{\lambda_1} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e^{\lambda_n} \end{bmatrix}, \]  

(6.4.4)

\( \exp(K \cdot t) \) can be easier solved if \( K \) is diagonalised as

\[ K = V \cdot D \cdot V^{-1}, \]  

(6.4.5)

where \( V^{-1} \) is the inverse of \( V \). In this form the solution of Equation 6.4.1 can be written as

\[ y(t) = \exp(K \cdot t) \cdot y(0) = \sum_{i=0}^{\infty} \frac{(K \cdot t)^i}{i!} \cdot y(0) \]

\[ = \sum_{i=0}^{\infty} V \cdot \frac{(D \cdot t)^i}{i!} \cdot V^{-1} \cdot y(0) = V \cdot \left( \sum_{i=0}^{\infty} \frac{(D \cdot t)^i}{i!} \right) \cdot V^{-1} \cdot y(0) \]

\[ = V \cdot \exp(D \cdot t) \cdot V^{-1} \cdot y(0) \]

\[ = V \cdot \begin{bmatrix} e^{\lambda_1 \cdot t} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2 \cdot t} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e^{\lambda_n \cdot t} \end{bmatrix} \cdot V^{-1} \cdot y(0). \]  

(6.4.6)
By observing that \( K = V \cdot D \cdot V^{-1} \Leftrightarrow K \cdot V = V \cdot D \Leftrightarrow K \cdot V^{(i)} = V^{(i)} \cdot D = \lambda_i \cdot V^{(i)} \), where \( V^{(i)} \) denotes the \( i \)th column of \( V \), we note that \( \lambda_i \) is an eigenvalue of \( K \) with \( V^{(i)} \) its corresponding eigenvector. It thus follows that the elements of \( D \) are the eigenvalues of \( K \) with the columns of \( V \) the corresponding eigenvectors.

The above result is very useful to show how the qualitative nature of solutions and critical points is determined by the eigenvalues of a linear system. This is best demonstrated if we consider a general 2-dimensional system

\[
\begin{bmatrix}
y_1(t) \\
y_2(t)
\end{bmatrix} =
\begin{bmatrix}
y_1(t) \\
y_2(t)
\end{bmatrix} = \mathbf{V} \cdot \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} \cdot \mathbf{V}^{-1} \cdot \begin{bmatrix} y_1(0) \\
y_2(0)
\end{bmatrix}
\]

(6.4.7)

where \( c_{ij} \) are functions of the elements of \( \mathbf{V} \). From 6.4.7 it is clear that analysing eigenvalues can provide very useful information when classifying critical values. For example, since the characteristic equation

\[
\det (K - \lambda I) = 0
\]

can be written as \( \lambda^2 - \tau \lambda + \Delta = 0 \), where \( \tau = \text{trace}(K) \) and \( \Delta = \det(K) \), and hence the eigenvalues determined by

\[
\lambda_{1,2} = \frac{1}{2} \left( \tau \pm \sqrt{\tau^2 - 4\Delta} \right),
\]

the nature and stability of a fixed point in a linear system can be determined just by studying the eigenvalues (and consequently the determinant and trace) of the coefficient matrix (Strogatz, 2001). For example if both eigenvalues are negative, both components of \( y(t) \) will decrease exponentially so that the critical point is classified as a stable node. See Strogatz, Figure 5.2.8 for a full critical point classification diagram.

**Complex eigenvalues.** Of special interest in this work is the case when eigenvalues are complex since such cases relate to oscillatory solutions. For example suppose

\[
\lambda = a + b \cdot i,
\]
with $Re(\lambda) = a$ the real part and $Im(\lambda) = b$ the imaginary unit. The exponential in Equation 6.4.7 now would take the form

$$e^{(a+b\cdot i)t} = e^{a \cdot t} \cdot e^{b \cdot t} = e^{a \cdot t} \cdot (\cos(b \cdot t) + i \cdot \sin(b \cdot t))$$ (by Euler’s formula).

which defines an oscillatory solution. Again there are a number of possible cases, depending on the value of $a$. If $a = 0$ the roots are pure imaginary and the solution is a standing wave with the fixed point classified as a centre. Otherwise the fixed point is a stable spiral when $a < 0$ (damped oscillations) and unstable spiral when $a > 0$. A Hopf bifurcation occurs when a stable spiral changes into an unstable spiral surrounded by a stable limit cycle. In other words, a Hopf bifurcation can be detected when a pair of complex conjugate eigenvalues have negative $a$, and $a$ crosses the imaginary axis of the complex plane (Figure 6.5).

![Figure 6.5: Complex plane. Example of how a conjugated pair of eigenvalues cross the imaginary axis $Im(\lambda)$ during a Hopf bifurcation.](image)

**Nonlinear systems**

Nonlinear systems, like the models introduced in Sections 6.2 and 6.3, are typically very hard or most often impossible to solve analytically. It is therefore rarely possible to determine the stability of a nonlinear systems’ critical points through a closed form solution as is the case for a linear system. Fortunately it turns out that we can use a linear approximation if we restrict ourselves to a small neighbourhood around the critical points of a nonlinear system, and then classify the corresponding critical points using the eigenvalue analyses techniques discussed in Section 6.4.1.
To demonstrate the idea, let us consider a two component nonlinear system

\[ \frac{dy_1}{dt} = f(y_1, y_2) \]
\[ \frac{dy_2}{dt} = g(y_1, y_2). \]

Suppose that \( f(y_1^*, y_2^*) = 0 \) and \( g(y_1^*, y_2^*) = 0 \) and hence \((y_1^*, y_2^*)\) is a critical point, i.e. \( \frac{dy_1^*}{dt} = 0 \) and \( \frac{dy_2^*}{dt} = 0 \). Let \( u_1 = y_1 - y_1^* \Leftrightarrow y_1 = y_1^* + u_1 \) and \( u_2 = y_2 - y_2^* \Leftrightarrow y_2 = y_2^* + u_2 \) indicate a small perturbation from the critical point. The rate of change at this perturbation in the \( u_1 \) direction then equals

\[
\frac{du_1}{dt} = \frac{dy_1}{dt} - \frac{dy_1^*}{dt} = \frac{dy_1}{dt} = f(y_1^* + u_1, y_2^* + u_2) \tag{by substitution} \]

expanding \( f \) as a Taylor series gives

\[
= f(y_1^*, y_2^*) + u_1 \frac{\partial f}{\partial y_1}(y_1^*, y_2^*) + u_2 \frac{\partial f}{\partial y_2}(y_1^*, y_2^*) + O(u_1^2, u_2^2, u_1u_2) \]

\[
\approx u_1 \frac{\partial f}{\partial y_1}(y_1^*, y_2^*) + u_2 \frac{\partial f}{\partial y_2}(y_1^*, y_2^*) \quad \text{(since \( u_1 \) and \( u_2 \) are very small)}. \]

Similarly in the \( u_2 \) direction we have

\[
\frac{du_2}{dt} = \frac{dy_2}{dt} - \frac{dy_2^*}{dt} = \frac{dy_2}{dt} \approx u_1 \frac{\partial g}{\partial y_1}(y_1^*, y_2^*) + u_2 \frac{\partial g}{\partial y_2}(y_1^*, y_2^*) \]

The linearisation of the entire system can thus be written as

\[
\begin{bmatrix}
\frac{du_1}{dt} \\
\frac{du_2}{dt}
\end{bmatrix} =
\begin{bmatrix}
\frac{\partial f}{\partial y_1} & \frac{\partial f}{\partial y_2} \\
\frac{\partial g}{\partial y_1} & \frac{\partial g}{\partial y_2}
\end{bmatrix}_{(y_1^*, y_2^*)}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix}, \quad \text{or}
\]

\[ u' = A \cdot u, \quad (6.4.8) \]

which can be analysed by the linear methods shown above in Section 6.4.1. The matrix \( A \) in 6.4.8 is the Jacobean matrix evaluated at \((y_1^*, y_2^*)\). In other words, in order to classify and evaluate the qualitative nature of a nonlinear system’s critical points, we study the eigenvalues of the Jacobean matrix in much the same way that the eigenvalues of the coefficient matrix can be studied in the case of a linear system. There are, however, borderline cases when linearisation gives inaccurate results (Strogatz, 2001). As we will show below, the d-HSC model’s trajectories are attracted to a stable spiral (Figure 6.6) and fortunately in this case the linearised analyses is exact.
6.4.2 Elasticities: a measure of hematopoietic regulation effectiveness

An interesting question that arises is: given two populations of a species that differ only in their hematopoietic architecture - one conventional and one d-HSC. Which one of these would have the evolutionary superior hematopoietic system? For example if two alpha males, one from each species and of equal strength, fight and injure each other so that both loose blood. Which one would heal the quickest and hence be dominant? Arguably the dominant one would be the hematopoietic architecture we observe in animals today. Hence answering this question can provide evolutionary support for or against the d-HSC hypothesis. It would thus be useful to mathematically quantify the evolutionary advantage or disadvantage of the d-HSC model vs. the conventional model in terms of injury repair efficiency. It turns out that elasticities, a well established measure mostly used in economics to quantify response of consumers on price, supply and demand interrelations (Case & Fair, 2008), can be very useful for this purpose.

All of the changes in cell populations defined by the equations in Sections 6.2 and 6.3 can be written in the general form

\[
\frac{dC_i}{dt} = R_i^+ - R_i^- ,
\]

(6.4.9)

where \( C_i \) is cell population \( i \), \( R_i^+ \) is the total birth flux (self-renew and precursor differentiation) and \( R_i^- \) is the total death flux (self differentiation and death) of cell population \( i \).

In order to quantify the response of a dynamical homeostasis model to perturbations we compute the elasticity matrix \( H \) of the model, with the elements of \( H \) defined by

\[
H_{ij} = \frac{C_j}{R_i^- / R_i^+} \frac{\partial (R_i^- / R_i^+)}{\partial C_j} = \frac{\partial \ln (R_i^- / R_i^+)}{\partial \ln C_j },
\]

(6.4.10)

so that \( H_{ij} \) is a normalised measure of how the ratio between proliferation and death of \( C_i \) respond to changes in the number of \( C_j \). Systems with higher elasticity return more rapidly to homeostasis. Here the primary focus was on calculating \( H_{Aj} \) (Table 6.2 and 6.3) since this is the only row where the d-HSC model differs form the conventional model. These are derived below.

6.4.3 Derivation of elasticities

Linear feedback model

\[
\frac{R_i^- / R_i^+}{R_A^- / R_A^+} = \frac{\delta_a + \gamma_a + k_0 A}{\sigma_a},
\]
so that
\[ \frac{\partial R^-_A}{\partial A} = \frac{k_0}{\sigma_a}, \quad \frac{\partial R^+_A}{\partial T} = 0, \quad \text{and} \quad \frac{\partial R^-_A}{\partial Z} = 0. \]
Hence
\[ H_{AA} = \frac{\sigma_a A}{(\delta_a + \gamma_a) + k_0 A}, \quad H_{AT} = 0, \quad \text{and} \quad H_{AZ} = 0. \]

Dormant model with linear feedback
\[ \frac{R^-_A}{R^+_A} = \frac{\phi_a A + k_0 A^2}{\sigma_a A + \sigma_d Da}, \quad \text{where} \quad \phi_a = \delta_a + \gamma_a, \]
so that
\[ \frac{\partial R^-_A}{\partial A} = \frac{(\phi_a + 2k_0 A)(\sigma_a A + \sigma_d Da) - \sigma_a(\phi_a A + k_0 A^2)}{(\sigma_a A + \sigma_d Da)^2}, \]
\[ \frac{\partial R^-_A}{\partial T} = 0, \quad \text{and} \quad \frac{\partial R^-_A}{\partial Z} = \frac{(\phi_a A + k_0 A^2)(\sigma_d Da)}{(\sigma_a A + \sigma_d Da)^2} \cdot \frac{sZ^{s-1}}{\theta^s + Z^s}. \]

After some algebra it follows
\[ H_{AA} = \frac{\phi_a + 2k_0 A}{\phi_a + k_0 A} - \frac{\sigma_a A}{\sigma_a A + \sigma_d Da} = \frac{k_0 A}{(\delta_a + \gamma_a) + k_0 A} + \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da}, \]
\[ H_{AT} = 0, \quad \text{and} \quad H_{AZ} = \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da} \cdot \frac{sZ^{s-1}}{\theta^s + Z^s} = \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da} \cdot \frac{s(D - Da)}{D}. \]

Saturating feedback model
\[ \frac{R^-_A}{R^+_A} = \frac{\phi_a(1 + m_0 A) + k_0 A}{\sigma_a(1 + m_0 A)} = \frac{\phi_a}{\sigma_a} + \frac{k_0 A}{\sigma_a(1 + m_0 A)}, \]
so that
\[ \frac{\partial R^-_A}{\partial A} = \frac{\sigma_a(1 + m_0 A)k_0 - k_0 A\sigma_a m_0}{\sigma_a^2(1 + m_0 A)^2} = \frac{k_0}{\sigma_a(1 + m_0 A)^2}, \]
\[ \frac{\partial R^-_A}{\partial T} = 0, \quad \text{and} \quad \frac{\partial R^-_A}{\partial Z} = 0. \]
Hence
\[ H_{AA} = \frac{\sigma_a(1 + m_0 A)A}{\phi_a(1 + m_0 A) + k_0 A} \cdot \frac{k_0}{\sigma_a(1 + m_0 A)^2} \cdot \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A}, \]
\[ H_{AT} = 0, \quad \text{and} \quad H_{AZ} = 0. \]
6.5. Results

Dormant model with saturating feedback

\[
\frac{\mathcal{R}_A^-}{\mathcal{R}_A^+} = \frac{\phi_a(1 + m_0 A)A + k_0 A^2}{(1 + m_0 A)(\sigma_a A + \sigma_d Da)} = \frac{\phi_a A + \phi_a m_0 A^2 + k_0 A^2}{\sigma_a A + \sigma_a m_0 A^2 + \sigma_d Da + \sigma_d m_0 A Da},
\]

so that

\[
\frac{\partial \mathcal{R}_A^-}{\partial A} = \frac{1}{(1 + m_0 A)^2(\sigma_a A + \sigma_d Da)^2} \times \left( \phi_a(1 + m_0 A)A + k_0 A^2)(\sigma_a A + \sigma_d Da) - (\phi_a(1 + m_0 A)A + k_0 A^2)(\sigma_a + 2\sigma_a m_0 A + m_0 \sigma_a Da) \right)
\]

\[
= \frac{\sigma_d Da (\phi_a(1 + m_0 A) + k_0 A)(1 + m_0 A) + k_0 A(\sigma_a A + \sigma_d Da)}{(1 + m_0 A)^2(\sigma_a A + \sigma_d Da)^2},
\]

after simplification. Partial derivatives with respect to \( T \) and \( Z \) are

\[
\frac{\partial \mathcal{R}_A^-}{\partial T} = 0, \quad \text{and}
\]

\[
\frac{\partial \mathcal{R}_A^-}{\partial Z} = \frac{(\phi_a(1 + m_0 A)A + k_0 A^2)(\sigma_d Da)}{(1 + m_0 A)(\sigma_a A + \sigma_d Da)^2} \cdot \frac{s Z^{s-1}}{\theta^s + Z^s}.
\]

After some algebra it follows

\[
H_{AA} = \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} + \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da},
\]

\[
H_{AT} = 0, \quad \text{and}
\]

\[
H_{AZ} = \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da} \cdot \frac{s Z^s}{\theta^s + Z^s} = \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da} \cdot \frac{s(D - Da)}{D}.
\]

6.5 Results

As a first measure of what impact introduction of the d-HSC population had on the conventional model of hematopoietic regulation we studied simulations and performed bifurcation analysis of the models (focusing only on the saturating feedback models since results for the linear feedback models are similar).

6.5.1 Simulations

Figure 6.6 shows a comparison between simulations of the conventional model (top panel) and the d-HSC model (bottom panel) for similar parameter values (given in Table 6.1) and arbitrary starting values. For each model, simulation output for \( A, T, Z; A \) and \( T \) in detail; and a \( T - Z \) phase plane are shown from left to right. It is immediately apparent that introduction of the d-HSC population has enabled oscillations to occur.
Initially the oscillations are damped and fade away as the system adjusts to homeostasis. The equilibrium point they are attracted to is known as a stable spiral in the dynamical analysis literature (Strogatz, 2001) and can be observed in the $T-Z$ phase plane of Figure 6.6.

Figure 6.7 shows the result when initial conditions are set to the steady state values of $A$, $T$, and $Z$ respectively, and each cell population perturbed at some point. Clearly visible is that both models are immediately attracted to their respective equilibrium states. The conventional model is, however, quite exact in the way it re-establishes homeostasis, very much in a manner like a human engineered system would. In contrast the d-HSC model slowly adjusts itself through damped oscillations until equilibrium is reached. This way of re-establishing homeostasis can be regarded as biologically more realistic. Also apparent is how $A$ of the d-HSC model is adjusted upwards (through activated d-HSCs) when either $T$ or $Z$ is perturbed.
6.5. Results

6.5.2 Bifurcation diagrams

Bifurcation diagrams visually depict the qualitative behaviour of stable solutions as a function of bifurcation parameter values. One of the finest examples on the use of bifurcation diagrams in systems biology can be found in Tyson et al. (2002) where dynamical principles of cell cycle control are uncovered using one-parameter bifurcation diagrams. Here we exclusively focus on two-parameter bifurcation diagrams which depicts changes in equilibrium solutions as a function of two bifurcation parameters simultaneously.

Oscillations

When the dormant population $D$ is increased, the amplitude of the damped oscillations also increase and they take longer to fade out (Figure 6.8, top panel) until $D$ reaches a bifurcation point where the system oscillates indefinitely (Figure 6.8, bottom panel). Since the equilibrium point changed from a stable spiral to an unstable spiral surrounded by a limit cycle (Figure 6.9), the bifurcation point is a Hopf-bifurcation (Strogatz, 2001).

Figure 6.7: Behaviour during injury. Initial conditions are set to steady state values and the re-adjustment to steady state by each model demonstrated for a perturbation in each of $A$, $T$, or $Z$. Parameter values are as in Figure 6.6.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure6_7}
\caption{Behaviour during injury.}
\end{figure}
Figure 6.8: Dynamic hematopoiesis model simulations showing self sustained oscillations. Top panel: oscillations take longer to damp out when $D$ is increased from 2 to 3. Bottom panel: when $D$ equals 4 a Hopf bifurcation has been reached and the system oscillates at a limit cycle. Panels from left to right and parameter values other than $D$ are as in Figure 6.6.

Figure 6.9: $T$ vs. $Z$ phase plane during oscillations. The initial stable spiral is now unstable and solutions are attracted to a newly formed stable limit cycle surrounding the unstable spiral. This behaviour is characteristic of a Hopf bifurcation.
6.5. Results

Figure 6.10: Two parameter bifurcation diagrams. Blue areas: oscillatory regions. Steady states where evaluated for parameter values as given in Table 6.1. The bifurcation diagrams suggest oscillations occur when the size of the d-HSC pool transcends a critical value, but only in the parameter regions where the d-HSCs are most sensitive to activation and deactivation.

Hopf bifurcations

Plotting two parameter bifurcation diagrams revealed interesting patterns (Figure 6.10). Although $Da(t)$ only indirectly affects $T$, the oscillating regions for $A$ and $T$ parameters in Figure 6.10 (top two panels) have similar shapes. Whilst it initially seemed as if the system can be driven to an oscillatory state by simply increasing $D$, it can be clearly seen in Figure 6.10 that parameter values exist for which $D$ can increase indefinitely without inducing oscillations. It turns out that the boundary of these D-independent-oscillation-free (DIOF) regions are determined by the net growth rate $\alpha = \sigma_a - \delta_a - \gamma_a$ and $\beta = \sigma_t - \delta_t - \gamma_t$ of $A$ and $T$ respectively. As $\alpha$ or $\beta$ increases, the steady state of $A^*$, $T^*$ or $Z^*$ will also increase. At some point $Z^*$ will reach a critical value larger than $\theta$ such that $Da(t)$ tend to zero irrespective of the value of $D$. As was the case in the models of
Johnston et al. (2007), unbounded growth will occur once $\alpha$ exceeds $\frac{k_0}{m_0}$ or $\beta$ exceeds $\frac{k_1}{m_1}$. Remarkably we don’t have extinction for negative $\alpha$ or $\beta$ as in the conventional models, rather larger d-HSC populations without inducing oscillations are now possible. The effect of a larger $Z^*$ is directly observable in the bifurcation diagram of $D$ vs. $\theta$ (bottom panel, right plot of Figure 6.10), where decreasing $\theta$ (imitating malfunction in the d-HSC activation pathway) results in a DIOF region. Conversely an oversensitive activation pathway (increasing $\theta$) seems to have very little effect on the Hopf bifurcation point. As can be expected there is a DIOF region for lower values of the steepness coefficient $s$ where d-HSCs will be less sensitive to activation or deactivation. Interestingly the $Z^*$ population parameter $\zeta$ also have a DIOF region for smaller $Z^*$ values. In such cases all of the d-HSCs would be activated most of the time, so that the model behaves in a dormant-free manner but with an increased $A$ population size. Taken together the bifurcation analyses suggest oscillations occur when the size of the d-HSC pool transcends a critical value but only in the parameter regions where the d-HSCs are most sensitive to activation and deactivation. These regions are mostly determined by the ratio of $D$ activation to the $A$ or $T$ net growth rate.

### 6.5.3 Evaluating model elasticities

We have argued in Section 6.5.1 that the d-HSC model exhibit biological more realistic behaviour compared to the conventional model. We have also shown in Section 6.4.2 how elasticities can mathematically formalise the effectiveness of this behaviour when homeostasis is disturbed. We subsequently computed elasticities $H_{AA}$, $H_{AT}$, and $H_{AZ}$ to measure how $A$ respond to changes in $A$, $T$, and $Z$ respectively. In Table 6.2 and 6.3 we summarise the results of the linear and saturating feedback models respectively, comparing the elasticity expressions of the conventional model with those of the d-HSC model.

For both linear and saturating feedback the d-HSC model $H_{AA}$ is larger by a factor of $\frac{\sigma_D D^+}{\sigma_A A^+ + \sigma_D D^+}$. This expression can be interpreted as the ratio of the d-HSC contribution to the total birth flux of $A$ and tends to 0 during homeostasis. It will be higher during injury of $Z$ so that the d-HSC model has a higher elasticity and hence will be more effective at re-establishing homeostasis. The expression for the d-HSC model $H_{AZ}$ is similar but adjusted by a fraction $\frac{D - Da}{D}$ of $s$. The interpretation of this elasticity is counter-intuitive in that it will tend to 0 during injury (i.e. $Da \to D$) but also during homeostasis (i.e. $Da \to 0$). To get a better understanding of $H_{AZ}$ we calculate its maximum with respect to $Z$. 
Table 6.1: Parameter values used in hematopoiesis simulations. These values were used for the plots in Figures 6.6 - 6.10 and although they are synthetic they do take some of the current understanding of the hematopoietic system into account (Bryder et al., 2006). For example the self-renew rate for the transit amplifying cells is larger than that of the stem cells. Similarly the death rate of the mature cells are significantly higher than that of the A or T population.

Maximum of $H_{AZ}$

The calculation can be simplified by first finding the derivative of $Da$. Recall (Equation 6.3.1) that

$$Da(t) = \frac{\theta^s}{\theta^s + Z(t)^s} D .$$

Differentiating with respect to $Z$ gives

$$\frac{d}{dZ} Da = \frac{-\theta^s D \cdot sZ^{s-1}}{(\theta^s + Z^s)^2} = - \frac{\theta^s}{\theta^s + Z^s} D \cdot sZ^{s-1} \frac{sZ^{s-1}}{\theta^s + Z^s} = - Da \frac{sZ^{s-1}}{\theta^s + Z^s} .$$

Now let $f \equiv \frac{\sigma_d Da}{\sigma_a A + \sigma_d Da}$ and $g \equiv \frac{s(D - Da)}{D} = \frac{sZ^s}{\theta^s + Z^s}$, so that $H_{AZ} = f \cdot g$, and hence

$$\frac{\partial}{\partial Z} H_{AZ} = \frac{\partial}{\partial Z}(f \cdot g) = \frac{\partial}{\partial Z} f \cdot g + f \cdot \frac{\partial}{\partial Z} g \quad \text{ (product rule)}. $$
### 6.5. Results

<table>
<thead>
<tr>
<th>$H_{AC_j}$</th>
<th>Conventional model</th>
<th>d-HSC model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{AA}$</td>
<td>$\frac{k_0A}{(\delta_a + \gamma_a) + k_0A}$</td>
<td>$\frac{k_0A}{(\delta_a + \gamma_a) + k_0A} + \frac{\sigma_dDa}{\sigma_aA + \sigma_dDa}$</td>
</tr>
<tr>
<td>$H_{AT}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$H_{AZ}$</td>
<td>0</td>
<td>$\frac{\sigma_dDa}{\sigma_aA + \sigma_dDa} \cdot \frac{s(D - Da)}{D}$</td>
</tr>
</tbody>
</table>

**Table 6.2: Elasticities for the linear feedback models.** $H_{AC_j}$ is a normalised measure of how $A$ respond to changes in $C_j$ with higher values indicating a more rapid adjustment to steady state. For both $H_{AA}$ and $H_{AZ}$ the d-HSC model has a higher value than the conventional model. $H_{AT} = 0$ for both models since neither has any $T$ dependant parameters.

We first compute the derivative of $f$ and $g$:

$$\frac{\partial}{\partial Z} f = \frac{1}{(\sigma_aA + \sigma_dDa)^2} \left[ -\sigma_dDa \frac{sZ^{s-1}}{\theta^s + Z^s} (\sigma_aA + \sigma_dDa) - \sigma_dDa \left( -\sigma_dDa \frac{sZ^{s-1}}{\theta^s + Z^s} \right) \right]$$

$$= \frac{1}{(\sigma_aA + \sigma_dDa)^2} \left[ -\sigma_dDa \frac{sZ^{s-1}}{\theta^s + Z^s} (\sigma_aA + \sigma_dDa - \sigma_dDa) \right]$$

$$= -\sigma_dDa \frac{\sigma_aA}{(\sigma_aA + \sigma_dDa)^2} \cdot \frac{sZ^{s-1}}{\theta^s + Z^s}$$

$$\frac{\partial}{\partial Z} g = \frac{1}{(\theta^s + Z^s)^2} \left[ s^2Z^{s-1}(\theta^s + Z^s) - s^2Z^sZ^{s-1} \right]$$

$$= \frac{1}{(\theta^s + Z^s)^2} \left[ s^2Z^{s-1}(\theta^s + Z^s - Z^s) \right] = \frac{\theta^s s^2Z^{s-1}}{(\theta^s + Z^s)^2}$$

After substitution and setting $\frac{\partial}{\partial Z} H_{AZ}$ equal to 0 we get

$$\frac{-\sigma_dDa \sigma_aA}{(\sigma_aA + \sigma_dDa)^2} \cdot \frac{sZ^{s-1}}{\theta^s + Z^s} \cdot \frac{sZ^s}{\theta^s + Z^s} + \frac{\sigma_dDa}{\sigma_aA + \sigma_dDa} \cdot \frac{\theta^s s^2Z^{s-1}}{(\theta^s + Z^s)^2} = 0$$

$$-\sigma_dDa \sigma_aA \cdot s^2Z^{s-1} + \sigma_dDa \cdot \theta^s s^2Z^{s-1}(\sigma_aA + \sigma_dDa) = 0$$

$$-\sigma_dD \frac{\theta^s}{\theta^s + Z^s} \cdot \sigma_aAs^2Z^{s-1} + \sigma_dD \frac{\theta^s}{\theta^s + Z^s} \theta^s s^2Z^{s-1}(\sigma_aA + \sigma_dD \frac{\theta^s}{\theta^s + Z^s}) = 0$$
6.5. Results

<table>
<thead>
<tr>
<th>Conventional model</th>
<th>d-HSC model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{AA}$</td>
<td>$H_{AT}$</td>
</tr>
<tr>
<td>[ \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} ]</td>
<td>[ \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} ]</td>
</tr>
<tr>
<td>[ \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} ]</td>
<td>[ 0 ]</td>
</tr>
<tr>
<td>[ \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} ]</td>
<td>[ \frac{1}{1 + m_0 A} \cdot \frac{k_0 A}{(\delta_a + \gamma_a)(1 + m_0 A) + k_0 A} ]</td>
</tr>
</tbody>
</table>

Table 6.3: Elasticities for the saturating feedback models. The same pattern as in Table 6.2 can be observed here for the saturating feedback models, namely that the d-HSC $H_{AA}$ and $H_{AZ}$ are larger than that of the conventional model.

Since typically $s \gg 2$ and $\frac{\sigma_d D}{\sigma_a A} \leq 1$, it follows that $\sqrt{1 + \frac{\sigma_d D}{\sigma_a A}} \to 1$ and hence the elasticity will be highest when $Z$ is larger than but very close to $\theta$ or $Da \to \frac{1}{2} D$ (from below). Interestingly this implies A is best positioned to adjust to changes in $Z$ when half of the d-HSC can be either activated or switched back to a resting state.

In summary thus, based on elasticities, although the d-HSC model is at least as efficient as the conventional model in maintaining homeostasis, it is guaranteed to be more efficient in re-establishing homeostasis after injury.
6.6 Conclusion

We have shown in the previous chapters that models of observed BrdU labelling dynamics strongly support heterogeneity in the bone marrow HSC population with a small slowly cycling portion of cells (d-HSCs) responsible for long-term label retention. BrdU is, however, merely an experimental technique for observing cellular turnover kinetics, and as we discussed in Chapter 3 special care has to be taken when the BrdU labelling process is modelled and interpreted. Here our modelling effort focused on homeostasis in the hematopoietic system conceptually, how it is maintained and how it is re-established after injury. We intentionally steered clear from the constraints sometimes imposed on models by experimental data (like BrdU content in DNA that our LRC models needed to account for). In particular we analysed what the effect of a d-HSC population that can reacquire dormancy after activation would be on a mathematical model of hematopoiesis. The d-HSC hypothesis postulated it to be unlikely for d-HSCs to significantly contribute to the daily maintenance of homeostasis, and suggested these cells rather serve as a resting reserve pool that is activated during injury repair. It turns out that this hypothesis is not only supported by the modelling of experimental data, but as we’ve shown here, also by a purely conceptual mathematical approach.

Numerical simulation of equilibrium solutions revealed that a model of dynamic hematopoiesis which includes a reversibly switching d-HSC population produces simulation output that can be regarded as biologically more realistic than that of the conventional model. Biological systems re-adjusting to homeostasis after a perturbation, for example, can be expected to behave as in Figure 6.7 (two right panels), rather than the “perfect adjustment” of Figure 6.7 (two left panels). Moreover, the Hopf bifurcation leading to oscillations identified in the d-HSC model provides novel insights into hematological diseases like cyclic neutropenia and periodic chronic myelogenous leukemia that are characterised by oscillatory peripheral cell numbers (Dale & Hammond, 1988; Fortin & Mackey, 1999). Whilst the precise origin of the oscillations remains disputed, many computational models suggest they originate from the HSC compartment (Colijn & Mackey, 2005b,a, 2007). This is in agreement with our analysis, but in addition, we learned that an imbalance in the d-HSC activation pathway (or too many d-HSCs in certain cases) might be at the root of peripheral cell oscillations. The d-HSC model thus has the desirable property, since it also retains the unbounded growth dynamics of the conventional model (as was reported in Johnston et al. (2007)), of being a single dynamic hematopoiesis model that can be interrogated to analyse both unbounded and oscillatory behaviour.

The question however remained whether d-HSC hematopoiesis is evolutionarily superior to the previously widely accepted dormant-free hematopoiesis. This question is extremely difficult to answer with human reasoning alone but here we show that a quanti-
tative treatment based on elasticities, a well established tool from the field of economics, provides some useful insights that support the d-HSC model as being more effective in re-establishing homeostasis after a perturbation.
Chapter 7

Summary and outlook

This thesis evaluated experimental and conceptual support for a new paradigm in hematopoietic regulation, namely the d-HSC hypothesis. The evaluation was carried out from a systems biology perspective, and the methodology and results of various computational models on HSC division kinetics and dynamic hematopoietic regulation were described. Our initial focus was on assessing the consistency of model predictions with observed experimental data, to test whether the conventional or d-HSC-based understanding of hematopoietic regulation is more appropriate, thus conforming to the first two aims of modelling in systems biology (Section 1.3). This comprehensive study introduced a diverse range of novel computational models on two related but independent datasets, and under our assumptions all the results strongly supported the d-HSC hypothesis. With our now established d-HSC frame of reference we subsequently defined a new conceptual model on dynamic hematopoietic regulation. This model allowed for “virtual experiments” in the form of parameter interrogation (bifurcation analysis) to be performed (third aim of modelling), in order to explore the qualitative nature of d-HSC hematopoiesis. In this thesis we thus made a distinction between the modelling of LRC-based experimental data and the modelling of dynamic maintenance and restoration of homeostasis in the hematopoietic system. The former assumed homeostasis and had constant kinetic parameters. The latter modelled homeostasis by including feedback through dynamic kinetic parameters and was a conceptual model in which homeostasis was observed as emergent rather than an assumption. The major motivation for not using the dynamic model for parameter estimation is the inherent data sparsity of the LRC results. We therefore defined the LRC model as simple as possible to limit the number of parameters that needed to be estimated.

Our analysis of experimental BrdU LRC data consistently showed that among the two classes of models we evaluated, only models that account for a slowly dividing subpopulation of cells (putative d-HSCs) can describe the observed data satisfactorily. When
trying to prove dormancy in cells using BrdU-based LRC assays, a paradoxical question invariably comes to mind. *If in what we observe there are dormant cells, and hence these dormant cells are all labelled during the pulse period, can these labelled dormant cells still be considered as dormant since only cells that have cycled, thus non-dormant cells, can be labelled in the first place?* The data of Wilson et al. (2008) showing that BrdU can indirectly induce activation of HSCs, provides an elegant explanation for this phenomenon which additionally raises the possibility that the d-HSCs can reversibly switch between active and dormant states (Figure 3.1).

In Chapter 3 we showed that an ODE model can successfully capture a heterogeneous HSC population structure, but needs to carefully address the multiple nonlinear relationship between the BDT and number of divisions during uptake and chase. The model’s analytical tractability makes it especially suitable for parameter estimation and consequently useful as a guide to new biological experiments (fourth aim of modelling). For example, our analysis of the chase day 10 predictions of Figure 3.6 in Section 3.5.3 enabled us to recommend more experiments between chase day 3 and 10. Parameter estimation indicated at least a third of the HSC population are d-HSCs that divide about once every 149-193 days with a-HSCs dividing once every 28-36 days. Using these parameter estimates we can predict the relative proportion of d-HSCs amongst BrdU$^+$ LRCs at any point during the LRC experiment (shaded area of Figure 7.1). This relative proportion is of practical interest when label retention is considered as a d-HSC marker, since it predicts that after 130 days of chase the a-HSC contribution to the label-retaining pool (BrdU$^+$ HSCs) is negligible as d-HSCs would constitute more than 90% of the remaining BrdU$^+$ cells. Notably, the curve asymptotes at 95%, implicating that a totally pure population of d-HSCs can never be purified by label-retaining assays alone, underlining the need to search for other d-HSC markers. Indeed, future biological studies will be focused on elucidating new surface markers that could be used to isolate d-HSCs without the need to perform long-term label retaining experiments. Once identified, putative d-HSC-specific markers may be utilised to screen tumours for potential cancer stem cells.

ODE models of cell division kinetics have been criticised for the capriciously short inter-division times allowed for by the implied exponential distribution of cell cycle time (De Boer & Perelson, 2005; Yates et al., 2007). One solution is to implement a more complex model of the cell cycle, like the Smith-Martin model (Smith & Martin, 1973) which views the cell cycle as having two distinct phases, a stochastic phase and a deterministic phase of finite length. However, some studies have found that a Smith-Martin implementation did not give significantly different results as their original ODE model (Asquith et al., 2006).

A further issue that could be raised is whether BrdU$^+$ HSCs undergo cell cycle arrest
Figure 7.1: Using label retention as a d-HSC marker. Kinetics of uptake and loss of BrdU within the phenotypic HSCs as determined experimentally (solid black line) overlaid with the relative proportion of d-HSCs and a-HSCs amongst BrdU+ phenotypic HSCs estimated by our ODE model (dark grey shaded curve). The populations that can be found at any time point of an LRC experiment amongst phenotypic HSCs are indicated in the light grey box. The time points of chase at which the d-HSCs represent 75% and 92.5% of the LRC phenotypic HSCs are indicated by solid black circles.

during chase due to the incorporated BrdU, thereby leading to the appearance of a slow cycling subset. Although we cannot completely exclude this possibility, there are two reasons why we think this is not the case. The first is that the BrdU-based observation of the slowly cycling HSC subset was confirmed by two other studies using H2B-GFP, a second non-chromosomal labelling technique (Wilson et al., 2008; Foudi et al., 2009). Secondly, our mathematical models predict that, at the onset of chase, half of the BrdU+ cells are a-HSCs (see Figure 7.1) which indeed have “normal” rates of cycling (28-36 days) similar to what was previously estimated for the entire HSC pool.
All the findings of the ODE model were supported by stochastic modelling in Chapter 4, which ruled out cellular level variation as a cause for observing long-term label retention. A master equation approach conveniently allowed us to estimate variation for a set of parameters whilst the average random walk remained true to the deterministic prediction. Future work should focus on exploring additional unexplained variation as proposed in Section 4.4.1.

The BDT estimate of the ODE model was confirmed with the stochastic and discrete agent-based model defined in Chapter 2. However, in comparison the agent-based approach is computational much more expensive and in Chapter 5 we described how we utilised a Grid supercomputer to explore its parameter space. Although not fully exploited in this thesis, agent-based models does have the advantage that individual cells can be tracked, even removed and placed in different environments, simulating replantation studies (Glauche et al., 2009). There are many potential future directions for using agent-based models of HSCs and their regulation as a first step in the quest to reconstruct a three-dimensional HSC niche \textit{ex vivo} (Wilson & Trumpp, 2006).

We concluded our study in Chapter 6 by defining models to investigate dynamic maintenance and restoration of homeostasis in the hematopoietic system. In particular our interests were in what the impact of the d-HSC population would be on hematopoietic regulation. Bifurcation analysis revealed that malfunction in d-HSCs are likely to play a central role in hematological diseases characterised by oscillatory peripheral cell numbers. A possible future direction to explore is to experimentally evaluate whether peripheral and progenitor blood cells do indeed exhibit damped oscillations during injury repair, as predicted by the d-HSC model. If they do, the period of these oscillations (before they dampen out) together with our LRC data can be used to infer real parameter values for the models defined here.

Using elasticities as a measure, we provided well-grounded arguments that d-HSC regulation of hematopoiesis is evolutionarily more likely than conventional (d-HSC free) regulation of hematopoiesis. The application of elasticities can be further extended to study the dynamics of noise and variation in HSC populations, since elasticities are the main determining factor of drift in the fluctuation-dissipation theorem (Paulsson, 2004, 2005).

It should be noted that our conclusions of Chapter 6 were made on the basis of comparing two classes of models each representing opposing views of hematopoietic regulation as dictated by current biological knowledge (Section 1.3.3). We endeavoured to have each model type capture the general essence of the biological hypothesis it represents by firstly incorporating a previously published model and then adopting minimalistic model equations. Another possible future direction is to extend more complex models of hematopoietic...
poiesis, like the delay differential equation model of Colijn & Mackey (2007) that already exhibits oscillations, to a d-HSC version. The qualitative and quantitative properties of the original and existing models can then be evaluated and compared in a similar fashion as we have done in Chapter 6. We are of the opinion that the discovery of the existence of the d-HSC population not only has important implications for future mathematical modelling efforts of HSCs specifically, but also adult stem cell systems in general, since putative dormant stem cell populations have recently been observed in the skin as well (Waghmare et al., 2008).

Biologically, the discovery of the d-HSC population has already made an impact on cancer therapy. Upon different stimuli such as 5-FU, G-CSF or IFN-α (Wilson et al., 2008; Essers et al., 2009), d-HSCs can exit dormancy and proliferate to the same rate as their active counterparts. As most current chemotherapeutic strategies target actively cycling cells, dormant and oncogenically mutated stem cells would be immune to such treatment, and could potentially cause leukemic relapse unless they can be pharmacologically stimulated to enter an active state. Hence the challenge is to understand which drugs should be combined with which chemotherapeutic agents in order to eradicate even the most primitive cancer stem cells (reviewed in Viale & Pelicci (2009)). To find such drug targets that activate dormant cells more needs to be known about the molecular machinery that controls activation. Here computational methods like the Bayesian models and stochastic search techniques described in Brown et al. (1998), can be applied to gene expression data to identify transcription factor binding sites from a collection of co-regulated genes and their nearby non-coding DNA sequences (Tadesse et al., 2004; Angelini et al., 2007). Moreover, it has been shown previously that computational models can help to optimise therapeutic strategies (Abkowitz et al., 1997, 2004; Obeyesekere et al., 2004). If the pharmacologically activated d-HSCs (and hence dormant cancer stem cells as well) also exhibit damped oscillations as predicted by our d-HSC model, elucidating the period of the oscillations can assist in optimising cancer stem cell exposure to drugs by carefully timing chemotherapy to be applied during active cycles.

7.1 Reflections

In this thesis we described computational models of biological data and dynamic biological systems. As mentioned before this work can thus be classified in the rapidly growing field of systems biology and hence was a multidisciplinary effort that needed to be done in close collaboration with biologists. For such a collaboration to be possible, mutual understanding of the respective fields involved were impeccable. High quality models cannot be developed if the underlying system to be modelled is not understood sufficiently.
Similarly biologists cannot fully understand results and properly convey their expertise if their understanding of the computational approach and terms is not deep enough.

In this regard the work presented here serves as an example of a successful multi-disciplinary collaboration built on several mutual visits to the respective laboratories. Nevertheless, we are of the opinion that systems biology is yet to reach its full potential. Even in this work we have to take note that the biological experiments were designed from a purely biological point of view constrained by the biologists’ frame of reference. The intention was for the results to be suitable for human interpretation, i.e. not taking into account the potential power and abilities of computational modelling. If either the biologists or model builders have a really deep understanding of each other’s field they can use this knowledge to influence the design and results of biological experiments so as to utilise the power of mathematical modelling much more effectively.

We are further of the opinion that the implications of our results had a higher than normal impact because it is supported by good quality data on a fairly high scale. The impression we got was that modelling and inference in biology on a higher scale (cellular turnover rates or individuals themselves for example) tends to have a more direct clinical impact than modelling systems on a lower scale (rate of molecular rates in a pathway for example). At the same time higher scale data tends to be less abundant and often more difficult to obtain which underlines how fortunate we were to have exclusive access to the LRC dataset given in Tables 3.1 and 3.2.
Appendix A

Acronyms and definitions

A.1 List of acronyms

a-HSC  Active hematopoietic stem cell.
AP     Active hematopoietic stem cell proportion.
ASC    Adult stem cell.
BDT    BrdU detection threshold.
BrdU   Bromodeoxyuridine.
CDF    Cumulative Distribution Function.
CE     Computing Element.
CFSE   Carboxyfluorescein diacetate succinimidyl ester.
CSC    Cancer stem cell.
d-HSC  Dormant hematopoietic stem cell.
DAG    Directed acyclic graph.
DIOF   D-independent-oscillation-free.
DNA    Deoxyribonucleic acid.
DP     Dormant hematopoietic stem cell proportion.
EGEE   Enabling Grids for E-sciencE.
ESC    Embryonic stem cell.
H2B-GFP Histone H2B-green fluorescent protein.
HIV    Human immunodeficiency virus.
HSC    Hematopoietic stem cell.
iPSC   Induced pluripotent stem cell.
JDL    Job Description Language.
JVM    Java Virtual Machine.
K-S    Kolmogorov-Smirnov.
LCG    Large Hadron Collider Computing Grid.
A.2 Biological nomenclature

Definitions of concepts not explained in the text are as given by Wikipedia or Princeton University’s WordNet®.

Apoptosis Programmed cell death or natural cell suicide, in contrast to necrosis which is premature death of cells usually due to injury or similar external factors.

Assay Evaluation of the biological activity of a substance by testing its effect on an organism and comparing the result with some agreed standard.

Carcinogenesis The process by which normal cells are transformed into cancer cells (the creation of cancer).

Enzyme Biomolecules, usually proteins, that catalyse (i.e., accelerate the rates of) chemical reactions.

Eukaryote Multicellular or single-celled organisms whose cells contains a membrane bound nucleus. Eukaryotes includes most organisms but excludes viruses, bacteria, archaea, etc. which are prokaryotes.

Ex vivo In an artificial environment outside the living organism.

Fixation of a cell A chemical process that preserve the state of the cell as much as possible, but usually results in the cell being killed.
Flow cytometry Automated technique for counting and sorting individual cells by detecting fluorescently dyed antibodies bound to the cells. Commonly referred to as FACS.

Genome The complete hereditary information (either DNA or, in some viruses, RNA) of an organism.

Hematopoiesis The formation of blood cells in the living body (especially in the bone marrow).

Homeostasis A state of stable equilibrium. In the context of hematopoiesis homeostasis refers to the equilibrium maintained in the daily production and loss of millions of blood cells by hematopoietic stem cells.

In silico An expression used to mean "performed on computer or via computer simulation."

In situ To examine phenomena exactly in place where it occurs (i.e. without moving it to some special medium). This usually means something intermediate between in vivo and in vitro.

In vitro Refers to the technique of performing a given procedure in a controlled environment outside of a living organism.

In vivo Refers to experimentation using a whole, living organism as opposed to a partial or dead organism, or an in vitro controlled environment.

Intraperitoneal injection The injection of a substance into the peritoneum (body cavity).

Isotope One of two or more atoms with the same atomic number but with different numbers of neutrons.

Lymphocytes A type of immune cell that normally makes up a quarter of the white blood cell count but increases in the presence of infection.

Metastasis The spreading of a disease (especially cancer) to another part of the body.

Pathway A metabolic pathway is a series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions.

Ploidy A term that refers to the number of complete sets of chromosomes in an organism. Human cells are diploid as they contain two complete sets of chromosomes, with the exception of sex cells which are haploid.
Proteome  The full complement of proteins produced by a particular genome.

Somatic cell  Adult, fully differentiated cell.

Stem cell  Special cells that has the unique ability to either self-renew (produce copies of themselves) or differentiate into mature specialised progeny (cells that the body needs to grow and maintain itself).

T-cell  T cells belong to a group of white blood cells known as lymphocytes, and play a central role in cell-mediated immunity.

Telomere  A short repeated DNA sequence at the end of a chromosome that protects the end of the chromosome from destruction and whose length is reduced with each cell division. The telomere shortening mechanism normally limits cells to a fixed number of divisions.
Appendix B

Grid deployment and example files

B.1 Grid deployment

In this section we relate some technical details about the EGEE Grid architecture and how it facilitates job submission. We further explain how individual jobs and resulting workflows are defined. Since the serialised objects and simulation output files of our agent-based model are relatively large, we conclude this section by describing how dataflow is managed by the gLite middleware layer.

B.1.1 EGEE Grid architecture

In order to submit jobs to the Grid, all users are required to be a member of a Virtual Organisation (VO) and have a valid EGEE Grid certificate. A VO comprises a sampling of Grid users sharing similar requirements and interests. VO groupings help to ensure the integrity of data stored on the Grid network and promote the linking of like minded research tools, algorithms and collaborations.

Middleware is a key component to any Grid computing effort for it serves as the communication layer which enables interaction across hardware and network environments. Initially the Large Hadron Collider Computing Grid (LCG - http://lcg.web.cern.ch) middleware stack was used on the EGEE infrastructure. Most of this stack was later re-developed and re-engineered into the current middleware solution, gLite. The gLite Grid services (http://glite.web.cern.ch) follow a service-oriented architecture advancing compliance with upcoming Grid standards and are currently widely deployed on hundreds of sites as part of the EGEE project. gLite provides a comprehensive suite of software tools that enable job submission to be oblivious to the underlying heterogeneity of the Grid.
Figure B.1: **Job submission** (adapted from Lingrand et al. (2009)).  
**A:** gLite Workload Management System (WMS) architecture: The WMProxy enables a connection to the Workload Manager which assigns a submitted job to the target Computing Element. This target site is determined with information received from the Resource Broker and its Information Supermarket. Job submission is finalised through the Job Adapter after which the Workload Manager delegates the processing to CondorC. The Log Monitor regulates job processing and intercepts events that affect the job state machine (see B), these events are logged by the Logging and Bookkeeping service.  
**B:** Life cycle state machine: a graphical interpretation of the state machine that internally controls a job’s life cycle.

**Job submission**

In order to submit a job a valid Grid certificate has to be uploaded to a Virtual Organisation Membership Services (VOMS) server which initiates a VOMS proxy to handle all security clearance for a defined period of time. A Grid certificate is obtainable from the EGEE Grid Certification Authority which requires valid personal identification as well as credible scientific motivation before a licence is issued (these licences are renewed yearly).

The job life cycle is internally controlled through a state machine displayed in Figure B.1. All job submission requests and controls are handled by the gLite Workload Management System (WMS) which are accessed through a web service component called WMProxy. To access the gLite WMS a client needs to create a VOMS proxy as well as a WMProxy and link them to each other through a delegation process. Once these proxies have been successfully initiated, jobs can be submitted to the client’s specific VO. Two service operations, `jobRegister` and `jobStart`, are called in sequence at actual job submission. When a `jobRegister` request arrives at the WMProxy and the client has the
rights to proceed, a set of specific attributes needed by the WMS for handling the request appropriately, together with a generated unique job identifier, are inserted into the job description. The requesting user is then mapped to a local user by means of Local Credential Mapping (LCMAP) which provides authorisation functionalities based on VOMS - resulting in the job, local directories, and files being created with appropriate ownership and permissions. When all the aforementioned steps have been successfully completed, the job with the generated job identifier and the enriched job description are registered to the Resource Broker (RB). From that point on the job is uniquely identified and can be monitored throughout the system and the various job states with its identifier (van der Wath et al., 2008). The RB will submit the job to a suitable Computing Element (CE) by means of adaptive scheduling algorithms which considers factors like CE availability, network transfer distance, user set requirements and other factors.

B.1.2 JDL design and job types

Job submission requires both a description of the job to be executed and a description of the needed resources. On submission the RB assigns the job to a suitable CE where it is in turn assigned to a Worker Node (WN) and actual execution takes place. These job descriptions and execution requirements are provided with a Grid-based high-level language called the Job Description Language (JDL). The JDL is based on Condor classified advertisements (Avellino et al., 2003) for describing jobs and aggregates of jobs. Such aggregates include DAG (directed acyclic graph), MPICH and parametric jobs. Jobs of type DAG represent a set of jobs, defined as nodes, where the input and output of one or more nodes are dependant on the execution of one or more other nodes. Dependency or workflow is controlled by the setting of a dependencies variable in the JDL file. For example, a specification of \{ nodeA, nodeB \} would restrict the execution of nodeB to the successful completion of nodeA. MPICH is a freely available implementation of the Message Passing Interface (MPI) standard which facilitates inter WN communication. The parametric job type allows a user to iterate through a parameter list of strings or consecutive numeric values where the starting (ParameterStart), increment (ParameterStep) and end values (Parameters) of parameters are defined. The submission of such a JDL will result in the generation of $N$ jobs, where $N = (\text{Parameters} - \text{ParameterStart})/\text{ParameterStep} + 1$. The substring _PARAM_ in the JDL file is replaced, on submission, with the parameters specified in the Parameters attribute. As mentioned in Section 5.2.2 our parallel search comprises of ten DAG jobs, one for each of the DP values within the range of 5% to 50% with incremental steps of 5%. Each one of these DAG jobs has three nodes whose execution depend on the completion of the previous node (except for node 1 of course). Node 2 defined a parametric job, iterating through twenty values of the BDT. The third
and last node calculates the best parameter set of each main DAG job by referencing the output files that were stored by node 2 on the Grid. An example JDL file that describes the DAG job for a 35% DP is shown in Section B.2.

Some CEs are configured for shorter jobs and hence abort jobs that exceeds their predefined time limit. Such CEs usually ends with an affix such as “sixhour”, “short”, “q30m” or “medium” and if required they can be excluded by specifying them in the Requirements attribute of the JDL file. Below is an extraction from the example JDL file in Section B.2 which shows the CEs that are excluded.

\[
\text{Requirements} = ( \\
\quad \text{other.GlueCEUniqueID} \neq \text{"gridba2.ba.infn.it:2119/jobmanager-lcgpbs-short"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"cclcgceli03.in2p3.fr:2119/jobmanager-bqs-medium"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"cclcgceli04.in2p3.fr:2119/jobmanager-bqs-medium"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"cclcgceli03.in2p3.fr:2119/jobmanager-bqs-short"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"cclcgceli04.in2p3.fr:2119/jobmanager-bqs-short"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"ce01.dur.scotgrid.ac.uk:2119/jobmanager-lcgpbs-q6h"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"gridgate.cs.tcd.ie:2119/jobmanager-pbs-sixhour"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"ce02.dur.scotgrid.ac.uk:2119/jobmanager-lcgpbs-q30m"} \\
\quad \&\& \text{other.GlueCEUniqueID} \neq \text{"gridgate.cs.tcd.ie:2119/jobmanager-pbs-one day"} \\
) ;
\]

In general the files that have to be moved between the user interface (from where the jobs are submitted) and the executing WN are defined in the JDL InputSandbox and OutputSandbox attributes. Files listed in the InputSandbox are copied to the WMS which downloads it onto a selected WN where actual execution takes place. All the nodes of a DAG job have access to each other’s output files but are limited by the user specified dependencies. The main DAG job has its own InputSandbox which is referenced as root. When many of the nodes access the same main files it is recommended to send these files only once to the WMS by using the root’s InputSandbox. The first file in this InputSandbox can be referenced by any of the nodes with the command root.InputSandbox[0], the second by root.InputSandbox[1] and so forth. Other files within the various nodes can be accessed by adding the correct prefix, for example root.nodes.phase1.description.OutputSandbox[0] will reference the first output file of node ‘phase 1’. All the output files that the user would like to retrieve are listed under the OutputSandbox attribute. These files are then downloaded back to the user interface by using the glite-wms-job-output command. Overall computing performance and Grid traffic are notably affected by very large file transfers. For this reason the sandboxes’ capacity are limited to 10 MB. In the next section we describe data management on the Grid for jobs with large input and output files.
B.1.3 gLite data management

When a workflow requires the move of large files or large quantities of files (as is the case for our parallel search), the files are specified in the InputData and OutputData JDL attributes which is another method of moving files to and from the WNs. In such cases resulting or initial files are stored on specified Grid Storage Elements (SE) and registered on a file catalogue. Jobs requesting these files are consequently allocated to WNs belonging to the CE related to the SEs on which the requested files are stored which reduces transfer times. The RB is able to detect what SEs contains the requested files by accessing the Logical File Catalogue (LFC) server which maps the logical file names of stored files to all their (possibly several) copies and related site URLs.

In order to specify SEs wisely it is possible to visually inspect the congestion status of all the CEs available within a particular VO by using the lcf-infosites command. This useful command allows the user to differentiate among the many CEs with respect to the amount of jobs that are currently running and waiting, the amount of free WNs, response time statistics, and also supplies the corresponding SEs’ addresses.

It is possible to distribute copies of input files to multiple SEs using the lcg-rep command. This is especially useful when many parametric jobs access the same input files and hence could easily congest a single SE. A list of all the replicates of a file can be obtained with the lcg-lr command which prints a list of all the replica storage URLs or physical file names to the screen. These storage URLs are printed in the general form <srm>://<SE_hostname>/<some_string> and since a virtual file system is used, this file name does not have any relation to its physical location. Unwanted files are removed with the lcg-del command.

Running Java applications on the Grid

Usually, on a normal workstation, a Java executable (class file) or application suite (archived jar file) is loaded into the preinstalled Java Virtual Machine (JVM) by the commands java classfile or java -jar jarfile respectively. For Java applications on the Grid, these commands have to be wrapped in an executable script file and then this script file defined as the job’s executable in the corresponding JDL file. These executable script files are also often used to transfer the files referenced by the InputData JDL attribute from the SE to the WN before actual computation commences. A typical script wrapper used in our parallel search are given below.

```
#!/bin/bash

export LFC_HOST=lfc-biomed.in2p3.fr
```
B.2. JDL file

The `lcg-cp` command is used to copy the file from a nearby SE to the current executing location (a WN) by referencing the file catalogue. Before the file catalogue can be referenced the LFC environment variables need to be set on the referencing WN.

B.2 JDL file

Below is the JDL code for the 35% DP main DAG job. The `max_nodes_running` attribute, with a default value of 10, limits the number of nodes running simultaneously. We set this attribute to 20 as our phase 2 is a parametric job which submits 20 jobs in parallel where we varied the BrdU detection threshold by iterating the parameter variable from one to twenty.

```plaintext
export LCG_CATALOG_TYPE=1fc
export GRID_PATH=/grid/biomed/Cells400
EXIT_VALUE=0

lcg-cp -v --vo biomed lfn:${GRID_PATH}/24h_35DP.ser file:$PWD/24h_35DP.ser
java -jar "$@

exit $EXIT_VALUE
```

[Type = "Dag";
VirtualOrganisation = "biomed";
max_nodes_running = 20;
InputSandbox = {
  "/home/files/HSC_divrate_sim_grid.jar",
  "/home/files/repast.jar",
  "/home/files/colt.jar",
  "/home/files/trove.jar"
};

Nodes = [
  phase1 = [
    Description = [
      JobType = "normal";
      Executable = "runjar_24h";
      Arguments = "-Xms50m -Xmx200m HSC_divrate_sim_grid.jar 24h parameters_24h.txt 0.35 1 24h_35DP 400 50";
      StdOutput = "24h_out_35DP.txt";
      StdError = "24h_err_35DP.txt";
      InputSandbox = {
        root.InputSandbox[0],
        root.InputSandbox[1],
        root.InputSandbox[2],
        root.InputSandbox[3],
      }
    ]
  ]
];
```
B.2. JDL file

```
B.2. JDL file 151

"/home/35DP/runjar_24h.txt",
"/home/35DP/parameters_24h.txt",
"/home/files/24h_expdata.txt"
);
OutputData = {
    OutputFile = "24h_35DP_states_50runs.ser"
    StorageElement = "pccms2.cmsfarm1.ba.infn.it"
    LogicalFileName = "lfn:/grid/biomed/Cells400/24h_35DP_states_50runs.ser"
};
OutputSandbox = {
    "24h_out_35D.txt",
    "24h_err_35D.txt",
    "24h_35D_data.txt",
    "24h_35D_parameters.txt",
    "24h_35D_bestParams_1.out",
    "24h_35D_RSS.out",
    "24h_35D_SIM_data.txt",
    "24h_35D_SIM_parameters.txt"
};
RetryCount = 3;
};
upch_chase = [
    Description = [
        JobType = "Parametric"
        Executable = "runjar_upch_chase"
        Arguments = "-Xms50m -Xmx300m HSC_divrate_sim_grid.jar upch_chase 24h_35DP_states_50runs.ser
        parameters_upch.txt parameters_chase.txt 0.35 _PARAM_ upch_chase_35DP 400 50"
    StdOutput = "upch_chase_output_PARAM_.txt"
    StdError = "upch_chase_error_PARAM_.txt"
    Parameters = 21;
    ParameterStart = 1;
    ParameterStep = 1;
    InputSandbox = {
        root.InputSandbox[0],
        root.InputSandbox[1],
        root.InputSandbox[2],
        root.InputSandbox[3],
        root.InputSandbox[4],
        "/home/35DP/runjar_upch_chase",
        "/home/35DP/parameters_chase.txt",
        "/home/35DP/parameters_upch.txt",
        "/home/files/upch_expdata.txt",
        "/home/files/chase_expdata.txt"
    }];
InputData = {
    "lfn:/grid/biomed/Cells400/24h_35DP_states_50runs.ser"
};
DataAccessProtocol = {"gridftp","rfio","gsiftp"};
OutputSandbox = {
    "upch_chase_output_PARAM_.txt",
```

OutputData = {
    [OutputFile = "upch_chase_35DP_UPCH_bestParams__PARAM_.out"; StorageElement = "pccms2.cmsfarm1.ba.infn.it"; LogicalFileName = "lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams__PARAM_.out";],
    [OutputFile = "upch_chase_35DP_CHASE_bestParams__PARAM_.out"; StorageElement = "pccms2.cmsfarm1.ba.infn.it"; LogicalFileName = "lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams__PARAM_.out";]
};

RetryCount = 3;
rank = (-other.GlueCEStateEstimatedResponseTime);
};

upch_chase_bob = [
    Description = [
        JobType = "normal";
        Executable = "runjar_upch_chase_bob";
        Arguments = "HSC_divrate_sim_grid.jar upch_chase_bob upch_chase_35DP 20";
        StdOutput = "upch_chase_bob_out.txt";
        StdError = "upch_chase_bob_err.txt";
        OutputSandbox = {
            "upch_chase_out_bob.txt",
            "upch_chase_err_bob.txt",
            "upch_chase_35DP_best_of_best.out"
        };
        InputSandbox = {
            root.InputSandbox[0],
            root.InputSandbox[1],
            root.InputSandbox[2],
            root.InputSandbox[3],
            root.InputSandbox[4],
            "/home/35PD/runjar_upch_chase_bob"
        };
    ];
    InputData = {
        "lfn:/grid/biomed/Cells400/24h_35DP_states_50runs.ser"
B.2. JDL file

```
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_1.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_2.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_3.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_4.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_5.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_6.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_7.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_8.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_9.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_10.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_11.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_12.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_13.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_14.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_15.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_16.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_17.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_18.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_19.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_UPCH_bestParams_20.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_1.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_2.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_3.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_4.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_5.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_6.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_7.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_8.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_9.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_10.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_11.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_12.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_13.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_14.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_15.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_16.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_17.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_18.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_19.out",
"lfn:/grid/biomed/Cells400/upch_chase_35DP_CHASE_bestParams_20.out"
];

DataAccessProtocol = {"gridftp","rfio","gsiftp"};

rank = (-other.GlueCEstadoEstimatedResponseTime);
Requirements = (other.GlueCEstadoUniqueID == "gridba2.ba.infn.it:2119/jobmanager-1cpbfs-short");
]
];

}
}

Dependencies = {[phase1, upch_chase],[upch_chase,upch_science_bob]};
rank = (-other.GlueCEstadoEstimatedResponseTime);
requirements = (other.GlueCEstadoStatus=="Production");
]
B.3 Repast parameter file

The Repast parameter file that was used for the 35% DP search is given below.

```plaintext
runs: 1
DorProbSelfrenew {
  start: 75
  end: 225
  incr: 30
  {
    runs: 1
    DorProbDifferentiate {
      start: 1
      end: 3
      incr: 1
      {
        runs: 50
        ActProbSelfrenew {
          start: 14
          end: 59
          incr: 15
        }
      }
    }
  }
}
```
List of Figures

1.1 The hematopoietic hierarchy (adapted from Bryder et al. (2006)). The hierarchical organisation of the hematopoietic system has long been recognised, with rarely-dividing multipotential HSCs producing rapidly dividing lineage-restricted transit-amplifying and committed progenitors which in turn will give rise to all differentiated cell types of the blood. HSCs with the strongest self-renewal potential is at the top and each successive node towards the bottom of the tree represents stages of differentiation. The proliferative index indicates division rate and assumes a water droplet shape with the leaf nodes (mature cells) not proliferating at all. ........................................ 6

1.2 Mitotic cell division. This is the process through which an eukaryotic cell duplicates itself. Firstly the genetic material is copied and doubled. Mitosis follows, during which the cell’s duplicated chromosomal material is separated and moved to opposite sides. Cytokinesis completes the process by dividing the cytoplasm to construct the two individual daughter cells. ........................................ 7

1.3 Bromodeoxyuridine. A: Normal hydrogen bond between adenine and thymine (top) and BrdU substituting for thymidine (bottom). B: The chemical structure of thymidine and BrdU. ........................................ 9

2.1 Chromosome segregation. Observed BrdU dynamics in stem cells (encircled with black) would be different for asymmetric and random chromosome segregation as depicted here (BrdU+ DNA strands indicated with light green, immortal strands in red). In the asymmetric case BrdU intensity remains constant after the first division during uptake and cells loose all labelling after the first chase division. In contrast, many different patterns are possible for both uptake and chase in the random segregation case. ........................................ 19

2.2 Chromosome implementation. Each DNA strand of a chromosome is represented by a boolean of which the value indicates whether the strand is BrdU+ or BrdU-. ........................................ 28
2.3 **Agent-based approach to time modelling.** The grey box at the top shows a toy example of BrdU dynamics for 4 cells in real time. Cell division events are shown as circles along the time axis with the amount of green fill indicating BrdU intensity. The period up to the red dotted line represents BrdU uptake (intensity increases) and the period thereafter represents BrdU chase (intensity decreasing). The white box at the bottom shows how this toy example would be modelled with our agent-based model. Each individual cell is tracked and at fixed discrete intervals of time (ticks) the probability of dividing is calculated for each cell agent. Some divisions might be missed (black arrows - tick 7) if more than one occurs during a tick ................................................................. 29

2.4 **The deterministic model by Kiel et al. (2007),** showing the model prediction for: BrdU uptake, asymmetric segregation BrdU chase, and random segregation BrdU chase on different detection thresholds. **A:** The prediction when using the original random segregation uptake equations (Equations 2.2.5 & 2.2.6). **B:** The prediction when using the alternative random segregation uptake equations (Equations 2.2.8 - 2.2.10). ............... 33

2.5 **The proportion of** \(y_2\) **and** \(y_1\) **during BrdU uptake.** **A:** The prediction of the original random segregation uptake equations (Equations 2.2.5 & 2.2.6). **B:** The prediction of the alternative random segregation uptake equations (Equations 2.2.8 - 2.2.10). ........................................ 34

2.6 **Homogeneous agent-based model simulation results.** **A:** Showing each of the 50 unique simulations along the z-axis. **B:** The average (of 50 simulations) predicted percentage of 400 cells that are BrdU positive after an uptake period of 10 days. A BDT of 20 chr means at least 20 chromosomes needs to be labelled before the agent is regarded as BrdU+. .. 35

2.7 **Homogeneous agent-based model predictions.** Showing individual simulations with their average and 95% confidence interval. **A:** Asymmetric segregation model predictions. **B:** Random segregation model predictions on a 20 chromosome BDT. .................................................. 36

2.8 **Heterogeneous agent-based model predictions assuming random chromosome segregation.** **A:** Single-phase turnover. **B:** Bi-phase turnover. ................................................................. 37

2.9 **Heterogeneous agent-based model predictions assuming random chromosome segregation.** Showing 50 individual simulations with their average and 95% confidence interval on a 20 chromosome BDT. **A:** Single-phase turnover. **B:** Bi-phase turnover. ................................................................. 38
2.10 **Heterogeneous agent-based model predictions assuming immortal-immortal chromosome segregation.**  
A: Single-phase turnover. B: Bi-phase turnover. ........................................... 39

2.11 **Heterogeneous agent-based model predictions assuming immortal-random chromosome segregation and a 20 chromosome BDT.**  
A: Single-phase turnover. B: Bi-phase turnover. ........................................... 39

3.1 **Conventional vs. dormant population HSC hierarchy.** Within the HSC population, two possible HSC hierarchy models can be envisaged. In the conventional model (left panel), the HSC population is homogeneous with respect to cell cycle entry with the entire HSC pool turning over every few weeks. In contrast, in the dormant HSC model (right panel), the hierarchical organisation of the hematopoietic system includes the phenotypic HSC pool, in which two subpopulations can be defined based on their relative turn-over frequencies. An active HSC (a-HSC) population is responsible for the day-to-day maintenance of the hematopoietic system, while a second population, the dormant HSC pool (d-HSC), cycles only a few times over the life span of the mouse in a homeostatic situation (dashed arrow) but is activated and participates in replenishment of the hematopoietic system after injury (solid arrow). LT: long-term. ................. 45

3.2 **Plot of experimental data in Tables 3.1 and 3.2.** Green line: BrdU uptake; Blue line: BrdU chase starting at day 0; red x: mean observed data. The long-term label retention is clearly visible and was found most interesting by biologists. .................................................. 46

3.3 **Mitosis schematic of BrdU uptake and chase.** Change in intensity in a population of cells during uptake is uniform, but in practice each intensity is transformed to a boolean classification of labelled or not labelled. It is thus not possible to directly know the number of times a cell has divided. During BrdU chase many different intensity patterns are possible, depending how chromosomes segregate. ................................. 50
3.4 **Average theoretical BrdU percentage of a single cell for a given number of divisions during uptake and chase.** Yellow: cell has divided once during uptake; green: cell has divided twice during uptake; red: cell has divided thrice during uptake; violet: cell has divided four or more times during uptake; red lines: detection thresholds. Slices of each uptake pie are cumulative and BrdU percentage is calculated in terms of the number of DNA strands that are labelled and not in terms of the proportions of nucleotides that are replaced by BrdU. Uptake divisions follow in a clockwise direction and chase divisions follow in an anticlockwise direction. Slices intersected by red lines indicate the number of divisions for the cell to fall below the detection threshold. ................................. 55

3.5 **Observed experimental data of BrdU uptake and chase.** Green: uptake; blue: chase; red x: mean observed data; vertical dotted line: time at which BrdU was removed; shaded area: stage 2 parameter estimation region. Plotting BrdU uptake on a timescale of hours and chase on a timescale of days clearly shows a markable change in kinetic slope at 5 time points (black arrows). ................................. 64

3.6 **Model predictions of BrdU content during stage 1 & 2.** Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. Predictions are shown on a timescale of hours and the BrdU chase period starts at 312 hours (13 days). Clearly both models can satisfactorily describe stage 1 & 2 data (uptake and first 10 days of chase). **A:** One-population model prediction (stage 1 MSE = 1.54, stage 2 MSE = 6.56). **B:** Red line: d-HSCs; green line: a-HSCs. Two-population model prediction with a 6 or 7 strand BDT and 40% DP (stage 1 MSE = 1.5, stage 2 MSE = 3.89). ................................. 69

3.7 **Model predictions of BrdU content during stage 3.** Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. **A:** One-population model predictions with a 9 strand BDT. (RSS = 391.9, stage 3 MSE = 27.54). **B:** Red line: d-HSCs; green line: a-HSCs. Two-population model predictions with a BDT of 6 or 7 strands and 40% DP. This model gave the best overall goodness-of-fit (RSS = 71.7, Stage 3 MSE = 3.22). Activation of the d-HSCs can be clearly seen in Figure 3.6B where both d-HSC and a-HSC are predicted to take BrdU up at the same rate (cycling about once every 10 days). During chase d-HSCs return to a dormant state and are predicted to divide about once every 165 days, whilst a-HSCs divide once every 31 days, diluting label much faster than the d-HSCs. ................................. 72
3.8 **Two-population model predictions with a 4 strand BDT and 30% DP.** Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. **A:** Stage 1 & 2 predictions (uptake and first 10 days of chase) on a timescale of hours. The effect of a smaller d-HSC population is clearly visible (stage 1 MSE = 1.51, stage 2 MSE = 9.42). **B:** Red line: d-HSC; green line: a-HSC. Uptake and chase predictions on a timescale of days (RSS = 113.8, stage 3 MSE = 3.94). ........................................... 73

3.9 **Proportion of c1 and c2 in BrdU$^+$ cells during uptake.** **A:** 6 & 7 BDT - 40% DP. Here $c_2$ represents cells that have divided 2 or more times. **B:** 4 BDT - 30% DP. In this case $c_2$ represents cells that have divided 3 or more times. ......................................................... 74

3.10 **Model predictions using homogeneous parameters.** Dashed line: HSC uptake; brown line: HSC chase; blue x: observed data. **A:** One-population model. **B:** Two-population model. .................. 75

4.1 **Example of a possible skew sample.** The distribution of the % BrdU$^+$ HSCs of a large mouse population would be approximately normal (blue line). If the mice that we examine (red circles) happen to fall in the tail of this normal distribution we might estimate actual BrdU retention higher (or lower) than what it really is. ....................................................... 79

4.2 **Markov process approach to time and event modelling.** The grey box depicts the same toy example as in Figure 2.3. The white box shows how the toy example would be modelled by an MP. Here we have a continuous implementation of time so that every event is accounted for. Each event initiates a jump to a new state, with the state determined by the label intensity of each cell. ....................................................... 81

4.3 **Variation estimates using the one-population model.** **A:** Dashed line: average uptake; brown line: average chase; shaded area: ± 3 standard deviations; blue x and error bars: observed data. **B:** Predicted density at chase day 306. .......................... 92

4.4 **Variation estimates using the two-population model.** **A:** Dashed line: average uptake; brown line: average chase; shaded area: ± 3 standard deviations; blue x and error bars: observed data. **B:** Predicted density at chase day 306. .......................... 92

5.1 **Parallel brute-force search strategy.** Here we show a flow diagram of one of the 10 main DAG jobs (35% DP). Our complete strategy encompasses 10 such DAG jobs. ............................. 101
5.2 Predictions results of best parameter set. Dashed line: average uptake; brown line: average chase; shaded area: \( \pm 3 \) standard deviations; blue x and error bars: observed data. ................................................. 107

6.1 Schematic showing the conventional model of hematopoietic regulation. a-HSCs (A) self-renew and differentiate into transit amplifying cells (T) which in turn also self-renew and differentiate into mature cells (Z). Feedback is indicated with dashed arrows. .............................. 112

6.2 Stability regions (from Johnston et al. (2007)) A: The linear feedback model (Equation 6.2.2) either predicts a finite cell population (\( \alpha \geq 0 \)) or extinction (\( \alpha \leq 0 \)). B: The saturating feedback model exhibits similar dynamics but can be driven to a state of unbounded growth if \( \alpha \) or \( \beta \) exceeds certain threshold values. ................................. 114

6.3 Activation function. The steepness of the activation function is determined by \( s \), with smaller values for \( s \) indicating more gradual activation (dark blue line) and larger values more abrupt activation (green line). .......... 115

6.4 Schematic of the d-HSC hematopoietic regulation model. Like in the conventional model, a-HSCs (A) self-renew and differentiate into transit amplifying cells (T) which in turn also self-renew and differentiate into mature cells (Z). The d-HSC model has a constant dormant pool \( D \) which, depending on the population level of Z, is reversibly activated to \( D_a \). When cells in \( D_a \) divide, one daughter enters the A pool whilst the other remain in \( D_a \) (and hence can switch back to the \( D \) pool). Feedback is indicated with dashed arrows. ......................... 116

6.5 Complex plane. Example of how a conjugated pair of eigenvalues cross the imaginary axis \( \text{Im}(\lambda) \) during a Hopf bifurcation. ................................. 120

6.6 Dynamic hematopoiesis model simulations. Time series simulation of the conventional (top) and d-HSC (bottom) models. Panels from left to right are: model representation; simulation output for A, T, and Z; simulation output for A and T in detail; T vs. Z phase portrait. Most notable are the damped oscillations of the d-HSC model when it adjusts to steady state as opposed to the non-oscillatory adjustment of the conventional model. The steady state attractor of the d-HSC model is known as a stable spiral and can be seen in the phase portrait on the right. Starting values are arbitrary and parameter values used in both simulations are given in Table 6.1. ................................. 125
6.7 **Behaviour during injury.** Initial conditions are set to steady state values and the re-adjustment to steady state by each model demonstrated for a perturbation in each of $A$, $T$, or $Z$. Parameter values are as in Figure 6.6.

6.8 **Dynamic hematopoiesis model simulations showing self sustained oscillations.** Top panel: oscillations take longer to damp out when $D$ is increased from 2 to 3. Bottom panel: when $D$ equals 4 a Hopf bifurcation has been reached and the system oscillates at a limit cycle. Panels from left to right and parameter values other than $D$ are as in Figure 6.6.

6.9 **T vs. Z phase plane during oscillations.** The initial stable spiral is now unstable and solutions are attracted to a newly formed stable limit cycle surrounding the unstable spiral. This behaviour is characteristic of a Hopf bifurcation.

6.10 **Two parameter bifurcation diagrams.** Blue areas: oscillatory regions. Steady states where evaluated for parameter values as given in Table 6.1. The bifurcation diagrams suggest oscillations occur when the size of the d-HSC pool transcends a critical value, but only in the parameter regions where the d-HSCs are most sensitive to activation and deactivation.

7.1 **Using label retention as a d-HSC marker.** Kinetics of uptake and loss of BrdU within the phenotypic HSCs as determined experimentally (solid black line) overlaid with the relative proportion of d-HSCs and a-HSCs amongst BrdU$^+$ phenotypic HSCs estimated by our ODE model (dark grey shaded curve). The populations that can be found at any time point of an LRC experiment amongst phenotypic HSCs are indicated in the light grey box. The time points of chase at which the d-HSCs represent 75% and 92.5% of the LRC phenotypic HSCs are indicated by solid black circles.

B.1 **Job submission** (adapted from Lingrand et al. (2009)). **A:** gLite Workload Management System (WMS) architecture: The WMProxy enables a connection to the Workload Manager which assigns a submitted job to the target Computing Element. This target site is determined with information received from the Resource Broker and its Information SuperMarket. Job submission is finalised through the Job Adapter after which the Workload Manager delegates the processing to CondorC. The Log Monitor regulates job processing and intercepts events that affect the job state machine (see B), these events are logged by the Logging and Bookkeeping service. **B:** Life cycle state machine: a graphical interpretation of the state machine that internally controls a job’s life cycle.
List of Tables

2.1 % BrdU$^+$ HSCs. Uptake (left table) and chase (right table) data from (Kiel et al., 2007, Figure 2d and 3d respectively). D: day; exp: experiment; SD: Standard deviation. ......................................................... 20

2.2 RSS values for the two versions of the Kiel et al. model and our agent-based model (homogeneous population). The best RSS for each model is indicated in bold. ................................................................. 36

2.3 RSS values for the agent-based model (heterogeneous population). Comparing the single phase to the bi-phase version. The best RSS values for each model is indicated in bold. ....................................................... 40

3.1 % BrdU$^+$ HSCs (uptake). Uptake data from (Wilson et al., 2008, Figure 7C). H: Hour; SD: Standard deviation. ................................................................. 47

3.2 % BrdU$^+$ HSCs (chase). Chase data from (Wilson et al., 2008, Figure 2D). D: Day; SD: Standard deviation. Day 0 corresponds to Hour 312 of Table 3.1. ................................................................. 48

3.3 BrdU detection threshold. BDT: BrdU detection threshold (minimum number of labelled DNA strands that can be detected). For each BDT we divide cells into two groups during the chase phase ($c_1$ and $c_2$) depending on how many times a cell has divided during BrdU uptake. We then give $n_{c_1}$ and $n_{c_2}$, the number of times cells in each group has to divide to go below the detection threshold. Consider a BDT of 4 strands for example: cells that have divided once or twice during uptake will dilute their BrdU in 4 divisions, whilst it will take 5 divisions for cells that have divided 3 or more times during uptake. This table was compiled using Figure 3.4. .................. 56
3.4 **Goodness-of-fit results and parameter estimates.** BDT: BrdU detection threshold, DP: Dormant proportion, RSS: Residual Sum of Squares, MSE: Mean Squared Error, superscripts indicate modelling stage 1, 2 or 3. The DP of 0% indicates the one-population model. Parameter values $\sigma_d$, $\delta_d$ and $\sigma_a$ are inverted to units of days. The best overall parameter set is highlighted in bold. Estimates for $\frac{1}{\sigma_d}$ are always equal to the lower bound $\frac{1}{24\times0.0002}$ which simply indicate that d-HSCs are practically not cycling during the first 24h.

5.1 **Parameter ranges and step sizes.** DP: Dormant proportion. OR: optimisation region. $\Delta$: step size. $n$ is the number of parameters considered. Total gives the total number of simulations required for each OR and is calculated as follows: $n_{\sigma_d} \times 3 \times n_{\sigma_a} \times \#\text{realisations} \times \#\text{BDTs}$. Optimisation regions 1 and 2 are given in a scale of hours while region 3 is given in a scale of days. The values for $\delta_d$ are not shown, since in all cases it iterated through one of three values: either 1.02, 1.33, or 2 times the value of $\sigma_d$.

5.2 **Processing times - phase 1.** DP: Dormant proportion, TST: Total submission time, TRT: Total running time. During phase 1 we evaluated parameter sets for the first 24 hours of BrdU administration. This table summarises the execution results with regard to the processing times on the Grid (hh:mm:ss).

5.3 **Processing times - phase 2.** BDT: BrdU detection threshold (in number of DNA strands), TST: Total submission time, TRT: Total running time. These are the submission and computation statistics for the twenty BDT values of the 35% DP DAG job. The BDT values were traversed using a parametric job and varying the _PARAM_ variable.

5.4 **Processing times - phase 3.** DP: Dormant proportion, TST: Total submission time, TRT: Total running time. These are the execution statistics of phase 3 which calculated the best BDT parameter sets for all 10 main DAG jobs. Here it is clearly demonstrated why the Grid is more suitable for resource hungry jobs since the submission time far exceeds the actual computation time.
5.5 **Goodness-of-fit results for the agent-based model.** DP: Dormant proportion, BDT: BrdU detection threshold, RSS: Residual Sum of Squares, MSE: Mean Squared Error, superscripts indicate modelling region 1, 2 or 3. Parameter values $\sigma_d$, $\delta_d$ and $\sigma_a$ are inverted to units of days. The best overall parameter set is highlighted in bold. The *40% DP row shows the prediction of the agent-based model using parameters optimised with the ODE model in Chapter 3.

6.1 **Parameter values used in hematopoiesis simulations.** These values were used for the plots in Figures 6.6 - 6.10 and although they are synthetic they do take some of the current understanding of the hematopoietic system into account (Bryder et al., 2006). For example the self-renew rate for the transit amplifying cells is larger than that of the stem cells. Similarly the death rate of the mature cells are significantly higher than that of the A or T population.

6.2 **Elasticities for the linear feedback models.** $H_{AC_j}$ is a normalised measure of how A respond to changes in $C_j$ with higher values indicating a more rapid adjustment to steady state. For both $H_{AA}$ and $H_{AZ}$ the d-HSC model has a higher value than the conventional model. $H_{AT} = 0$ for both models since neither has any T dependant parameters.

6.3 **Elasticities for the saturating feedback models.** The same pattern as in Table 6.2 can be observed here for the saturating feedback models, namely that the d-HSC $H_{AA}$ and $H_{AZ}$ are larger than that of the conventional model.
Bibliography


