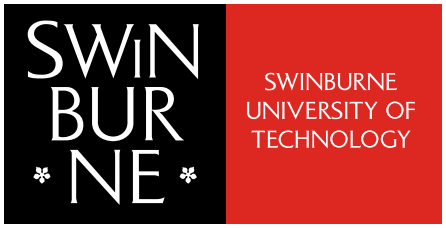
Politecnico di Torino

Andrea Stocchi – s301886

Corso di Laurea in Ingegneria Matematica

A.A. 2023/2024 - Settembre 2024

Politecnico di Torino



Understanding and modelling T-cells reproduction in vitro

Relatori:

Prof. DELITALA MARCELLO EDOARDO

Prof. FRASCOLI FEDERICO

Abstract

This thesis investigates the dynamics of T-cell populations during clonal expansion, utilizing mathematical models to replicate and analyze these processes. By focusing on single-cell resolution data from various cell families, the study aims to uncover heterogeneity in cell behavior. Comparisons among different approaches -Cyton, ODE and stochastic- highlight the Cyton model's superior adaptability in capturing complex dynamics. The work emphasises the need for nuanced models incorporating biological rules to better represent immune system responses.

Contents

1 Literature Review **5**

1.1 Biological problem 5

1.2 Mathematical models in cellular-scale immunology 18

2 The experiment, its material and methods **22**

2.1 Experimental setting and data collection 22

2.2 Available data, its structure and characteristics 28

3 Introduction to the Cyton Model **33**

3.1 The equations of the model 34

3.2 Further considerations about the model 38

4 Analysis of results **43**

4.1 Fitting the model 43

4.1.1 Optimising the Cyton model for our experimental data 44

4.2 Results from the fit 54

5 Comparison with ODE model **65**

6 A Stochastic model **73**

6.1 Results 74

7 Conclusions and open questions **80**

References **84**

A Plot of single generation of fitted Cyton model on benchmark data 90

B Plot of single generation of fitted Cyton model on original data 93

C Plot of single generation of fitted ODE model on benchmark data 98

D Plot of single generation of fitted ODE model on original data 101

E Example of fitting pipeline 105

F Code repository 109

Introduction

This work focuses on analyzing data regarding the growth of an immune system cell population (T-cells) in response to the detection of a pathogen and on identifying a mathematical model that describes and synthetically reproduces the process. The immune system comprises various type of cells, among which T and B lymphocytes are primarily responsible for the adaptive immune response, the component that develops during an individual's lifetime. Antigen-presenting cells, as their name suggests, are instead those classes of cells that present the antigen of the detected pathogen to the naïve immune cell. An antigen is any molecule capable of binding to an antibody and being recognized by the immune system as either self, tolerated, or non-self, meaning foreign and potentially dangerous, thus triggering an immune response. This triggers a cascade of metabolic events in the T cell, leading it to divide and reproduce, creating a population of specialized cells. These continue to reproduce until the infection is eradicated, after which the T cell population regresses, leaving behind only a set of modified cells known as memory T cells, capable of responding quickly to a new infection by the same pathogen. This phenomenon has often been studied in laboratory and "in vivo" over the years at the population level, characterising the total or average response of the cells. After a series of advancements in research, recent years have seen a focus on analyzing behavior at the single-cell level, significantly increasing the resolution of the study in terms of depth and precision. Despite the principles of this process being well known for some time, the description of the heterogeneity of dynamics among different cell pedigrees remains to be investigated. There-

fore, we analysed proprietary data related to clonal expansion at the single-cell resolution for 16 different "families" of T cells.

Additionally, understanding the characteristics that cells inherit from their predecessors is key to determining if it is possible to predict in advance, prior to the clonal phase, the effectiveness of a certain cell pedigree in combating an infection. For example, one possible application could be the ability to "preselect" which cell strains are preferable for culturing in preparation for infusion into an immunodeficient patient. This is particularly relevant for current employed immunotherapies in oncology.

This work was conducted in collaboration with the cell biology research group at Swinburne University of Technology based in Melbourne, Australia, employing original data from a research project by Kajal Zibaei, as part of her PhD thesis in Biochemistry at the same university.

Chapter 1

Literature Review

We will present a brief discussion of the biological process in action, recalling the concept of activation of naïve T cells and clonal expansion and introducing the main theme of heterogeneity between siblings' response and diversification of cells across generations. The discussion about the biological aspects of the problem is inspired by the work of Kajal Zibaei [47].

1.1 Biological problem

T cells play a pivotal role in the cell-mediated immune response, recognising and eliminating a variety of pathogens based on the antigens presented on their surface [16]. T cells originate from hematopoietic cells in bone marrow as non-specialized cells: after migration in the thymus, the cells undergo a complex process called "maturation" following biochemical pathways and become specialised in targeting a particular pathogen. From that moment every T cells has on its own surface more copies of a "unique" receptor (T cell receptor, TCR) capable of attaching to a very limited set of antigens. In general each cell has only one type of TCR, slightly different from that of other cells.

The T lymphocytes' process of differentiation is determined by the expression of certain membrane proteins among which the most important ones are

CD4 and CD8 [40]. In the first step, cells do not present on their surface neither CD4 and CD8 proteins and for this reason are call Double Negative [40],whereas at the end of the maturation process the cells are divided in two groups, namely the T-helper (showing on their surface CD4) and T-killer (showing CD8) [16]. When cells are mature, the variety of T-cell receptors (TCR) in the resulting population represent the fundamental factor for having an immune response able to counteract a broad range of pathogens (with associated antigens' variability) [36]. The specificity of the receptor exposed by each mature T cell implies a narrowness of the response, since a single cell can provide a response to only limited variations of antigen in diverse setting of infections [25]. To react in presence of new antigens, the immune system uses naïve T cells able to adapt to the new pathogen in the bloodstream, giving start to an adaptative response.

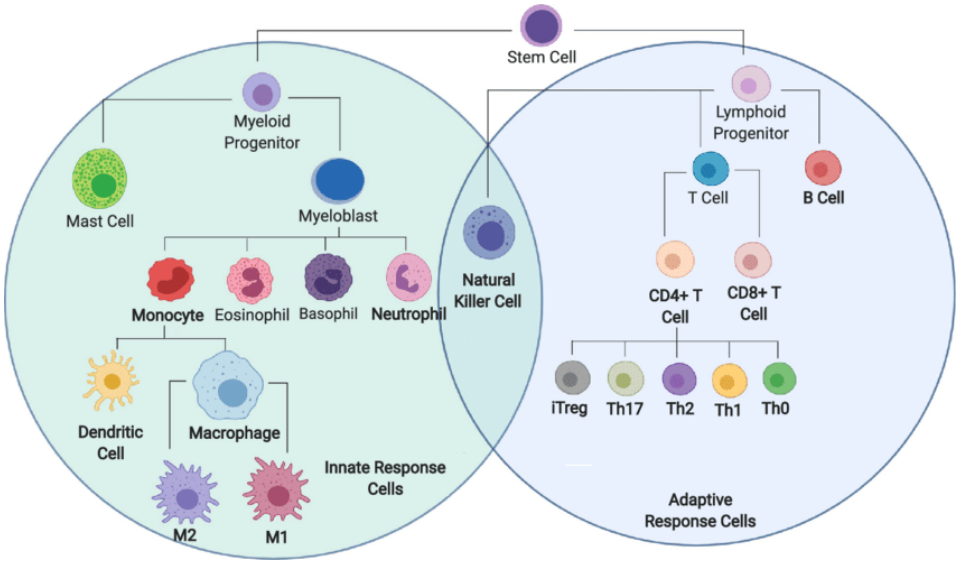


Figure 1.1: The image shows the complex variety of cells of the immune system. Image taken from Ref. [43]

In the human body, mature naïve CD8+ T cells migrate to lymphnodes to search for their cognate antigen [18]. Once naïve CD8+ T cells encounter their cognate antigen, presented to them by antigen presenting cells (APCs),

an activation process starts, necessary for priming the adaptive immune response [18]. APCs can be divided in two types, professional and "somatic cells": almost every cell can display antigens via the major histocompatibility complex (MHC) but some cells are completely dedicated to this task. Within the professional APC classes we can find macrophages, B cells and dendritic cells (DC), with the latter being the most common ones. DCs are the cells used in this particular experiment [2], which is at the centre of this work.
The interface that forms between APCs and naïve T cell during the activation process is called the "immunological synapse" (IS). The IS is a platform that, through stabilised adhesion, permits TCR signaling activation and signal termination.
The MHC is a large locus on DNA that codes for cell surface proteins, essential for the adaptive immune system. These cell surface proteins are called MHC molecules. The role of these proteins is to permit the recognition of self-cells through the binding to a self-antigen, such cells will not be targeted by the immune system. On the contrary, when a non-self cell is found an immune response is triggered, since they are considered harmful. MHC is at the heart of compatibily in organs trasplants (where the name comes from) and influences susceptibility to autoimmune diseases.

When T cells encounter a pathogen, they enter an initial phase of rapid expansion, resulting in a large population of progeny cells.

This phase is called clonal expansion and is characterised by the repeated duplication of cells, starting from naïve T cells. These cells are considered the origin of the population and are here labeled as generation 0. Each time a cell divides, it is considered no longer alive, counting as -1 to the number of cells in the generation to which it belongs. The two daughter cells are considered newborns and each count as +1 in the generation following the mother's.

Once the pathogen is cleared, a contraction phase occurs, reducing the

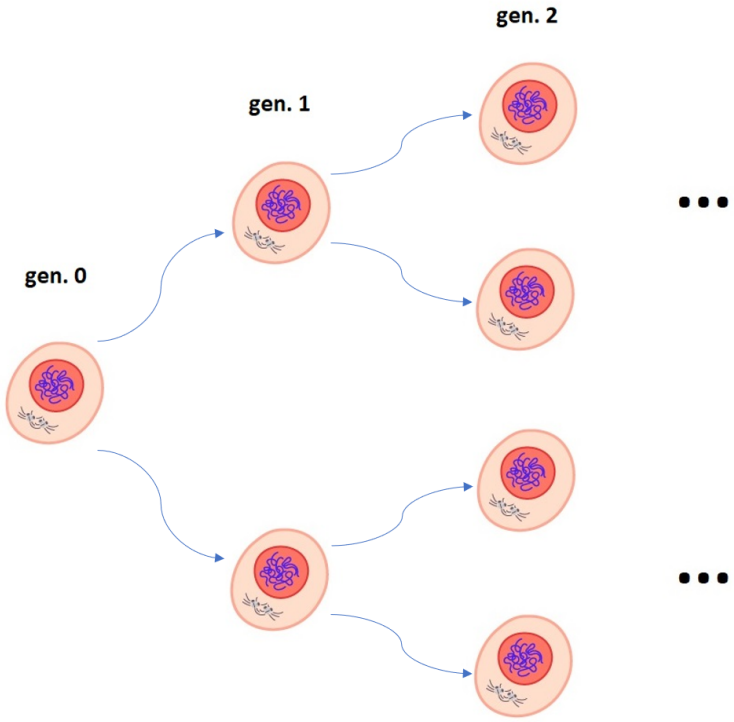


Figure 1.2: Example sketch of a prototypical cells subdivision process and counting through generations

cell count by 90-95%, employing immunosuppressor cells such as regulatory T cells ($T_{\texttt{reg}}$). The remaining cells are now only a small population of long-lived memory cells [41]. These memory cells can persist in the body for several years and be quickly reactivated upon re-exposure to the same antigen. This reactivation happens faster in memory cells compared to naïve cells. Memory cells are diverse, including central memory T cells (TCM), reside in lymphnodes and can proliferate, but do not have immediate effector functions. On the contrary, effector memory T cells (TEM), which are found in peripheral tissues, can immediately respond upon re-exposure to a specific antigen.

The clonal heterogeneity in T cell subtypes is a hallmark of adaptive immunity. T cells tend to exhibit varied functional and phenotypical characteris-

tics, allowing them to evolve into effector T cells or differentiate into different memory cell subtypes. Although antigen-specific T cells may display similar responses at the population level, individual cells show significant variability in their proliferation capacity and differentiation patterns [6]. Understanding how the population balances effector and memory T cells to ensure effective pathogen clearance and the availability of adequate memory cells for secondary responses remains a critical question in immunology. Moreover, the relationship between effector and memory T cells and the origin of memory T cell differentiation continues to be a subject of extensive research [1, 6].

Understanding the fate of division and differentiation is central and two main hypotheses were proposed regarding the development of memory cells from naïve cells. One theory suggested that a single naïve cell could become either an effector or a memory T cell. Another theory proposed that a single naïve cell could give rise to both effector and memory cells. Studies favoring the first scenario suggested that individual naïve CD8+ T cells are predetermined to become either effector or memory T cells, with some researchers attributing this determination to the initial priming by antigen-presenting cells (APCs) [21] and the timing of receiving signals during an infection. However, more recent research, including single-cell adoptive transfer experiments by Stemberger et al., has demonstrated that a single naïve cell can indeed differentiate into both effector and memory T cells [41]. Tagging individual naïve cells allowed for high-throughput analysis of naïve CD8+ T cells and, following their progeny upon activation [38], researchers have been able to track cellular lineage and fate, confirming that a single naïve T cell can produce both short-lived effector and long-lived memory T cells.

These findings support the "single cell, multiple fate" scenario, though different mechanisms for memory differentiation have been proposed [21, 42]. One model, the "*linear differentiation*" model, suggests instead that memory

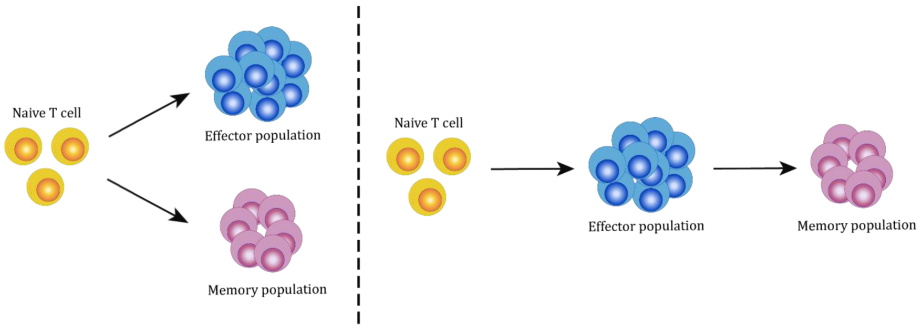


Figure 1.3: The two basic theories on how T cells differentiate in memory cells. On the left the single-cell, single-fate hypothesis; on the right the single-cell, multiple-fate hypothesis.

cells are direct progeny of effector cells and that differentiation into memory cells occurs only after the antigen is cleared or significantly reduced. This model posits that higher TCR affinity might allow some cells to escape apoptosis (i.e. death) and differentiate into memory cells. Another model, the "*progressive differentiation*" model [28], proposes that naïve T cells acquire proliferative, survival, and cytotoxic capabilities based on the strength of the signals they receive (the cytotoxicy, is the ability to harm another cell potentially provoking its lysis, e.g. its destruction). The stochastic nature of T cell-APC interactions leads to varied signal strengths among individual T cells, resulting in progeny with different differentiation potentials. According to this model, stronger signals produce progeny better equipped to survive and differentiate into TCM or TEM [28, 27]. In contrast, another set of studies suggests that prolonged or intermediate signaling is unnecessary for initiating clonal expansion and differentiation [21]. Once naïve T cells receive the initial signal, they and their progeny follow a preprogrammed differentiation path, independent of subsequent signal strength or duration. For example, Van Stipdonk et al. demonstrated that naïve T cells could commit to clonal expansion and differentiation even with brief stimulation in vitro, though longer stimulation was required in vivo for optimal response [42].

A third model, the "*decreasing potential model*", finally suggests that re-

peated stimulation diminishes effector cell functionality, making them more prone to apoptosis. Consequently, the death of effector cells reduces the number of developed memory cells [22]. This model aligns with observations that prolonged inflammatory signals or sustained activation can adversely affect effector cell differentiation. Although both the "progressive differentiation" and "decreasing potential" models agree on the "one naïve T cell, multiple fate" scenario, they propose different mechanisms for effector and memory differentiation.

Despite these various models, the exact factors influencing T cell fate decisions remain unclear. It is not yet fully understood how and when cells diverge in their differentiation paths. Single-cell studies hold promise for elucidating these mechanisms, providing insights into the timing and determinants of effector and memory T cell differentiation.

In this regard, one proposed mechanism for heterogeneity among generations is asymmetric cell division, as suggested by Chang et al. During the first mitotic division of CD8+ and CD4+ T cells, signaling molecules co-localise with the microtubule-organizing center (MTOC), maintaining this localisation and potentially influencing the fate of the daughter cells [9]. Asymmetric division may thus be a critical factor in generating inter-clonal heterogeneity, with implications for the functional diversity observed among T cell populations.

The phenomenon of unequal partitioning during mitosis has been closely observed in T cells, with researchers noting the importance of prolonged contact between T cells and antigen-presenting cells (APCs) in ensuring asymmetric cell division (ACD). This interaction is critical for the development of daughter cells with distinct phenotypic markers and functionalities, specifically effector and memory cells.

Similarly, Lin and colleagues proposed that the differential inheritance of

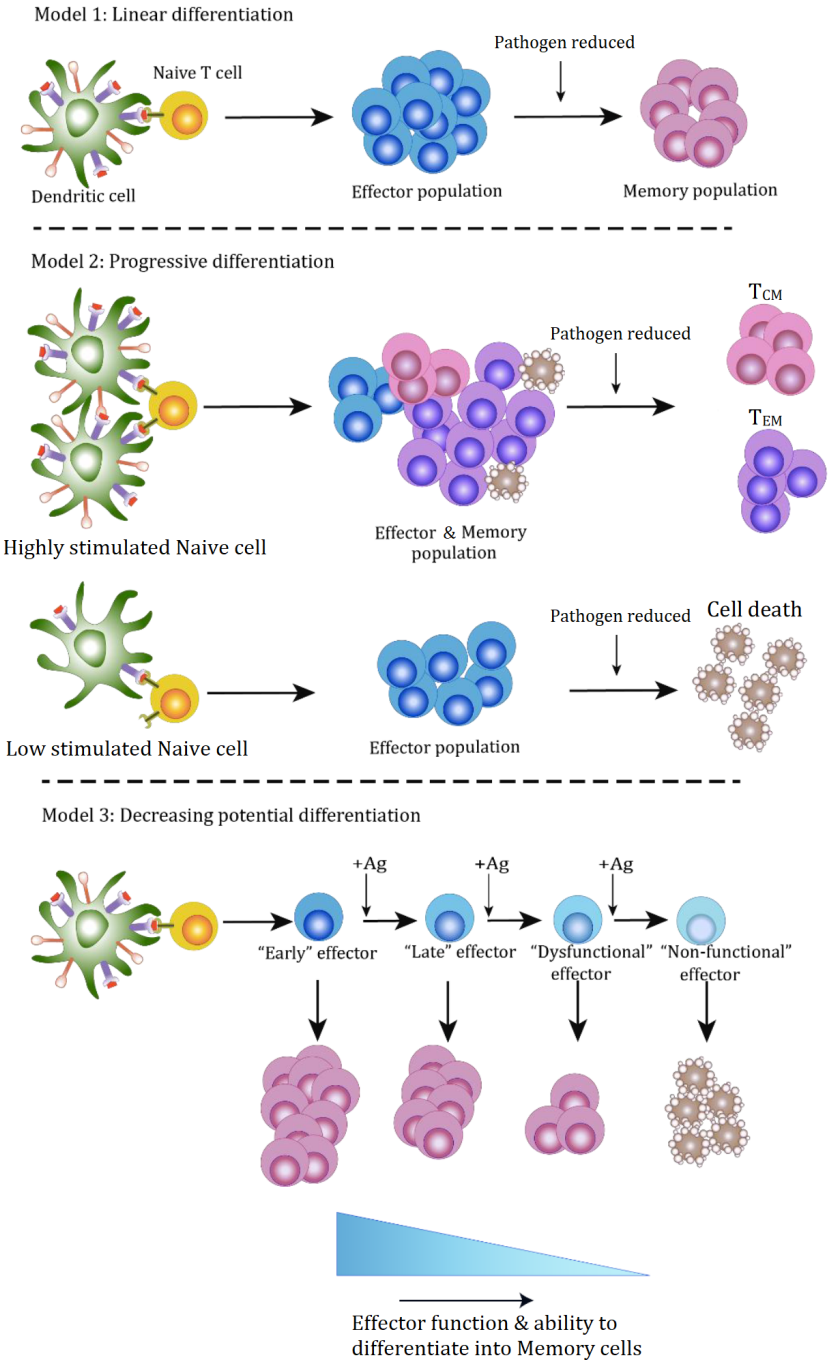


Figure 1.4: The three models on how once that the immune response is triggered the population differentiate into effector and memory T cells. All three models support the hypothesis of single-cell, multiple-fate. Image modified from Ref. [47]

proliferation-promoting factors during the first division results in the creation of two distinct cell populations. These populations, over the first three generations, diverge into more activated or more quiescent subsets, eventually forming the effector and memory cell groups, respectively [32]. Despite evidence supporting ACD's role in lymphocyte fate determination, its significance remains a topic of debate.

An alternative hypothesis suggests instead that the functional fate of individual T cells is predetermined prior to their first division. For instance, in vivo tracking of clonal progeny of CD8+ T cells indicated that the intra-clonal diversity might stem from variations in signaling during the interaction between naïve T cells and dendritic cells (DCs). This implies that the functionality of CD8+ T cells could be established before the initial cell division [5]. Alternatively, another model posits that T cell differentiation begins before the first division and is amplified during subsequent divisions. This was supported by a study from Lemaitre et al. [30], which involved isolating naïve CD8+ T cells in micro-arrays and stimulating them with DCs. The analysis of progeny from wells containing a single T cell revealed significant size disparities within and between clones. Further examination of phenotype diversity, particularly the expression levels of CD25 and CD62L, suggested that this diversity was independent of family size [8]. These findings support the idea that clonal differentiation patterns are largely unaffected by stimulation conditions. Comparisons of progeny from two daughter cells after the first division also indicated that activation-induced division could be a source of intra-clonal diversity [30].

These studies have illuminated various aspects of clonal diversification in T cells, highlighting how different factors contribute to heterogeneity. Activation signals received by naïve T cells from APCs during activation and additional stimulation during amplification, along with asymmetric partitioning of signaling molecules during the first division, can all influence clonal disparity.

However, the specific roles of individual cells and their impact on the overall immune response remain incompletely understood. While population-level studies have provided significant insights into cell differentiation and proliferation during immune responses, many questions can only be addressed through single-cell analysis. The ability to track and study individual cells offers unique insights into their developmental processes.

Recent advances in experimental technologies have indeed made single-cell analysis feasible both in vivo and in vitro. Although studies at the single-cell level have been highly informative, they are less common due to their labor-intensive and time-consuming nature. Tracking individual cells during activation and amplification could clarify many aspects of the immune response, including the development of effector and memory cells. In recent years, researchers have employed single-cell approaches such as single-cell transfer to trace the lineage of individual cells. However, these techniques often lack information about the relationships between cells within individual clones, which is crucial for evaluating hypotheses about lymphocyte fate determination.
Furthermore, intra-clonal disparity and burst size have been significant areas of research. Studies have explored the patterns of differentiation and phenotype heterogeneity within clones. The activation signals and environmental factors influencing these processes are complex and multifaceted. The diverse outcomes observed in clonal populations highlight the importance of both intrinsic cellular mechanisms and extrinsic environmental influences. Single-cell studies, despite their challenges, have the potential to unravel the intricate dynamics of T cell differentiation and proliferation, providing deeper insights into immune responses and the development of cellular immunity.

The debate over the importance of ACD versus pre-determined functionality prior to the first division highlights the complexity of lymphocyte fate determination. Single-cell analyses offer promising avenues for understanding these processes at a finer resolution, although they come with significant

methodological challenges. Advances in single-cell technologies continue to push the boundaries of our knowledge, offering new perspectives on immune cell development and function. These insights are critical for developing more effective immunotherapies and vaccines by leveraging our understanding of T cell biology.

As already mentioned, the numerical output of individual T cells exhibits a marked disparity from one clone to another [15, 30]. The importance of heterogeneity is exemplified by the fact that the number of progeny generated by an original cell, or the number of division cycles individual lymphocytes undergo before ceasing to proliferate, is a key determinant of the strength of an immune response. Numerous studies have explored how various intrinsic and extrinsic signals received by a cell influence the programming of immune response outcomes. For instance, it has been shown that individual naïve T cells can produce a highly variable number of offspring, regardless of the stimulation conditions [15, 30]. It is interesting to note that despite expectations, a delay in activation during infection would not affect differentiation [41, 11]: indeed in vivo lineage tracking of barcoded naïve CD8+ T cells revealed that both early and late activated T cells produce a heterogeneous number of progeny [15].

Both B cells and T cells have been subjects of quantitative studies examining their proliferation upon stimulation and this will be one of the focuses of the mathematical investigations undertaken in this Thesis. Two critical factors, the concentration of the stimulus and the duration of stimulation, were shown to impact population growth. For varying concentrations of stimuli and stimulation durations, a similar growth profile was observed: following priming, cell numbers increased until a point where the population stopped growing and entered a contraction phase. The main difference was that higher concentrations of stimuli and longer priming times resulted in more cells being produced. However, it was noted that there is an intrinsic limit to the num-

ber of offspring each B cell can produce, which cannot be altered by changing stimulation conditions [44, 19]. This hypothesis was further supported by experiments that interrupted expansion for a short period following stimulation. Upon resuming previous conditions, cells started expanding again, but those returned to the media only proliferated less compared to those that continued receiving stimuli. Restimulating the cells after interruption caused them to expand again and reach the maximum cell number, though at a later time point compared to non-interrupted cultures [44].

A key conclusion from these studies is that the intrinsic mechanism controlling cell proliferation is likely division-based rather than time-based, meaning there is a limit to the number of progeny each cell can produce rather than a time period within which a cell can proliferate. Further studies on B cells also indicated that the differentiation process could reprogram the division destiny of cells [44].

Population-based in vitro studies of T cell expansion showed that cell proliferation starts immediately upon activation, followed by a gradual increase in the number of quiescent cells, which reaches nearly 75% by day seven when expansion ceases. In vivo, the expansion profile followed a similar pattern, with the main difference being the division destiny and subsequent lifespan of the cells: lower division destiny and shorter lifespan were observed in vitro compared to in vivo [34]. That study also suggested that early signals received by T cells program their capacity to produce offspring, while later during cell division, extrinsic signals can modify this intrinsic feature [34].

Under different stimulation conditions, lymphocytes generate heterogeneous immune responses. Therefore, other deterministic factors likely influence the varying capacity of clones to produce different progeny. The population-based analysis of proliferation patterns of T cells and B cells revealed common results indicating intra-clonal diversification of burst size (i.e. the total number of cells produced during the clonal expansion phase). However, the regulation of disparity between families remains unclear: some hypotheses have

been proposed about how different signals or even intrinsic characteristics of the founder cell before priming can contribute to the outcome of an immune response. Tracking individual cell fate and analyzing its expansion profile is necessary to understand the role of different contributing factors in the final response profile.

In this regard, a combination of multiple division tracking dyes has enabled researchers to track multiple founder cells' families while they expand in a single culture environment [35]. This research suggested that the division capacity of naïve CD8+ T cells is programmed at the beginning of the immune response and that the division destiny of clones is highly similar in response to different stimuli [35]. Notably, more than half of the progenies are produced by a small fraction of founder cells, indicating that tracking the fate of these progenitors [7] over a long time can elucidate the determinant factors controlling the response.

T cells tend to form clusters during their activation and interaction with APCs, making it challenging to follow individual cells over extended periods. However, recent advances in microscopy techniques, particularly time-lapse imaging and the development of platforms that facilitate isolating single cells with minimal disturbance, have made it possible to monitor individual primary lymphocytes during activation and subsequent clonal expansion. Quantitative analysis of survival features like time to divide, time to die, and division destiny of individual families have been modeled using the Cyton model [19], which is one of the major mathematical focus of the present work. This model allowed to put data from many single-cells experiments into perspective. For instance, consistently with previous observations, the time to the first division was longer than subsequent division rounds, and variation in division times was independent of the generation number (or equivalently, the division rounds the cell underwent).

The observation that death data did not fit an exponential model led authors to suggest that an internal age-dependent mechanism regulates lympho-

cyte death. Cells that died showed no signs of growth (or very little growth) upon their last division, indicating that they lacked the capacity for further division [20]. Interestingly, there was significant concordance in division destiny within individual families, meaning most descendants died in the same or close round of division, suggesting descendants might inherit this feature from their founder cell. Specifically, the size of founder cells before the first division was highly correlated with the division destiny of its descendants. Siblings also showed high correlation in their time to division and death, reinforcing the idea that some features of founder cells are inherited by their progeny.

Overall, it is now established in experimental immunology that quantitative information and advance microscopy can reveal aspects of immune response that cannot be discerned otherwise. This is where mathematical biology shows its value.

1.2 Mathematical models in cellular-scale immunology

The behaviour of naïve T cells in response to encountering a pathogen/antigen shows a wide variability, particularly in the size of the progeny generated, since it can vary by several orders of magnitude. This variability has not been experimentally clarified, but it can be attributed to the heterogeneity of the naïve cell population and the contribution of the external stimuli, which vary in terms of duration and intensity, as well as in the way they are passed on to descendants. As mentioned, asymmetric cell division could be another element contributing to the variability of clonal expansion.

To delve deeper into these phenomena, population studies and in-vitro experiments have typically been conducted, but that approach loses resolution at the single-cell level and misses the temporal resolution of individual events. For this reason, an individual analysis approach of clone families was chosen

for the present work. This allowed for the exact quantification of the number of cells over time and the evaluation of birth and death times, establishing a genealogical relationship between the cells participating in the experiment. The total response variability was investigated starting from the single mother cells, providing insight into clonal heterogeneity and how this characterises the overall immune response.

The key goal from a quantitative perspective is a mathematical model that describes how the lifespan of a single cell can influence the advanced phase of population expansion, also trying to predict how the duplication rate varies over time and among families derived from different naïve cells.

In general, the reader should appreciate that there are various types of models that describe the dynamics of the immune system at different scales, starting from the molecular scale, attempting to model the biochemical regulatory cycles of cells, up to tissue-level or multiscale models, as extensively described in the Review by Eftimie et al. [14]. We will only focus on the mathematical models used to describe and study growth, reproduction and movement of cells taking part in immune response, at the cellular level.

The majority of the models developed to study different aspects of T lymphocyte and immune system dynamics make extensive use of ordinary differential equations (ODEs) and concentrate on quantifying T and B cell turnover, T cell movement, differentiation patterns, immune responses to infections and cancer (Luo et al. [33], Bains et al. [4], Boer and Perelson [13], Reynolds et al. [37], Ankomah and Levin [3]). Despite ODEs systems can be easily fitted to experimental data (thus avoiding overfitting issue due to too many parameters), there exist also more complex approaches for cell-level dynamics, which are used for theoretical investigation of various aspects of the immune system. Among them, there are delay differential equations (DDEs) models to account for time delays in immune responses, as seen in the work by Lee et al. [29] and Kim et al. [24], probabilistic and stochastic models like those

by Davis and Adler [12] and Choo and Murali-Krishna [10], which instead describe cell proliferation and death probabilities or analyze memory CD8 T cell population dynamics.

The complexity of these models varies from simple equations to extensive frameworks for exploring non-trivial behavior.

Model Type	**Description**	**Example Studies**
ODE (Ordinary Differential Equations)	Used for quantifying cell turnover, movement, differentiation patterns, and immune responses to infections, autoimmune diseases, and cancer.	Bains et al. 2009; Boer and Perelson 2013; Ankomah and Levin 2014; Reynolds et al. 2013
DDE (Delay Differential Equations)	Account for time delays in immune responses, such as the delay between viral infection and immune response, or between initial T cell stimulation and full activation.	Lee et al. 2009; Kim et al. 2007
Stochastic Models	Analyze the random nature of processes, such as the proliferation and maintenance of memory T cell populations. These models often incorporate randomness to describe biological variability.	Choo and Murali-Krishna 2010
Probabilistic Models	Describe probabilities of different biological events, like cell proliferation and death, to capture the variability and uncertainty in biological processes.	Davis and Adler 2013

Table 1.1: A selection of typical examples of mathematical models used to study the immune system. Note that this is a list of only few, representative models.

In particular and in regards to the problem at the heart of this Thesis,

the Cyton model was proposed by Hawkins et al. [19] in 2007 as extention of a previous work published by Gett and Hodgkin [17]. The model utilizes a system of ODEs with the use of convolution terms over time to describe a population of lymphocytes with structure, i.e., divided by generations to which each cell belongs, starting from the activation and clonal expansion phase. The Cyton model was developed with a top-down approach, aiming to model the cellular processes necessary to describe the data, thereby choosing the two fundamental processes of division and death, with the addition of a term for the reduction of activity [19].

We will discuss the Cyton model in more detail in Chapter 3, describing it comprehensively.

Chapter 2

The experiment, its material and methods

2.1 Experimental setting and data collection

As already mentioned, the data used for the analysis in this thesis were obtained by Dr. Kajal Zibaei, a doctoral student at Swinburne University of Technology from 2016-19, based on previous research work. All data discussed here are original.

The data pertain to the division and death times of primed CD8+ naïve T cells, with single-cell granularity, obtained through recordings of the wells where each cell was placed. In practice, a T cell was placed in a well, and its behaviors were then observed under a microscope. At the time of division, one of the two daughter cells was individually transferred to a new well. Repeating this procedure at each cell division allows precise determination of a "genealogical structure" of the population and measurement of the intervals between divisions.

The procedure was continued identically up to the fifth generation, when only 2 cells of the genealogical tree were selected, continuing with this procedure while all the others were kept together, losing the genealogical characterisation

and birth-division time measurements but maintaining a cell count over time.

The cells were observed and photographed according to two distinct procedures:
- The first aimed to photograph the Immunological Synapse (IS), which forms during antigen presentation by Dendritic cells to naïve T cells, from formation to signaling to termination.
- The second aimed to record the entire cell division process at the population level.

High-magnification images and images of cells in epifluorescence were captured using a "Super Resolution" Nikon STORM microscope. The Nikon STORM is a laser microscope that allows the reconstruction of the position in a 3D environment of a fluorescent molecule by combining various information, enabling the analysis of single molecules in interacting complexes like those formed in the IS. This aspect of the experiment will not be further detailed in this work, as the collected data mainly refer to a qualitative description of the observable characteristics of the cells and the IS formed, rather than the timing of differentiation processes. It is worth mentioning that laser is not equivalent to optical imaging and this can influence the biological process of the cells, which tend to have a shorter lifespan in this setting.

Regarding the data on tracking, the entire response process of naïve T cells to a pathogen, they were obtained by following the population development for a period of 2 weeks, using an inverted optical microscope combined with a confocal unit and a camera. naïve T cells were isolated from OT-1 transgenic mice; these cells expressed a fluorescent protein useful for tracking their position over time. The antigen-presenting cells (APCs), in this case, dendritic cells (DCs), were extracted from the bone marrow of C57BI/6 mice, then brought into contact with the antigens and placed in microwells. The

next day, naïve T cells were added to the wells, averaging one cell per well. Images were initially captured at 10-minute intervals.
The recording starts from the interaction between DCs and naïve T cells and continues uninterrupted until the mother cell's first division. At that point, recording is stopped to separate the daughter cells into two wells using a micromanipulator, and it is not possible to record the cytokinetic processes during this phase. From that point on, the cells were left free to duplicate up to the 4th generation, measuring the lifespan of each. Upon reaching 8 cells per well, 2 cells per well were isolated into new wells, allowing detailed tracking of two branches of the genealogical tree. The others were left in the well where they developed, providing only information on the total population quantity. From the first to the fourth generation, the interval between frames was set to 30 seconds, and after the relocation of the two cells, from the $5th$ generation onward, to 2 minutes.

Therefore, single-cell resolution data are complete up to the fifth generation, beyond which only partial analysis of a part of the progeny of the naïve T cell can be conducted. However, the quality of the total cell count data for each pedigree remains unchanged.

The analysis of the frames was conducted using TACTICs, a software developed in Matlab by Dr. Raz Shimoni at Swinburne University. The processing of the frames consists of a series of steps that analyze the previously obtained fluorescence images.
The procedure used is as follows: first, an attempt was made to reduce the noise present in the obtained images, often due to exogenous elements such as light characteristics or the sensor. Next, the background of the image was subtracted, and the contours of the foreground elements were better defined. The next phase was segmentation, where small objects were removed from the image, or holes created during filtering were filled. Commonly to similar

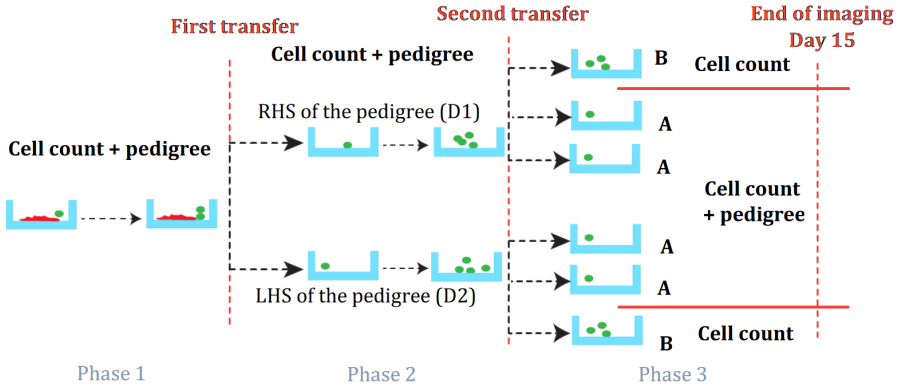


Figure 2.1: Cells subdivision and transfer in wells. Image modified from Ref. [47]

tracking procedures, one of the problems to be solved is the possible image overlap of two cells. This can be managed automatically via an algorithm or manually.

Each cell must be assigned an ID, which must be maintained in consecutive frames. The crucial point of this phase is the "assignment problem"; the algorithm used to assign the ID to the cell is based on recognizing a minimal distance traveled by a cell after division. When the cells do not respect this distance and remain very close to each other, errors may occur, causing cells to change IDs. To solve this problem, a manual analysis was conducted, frame by frame, additionally labeling the frames where cells underwent death or division.

Cells that did not exhibit division or death during the data collection period were ultimately annotated as having an "unknown fate".

This entire process, which, incidentally, can still give rise to errors, allows for tracking and recognizing cell families across various generations, not without some hiccups.

The data were then cross-referenced to reconstruct family lines, namely pedigrees, starting from each individual naïve mother cell. The data from the

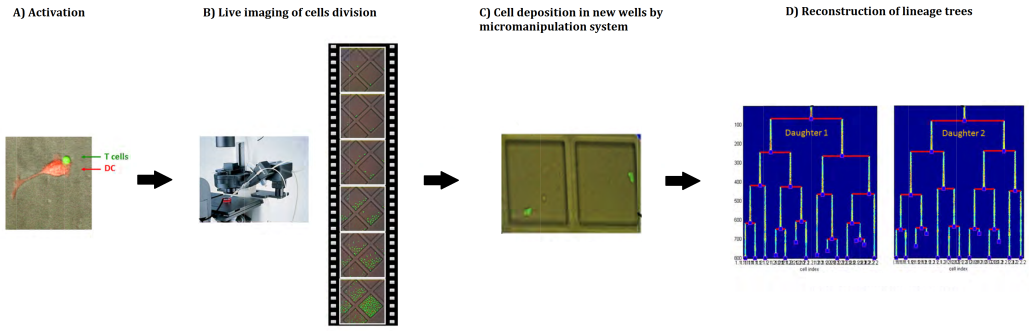


Figure 2.2: This image shows the process of data gathering from the activation of naïve T cells to the reconstruction of the pedigrees. Image from Ref. [39]

first two experimental phases, those preceding the isolation of only two cells, constitute the first four generations of the pedigree. From phase 3, only the isolated cells were incorporated into the cell's genealogical structure, while those left in culture together were considered only in the total cell count generated by that particular mother cell.

The data structures at the end of the analysis are of two types: a genealogical tree that includes morphology and division and death times for each tracked cell (thus allowing generation analysis) and a second data set that represents the total number of cells associated with that pedigree over time.
The latter measurement results from further data processing, as the frames of the two environments, the well with the set of cells in culture for counting and the wells containing the cells for genealogical analysis, were taken at different time intervals, thus necessitating synchronization of the counts to obtain a stable and reliable measure.

The activation and subsequent duplication process of naïve T cells, although known, is not yet fully understood in all its details. As mentioned, once T cells come into contact with DCs, an interface called the immunologi-

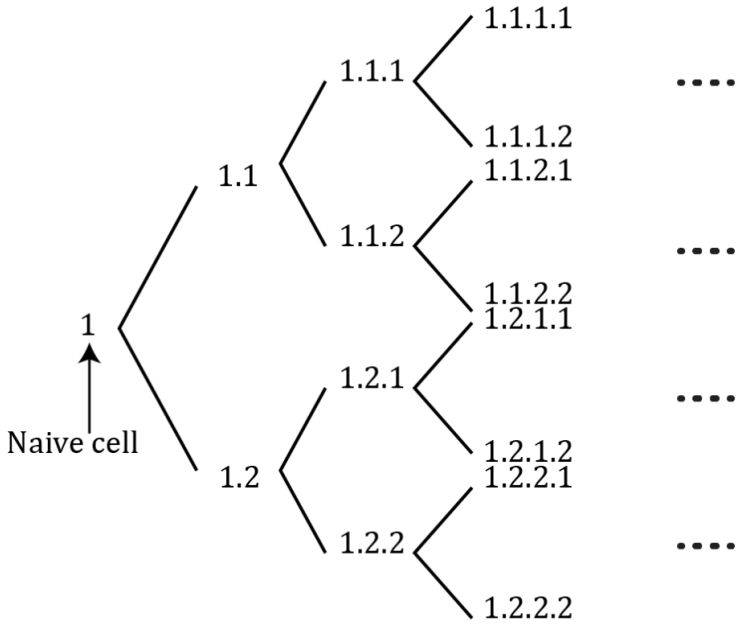


Figure 2.3: This image shows how the cells were tracked, with each naïve T cell having a different initial number and the progeny addressed accordingly. Image from Ref. [47]

cal synapse (IS) forms between them, acting as a platform for translating the chemical signal. In the T-DC conjugation, the surface molecules of both cells rearrange to stabilise the synapse. Research results suggest that the space between the cell and the synapse is a significant factor in determining the level of TCR signaling, either inhibiting or promoting it. T cell activation leads to proliferation and differentiation; one of the events generated in this phase is cell growth. Although the metabolism of activated naïve T cells is known, little is known about how the growth of such cells influences the ultimate fate of the progeny. One feature that is known to influence replication is size, since it impacts the division time of daughter cells [45]. For example, the size of founder B lymphocytes at the time of the first division appears to influence the maximum number of generations a particular family can undergo before reaching a quiescent state. Since each naïve T cell generates an entire family of clones that contributes to the immune response, knowledge about activa-

tion dynamics and the links between physical characteristics and population development can help better understand important aspects of the immune response.

2.2 Available data, its structure and characteristics

The in vitro experiments discussed in this work are designed to resemble in-vivo dynamics as closely as possible.
The possible outcomes for each cell are: division, death or survival, where the latter occurs if, at the end of the experiment (15 days), the cell is still alive but not undergoing any division or death processes, making its fate unknown.

The first aspect studied was the size of the progeny derived from a single naïve cell. It was observed that in the first few days, there is a numerical expansion due to the pure duplication of cells, until about day three or four. After this period, some families slow down in duplication and stop growing, which may be linked to a loss of synchronization in cell divisions. Most families reach a population peak between days 4 and 5, showing heterogeneity in the number of individuals per family and the time required for the population to decrease.

It is then possible to compare the progression of a single family with that of others to see how each contributes to the bulk response: it is observed that only a small fraction of the families, 5 or 6 out of 16, contribute from half to over two-thirds of the total response.

Analyzing the life spans across generations, it becomes evident that division and death rates are not static over time and between generations, but

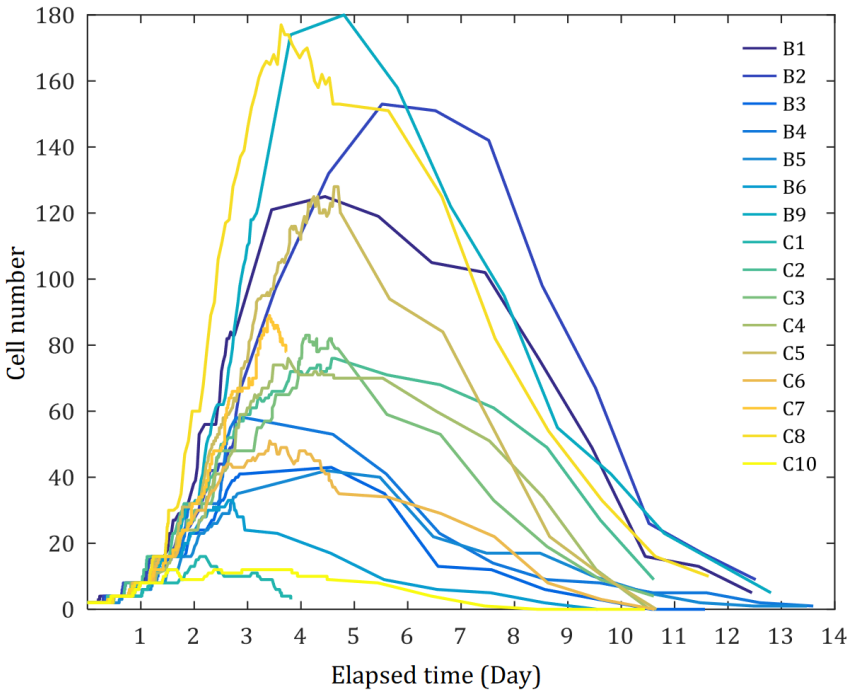


Figure 2.4: This image shows the number of cells for each pedigree during various timestamps. Every pedigree is associated to a different alphanumeric code and color and it is experimentally obtained in an independent way. Image from Ref. [47]

exhibit observable differences between families. The most notable difference is in the time required to divide, which is much longer for generation 0 cells (naïve T cells) compared to subsequent generations. For mother cells, this time ranges from one to one and a half days, while for subsequent generations, it averages half a day with right-skewness up to two days. An example of this is shown in figure 2.7. Cells in the dataset begin to die starting from generation five, and in advanced generations, there is an increase in cell lifespan leading to an unknown fate.

The anomalous behavior of these cells results in a "defect" in the data meant to describe division and death rates, as the absence of observable events makes accurate quantification impossible. The completeness of records would

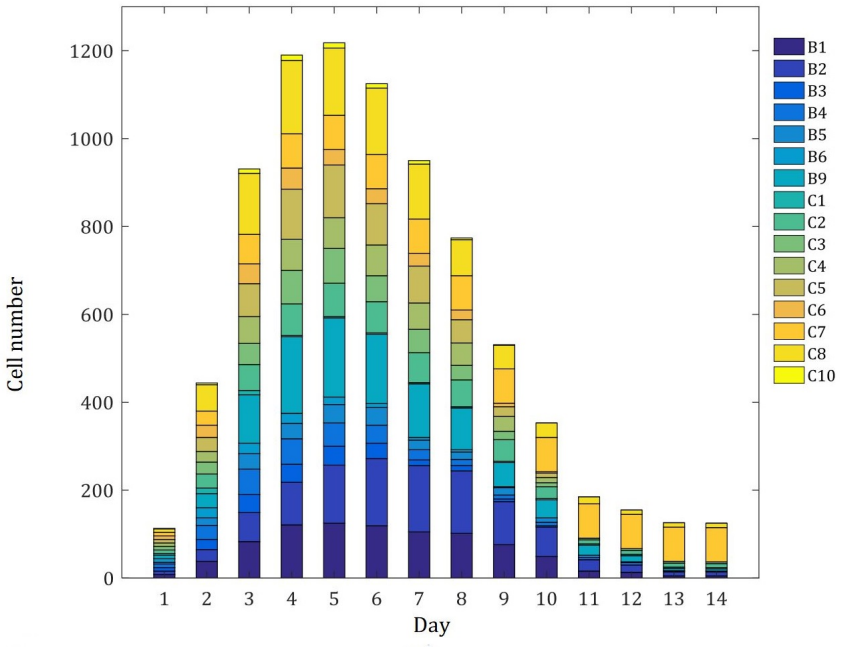


Figure 2.5: The total number of cells from each pedigree was quantified at 24 hour time intervals. Plots consist of stacked bar chart showing how the 16 clones together give rise to a cumulative expansion and contraction that represents how a bulk culture might behave. Image from Ref. [47]

be very important, in fact, the data could be highly informative for a detailed description of the biological phenomenon as a whole, potentially indicating if these cells have become quiescent or have differentiated into memory cells. Thus, the data are considered censored, meaning there is a lack of information due to no events occurring. This limits the extension of some of our results, as we will see.

Several approaches have been developed over time to handle "censored" data, with the most common being KM and CR-analysis. Unfortunately, neither is appropriate in this case as both assume homogeneous survival probability distributions over time, which is not correct here.

Also note that cells that lived for more than 18 and a half hours and exhibited anomalous behavior leading to an unknown fate were defined as "Long

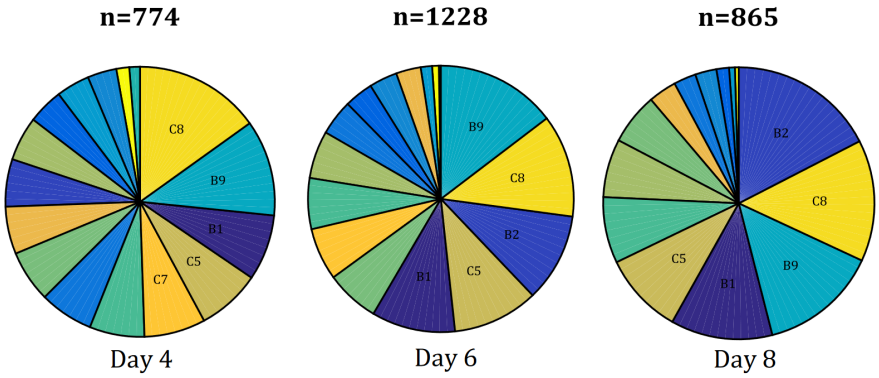


Figure 2.6: Here are the pie charts, at different times, of the relative abundance of cells divided in pedigrees. Image from Ref. [47]

Lived".

The initial phase of clonal expansion can be described by specific functional forms, such as exponential, Gompertz, or sigmoid functions, with the latter being the best at describing the process, in this case. This characterisation also enables the analysis of the expansion rate over time, showing no strong correlation between the initial expansion rate and the maximum number of cells reached during the full immune response of a pedigree. Most pedigrees instead display the same initial expansion rate, which diverges after one or two days: this highlights that no characteristic in the early dynamics is yet known to define a high-clone-producing pedigree. These considerations are expanded and discussed in Ref. [47]

Finally it is worth mentioning that the data show a very strong correlation between the lifespans of two daughter cells from the same mother, regardless of their generation. This may suggest a synchronization in cell cycle control processes established at the time of individual division, not inherited vertically through intergenerational passage, but only affecting few generations [47].
It is difficult to establish which characteristics inherited by daughter are deter-

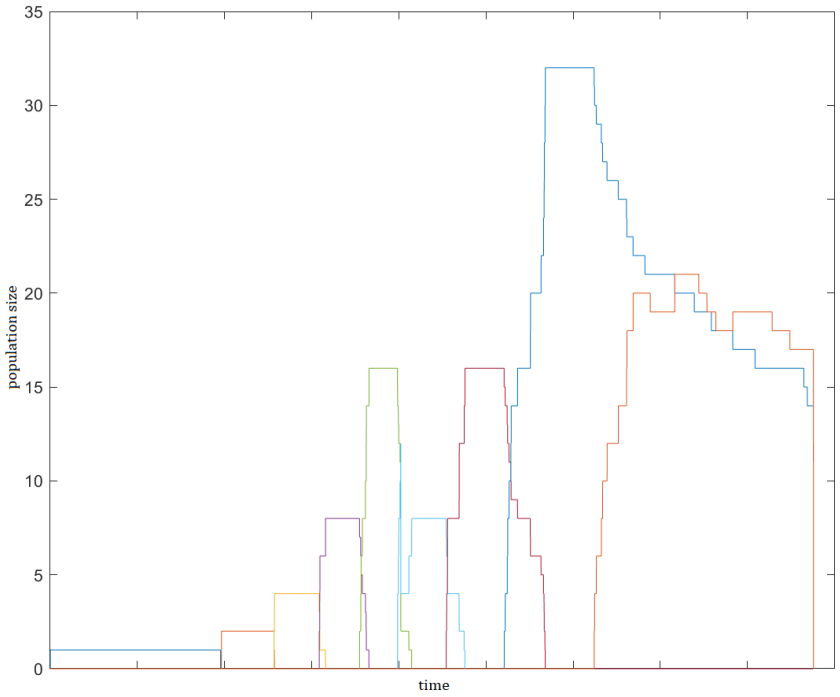


Figure 2.7: An example of plot from Kajal's data, showing the number of cells of a pedigree in each generation in time. The population develops starting from a single cell and stays still until the mother cell divides. After this slow start, dynamics become faster with narrow increase and decrease front. As said in the text not the entire population is described here and part of the data is lost, this is caused by the measurement procedure. The effects can be seen in the reduction of the cell numbers between 5th and 6th generation, this breaks a pattern of monotonous population growth through the generations.

minant of a prolific pedigree and this is why a number of refined mathematical models have been proposed. We describe the most cited and used one, the Cyton model, in the next chapter. We will then use that model to describe the data we have.

Chapter 3

Introduction to the Cyton Model

My original work initially focused on implementing the Cyton model, a model of integro-differential equations proposed in 2007 by the research group of Philip Hodgkin, a group based in Melbourne. I first studied the equations that make up the model.
The Cyton model is a mean-field model, that describes the behavior of a phenomenon related to a "large" population, devoid of the elements of stochasticity. Therefore, it is not possible to reconstruct the variability of the process itself from this model, it only provides an average description of signals. In fact, the phenomenon that we are trying to describe is one in which the average over repetitions of this phenomenon in an individual and the average over a phenomenon characterised by a very large number of participating individuals are the same. Or in other words we assumes the hypothesis that the dynamics driving the entire population are the same as those driving the single individual, and therefore the microscopic process is essentially the same as the macroscopic one.

Analyzing the Cyton model technically, it can be said that it is a structured population model. The population structure is organized into sub-populations

that reflect the generations to which each cell belongs, which is given by the number of divisions that have occurred between the naïve cell and the daughter cell we are analysing. The population that describes each generation increases depending on the divisions of the cells from the previous generation and decreases due to the effects of division and death phenomena within the generation itself.

3.1 The equations of the model

The equations of the Cyton model are as follows:

$$\begin{cases} \dot{n}_0(t) = -n_0^B(t) - n_0^D(t) \\ \dot{n}_i(t) = +2n_{i-1}^B(t) - n_i^B(t) - n_i^D(t) \qquad \text{with } i = 1, 2, ... \end{cases}$$

where the terms n_i^B and n_i^D represent, respectively, the number of cells in the i-th generation that divide and die at time t, and are defined as follows

$$n_i^B(t) = n_i(0)q_i^B(t) + \int_0^t 2n_{i-1}^B(\tau)q_i^B(t-\tau)\,d\tau$$
$$n_i^D(t) = n_i(0)q_i^D(t) + \int_0^t 2n_{i-1}^B(\tau)q_i^D(t-\tau)\,d\tau$$

both variables n_i^x are formed by two additive terms:

• The first term represents the effect of the division/death phenomenon on the initial population of cells belonging to the i-th generation.

• The second term, on the other hand, represents the effect of division/death brought by the cells that gradually divide in the previous (i-1)-th generation

and enter the current i-th generation, subsequently dividing or dying.

Let us now focus on the coefficients of the model. These are time-varying and represented by the functions $q_i^{B/D}(t)$. Each of these functions represents the instantaneous/in time rate of division (B) and death (D) for cells belonging to the i-th generation. Such rates actually have an underlying structure dependant on the probability distributions related to the birth and death processes.

$$q_i^B = \gamma_i \, p_i^B(t) \left(1 - \int_0^t p_i^D(t)\, d\tau\right) = \hat{p}_i^B(t) \; \bar{P}_i^D(t)$$
$$q_i^D = p_i^D(t) \left(1 - \gamma_i \int_0^t p_i^B(t)\, d\tau\right) = p_i^D(t) \; \bar{\hat{P}}_i^B(t)$$

The rates q_i^B therefore represent the effects of division after taking into account the effect of death, while q_i^D is the analogous for the death process. Assuming that the two processes are concurrent and independent (following the experimental results), the structure of the division rates is nothing more than the instantaneous probability/probability density function of dividing at time t, reduced by the value of the Progressor Fraction γ_i, multiplied by the cumulative probability of not having undergone the death process by time t. For the death rates, a similar calculation is performed, considering the cumulative probability of the division process and the probability density function for death.

The probability distributions $p_i^{B/D}(t)$ are defined based on the analysis of biological data obtained from experiments with lymphocytes. In this study we are concerned about the dynamics of T lymphocytes only, but the response of B cells has also been investigated deeply [46, 23]. Since they are not directly observable, as the phenomena censor each other, it is necessary to modify the behaviors of the cells ad hoc to suppress the biochemical cycles of the division

and, alternatively, death processes.

Given that these processes are right-skewed, meaning they sometimes occur with long delays compared to the population mean, asymmetric distributions such as the gamma, Weibull, or log-normal are used to describe them. The log-normal distribution has proven to be the most appropriate and has therefore been chosen as the functional form for the "basic" distributions (p_i) of the model [26, 31].

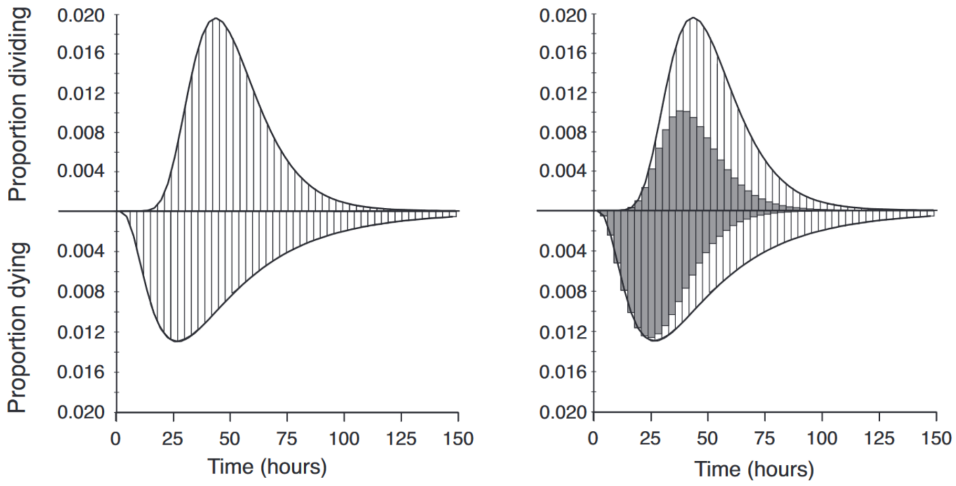


Figure 3.1: Probability density function used to model the phenomena of death and division, images from Ref. [19]

Returning to the terms $n_i^{B/D}$ of the model, the second term is a convolution integral, with the variable $2n_{i-1}^B(\tau)$, which represents the cells born at time τ, interacting with the parameter q(t) delayed by τ. This represents, for each instant of the continuous signal n_{i-1}^x, the comparison of the same with the entire probability distribution q(t), starting from the moment of "birth".

The differential equation for the first generation is then expressed as follows:

$$\dot{n}_0(t) = -n_i(0)q_i^B(t) - n_i(0)q_i^D(t) = -n_i(0) \cdot (q_i^B(t) + q_i^D(t))$$

While those for subsequent generations (taking generation number 1 for example) become:

$$\dot{n}_1(t) = +2n_0^B(t) - n_1(0)q_1^B(t) - \int_0^t 2n_0^B(\tau)q_1^B(t-\tau)\,d\tau$$

$$-n_1(0)q_1^D(t) - \int_0^t 2n_0^B(\tau)q_1^D(t-\tau)\,d\tau$$

For the subsequent generations (say generation y), the terms can be explicitly expressed as compositions of terms from the previous generation (n_{y-1}) interacting with the distributions of the current generation (q_y).
By recursively repeating this operation, it is possible to connect back to the terms related to the initial generation indicated as n_0, linked to a combination of all terms $q_i^{B/D}$ from the intermediate generations (between y and 0). Thus, the model exhibits a structure alike that shown in figure below.

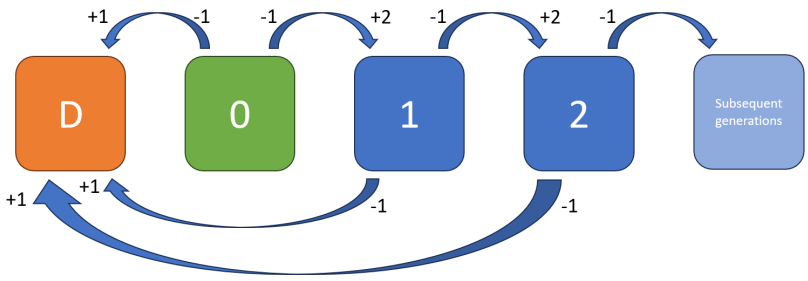


Figure 3.2: Structure of the populations model and scheme of transitions in an equivalent discrete system

Cells start from generation 0 (in green) and with each division progress into subsequent generations (in blue). When a cell dies, it enters the "population of the dead" (in orange). In the diagram, the arrows pointing to the right indicate transitions for the division process, while those pointing to the left (which necessarily end in the "population of the dead", the sink node of the graph) indicate the process of cell death.

3.2 Further considerations about the model

The Cyton model considers only two phenomena: cell division and cell death, treating them as independent and competing processes in determining the fate of the cell. Technically, this does not imply that at any given moment the Cyton model assumes only one of these actions occurs, but rather it imagines that each phenomenon can be described by independent clocks. A sample is drawn from each of the two distributions, and the action that occurs first is the one that will actually be observed to unfold at the extracted time.
Another hypothesis of the Cyton model is that division events reset the mechanisms responsible for "vital" processes, necessitating the extraction of new concurrent division and death times after each division.

To describe the mechanism of decreased cell activity as generations progress, and thus reduced involvement in the division process, the Cyton model introduces a variable called the *Progressor Fraction* (gamma) for each generation. This identifies the fraction of the population that, on average, participates in the division process. From the perspective of an individual cell, it represents the probability of actually undergoing division. The two+1 phenomena, i.e., division, death, and actual participation in the process (latter identified by gamma) appear to depend on the number of division rounds a cell undergoes since the stimulation of the naïve cell. Hence, there is a dependence of the probability distribution values and gamma on the generation to which the cell belongs. Gamma, in particular, is expected to decrease as generations advance, reflecting biological phenomena.

Indeed, another important aspect highlighted in the Cyton model, particularly in the 2011 paper Ref. [46], is how gamma is defined. The authors proposed a calculation involving a normal distribution, which represents the distribution of generations at which cells die. It appears that the underlying hypothesis is that cells die around a predetermined/fixed generation i on av-

erage, with uniform variability across preceding and succeeding generations, thus exhibiting a unimodal and symmetric process centered around the mean. This is biologically sound but, as far as we are aware, needs to be tested more thoroughly.

$$\begin{cases} \gamma_0 & \text{if } i = 0 \\ \gamma_i = \frac{\int_{i+1}^{+\infty} \mathcal{N}(n,\mu_{de},\sigma_{de})\, dn}{\int_{i}^{+\infty} \mathcal{N}(n,\mu_{de},\sigma_{de})\, dn} & \text{if } i = 1, 2, .. \end{cases}$$

The value of gamma associated with generation 0 is considered a free parameter, thus it can vary independently of the definition of subsequent gamma values. The latter are instead related to the ratio of two integrals, where the numerator represents the survival function from $i+1$ to infinity of a normally centered distribution with parameters μ_{de} and σ_{de}. The second is the same survival function but evaluated at i.

$$\gamma_i = \frac{\bar{N}(i+1)}{\bar{N}(i+1) + \int_i^{i+1} \mathcal{N}(n, \mu_{de}, \sigma_{de})\, dn} = \frac{\bar{N}(i+1)}{\bar{N}(i)} \tag{3.1}$$

The normal distribution $\mathcal{N}$ represents the probability density function of death, with respect to the variable n which varies among integers and represents the generation at which a cell dies. This complicated formulation is derived from experimental observations and meticulous data modeling efforts aimed at reducing the number of parameters used to define the trend of gamma across generations. Recognising the work done by the researchers, in this thesis we questioned whether a simpler description of this trend might be adequate. To perform this analysis, we first sought the optimal gamma values for describing our data and subsequently evaluated how well these values adhered

to the aforementioned definition.

In one of the version of the Cyton model [46], log-normal distributions are used to describe the processes of division and death. A log-normal distribution is characterised by its mean and standard deviation, which uniquely define it. Experimentally, a significant difference in division times has been observed between undifferentiated naïve cells and the populations generated during clonal expansion. Thus, the Cyton model adopts a pair of distributions (division/death) for generation zero, representing naïve cells, and another pair (division/death) for the development of all subsequent generations. This is an important feature.

Another important aspect of the Cyton model is that what are called the probability distributions of division/death at first glance seem to be merely event rates over time. If this were the case, they would not necessarily sum to 1. However, it is crucial to note that they are descriptions of how the division or death process unfolds over time, and thus they function as probability density functions (pdf) for a cell belonging to that generation. Conceptually, these must integrate to 1, as the idea is that within some time t, at most infinite, there is certainty that this phenomenon will occur, though it can be distributed between the initial moment and the maximum moment in any manner.

The observability of division and death phenomena allows these processes to be described over time, characterised either by rates or alternatively by a probabilistic description (which integrates to 1 over time). However, it is not possible to individually observe the effects of division and death, as the coexistence of the two means that the occurrence of one limits the free observability of the opposite event. Thus, the two phenomena are always observed net of each other, (which does not mean the net effect of population increases or decreases, as that is represented by the sum of the two "net" phenomena).

It is possible, however, to alter the biochemistry of a cell by exposing it to certain simple molecules and proteins, which can induce a cessation of the division process, as done in Ref. [19]. At this point, the phenomenon that can be freely observed is death. However, it must be considered that this altered state may also influence the normal functionality of the death process, and therefore the distribution found may not be entirely accurate.

We now conclude this section with some general consideration on the Cyton model and its characteristics.

The system lacks disturbance or perturbation terms, and the complete heterogeneity of the population (thus, the stochastic element defining how different agents divide, reproduce, and act during the immune response) is entirely encapsulated within the probability distributions of division and death.

This model is "sophisticated" from a certain perspective, as it encompasses various elements of stochasticity arising from separate phenomena. These can include environmental variability in which the population is immersed, "randomness" due to individuals with different characteristics within the population, and even individual internal variables of the cell (e.g. phases of the cell cycle). All these sources of variability are described and included in the distributions (we keep mentioning). For this reason, we can say that the model adopts a phenomenological approach.

The cyton model is thus a phenomenological model, meaning it is based on the description of measurable processes concerning the cell population and does not attempt to infer the microscopic dynamics that drive these processes. It is, of course, possible to make hypotheses about the biochemical causes of the phenomena, but it is important to remember that the real element is the observable measurements.

The Cyton model is an unconstrained model, meaning there are no explicit constraints on the solutions of the model, which is entirely based on

the parameters that define it. Saying that the Cyton model is unconstrained expresses that the constraints of "mass" conservation and the admissibility of state variable values are not intrinsic to the system's equations but depend solely on how the division and death distributions are chosen.

These parameters are determined based on the death and division distributions. Once the probability distributions of death and division for each generation are defined, the entire set of system coefficients is determined and thus not influenced by the dynamics.

One final element to note is the fact that in evolution, a few very fit/good individuals contribute the most to the advancement of the species against selection. This process is thus a succession of continuous bottlenecks, which is something also observed in our data (as seen in figure 2.6, in the previous chapter).

Note that the shape and parameters of the aforementioned distributions must be experimentally assessed, as they fully characterise the solution of our system of differential equations. Fitting the model involves adjusting the dynamics by varying the distributions and their parameters. The final analysis involves comparing the optimal distributions after regression with the actual distributions experimentally found for the division and death processes of individual cells or the population.

Chapter 4

Analysis of results

In this chapter, we will first address and explain the fitting process of the Cyton model and the procedure used. This will include an analysis of how the data was handled and the structure of the stochastic optimization algorithm. Next, we will examine the results obtained from the fitting process, evaluating how well the model describes both the benchmark data from Ref. [46] and the original data in our possession. For the latter, we will compare the models fitted for all pedigrees together and for the pedigrees divided into groups. Finally, the results of a sensitivity analysis, assessing the model's response to parameter variations, will be presented.

4.1 Fitting the model

The process of studying the model and analyzing its correct implementation was carried out using real data. The first dataset, which we will call the benchmark, consists of the data presented in the Ref. [46], where a thorough description of the Cyton model is given. These data represent the response of a set of B cells to an infection.

These data were analysed by fitting the model. Once the fitting was achieved in-depth exploration of the biological process was conducted, using

original data for the clonal expansion of CD8+ T cells in response to interaction with a pathogen.

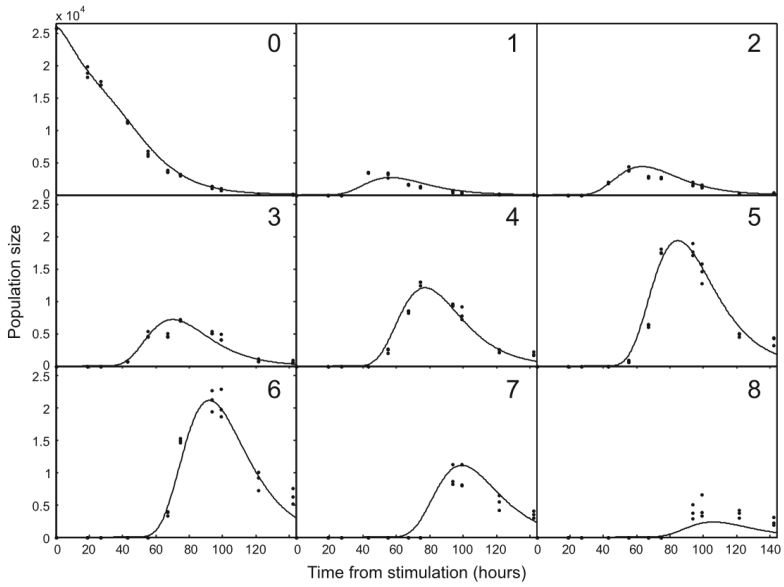


Figure 4.1: Set of benchmark data used for the model evaluation phase in the present thesis. Image from Ref. [46]

4.1.1 Optimising the Cyton model for our experimental data

The fitting procedure

To analyse the data using the Cyton model, one must first fit the model to the data. This entails adjusting the model's free parameters, namely the mean and variance of the distributions related to division/death processes, as well as the progenitor fraction gamma for each generation. Varying these parameters allows the simulated model results to approximate the actual data trend. "Matching" was done by measuring the distance between the data point and the simulated curve point; that distance was then appropriately modified to obtain a score. After testing different formulas for score calculation, including

linear, exponential, and polynomial functions, a quadratic form was chosen, as follow

$$Score : S = (d \cdot 20)^2 \tag{4.1}$$

with d representing the difference between the values predicted by the model and the actual data.

$$Distance : d = \sum_{i=0}^{N_p} (y_i^D - y_i^M)$$

This choice was made so that the obtained models, while tending towards a low mean error (as in the case of the linear score), would favor a lower variability in the fit error. This way, all the models stay close to the target points, instead of giving equal importance to both small and large errors, which is what happens with linear forms.

The final goal was to push the model curves representing population dynamics as close as possible to the points present in the data. The fitting procedure was refined through successive attempts, modifying and adding steps to the pipeline. A standard protocol was eventually established, which was subsequently applied in all fitting attempts. This protocol is composed of two sections:

• The first section, in which the parameters of the distributions are varied, is further divided into three steps. 1) Initially, models are simulated by exploring the parameter space using a grid search. A grid search is a systematic method of searching through a manually specified subset of the hyperparameter space of a learning algorithm. Based on the best models obtained from the grid search (best in terms of score, which will be assumed from here on), the next two steps of stochastic optimisation are performed. 2) The next step involves optimising only six parameters: the mean and variance of the death rate kept homogeneous for all generations, the mean and variance of the division rate for the 0 generation of naïve cells, and finally, the mean and variance of the division rate for generations following the first, which are the ones characterised by clonal expansion. 3) The second version of stochastic optimisation

further includes the subdivision of death rate distributions similar to what is done for the division rates, thus having one distribution for the 0 generation and another distribution for generations from 1 onwards. In this first phase of the pipeline, only the division and death rates were varied, while the progressor fraction gamma was kept static/invariant to reduce the dimensionality of the problem. The gamma values used for each generation were predefined based on some biological hypotheses about the data: the average generation of death being the $5th$ and deaths being normally distributed with a variance of 1. Following the formula defined in Wellard et al. [46] for γ_i, individual values for each generation were obtained. This dimensionality reduction procedure practically meant only searching for the best distributions of rates conditioned to the fixed gamma values.

• In the second phase, starting from the best models obtained in the first phase and fixing the parameters that describe them (mean and variance of the distributions), the value of gamma (progressor fraction) for each generation was varied freely. In this phase, the goal was to find the best trend of gamma, conditioned on the set of parameters found in the first optimisation phase.

It should be noted that this procedure could lead us to the global optimum only under the assumption that the optimisation of the mean/variance parameters of the distributions and the trend of gamma across generations are independent. This means that whether we optimise them together or sequentially does not affect the "reciprocal" behavior of the optimisation dynamics. However, this assumption is too strong and is a limitation of the study, made necessary by the limited computational resources. To further investigate this analysis in the future, we could move to a single optimisation phase that includes all parameters, or we could repeat this optimisation procedure cyclically, seeking convergence as the optimisation iterations progress.

At the end of the fitting phase, the parameters obtained from stochastic

optimisation are analyzed. If these parameters exhibit limited variability, it can be said that the model is in a local optimum, albeit within a "limited zone". On the other hand, if the expressed variability is wide, it is likely that the different optimisation dynamics have led the models to "scattered" local optima.

An additional element considered is that the variability expressed by the parameters post-optimisation can be useful in understanding the importance of the parameter itself. If high variability does not result in a significant difference in scores between models, it can be inferred that the model is not particularly sensitive to that parameter's value. Conversely, if a parameter exhibits low variability (the value remains the same across different optimisation dynamics), it is likely that this parameter is a defining characteristic of the phenomenon described by the Cyton model.

Details on the Stochastic Optimisation Algorithm

All the stochastic optimisation algorithms described here were implemented from scratch in MATLAB and operate as follows: starting from a fixed set of parameters (whether means and variances or gamma values for each generation), the algorithm adds or subtracts a random quantity between 0 and 0.5, with a decreasing optimisation step starting from 0.1.

death mean	**death std. dev.**	**division mean**	**division std. dev.**
$\mu = 6.0$	$\sigma = 1.5$	$\mu = 7.5$	$\sigma = 0.9$
$\mu = 6.0$	$\sigma = 0.9$	$\mu = 2.0$	$\sigma = 2.2$
$\mu = 7.0$	$\sigma = 2.0$	$\mu = 4.0$	$\sigma = 0.4$

Table 4.1: Table with examples of parameters for initial models before the stochastic optimisation

In the initial optimisation steps, this adjustment occurs for multiple pa-

rameters simultaneously, then for only one parameter per step (i.e. in each optimisation step, only one parameter is modified), using a variable and decreasing optimisation step size during the procedure. Once the new set of parameters is obtained, the algorithm simulates the model, calculates the distance/score relative to the real signal, and if the new score is better than the previous one, the parameter modification is retained. If the score is worse, the probability of retaining the modification decreases as the score worsens, this means that in case of worsening of the score the acceptance of the modification has a probabilistic nature.
The algorithm is based on the idea of simulated annealing with stochastic corrections of the variable to be optimised and probabilistic acceptance of these variations.

How to fit the original data

Moving on to the analysis of the original data, these are divided by families, meaning by the originating naïve cell, and, for each family, the number of cells in each generation over time is tracked. As a first step, we used the fitting pipeline on a "fictitious" population, created by combining the time-course data of all the pedigrees, starting from a 0 generation composed of 16 cells and advancing generation by generation up to 8. We stopped our analysis at generation $5th$ because in that generation the data are "truncated", as only 4 of the 16 clones present were continuously monitored, while the others from that moment on were only used for counting. This experimental decision resulted in the loss of generation-wise data, meaning that, for the remaining clones, we only have the total number of cells without a breakdown by generations, so we cannot fully apply the Cyton model. Due to this reset in the data (around the equivalent time 8000) starting from generation 5 we decided to fit the Cyton model only up to the fifth generation only, excluding the subsequent ones.

Firstly, we considered the variability of the dynamics exhibited by different

cell pedigrees, with some families producing a large progeny and others much less.

We then decided to divide the 16 pedigrees at our disposal into three groups, based on the total cell count data for each pedigree. The families that reached the highest cell peaks were grouped as 'best individuals', the ones that produced the fewest cells were called 'worst individuals', and the remaining families that did not stand out either positively or negatively were labeled 'middle'. This gave us an idea of typical outcomes for clonal expansions.

The question we posed was whether the phenomenon of response to a pathogen, with the activation of naïve cells and subsequent expansion, is homogeneous among the pedigrees involved. We wanted to determine if all processes can be described in the same manner, considering intrinsic variability, or if the macroscopic process is an average of structurally different processes. Our idea was to compare the fitted models for the three groups (best, worst, and middle) and the population containing all pedigrees, evaluating whether the parameters of the distributions for division and death phenomena differ significantly or reflect the same pattern across all groups. We fitted the Cyton models using the described pipeline on each group.
The data structure for each group is the sum, for each generation, of the number of cells belonging to the pedigrees of that group over time.

Another analysis we considered is fitting the model to a single pedigree, whether it is an "outlier" (either high or low) or an average one. However, this approach presents technical limitations, as the early generations consist of a very small number of cells (1, 2, 4, ...), resulting in very irregular data signals with abrupt vertical jumps. This poses a problem for the Cyton model, which was developed as a mean-field model and thus is less effective at representing individual cases (for this task, a stochastic agent-based model could be more

appropriate). For this reason, we decided to use the "group" fitting approach (best, worst, middle) to have more data and thus slightly smoother and less irregular signals.

The question of whether cellular variability is simply intrinsic to the process or requires the addition of external factors is interesting because it explores the microscopic dynamics inherent to each individual rather than just the average trend across the entire population. This approach aims to be as grounded as possible in the biochemical reality of the processes behind such complex phenomena, without losing focus on observable variables.

Data manipulation

The data extracted from the 2011 Cyton paper [19] contained multiple measurements (values on the y-axis) for each time point. To achieve a smoother signal, we decided to use a transformation of the measurements as the actual signal for the fit. Both the mean and the median of the measurements were tested, and the mean was ultimately chosen because we expect that, given how the distance score is constructed, the model tends to converge towards the mean of the measurements.

Data manipulations were performed to evaluate the results of the fitting pipeline with more complete and accurate data. We tried inserting "fictitious" points into the data dynamics in addition to the actual ones using a simple averaging technique on both the x-axis and y-axis, but this did not yield significant advantages.

Details of the Grid-search procedure

The grid search on the model parameters (μ, σ, and γ_0) was not performed in a single pass spanning all possible values of the parameter state space.

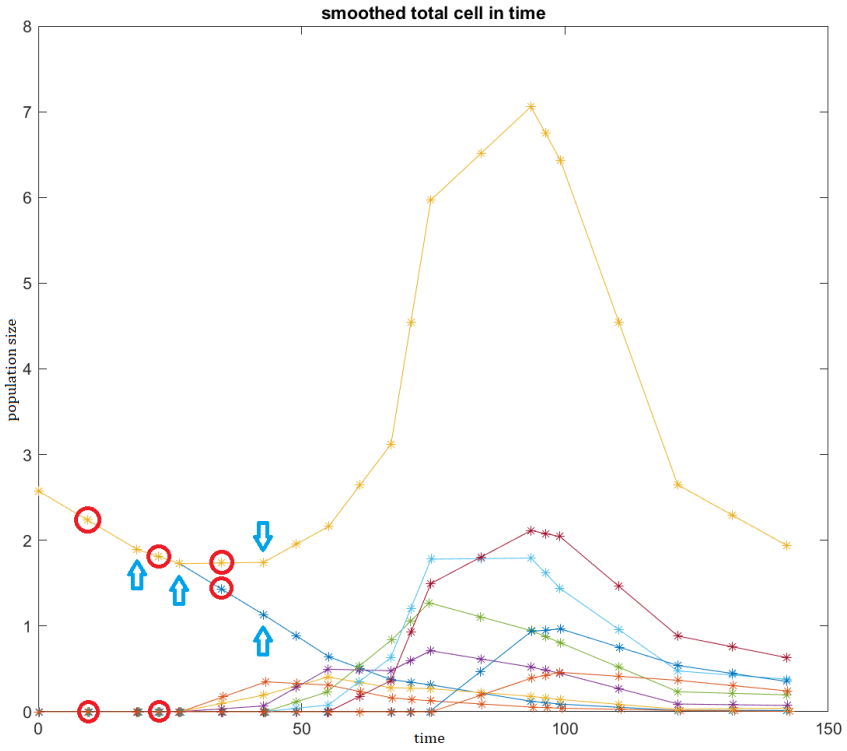


Figure 4.2: The image shows the total number of cells data (the curve in yellow above) and the same data divided by generation with original points highlighted in blue and added points highlighted in red

Instead, it was conducted using four different "offsetted" grids. This approach allowed for a reduction in search time while maintaining a wide variability of evaluated parameters. The grids in various searches were always "offset" relative to each other, thus covering different values of the spectrum. This strategy helped reduce the problem's dimensionality and, consequently, the required computational time. The idea behind this strategy is that by selecting the 35 best points and the 5 worst points for each grid search and subsequently applying stochastic optimisation techniques, the initial set of selected values is broad enough to avoid local optima.

Introduction to sensitivity analysis

To evaluate whether the model that describes the data, found at the end of the optimisation pipeline, is unique, meaning the parameters defining it are truly characteristic of the curves of the described dynamics, we proceed to assess the sensitivity of those curves to the variation of these same parameters. We conducted an analysis allowing the obtained parameters to vary by 10% and then analyzed the variation in the score obtained. It was observed that the score had oscillations greater than 10%, resulting in "elasticity" and indicating that for certain parameters, the sensitivity is high. This means it is important that the parameters are as they are, allowing us to say that the correct model is "unique" and not equivalent to other models with different parameters values.

Algorithm 1 Pseudo-code of stochastic optimization

≫ Research for starting points:

> Run the 4 "offsetted" grid searches, evaluating the model defined by the parameters present in each of the grid points

> From each grid search select the 35 models with the lowest score and 5 with the worst score

≫ Stochastic optimization of PARAMETERS DISTRIBUTIONS on the 160 selected points:

> 1st algorithm: optimize 3 log-normal distributions, gen0+ death, gen0 division, gen1+ division

> 2nd algorithm: optimize 4 log-normal distributions, gen0 death, gen1+ death, gen0 division, gen1+ division

> From each algorithm select the 50 resulting models with the best fitting score

≫ Stochastic optimization of the PROGRESSOR FRACTION γ on the 100 selected points:

> algorithm: optimizes the gamma values for each generation, keeping fixed the values of the 4 log-normal distributions found in the previous steps

≫ Analysis of results: best model and parameter variability

4.2 Results from the fit

Let us now analyze the results obtained at the end of the fitting process.

Fit of the benchmark data

The data used to validate the model are well described/fitted by the model. Despite the limited number of points available, the simulated curves closely match the dynamics present in the data, adequately following the data both in the steeper growth phases and in maintaining the populations over longer timescales (as seen in generations 1 and 2).

Plots of the individual generations are included in Appendix A for better observation.

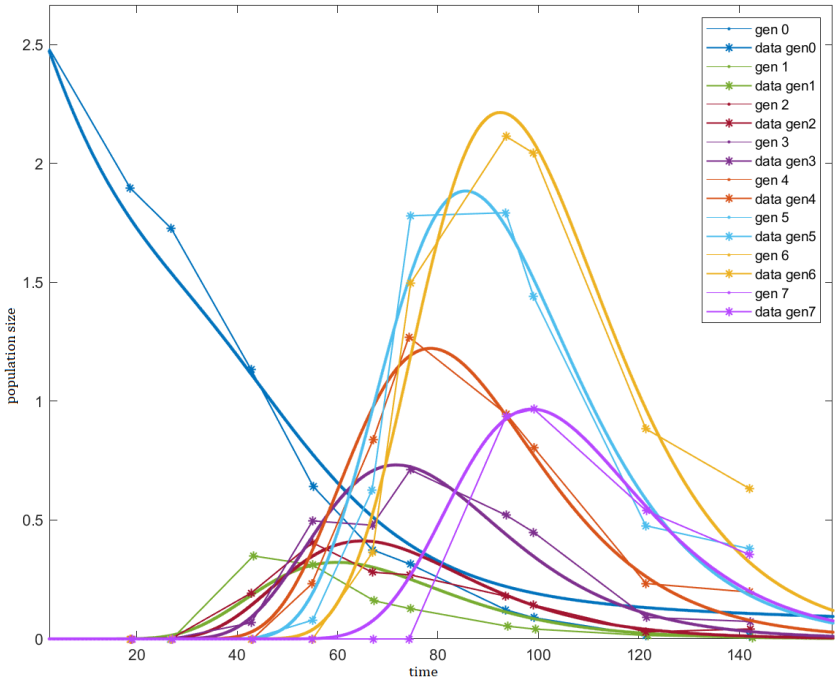


Figure 4.3: Fitted dynamics of the best Cyton model obtained at the end of the fitting procedure over benchmark data. Each color represent a generations from the 0th to the 7th, the smooth curves are the predicted dynamics.

By using a score that evaluates the distances of data points from the simulation in an aggregated manner (recall score from equation 4.1), the mean distances were minimized, leading the curves to pass close to the data points without having to "exactly cross" them.
From the image of the simulated curves, it is evident that the Cyton model describes well the phenomenon under analysis. We will compare these performances with those of an ODE model and a stochastic model (which start from the same hypotheses as the Cyton) in subsequent chapters 5 and 6.

Now, we present the theoretical progressor fraction (gamma) curve, obtained from the data using the form stated in the Cyton model paper, formula 3.1 and we compare it with the gamma curve obtained post-optimisation.

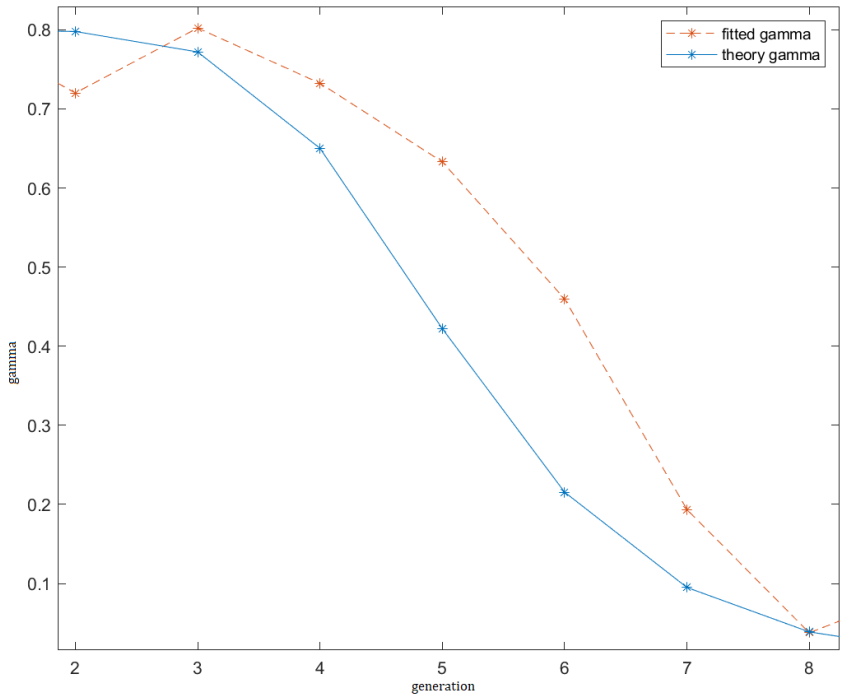


Figure 4.4: The profiles of gamma over generation, from the second generation (n°1) to the eighth (n°7). In blue the theoretical values from the formulation equation 3.1, in red the values obtained at the end of the fitting procedure on benchmark data.

The dynamics obtained from the optimization process are similar in form to the theoretical ones, giving us the impression that the idea behind equation 3.1 used to describe the behavior is sensible, but it shows a delay in the decrease of values. In particular, it is noted that the curve found through the optimization process, starts with a lower gamma value for the first generation, this results in higher values for all subsequent generations compared to the theoretical curve. The distance between the two curves increases up to the 6th generation and then decreases, becoming zero at the 8th generation, likely due to the "extinction constraint" of the population.

To evaluate the effectiveness of the overall Cyton model and understand also how well a model with specific parameters describes the data, a sensitivity analysis was performed. This process involved taking a definite model, obtained at the end of the fitting pipeline, and varying its parameters within a range of -10% to +10% of their value. We defined and evaluated two KPI parameters: 1) the mean of the absolute value of the difference between the scores ($mean(|\tilde{s} - s^*|)$). 2) the mean of the ratio of the absolute value of the difference in scores to the optimal score value ($mean(|\tilde{s} - s^*|/s^*)$). The result is shown in the table below.

varied parameter	$mean(\lvert\tilde{s} - s^*\rvert)$	$mean(\lvert\tilde{s} - s^*\rvert/s^*)$
death gen 0 mean	3133.7	4.674
death gen 0 std.dev.	34.8	0.050
death gen 1+ mean	773.2	1.046
death gen 1+ std.dev.	14.9	0.021
division gen 0 mean	8401.1	13.591
division gen 0 std.dev.	39.0	0.041
division gen 1+ mean	293.0	0.420
division gen 1+ std.dev.	2.9	0.004

Table 4.2: Table shows the mean difference in fit error between the optimal model and a model with one parameter varied in a (-10%, +10%) range. The first number is the absolute difference, the second one is the relative variation in fitting score.

From this data, it is understood that some parameters are more important than others. In fact, varying them leads to a significant change in the score, while others seem to have a minimal impact, at least within the range of variations considered. In particular, it seems that the greatest impact on the model comes from the means of the distributions, which always exhibit greater sensitivity compared to the associated standard deviations. A possible explanation for this could be that the absolute values of the standard deviations are one to two orders of magnitude smaller than those of the means, and thus their relative variation does not significantly affect the shape of the log-normal distribution under examination.

It was decided not to show the graphs containing more details of the sensitivity analysis, as they are not the focal point of this thesis and would not add anything to the summarized data presented in the table above.

Fit of original data - all pedigrees together

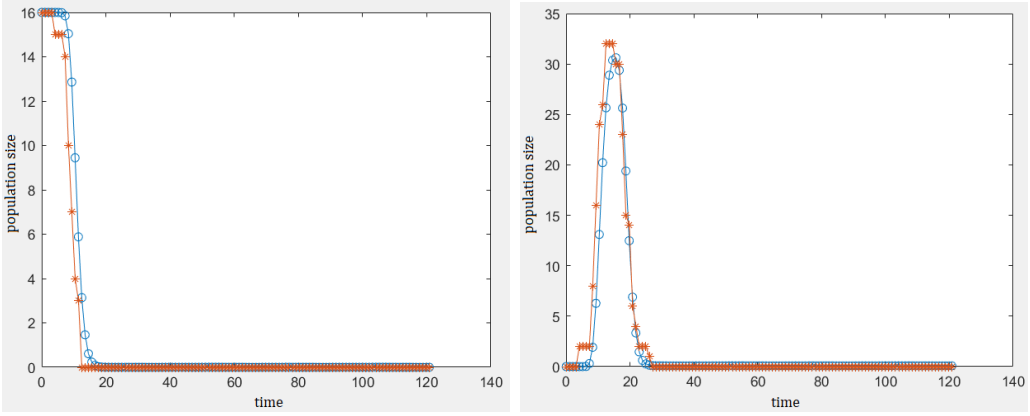
The fit on original Kajal's data was performed following the same procedure previously described and used for the benchmark data. It is important to remind that only the first 5 generations were subject to fitting due to the limitations of the data collection procedure, and the data for the more advanced generations were not considered.

A first fit was performed on the entire set of summed pedigrees, mimicking a population of 16 naïve T cells that activate and enter the clonal phase. The best model resulting from this fit presents the following parameters:

distribution	**mean**	**std.dev.**
death gen 0	$\mu = 7.559$	$\sigma = 0.329$
death gen 1+	$\mu = 5.768$	$\sigma = 0.7535$
division gen 0	$\mu = 4.792$	$\sigma = 0.111$
division gen 1+	$\mu = 4.718$	$\sigma = 0.087$

Table 4.3: Table with the parameters of the best fitting Cyton model for the "all-pedigree" group. It is possible to compare these values with the ones below obtained fitting each of the other groups of pedigree.

Here we present the plot of the very same model:



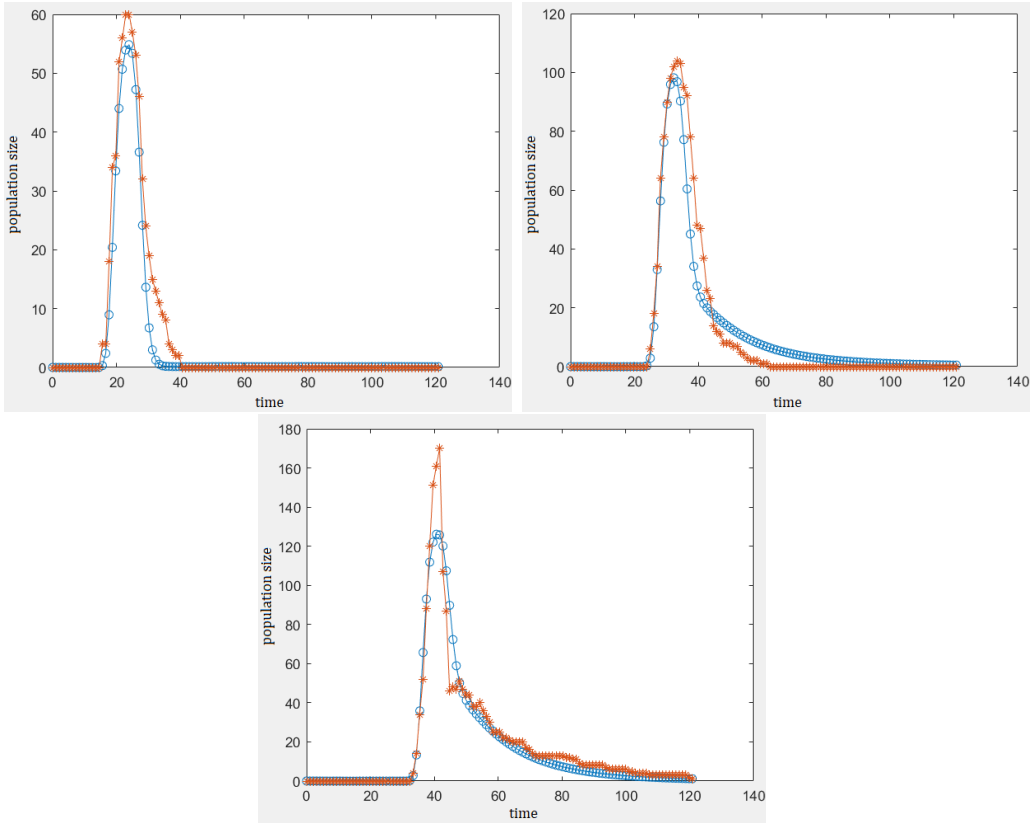


Figure 4.5: All pedigree cells group - Plot of generation 0 - 4, in red the real data, in blue the simulated dynamics with the parameters of the best Cyton model

It is observed that already in generation 4 (the fifth), the change in cell count induces an effect that narrows the curve, resulting in a steeper decrease in the number of cells. This might be why the model cannot follow the signal in the higher part or in the tail, differently than the previous generations (which describe the natural process and are not affected by the counting procedure). A similar effect from this problem will also be noticed in generation 4 of the other groups, in particular in the "worst pedigree" group.

Fit of original data - the different groups

Subsequently, fits were performed on the three previously created pedigree groups: worst, middle, and best.

Now we analyze the parameters of the best model obtained at the end of the fitting pipeline for these three groups.

parameter	worst (4gen)	worst	middle	best
death gen 0 mean μ	5.615	7.280	7.222	5.226
death gen 0 std.dev. σ	0.278	0.276	0.294	0.220
death gen 1+ mean μ	5.615	5.846	5.799	5.226
death gen 1+ std.dev. σ	0.278	0.780	0.691	0.220
division gen 0 mean μ	4.835	4.775	4.737	4.900
division gen 0 std.dev. σ	0.241	0.0001	0.100	0.109
division gen 1+ mean μ	4.937	5.015	4.730	4.643
division gen 1+ std.dev. σ	0.145	0.284	0.085	0.021

Table 4.4: Parameters value of the best model for each cells group. We also added the results of the fit of the worst group on 4 generations instead of 5, trying to eliminate the impact of the variation in counting procedure.

The plots of the best model for each group have been included in Appendix B for space reasons.

Here we only present a brief comparison between the last fitted generation (the 5th) of each group.

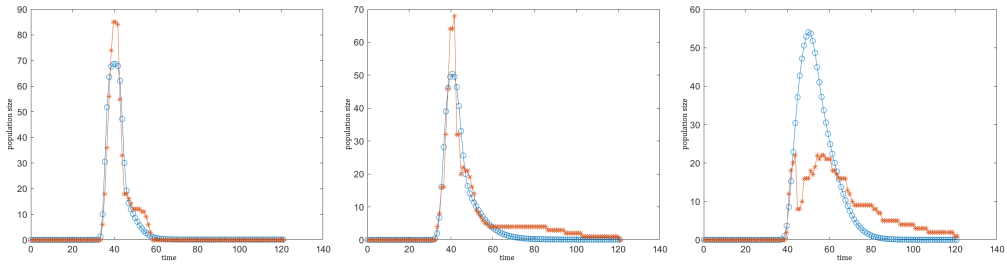


Figure 4.6: Comparison between the 5th generation of each group. Starting from left we show the best group, the mid group, the worst group.

In all the models the change in the counting procedure happened during the data gathering phase has an impact on how the dynamics of the population is affected. The most relevant effect is the insertion of a unnatural bi-modal dynamic in the worst group.
The Cyton model fit better the best group and worse the worst group as an effect of this twist in the data.

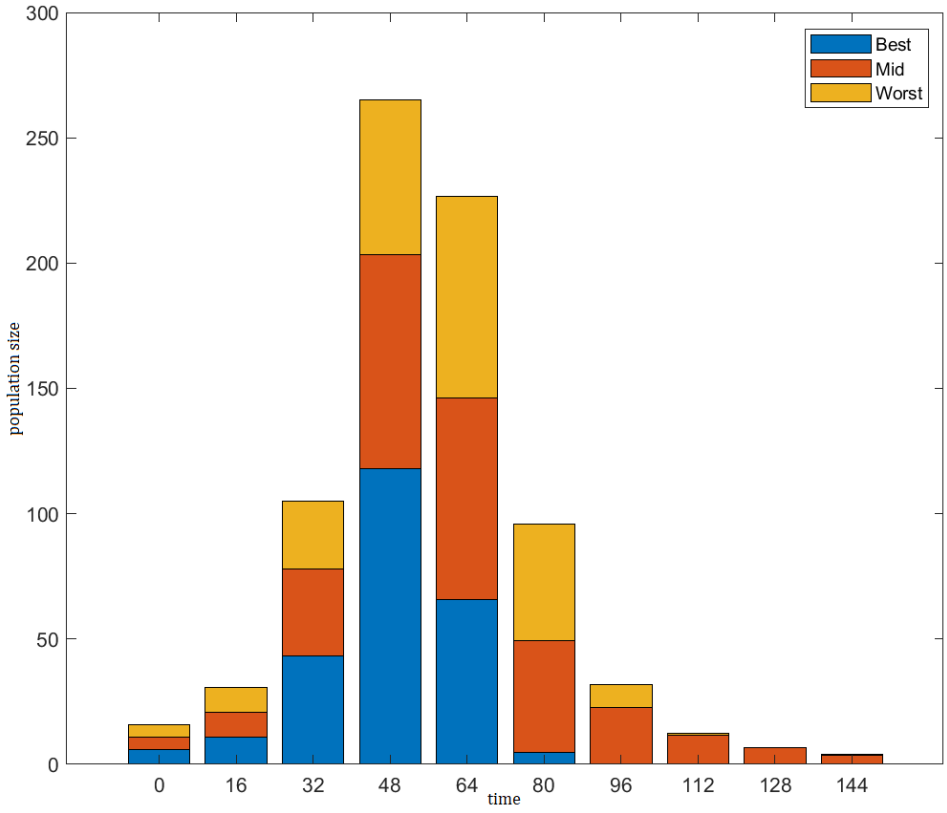


Figure 4.7: Plot of the cell population trends over time for the three groups. The overall shape is a skewed bell curve, but each group has different timing for the rise and decrease phases.

To analyze how the dynamics of the Cyton model, fitted to the data for the various groups, contribute differently to the total cell population, we present an image similar to the one shown in chapter 2, illustrating how the population fraction of each group (best, mid, worst) changes over time. It was

not possible to replicate the precision of the original image, which included the breakdown by individual pedigrees, as the model was fitted to the groups rather than individual families. The x-axis represents time measured in hours, while the y-axis shows the number of alive cells for each group at the time of measurement.

It is evident that the best group has the highest absolute numbers among the three groups, while the mid group exhibits a decline that extends for a longer period. Notably, the pedigrees divided in this manner do not show an absolute dominance of one group over another throughout all time, but rather dynamics with faster evolution times. In fact, it can be observed that by time t=80 hours, the best group has already gone extinct, while the worst group still persists and is, at that point, the majority of population.

In the figure below we show the profile of the progressor fraction γ for each group.

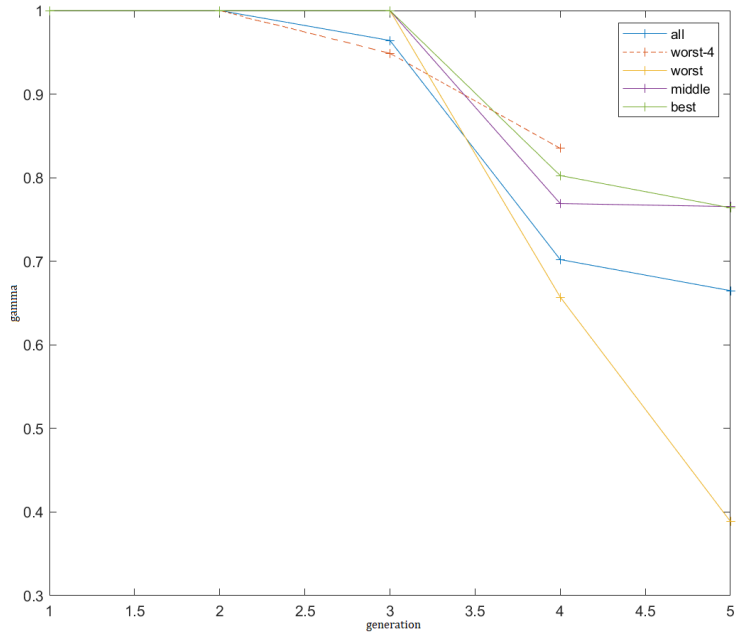


Figure 4.8: Values of the progressor fraction in the various generations fitted (the first five). For each group it is shown the profile of γ of the best model obtained at the end of the fitting procedure

In this case, the values do not exhibit the convex shape expected from the first part of the profile calculated with formula 3.1. However, the decreasing profile of the progressor fraction remains, with different values depending on the group of naïve cells observed. Excluding the "worst-4" group, whose values can be shifted by the fact that the 5th generation is not considered in the fitting (this should make us aware of how changes in the counting process in the data might affect the values), the others groups show a clear order, coherent with the different trend in the groups, with the best over the others and the worst at the bottom.

In the end, we consider the variability of the parameters resulting from the fitting procedure, analyzing the values of the top 20 models for each group. First, we calculated the mean value across the 20 models for each parameter, and then estimated the standard deviation of these values from the mean. As explained before, high variability in the parameters indicates that there are many models that fit our data well, suggesting the existence of multiple local optima from an optimisation perspective.

parameter	**worst(4g.)**	**worst**	**middle**	**best**	**all-ped**
death gen0 μ	6.162 (0.60)	6.274 (0.81)	6.207 (0.85)	6.468 (0.92)	6.193 (0.87)
death gen0 σ	0.432 (0.13)	0.616 (0.23)	0.464 (0.30)	0.590 (0.32)	0.589 (0.14)
death gen1+ μ	5.973 (0.27)	5.860 (0.03)	5.508 (0.23)	6.011 (0.53)	5.695 (0.11)
death gen1+ σ	0.500 (0.18)	0.747 (0.06)	0.496 (0.20)	0.812 (0.39)	0.714 (0.10)
div. gen0 μ	4.849 (0.07)	4.649 (0.14)	4.867 (0.09)	4.822 (0.08)	4.849 (0.07)
div. gen0 σ	0.222 (0.05)	0.047 (0.08)	0.066 (0.05)	0.094 (0.02)	0.102 (0.03)
div. gen1+ μ	4.929 (0.03)	5.044 (0.04)	4.678 (0.04)	4.676 (0.03)	4.697 (0.03)
div. gen1+ σ	0.137 (0.05)	0.275 (0.01)	0.106 (0.03)	0.029 (0.03)	0.088 (0.02)

Table 4.5: Mean and standard deviation in the format mean (std. dev.) of the parameters of the 20 models with the best fitting score for each group (the parameters are μ and σ of the 4 distributions of the Cyton).

A point that was brought to my attention by the cell biology research group at Swinburne University is that the difference between the sensitivity analysis conducted for the benchmark data and this variability analysis of the results of optimization process, lies in their focus. The former (sensitivity study) examines whether random variations in good models can lead to either good or poor models (thus assessing whether the parameter space is essentially "flat"). However, it does not allow for a complete evaluation of the existence of multiple suitable models to describe our dynamics (i.e., the existence of multiple equally optimal local optima in a non-flat space). This latter aspect is highlighted by a results variability study.

Chapter 5

Comparison with ODE model

In this chapter, we introduce a simpler ODE model compared to the Cyton model to evaluate the differences in descriptive capabilities for the clonal expansion process. Following a similar structure to the previous chapter, we present the fitting method and the results it yielded on the benchmark data from Ref. [46], followed by the results on our original data.

The ODE model

To evaluate the accuracy of the Cyton model in fitting the biological data at our disposal, we compared the results with those obtained by fitting a simpler mathematical model, a system of ordinary differential equations (ODE), which is introduced below.

The system, starting from the same assumptions as the Cyton model, simulates the phenomena of division and death, using a different coefficient for each generation. It is not necessary to model the behavior of the progressor fraction (gamma) as this is implicitly included in the division rate.

The equations of the system are therefore:

$$\begin{cases} \dot{n}_0(t) = -a_0 n_0(t) - m_0 n_0(t) \\ \dot{n}_1(t) = +b_1 n_0(t) - a_1 n_1(t) - m_1 n_1(t) \\ \dot{n}_i(t) = ... \qquad with\ i = 2, 3, ... \end{cases}$$

Where a_i represents the division rates, m_i the death rates, and the birth terms b_i are exactly double the division terms $b_i = 2a_{i-1}$.

Fitting procedure

Also in this case, the model was initially fitted to the data from the Cyton paper [46].

The fitting procedure was modified due to the structural differences from the previous model, but the stochastic nature of the optimisation algorithm was retained. Division and death rates for each generation were varied, starting from an initial set of values where all division rates were set to 2 and all death rates to 1.5. Subsequently, at each optimisation step, a random variation was applied to one of the parameters, mediated by a variable optimisation step size (θ), whose value was made to oscillate during the optimisation process, following a decreasing trend. This approach was inspired by the concept of a "thermal hysteresis cycle" typical of simulated annealing algorithms.

Fitting results on the benchmark data

The fitting results, however, showed a poorer adaptation to the data compared to the Cyton model. The simulated curves were unable to rapidly follow the changes in dynamics observed in the data and completely missed the delay effect between successive generations (which is strongly visible in generations 0, 1, and 2).

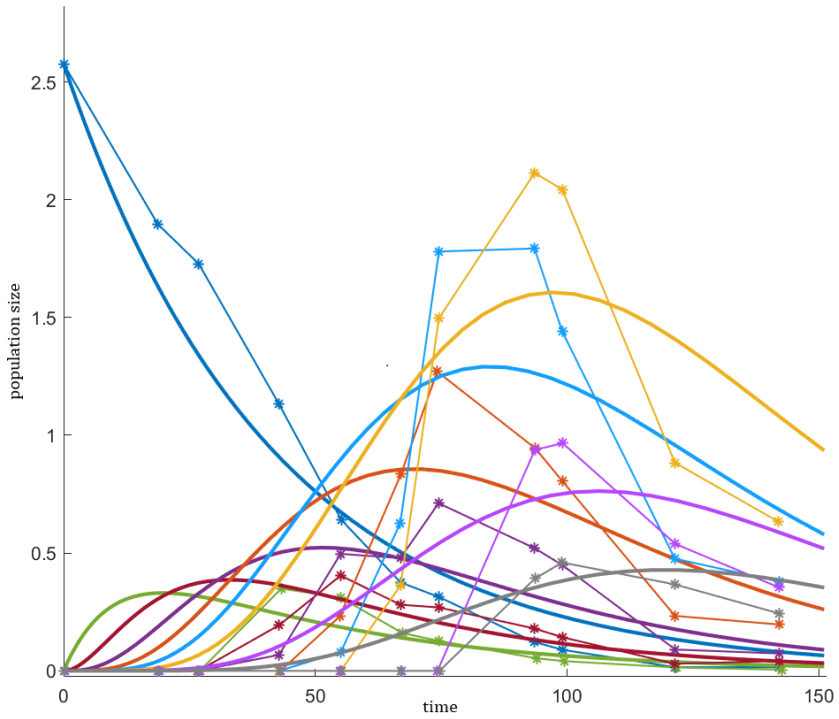


Figure 5.1: Fitted dynamics of the best ODE model obtained at the end of the fitting procedure over benchmark data. Plots of the individual generations are included in Appendix C.

We now present the parameters resulting from the fitting process over 100 repetitions and their variability:

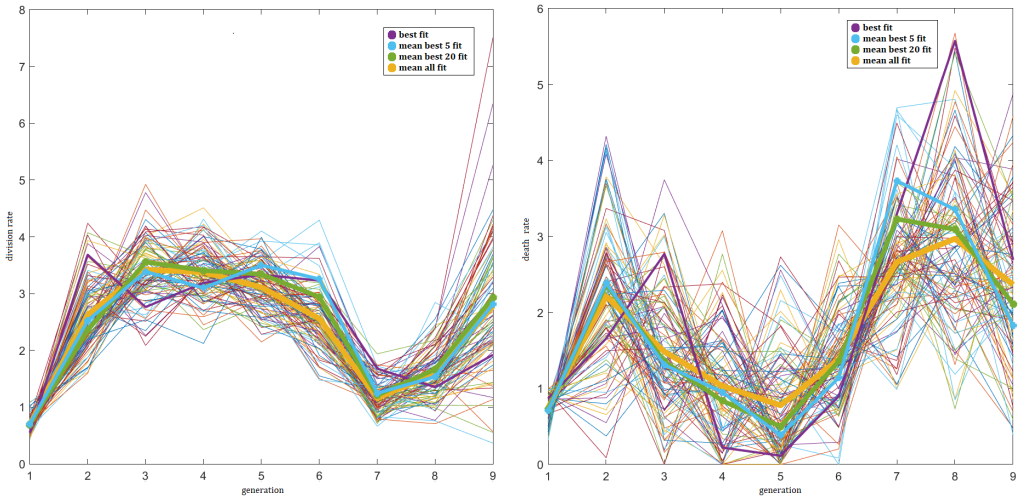


Figure 5.2: On the left there are the division rates and on the right the death rates for each generation obtained at the end of the fitting procedure. We also show the variability in the results plotting every set of parameters obtained by different iterations of the optimisation pipeline.

It can be observed that the division rates exhibit a clearer dynamic, with parameters whose values remain more stable across repetitions. In contrast, the death rates show greater dispersion in the results.
Additionally, the plot includes signals representing the average parameters across all obtained models (in yellow), the average among the best 20 resulting models (in green), and the best 5 (in blue), and highlighting the values of the best model, which had the lowest fitting error.

To better illustrate these trends, we compare them, isolating them from the results of all 100 simulations:

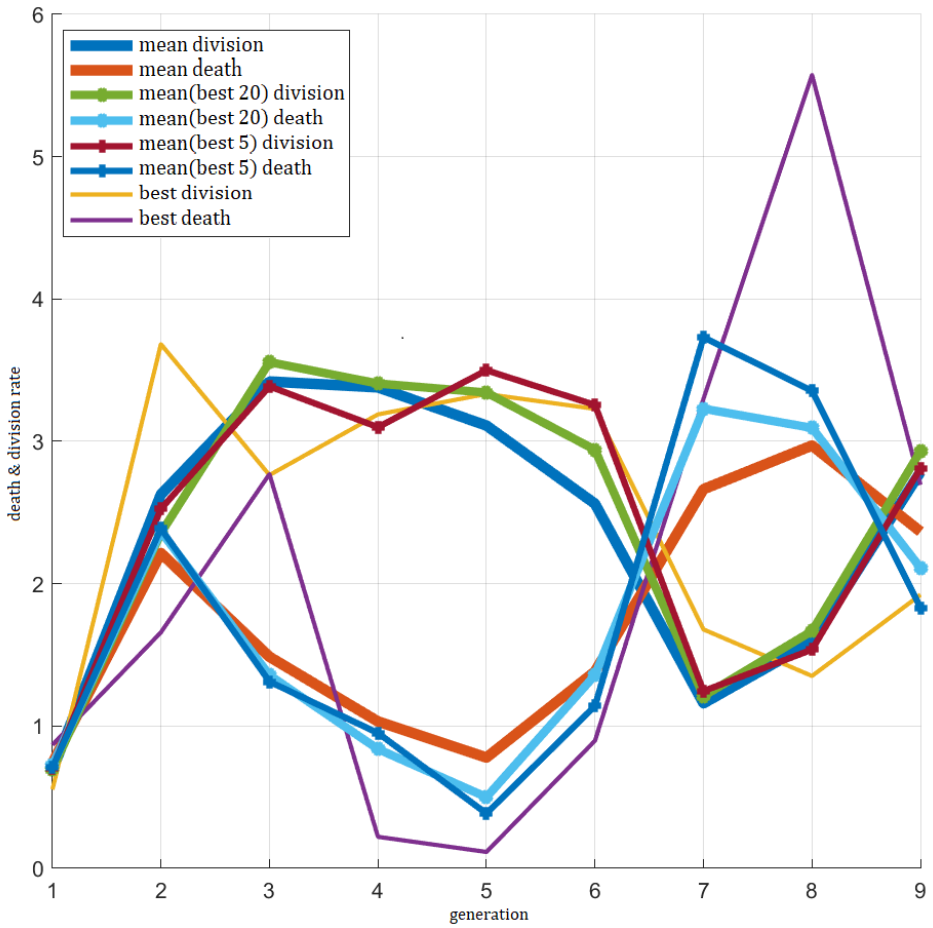


Figure 5.3: Resume and comparison of the trends of division and death parameters obtained by the optimisation pipeline. Here are shown the parameters of the best model and various aggregated results such as the mean over each independent iteration, the mean over the 20 best models, and over the best 5.

Despite the previously shown variability, the results demonstrate common trends. Indeed, the overall average trends and those considering only the best models follow similar dynamics. This suggests that the model and the optimisation algorithm are reliable and stable, capable of achieving an optimal fit for the considered data. Also for this reason, a behaviour common to all the models that we would like to point out is the exchange of position of the two curves between the 6th and 7th generation, a crucial point in which the death process become more important for the decaying of the population.

Fitting results on the original data

Here we present the plots for the best ODE model used to describe the population named "all-ped", which includes the combined data from all the available cell families.

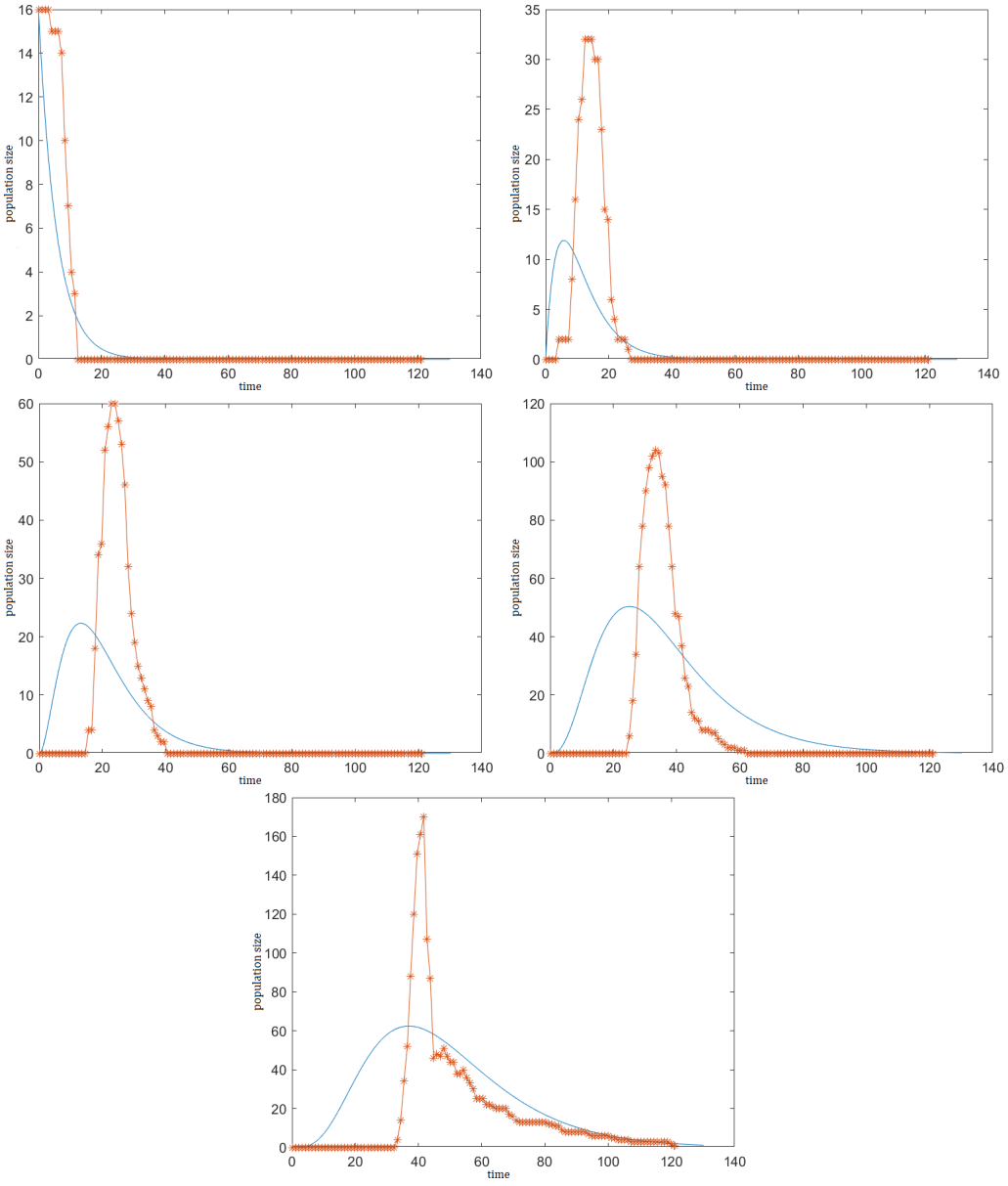


Figure 5.4: All pedigree - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

It is easy to notice that the model is much less capable of following the data dynamics, both in terms of amplitude (the simulated dynamics do not reach the values present in the data) and in terms of the timing of the signal's rise and fall, the simulated dynamics with the ODE model are in fact much slower to change trends with respect to the Cyton model.

We have placed the plots of the three pedigree groups (worst, mid, best) in Appendix D for space reasons.

Let us compare the four groups by observing the trends of the parameters belonging to the best ODE model for each of them.

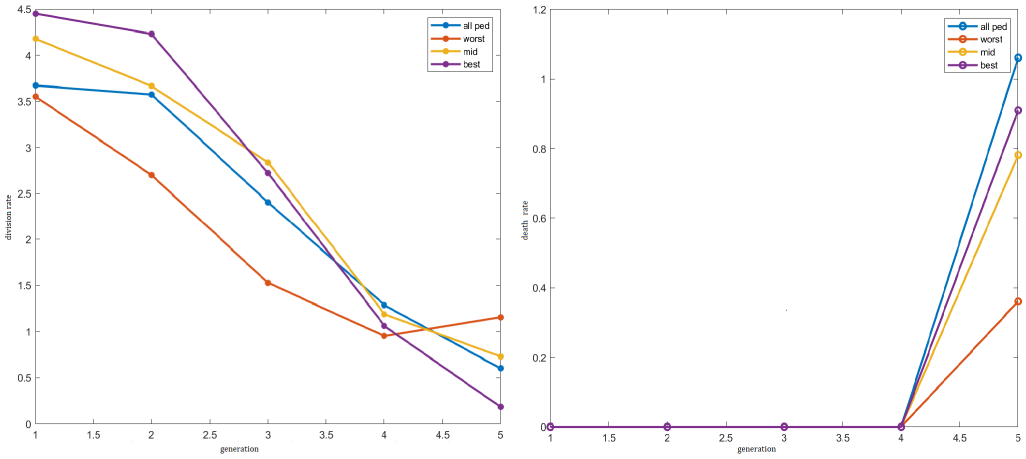


Figure 5.5: The image shows on the right the values of the division rates for each generation, on the left the death rates for each generation for the different groups of fitted pedigrees.

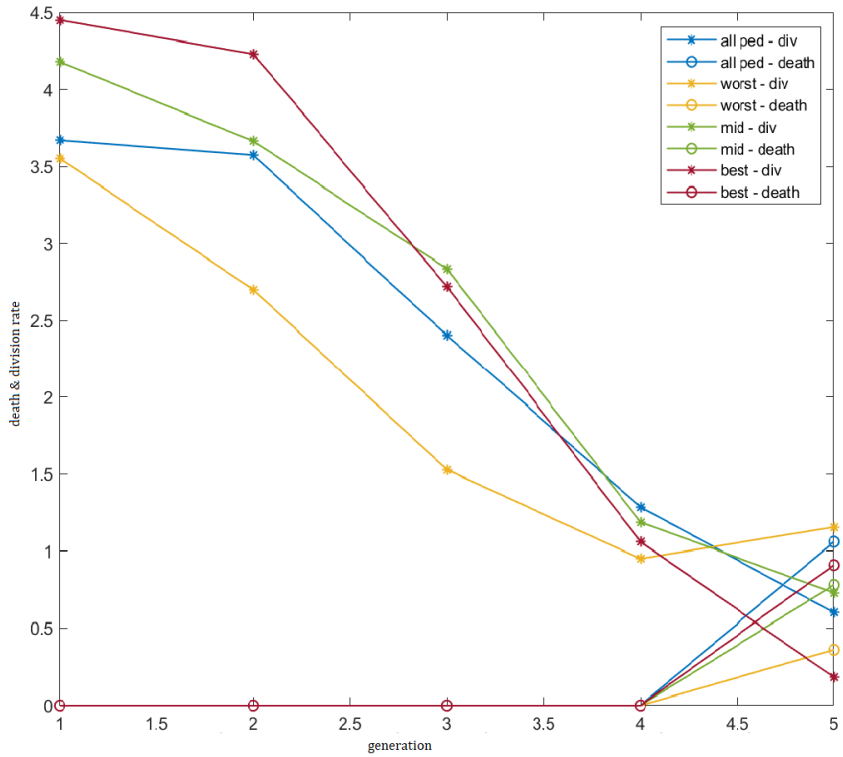


Figure 5.6: Comparison of the rates of the two processes for the various groups

It is notable that the division parameters exhibit the typical decreasing trend of reduced participation in division processes, previously modeled through the progressor fraction. Furthermore, the death rates remain at zero for the first 4 generations, taking on positive values only from the 5th generation onward. Also here we can see the generation where the two curves cross each other and the death process overcome the division one.

Chapter 6

A Stochastic model

In this chapter we present a stochastic agent-based model we implemented with the aim of reproducing in silico the behavior observed experimentally in a cell population.

To achieve this, we based the stochastic model on a birth-death process, with a structure similar to that of the Cyton and ODE models, as illustrated in the figure 3.2 in Chapter 3. Indeed, by taking this stochastic model to the continuous limit, one should obtain the results of the ODE model.
The structure of the model considers two probability values, one for the death phenomenon and one for the division phenomenon, for each of the generations to which a cell can belong, similar to what is done in the ODE model. Implementing this model, we adopted one of the assumptions from the Cyton model, where the probability of death varies with the generations following a Gaussian distribution centered on the fifth generation, while the division probability remains constant through all the generations.

At the start of the simulation, the cells are in generation 0. At each time interval ΔT, a random number is drawn for each living cell at that moment. This number determines the behavior of each cell during that time interval, defining which state transition the system will undergo. If the number drawn

satisfies a certain condition (i.e. rng $<$ div_rate), the cell duplicates; if it satisfies another similar condition, the cell dies; if neither condition is satisfied, the cell remains inactive for that time step.

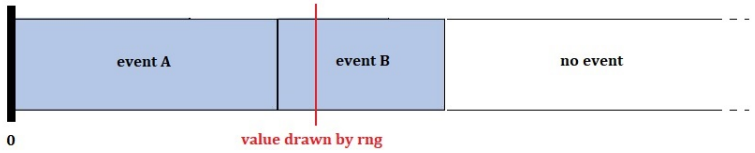


Figure 6.1: In this figure is shown the process explained before, in which a random number is extracted by an RNG algorithm and, depending from the region where it belongs to, a different event is chosen.

As in other models, the duplication of a cell results in its disappearance from the current generation and the appearance of two new cells in the next generation (see figure 1.2 of clonal expansion).

We use this structure to evaluate whether a stochastic agent-based model can reflect the variability observed in the data, in terms of differences between the dynamics.

6.1 Results

We performed stochastic simulations aiming to replicate the behaviors observed in both the benchmark data and our original data.
For our data, we specifically aimed to closely approximate the simulation to the group containing data from all the pedigrees at our disposal, effectively representing the entire cell population.
The initial results obtained from the stochastic models set up in this way show population dynamics with a trend opposite to the right-skewed bell curve that we would expect from this type of process. This discrepancy is likely due to the

relationship between the division and death rates, which takes the following form:

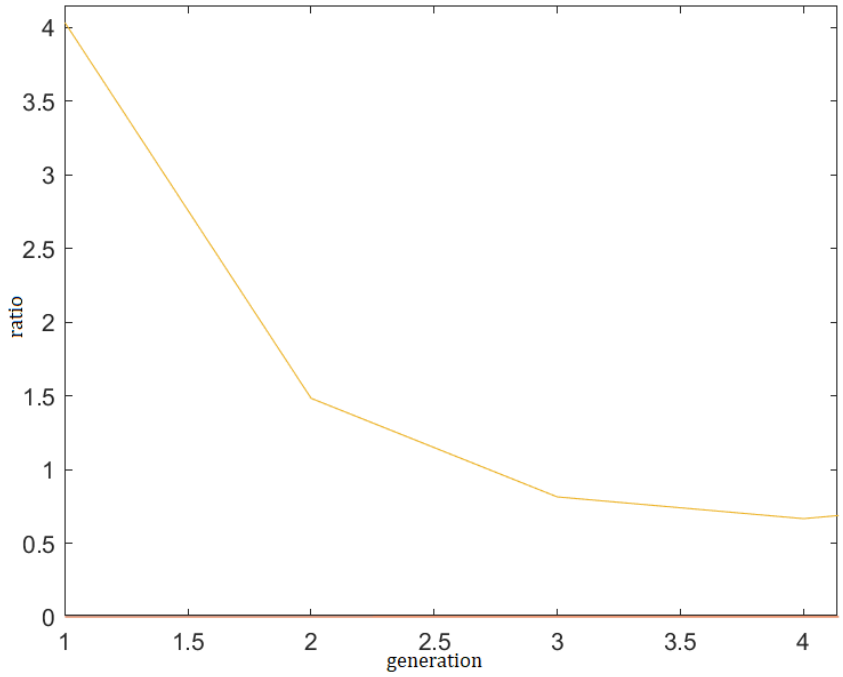


Figure 6.2: Plot of the ratio between division and death rates across generations, with a constant division rate and a death rate that follows a Gaussian distribution.

We then attempted to "invert" the Gaussian distribution representing the death rates by applying a "ones' complement" transformation ($Rate = 1 - Gaussian$). This results in an increasing trend in the ratio between the rates, which is the opposite of the previous trend.

At this point, the simulations of the resulting population dynamics better reflect the expected trends, as shown in the following figure.

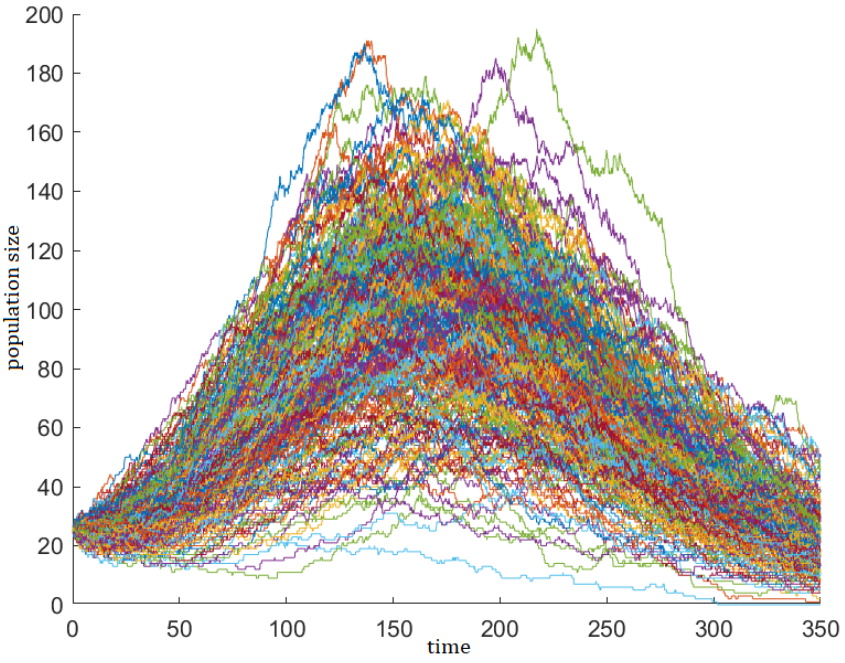
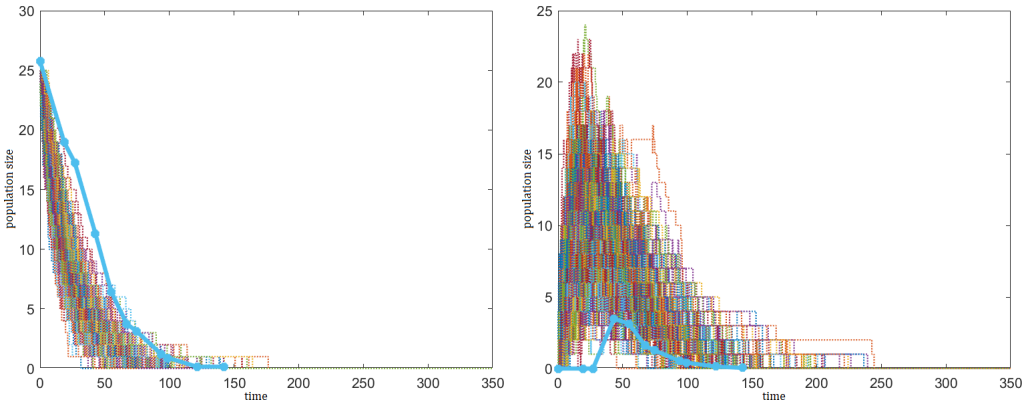
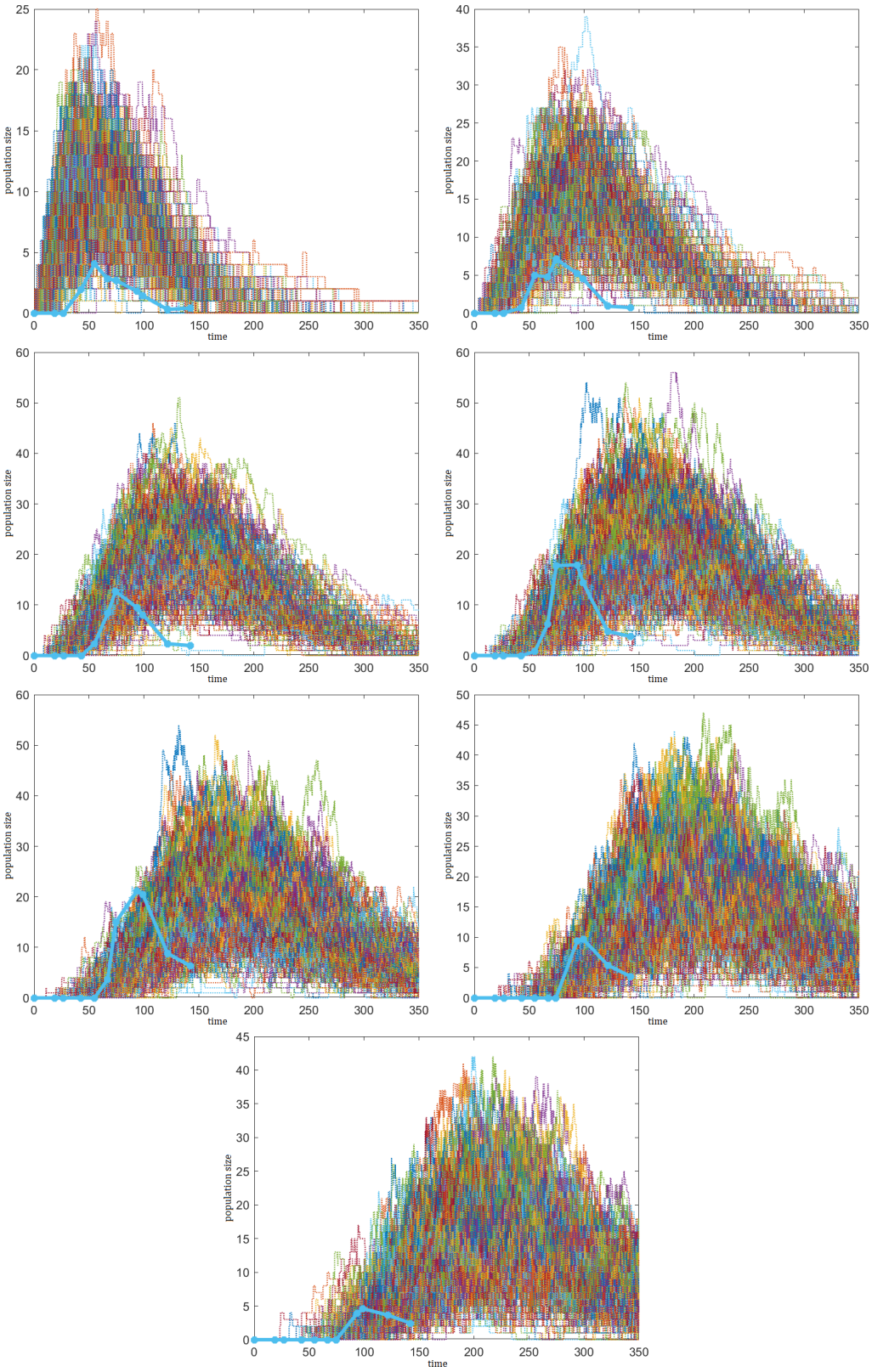


Figure 6.3: Plot of the total population for 250 independent stochastic simulations. The characteristic bell-shaped skewed form, typical of biological clonal expansion phenomena, can be observed.

We then proceeded to compare the simulated dynamics generation by generation, first with the benchmark data and then with our original data.

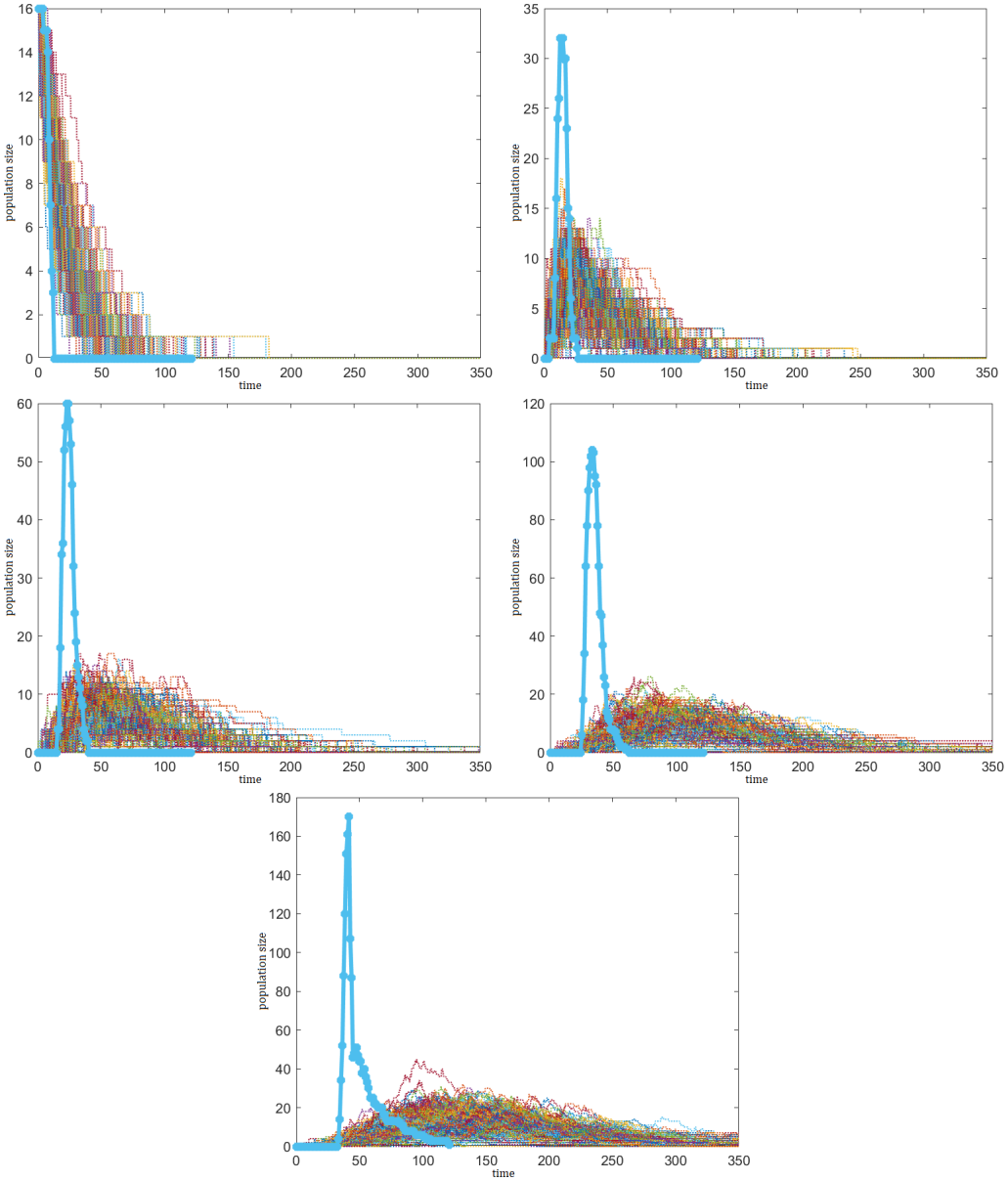
Benchmark data




population size
time
0
50
100
150
200
250
300
350

In these graphs, it can be observed that although the dynamics shown by the data are overestimated by the simulations, the timing of growth and the shape of the curves for each generation to some extent resemble those measured. The data used generally fall within the range of possible simulations, though they are often far from the average. Additionally, the typical skewed shapes with increasingly delayed growth as the generations progress can be seen.

Comparing Kajal's data



The opposite situation occurs when comparing the simulations with our original data. In the data, the dynamics are significantly faster than those simulated (an effect that was already observed for the ODE model but is now even more pronounced). The simulations, however, extend over much longer periods of time and fail to reach the high amplitude peaks observed in our data. The model is not able to mimic the process we observe in the experiments.

Chapter 7

Conclusions and open questions

We feel confident to say that our procedure, which involves training the model on the entire pedigree population and subsequently on subdivisions of this population, such as the worst, mid, and best groups, represents an effective method for extracting "objective" and concentrated data from the "broad" dynamics of population curves, provided that the fitting is performed correctly. The first step aims to perform a fit under conditions as similar as possible to those of other published experiments that use the Cyton model on very large populations. This approach allows for a quicker comparison based on quantitative components that serve as proxies for the qualitative differences in the observed dynamics.
Indeed, using this method in combination with more detailed data on the processes analyzed can help highlight the differences in data gathered from processes with different characteristics.

We did not gain significant biological insights from our simulation experiments, as our objective was to compare how different models could effectively describe the data regarding immune cell populations during their clonal expansion phase. It was found that the Cyton model, thanks to its greater adaptability, was able to closely reproduce the dynamics observed in the data, which is due to its extremely high degrees of freedom. The pure ODE model,

without the introduction of additional constraints, is too limited to capture the highly heterogeneous dynamics observed in nature, while the stochastic model is useful for evaluating the variability of the simulations and whether the range of possibilities matches the actual measured data. It is important to highlight that there are different levels of detail for comparing simulations and real data. These can be at the level of the total population or, more specifically, by observing cell behaviors, for example by dividing them into generations or pedigree groups. Therefore, the analysis we conducted aims to be less coarse and more comprehensive in studying the phenomenon in question, although it is undoubtedly more challenging.

It was observed, following the computational experiments, that it is still not possible to create fully agnostic models (i.e., models that start solely from basic "first principles" of cellular activity) and data-driven models that can effectively simulate the behavior of cell populations observed in vivo. This is because the biology of these processes is not yet fully understood in detail. Contributing to this challenge are the difficulties in measuring and studying the details behind complex behaviors such as cell division or apoptosis, particularly phenomena like phase transitions in the cell cycle and interactions between organelles and proteins within the cell. A systemic approach to the biology of these events is therefore not yet feasible, and it is necessary to introduce knowledge in the form of top-down rules in the models to be simulated. Our mental representations of how processes function within cells can be translated into the choice of specific ad-hoc distributions for defining a phenomenon or the adoption of certain types of equations over others.

Questo è il momento di fermarsi un attimo e prendere il tempo per dire grazie a tutte le persone che mi hanno da sempre aiutato in qualche modo fino ad oggi.

Grazie a tutta la mia famiglia, per il tempo che mi avete dato, per il supporto, per le discussioni. Per avermi cresciuto con i vostri valori cher cerco di interpretare al meglio di come posso. A Mamma, Papà grazie per i valori, ciò che mi avete insegnato, per le esperienze che avete permesso io facessi, alcune con voi, altre senza; dandomi il mio tempo per crescere e capire quale strada intraprendere. Mi avete guidato, senza imporvi, sul sentiero che per voi era più giusto. A Gaia, sono entusiasta della persona che stai diventando, ti ringrazio perchè sentire che guardavi a me come un modello ha fatto scaturire un senso di responsabilità nel cercare di non deluderti, spero di esserci riuscito. Alle nonne DoDò e CriCrì, grazie per avermi spinto avanti e incoraggiato, vedervi così orgogliose di me e dei miei piccoli risultati me li ha fatti percepire più grandi e mi ha permesso di affrontarli con più energia. A zio Gigi e a Matteo e Bianca, al secondo papà che ho e i suoi piccoli; mi hai sempre accompagnato e mi hai insegnato molto ma non ho avuto spesso la possibilità di ringraziarti e dirti quanto sono grato di quelli che hai fatto per me/del nostro rapporto. A tutti quelli che non ci sono più ma con le loro azioni o indirettamente mi hanno cresciuto e aiutato.

Un enorme ringraziamento vai ai miei due relatori Federico Frascoli e Marcello Delitala che, instancabilmente, hanno dedicato tantissimo tempo nel seguirmi e indirizzarmi, proponendomi punti di vista diversi e nuovi, stimolandomi ad andare sempre un passo oltre nel ragionamento e nella comprensione. Un ulteriore grazie è per i consigli che mi avete dato al di fuori dell'aspetto puramente accademico e per l'opportunità di fare una delle esperienze più belle della mia vita.

Ai miei amici, a tutte le nostre chiacchierate notturne, senza di voi non sarei dove sono oggi, soprattutto come persona. Ad Agnese un po' mamma, un po' nonna del gruppo, per sopportarci ogni giorno e prendere comunque parte ai

nostri piani improvvisati. A Jacopo per tutte le riflessioni sempre geniali, e le esperienze che mi hanno portato a ragionare sugli altri e su di me. A Said per essere un elemento fondamentale del gruppo e per essere stato con il cuore vicino a noi in questo percorso. A Chib, grazie per il continuo ed ininterrotto confronto che ha accompagnato questo viaggio, sono felice di aver potuto percorrere ancora un pezzo di strada insieme. A Christian per la fiducia che riesci a infondere con il tuo punto di vista sempre solare e pronto a cogliere il buono in tutto. A Edoardo, Ilaria e Michele, grazie per la aver alleggerito questo viaggio a volta anche solo con due chiacchiere o una battuta, è stata una fortuna avervi al fianco durante tante peripezie. Non vedo l'ora di venirvi a trovare tutti in giro per il mondo. Un grazie va anche a tutti i ragazzi e le ragazze incontrati in questi anni al Poli: Carlo, Carlotta, Gaia, Agnese solo per nominare qualcuno ma siete molti di più, ognuno di voi mi ha dato qualcosa e ve ne sono grato.

E ovviamente un grazie è d'obbligo per tutti i ragazzi che in Australia mi hanno accolto e fatto sentire in famiglia dal primo all'ultimo momento. Grazie a Edo e Solvej per le lezioni di cucina e di vita, vi devo delle pizze e forse qualcosa di più. Grazie a mamma Dario e a Simone che hanno aperto le porte di casa loro per farmi sentire a "casa" anche a 16'000 km da Torino. Grazie a Irene, non pensavo che Natale lontano dalla famiglia sarebbe potuto essere così caloroso, so di aver trovato un'amica. E infine Riccardo, non credo di aver mai legato tanto velocemente con qualcuno, grazie per aver reso l'Australia il posto magico che mi è sembrato.

Per ultima devo ringraziare Elisa che in tutti questi anni mi ha supportato, standomi vicina nelle mie scelte, ispirandomi ad essere ogni volta la versione migliore di me. Tanti anni fa, quando eravamo piccoli e tutto questo stava solo cominciando non mi sono chiesto fin dove saremmo potuti arrivare e oggi siamo qui, vorrei andare ancora oltre con la stessa naturalezza che ci ha spinti fin ora. Sono felice di averti nella mia vita, sei meravigliosa.

References

[1] Rafi Ahmed et al. "The precursors of memory: models and controversies". In: *Nature Reviews Immunology* 9.9 (2009), pp. 662–668.

[2] Balbino Alarcón, David Mestre, and Nuria Martínez-Martín. "The immunological synapse: a cause or consequence of T-cell receptor triggering?" In: *Immunology* 133.4 (2011), pp. 420–425.

[3] Peter Ankomah and Bruce R Levin. "Exploring the collaboration between antibiotics and the immune response in the treatment of acute, self-limiting infections". In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8331–8338.

[4] Iren Bains et al. "Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output". In: *The Journal of Immunology* 183.7 (2009), pp. 4329–4336.

[5] Hélène Beuneu et al. "Visualizing the functional diversification of CD8+ T cell responses in lymph nodes". In: *Immunity* 33.3 (2010), pp. 412–423.

[6] Veit R Buchholz, Ton NM Schumacher, and Dirk H Busch. "T cell fate at the single-cell level". In: *Annual review of immunology* 34 (2016), pp. 65–92.

[7] Veit R Buchholz et al. "Disparate individual fates compose robust CD8+ T cell immunity". In: *Science* 340.6132 (2013), pp. 630–635.

[8] Eugene C Butcher and Louis J Picker. "Lymphocyte homing and homeostasis". In: *Science* 272.5258 (1996), pp. 60–67.

[9] John T Chang et al. "Asymmetric T lymphocyte division in the initiation of adaptive immune responses". In: *science* 315.5819 (2007), pp. 1687–1691.

[10] Daniel K Choo et al. "Homeostatic turnover of virus-specific memory CD8 T cells occurs stochastically and is independent of CD4 T cell help". In: *The Journal of Immunology* 185.6 (2010), pp. 3436–3444.

[11] Warren N D'Souza and Stephen M Hedrick. "Cutting edge: latecomer CD8 T cells are imprinted with a unique differentiation program". In: *The Journal of Immunology* 177.2 (2006), pp. 777–781.

[12] Courtney L Davis and Frederick R Adler. "Mathematical models of memory CD8+ T-cell repertoire dynamics in response to viral infections". In: *Bulletin of mathematical biology* 75 (2013), pp. 491–522.

[13] Rob J De Boer and Alan S Perelson. "Quantifying T lymphocyte turnover". In: *Journal of theoretical biology* 327 (2013), pp. 45–87.

[14] Raluca Eftimie, Joseph J Gillard, and Doreen A Cantrell. "Mathematical models for immunology: current state of the art and future research directions". In: *Bulletin of mathematical biology* 78 (2016), pp. 2091–2134.

[15] Carmen Gerlach et al. "Heterogeneous differentiation patterns of individual CD8+ T cells". In: *Science* 340.6132 (2013), pp. 635–639.

[16] Ronald N Germain. "T-cell development and the CD4–CD8 lineage decision". In: *Nature reviews immunology* 2.5 (2002), pp. 309–322.

[17] Amanda V Gett and Philip D Hodgkin. "A cellular calculus for signal integration by T cells". In: *Nature immunology* 1.3 (2000), pp. 239–244.

[18] Arash Grakoui et al. "The immunological synapse: a molecular machine controlling T cell activation". In: *Science* 285.5425 (1999), pp. 221–227.

[19] Edwin D Hawkins et al. “A model of immune regulation as a consequence of randomized lymphocyte division and death times”. In: *Proceedings of the National Academy of Sciences* 104.12 (2007), pp. 5032–5037.

[20] Edwin D Hawkins et al. “A single-cell pedigree analysis of alternative stochastic lymphocyte fates”. In: *Proceedings of the National Academy of Sciences* 106.32 (2009), pp. 13457–13462.

[21] Susan M Kaech and Rafi Ahmed. “Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naive cells”. In: *Nature immunology* 2.5 (2001), pp. 415–422.

[22] Susan M Kaech, E John Wherry, and Rafi Ahmed. “Effector and memory T-cell differentiation: implications for vaccine development”. In: *Nature Reviews Immunology* 2.4 (2002), pp. 251–262.

[23] Andrey Kan et al. “Stochastic measurement models for quantifying lymphocyte responses using flow cytometry”. In: *PloS one* 11.1 (2016), e0146227.

[24] Peter S Kim, Peter P Lee, and Doron Levy. “Modeling regulation mechanisms in the immune system”. In: *Journal of theoretical biology* 246.1 (2007), pp. 33–69.

[25] Ludger Klein et al. “Positive and negative selection of the T cell repertoire: what thymocytes see (and don’t see)”. In: *Nature Reviews Immunology* 14.6 (2014), pp. 377–391.

[26] Arthur L Koch. “The logarithm in biology 1. Mechanisms generating the log-normal distribution exactly”. In: *Journal of theoretical biology* 12.2 (1966), pp. 276–290.

[27] Anja Langenkamp et al. “T cell priming by dendritic cells: thresholds for proliferation, differentiation and death and intraclonal functional diversification”. In: *European journal of immunology* 32.7 (2002), pp. 2046–2054.

[28] Antonio Lanzavecchia and Federica Sallusto. “Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells”. In: *Science* 290.5489 (2000), pp. 92–97.

[29] Ha Youn Lee et al. “Simulation and prediction of the adaptive immune response to influenza A virus infection”. In: *Journal of virology* 83.14 (2009), pp. 7151–7165.

[30] Fabrice Lemaıtre et al. “Phenotypic CD8+ T cell diversification occurs before, during, and after the first T cell division”. In: *The Journal of Immunology* 191.4 (2013), pp. 1578–1585.

[31] Eckhard Limpert, Werner A Stahel, and Markus Abbt. “Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question”. In: *BioScience* 51.5 (2001), pp. 341–352.

[32] Wen-Hsuan W Lin et al. “CD8+ T lymphocyte self-renewal during effector cell determination”. In: *Cell reports* 17.7 (2016), pp. 1773–1782.

[33] Shishi Luo et al. “The impact of host immune status on the within-host and population dynamics of antigenic immune escape”. In: *Journal of The Royal Society Interface* 9.75 (2012), pp. 2603–2613.

[34] Julia M Marchingo et al. “Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion”. In: *Science* 346.6213 (2014), pp. 1123–1127.

[35] Julia M Marchingo et al. “T-cell stimuli independently sum to regulate an inherited clonal division fate”. In: *Nature communications* 7.1 (2016), p. 13540.

[36] Janko Nikolich-Žugich, Mark K Slifka, and Ilhem Messaoudi. "The many important facets of T-cell repertoire diversity". In: *Nature Reviews Immunology* 4.2 (2004), pp. 123–132.

[37] Joseph Reynolds et al. "Mathematical model of naive T cell division and survival IL-7 thresholds". In: *Frontiers in Immunology* 4 (2013), p. 434.

[38] Koen Schepers et al. "Dissecting T cell lineage relationships by cellular barcoding". In: *The Journal of experimental medicine* 205.10 (2008), pp. 2309–2318.

[39] Raz Shimoni. "TACTICS for bioimaging informatics and analysis of T Cells". In: *Swinburne University of T,(ed)* (2014).

[40] Timothy K Starr, Stephen C Jameson, and Kristin A Hogquist. "Positive and negative selection of T cells". In: *Annual review of immunology* 21.1 (2003), pp. 139–176.

[41] Christian Stemberger et al. "A single naive CD8+ T cell precursor can develop into diverse effector and memory subsets". In: *Immunity* 27.6 (2007), pp. 985–997.

[42] Marianne JB van Stipdonk, Edward E Lemmens, and Stephen P Schoenberger. "Naive CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation". In: *Nature immunology* 2.5 (2001), pp. 423–429.

[43] Arezo Torang, Paraag Gupta, and David J Klinke. "An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets". In: *BMC bioinformatics* 20 (2019), pp. 1–15.

[44] Marian L Turner, Edwin D Hawkins, and Philip D Hodgkin. "Quantitative regulation of B cell division destiny by signal strength". In: *The Journal of Immunology* 181.1 (2008), pp. 374–382.

[45] John J Tyson and Odo Diekmann. “Sloppy size control of the cell division cycle”. In: *Journal of theoretical biology* 118.4 (1986), pp. 405–426.

[46] Cameron Wellard et al. “The cyton model for lymphocyte proliferation and differentiation”. In: *Mathematical Models and Immune Cell Biology*. Springer, 2011, pp. 107–120.

[47] Kajal Zibaei. “Studying CD8+ T cell immune response using in vitro live cell imaging”. PhD thesis. Peter MacCallum Cancer Centre, 2019.

Appendix A

Plot of single generation of fitted Cyton model on benchmark data

Appendix of Chapter 4.

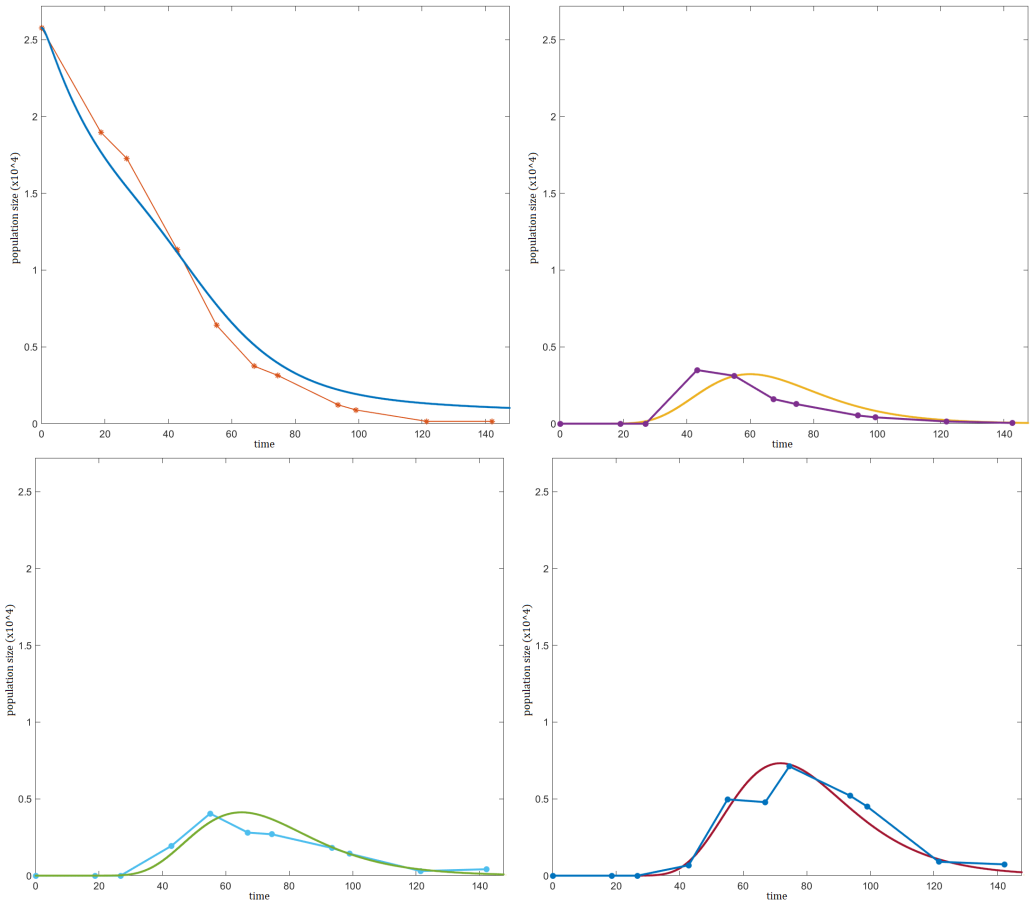


Figure A.1: Data and fitted curve referring to the Cyton model of the first four generations (0-3) of the population from benchmark data. The smooth curves represent the best model obtained from fitting.

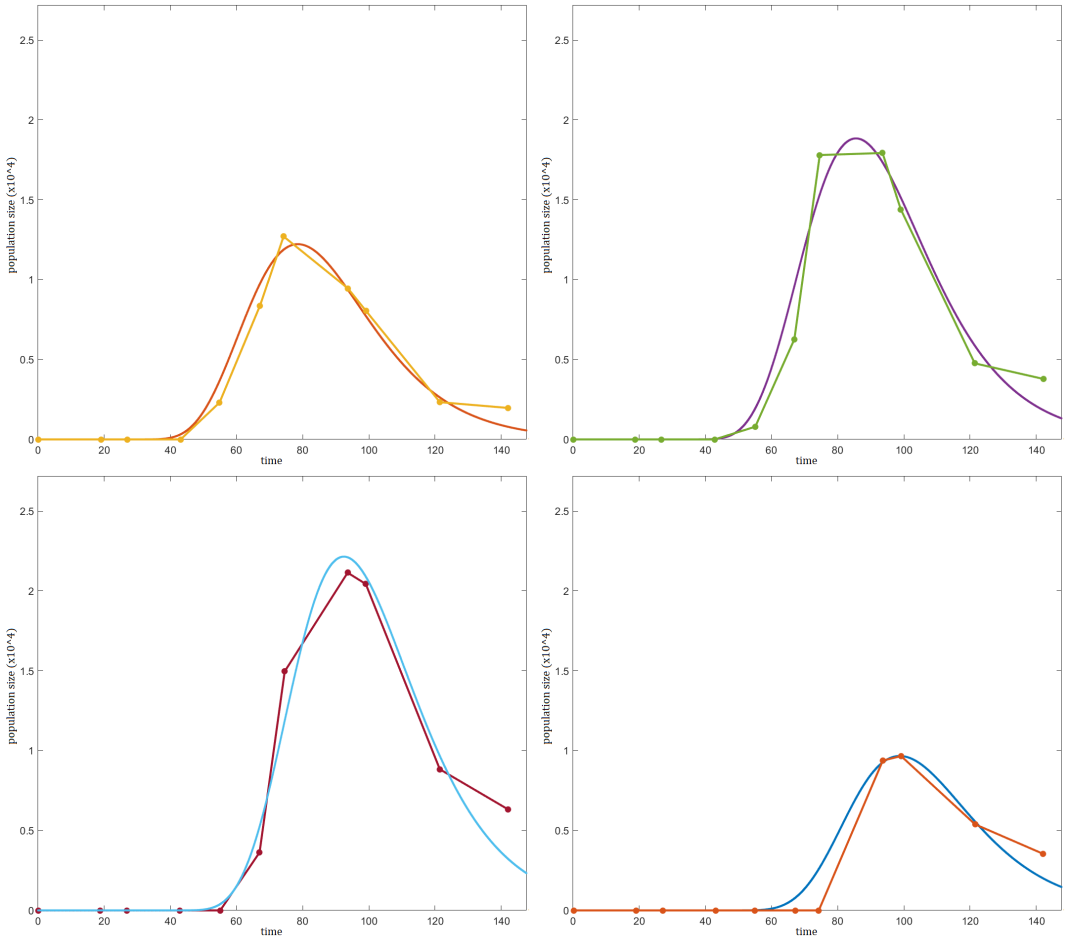


Figure A.2: Data and fitted curve referring to the Cyton model of the last four generations (4-7) of the population from benchmark data.

Appendix B

Plot of single generation of fitted Cyton model on original data

Appendix of Chapter 4.

Plot of the fit on to the best pedigree group

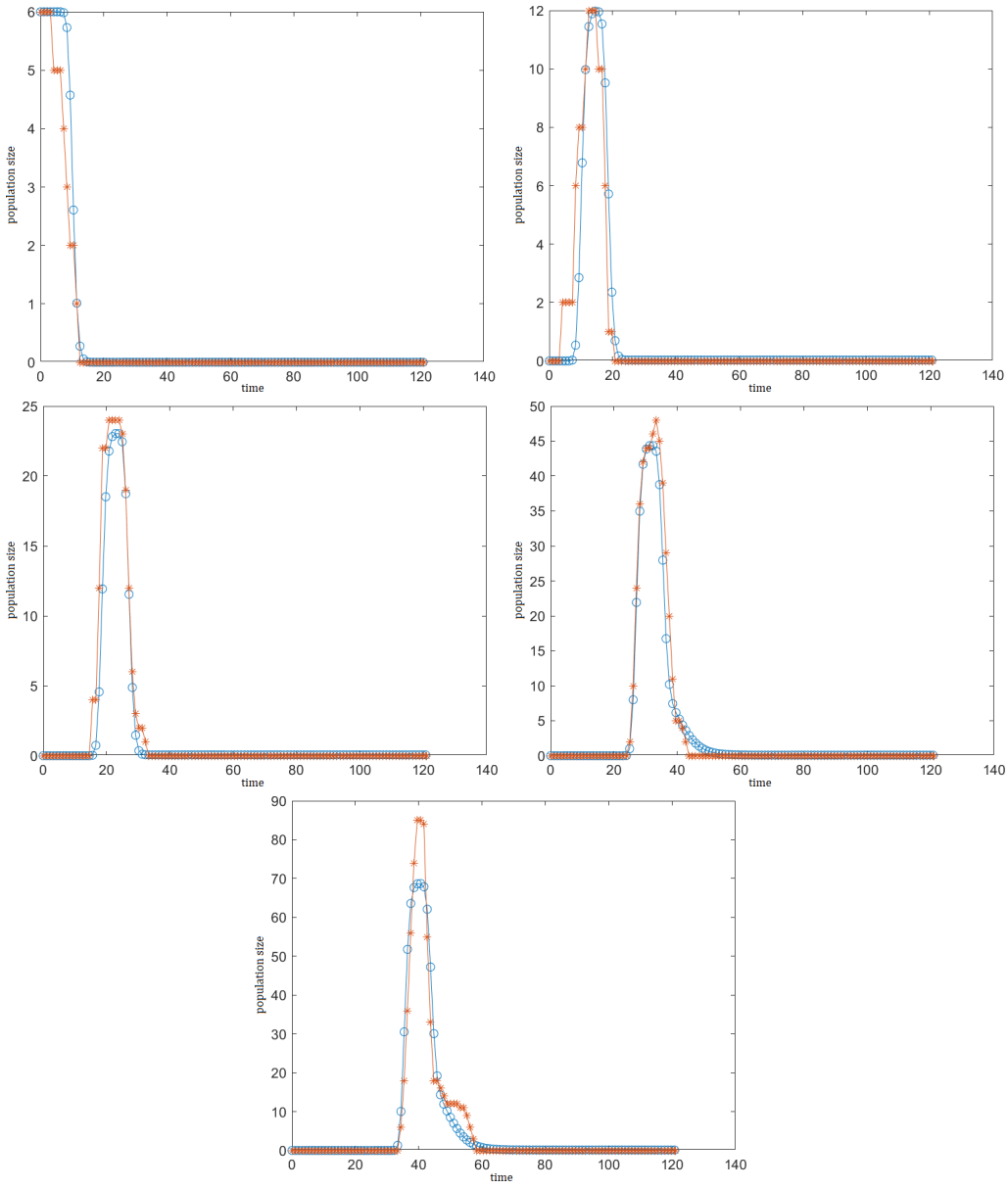


Figure B.1: Best Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

Plot of the fit on to the mid pedigree group

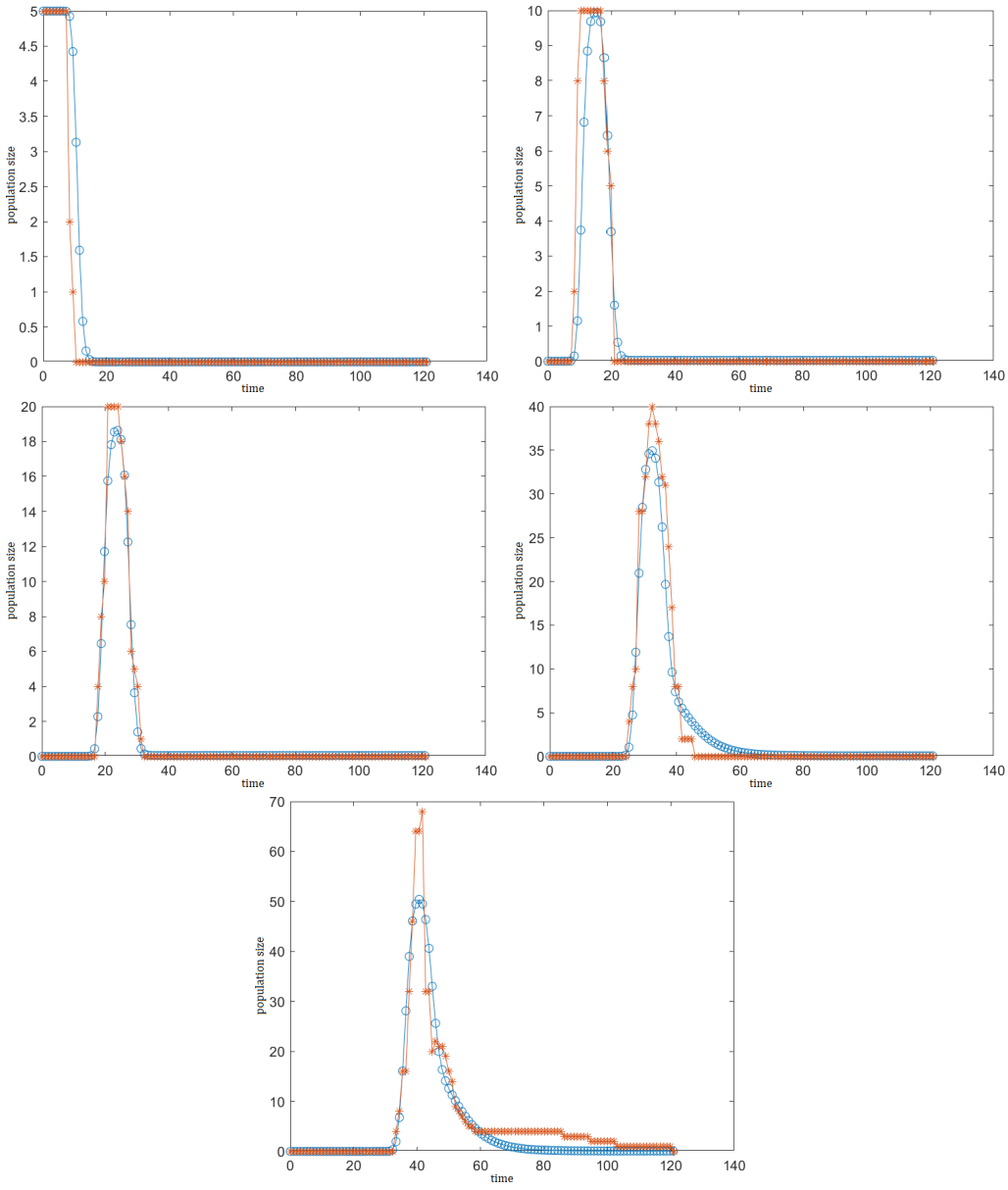


Figure B.2: Mid Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

Plot of the fit on to the worst pedigree group

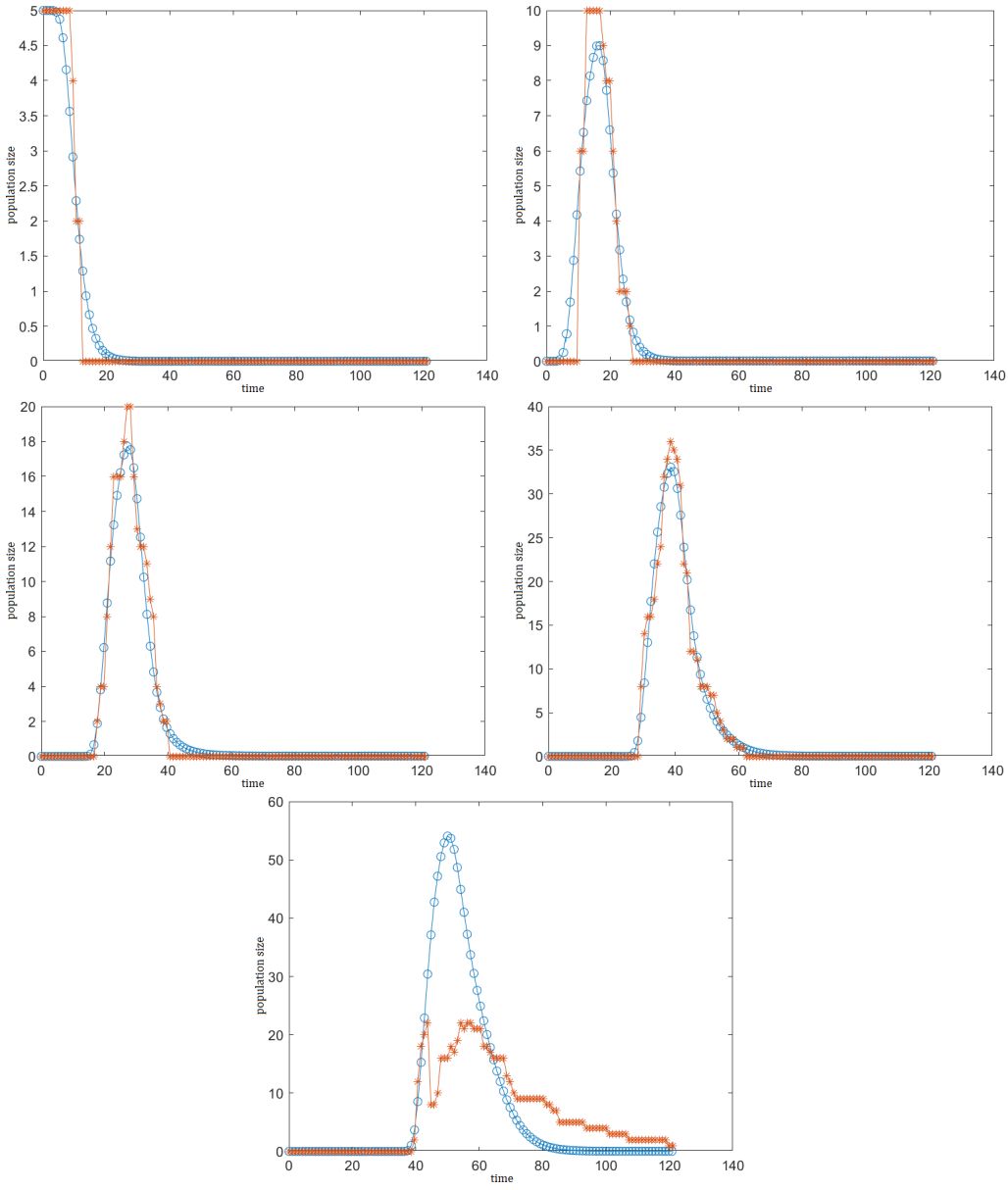


Figure B.3: Worst Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

In this last group is clearly showed how the last generation is affected by the change in the way of tracking cells, that happen in the middle of the replicative phase.

For these reason, it has been decided to not take into account the generation number 4 of this group during the fit of the cyton model. The blue curve stands for the simulated dynamics without the intervention of the modification to the data.

Appendix C

Plot of single generation of fitted ODE model on benchmark data

Appendix of Chapter 5.

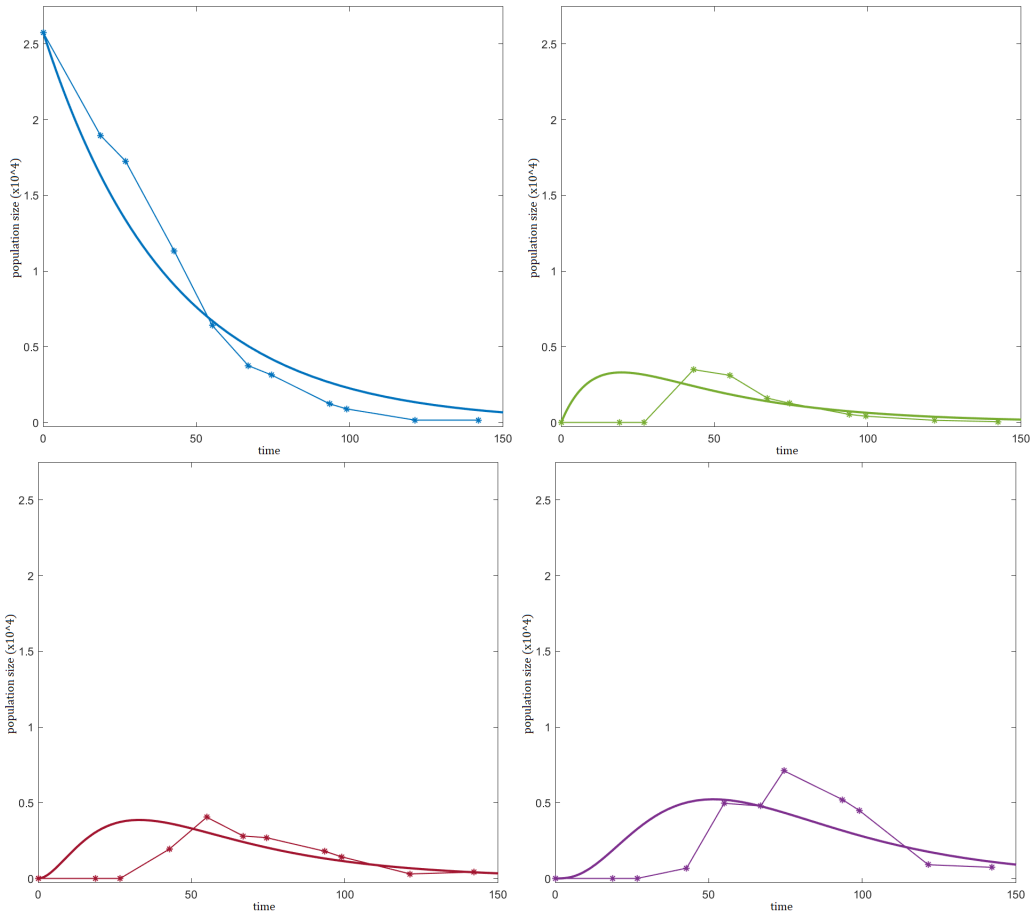


Figure C.1: Data and fitted curve referring to the ODE model of the first four generations (0-3) of the population from benchmark data. The smooth curves represent the best model obtained from fitting.

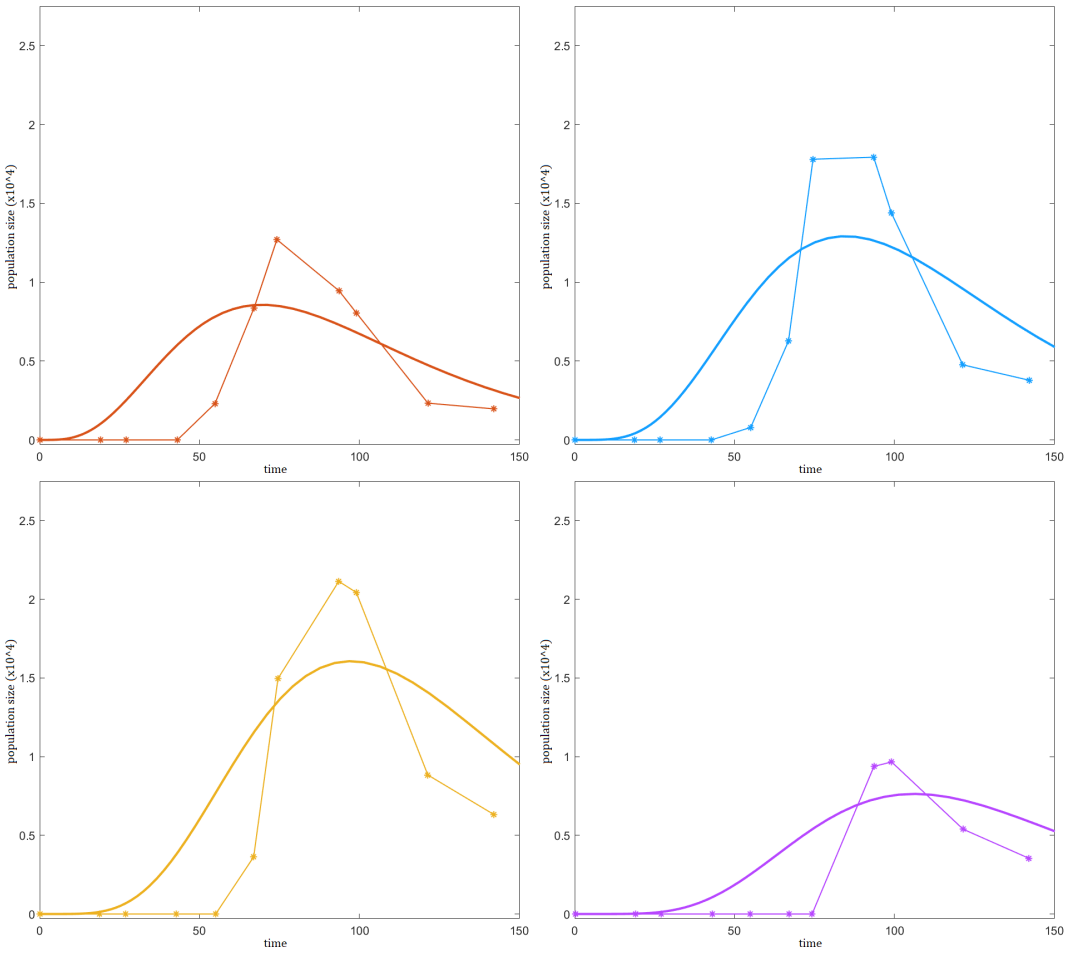


Figure C.2: Data and fitted curve referring to the ODE model of the last four generations (4-7) of the population from benchmark data. Note how different the rise and fall of the curves of the ODE model are from the real data.

Appendix D

Plot of single generation of fitted ODE model on original data

Appendix of Chapter 5.

Plot of the fit on to the best pedigree group

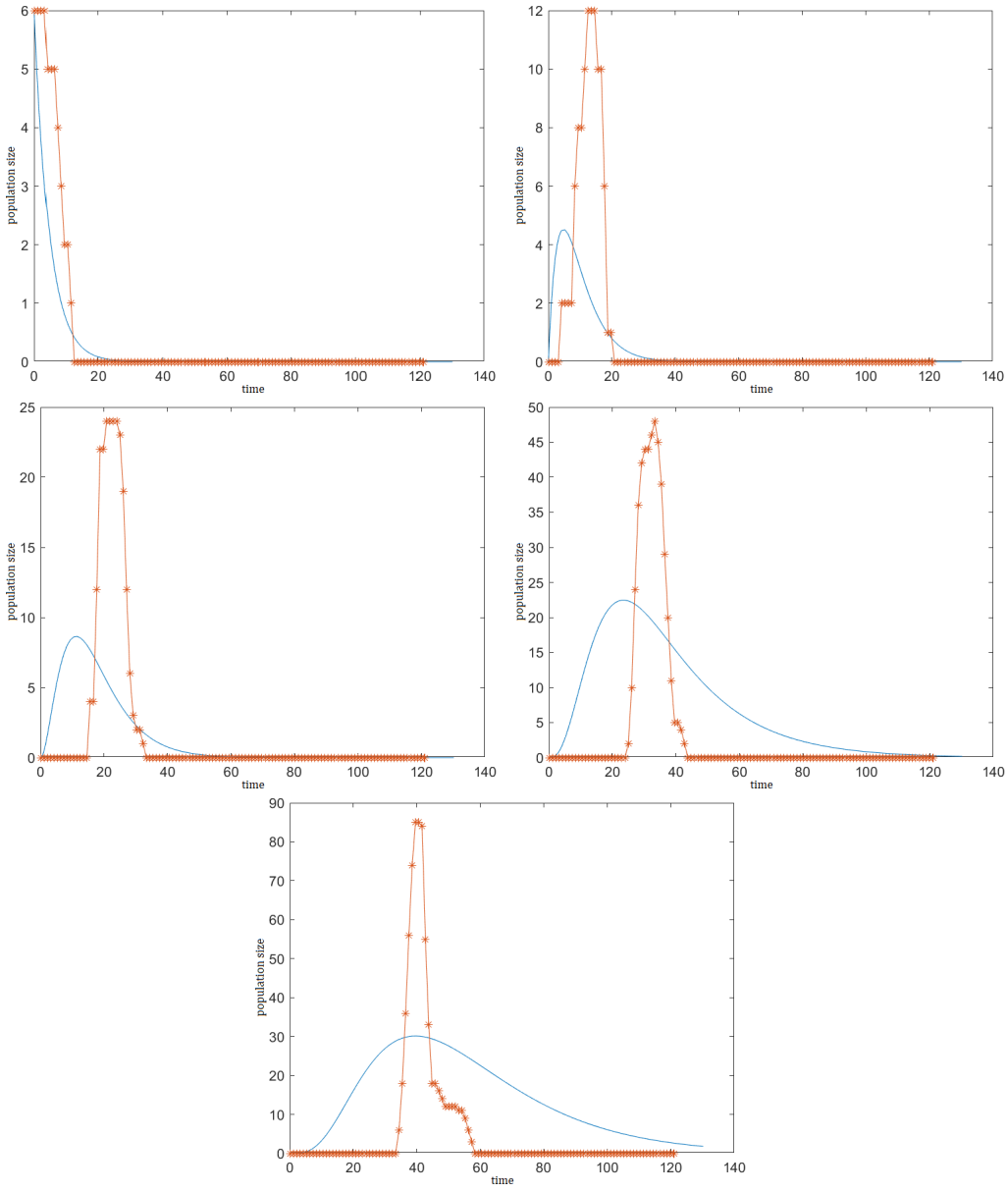


Figure D.1: Best Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

Plot of the fit on to the mid pedigree group

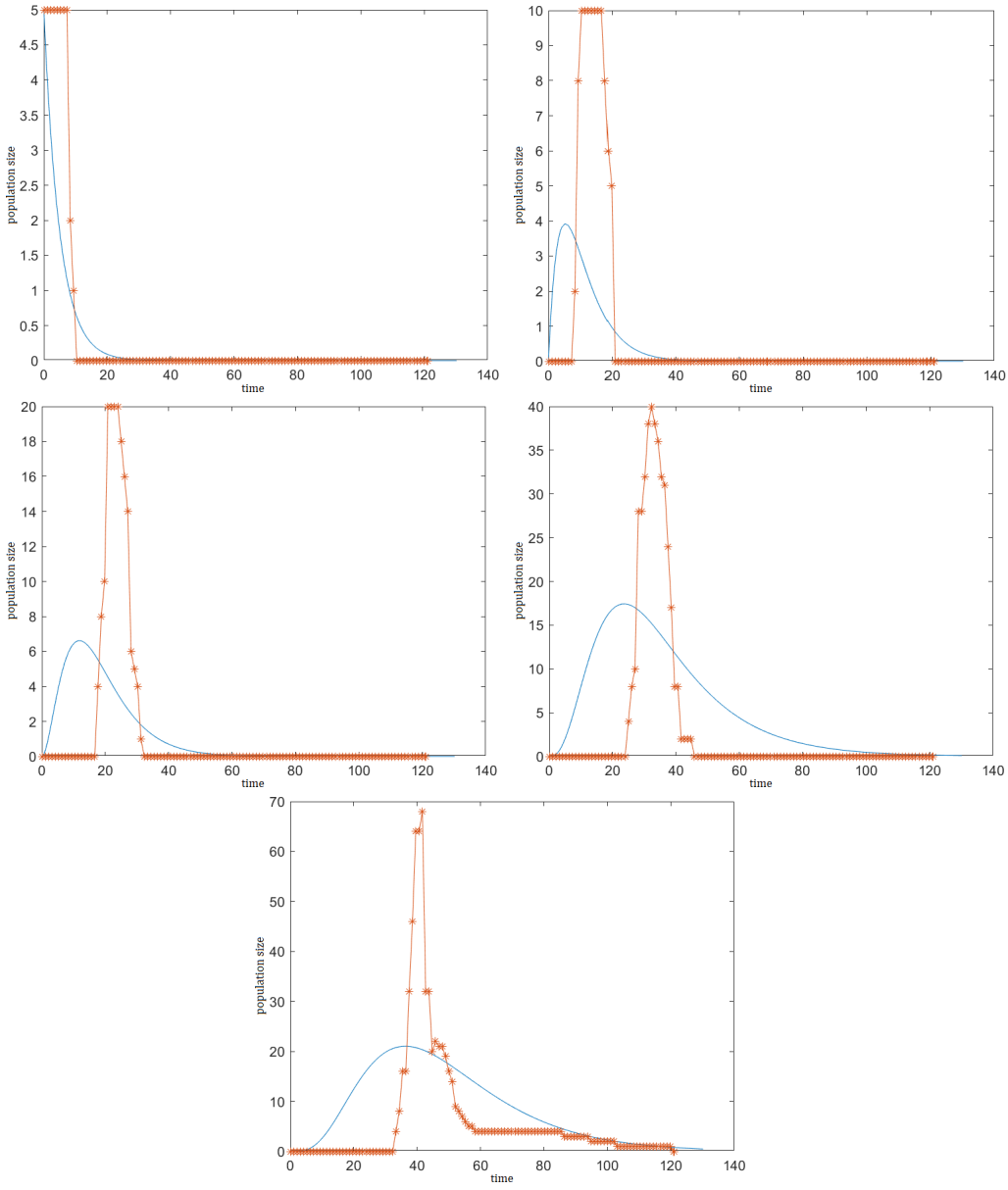


Figure D.2: Mid Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

Plot of the fit on to the worst pedigree group

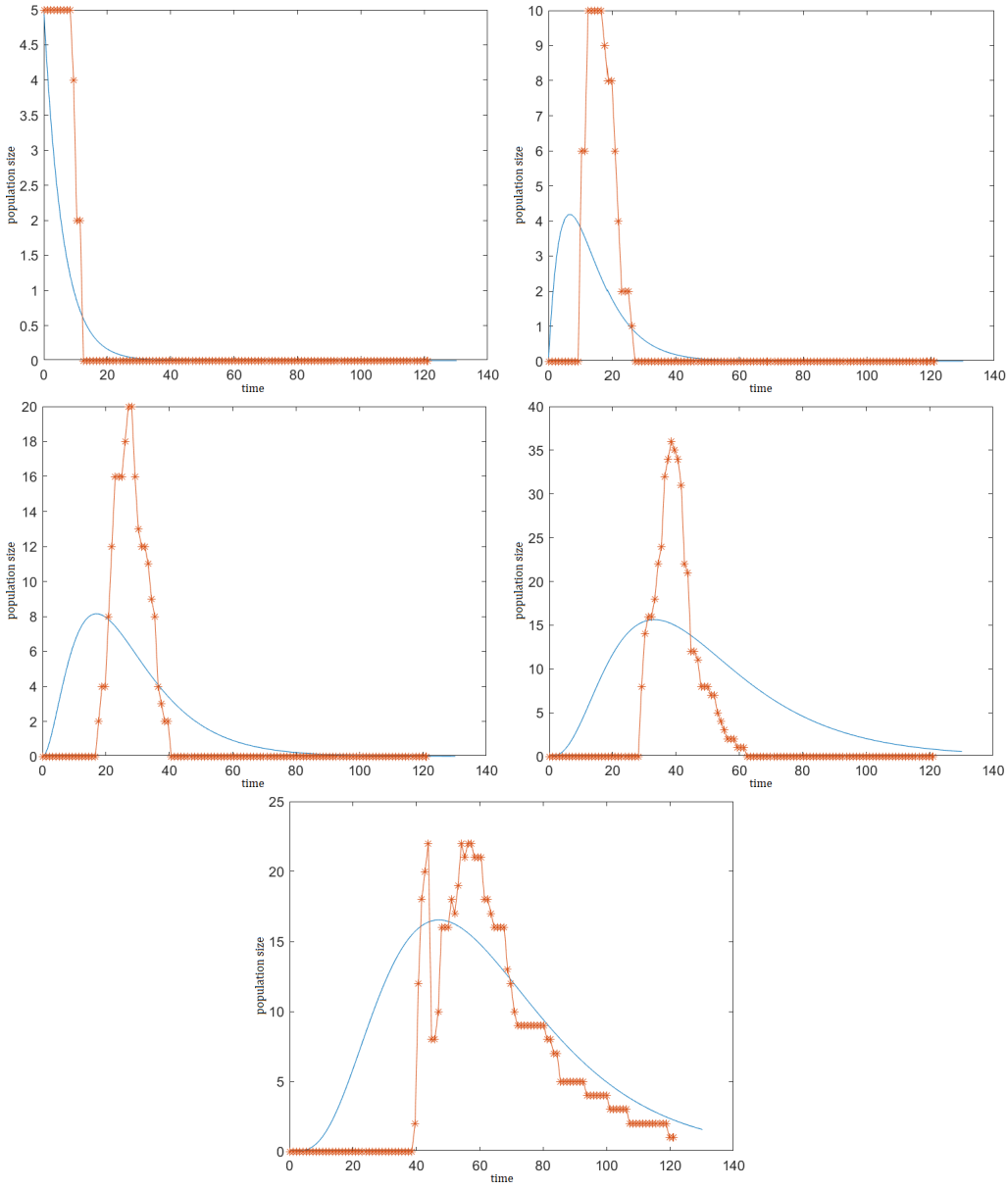


Figure D.3: Worst Pedigrees - Plots for generations 0 to 4, where the real data points are shown in red, and the simulated trends using the best-fit parameters for the ODE model are displayed in blue.

As discussed in the main text, the generation 4 is influenced by the issue with "censored" data.

Appendix E

Example of fitting pipeline

A graphical presentation of the process and progression of a Cyton model fitting pipeline is presented.

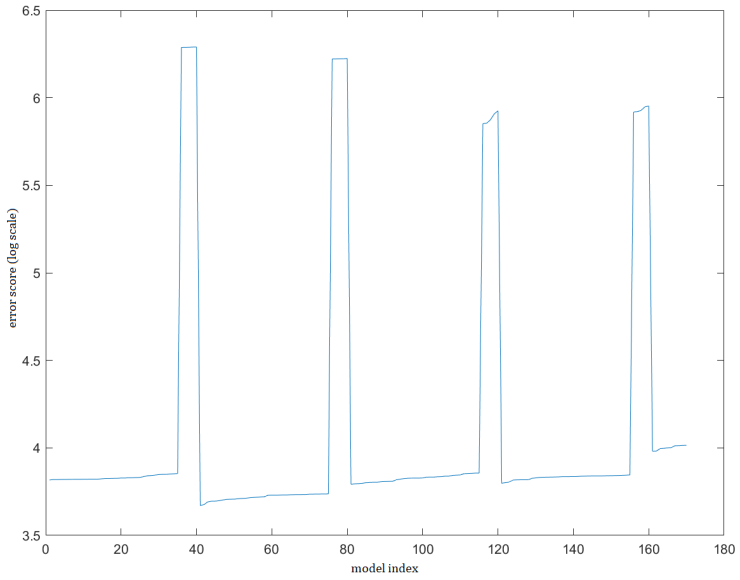


Figure E.1: Logarithmic scale score of the points resulting from the grid search, which will be used as the starting points for the stochastic optimization procedure.

After performing the 4 grid-search and selecting 40 points from each (the 35 best and 5 worst evaluated), we proceed to the first step of the stochastic optimization. This involves optimizing the parameters of the probability dis-

tributions in two variants. The first variant optimizes 3 different distributions (death gen 0+, division gen 0, division gen 1+), while the second variant optimizes 4 distributions (splitting the death process into two distributions: gen 0 and gen 1+).

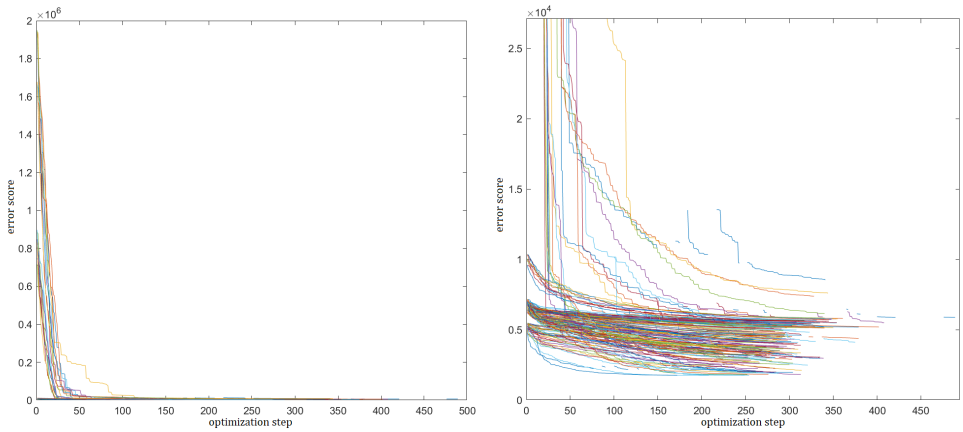


Figure E.2: Dynamics of the scores during the optimization process for each starting point, on the right a zoom on the scores below the value 10^4.

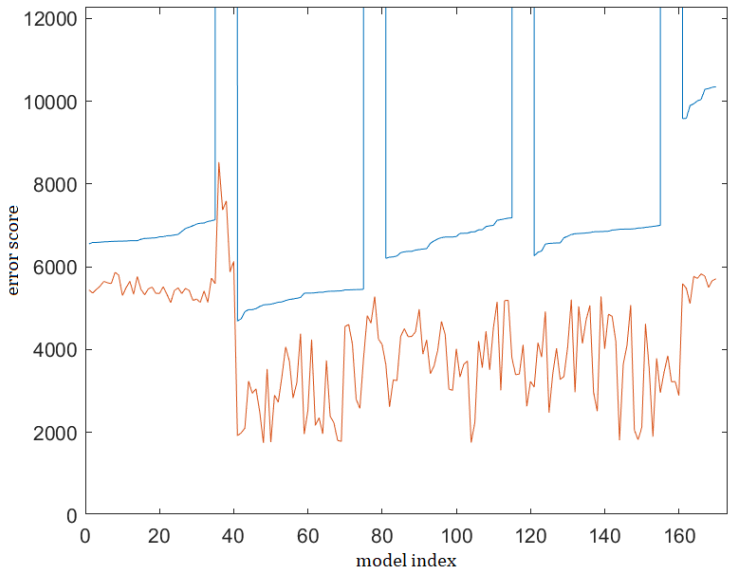
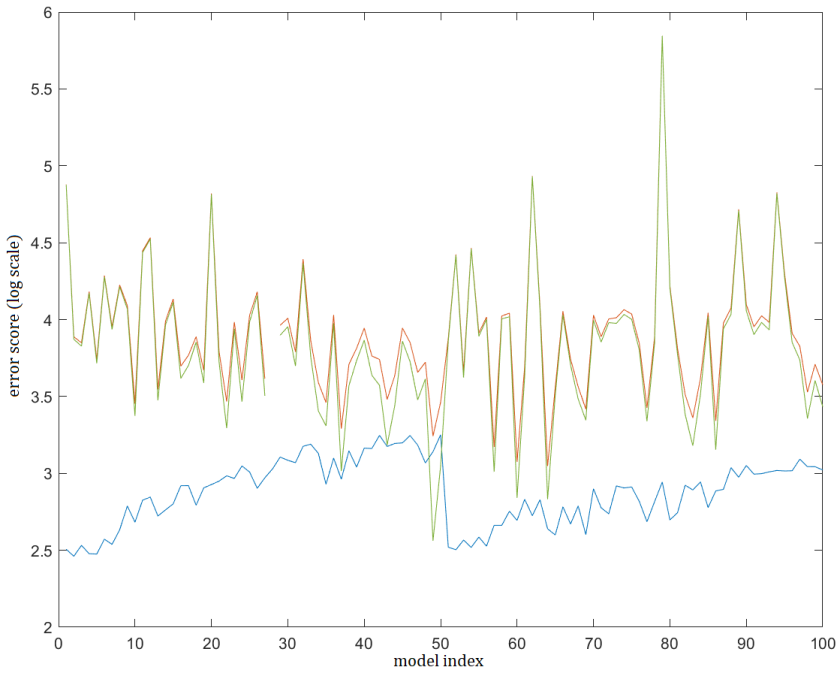


Figure E.3: Comparison of the scores between the starting points of the stochastic optimization, obtained from the grid-search, and the scores of the points post-optimization with 4 distributions.

It can be observed that the optimization leads to an improvement for each starting point, regardless of the initial score. Moreover, the resulting scores are typically within the 2000-5000 interval.

Subsequently, the second step of optimization is performed, aiming to find the optimal values of the progressor fraction γ_i for each generation.



In the figure above, a comparison is shown (on a logarithmic scale due to the wide range of the interval) between the scores of the models at the end of the optimization (in blue) and the scores of the same models after applying a variation to the parameters (in red). This variation ranges from 0 to 10% of the parameter's value.
In green, the logarithm of the difference between the two scores is shown.
This procedure was conducted to evaluate the quality of the models through sensitivity analysis, aiming to understand whether the model is specific in accurately describing the data or if it is not significantly different in performance from a model with different parameters.
As can be observed, the variations imposed on the model cause the score to

change by several orders of magnitude, demonstrating the high specificity of the obtained models.

Appendix F

Code repository

The entire Matlab code implemented for this project and used to produce the described results has been archived in the GitHub repository:
https://github.com/AStocchi/Tesi_cyton-model_codes