THE UNIVERSITY
of EDINBURGH

Applications of Branching Processes to Cancer Evolution and Initiation

Michael David Nicholson

Doctor of Philosophy
The University of Edinburgh
June 2018

Lay Summary

The generation of diversity is a hallmark of biological systems. In systems causing disease, this diversity can have dire consequences. To understand this universal feature, we must consider the processes that created it. Due to the extraordinary complexity of biological systems, quantitative reasoning is required. In this thesis, we adopt a mathematical approach and consider how diversity is generated for idealised, conceptual models, developed under the motivation of describing cancer. First, we characterise the expected sizes of clusters of cells, when the clusters are offshoots from an originating cellular population. Second, for a population of cells which can replicate and evolve, we enquire as to when, and how, a particular type of cell emerges. The aim is to capture the essence of the biological processes in our idealised models, understand the key quantities in our mathematical world and then to reinterpret these in terms of our original, biological scenario.

Abstract

There is a growing appreciation for the insight mathematical models can yield on biological systems. In particular, due to the challenges inherent in experimental observation of disease progression, models describing the genesis, growth and evolution of cancer have been developed. Many of these models possess the common feature that one particular type of cellular population initiates a further, distinct population. This thesis explores two models containing this feature, which also employ branching processes to describe population growth.

Firstly, we consider a deterministically growing wild type population which seeds stochastically developing mutant clones. This generalises the classic Luria-Delbrück model of bacterial evolution. We focus on how differing wild type growth manifests itself in the distribution of clone sizes. In our main result we prove that for a large class of wild type growth, the long-time limit of the clone size distribution has a general two-parameter form, whose tail decays as a power-law.

In the second model, we consider a fully stochastic system of cells in a growing population that can undergo birth, death and transitions. New cellular types appear via transitions, examples of which are genetic mutations or migrations bringing cells into a new environment. We concentrate on the scenario where the original cell type has the largest net growth rate, which is relevant for modelling drug resistance, due to fitness costs of resistance, or cells migrating into contact with a toxin. Two questions are considered in our main results. First, how long do we wait until a cell with a specific target type, an arbitrary number of transitions from the original population, exists. Second, which particular sequence of transitions initiated the target population. In the limit of small final transition rates, simple, explicit formulas are given to answer these questions.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in [68].

(*Michael David Nicholson, June 2018*)

Acknowledgements

Firstly, I thank my supervisor Tibor Antal, who has provided encouragement and guidance, as well as creating a relaxed working environment where passion and interest is the norm. I am also grateful to my second supervisor Bartek Waclaw, for continually attempting to inject realism into my work, offering biological inspiration and general advice.

Both my learning and enjoyment have been greatly enhanced by discussions with my fellow students, David Cheek and Stefano Avanzini. Similarly, the population dynamics, and statistical physics and complexity groups in physics, and the probability working seminar in mathematics have provided informal settings with which to receive feedback and learn, which has significantly improved my PhD experience.

I would not have been able to reach this point without my friends and family. In particular the belief and support I have received from my parents. Finally, I thank my wife, Rachael, for her support, encouragement and helping make these last 4 years an amazing time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

This thesis is concerned with populations of cells which seed further cellular populations. The meaning of 'seed' is context dependent, but examples are genetic mutations initiating a drug resistant subpopulation or cells migrating into a new environment. Mathematical models to study such a scenario abound [1, 13, 66], and have been applied to numerous settings including the emergence of drug resistance [14, 57], inference on genomic data [16, 91], the initiation of metastasis [25, 37] or microlesions [69] and the impact of poor drug penetration [33, 36, 65]. Motivated by cancer, and bacterial and viral infections, we will focus on the case where the populations under consideration grow, by which we mean, increase in number over time.

Perhaps the prototypic biological example for the initiation of a growing population from another is found in the Nobel prize winning work of microbiologist Salvador Luria and theoretical physicist Max Delbrück [62]. Due its relevance on this thesis, its simplicity and its profundity, we shall discuss this example in some detail. The motivation for Luria and Delbrück's work was discerning the manner in which mutations arose in bacterial populations. At the time, it was known that if a large population of bacteria (say 10^9 cells) was mixed with a bactericidal toxin, a fraction of the bacteria often survived. Further, the offspring of the surviving bacteria remained resistant, in the sense that they could survive further exposure to the toxin. Two competing hypotheses existed to explain this

phenomenon. The first, known as the adaption theory, proposed that, by some unknown mechanism, upon exposure to the toxin individual bacterium could alter their state, and this new state was heritable. In essence, population variation is induced in response to the environment. The second hypothesis, known as the mutation theory, posited that from a replication event, a bacterium with the resistant state could arise spontaneously. It is worth noting that at the time it was unknown even whether bacteria possessed a genome. An accessible historical overview of these matters is given by Robbins in the foreword to [62] as part of the electronic scholarly publishing project [72].

By combining theory and experiment, Luria and Delbrück provided the first strong evidence for the mutation theory hypothesis. To do so, two idealised models were proposed for the two hypotheses. For the adaption theory it was supposed that each of the large number of bacteria, say N, independently adapted upon exposure to the toxin with a small probability p. Therefore the number of bacteria surviving exposure, known as the mutants, would be binomially distributed with mean Np. In this case, the ratio of the variance to the mean mutant number would be $(1-p)$, and so, for a small adaption probability the mean and variance are approximately equal.

For the mutation theory hypothesis the following model was proposed. Starting from an initial cell, the number of toxin sensitive bacteria (often referred to as the wild type), say $N_s(t)$, grow deterministically and exponentially. If mutation events occur at the replication of sensitive bacteria, then in a short time interval $[t, t+dt]$, the chance of a mutation occurring should be proportional to the number of sensitive cells, $N_s(t)$ and the per replication mutation probability, say ν, which is again assumed small. This defined when mutations occurred, and to complete the model they supposed the mutant population grew in a similar fashion to the sensitive cells.

The key insight is that, under the mutation theory, early mutation events are relatively unlikely (as $\nu N_s(t)$ is small), however when they occur, the mutant number will be large (as the mutant population has longer to exponentially grow). Intuitively this results in a larger variance in the mutant number. Confirming this, under their mutation model (as described in the previous paragraph), Delbrück demonstrated that the ratio of the variance to the mean number of mutants was approximately $N/\log N$ (omitting prefactors also). N is again the total population size when the toxin is added, and as it is assumed large, the variance to mean number of mutants is considerably larger than 1 (as it was under the

adaption theory model). This result will be given as a specific example in Chapter 2.

Each hypothesis gave a distinct outcome in terms of the relationship of the mean and variance of the mutant numbers, and so it remained to experimentally determine these quantities. By taking replicate cultures of bacteria, exposing these to toxins, and counting the number of surviving mutant cells, Luria obtained estimates of the mean and variance. The sample variance was found to be roughly two orders of magnitude greater than the sample mean. This provided strong evidence for the mutation theory, and resulted in a paradigm shift in the study of bacteria [77].

Calculation of the mean and variance in the number of mutants was sufficient to distinguish between the competing hypotheses in the work of Luria and Delbrück. They also established the probability that no resistant bacteria would exist and derived one of the first estimators of the mutation probability (the probability of a resistance conferring mutation per cell division). However several questions remained. The distribution of the number of mutants was unknown, as was the effect of stochastic growth, cell death and non-neutral mutations (the mutant populations grew at the same rate in [62]).

These extensions to the original Luria-Delbrück mutation theory model are not simply mathematical curiosities. For example, knowledge of the mutant distribution permits maximum likelihood estimation of the mutation probability, which, within a Luria-Delbrück formalism, is the current state of the art estimation method [74, 94]. As an aside, the event of interest need not be a resistance conferring point mutation, the probability of gene amplifications per cell replication has also been inferred using this framework [83]. While in the setting of cancer, if a tumour of a fixed size is observed, neglecting cell death leads to an underestimation of the probability of resistance. This is in turn due to an underestimation of the number of replication events.

Thus a large effort has been made to understand the mutant distribution when some or all of the extensions mentioned above are included [3, 7, 22, 46, 51, 53, 58, 60, 64, 93]. The level of rigor, techniques used, and applications of interest vary across these works. The first breakthrough came from Lea and Coulson [60] who introduced neutral, stochastic mutant growth but kept the wild type growth deterministic. The most recent complete description is found in [5] who deduced the exact solution of the mutant distribution at a finite time, while

including all the previously mentioned extensions. Interestingly, in the often biologically relevant regimes of large time/population size and small mutation rate, the mutant distribution coincides with that given by a deterministically, exponentially growing wild type population, although a random amplitude is required in the time setting [16]. The meaning of 'random amplitude' will be given later in this chapter.

Common to all the works discussed above is that on average the populations grow exponentially. There are at least two reasons for this. Firstly, in the primary applications, namely cancer and bacterial infections, growth often does appear exponential, at least initially [8, 73]. Secondly, in many models, the exponential growth arises from assuming cells behave independently, and this independence offers far greater tractability. However due to environmental restrictions, e.g. space limitation, exponential population growth cannot continue indefinitely. For the types of questions considered thus far, moving away from the assumption of exponential growth has received limited previous consideration, as discussed in [31]. Introducing environmental restrictions may be accomplished in a variety of manners, for example using logistic branching processes, but motivated by the success of the semi-deterministic approach of Lea and Coulson discussed above a first step could be to consider non-exponential but deterministic wild type growth. This motivates the work presented in Chapter 2 of this thesis.

Note that the above discussion thus far has concerned the wild type population seeding a secondary mutant population, where mutant cells are one transition from the wild type. A natural extension is when the population of interest is several transitions from the original population. This would be the scenario if our interest is in multidrug resistance, where it is often necessary to have two separate genes, coding for two target proteins, mutated. As in the setting with only two populations, a standard question is: starting with a given number of cells of the original population type, do cells of a given number of transitions exist when the total population is observed at a later time or a given size? For differing applications, the case of two transitions being required has been treated in [14, 18, 38]. Motivated by the accumulation of driver mutations in cancer, the time until an arbitrary number of transitions has occurred, when each transition increases the reproductive potential of cells was given in [27]. An efficient numerical scheme coupled with an extensive biological discussion on multidrug resistance was presented in [57] for an arbitrary number of neutral transitions.

However, the setting of transitions leading to a decrease in reproductive ability within a growing population framework has not been fully considered. Exemplar reasons for contemplating such a scenario are that acquiring resistance often comes with a fitness 'cost' [2] and, interpreting transitions as migration events, the effect of drugs on migrated, sensitive cells. When several transitions are required for a cell type of interest to occur, a natural question is in which sequence do these transitions occur? For example do cells acquire resistance to drug A then B or vice versa. For varied biological scenarios this question has been considered in [33, 41, 65]. However except for [65] only two transitions were required. We attempt to address some of the questions indicated here in Chapter 3. It is worth reiterating that we continue to focus on scenarios with population growth. Within a fixed size population, the time until an arbitrary number of fitness reducing transitions has been considered in [35, 89]. Similarly, again for an approximately fixed size population, which sequence of transitions is most likely when transitions reduce reproductive ability was computationally explored in [80], with phenotypic switching the transitions of key interest.

Our introduction thus far has focused on models arising from biological applications. However many of these models are specific examples of the class of stochastic processes known as branching processes [6]. The questions considered in this thesis, for instance the large time small mutation limit discussed above, are not standard in branching process theory. However it will be seen that the well developed asymptotic theory [6, 47] provides useful tools with which to answer biologically motivated questions.

Having provided motivation for the work presented in this thesis, we now give a brief outline, detailing the contributions.

1.2 Thesis overview and summary of results

The outline of the remainder of this work is as follows. In the subsequent section a short introduction to the linear birth death process, non-homogeneous Poisson processes and Cox processes is given. Our reasoning for doing so is that these mathematical objects are common to both central chapters of the thesis. We then turn to the main content of this work.

In Chapter 2, similar to the classic setting of Luria and Delbrück, we consider

wild type cells initiating a mutant population. It is near ubiquitous to assume exponential growth of the wild type population and we investigate how deviations from this assumption alter the process. While we continue to use the language of wild-type/mutant cells, our main motivation is that of a primary tumour (wild-type) seeding metastases (mutant clones). Due to this, differing from previous studies, our interest is in the mutant clone size distribution, which represents the distribution of metastases sizes in the cancer setting. Results regarding the clones can be mapped to the total mutant number. In our main result we prove that the large time limit of the clone size distribution has a general two-parameter form for a large class of population growth. The limiting clone size distribution always has a power-law tail, and for slower than exponential wild type growth the probability of a given clone size is inversely proportional to the clone size. We support our results by analysing a dataset on metastasis sizes, and we find that a power-law tail is more likely than an exponential one, in agreement with our predictions.

Turning to Chapter 3, as discussed above, motivated by numerous applications we consider cells in a growing population that can undergo birth, death and transitions. Transitions, which may affect the genotype or environment of the cells, result in new cellular types. Motivated by the measured fitness cost of resistance in bacteria and sensitive cells in contact with drugs, we focus on the setting where the initial cellular type has the highest reproductive ability. We answer two questions. Firstly, for a target type of interest, how long do we wait until a cell with that type exists? Secondly, what is the probability that a particular sequence of transitions led to the target type? Our answers contain simple, explicit formulas that can be used to gain biological insight. We demonstrate this by considering several applications including the effect of imperfect drug penetration and the orderings of mutations.

We finish with some concluding remarks, summarising the presented work.

1.3 Mathematical Preliminaries

As discussed above we concern ourself with modelling the development and initiation of populations. We aim to use a mathematical framework which captures the key ingredients of the systems under consideration while remaining analytically tractable. The basis for this framework throughout will be the linear

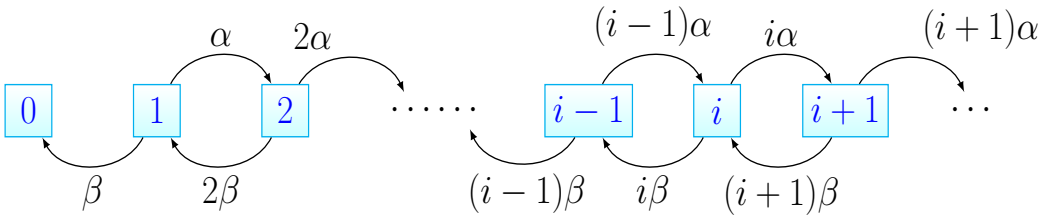


Figure 1.1 *Population level dynamics of the linear birth-death process with birth rate α and death rate β, see Section 1.3.1.*

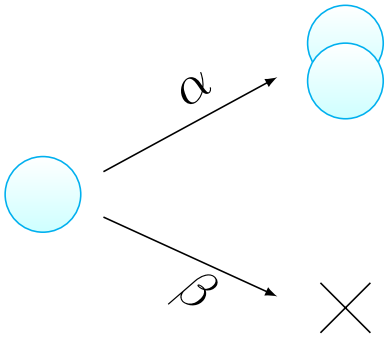


Figure 1.2 *Cell level dynamics of the linear birth-death process. Cells replicate with rate α and die at rate β, see Section 1.3.1.*

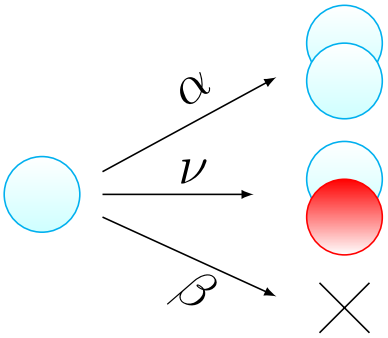


Figure 1.3 *Cell level dynamics of the linear birth-death process with mutations. Cells replicate with rate α, die at rate β and mutate at rate ν, see Section 1.3.2.*

birth-death process for population growth, and Poisson processes for initiation. Due to their recurrent use throughout this work, we now introduce these objects. Our intention is to provide a brief, informal outline. Comprehensive rigorous treatments of these objects may be found in [6, 20].

1.3.1 Stochastic growth: the linear birth-death process

The linear birth-death process $(Z_t)_{t\geq 0}$ is a one dimensional continuous time Markov process with the state space of the non-negative integers. It is one of the simplest examples of a continuous time Markovian branching process [6]. For each time t, Z_t represents the number of cells in a population. Each cell independently gives birth at a rate α and dies at a rate β. Thus, at the population level the

transition probabilities satisfy, as $\tau \to 0$,

$$\mathbb{P}(Z_{t+\tau} = j | Z_t = i) = \begin{cases} i\alpha\tau + o(\tau) & \text{for } j = i+1 \\ i\beta\tau + o(\tau) & \text{for } j = i-1 \\ 1 - i(\alpha+\beta)\tau + o(\tau) & \text{for } j = i \\ o(\tau) & \text{otherwise.} \end{cases} \tag{1.1}$$

The process at the cell level and population level is illustrated in Figures 1.2 and 1.1.

The marginal distribution starting from a single cell $\mathbb{P}(Z_t = k | Z_0 = 1)$ will be used repeatedly throughout this thesis. It is most easily found by the backward Kolmogorov equation, which can be derived using (1.1) and considering the fate of the initial cell in a small time period. Using the notation $\mathbb{P}^{(i)}(\cdot) = \mathbb{P}(\cdot | Z_0 = i)$, the backward Kolmogorov equation for the linear birth-death process is

$$\frac{d}{dt}\mathbb{P}^{(1)}(Z_t = k) = \alpha\mathbb{P}^{(2)}(Z_t = k) + \beta\mathbb{P}^{(0)}(Z_t = k) - (\alpha+\beta)\mathbb{P}^{(1)}(Z_t = k). \tag{1.2}$$

The terms on the right hand side of (1.2) correspond to the following events occurring in a small time period after $t = 0$: the initial cell replicates, the initial cell dies, and the initial cell neither replicates nor dies. Note that due to 0 being an absorbing state for the process, $\mathbb{P}^{(0)}(Z_t = k) = \delta_{k,0}$, where $\delta_{i,j}$ is the Kronecker delta function.

To obtain $\mathbb{P}^{(1)}(Z_t = k)$ from (1.2), we use the generating function for Z_t which we define as

$$\mathcal{Z}_t^{(i)}(s) = \mathbb{E}(s^{Z_t} | Z_0 = i)$$

The above definition holds for $s \in \mathbb{C}$ such that the expectation exists. The key feature of the generating function used here is that, as the populations initiated by each of the Z_0 initial cells are independent of each other,

$$\mathcal{Z}_t^{(i)}(s) = \left(\mathcal{Z}_t^{(1)}(s)\right)^i.$$

Now, multiplying both sides of (1.2) by s^k and summing over all non-negative integer k, we obtain the corresponding backward Kolmogorov equation in terms of generating functions as

$$\frac{d}{dt}\mathcal{Z}_t^{(1)}(s) = \alpha\left(\mathcal{Z}_t^{(1)}(s)\right)^2 + \beta - (\alpha+\beta)\mathcal{Z}_t^{(1)}(s). \tag{1.3}$$

As (1.3) contains only the generating function when the process is initiated from a single cell, $\mathcal{Z}_t^{(1)}(s)$, we now fix $Z_0 = 1$ and cease the use of superscripts (i.e. henceforth $\mathbb{P}(\cdot) = \mathbb{P}^{(1)}(\cdot)$ etc). Equation (1.3) can be explicitly solved and its solution, given on page 76 of [7], is

$$\mathcal{Z}_t(s) = \mathbb{E}(s^{Z_t}) = 1 - \frac{\lambda/\alpha}{1 - \xi e^{-\lambda t}}, \text{ where } \xi = \frac{\beta/\alpha - s}{1 - s}, \lambda = \alpha - \beta. \quad (1.4)$$

In this thesis λ will be referred to as the growth rate, or fitness of the population. By expanding the generating function around $s = 0$, we obtain that the probability of the population size being k follows a geometric distribution with a modified zero term

$$\mathbb{P}(Z_t = k) = \begin{cases} \frac{\beta}{\alpha \mathcal{S}_t} & k = 0 \\ (1 - \beta/(\alpha \mathcal{S}_t))(\mathcal{S}_t - 1)\, \mathcal{S}_t^{-k} & k \geq 1, \end{cases} \quad (1.5)$$

where

$$\mathcal{S}_t = \frac{1 - e^{-\lambda t}\beta/\alpha}{1 - e^{-\lambda t}}. \quad (1.6)$$

Thus the mode of the distribution is either at 0 or 1, and the mass function decays geometrically for $k > 1$. For the particular case of a critical process, i.e. when birth and death rates are equal, the above probabilities are simplified by observing

$$\lim_{\beta \to \alpha} \mathcal{S}_t = \frac{\alpha t + 1}{\alpha t}. \quad (1.7)$$

Immediately from (1.5), and the large time limit of $\mathcal{S}_t$, we see that the extinction probability is

$$\lim_{t \to \infty} \mathbb{P}(Z_t = 0) = \begin{cases} 1 & \text{for } \beta \geq \alpha \\ \beta/\alpha & \text{for } \alpha > \beta. \end{cases} \quad (1.8)$$

When $\beta = 0$ and so the stochastic proliferation follows a pure birth or Yule process, the cells will be denoted immortal.

From the generating function (1.4) the mean number of cells is obtained as

$$\frac{d}{ds}\mathcal{Z}_t(1) = \mathbb{E}[Z_t] = e^{\lambda t}.$$

We can go further and obtain a pathwise description of the growth at large times by observing that $(e^{-\lambda t} Z_t)_{t \geq 0}$ is a martingale. To show this, note that the number of cells at time $t + u$ is comprised of the descendants of all existing cells at time

t. Therefore, for $u > 0$,

$$Z_{t+u} = \sum_{i=1}^{Z_t} Z_u^{(i)}$$

where conditional on Z_t, all $Z_u^{(i)}$ are iid with distribution equal to Z_u. From this representation we have

$$\begin{aligned}
\mathbb{E}[e^{-\lambda(t+u)} Z_{t+u} | (e^{-\lambda s} Z_s)_{0 \le s \le t}] &= \mathbb{E}[e^{-\lambda(t+u)} Z_{t+u} | e^{-\lambda t} Z_t] \\
&= \mathbb{E}\left[e^{-\lambda t} \sum_{i=1}^{Z_t} e^{-\lambda u} Z_u^{(i)} \middle| e^{-\lambda t} Z_t \right] \\
&= e^{-\lambda t} Z_t.
\end{aligned}$$

Which demonstrates the martingale property. As $(e^{-\lambda t} Z_t)_{t \ge 0}$ is a non-negative martingale, the martingale convergence theorem [90] (Theorem 11.5) ensures the existence of a non-negative random variable W such that, with probability one

$$\lim_{t \to \infty} e^{-\lambda t} Z_t = W. \tag{1.9}$$

Before offering an interpretation of the above equation let us consider the distribution of W.

The distribution of W can be deduced as follows. In the case of $\beta \ge \alpha$ as extinction occurs with probability 1, then $W = 0$ with probability one also. For $\alpha > \beta$, by the continuity theorem [90] (Theorem 18.1), it is enough to consider the limit of $\mathcal{Z}_t(e^{-se^{-\lambda t}})$ which is the Laplace transform of $e^{-\lambda t} Z_t$. From (1.4) the only time dependent part of $\mathcal{Z}_t(s)$ is $\xi e^{-\lambda t}$, whose limit under the mapping $s \mapsto e^{-se^{-\lambda t}}$ is

$$\lim_{t \to \infty} \frac{\beta/\alpha - e^{-se^{-\lambda t}}}{e^{\lambda t}(1 - e^{-se^{-\lambda t}})} = \frac{\beta/\alpha - 1}{s}.$$

Thus we have

$$\lim_{t \to \infty} \mathcal{Z}_t(e^{-se^{-\lambda t}}) = \beta/\alpha + (1 - \beta/\alpha) \frac{1}{1 + s/(1 - \beta/\alpha)}.$$

If we let χ be Bernoulli$(1 - \beta/\alpha)$ and $\tilde{W}$ an Exponential$(1 - \beta/\alpha)$ variable, with χ and $\tilde{W}$ independent of each other, then the right hand side of the above is the Laplace transform of $\chi \tilde{W}$. Hence

$$\mathbb{P}(W \le x) = \mathbb{P}(\chi \tilde{W} \le x) = \beta/\alpha + (1 - \beta/\alpha)(1 - e^{-x(1-\beta/\alpha)}). \tag{1.10}$$

Therefore the interpretation of (1.9) is that there are two options for the asymptotic behavior of $(Z_t)_{t\geq 0}$ pathwise. The process goes extinct with probability β/α. If extinction is avoided, the population grows as $\tilde{W}e^{\lambda t}$. This is the meaning of deterministic growth with a random amplitude mentioned previously.

1.3.2 Stochastic initiation: non-homogeneous Poisson processes and Cox processes

We shall be interested in the initiation of populations and will use non-homogeneous Poisson/Cox processes to model the initiation times. Before introducing these processes, we give a brief motivation by modifying the birth-death process discussed above. Suppose that in addition to replicating and dying at rates α, β, cells may also mutate at a rate ν, as is illustrated in Figure 1.3. By independence, similar to (1.1), in a small time increment τ a mutation event occurs with probability $\nu Z_t \tau + o(\tau)$. Analogously to the time homogeneous Poisson process, mutation events accrue with initiation rate νZ_t. Considering this process for a fixed realisation of $(Z_t)_{t\geq 0}$ gives the motivation to introduce the time-dependent, or non-homogeneous Poisson process.

The initiation times of populations will always be assumed to form a counting process. A stochastic process $(K_t)_{t\geq 0}$ is a counting process if it takes values in $\mathbb{Z}_{\geq 0}$, $K_0 = 0$ and it is non-decreasing in t with probability one. Two particular forms of counting processes will be considered.

Suppose we have a locally integrable non-negative, real valued function, $(n(t))_{t\geq 0}$, where by locally integrable it is meant that for each closed and bounded interval $[a, b] \subset \mathbb{R}_{\geq 0}$

$$\int_a^b n(t)\, dt < \infty.$$

The above requirement is motivated by the desire to exclude infinite arrivals over a finite time interval. For notational simplicity, we will often abuse notation by referring to $n(t)$ as a function in place of $(n(t))_{t\geq 0}$. The counting process $(K_t)_{t\geq 0}$ is a non-homogeneous Poisson process with rate $n(t)$ if:

- $(K_t)_{t\geq 0}$ has independent increments.

- For each t, as $\tau \to 0$

$$\mathbb{P}(K_{t+\tau} - K_t = k) = \begin{cases} 1 - n(t)\tau + o(\tau), & k = 0 \\ n(t)\tau + o(\tau), & k = 1 \\ o(\tau), & k \geq 2. \end{cases}$$

Independent increments means that for any disjoint intervals, i.e. $[s_1, t_1]$ and $[s_2, t_2]$ with $[s_1, t_1] \cap [s_2, t_2] = \emptyset$, $K_{t_1} - K_{s_1}$ and $K_{t_2} - K_{s_2}$ are independent. The function $n(t)$ is called also the intensity of the process. The above implies that the number of arrivals in any interval $[a, b]$ is a Poisson random variable with mean $\int_a^b n(s)\, ds$. The quantity $m(a, b) = \int_a^b n(s)\, ds$ is referred to as the mean measure. In particular K_t is Poisson with mean

$$\mathbb{E}[K_t] = \int_0^t n(s)\, ds.$$

Note that the standard homogeneous Poisson process is the special case of constant $n(t)$. Here we have given the definition with domain $\mathbb{R}_{\geq 0}$, but this may be altered to arbitrary subsets of $\mathbb{R}$ (or indeed more general spaces).

Returning to our example of a birth death process with mutations, with our toolkit of non-homogeneous Poisson processes we can now describe the accrual of mutations for a fixed realisation of $(Z_t)_{t\geq 0}$. However $(Z_t)_{t\geq 0}$ is itself a stochastic process and this motivates extending our discussion to Cox processes. Cox processes are extensions of non-homogeneous Poisson processes achieved by letting the intensity itself be random. In particular if $(n_t)_{t\geq 0}$ is a random real valued function, locally integrable with probability one, then $(K_t)_{\geq 0}$ is a Cox process if, $(K_t)_{\geq 0}|(n_t)_{t\geq 0}$ is a non-homogeneous Poisson process with intensity $(n_t)_{t\geq 0}$. Here and throughout this thesis $X|Y$ should be read as X conditioned on Y.

For our running example, if we let K_t be the number of mutations that have occurred by t, then the probability of no mutations by t is

$$\mathbb{P}(K_t = 0) = \mathbb{E}[\mathbb{P}(K_t = 0|(Z_s)_{s\geq 0})] = \mathbb{E}\left[\exp\left(-\nu \int_0^t Z_s\, ds\right)\right]. \tag{1.11}$$

In the first equality the tower rule was applied and in the second we used the fact that $K_t|(Z_s)_{s\geq 0}$ is Poisson with mean $\nu \int_0^t Z_s\, ds$. Equation (1.11) will be critical in Chapter 3. Other properties of these objects shall be introduced as required.

Chapter 2

Clone Sizes for Growing Populations

2.1 Introduction and Model

2.1.1 Introduction

Cancerous tumours spawning metastases, bacterial colonies developing antibiotic resistance or pathogens kick-starting the immune system are examples in which events in a primary population initiate a distinct, secondary population. Regardless of the scenario under consideration, the number of individuals in the secondary population, and how they are clustered into colonies, or *clones*, is of paramount importance. An approach which has offered insight has been to bundle the complexities of the initiation process into a mutation rate and assume that the primary, or *wild-type*, population seeding the secondary, or *mutant*, population is a random event.

As discussed in the previous chapter, this method was pioneered by microbiologist Salvador Luria and theoretical physicist Max Delbrück [62]. In their original model both wild-type and mutant populations grew deterministically, with mutant initiation events being the sole source of randomness. Lea and Coulson [60] generalised this process by introducing stochastic mutant growth in the form of the pure birth process and were able to derive the distribution of the number of mutants for neutral mutations. This was again extended by Bartlett and later Kendall [7, 53], who considered both populations developing according to a birth process. An accessible review discussing these formulations is given by Zheng

[93].

Recent developments have focused on cancer modelling, where usually mutant cell death is included in the models. The main quantity of interest in these studies has been the total number of mutant cells. Explicit and approximate solutions appeared for deterministic, exponential wild-type growth, corresponding to a fixed size wild-type population [3, 22, 46, 51, 58], and fully stochastic wild-type growth either at fixed time or fixed size [5, 27, 54]. An exciting recent application has been to model emergence of resistance to cancer treatments [12, 14, 55]. The current study continues in this vein with our inspiration being primary tumours (wild-type) seeding metastases (mutant clones).

Interestingly, in the large-time small-mutation rate limit, the clone size distributions at a fixed wild-type population size coincide for stochastic and deterministic exponential wild-type growth [51, 54]. This was recently rigorously proved in [16]. The intuition behind this observation is that a supercritical birth-death branching process converges to exponential growth in the large time limit, and, for a small mutation rate, mutant clones are initiated at large times. So asymptotically the two methods are equivalent, but the deterministic description of the wild-type population has twofold advantages: (i) the calculations are much simpler in this case [51], and (ii) the method can be easily generalised to arbitrary growth functions. This is the programme that we develop in the present chapter.

The present work differs from previous approaches in two ways. Firstly, motivated by populations with environmental restrictions, we move away from the assumption of exponential wild-type growth, a setting which, as already noted, has received limited previous consideration [31]. Explicit solutions are given for exponential, power-law and logistic growth, which we then extrapolate from to present general results which are valid for a large class of growth functions. This extends the classic results found in [6, 50, 52, 81] and recent work in [42, 84] who considered the wild-type population growth rate to be time-dependent but coupled with the mutant growth rate. Secondly, rather than the total number of mutants, our primary interest is on the distribution of mutant number in the clones initiated by mutation events. This complements [40], which allowed deterministic wild-type and mutant growth, and the treatment of clone sizes for constant wild-type populations found in [23]. Whilst we focus on clone sizes, we demonstrate that the distribution for the total number of mutants follows as a consequence.

The outline of this chapter is as follows. Firstly, we define our model and demonstrate a mapping between the mutant clone size distribution and the distribution for the total number of mutants. In order to gain some intuition, this is followed by a short digression on the reduced model when the mutants grow deterministically also. The exact time-dependent size distribution is given for exponential, power-law and logistic wild-type growth function in Section 2.2 and some general finite time features are discussed. Section 2.3 pertains to the asymptotic clone size distribution for general wild type growth and contains our most significant results. At a fixed clone size, for a large class of wild-type growth functions, we demonstrate a general two parameter distribution for clone sizes at large times. The distribution has power-law tail behaviour which corroborates previous work [27, 46, 91]. Large time results are also given for the mean and variance of the clone sizes under general wild-type growth. We follow this with a characterisation of the large clone sizes at long times. Adopting the interpretation of the wild-type population as the primary tumour and mutant clones as metastases, we explore our results regarding the tail of the distribution on an empirical metastatic data in Section 2.5. In our discussion section, we indicate how some contemporaneous work relates to our results, review shortcomings of the work and areas for progression, and then give some concluding remarks.

We remark that some of this chapter has been published in [68], but the work has been revised, and in places strengthened, for presentation here.

2.1.2 Model

In our model a wild-type population gives rise to mutants during reproduction events. The arisen mutant also reproduces and so mutant clones stem from the original initiating mutant's progeny. We shall be focusing on the case where the wild-type population's growth is deterministic for two reasons. Firstly, in many application the wild-type population is significantly larger than the mutant clones. Secondly, if we wished to consider clone sizes under stochastic wild type growth then an appropriate first step would be to obtain results conditioned on the wild type growth, which is equivalent to the setting studied in this chapter. For logistic wild-type growth a sample realisation of the process is shown in Figure 2.1. Further details of the model are now given.

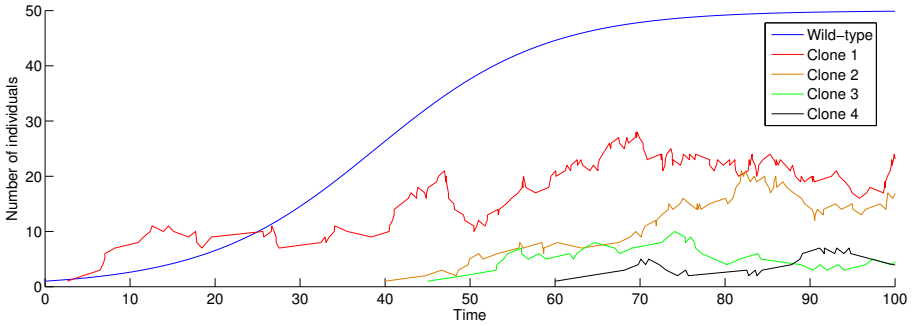


Figure 2.1 *A sample realisation for deterministic logistic wild-type growth, with a carrying capacity of 50, and stochastic mutant growth. Note that we are typically motivated by the case when the wild-type population is much larger than individual clones.*

Mutant clone size distribution

Stochastic growth of mutants will follow a birth-death branching process [6], as discussed in 1.3. We follow the notation introduced there, in particular letting Z_t be the number of cells at time t in a birth-death process initiated with a single cell, and $\mathcal{Z}_t$ its generating function. In this chapter, we shall scale time such that each mutant has unit birth rate and death rate β. If we let the timescale in Section 1.3 be t', with birth and death rates α', β', this is accomplished by defining the timescale in this chapter as $t = \alpha' t'$. This in turn implies that all rates under t are given by dividing the corresponding rate under t' by α', e.g. $\beta = \frac{\beta'}{\alpha'}$. This will simplify the presentation.

We will be interested in statistics of the system at a fixed observation time t. The manner of growth of the wild type until t is critical and to discuss this we let the number of wild-type individuals be denoted by n_τ for $0 \leq \tau \leq t$. Since we suppose mutants are produced by wild-type individuals, the rate of mutant clone initiations will be proportional to the product of n_τ and the mutation rate ν. More precisely, the process of clone initiations is a non-homogeneous Poisson process with intensity νn_τ as in 1.3.2. Let the Poisson random variable K_t denote the number of clones that have been initiated by t, which has mean

$$\mathbb{E}(K_t) = \int_0^t \nu n_\tau \, d\tau. \tag{2.1}$$

Note due to the local integrability property of n_τ given in Section 1.3.2, for any t, $K_t < \infty$ with probability one.

We comment that, if one adopts the interpretation of using deterministic wild-type growth as an approximation to an underlying stochastic process, then many different processes may map to the same deterministic n_τ. To illustrate this point, let us suppose that the wild-type population truly follows the linear birth-death process with mutation of Section 1.3 with birth, death and mutation rates α_w, β_w and ν. We approximate the number of wild type cells as $n_\tau = e^{\delta\tau}$ with $\delta = \alpha_w - \beta_w$. It is clear that for any c, altering the birth and death rates to $\alpha_w + c$, $\beta_w + c$ will lead to the same deterministic approximation n_τ. Reassuringly, the expected number of clones by a given time, (2.1), will also be unchanged under these altered death rates (equation 7 in [5]). However, if we instead defined the linear birth-death with mutations as: replications appear at a rate α_w, and one of the daughter cells is a mutant with probability μ, then altering the birth rate α_w would change the number of clones initiated by a given time, even if the mean number of wild-type cells is the same. Thus care must be taken when using a deterministic approximation to the number of wild-type cells when a particular underlying stochastic process captures the true wild-type dynamics.

At time $t > 0$, assuming $K_t > 0$, by the order statistics property for Poisson processes ([50] Chapter 13) the unordered clone initiation times before t, $(T^{(i)})_{i=1}^{K_t}$ will be *iid* with shared distribution

$$f_T(\tau) = \frac{\nu n_\tau}{\mathbb{E}(K_t)} = \frac{n_\tau}{a_t}, \tag{2.2}$$

where $T = T^{(1)}$ and

$$a_t = \frac{\mathbb{E}(K_t)}{\nu} = \int_0^t n_\tau \, d\tau.$$

Note a_t is the expected number of clones seeded when the mutation rate is unity. To each initiation time $T^{(i)}$ we associate an *iid* linear birth-death process $(Z_s^{(i)})_{s\geq 0}$. Then the unordered clone sizes at time t is given by the *iid* variables $(Z_{t-T^{(i)}}^{(i)})_{i=1}^{K_t}$. We aim to characterise the shared distribution of these clone sizes and so will be considering

$$Y_t = Z_{t-T}^{(1)},$$

where the first member of the unordered clone sizes has been chosen arbitrarily (thus Y_t is equal in distribution to any of the $Z_{t-T^{(i)}}^{(i)}$ due to their *iid* nature). We henceforth drop the superscripts as the linear birth-death processes $(Z_s^{(i)})_{s\geq 0}$ are all *iid*. As we identified each of the K_t initiated clones by their associated unordered arrival times, and as each of the $T^{(i)}$ is equal in distribution to a

uniformly chosen sample from the ordered clone arrival times $(T_i)_{i=1}^{K_t}$, Y_t may have been equivalently defined as the size of a mutant clone uniformly sampled from the K_t initiated clones.

By conditioning on the initiation time of Y_t, we see that

$$\mathbb{P}(Y_t = k) = \frac{1}{a_t} \int_0^t n_\tau \mathbb{P}(Z_{t-\tau} = k)\, d\tau. \tag{2.3}$$

In turn the generating function of the clone size is given by

$$\mathcal{Y}_t(s) = \mathbb{E}(s^{Y_t}) = \frac{1}{a_t} \int_0^t n_\tau \mathcal{Z}_{t-\tau}(s)\, d\tau, \tag{2.4}$$

where the expectation and integral may be interchanged as all terms are non-negative.

We make the following remarks on the above: (i) The mutation rate ν does not appear in the density for initiation times in (2.2), hence mutant clone sizes are independent of the mutation rate and thus all following results regarding clone sizes will be also. (ii) The integral in (2.3) is a convolution and as convolutions commute we may swap the arguments of the integrand functions ($n_\tau \mathcal{Z}_{t-\tau} \leftrightarrow n_{t-\tau}\mathcal{Z}_\tau$). (iii) If we start with n_0 wild-type individuals, so the wild-type follows $m_\tau = n_0 n_\tau$, then both the numerator and denominator in (2.2) will have a factor of n_0, which cancel. So henceforth, apart from when $n_0 = 0$ (used occasionally for analytic convenience), we set $n_0 = 1$ without loss of generality. (iv) By similar logic, a positive random amplitude for the wild-type growth function, i.e. $m_\tau = X n_\tau$ for a general positive random variable X, would also cancel and so our results on clone sizes hold in that case also. Note this includes the asymptotic behaviour of a birth-death process as discussed in Section 1.3.1. (v) If only clones above a certain size are of interest, say clones larger than k^* cells, then as

$$\mathbb{P}(Y_t > k | Y_t \geq k^*) = \frac{\mathbb{P}(Y_t > k)}{\mathbb{P}(Y_t \geq k^*)}, \quad k \geq k^*$$

we see that only the amplitude of the mass function is altered in this case. This comment is particularly relevant to Section 2.5, where we consider metastases, which are typically only observable when greater than 10^7 cells [14]. However the analysis of Section 2.5 will focus on the shape of the mass function, and so the altered amplitude of the conditioned clone size distribution will not be important.

2.1.3 Mapping distributions: clone size to total mutant number

This section is related to the classic Luria-Delbrück problem. Let B_t be the total number of mutants existing at time t. Then B_t is the sum of K_t generic clones

$$B_t = \sum_{i=1}^{K_t} (Y_t)_i \,,$$

where all $(Y_t)_i$ are *iid* random variables specifying the clone sizes. As such, B_t is a compound Poisson random variable, and hence its generating function is

$$\mathcal{B}_t(s) = \mathbb{E}(s^{B_t}) = e^{\mathbb{E}(K_t)[\mathcal{Y}_t(s)-1]}, \tag{2.5}$$

which can be derived by conditioning on K_t. It follows that

$$\mathbb{E}(B_t) = \mathbb{E}(K_t)\mathbb{E}(Y_t) \text{ and } \mathrm{Var}(B_t) = \mathbb{E}(K_t)\mathbb{E}(Y_t^2). \tag{2.6}$$

Notice that the ratio of the variance to the mean is dependent only on the clone size distribution, and thus we will recover the result of Delbrück discussed at the beginning of Chapter 1 for the case of exponential wild type growth. The link between the mass functions of the mutant clone size, Y_t, and the total number of mutants, B_t, is given by

$$\mathbb{P}(B_t = n) = \begin{cases} e^{\mathbb{E}(K_t)(\mathbb{P}(Y_t=0)-1)} & n = 0 \\ \mathbb{E}(K_t)\sum_{k=0}^{n-1} \frac{n-k}{n}\mathbb{P}(B_t = k)\mathbb{P}(Y_t = n-k) & n \geq 1. \end{cases}$$

This is a consequence of the following lemma, which is stated without proof in [93] (Lemma 2). For lack of a reference we provide a proof.

Lemma 2.1.1. *Consider generating functions* $F(s) = \sum_{n\geq 0} p_n s^n$ *and* $G(s) = \sum_{n\geq 0} q_n s^n$ *where* $F(s) = e^{G(s)}$. *Then* $p_0 = e^{q_0}$ *and for* $n \geq 1$ *the following recursion holds*

$$np_n = \sum_{k=0}^{n-1} (n-k) p_k q_{n-k} \,.$$

Proof. Clearly $p_0 = e^{q_0}$ from $F(0) = e^{G(0)}$. By differentiating $F(s)$ we obtain $F'(s) = F(s)G'(s)$, and in general, here with superscript (n) denoting the nth

derivate,

$$F^{(n)}(s) = \sum_{k=0}^{n-1} \binom{n-1}{k} F^{(k)}(s) G^{(n-k)}(s)$$

which can be shown by induction using Pascal's formula for binomial coefficients. Evaluating the above equation at $s = 0$ and using that $F^{(m)}(0) = m!p_n$ and $G^{(m)}(0) = m!q_n$ we arrive at the announced recursion. □

Hence while we may initially work in the setting of size distribution of a single clone, by the above discussion, results are transferable to the total number of mutants case.

Often long-time results are sought, which significantly reduces the complexity of the distributions. For any fixed positive mutation rate, in the long-time limit, an infinite number of clones will have been initiated, resulting in $\lim_{t\to\infty} \mathbb{P}(B_t < \infty) < 1$. A common solution to this problem is the *Large-Population Small-Mutation* limit [51]. The interpretation of this limit is that we are interested in the number of mutants at a large observation time t, and the mutation rate for the process (which unfolds over the time interval $[0, t]$) is chosen to be $\nu = \theta/n_t$. Then for exponential wild-type growth (or exponential-type, see Section 2.3.1) the expected number of initiated clones, $\mathbb{E}(K_t)$, tends to θ for large times. Hence we see that

$$\lim_{\substack{t\to\infty \\ \theta \text{ constant}}} \mathcal{B}_t(s) = \exp\left[\theta(\lim_{t\to\infty} \mathcal{Y}_t(s) - 1)\right],$$

demonstrating that the limit of the clone size distribution is of primary concern.

Furthermore, if the expected number of initiated clones is small, we have the following proposition.

Proposition 2.1. *For a small expected number of initiated clones, conditioned on survival, the size of a single clone and the total number of mutants are approximately equal in distribution. That is,*

$$\mathbb{P}(B_t = k | B_t > 0) = \mathbb{P}(Y_t = k | Y_t > 0) + O(\mathbb{E}(K_t)), \quad \textit{as} \quad \mathbb{E}(K_t) \to 0.$$

Proof of Proposition 2.1. using generating functions and the relevant Taylor

expansions,

$$\begin{aligned}\mathbb{E}(s^{B_t}|B_t>0) &= \frac{\mathcal{B}_t(s)-\mathcal{B}_t(0)}{1-\mathcal{B}_t(0)} = \frac{e^{\mathbb{E}(K_t)(\mathcal{Y}_t(s)-1)}-e^{\mathbb{E}(K_t)(\mathcal{Y}_t(0)-1)}}{1-e^{\mathbb{E}(K_t)(\mathcal{Y}_t(0)-1)}} \\ &= \frac{\mathbb{E}(K_t)((\mathcal{Y}_t(s)-1)-(\mathcal{Y}_t(0)-1))}{-\mathbb{E}(K_t)(\mathcal{Y}_t(0)-1)} + O(\mathbb{E}(K_t)) \\ &= \frac{\mathcal{Y}_t(s)-\mathcal{Y}_t(0)}{1-\mathcal{Y}_t(0)} + O(\mathbb{E}(K_t)) = \mathbb{E}(s^{Y_t}|Y_t>0)+O(\mathbb{E}(K_t)).\end{aligned}$$

□

One immediate consequence of this result is that for immortal mutants ($\beta = 0$) and $\mathbb{E}(K_t) \ll 1$ we have

$$\mathcal{B}_t(s) \approx (1-e^{-\mathbb{E}(K_t)})\mathcal{Y}_t(s) + e^{-\mathbb{E}(K_t)} \implies \mathbb{P}(B_t=k) \approx \mathbb{E}(K_t)\mathbb{P}(Y_t=k) \quad \text{for } k \geq 1.$$

This agrees with intuition as for small enough $\mathbb{E}(K_t)$, we expect only 0 or 1 clones to be initiated and hence the total number of mutants will be dictated by the clone size distribution. With exponential wild-type growth this approximation was used in [46] to investigate drug resistance in cancer.

2.1.4 A first approximation: deterministic mutant growth

Before discussing some exact results for the above model, we take a short digression to discuss the simpler model where the size of each mutant clone grows deterministically. That is we consider the same process as defined above but the size of a mutant clone, at a time s after initiation will follow

$$Z'_s = e^{\lambda s}.$$

The equivalent random variables to those introduced thus far will be given a prime to indicate we work with a reduced model. Many of the characteristics of the more complex model are present in this setting. The simple calculations presented here should facilitate understanding of the later discussion.

In this setting the size of a randomly sampled clone, Y'_t, for $1 \leq y \leq e^{\lambda t}$ is

$$\mathbb{P}(Y'_t > y) = \mathbb{P}(e^{\lambda(t-T')} > y) = \int_0^{t-\lambda^{-1}\log(y)} \frac{n_\tau}{a_t}\, d\tau$$

For the particular case of exponential wild type growth, that is $n_\tau = e^{\delta\tau}$ with $\delta > 0$, this yields

$$\mathbb{P}(Y'_t > y) = \frac{e^{\delta t}}{e^{\delta t} - 1}\left(y^{-\delta/\lambda} - e^{-\delta t}\right)$$

Of note is that, for large times, the clone size density will be a power law decaying as $y^{-\delta/\lambda - 1}$. We contrast this with the case of a constant wild type population $n_\tau = 1$, which results in

$$\mathbb{P}(Y'_t > y) = 1 + \frac{\log(y^{-1})}{\lambda t}$$

Thus the density for clone sizes will decay as $(y\lambda t)^{-1}$ in this case.

We make two observations. We firstly note that the power-law densities derived are a consequence of 'early arriving clones', by which we mean clones initiated close to the beginning of the process. In the exponential case these early arrivals are rare but when they occur, they are likely to be large. Using (2.6), we see this will result in a high variance for the total number of mutants. This observation was the means by which Luria and Delbrück were able to distinguish spontaneous from adaptive mutations [62]. Furthermore by examining the density of initiations (2.2) in each case, early arriving clones are more likely for constant wild type growth than exponential. This explains the 'heavier tail' in the constant setting.

2.2 Finite time clone size distributions

We now return to our original model where each mutant clone grows according to a linear birth-death process and examine some particular cases where the clone size distribution may be computed exactly. This will allow us to see how the distribution is altered for different wild type growth but is also motivated by the fact that in the exponential growth case, for the total number of mutants, the full distribution can be obtained with discrete mutant growth but remains elusive when the reduced model used in the previous section (deterministic mutant growth) is employed [93]. Additionally, these results will hold regardless of the sign of λ (later we will focus on $\lambda > 0$).

Three particular cases of wild-type growth function, n_τ, will be considered namely: exponential, power-law and logistic. Exponential and logistic growth

are widely used in biological modelling [66]. For the power-law cases, under the assumption that the radius of a spherical wild-type population is proportional to time, quadratic and cubic power-law growth represents mutation rates proportional to the surface area and volume respectively. In each case an 'explicit' representation is given for the generating function. These are displayed in Table 2.1. As the derivation of these generating functions mostly consists of manipulating special functions, we relegate the derivation to Appendix A.2.

Wild type growth function	Clone size generating function
Exponential $n_\tau = e^{\delta\tau}$	$\mathcal{Y}_t(s) = 1 + \frac{\lambda}{1-n_t^{-1}}\left[n_t^{-1}\, F\begin{pmatrix}1,\gamma\\1+\gamma\end{pmatrix}; \xi n_t^{-1/\gamma}\right) - F\begin{pmatrix}1,\gamma\\1+\gamma\end{pmatrix};\xi\right)\Bigg]$ with $\gamma = \delta/\lambda$
Power law $n_\tau = \tau^\rho$	$\mathcal{Y}_t(s) = \beta + \lambda(\rho+1)!\left[\frac{(-1)^\rho \mathrm{Li}_{\rho+1}(\xi e^{-\lambda t})}{(t\lambda)^{\rho+1}} + \sum_{i=0}^{\rho}\frac{(-1)^{i+1}\mathrm{Li}_{i+1}(\xi)}{(\rho-i)!(t\lambda)^{i+1}}\right]$
Logistic $n_\tau = \frac{Ke^{\lambda\tau}}{K+e^{\lambda\tau}-1}$	$\mathcal{Y}_t(s) = 1 + \frac{\lambda e^{\lambda t}}{[e^{\lambda t}+(K-1)\xi]\log(\frac{e^{\lambda t}}{n_t})}\log\left(\frac{n_t(1-\xi)}{e^{\lambda t}(1-\xi e^{-\lambda t})}\right)$

Table 2.1 *The clone size generating function for special cases of the wild type growth function. Here* $F\begin{pmatrix}a,b\\c\end{pmatrix};z\right)$ *is Gauss's hypergeometric function,* $\mathrm{Li}_i(s)$ *is the polylogarithm of order* i*, and* ξ *is defined in* (1.4)

The associated probability mass functions for the generating functions given in Table 2.1 are shown in Figure 2.2 at short, medium and long times. Of note is that for short times all mass functions have an exponential cut-off, for the chosen parameters, occurred for clone sizes approximately 20. Also at large times the exponential case has distinct behaviour from all other growth functions. While exact expressions for the mass functions are given in Appendix A.2, the numerical values displayed in Figure 2.2 were obtained via Cauchy's integral formula, which may be efficiently computed using the fast Fourier transform. For a brief discussion on implementation see [4] and references therein.

The mass functions are particularly simple in the following two cases, which we highlight so as to compare with the discussion of Section 2.1.4. For exponential wild type growth ($n_\tau = e^{\delta\tau}$) and immortal, neutral mutants ($\delta = \lambda$, $\beta = 0$) we

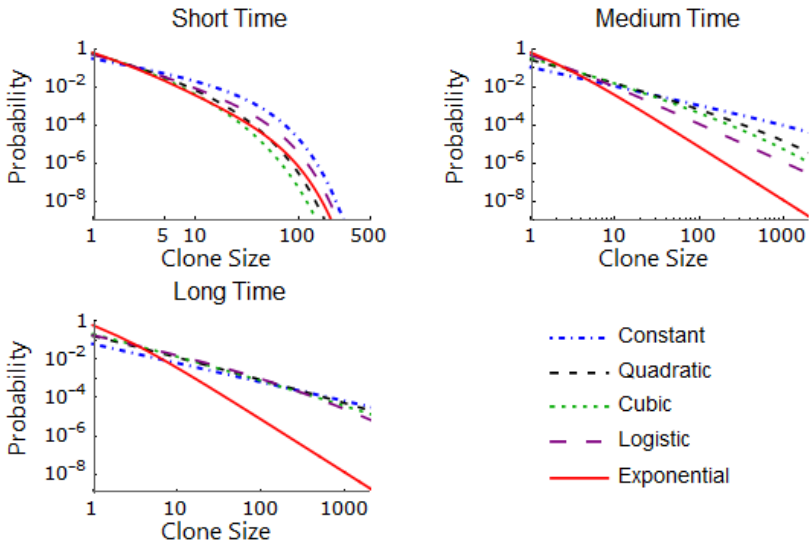


Figure 2.2 *Clone size distribution for different wild type growth for short, medium and long times obtained using the generating functions given in Table 2.1. Parameters constant across plots:* $\delta = 1.8$, $\lambda = 1$, $K = 20000$. *Times used:* $t = 3$, $t = 9$, $t = 15$.

have

$$\mathbb{P}(Y_t = k) = \frac{\phi^{k-1}}{k} - \frac{\phi^k}{k+1} \text{ or } \mathbb{P}(Y_t > k) = \frac{\phi^k}{k+1} \text{ with } \phi = 1 - e^{-\delta t}.$$

While for a constant wild type population ($n_\tau = n_0$)

$$\mathbb{P}(Y_t = k) = \begin{cases} 1 + t^{-1}\log(1 - \mathcal{S}_t^{-1}) & k = 0 \\ \frac{1}{tk}\,\mathcal{S}_t^{-k} & k \geq 1. \end{cases} \tag{2.7}$$

where as in (1.6) $\mathcal{S}_t = \frac{1-e^{-\lambda t}\beta}{1-e^{-\lambda t}}$ (when comparing recall $\alpha = 1$ in this chapter).

Monotone distribution and finite time cut-off

We conclude this section by demonstrating general features that exist in the clone size distribution at finite times, namely a monotone decreasing mass function and an exponential cut-off. These features can be seen Figure 2.2 at 'short times'. Note that the mass function for the birth-death process is monotone decaying and the following short proposition is a consequence of this.

Proposition 2.2. *As long as* n_τ *is positive for some subinterval of* $[0, t]$, *then*

for $k \geq 1$ we have $\mathbb{P}(Y_t = k+1) < \mathbb{P}(Y_t = k)$ for any finite $t > 0$.

Proof of Proposition 2.2. Using (2.3), we see that for $k \geq 1$

$$\mathbb{P}(Y_t = k+1) - \mathbb{P}(Y_t = k) = \frac{1}{a_t} \int_0^t n_{t-\tau} \left[\mathbb{P}(Z_\tau = k+1) - \mathbb{P}(Z_\tau = k)\right] d\tau.$$

Now from (1.5) it is clear that the integrand is negative for finite, positive τ giving the result. □

Whether $\mathbb{P}(Y_t = 0) \geq \mathbb{P}(Y_t = 1)$ depends on n_τ and t. Note that in contrast, the mass function of the total number of mutants is not monotone in general [51].

Now as an example consider the mass function when the size of the wild-type population is constant, which is given by (2.7), and specifically for $k \geq 1$ and $\lambda > 0$. For any moderate t, $\mathcal{S}_t^{-1}$ is typically close to (but strictly less than) one but for large k, $\mathcal{S}_t^{-k}$ will become the dominant term in the mass function, dictating exponential decay. We term this a cut-off in the distribution. To determine where the cut-off occurs, note, with $1 > \epsilon > 0$, we can write $\mathcal{S}_t^{-k} = (1-\epsilon)^k$ which is approximately 1 for $\epsilon k \ll 1$. However this approximation breaks down for $k \approx \epsilon^{-1}$. Therefore we can deduce the cut-off occurs at approximately $k \approx e^{\lambda t}/\lambda$. It is an artifact of the mass function for the birth-death process (1.5). Hence we will have (at least) two behaviour regimes for the mass function for finite times. Here we show that this cut-off exists generally for finite times.

Theorem 2.1. *Let n_τ be continuous and positive for $\tau \in [0, t]$. Then*

$$\mathbb{P}(Y_t = k) = \mathcal{S}_t^{-k} \Theta_t(k),$$

where $\Theta_t(k)$ is an unspecified subexponential factor, i.e. $\limsup_{k\to\infty} \sqrt[k]{\Theta_t(k)} = 1$, and $\mathcal{S}_t$ is given by (1.6).

We interpret the fact that $\limsup_{k\to\infty} \sqrt[k]{\Theta_t(k)} = 1$, as; regardless of the precise form of $\Theta_t(k)$, for large k it neither grows nor decays at a geometric rate or faster. The proof is an application of Theorem IV.7 in [30], which we now state using the notation that for any analytic function $f(z) = \sum_{n\geq 0} f_n z^n$, the nth coefficient is denoted as $[z^n]f(z) = f_n$.

Theorem 2.2 ([30]: Exponential Growth Formula)**.** *If $f(z)$ is analytic at 0 and R is the modulus of singularity nearest the origin in the sense that $R := \sup\{r \geq$*

$0|$ *f is analytic in $|z| < r\}$. Then the coefficient $[z^n]f(z)$ satisfies $f_n = R^{-n}\Theta(n)$ where* $\limsup_n \sqrt[n]{|\Theta(n)|} = 1$.

Proof of Theorem 2.1. Due to Theorem 2.2 we must only seek the closest to the origin singularity of

$$I_t(s) = \int_0^t n_\tau \mathcal{Z}_{t-\tau}(s)\, d\tau = \int_0^t n_{t-\tau} \mathcal{Z}_\tau(s)\, d\tau$$

which is claimed to be at $\mathcal{S}_t$. Indeed, we note that for $|s| < \mathcal{S}_t$, $\mathcal{Z}_\tau(s)$ is analytic for all τ and as n_τ is continuous we can conclude that the $I_t(s)$ is analytic in this region also (Chapter 2, Theorem 5.4 in [79]). As $n_\tau > 0$ there exists $\epsilon > 0$ such that

$$|I_t(s)| \geq \epsilon \left| \int_0^t \mathcal{Z}_\tau(s)\, d\tau \right| = \epsilon \left| \beta t + \log \left[\frac{\lambda}{1 - \beta e^{-\lambda t} - s(1 - e^{-\lambda t})} \right] \right|.$$

The rightmost expression can be seen to have closest to origin singularity at $\mathcal{S}_t$ and as $a_t \mathcal{Y}_t(s) = I_t(s)$, by Theorem 2.2, we can conclude Theorem 2.1.

□

When $\lambda > 0$, $\mathcal{S}_t > 1$ for finite t, and $\mathcal{S}_t \to 1$ exponentially fast for large t. Hence the cut-off will disappear for long times and the subexponential factor, discussed in detail in Section 2.3, will completely determine the tail of the distribution. Also notice that the power-law cases, $n_\tau = \tau^\rho$, for $\rho \geq 1$ are not covered as, to make the analysis tractable, they artificially start at $n_0 = 0$. However the generating function in this case, giving in Table 2.1, also has its closest to origin singularity at $\mathcal{S}_t$ so the cut-off exists there also.

2.3 Large time features: results

In this section we give results regarding the large time behaviour of our model. In many applications the cut-off location ($k \approx e^{\lambda t}/\lambda$) is so large that the distribution at or above this point is of little relevance and hence for this purpose the limiting approximations now discussed are of particular interest. Using the notation of Theorem 2.1, this section investigates the large time form of $\Theta_t(k)$. In particular we can explain the collapse of the mass functions for non-exponential wild type growth at long times seen in Figure 2.2. We highlight the power-law decaying,

'fat' tail found in each case, echoing the analysis of Section 2.1.4. Henceforth we assume $\lambda > 0$, i.e. a supercritical birth-death process.

The analysis will be split in two parts. The first concerns the asymptotic distribution of $\mathbb{P}(Y_t = k)$ at a fixed k, while in the second we examine $\mathbb{P}(Y_t = k)$ for k of order $e^{\lambda t}$. The techniques used for each section are disjoint, and so that we can present a coherent narrative we delay the proofs for this section until Section 2.4.

2.3.1 Fixed clone size

In this section we consider the long time limit for the clone size distribution at a fixed size. We are able to characterise the limit of $\mathbb{P}(Y_t = k)$ for a large class of wild type growth functions, which we now introduce.

General wild type growth functions

To give general results we introduce the following assumption which will be assumed to hold for the remainder of this section.

Assumption 1. *For wild-type growth function n_τ we assume*

(i) *n_τ is positive for all $\tau > 0$ and $n_\tau = 0$ for $\tau < 0$.*

(ii) *n_τ is right continuous with left limits (cadlag).*

(iii) *For $x \geq 0$ the limit $\lim_{\tau\to\infty} n_{\tau-x}/n_\tau$ exists, is positive and less than or equal to 1.*

We note that the cases discussed in Section 2.2 are all covered by Assumption 1. The reason for the seemingly arbitrary limit assumed in (iii) becomes clear with the following result which is an application of the theory of regular variation found in [10].

Lemma 2.1. *For $x \geq 0$*

$$\lim_{t\to\infty} \frac{n_{t-x}}{n_t} = e^{-x\delta^*}, \text{ where } \lim_{t\to\infty} \frac{\log n_t}{t} = \delta^* \geq 0.$$

Often the long-time behaviour of the clone size distribution may be separated into $\delta^* > 0$ and $\delta^* = 0$, and so we give the following definition [30].

Definition 2.1. Consider a real valued function $f(x)$ such that

$$\lim_{x\to\infty} \frac{\log f(x)}{x} = \delta^*$$

holds for some $\delta^* \in \mathbb{R}$. Then $f(x)$ is of *exponential-type* for $\delta^* \neq 0$ and is *subexponential* for $\delta^* = 0$.

Simple examples of subexponential functions are $e^{\sqrt{t}}$, $\log(t)$, t^ρ, while $e^{\delta t}$, $e^{\delta t}t^\rho$ are of exponential-type, with $\delta, \rho \in \mathbb{R}$.

Mean and variance

We now address the asymptotic properties of the clone size distribution by first discussing its mean and variance.

Theorem 2.3. *There exists subexponential function $s_1(t), s_2(t), s_3(t), s_4(t)$ such that, as $t \to \infty$,*

$$\mathbb{E}(Y_t) \sim \begin{cases} \frac{\delta^*}{\delta^*-\lambda} & \lambda < \delta^* \\ s_1(t) & \delta^* = \lambda \\ e^{(\lambda-\delta^*)t}s_2(t) & \delta^* < \lambda \end{cases}$$

$$\mathrm{Var}(Y_t) \sim \begin{cases} \frac{\delta^*}{\lambda}\left(\frac{2}{\delta^*-2\lambda} - \frac{2-\lambda}{\delta^*-\lambda}\right) - \left(\frac{\delta^*}{\delta^*-\lambda}\right)^2 & 2\lambda < \delta^* \\ s_3(t) & \delta^* = 2\lambda \\ e^{(2\lambda-\delta^*)t}s_4(t) & \delta^* < 2\lambda, \end{cases}$$

and $s_1(t), s_3(t) \to \infty$.

The leading asymptotic behaviour which has different regimes dependent on δ^*/λ is illustrated in Figure 2.3. As an example, for the exponential case $n_\tau = e^{\delta\tau}$, by using (2.6) and the results found in [51], then $\delta^* = \delta$, $s_1(t) = \lambda t$, $s_2(t) = \frac{\delta}{\lambda-\delta}$, $s_3(t) = 4t$ and $s_4(t) = \frac{2\delta}{\lambda(2\lambda-\delta)}$. If we let $\delta = \lambda$ and stop the process when the wild type reaches $N = e^{\delta t}$, then $\mathrm{Var}(Y_t)/\mathbb{E}(Y_t) \sim 2N/(\delta \log N)$, recovering Delbrück's result as advertised in Chapter 1.

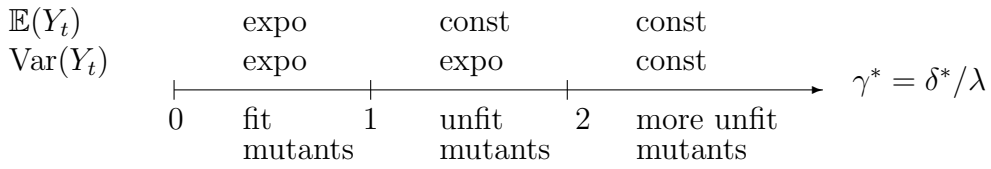


Figure 2.3 *Illustration of the asymptotic behaviour of the mean and variance as given in Theorem 2.3.*

Large time clone size distribution

Turning to the distribution, we have the following result regarding the generating function at large times.

Theorem 2.4. *Let $\gamma^* = \delta^*/\lambda$. Then for $|s| < 1$*

$$\lim_{t\to\infty} \frac{a_t}{n_t}(\mathcal{Y}_t(s) - \beta) = \frac{1}{\gamma^*}\left[1 - F\left(\begin{matrix}1, \gamma^* \\ 1+\gamma*\end{matrix}; \xi\right)\right] = -\sum_{k\geq 1} \frac{\xi^k}{\gamma^* + k}.$$

This result is made clearer in the next corollary, in which the cases of exponential-type and subexponential growth are separated. We separate these cases as, for $\delta^* > 0$,

$$\lim_{t\to\infty} \frac{n_t}{a_t} \to \delta^*.$$

For a proof see Lemma 2.4.1. Consequently in the exponential-type setting, the limiting result is a proper probability distribution, whilst in the subexponential case it is not. We can interpret this as the clone sizes staying finite in the exponential case but grow to infinity for subexponential cases at large times. Unpacking Theorem 2.4 yields the following.

Corollary 2.1. *For $|s| < 1$,*

$$\begin{aligned}
&\lim_{t\to\infty}(\mathcal{Y}_t(s) - \beta) = \lambda\left[1 - F\left(\begin{matrix}1, \gamma^* \\ 1+\gamma*\end{matrix}; \xi\right)\right] && \gamma^* > 0, \\
&\lim_{t\to\infty} \frac{a_t}{n_t}(\mathcal{Y}_t(s) - \beta) = \log(1-\xi) && \gamma^* = 0,
\end{aligned}$$

Then for $t \to \infty$ the probabilities for exponential-type growth $\gamma^* > 0$ are

$$\mathbb{P}(Y_t = k) \sim \begin{cases} 1 - \lambda\, F\left(\begin{matrix}1, \gamma^* \\ 1+\gamma^*\end{matrix}; \beta\right) & k = 0 \\ \frac{\delta^*\Gamma(k)}{(\gamma^*+1)_k}\, F\left(\begin{matrix}k, \gamma^* \\ 1+\gamma^*+k\end{matrix}; \beta\right) & k \geq 1, \end{cases}$$

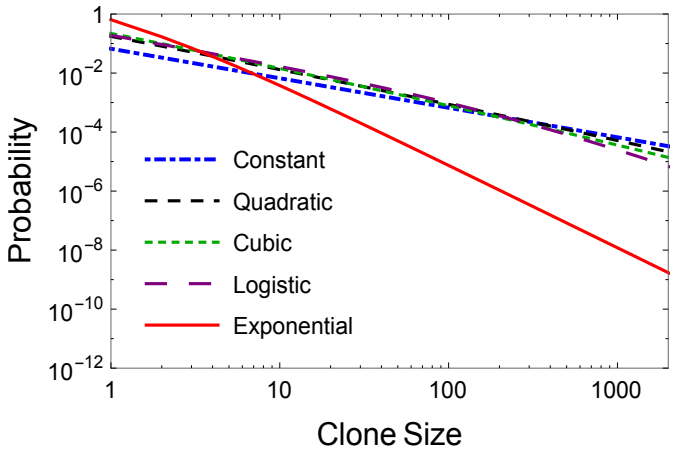


Figure 2.4 *Transition to the asymptotic regime as described in Corollary 2.1. For subexponential wild-type growth the mass functions tend to k^{-1} behaviour, while for exponential-type it tends to $k^{-1-\gamma^*}$. Same as 'long time' in Figure 2.2.*

and for subexponential growth ($\gamma^* = 0$)

$$\mathbb{P}(Y_t = k) \sim \begin{cases} \beta + \frac{n_t \log(\lambda)}{a_t} & k = 0 \\ \frac{n_t}{a_t k} & k \geq 1. \end{cases}$$

The expressions obtained in the $\delta^* > 0$ case also appeared as an approximation in [54] for the total number of mutants with stochastic wild-type and mutant growth when the mean number of clones is small. The agreement between the total number of mutants and the clone size distribution is to be expected by Proposition 2.1. That the above gives the clone size distribution when the wild type also grows as a birth-death process will be discussed shortly.

The case of immortal mutants does not simplify the above expressions for subexponential growth, but for exponential-type growth, by applying (A.2) then (A.1) to the limiting generating function, we have the following link to the Yule-Simon distribution which appears often in random networks [59, 76].

Corollary 2.2. *For immortal mutants with exponential-type wild-type growth the clone size distribution Y_t follows a Yule-Simon distribution with parameter δ^* for large times. That is, for $\beta = 0$, $\delta^* > 0$,*

$$\lim_{t\to\infty} \mathcal{Y}_t(s) = \frac{s\delta^*}{\delta^* + 1} F\left(\begin{matrix} 1, 1 \\ 2 + \delta^* \end{matrix}; s\right),$$

and thus, for $k \geq 1$,

$$\lim_{t\to\infty} \mathbb{P}(Y_t = k) = \frac{\delta^* \Gamma(k)}{(\delta^* + 1)_k}.$$

With immortal, neutral ($\delta^* = 1$) mutants we have

$$\lim_{t\to\infty} \mathbb{P}(Y_t = k) = \frac{1}{k(k+1)}.$$

which is in agreement with the long time limit of (2.2). As a point of interest, note in the immortal, neutral case the mode of the limiting distribution is at $k = 1$, but the expected value is infinite.

For immortal mutants and exponential-type growth, as the clone size distribution tends to a Yule-Simon distribution, we expect power-law tail behaviour at large times [67]. Interestingly, we see that this behaviour holds when we include mutant death and have general wild-type growth.

Corollary 2.3. *At large times, the tail of the clone size distribution follows a power-law with index* $1 + \gamma^*$. *More precisely,*

$$\lim_{k\to\infty} \lim_{t\to\infty} \frac{k^{\gamma^*+1} a_t}{n_t} \mathbb{P}(Y_t = k) = \frac{\Gamma(1+\gamma^*)}{\lambda^{\gamma^*}}.$$

Observe that the exponent in the power-law tail for the constant and exponential growth is the same as that obtained with the deterministic approximation in Section 2.1.4. Further note the ordering of the limits cannot be interchanged as for fixed t, due to the cut off shown in Theorem 2.1, $\lim_{k\to\infty} k^{1+\gamma^*} \mathbb{P}(Y_t = k) = 0$.

An implication: clone sizes in the fully stochastic Luria-Delbrück model

We briefly highlight an implication of the results of Section 2.3.1 when the wild type growth function is a stochastic process. In particular, consider the scenario when clones are initiated as a Cox process with intensity $(\nu A_s)_{s\in\mathbb{R}}$, where $(A_s)_{s\geq 0}$ is a linear birth-death process with growth rate δ, and $A_s = 0$ for $s < 0$. This is the setting where the clone initiation times are generated by the birth-death mutation process of Section 1.3.2. The birth-death processes dictating clone growth are independent of $(A_s)_{s\geq 0}$. This corresponds to the clone sizes in the fully stochastic Luria-Delbrück, or two type branching process, model [5, 54].

Fix a realisation of the the process, ω, such that $A_t(\omega) > 0$ for all t, that is the wild type does not go extinct. Then we see that $(A_t(\omega))_{t\in\mathbb{R}}$ satisfies the first two conditions of Assumption 1. For the third, by the martingale convergence given in (1.9), with $x > 0$,

$$\lim_{t\to\infty} \frac{A_{t-x}(\omega)}{A_t(\omega)} = e^{-\delta x} \lim_{t\to\infty} \frac{e^{-\delta(t-x)}}{e^{-\delta t}} \frac{A_{t-x}(\omega)}{A_t(\omega)} = e^{-\delta x}.$$

Therefore, for each ω such that the wild type does not go extinct the asymptotic clone size distribution is given by Theorem 2.4. Hence the asymptotic clone size distribution in the fully stochastic Luria-Delbrück model, conditioned that the wild type does not go extinct is given by Theorem 2.4. This was also shown in [16], as a consequence of a general result on the convergence of mutation times, but still using the martingale convergence of (1.9).

2.3.2 Large clone size

Thus far we have seen an exponential cut-off exists around $k \approx e^{\lambda t}/\lambda$ and for a fixed k we can describe the clone size distribution, $\mathbb{P}(Y_t = k)$, at large times. For some applications, in particular cancer where the observable metastases (the clones) are of sizes larger than 10^6 cells, it is the large clones we are interested in. It is unclear whether the results of the previous section sufficiently describe the distribution of large clones. We have not been able to answer this question fully, but in this section we show some progress in characterising the clones of size around the cut-off. For consistency we again assume Assumption 1.

We follow [81] to describe the size distribution for large clones at long times via a Poisson process. The idea is to obtain a description of the time-ordered clones. To this end let $(T_i)_{i\geq 1} \subset (0, \infty)$ be the clone initiation times taken from points of a Poisson process with intensity νn_τ, for $\tau \geq 0$. Further let $\tilde{Y}_i(t)$ be the size of the ith initiated clone at time t, where if $t < T_i$, $\tilde{Y}_i(t) = 0$. To connect to the variables discussed thus far $K_t = |\{T_i\}_{i\geq 1} \cap [0, t]|$ and Y_t is uniformly sampled from $\{\tilde{Y}_i(t)\}_{i=1}^{K_t}$.

Then the following theorem gives a Poisson process description of the clone sizes.

Theorem 2.5. *Let $(n_\tau)_{\tau\geq 0}$ satisfy Assumption 1. Then with probability one*

$$\lim_{t\to\infty} e^{-\lambda t}(\tilde{Y}_i(t))_{i\geq 1} = (e^{-\lambda T_i} W_i)_{i\geq 1}$$

where each W_i is iid, with distribution given by (1.10)*. Furthermore, defining*

$$\tilde{K}(u) = |\{e^{-\lambda T_i} W_i\}_{i \geq 1} \cap (0, u)|$$

then $(\tilde{K}(u))_{u>0}$ is a non-homogeneous Poisson process with mean measure

$$m(x, \infty) = \nu\lambda \int_x^\infty n_{\lambda^{-1}\log(s/x)} \frac{e^{-\lambda s}}{s}\, ds < \infty, \quad x > 0. \tag{2.8}$$

For the Poisson process of Theorem 2.5, as for $\epsilon > 0$, $m(\epsilon, \infty) < \infty$ we have that the number of $e^{-\lambda T_i} W_i > \epsilon$ is finite almost surely. Thus we may sample uniformly from this set (i.e. $\{e^{-\lambda T_i} W_i\}_{i \geq 1} \cap (\epsilon, \infty)$) and construct a random variable $Y^{(\epsilon)}$ with distribution

$$\mathbb{P}(Y^{(\epsilon)} > x) = \frac{m(x, \infty)}{m(\epsilon, \infty)}, \quad x \geq \epsilon.$$

The new variable $Y^{(\epsilon)}$ can be related to the previously considered random variable Y_t by the following result.

Theorem 2.6. *For $\epsilon > 0$, with $Y^{(\epsilon)}$ as above,*

$$\lim_{t \to \infty} \mathbb{P}(Y_t e^{-\lambda t} > x | Y_t e^{-\lambda t} > \epsilon) = \mathbb{P}(Y^{(\epsilon)} > x), \quad x \geq \epsilon.$$

Theorem 2.6 only concerns the asymptotic distribution of clones larger than $\epsilon e^{\lambda t}$ and so pertains to large clones. Of note is the reappearance of power-law behaviour with a cut-off in the density of $Y^{(\epsilon)}$. For example in the constant wild-type case, $n_\tau = 1$, the density, using (2.8), is given by

$$f_{Y^{(\epsilon)}}(x) = \frac{d}{dx}\mathbb{P}(Y^{(\epsilon)} \leq x) = \frac{e^{-\lambda x}}{x\Gamma(0, \lambda\epsilon)}, \quad x \geq \epsilon.$$

For exponential growth with neutral mutants, $n_\tau = e^{\lambda\tau}$,

$$f_{Y^{(\epsilon)}}(x) = \frac{e^{-\lambda x}}{x^2}(1 + \lambda x)\epsilon e^{\lambda\epsilon}, \quad x \geq \epsilon.$$

Note that the exponents in the power-law terms is equal to that given in Corollary 2.3, indicating the two approaches are complimentary. However we have not been able to characterise this distribution in a general fashion as in Theorem 2.4.

2.4 Large time features: proofs

Here we present proofs for the asymptotic results given in the previous section.

2.4.1 Fixed clone size

Our aim is to characterise the long time behaviour of $\mathbb{P}(Y_t = k)$ for a class of wild type growth functions. It will be seen that that Assumption 1 (iii) allows us to use the theory of regularly varying functions. The study of such functions is a sub-branch of real analysis but they may typically be encountered as assumptions on the manner in which the tail of a cumulative distribution function decays. The reason for this is that these functions can be thought of as 'asymptotically behaving as a power-law', and power-law decay of a tail gives information on which moments exist of a distribution. However in our setting, in Figure 2.2 we saw that exponential type growth gave a distinctive clone size distribution relative to the other growth functions. By a variable change we can adapt the theory of regular variation for our purposes.

We start by collecting some definitions and required properties from [10]

Definition 2.2. [10] A Lebesgue measurable function $f : \mathbb{R}^+ \mapsto \mathbb{R}$ that is eventually positive is *regularly varying* (at infinity) if for some $\kappa \in \mathbb{R}$,

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\kappa, \quad x > 0.$$

The notation $f \in RV_\kappa$ will be used and we will denote $f \in RV_0$ as *slowly varying functions.*

Theorem 2.4.1 ([10]: Characterisation Theorem)**.** *Suppose $f : \mathbb{R}^+ \mapsto \mathbb{R}$ is measurable, eventually positive, and $\lim_{t\to\infty} \frac{f(tx)}{f(t)}$ exists, and is finite and positive for all x in a set of positive Lebesgue measure. Then, for some $\kappa \in \mathbb{R}$,*

(i) $f \in RV_\kappa$.

(ii) $f(y) = y^\kappa l(y)$ *where* $l \in RV_0$.

Proposition 2.4.1 ([10]: Proposition 1.3.6)**.** *For $f \in RV_\kappa$ $\lim_{t\to\infty} \frac{\log f(t)}{\log t} = \kappa$.*

Theorem 2.4.2 ([10]: Karamata's Theorem)**.** *If $f \in RV_\kappa$, X sufficiently large such that $f(y)$ is locally bounded in $[X, \infty)$, and $\kappa > -1$, then*

$$\int_X^y f(s)\, ds \sim \frac{yf(y)}{\kappa + 1} \quad as \quad y \to \infty.$$

Proposition 2.4.2 ([10]: Proposition 1.5.9.a)**.** *Let $l \in RV_0$ and choose X so that l is locally integrable on $[X, \infty)$. Then*

(i) $\int_X^x \frac{l(t)}{t}\, dt \in RV_0$.

(ii) $\frac{1}{l(x)} \int_X^x \frac{l(t)}{t}\, dt \to \infty$ *as* $x \to \infty$.

Note cadlag functions satisfy the local boundedness and local integrability properties, as they are bounded over compact intervals [9] (chapter 3). We now turn to the proofs.

Proof of Lemma 2.1. Choose $x \geq 0$ and let $y = e^t$, $c = e^{-x}$. Consider the function $g(z) = n_{\log(z)}$. Then Theorem 2.4.1(i) yields

$$\lim_{t\to\infty} \frac{n_{t-x}}{n_t} = \lim_{y\to\infty} \frac{g(yc)}{g(y)} = c^{\delta^*} = e^{-x\delta^*}.$$

Further, Proposition 2.4.1 gives

$$\lim_{y\to\infty} \frac{\log g(y)}{\log y} = \lim_{t\to\infty} \frac{\log(n_t)}{t} = \delta^* \geq 0.$$

The non-negativity of δ^* follows from Assumption 1 (iii). □

To prove Theorem 2.3 we require the following:

Lemma 2.4.1. *There exists subexponential functions $s_1(t)$, $s_2(t)$ such that*

(i) $n_t = e^{t\delta^*} s_1(t)$.

(ii) For $\eta \geq 0$, $C > 0$

$$\int_0^t n_\tau e^{-\eta\tau}\, d\tau \sim \begin{cases} \frac{e^{(\delta^*-\eta)t} s_1(t)}{\delta^*-\eta} & \eta < \delta^* \\ s_2(t) & \delta^* = \eta \\ C & \delta^* < \eta \end{cases} \quad as\ t \to \infty.$$

We highlight that neither subexponential function depends on η.

Proof. (i) For $y = e^t$, $n_{\log y} = g(y)$. Now $g \in RV_{\delta^*}$ hence $g(y) = y^{\delta^*} l(y)$ with $l \in RV_0$ by Theorem 2.4.1(ii). Setting $s_1(\log y) = l(y)$, by Lemma 2.1, $s_1(t)$ is subexponential. (ii) Let $g(y)$ be as above. With $\delta^* > \eta \geq 0$ and using the change of variables $\tau = \log s$ we have

$$\int_0^t n_\tau e^{-\eta\tau}\, d\tau = \int_1^y g(s) s^{-1-\eta}\, ds \sim \frac{y^{-\eta} g(y)}{\delta^* - \eta} = \frac{e^{(\delta^*-\eta)t} s_1(t)}{\delta^* - \eta} \tag{2.9}$$

where the asymptotic equivalence is due to Theorem 2.4.2 applied to $g(y)y^{-1-\eta} \in RV_{\delta^*-\eta-1}$ and the final equality is by part (i). For $\delta^* = \eta$, by Theorem 2.4.1(ii) the integrand will be a subexponential function. Applying the same change of variables as in (2.9) we see by Proposition 2.4.2(i) that the integral is a slowly varying function in y and hence is subexponential in t, which we denote $s_2(t)$. Now for $\delta^* < \eta$, by Lemma 2.1 we may choose t large enough such that $n_t^{1/t} < e^{(\delta^*+\eta)/2}$ which by a basic result for Laplace transforms, see, e.g., Theorem 1.11 in [75], ensures convergence to a finite, positive constant. □

As an example, which will be useful for the next proof, we apply the above lemma to a_t. With $s_1(t)$, $s_2(t)$ subexponential functions

$$a_t = \int_0^t n_\tau\, d\tau \sim \begin{cases} \frac{e^{\delta^* t} s_1(t)}{\delta^*} & \delta^* > 0 \\ s_2(t) & \delta^* = 0 \end{cases} \quad \text{as } t \to \infty.$$

Proof of Theorem 2.3. We require the first and second moments of Z_t which may be found by differentiating (1.4). Using this we have

$$\mathbb{E}(Y_t) = \frac{1}{a_t} \int_0^t n_\tau \mathbb{E}(Z_{t-\tau})\, d\tau = e^{\lambda t} \frac{\int_0^t n_\tau e^{-\lambda\tau}\, d\tau}{\int_0^t n_\tau\, d\tau} \tag{2.10}$$

and

$$\begin{aligned} \mathbb{E}(Y_t^2) &= \frac{1}{a_t} \int_0^t n_\tau \mathbb{E}(Z_{t-\tau}^2)\, d\tau \\ &= \frac{e^{2\lambda t}}{a_t \lambda} \left(2 \int_0^t n_\tau e^{-2\lambda\tau}\, d\tau - (2-\lambda) e^{-\lambda t} \int_0^t n_\tau e^{-\lambda\tau}\, d\tau \right). \end{aligned} \tag{2.11}$$

Throughout let s_t be a generic subexponential function and it will be helpful to observe that the reciprocal or constant multiples of a subexponential function

are subexponential. We first consider the mean. For the cases $\delta^* \neq \lambda$, applying Lemma 2.4.1(ii) to (2.10) with $\eta = \lambda$ for the numerator and $\eta = 0$ for the denominator proves the claim. For $\delta^* = \lambda$, using Lemma 2.4.1(i) then (ii), we have

$$\mathbb{E}(Y_t) = \frac{e^{\delta^* t}\int_0^t e^{\delta^*\tau} s_\tau e^{-\delta^*\tau}}{\int_0^t e^{\delta^*\tau} s_\tau \, d\tau} \sim \frac{\delta^* \int_0^t s_\tau \, d\tau}{s_t} = s_1(t). \tag{2.12}$$

That $s_1(t)$ diverges can be seen by applying the standard change of variables $t = \log(y)$, $\tau = \log(s)$ coupled with Proposition 2.4.2(ii). Turning to the variance, with $\mathrm{Var}(Y_t) = \mathbb{E}(Y_t^2) - \mathbb{E}(Y_t)^2$, when $\delta^* > 2\lambda$ we apply Lemma 2.4.1(ii) term by term to (2.11). For $\delta^* < \lambda$ all integrals converge and so, with C_1, C_2 constants,

$$\mathrm{Var}(Y_t) \sim \frac{e^{2\lambda t}}{a_t}\left(C_1 - \frac{C_2}{a_t}\right) \sim C_1 \frac{e^{2\lambda t}}{a_t}.$$

The last relation is due to the monotonicity of a_t, a consequence of Assumption 1 (i). The desired representation is obtained by applying Lemma 2.4.1(ii) to a_t and absorbing C_1 into $s_4(t)$. When $\lambda \leq \delta^* < 2\lambda$ the same argument holds as long as we note that

$$e^{-\lambda t}\int_0^t n_\tau e^{-\lambda t}\, d\tau = e^{-\lambda t}\int_0^t e^{(\delta^*-\lambda)\tau} s_\tau \, d\tau \leq e^{-(2\lambda-\delta^*)t}\int_0^t s_\tau \, d\tau.$$

By Proposition 2.4.2(i) the rightmost integral is a subexponential function and as we may always choose t sufficiently large such that $s_t^{1/t} < e^{2\lambda-\delta^*}$, we find

$$e^{-\lambda t}\int_0^t n_\tau e^{-\lambda t}\, d\tau \to 0.$$

Applying Lemma 2.4.1 to $a_t^{-1}\int_0^t n_\tau e^{-\lambda t}\, d\tau$ demonstrates the contribution from the mean squared is negligible. For $\delta^* = 2\lambda$ we apply the same argument as in (2.12) to each term and this completes the proof. □

In order to prove Theorem 2.4 we require the following lemma.

Lemma 2.4.2. *For $|s| < 1$, $\beta \in [0, 1)$, $u \in [0, 1]$ and ξ as in (1.4), we have*

$$\left|\frac{\xi}{1-\xi u}\right| \leq \left|\frac{\beta - s}{1 - \max\{\beta, |s|\}}\right|.$$

Proof. By the definition of ξ,

$$\frac{\xi}{1-\xi u} = \frac{\beta - s}{1 - s - (\beta - s)u}.$$

The triangle inequality yields

$$|1 - s - \beta u + su| = |1 - (\beta u + s(1-u))| \geq |1 - |\beta u + s(1-u)||$$

and

$$|\beta u + s(1-u)| \leq u\beta + (1-u)|s| \leq \max\{\beta, |s|\}.$$

The claimed inequality now follows. □

Proof of Theorem 2.4. Taking the generating function for Y_t from equation (2.4) we apply the change of variables $u = e^{-\lambda\tau}$ which gives

$$\mathcal{Y}_t(s) - \beta = -\frac{1}{a_t}\int_{e^{-\lambda t}}^{1} n_{t+\frac{\log u}{\lambda}} \frac{\xi}{1-\xi u}\, du.$$

Now recalling $n_\tau = 0$ for $\tau < 0$ and multiplying both sides by $\frac{a_t}{n_t}$ yields

$$\frac{a_t}{n_t}(\mathcal{Y}_t(s) - \beta) = -\int_0^1 \frac{n_{t+\frac{\log u}{\lambda}}}{n_t} \frac{\xi}{1-\xi u}\, du.$$

Lemma 2.1 yields $\lim_{t\to\infty} n_{t+\frac{\log u}{\lambda}}/n_t = u^{\delta^*/\lambda}$. Hence there exists t^*, such that for all $t > t^*$, $n_{t+\frac{\log u}{\lambda}}/n_t \leq u^{\delta^*/\lambda} + 1$. For $t \leq t^*$ the above is bounded, as cadlag functions are over finite intervals. Coupling this with Lemma 2.4.2 shows the integrand may be dominated. By assumption the integrand converges and therefore, using the dominated convergence theorem, we have

$$\begin{aligned}\lim_{t\to\infty} \frac{a_t}{n_t}(\mathcal{Y}_t(s) - \beta) = -\int_0^1 u^{\delta^*/\lambda} \frac{\xi}{1-\xi u}\, du &= \frac{-1}{\xi^{\gamma^*}} \mathrm{B}_\xi(\gamma^* + 1, 0) \\ &= \frac{1}{\gamma^*}\left[1 - F\left(\begin{matrix}1, \gamma^* \\ 1+\gamma^*\end{matrix}; \xi\right)\right].\end{aligned}$$

The final equality follows from applying (A.4) then (A.2). □

Proof of Corollary 2.1. The first statement is given by applying Lemma 2.4.1(ii) to a_t/n_t. Then taking the limiting generating function in Theorem 2.4 we firstly apply (A.2) then (A.3) which yields generating function representation for $\gamma^* = 0$. The mass function for $\gamma^* = 0$ is simply a logarithmic expansion. For $\gamma^* > 0$ we use the expression given in Appendix A of [51] to obtain a series expansion for the generating function in terms of s and the coefficients of the expansion give the mass function. □

Proof of Corollary 2.3. The analysis involves expanding the limiting generating function in Theorem 2.4 around its singularity at $s = 1$ and exactly mirrors that given in section 6 of [51] and so is omitted.

□

2.4.2 Large clone size

We start by proving the Poisson description of clone sizes.

Proof of Theorem 2.5. The proof of the first statement follows Theorem 3 in [81]. The general idea can be seen by considering the limit of the first clone. Let $Z^{(1)}(t)$ be a linear birth-death process, independent from the Poisson process $(K_t)_{t\geq 0}$. We have that $\tilde{Y}_1(t) = Z^{(1)}(t-T_1)\mathbb{I}_{\{t\geq T_1\}}$. By the martingale limit in Section 1.3.1, in particular equation (1.9), we have,

$$e^{-\lambda t}Z^{(1)}(t-T_1) = e^{-\lambda T_1}e^{-\lambda(t-T_1)}Z^{(1)}(t-T_1) \to e^{-\lambda T_1}W_1 \, a.s.$$

Similar statements hold for each $e^{-\lambda t}\tilde{Y}_i(t)$, which implies convergence of the sequence.

We now turn to the Poisson description, which is similar to the proof of Theorem 4 in [81]. We'll aim to show $\tilde{K}(a,b) = \tilde{K}(b) - \tilde{K}(a)$ is Poisson with mean $m(a,b)$. First though let us take A, B to be open subsets of $\mathbb{R}_{>0}$ and consider $\tilde{N}(A\times B) = |\{(T_i, W_i) \in A\times B\}|$, that is the number of points that fall in the (open) rectangle $A \times B$. We already know that the number of T_i falling in A is Poisson with mean $\int_A \nu n_\tau \, d\tau$, thus for $\tilde{N}(A \times B)$ we simply do not count those T_i, whose associated W_i does not fall in B. This is a thinning of our original Poisson process $(K_t)_t \geq 0$ (dictating the T_i), and hence $\tilde{N}(A\times B)$ is Poisson with mean $\mathbb{P}(W_i \in B)\int_A \nu n_\tau \, d\tau$. Therefore, for any open $S \subset \mathbb{R}^2_{>0}$ (which can be decomposed into open rectangles), $\tilde{N}(S)$ is Poisson with mean

$$\mathbb{E}[\tilde{N}(S)] = \int_S \int \nu n_\tau \lambda^2 e^{-\lambda y} \, dy d\tau.$$

The reason for the 'extra' λ (as opposed to just the exponential density) is due to $\mathbb{P}(W_i > 0) = \lambda$. Arguing similarly (using the independent increments on the T_i) for sets S_1, S_2, $\tilde{N}(S_1), \tilde{N}(S_2)$ are independent Poisson variables.

Turning to $\tilde{K}(u)$, we observe that

$$\tilde{K}(a,b) = \tilde{N}\left(\{(\tau,w) : a < e^{-\lambda\tau}w < b\}\right).$$

This implies $\tilde{K}(a,b)$ is Poisson with mean

$$\begin{aligned} m(a,b) = \mathbb{E}[\tilde{K}(a,b)] &= \int_0^\infty \int_{ae^{\lambda\tau}}^{be^{\lambda\tau}} \nu n_\tau \lambda^2 e^{-\lambda y}\, dy d\tau \\ &= \lambda \int_0^\infty \nu n_\tau (e^{-\lambda a e^{\lambda\tau}} - e^{-\lambda b e^{\lambda\tau}})\, d\tau. \end{aligned}$$

Independent increments follows as a consequence of the independent sets property for $\tilde{N}$. The given representation of $m(x,\infty)$ is achieved by a change of variables. That $m(x,\infty) < \infty$ is due to the above integrals being convergent, as by Lemma 2.1, for sufficiently large t, $n_t^{1/t} < 2e^{\delta^*}$. □

Proof of Theorem 2.6. Immediately from (2.3) we see

$$\mathbb{P}(Y_t > xe^{\lambda t} | Y_t > \epsilon e^{\lambda t}) = \frac{\mathbb{P}(Y_t > xe^{\lambda t})}{\mathbb{P}(Y_t > \epsilon e^{\lambda t})} = \frac{\int_0^t n_{t-\tau}\mathbb{P}(Z_\tau > xe^{\lambda t})\, d\tau}{\int_0^t n_{t-\tau}\mathbb{P}(Z_\tau > \epsilon e^{\lambda t})\, d\tau}.$$

It is enough to examine the numerator. As, from (1.5),

$$\mathbb{P}(Z_t > k) = (1 - \beta \mathcal{S}_t^{-1})\mathcal{S}_t^{-k}, \quad k \geq 0$$

we have

$$\int_0^t n_{t-\tau}\mathbb{P}(Z_\tau > xe^{\lambda t})\, d\tau = \int_0^t n_{t-\tau}\mathcal{S}_\tau^{-\lfloor xe^{\lambda t}\rfloor}\, d\tau - \beta \int_0^t n_{t-\tau}\mathcal{S}_\tau^{-\lfloor xe^{\lambda t}\rfloor - 1}\, d\tau \tag{2.13}$$

Here $\lfloor a \rfloor$ denotes the integer part of a, and is necessary as $\mathbb{P}(Z_t > k)$ is defined on the non-negative integers. Focusing on the first term from the right hand side of (2.13) and using the definition of $\mathcal{S}_\tau$ (1.6) gives

$$\int_0^t n_{t-\tau} \exp\left(-\lfloor xe^{\lambda t}\rfloor(\log(1-\beta e^{-\lambda\tau}) - \log(1-e^{-\lambda\tau}))\right)\, d\tau.$$

Now we change variables to $s = xe^{\lambda(t-\tau)}$ and note that the resulting integrand can be dominated by $\frac{n_{\lambda^{-1}\log(s/x)}}{\lambda s} e^{\lambda(1-s)}$ which is integrable over $[x,\infty)$ by $m(x,\infty) < \infty$ (from Theorem 2.5). Hence by the dominated convergence theorem we can conclude

$$\lim_{t\to\infty} \int_0^t n_{t-\tau}\mathcal{S}_\tau^{-\lfloor xe^{\lambda t}\rfloor}\, d\tau = \lambda^{-1} \int_x^\infty n_{\lambda^{-1}\log(s/x)} \frac{e^{-\lambda s}}{s}\, ds.$$

The second integral from the right hand side of (2.13) can be treated analogously and yields an identical result with β as a prefactor. Hence

$$\lim_{t\to\infty}\int_0^t n_{t-\tau}\mathbb{P}(Z_\tau > xe^{\lambda t})\,d\tau = \int_x^\infty n_{\lambda^{-1}\log(s/x)}\frac{e^{-\lambda s}}{s}\,ds = (\lambda\nu)^{-1}m(x,\infty),$$

and the claimed result follows. □

2.5 An application: tail behaviour in empirical metastatic data

Given the discussion thus far we expect, for a large class of wild-type growth functions, to see power-law tail behaviour on approach to the exponential cut-off in the clone size distribution. We take the first steps to verify this theoretical hypothesis by analysing an empirical metastatic data. In this setting the wild-type population is the primary tumour, mutant clones are the metastases and the mutation rate is the rate at which metastases are seeded from the primary. Our aim here is to explore whether adapting our model to this setting is at all reasonable (rejection of a power law tail in the metastasis sizes would imply it is not).

The data is sourced from the supplementary materials in [14]. This data is taken from 22 patients; 7 with pancreatic ductal adenocarcinomas, 11 with colorectal carcinomas, and 6 with melanomas. One patient had only a single metastasis so we discard this data. Of the 21 remaining patients the number of cells in a single metastasis ranged from 6×10^7 to 2.23×10^9. Our theoretical model predicts a cut-off in the distribution around $k = e^{\lambda t}/\lambda$. Taking some sample parameters from the literature, namely $\lambda = 0.069$/day [24], and $t = 14.1$ years [92], this leads to a cut-off around $k \approx 10^{154}$ cells. Due to the enormity of this value we ignore the cut-off here. Additionally, as the minimum observed metastasis size is 6×10^6 cells, we assume that all data points are sampled from the tail of the distribution.

2.5.1 Likelihood methodology

We broadly follow the methodology outlined in [17]. For each of the datasets, one from each patient, we examine the likelihood ratio to determine if the data is more

likely sampled from a power-law decaying or geometrically decaying distribution. In particular for each metastasis dataset $\mathbf{y} = (y_1, \ldots, y_N)$ we wish to test the hypothesis that the elements of $\mathbf{y}$ are drawn *iid* from $\mathbb{P}_1(Y_t = y; \psi) = C_1 y^{-\psi}$ for $y \geq k^*$, versus that that they are sampled from $\mathbb{P}_2(Y_t = y; p) = C_2 p(1-p)^y$ for $y \geq k^*$. Here C_1, C_2 are normalising constants and k^* is chosen as the 'beginning of the tail'. We test the hypothesis by examining the log-likelihood ratio

$$\hat{\mathcal{R}} = \sum_{i=1}^{N} [\log \mathbb{P}_1(Y_t = y_i; \hat{\psi}) - \log \mathbb{P}_2(Y_t = y_i; \hat{p})],$$

where $\hat{\psi}$ is the maximum likelihood estimate (MLE) for ψ assuming the data is drawn from $\mathbb{P}_1(Y_t = k; \psi)$ and $\hat{p}$ is the MLE assuming $\mathbb{P}_2(Y_t = k; p)$. We see that $\hat{\mathcal{R}} > 0$ gives support to the hypothesis that the data is drawn from the power-law decaying distribution relative to a distribution with a geometric tail.

Given a set of metastases sizes, $\mathbf{y}$, to calculate $\hat{\mathcal{R}}$ it remains to obtain the MLEs. We derive the MLE in detail for geometric tail decay, the power-law case is analogous but for the final step. Here $\mathbf{y}$ is assumed to be a realisation of the random vector $\mathbf{Y}_t = (Y_t^{(1)}, \ldots, Y_t^{(N)})$. All $Y_t^{(i)}$ are *iid* with distribution $\mathbb{P}_2(Y_t = y; p) = C_2 p(1-p)^y$. Let $A = \mathbb{P}_2(Y_t < k^*; p)$ (the weight of the distribution contained before the beginning of the tail). Then

$$\sum_{y \geq k^*} \mathbb{P}_2(Y_t = y; p) = \sum_{y \geq k^*} C_2 p(1-p)^y = 1 - A \implies C_2 = (1-A)(1-p)^{-k^*}.$$

The log-likelihood is now given by

$$\begin{aligned} \log \mathfrak{L}(p) &= \log \left[\prod_{i=1}^{N} \mathbb{P}_2(Y_t^{(i)} = y_i; p) \right] \\ &= N \log(1-A) - k^* N \log(1-p) + N \log(p) + \log(1-p) \sum_{i=1}^{N} y_i. \end{aligned}$$

The MLE is found by maximizing the likelihood $\mathfrak{L}(p)$. Setting $\partial_p \log \mathfrak{L}(p) = 0$ we solve to find the MLE

$$\hat{p} = \frac{N}{\sum_{i=1}^{N} y_i + N(1-k^*)}.$$

The power-law case is analogous. There $C_1 = \frac{1-A}{\zeta(\psi, k^*)}$, where ζ is the Hurwitz zeta function [26]. No closed form expression is found for the MLE and instead we

have the approximation [17] for large k^*

$$\hat{\psi} \approx 1 + N\Big[\sum_{i=1}^{N} \log\Big(\frac{y_i}{k^* - \frac{1}{2}}\Big)\Big]^{-1}.$$

We must select k^* as the start of the tail behaviour of each of the given distributions. As we assume all data is drawn from the tail, we choose $k^* = \min \mathbf{y}$. Note we also assume the weight of the distribution prior to entering the tail behaviour, A, is the same in both cases.

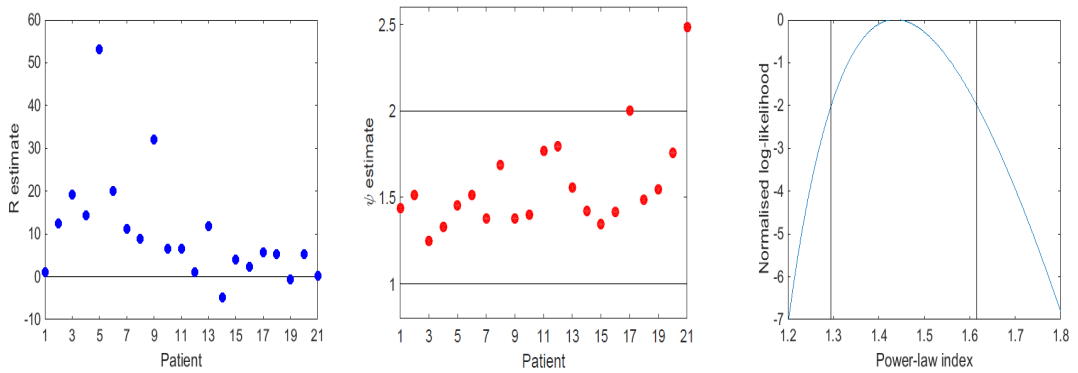
2.5.2 Likelihood analysis results

The estimates of the log-likelihood ratio for each dataset is shown in Figure 2.5a. We see that 19 of the 21 dataset return the power-law hypothesis as more plausible ($\hat{\mathcal{R}} > 0$) than a geometric tail.. Assuming a power-law distribution, the maximum likelihood estimates for the exponent ψ for each dataset are given in Figure 2.5b. For 20 of the 21 datasets we find the point estimate of the power-law index, $\hat{\psi}$, lies in $[1, 2]$, with an average of 1.57. The outlier comes from the smallest dataset (3 metastases). Due to the small sample size of our datasets and the high variance in the distribution, we do not derive confidence intervals via normal distribution approximations. Instead we show the normalised log-likelihood, $\log \mathfrak{L}(\psi)/\mathfrak{L}(\hat{\psi})$, for our best dataset, with $N = 30$, in Figure 2.5c, where $\mathfrak{L}(\psi)$ is the likelihood function. Also, following [43], we demonstrate the likelihood interval defined as

$$I(\psi) = \{\psi : \log \frac{\mathfrak{L}(\psi)}{\mathfrak{L}(\hat{\psi})} \geq -2\}.$$

If a large sample size was possible this interval would correspond to a 95.4% confidence interval. For the dataset with $N = 30$ we numerically determined $I(\psi) = [1.295, 1.616]$, illustrated as the domain between the vertical bars in Figure 2.5c.

We continue to assume the data is drawn from a power-law decaying distribution. The results of Sections 2.3.1 and 2.3.2, imply that given the data is sampled from a distribution with a power-law decaying tail, the exponent of the decay conveys information on the relative growth rate of the primary to the metastases. As the estimated exponents are all greater than 1, which would be the exponent for subexponential primary tumour growth, we focus on the case where the primary tumour grows exponentially. From Corollary 2.1 we see $\hat{\psi} - 1$ is a point estimate



(a) *Likelihood ratio* $\hat{\mathcal{R}}$ *for each dataset. Points above the horizontal line suggest dataset is from a power-law distribution over a geometric distribution.*

(b) *Estimate* $\hat{\psi}$ *for each dataset, determined via maximum likelihood.*

(c) *Normalised log-likelihood function for first best set. Vertical bars show the likelihood interval.*

Figure 2.5 *Likelihood analysis results: Patients are sorted left to right by number of metastases with patient 1 having 30 mets to patient 21 having 3. Hence values to left of figures are more significant.*

of γ/λ, the ratio of the primary growth rate to the metastases growth rate. As the average estimate of this growth rate ratio is 0.57, under the exponential growth model, the metastases grew almost twice as fast as the primary tumours. Note that each patient may have different primary and metastasis growth rates, i.e. we do not assume a uniform δ across the patients. We can compare our estimate with previous studies who have reported estimates for the ratio of the primary growth rate to the metastatic growth rate. For melanoma [15] reports 0.44 while [78] gives 0.66 for epidermoid cancer types. These values are computed as the average primary growth divided by the average metastatic growth, whereas we give the average of the primary growth rate to the metastatic growth rate. We can see that our estimates are reasonable.

2.6 Discussion

Before giving some concluding remarks we contrast the work presented above with two contemporaneous studies.

Time-dependent rate parameters

Two recent works [42, 84] have also investigated the impact of relaxing the assumption of exponential growth in Luria-Delbrück models of mutant generation. In both cases they considered the case where all rates in the system are multiplied by a time-dependent function, say $z(\tau)$. This is relevant in the scenario where both the wild-type and mutant populations have their growth restricted simultaneously by environmental factors, for example exposure to a chemotherapeutic agent. We observe that under a change of timescale this system is equivalent to our setting with exponential wild-type growth. This is due to the following argument.

We firstly examine the initiation process. In this setting the wild-type population is governed by

$$\frac{dn_\tau}{d\tau} = \delta z(\tau) n_\tau$$

which implies

$$n_\tau = \exp\left(\delta \int_0^\tau z(s)\, ds\right).$$

Mutant clones are now initiated at a rate $\nu z(\tau) n_\tau$ and we let $\widehat{K}_t$ be the number of initiations by t. Let

$$F(\tau) = \int_0^\tau z(s) ds$$

and define the new timescale as $r(\tau) = F^{-1}(\tau)$. The expected number of initiations under this timescale is

$$\begin{aligned}
\mathbb{E}[\widehat{K}_{r(t)}] &= \int_0^{r(t)} \nu z(s) \exp\left(\delta \int_0^s z(u)\, du\right) ds \\
&= \frac{\nu}{\delta} \int_0^{r(t)} \frac{d}{ds} \exp\left(\delta \int_0^s z(u)\, du\right) ds \\
&= \frac{\nu}{\delta}\left[\exp\left(\delta \int_0^{r(t)} z(u)\, du\right) - 1\right] = \frac{\nu}{\delta}(e^{\delta t} - 1).
\end{aligned}$$

Thus the initiation of mutant clones is equivalent to the exponential wild type case. We now turn to mutant growth.

Let $\widehat{Z}_t$ be the size of a mutant population governed by the birth-death process with time-dependent rates. Once initiated, the size distribution obeys the forward

Kolmogorov equation for time-dependent stochastic mutant proliferation

$$\frac{d}{dt}\mathbb{P}(\widehat{Z}_t = k) = z(t)(k-1)\mathbb{P}(\widehat{Z}_t = k-1) \\ + \beta z(t)(k+1)\mathbb{P}(\widehat{Z}_t = k+1) - (1+\beta)z(t)k\mathbb{P}(\widehat{Z}_t = k).$$

Similar to the initiation process, the chain rule shows that the forward Kolmogorov equation for $\widehat{Z}_{r(t)}$ is equal to that of the time independent linear birth death process. Thus, under a time-rescaling, all dynamics are equivalent to the system with exponential wild-type growth and stochastic mutant proliferation with constant birth and death rates, as studied in this chapter or in [51].

2.6.1 Shortcomings and future work

From both a biological and mathematical viewpoint, there are several aspects of this study which should be examined further. We will use the approximate model of Section 2.1.4 to illustrate future areas of interest. Recall in this setting, mutant clones grow deterministically and so clones may take values in $[1, e^{\lambda t}]$

Firstly, while the asymptotic distribution has been explored for clones of a fixed size, and of size around $e^{\lambda t}$ (Sections 2.3.1 and 2.3.2), these might not be the most relevant limits. For example, taking power-law wild type growth $n_\tau = \tau^\rho$, we see that if we examine the distribution around $y = e^{a\lambda t}$ for $0 < a < 1$ we have

$$\mathbb{P}(Y'_t > y) = \mathbb{P}(e^{\lambda(t-T')} > e^{a\lambda t}) = \frac{(\rho+1)\int_0^{t(1-a)} \tau^\rho \, d\tau}{t^{\rho+1}} = (1-a)^{\rho+1}.$$

This implies that $(Y'_t)^{1/(\lambda t)}$ which takes values in $[1, e]$ possesses a proper probability distribution for all t. Exploring this observation might lead to more succinct proofs but should not change the general conclusions of the above.

Relatedly, in the context of cancer one of our motivations for considering wild type growth other than exponential was that primary tumours growth will be restricted eventually due to environmental constraints. However this will also be true of large enough metastases. A size dependent branching process could be considered, but in light of the insight that has been provided in the exponential case, a first step would be to have mutant growth follow a generic deterministic function g_τ. For example, when both the wild-type and mutant populations grow deterministically as τ^ρ, we see that for large times the clone size distribution still

displays a power-law tail, $\lim_{t\to\infty} t f_{Y_t'}(y) = \frac{\rho+1}{\rho} y^{1/\rho-1}$. It would be interesting to explore how general power law tails are in growing populations seeding growing populations.

A further biological criticism is the fact that each clone has the same birth and death rates. However the growth rates of individual metastases are likely to be dependent on the genotype of the founding cell and its local environment. For example the per day growth rates of metastases given in [78] yield a 95% range of 0.0013-0.0866 against a mean of 0.0107. While this data pools the growth rates across patients, we still expect fluctuations at the patient level. To address this we could modify our process such that we associate an *iid* real valued random variable Λ_i for each arrival in the non-homogeneous Poisson process $(K_t)_{t\geq 0}$. For a fixed death rate, the birth rate of each clone would then be $\delta + \beta + \Lambda_i$. This is similar to the model of [28], where the impact of random fitnesses on the total number of mutants was explored. The exploration of this model is left for future work.

For the metastasis application of this chapter, it was assumed that metastases were initiated at a rate proportional to the primary tumours size. Implicitly we are supposing all (or a time independent fraction) of the primary cells can lead to metastases. However, it is often thought that cells in the primary tumour must acquire a mutation to become 'metastatically enabled', and it is only this subpopulation that seeds the metastases. While some models of metastases [25, 37] consider this extra complication, we have not included it for two reasons. Firstly, the evidence for whether a metastatically enabling mutation exists currently appears inconclusive [49, 85]. Secondly if such a mutation is required, for exponential primary growth, the resultant subpopulation would also eventually grow exponentially. For similar reasons, we do not treat the case of metastases initiating further metastases

A technical criticism regarding the metastasis application is our use of the likelihood ratio as a model selection tool. While the approach is intuitive and it is that recommended in [17], it may be simpler to place the model selection question within a Bayesian framework. The advantage of this approach would be to allow confidence intervals to be estimated via simulation for the corresponding quantity to the likelihood ratio. This was avoided here, as the confidence intervals for the likelihood ratio are based on large sample asymptotic normality which was not applicable, although a simulation based approach conditional on the MLE estimates would be possible.

A different application of interest concerns the sizes of clones of mutated skin cells, where the mutation is induced by exposure to ultraviolet radiation. This was the setting considered in [56]. Prolonged exposure to UVB radiation leads to the initiation of $p53$ mutant clones (PMC), that is colonies of cells lacking sufficient, functional amounts of the protein $p53$, which is known to suppress cancer initiation. The observed size distribution of PMCs is remarkably similar to that shown Figure 2.2 at short times with a constant wild type growth function. Namely a power-law with approximate exponent -1 was found, followed by an exponential cut off. Hence, as was concluded in [56], a model of stochastically (birth-death process) growing clones initiated at a constant rate is consistent with the data. Due to Corollary 2.1, this will also be true for subexponential type initiation of clones, for a large enough time. This does not change the conclusions of [56], whose main aim was challenging that clones grew according to a power-law, which would result in a different PMC size distribution. However it does indicate further applications for some of the work presented here.

2.6.2 Concluding remarks

In this study we focus on the size distribution for mutant clones initiated at a rate proportional to the size of the wild-type population. The size of the wild-type population is dictated by a generic deterministic growth function and the mutant growth is stochastic. This shifts the focus from previous studies which have mostly been concerned with exponential, or mean exponential, wild-type growth, and considered the total number of mutants. Results for the total number of mutants are, however, obtainable from the clone size distribution.

Motivated by their extensive use in applications the generating function for the special cases of exponential, power-law and logistic wild-type growth was given. Regardless of the growth function, the mass function is monotone decreasing and the distribution has a cut-off for any finite time. This cut-off goes to infinity for large times and is often enormous in practical applications, hence we focused on the approach to the cut-off.

We found that the clone size distribution behaves quite distinctly for exponential-type versus subexponential wild-type growth. Although the probability of finding a clone of any given size stays finite as $t \to \infty$ for exponential-type growth, it tends to zero for subexponential type. Despite these differences, with an appropriate scaling, for a large class of growth functions we proved that the clone

size distribution has a universal long-time form. This long-time form possesses a power-law, heavy tail which decays as $1/k$ for subexponential wild-type growth, but faster for exponential-type growth. This can be intuitively understood as the tail distribution represents clones that arrive early, and the chance that a clone is initiated early in the process is larger for a slower growing wild-type function. Hence we expect a heavier tail in the subexponential case.

An underlying motivation for this work is the scenario of primary tumours seeding metastases in cancer. From our theoretical results it was predicted that regardless of the primary tumour growth, a power law tail was expected in the metastasis size distribution. As a first step to explore this, likelihood analysis was applied to a dataset of metastasis sizes. Asking whether the data was more likely sampled from a power-law or exponentially decaying distribution, the power-law was returned as more plausible for 19 of 21 datasets. This allowed estimation of the relative growth rate of primary tumour to the metastases, which was approximately 0.57 (metastases grew almost twice as fast). For medical applications, there are superior methods to obtain this quantity. For example, tracking the growth of primary tumours and metastases in images. However, given the array of complexities not considered by our model, that the general prediction of the model (power-law tail in the clone size distribution) was not inconsistent with the data gives some support to the reasonableness of our approach.

Chapter 3

Competing Pathways in Growing Populations over Fitness Valleys

3.1 Introduction and Model

3.1.1 Introduction

In the previous chapter we considered the clone size distribution of mutants initiated from a wild type population. A single transition was required to alter wild type cells into mutants. In this chapter we discuss the natural extension where the cell type of interest, mutants in Chapter 2, is an arbitrary number of transitions from the initial cell type.

Transitions might be the acquisition of (epi)genetic alterations or the migration into new environments. Suppose we have a cellular state of interest, for example; a given genotype, a spatial location, or a combination of both. Will this state ever be reached? If it is, when is it reached? And by which sequence of intermediate states?

Consider the example illustrated in Figure 3.1. Each vertex represents a cellular type, in this case its resistance profile. The edges represent cell transitions via mutation upon replication. How long does it take for multidrug resistance to emerge? En route, is resistance to drug A or drug B more likely to arise first?

As discussed in Chapter 1 similar questions arise frequently in applications,

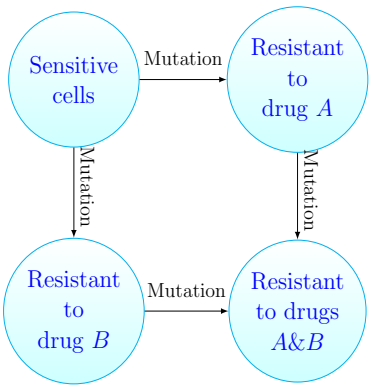


Figure 3.1 *Motivating example: consider a population of growing cells. Each vertex represents a cell type and cells can mutate to acquire resistance. Starting from one sensitive cell that can replicate, die or mutate, how long does it take to traverse the graph until multidrug resistance emerges? Is resistance to drug A then B more likely than the converse?*

where the meaning of transition is determined by context. A common method is to model the growing population as a continuous time birth-death (branching) process. In this chapter we offer a detailed treatment of this model when all the intermediate states have reduced reproductive ability (fitness). We focus on this due to applications in cancer and bacterial modelling.

Our main contribution is to provide simple intuitive formulas that answer the questions posed above for a moderately general setting. These formulas show explicitly the contribution of the model parameters, e.g. transition rates and fitness reductions, which allow them to be probed for biological insight. This provides relationships which might be difficult to deduce from simulations alone.

The outline of this chapter is as follows. Firstly, we precisely describe the process we consider in Section 3.1.2. We next appeal to recent progress in branching process theory [47] in order to understand the population growth of intermediate types. This is then used to characterise the time it takes for a population of a given type to arise. Both our accounts of the population growth and time involve the entries of an eigenvector of a given matrix. A fuller picture of the impact of all the system's parameters becomes apparent when we specialise our graph to be acyclic which we do in Section 3.2.3. Continuing with the acyclic setting, we follow this by considering which sequence of transitions leads to this target population arising, in particular the probability of a given sequence is derived. How these results carry into the cyclic setting is next considered and we remark on some extensions to the initial model for which our discussion still holds. In

section 3.3 we demonstrate the utility of our results with some applications. The scenarios of imperfect drug penetration and the orderings of resistance conferring mutations are considered at length. We hope this exemplifies how readers may apply the results of Section 3.2. We close with a concluding discussion.

3.1.2 Model

We start by defining the specific framework our results hold for.

Our model is a specific form of multitype branching process [6] in which each population will evolve according to a Markovian linear birth-death process with transitions. We will assume each population is comprised of cells. As a conceptual framework we associate the multitype branching process with a simple, finite, rooted, directed, graph $G = (V, E)$ containing N vertices ($N = |V|$). Labels of the vertices take values in $\{1, \ldots, N\}$ and E is a subset of the set of ordered pairs $\{(i, j) : i, j \in V, i \neq j\}$. We shall often refer to vertex 1 as the root and vertex N as the target. Each vertex is reachable from the root and the target is reachable from any other vertex. We further assume the root is a source vertex and the target vertex is a sink. Letting the set of incoming neighbours for vertex i be $\mathcal{N}^-(i) = \{k \in V : (k, i) \in E\}$ and the set of outgoing neighbours $\mathcal{N}^+(i) = \{k \in V : (i, k) \in E\}$, the previous assumption is $\mathcal{N}^-(1) = \mathcal{N}^+(N) = \emptyset$. Each type in the branching process is uniquely mapped to a vertex, and so the number of types is N. Hence any cell may be described by its type or the vertex associated to that type.

Take any cell at vertex x. We assume this cell gives birth at rate $\alpha(x)$, dies at rate $\beta(x)$ and transitions (at replication) to a cell of type y at rate $\nu(x, y)$ if edge (x, y) is contained in E. The growth rate of type x will be denoted $\lambda(x) = \alpha(x) - \beta(x)$. The parameters associated with the type 1 population feature prominently and so for convenience we let $\alpha = \alpha(1)$, $\beta = \beta(1)$ and $\lambda = \lambda(1)$. All cells are independent. We will focus on the setting where the type 1 population is the most fit and has positive growth rate. Therefore, we henceforth assume that $\lambda > 0$ and for $2 \leq x \leq N - 1$, $\lambda(x) < \lambda$. We do not specify the relative fitness of the type N population, but in most applications $\lambda(N) > 0$. The cell level

dynamics may summarised as

$$(x) \to \begin{cases} (x),(x) & \text{rate } \alpha(x) \\ \varnothing & \text{rate } \beta(x) \\ (x),(y) & \text{rate } \nu(x,y) \text{ if } (x,y) \in E \end{cases}$$

where (x) represents a cell of type x and $\varnothing$ symbolises a dead cell. At a population level, the number of cells of type x at time t will be denoted $Z_x(t)$. We shall always assume $Z_x(0) = z\delta_{x,1}$, where $\delta_{x,y}$ is the Kronecker delta function. That is we initiate the system with only type 1 cells, at a quantity z. The population growth of the initial and intermediate types is crucial and so the notation $\mathcal{Z}(t) = (Z_x(t))_{1 \le x \le N-1}$ will be used.

We have two primary questions. The first is, having initiated the system with z type 1 cells, how long does it take for the population at the target vertex to arise? That is we concern ourself with the distribution of

$$T = \inf\{t \ge 0 : \ Z_N(t) > 0\}. \tag{3.1}$$

Now assuming that the target vertex is populated by a founding cell, we ask from which path through the graph G did this founding cell come? This gives rise to a distribution over the set of paths (or walks) from the root vertex to the target

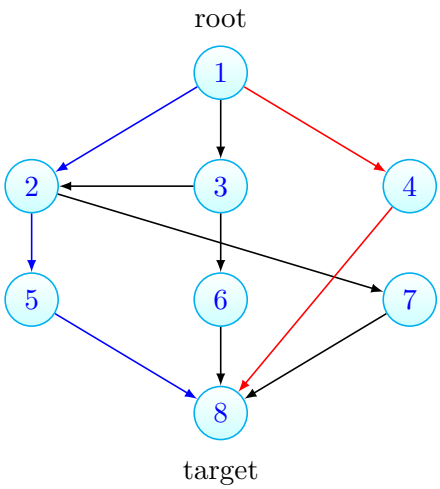


Figure 3.2 *Example of the process when $N = 8$. Vertices represent cellular types. Cells of type x can transition to type y if the edge (x, y) exists. Starting with cells at the root vertex, how long until the target is populated? Is the blue or red path more likely to initiate the target population?*

vertex, which we aim to characterise. This second question is more precisely formulated in Section 3.2.4, but is also illustrated in Figure 3.2.

Often a slightly modified set of questions considered, where instead we wait for the first type N individual to arise whose progeny does not ultimately go extinct. If this is the setting of interest then the entirety of the discussion below holds under the minor alteration that the transitions rates into vertex N should be mapped $\nu(x, N) \mapsto \nu(x, N)\lambda(N)/\alpha(N)$ for $x \in \mathcal{N}^-(N)$. This is as $\lambda(N)/\alpha(N)$ is the probability of non-extinction for a population initiated with a single Nth type cell (see (1.8)). We would also like to stress that multiple types can exist at any time, contrasting the often studied case of strong selection weak mutation regime for a fixed size population [34].

3.1.3 A first approximation: deterministic growth for $N = 2$

As in Chapter 2, before proceeding to a discussion of the full model as given above, we take a short digression to discuss a simpler model. In particular, let us consider the case when $N = 2$, that is there are only two vertices and so we only wait for the first mutation from the initial population. Further, let us suppose that the type 1 population grows deterministically as $Z'_t = e^{\lambda t}$ (as in Chapter 2, we will use the same variables as given above but with primes to indicate we consider the reduced model). Transitions from the type 1 population to the target population occur at rate $\nu(1, 2)Z'_t$. Note this is exactly the initiation process of mutants considered by Luria and Delbrück [62] as discussed in Chapter 1.

In this setting, as the target is initiated as a non-homogeneous Poisson process,

$$\mathbb{P}(T' > t) = \exp\left(-\nu(1,2)\int_0^t Z'_s\, ds\right) = \exp\left(-\frac{\nu(1,2)(e^{\lambda t} - 1)}{\lambda}\right).$$

The above distribution is of the Gumbel class, and so we can obtain the mean of T', but instead note that, for fixed $\nu(1, 2)$, it is only when $\nu(1, 2)e^{\lambda t} \approx 1$ that T' is likely to occur. In fact, from direct computation, we see that

$$\lim_{\nu(1,2)\to 0} \mathbb{P}(T' - \mu' > t) = e^{-e^{\lambda t}/\lambda}, \text{ where } \mu' = \lambda^{-1}\log[\nu(1,2)^{-1}]. \tag{3.2}$$

Hence T' is concentrated around the variable μ'. A similar scaling will be presented for the more complex model.

In the full model, when we turn to which sequence of transitions occurred (the second question from the previous section) the following observation provides the intuition. Observe that we can informally state (3.2) as, for small $\nu(1,2)$,

$$\mathbb{P}(T' > t) = \mathbb{P}(T' - \mu' > t - \mu') \approx \exp\left(-\nu(1,2)e^{\lambda t}/\lambda\right).$$

In turn, this implies

$$\mathbb{P}(e^{\lambda T'} > t) \approx \exp\left(-\nu(1,2)t/\lambda\right).$$

Thus by rescaling time, the time from the root vertex population to the target population arising is approximately exponentially distributed. Here there is only one transition that can occur, and hence no questions about the orderings of the transitions is possible. However in the full model, by arguing in a similar manner to the above, the scaled time it takes for a particular sequence of transitions will again be exponential. From this, which sequence of transitions occurs first may be deduced by the orderings of exponential random variables, which is well studied.

3.2 Results

Where possible, we aim to present our results in tandem with heuristic derivations, similar to that of the previous section, to convey the general idea. Key is the understanding of the asymptotic behaviour of the initial and intermediate population $\mathcal{Z}(t)$ at long times.

3.2.1 Long time population size

In the above discussion of the reduced model , the distribution of T' was accessible due to our assumption that the root population grew deterministically and exponentially. Returning to the full model, the target vertex is seeded by the neighbouring populations, whose growth is stochastic. However the asymptotic theory given in [47], whose relevant statements we give below, shows that the long time behaviour of the initial and intermediate population $\mathcal{Z}(t)$ is exponential growth with a random amplitude. Therefore the situation in the full case is similar to that considered in the reduced model.

The main purpose of this subsection is to present the required statement from [47]. However to do so we first introduce some further notation and then give a short lemma detailing restrictions on G such that the asymptotic result is guaranteed to hold.

As defined above the process $\mathcal{Z}(t) = (Z_1(t), \ldots, Z_{N-1}(t))$ is a multitype branching process, and we further specify that each $Z_i(t)$ be right-continuous. It will become apparent that the stochasticity in $\mathcal{Z}(t)$ is controlled by the type 1 population. As vertex 1 is a source vertex, and thus no type transitions into a type 1 cell, the long term behaviour of $Z_1(t)$ is as described in Section 1.3.1, which we restate here in a form convenient for our later discussion. Recall that z is the initial number of type 1 cells.

Lemma 3.1.

$$\lim_{t\to\infty} Z_1(t)e^{-\lambda t} = W \quad a.s.$$

W is a non-negative random variable and satisfies

$$W \stackrel{d}{=} \sum_{i=1}^{z} \chi_i \xi_i$$

where $\chi_i \sim$ Bernoulli(λ/α), $\xi_i \sim$ Exponential(λ/α) and are independent.

In terms of notation, for real random variables X, Y, that $X \stackrel{d}{=} Y$ means $\mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x)$ for each $x \in \mathbb{R}$.

To obtain a similar result for $\mathcal{Z}(t)$, let us introduce the $N-1 \times N-1$ matrix A, with entries

$$a_{ij} = \begin{cases} \lambda(i) & j = i \\ \nu(j,i) & \text{if } (j,i) \in E \\ 0 & \text{otherwise.} \end{cases}$$

A controls the mean behaviour and as is seen from the forward Kolmogorov equations, or Lemma 9.2 in [47], we have

$$\mathbb{E}[\mathcal{Z}(t)] = \exp(At) z \mathbf{e}_1.$$

The notation $\mathbf{e}_i$ is used for the $(N-1) \times 1$ vector with ith entry equal to 1 and 0 otherwise. Also note that the off diagonal elements of the transpose of A are positive whenever the corresponding element of the adjaceny matrix of G is one.

In order to use results from [47], we wish to discuss the case when the largest

real eigenvalue of A, λ^*, is simple (i.e. has algebraic multiplicity 1) and equal to λ. Below we give two sufficient conditions for this to be true. The second is motivated by applications when typically all transition rates are small.

Lemma 3.2. *Sufficient conditions such that $\lambda^* = \lambda$ and is simple are: 1) G is acyclic. 2) For all $2 \leq i \leq N-1$, $\lambda(i) + \sum_{k \in \mathcal{N}^-(i)} \nu(k,i) < \lambda$.*

Proof. 1) As G is directed and acyclic, a relabelling of the vertices exists such that the underlying adjaceny matrix of the relabelled graph is upper triangular. This is shown in Appendix B. We use this relabelling with vertex 1 as first in the labelling, π. Thus π is a permutation of $\{1, \ldots, n\}$ with $\pi(1) = 1$. The permutation matrix associated with π and its transpose are

$$P_1 = \begin{bmatrix} \mathbf{e}_{\pi(1)}^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_{\pi(n)}^{\mathrm{T}} \end{bmatrix}, \quad P_1^{\mathrm{T}} = \begin{bmatrix} \mathbf{e}_{\pi(1)}, \cdots, \mathbf{e}_{\pi(n)} \end{bmatrix}.$$

Under π we have the new matrix

$$A' = P_1 A P_1^{-1}.$$

By Lemma B.2 A' is lower triangular and has the $\lambda(i)$ as diagonal elements. Thus λ is the largest eigenvalue of A'. As A and A' are similar matrices, and hence share eigenvalues, we can conclude the statement.

2) By considering the left eigenvector $\mathbf{e}_1^{\mathrm{T}}$, λ is indeed an eigenvalue. We now demonstrate all other eigenvalues have real part smaller than λ. Observe that the assumed condition implies the bound on the row sums

$$\sum_{j=1}^{N-1} a_{i,j} < \lambda. \tag{3.3}$$

As A is reducible (due to the root being a source vertex) there exists a permutation matrix P_2 such that

$$P_2 A P_2^{-1} = \begin{pmatrix} \lambda & & & \\ R_{2,1} & R_{2,2} & & 0 \\ \vdots & & \ddots & \\ R_{M,1} & \ldots & & R_{M,M} \end{pmatrix}$$

with each submatrix $R_{i,i}$ square and irreducible and $2 \leq M \leq N-1$ (see [86] section 2.3). The eigenvalues of A comprise λ and the eigenvalues of the $R_{i,i}$. For any of the $R_{i,i}$ let $r_j^{(i)}$ be sum of the jth row of $R_{i,i}$. By Gershgorin's Disc Theorem ([86] Theorem 1.5) the real part of the eigenvalues of $R_{i,i}$ are bounded by $\max_j\{r_j^{(i)}\}$. Further, the bound (3.3) is preserved under permutations and so each $r_j^{(i)} < \lambda$. As this holds for each $R_{i,i}$ the claim is shown. □

Henceforth we assume either of the conditions given in Lemma 3.2 hold and thus $\lambda^* = \lambda$ and is simple. Due to this assumption, results of [47] show that the type 1 cells drives the entire population growth. To state this result, we introduce the vector ψ with

$$\psi_i = \alpha(i) + \beta(i) + \sum_{k \in \mathcal{N}^+(i)} \nu(i,k). \tag{3.4}$$

Now let $\tilde{u}, \tilde{v}$ be the left and right eigenvectors of A corresponding to λ, scaled such that

$$\psi \cdot \tilde{v} = 1, \quad \tilde{u} \cdot \tilde{v} = 1. \tag{3.5}$$

Note, from the structure of A, only the first entry of $\tilde{u}$ is non-zero, that is $\tilde{u}_i > 0$ if $i = 1$ and 0 otherwise. Further let

$$\phi_x = \tilde{u}_1 \tilde{v}_x. \tag{3.6}$$

Then the following holds.

Theorem 3.1 (Janson [47])**.** *With W as in Lemma 3.1,*

$$\lim_{t\to\infty} e^{-\lambda t}\mathcal{Z}(t) = W(\phi_x)_{x=1}^{N-1} \quad a.s.$$

Proof. We demonstrate how to apply the result in [47] to our setting. From Theorem 3.1 in [47], with probability one

$$\lim_{t\to\infty} e^{-\lambda t}\mathcal{Z}(t) = \hat{W}\tilde{v}$$

where $\hat{W}$ is a non-negative random variable, with as yet unknown distribution. However from Lemma 3.1 almost surely $e^{-\lambda t}Z_1(t) \to W$, with the distribution of W given in the lemma. This implies $\hat{W} = W/\tilde{v}_1 = \tilde{u}_1 W$ with the last equality following from (3.5) and that only the first entry of $\tilde{u}$ is positive. □

It is perhaps surprising that at the scale of $e^{\lambda t}$ the only randomness of the intermediate types (types 2 to $N-1$) is inherited from the random amplitude of

the type 1 population. The same result for the case $N = 3$ was demonstrated in [16].

W, as given in the above theorem, is positive only when the type 1 population avoids extinction (almost surely). It will be convenient to define this event and so we let

$$S_1 = \{Z_1(t) > 0 \text{ for all } t \geq 0\}. \tag{3.7}$$

Immediately from Lemma 3.1 (when $W = 0$) or (1.8) we have

$$\mathbb{P}(S_1) = 1 - (\beta/\alpha)^z. \tag{3.8}$$

Now due to Theorem 3.1, we find that the frequencies of the initial and intermediate types converge deterministically as detailed in the following.

Corollary 3.1. *Conditioned on S_1,*

$$\lim_{t\to\infty} \frac{\mathcal{Z}(t)}{\sum_{i=1}^{N-1} Z_i(t)} = \frac{(\phi_x)_{x=1}^{N-1}}{\sum_{i=1}^{N-1} \phi_i} \quad a.s.$$

where the ϕ_x are as given in (3.6).

3.2.2 Time until target vertex is populated

We now turn our attention to T, the time at which the type N population arises, as defined in (3.1). In particular we focus on the distribution of T when the transition rates into the target are small. For each neighbour of the target $x \in \mathcal{N}^-(N)$, the vertex x seeds vertex N with intensity, or seeding rate, $\nu(x, N)Z_x(t)$. This is similar to the motivating example of the linear birth death process with mutations discussed in Section 1.3.2. Assume the transition rates into vertex N are small enough such that when T is likely to occur sufficient time will have passed that each $Z_x(t)$ can be well approximated by its asymptotic form. Then the seeding rate from vertex x into the target vertex will be approximately $\nu(x, N)We^{\lambda t}\phi_x$. Hence we might expect

$$\begin{aligned}\mathbb{P}(T > t) &= \mathbb{E}\left[\exp\left(-\int_0^t \sum_{x\in\mathcal{N}^-(N)} \nu(x, N)Z_x(s)\, ds\right)\right] \\ &\approx \mathbb{E}\left[\exp\left(-\int_0^t \sum_{i\in\mathcal{N}^-(N)} \nu(x, N)We^{\lambda s}\phi_x\, ds\right)\right].\end{aligned}$$

The first equality arises as the target vertex is initiated according to a Cox process and should be compared with the discussion leading to (1.11) in Section 1.3.2. Note that the final expectation can be computed due to our knowledge of W provided by Lemma 3.1. To formalise this argument we introduce some further notation.

In order to ensure the transition rates are small enough such that the asymptotic regime is reached we define

$$\nu^* = \max_{i \in \mathcal{N}^-(N)} \nu(i, N)$$

i.e. the largest of the transition rates into the target. As $\nu^* \to 0$, $T \to \infty$. To describe the manner in which T increases in this limit we define

$$\mu = \lambda^{-1} \log \left(\lambda^2 \left[\alpha \sum_{i \in \mathcal{N}^-(N)} \nu(i, N) \phi_i \right]^{-1} \right).$$

Then the following holds.

Theorem 3.2. *Let W be as in Theorem 3.1, then with probability one*

$$\lim_{\nu^* \to 0} \mathbb{P}(\lambda(T - \mu) > t | (\mathcal{Z}(t))_{t \geq 0})) = \exp\left(-e^t W \lambda / \alpha\right).$$

It follows that

$$\lim_{\nu^* \to 0} \mathbb{P}(\lambda(T - \mu) > t) = \left(\frac{\lambda/\alpha}{1 + e^t} + \beta/\alpha \right)^z$$

Proof. Let

$$\sigma = \sum_{i \in \mathcal{N}^-(N)} \nu(i, N) \phi_i.$$

Then, as for given $(\mathcal{Z}(t))_{t \geq 0}$ immigration into vertex N occurs as a non-homogeneous Poisson process,

$$\mathbb{P}(T - \lambda^{-1} \log(\sigma^{-1}) > t | (\mathcal{Z}(t))_{t \geq 0}) = \exp\left(- \int_0^{t + \lambda^{-1} \log(\sigma^{-1})} \sum_{i \in \mathcal{N}^-(N)} \nu(i, N) Z_i(s) \, ds \right). \tag{3.9}$$

The variable change $u = s - \lambda^{-1} \log(\sigma^{-1})$ yields

$$\exp\left(- \int_{-\lambda^{-1} \log(\sigma^{-1})}^{t} \sum_{i \in \mathcal{N}^-(N)} \nu(i, N) Z_i(u + \lambda^{-1} \log(\sigma^{-1})) \, du \right). \tag{3.10}$$

Now observe that, by multiplying each summand by $1 = \frac{e^{\lambda(u+\lambda^{-1}\log(\sigma^{-1}))}}{e^{\lambda(u+\lambda^{-1}\log(\sigma^{-1}))}}$,

$$\sum_{i\in\mathcal{N}^-(N)} \nu(i,N) Z_i(u+\lambda^{-1}\log(\sigma^{-1})) = \sum_{i\in\mathcal{N}^-(N)} \frac{Z_i(u+\lambda^{-1}\log(\sigma^{-1}))}{e^{\lambda(u+\lambda^{-1}\log(\sigma^{-1}))}} \frac{\nu(i,N)e^{\lambda u}}{\sigma}. \tag{3.11}$$

Consider the limit of the summands as $\nu^* \to 0$, which results in $\log(\sigma^{-1}) \to \infty$. As the $(Z_i(t))_{i\in\mathcal{N}^-(N)}$ are independent of $(\nu(i,N))_{i\in\mathcal{N}^-(N)}$, Theorem 3.1 provides the limit for the first term in the product of each summand. Upon considering $(\sigma/\nu(i,N))^{-1}$, we see $\nu(i,N)/\sigma < \infty$ for each i. Hence taking limits across (3.11) yields, with probability one,

$$\begin{aligned}\lim_{\nu^*\to 0} \sum_{i\in\mathcal{N}^-(N)} \frac{Z_i(u+\lambda^{-1}\log(\sigma^{-1}))}{e^{\lambda(u+\lambda^{-1}\log(\sigma^{-1}))}} \frac{\nu(i,N)e^{\lambda u}}{\sigma} &= \sum_{i\in\mathcal{N}^-(N)} We^{\lambda u}\phi_i \lim_{\nu^*\to 0} \frac{\nu(i,N)}{\sigma} \\ &= We^{\lambda u} \lim_{\nu^*\to 0} \sum_{i\in\mathcal{N}^-(N)} \frac{\nu(i,N)\phi_i}{\sigma}.\end{aligned}$$

By the definition of σ this last sum is one and hence

$$\lim_{\nu^*\to 0} \sum_{i\in\mathcal{N}^-(N)} \nu(i,N) Z_i(u+\lambda^{-1}\log(\sigma^{-1})) = We^{\lambda u} \quad a.s. \tag{3.12}$$

We seek to apply the dominated convergence theorem to the integral in (3.10). The above convergence implies that for any realisation, we may find x such that for $\nu^* \leq x$

$$\sum_{i\in\mathcal{N}^-(N)} \nu(i,N) Z_i(u+\lambda^{-1}\log(\sigma^{-1})) \leq 2We^{\lambda u}$$

which is integrable over $(-\infty, t]$. If $\nu^* > x$ then $\sigma > 0$. Thus the interval $[\lambda^{-1}\log(\sigma^{-1}), t]$ is finite. As each $Z_i(t)$ is cadlag, it is bounded over finite intervals. Hence for all (ν^*, u) the integrand is dominated by an integrable function.

Therefore, taking the limit of (3.10) and using (3.12) leads to, with probability one,

$$\lim_{\nu^*\to 0} \mathbb{P}(T - \lambda^{-1}\log(\sigma^{-1}) > t | (\mathcal{Z}(t))_{t\geq 0}) = \exp\left(-We^{\lambda t}/\lambda\right).$$

Using that

$$\mu = \lambda^{-1}\log(\lambda^2/\alpha) + \lambda^{-1}\log(\sigma^{-1})$$

so that we centre appropriately, leads to the first stated result.

For the second statement, we must derive an expression for

$$\mathbb{E}\left[\exp\left(-e^t W\lambda/\alpha\right)\right]$$

From Lemma 3.1

$$W\lambda/\alpha \stackrel{d}{=} \sum_{i=1}^{z} \chi_i \xi_i'$$

where $\chi_i \sim$ Bernoulli(λ/α), $\xi_i' \sim$ Exponential(1) and are independent. Thus

$$\mathbb{E}\left[\exp\left(-e^t W\lambda/\alpha\right)\right] = \left(\mathbb{E}\left[\left(\exp(-e^t\chi_1\xi_1'\right)\right]\right)^z. \tag{3.13}$$

Using a one-step conditioning argument on χ_1 and considering the Laplace transform for a unit rate exponential variable leads to the stated result. □

The first statement in the above theorem (3.2) will be key to understanding the path distribution, discussed in Section 3.2.4, which justifies its prominence. It may be interpreted as indicating that if time is rescaled and we look at the distribution of T around μ, an exponential distribution appears. We also remark that our reason for taking only the transition rates into the target vertex is as discussed directly under (3.11), namely that the growth of $\mathcal{Z}(t)$ is independent of these transition rates, which allows us to use Theorem 3.1 to show (3.12).

It is natural to consider the distribution of T conditioned that it is finite (we cannot discuss which path populated the target vertex if $T = \infty$). However it is technically more convenient to condition on S_1, defined in (3.7). The following proposition states that in many relevant cases, namely large initial population, low death rate or small transition rates leaving vertex 1 these events are similar. Below, the superscript c denotes the complement of the set.

Proposition 3.1.

$$\mathbb{P}(\{T < \infty\}^c \cap S_1) = 0.$$

$$\mathbb{P}(\{T < \infty\} \cap S_1^c) \leq \begin{cases} O((\beta/\alpha)^z), & \beta \to 0 \\ O((\beta/\alpha)^z), & z \to \infty \\ O(\nu_1), & \nu_1 \to 0 \end{cases}$$

where $\nu_1 = \sum_{x \in \mathbb{N}^+(1)} \nu(1, x)$.

Proof. As $\mathbb{P}(\{T < \infty\}^c \cap S_1) = \mathbb{P}(T = \infty | S_1)\mathbb{P}(S_1)$ and $\mathbb{P}(S_1) > 0$ we initially

demonstrate

$$\mathbb{P}(T < \infty | S_1) = 1. \tag{3.14}$$

Firstly recall that observation that with W as in Lemma 3.1, up to null events, $\{W > 0\} = S_1$. Now T is the first arrival in a Cox process with intensity $\sum_{x \in \mathcal{N}^-(N)} \nu(x, N) Z_x(s)$, and hence $T < \infty$ with probability 1 on S_1 if

$$\int_0^\infty \sum_{x \in \mathcal{N}^-(N)} \nu_{x,N} Z_x(s) ds = \infty \tag{3.15}$$

with probability 1 on S_1. Observe that (3.15) is true as, by Theorem 3.1, for any realisation we may find $t_1 \geq 0$ such that

$$\int_{t_1}^\infty \sum_{x \in \mathcal{N}^-(N)} \nu(x, N) Z_x(s) ds \geq 2^{-1} \int_{t_1}^\infty \sum_{x \in \mathcal{N}^-(N)} \nu(x, N) W \phi_x e^{\lambda s} ds.$$

As each $\phi_x > 0$, due to each entry of $\tilde{v}$ also being positive [47], the right hand side of the above inequality is infinite a.s. on S_1 which gives (3.14). It remains to bound $\mathbb{P}(\{T < \infty\} \cap S_1^c)$.

Let us introduce $Z^*(t) = \sum_{i \in \mathcal{N}^+(1)} Z_i(t)$ and

$$\tau = \inf\{t \geq 0 : Z^*(t) > 0\}.$$

Note that we cannot hope to populate the target vertex without $Z^*(t)$ becoming positive. Therefore, $\{T < \infty\} \subseteq \{\tau < \infty\}$ which implies

$$\mathbb{P}(\{T < \infty\} \cap S_1^c) \leq \mathbb{P}(\{\tau < \infty\} \cap S_1^c).$$

Hence our interest turns to the right hand side of the preceding inequality. The distribution of τ is invariant to the growth of the types $i \in \mathcal{N}^+(1)$, and so we let $\alpha(i) = \beta(i) = 0$ for such i. The process $(Z_1(t), Z^*(t))$ is a two-type branching process where cells of the second type (that contribute to $Z^*(t)$) simply accumulate. For now assume $z = 1$. As

$$\begin{aligned} \mathbb{P}(\{\tau < \infty\} \cap S_1^c) &= (1 - \mathbb{P}(\{\tau = \infty\} | S_1^c)) \mathbb{P}(S_1^c) \\ &= \mathbb{P}(S_1^c) - \lim_{t \to \infty} \mathbb{P}(Z_1(t) = 0, Z^*(t) = 0), \end{aligned} \tag{3.16}$$

and from (3.8) we know $\mathbb{P}(S_1^c) = \beta/\alpha$, we now solve for the joint distribution of $(Z_1(t), Z^*(t))$.

We introduce the generating functions

$$G_1(x,y,t) = \mathbb{E}[x^{Z_1(t)}y^{Z^*(t)}|(Z_1(0),Z^*(0)) = (1,0)]$$
$$G^*(x,y,t) = \mathbb{E}[x^{Z_1(t)}y^{Z^*(t)}|(Z_1(0),Z^*(0)) = (0,1)].$$

Immediately $G^*(x,y,t) = y$ for all t. The backward Kolmogorov equation for G_1 is

$$\frac{\partial}{\partial t}G_1 = \alpha G_1^2 + \beta + \nu_1 G_1 G^* - (\alpha + \beta + \nu_1)G_1.$$

By solving this equation (separation of variables with a partial fraction decomposition), and using the initial condition we find

$$\mathbb{P}(Z_1(t) = 0, Z^*(t) = 0) = G_1(0,0,t) = \frac{r_2(1 - e^{\alpha_2(r_2 - r_1)t})}{1 - r_2 e^{\alpha_2(r_2-r_1)t}/r_1}.$$

With

$$r_1 = \frac{1}{2}\left(1 + q_1 + \nu_1' + \sqrt{(1 + q_1 + \nu_1')^2 - 4q_1}\right)$$
$$r_2 = \frac{1}{2}\left(1 + q_1 + \nu_1' - \sqrt{(1 + q_1 + \nu_1')^2 - 4q_1}\right)$$

where $q_1 = \beta_1/\alpha$ and $\nu_1' = \nu_1/\alpha$. Observe that $(1 + q_1 + \nu_1')^2 - 4q_1 > (1 - q_1)^2$ which implies $r_2 - r_1 < 0$. Hence

$$\lim_{t\to\infty} \mathbb{P}(Z_1(t) = 0, Z^*(t) = 0) = r_2.$$

We now have expressions for the final terms in (3.16) for $z = 1$. For $z \geq 1$, we can use independence, and this leads to

$$\mathbb{P}(\{\tau < \infty\} \cap S_1^c) = q_1^z - r_2^z.$$

By examining the leading order with respect to the relevant variables we can conclude. □

However we note that if we instead wish to look at the case when the vertex N population arises and does not go extinct then $\mathbb{P}(T = \infty) = \mathbb{P}$(All populations go extinct). The probability of population extinction is known to be given by an implicit equation involving the offspring distribution's generating function. This was the approach taken by [45]. The approximate

solution for small transition rates was derived and agrees with the result presented above.

The above proposition yields an approximation for the probability that the type N population will ever arise, that is

$$\mathbb{P}(T < \infty) \approx \mathbb{P}(S_1) = 1 - (\beta/\alpha)^z. \tag{3.17}$$

The distribution of T given S_1 can now be stated.

Corollary 3.2. *With the notation of Theorem 3.2,*

$$\lim_{\nu^* \to 0} \mathbb{P}(\lambda(T-\mu) > t | S_1) = \frac{[\lambda/\alpha(1+e^t)^{-1} + \beta/\alpha]^z - (\beta/\alpha)^z}{1-(\beta/\alpha)^z}$$

Proof. Starting from the first statement in Theorem 3.2, we only need apply $\mathbb{E}[\cdot|S_1]$. The distribution of $W\lambda/\alpha$ can be rewritten as

$$W\lambda/\alpha \stackrel{d}{=} \sum_{i=1}^{B} \xi_i'$$

where ξ_i' are rate one exponential variables, and $B \sim \text{Binomial}(z, \lambda/\alpha)$, independent of all ξ_i'. Conditioning on S_1 ensures that $B > 0$. Thus as in (3.13) we must calculate

$$\mathbb{E}\left[\exp\left(-e^t W\lambda/\alpha\right)|S_1\right] = \mathbb{E}\left[(1+e^t)^{-B}|B>0\right].$$

The binomial theorem leads to the stated expression. □

We remark that while Theorem 3.2, is not conditional on S_1, the rare trajectories when S_1^c occurs (the type 1 population dies) and $T < \infty$ are not captured by the current scaling. This is due to the fact that in these cases T will be much larger than μ.

From the above we can see that T is concentrated around the variable μ. The corollary below states that when we initiate with one cell, and the type 1 population survives, μ is the asymptotic median time to reach the target. If $z > 1$ the median is instead μ minus a z-dependent term.

Corollary 3.3. *Let $t_{1/2}$ be the median time for the type N population to arise,*

conditioned on S_1. That is $t_{1/2}$ satisfies

$$\mathbb{P}(T > t_{1/2}|S_1) = \frac{1}{2}. \tag{3.18}$$

Then

$$\lim_{\nu^* \to 0} |t_{1/2} - \mu - h(z)| = 0$$

where

$$h(z) = \lambda^{-1} \log \left(\frac{\lambda/\alpha}{2^{-1/z}(1 + (\beta/\alpha)^z)^{1/z} - \beta/\alpha} - 1 \right).$$

This also implies $t_{1/2} \sim \mu + h(z)$. Regarding $h(z)$ we have

$$h(1) = 0, \quad h(z) = \lambda^{-1} \log \left(\frac{\alpha \log(2)}{z\lambda} \right) + O(z)^{-1}, \quad z \to \infty.$$

Proof. We firstly note, as is apparent from the cumulative distribution function of T (see (3.9) for instance), that $t_{1/2}$ is monotone increasing as ν^* decreases. By (3.18)

$$\frac{1}{2} = \mathbb{P}(\lambda(T - \mu) > \lambda(t_{1/2} - \mu)|T < \infty).$$

Taking the $\nu^* \to 0$ limit we find $\lambda(t_{1/2} - \mu) \to x$, where the convergence is guaranteed as both $t_{1/2}$, μ are monotone increasing. Further, from Corollary 3.2, the limit x satisfies

$$\frac{[\lambda/\alpha(1 + e^x)^{-1} + \beta/\alpha]^z - (\beta/\alpha)^z}{1 - (\beta/\alpha)^z} = \frac{1}{2}.$$

Solving for x leads to $x = \lambda h(z)$ as stated. For the asymptotic result we firstly consider

$$\begin{aligned} d(z) &= \frac{\lambda/\alpha}{2^{-1/z}(1 + (\beta/\alpha)^z)^{1/z} - \beta/\alpha} - 2^{\frac{\alpha}{\lambda z}} \\ &= \frac{2^{1/z}\lambda/\alpha - 2^{\frac{\alpha}{\lambda z}}(1 + (\beta/\alpha)^z)^{1/z} + 2^{\frac{1}{z} + \frac{\alpha}{\lambda z}}\beta/\alpha}{(1 + (\beta/\alpha)^z)^{1/z} - 2^{1/z}\beta/\alpha}. \end{aligned}$$

Expansions on the numerator and denominator show that $d(z)$ is $O(z^{-2})$. Then

$$\begin{aligned} \lambda h(z) &= \log(2^{\frac{\alpha}{\lambda z}} - 1 + d(z)) \\ &= \log(2^{\frac{\alpha}{\lambda z}} - 1) + \frac{d(z)}{2^{\frac{\alpha}{\lambda z}} - 1} + \dots \end{aligned}$$

As $(2^{\frac{\alpha}{\lambda z}} - 1)^{-1} = O(z)$ we have shown the claimed asymptotic behaviour.

□

3.2.3 Path costs/benefits and the case of acyclic graphs

The sequence $(\phi_x)_{x=1}^{N-1}$ features prominently in both the main results of the previous section, Theorems 3.1 and 3.2. In applications we wish to have explicit representations for $(\phi_x)_{x=1}^{N-1}$ to assess the impact of parameters. It will become apparent that in the setting where G is acyclic, the ϕ_x can be written in terms of the intuitively comprehensible path costs and benefits which we now introduce.

Path costs and benefits

Consider any path from the root to the target vertex. Intuitively, if the transition rates encountered along this path are larger, the target population will be seeded from this path quicker. Conversely the smaller the growth rates along this path are, the target will be reached slower. This motivates the concept of the cost and benefit of paths which we now define for paths between the root and any vertex.

Let the set of paths between the root and vertex x be denoted $\mathcal{P}_{1,x}$. So that we can discuss the root we let $\mathcal{P}_{1,1} = \{(1)\}$. Any $p \in \mathcal{P}_{1,x}$, will be a sequence $p = (p_1, p_2, \ldots, p_l, p_{l+1})$, where each $1 \leq p_i \leq N$, l is the path length, all p_i are distinct, and $p_1 = 1$, $p_{l+1} = x$. Along a path p, let us call the difference between the growth rate at the root and the vertices encountered the fitness costs. The path benefit, $b(p)$ and cost, $c(p)$, are defined in terms of the transition rates and fitness costs encountered on the path in the following way

$$b(p) = \prod_{i=1}^{l} \nu(p_i, p_{i+1}), \quad c(p) = \prod_{i=2}^{l} (\lambda - \lambda(p_i)) \,. \tag{3.19}$$

Throughout the empty product is set to 1.

Path graphs

To see the reason for introducing these quantities let us consider the case when G is a path graph. Path graphs, or linear graphs, are graphs where every vertex has indegree and outdegree 1, except vertex 1 (indegree 0 and outdegree 1) and vertex N (indegree 1 and outdegree 0). An example is shown in Figure 3.3.

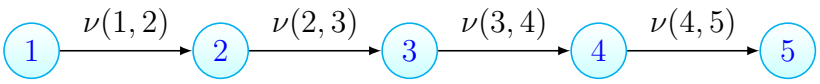


Figure 3.3 *A path graph on 5 vertices*

Let the path between the root and the target vertex be denoted $\tilde{p}$. Further let the first x terms in $\tilde{p}$ be denoted

$$\tilde{p}_{1:x} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_x).$$

Observe that for $x \geq 2$, $\mathcal{P}_{1,x} = \{\tilde{p}_{1:x}\}$, and

$$b(\tilde{p}_{1:x}) = \prod_{i=1}^{x-1} \nu(i, i+1), \quad c(\tilde{p}_{1:x}) = \prod_{i=2}^{x-1} (\lambda - \lambda(i)).$$

In this setting the matrix A from Section 3.2.1 has entries

$$a_{ij} = \begin{cases} \lambda(i) & j = i \\ \nu(i-1, j) & j = i-1 \\ 0 & \text{otherwise,} \end{cases}$$

and the vector ψ has entries $\psi_i = \alpha(i) + \beta(i) + \nu(i, i+1)$ for $i = 1, \ldots, N-1$. We want a useful representation of the sequence $(\phi_x)_{x=1}^{N-1}$, which we recall is given in terms of the entries of the suitably normalised left and right eigenvectors of A. Because they are easier to work with (i.e. check they indeed are eigenvectors) we first introduce the unnormalised left and right vectors u', v' with entries

$$u'_i = \begin{cases} 1, & i = 1 \\ 0 & 2 \leq i \leq N-1 \end{cases} \qquad v'_i = \prod_{j=i+1}^{N} \frac{\lambda - \lambda(j)}{\nu(j-1, j)}, \quad 1 \leq i \leq N-1$$

It may be directly verified that these are left and right eigenvectors for A with eigenvalue λ. Now we normalise these, so that they agree with (3.5) via

$$\tilde{u} = \frac{\psi \cdot v'}{v'_1} u', \quad \tilde{v} = \frac{1}{\psi \cdot v'} v'.$$

The following corollary demonstrates how the results of the previous section on population growth and time to reach the target simplifies in the path graph setting.

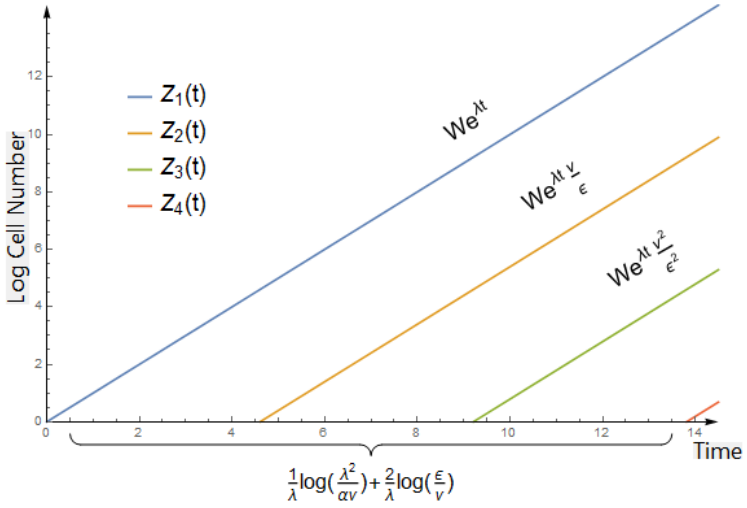


Figure 3.4 *Illustration of the asymptotic behaviour of the population growth and the median time until the 4th population arises in a path graph. Here all transition rates are ν and all fitness costs are ϵ, as in* (3.20)

Corollary 3.4. *The results of Sections 3.2.1 and 3.2.2 hold with the explicit representations*

$$\phi_x = \frac{b(\tilde{p}_{1:x})}{c(\tilde{p}_{1:x})}$$

$$\nu^* = \nu(N-1, N),$$

and

$$\mu = \lambda^{-1} \log\left[\lambda^2 (\alpha\phi_N)^{-1}\right].$$

Let us explicitly consider the case of equal transition rates and fitness costs, say $\nu = \nu(j-1, j)$ and $\epsilon = \lambda - \lambda(j)$. We see for path graphs, asymptotically $Z_{i+1}(t)$ is related to $Z_i(t)$ by a multiplicative factor of $\frac{\nu}{\epsilon}$. For $z = 1$ we have the following approximation for the median time to reach the target. For small ν

$$t_{1/2} \approx \frac{1}{\lambda} \log\left(\frac{\lambda^2}{\alpha n \nu}\right) + \frac{N-2}{\lambda} \log\left(\frac{\epsilon}{\nu}\right). \tag{3.20}$$

The first term on the right hand side comes from waiting for the first transition from the type 1 population, while the second is due transitions between the remaining types. The first term is distinct as the type 1 growth is the main cause of stochasticity, as mentioned in the discussion after Theorem 3.1. The population growth and time until initiation of the type N cells is illustrated for this case in Figure 3.4.

Acyclic graphs

We now consider the case where G is acyclic. Representations of the sequence $(\phi_x)_{x=1}^{N-1}$ are obtainable via algebraic methods, but instead we deduce the same representations via the concept of vertex lineages. This construction is more intuitive and will be required when discussing which path the target vertex is seeded from.

For any existing cell we associate a vertex lineage $p = (p_1, \ldots, p_{l+1})$, where each $p_i \in V$ and $l+1 = |p|$. This represents the types of cells in a given cells ancestral lineage. For example we may have a type-2 cell with vertex lineage $p = (1, 3, 4, 2)$. Each lineage begins with 1 and ends with the present type of the cell. It is clear a vertex lineage is a path in G of length l. As above we let the set of paths from vertex 1 to vertex x be $\mathcal{P}_{1,x}$. We denote the number of cells at time t with vertex lineage p as $X(p,t)$.

The number of cells at a vertex is related to the cells with vertex lineages ending at that vertex via

$$Z_x(t) = \sum_{p \in \mathcal{P}_{1,x}} X(p,t). \tag{3.21}$$

For any path p, as above we let the first j terms in p be denoted $p_{1:j}$. The process $(X(p_{1:j},t))_{j=1}^{l+1}$ is a multitype branching process, and $X(p_{1:j},t)$ represents the number of cells that have progressed to step j along path p. Births occur to individuals of type i at at rate $\alpha(p_i)$, deaths at rate $\beta(p_i)$ and individuals of type i transition to type $i+1$ at rate $\nu(p_i, p_{i+1})$. Thus we may use our results for path graphs with vertices labelled $1, \ldots, l+1$. Hence via Theorem 3.1 and Corollary 3.4 we have the following.

Corollary 3.5.

$$\lim_{t\to\infty} e^{-\lambda t} X(p,t) = \frac{Wb(p)}{c(p)} \quad a.s.$$

where $b(p)$, $c(p)$ are the path benefit and cost given in (3.19) *and W is as in Lemma 3.1.*

Analogously to Corollary 3.4 we have the following statement regarding the growth of the initial and intermediate population types.

Corollary 3.6. *The results of Section 3.2.1 hold with the explicit representations*

$$\phi_x = \sum_{p \in \mathcal{P}_{1,x}} \frac{b(p)}{c(p)} \quad 1 \leq x \leq N-1,$$

$b(p)$, $c(p)$ are the path benefit and cost given in (3.19)*.*

Note that as the empty product is defined to be 1, $\phi_1 = 1$. Before giving the proof of the above corollary we give a needed lemma.

Lemma 3.3. *Let $A_1, \ldots, A_n$ be events such that $\mathbb{P}(A_i) = 1$ for every i. Then $\mathbb{P}(\cap_i^n A_i) = 1$.*

Proof.

$$\mathbb{P}(\cap_{i=1}^n A_i) = 1 - \mathbb{P}(\cup_{i=1}^n A_i^c) \geq 1 - \sum_{i=1}^n \mathbb{P}(A_i^c) = 1$$

□

Proof of Corollary 3.6. Due to Theorem 3.1 we must demonstrate

$$\lim_{t \to \infty} e^{-\lambda t} Z_x(t) = W \sum_{p \in \mathcal{P}_{1,x}} \frac{b(p)}{c(p)} \quad a.s.$$

By (3.21) it is enough to show

$$\mathbb{P}\left(\bigcap_{p \in \mathcal{P}_{1,x}} \lim_{t \to \infty} e^{-\lambda t} X(p,t) = W \right) = 1.$$

Let $A_p = \{\lim_{t \to \infty} e^{-\lambda t} X(p,t) = W\}$ and use Lemma 3.3 to conclude. □

The results of Section 3.2.2 (regarding the time to reach the target) hold also but they are simpler to state after some further notation has been introduced. This notation will also be useful for discussing the distribution of paths to the target.

Note that $|\mathcal{P}_{1,N}|$ is finite, and so we let

$$n = |\mathcal{P}_{1,N}|.$$

The elements of $\mathcal{P}_{1,N}$ can be enumerated and we use the following notation to do

so

$$\mathcal{P}_{1,N} = \{p^{(1)}, \ldots, p^{(n)}\}.$$

For each $p^{(i)} \in \mathcal{P}_{1,N}$ we let $l^{(i)} = |p^{(i)}| - 1$, i.e. the path lengths and $u^{(i)} = \nu(p^{(i)}_{l^{(i)}}, p^{(i)}_{l^{(i)}+1})$ be the final transition rates. Note that $\nu^* = \max_{1 \le i \le n}\{u^{(i)}\}$. Then we have the following.

Corollary 3.7. *The results of Section 3.2.2 hold with the explicit representation*

$$\mu = \lambda^{-1} \log \left(\lambda^2 \left[\alpha \sum_{i=1}^{n} \frac{b(p^{(i)})}{c(p^{(i)})} \right]^{-1} \right).$$

As the time to reach the target T is concentrated around μ it is reassuring that as the path benefits increase, μ decreases, while if the path costs increase, μ increases.

3.2.4 Path to target

We now consider which path leads to the type N population arising. We expect paths with higher benefit to cost ratios to be more likely and this is confirmed. If we only consider the time to reach N along path i, then our result on path graphs, Corollary 3.4, describes this. Therefore to determine whether the target is reached first along path i, we must only check if this time is the minimum compared to the other paths.

Consider $p^{(i)} \in \mathcal{P}_{1,N}$ and recall that $X(p^{(i)}_{1:j}, t)$ represents the number of cells that have progressed to step j along path i. If we apply the first statement in Theorem 3.2 to the branching process $(X(p^{(i)}_{1:j}, t))_{j=1}^{l^{(i)}+1}$, then as in the reduced model of Section 3.1.3, we can deduce that under a time rescaling, the time to traverse path i is approximately exponentially distributed, conditional on the population growth along that path. Therefore the question of which path the target population arises from becomes equivalent to the minimum of a set of exponential random variables. To state this more precisely, let the time taken to reach the target via path i be

$$T^{(i)} = \inf\{t \ge 0 : \, X(p^{(i)}_{l^{(i)}+1}, t) > 0\}.$$

Connecting to the earlier discussion, note that

$$T = \min_{1 \le i \le n} \{T^{(i)}\}.$$

We may have occasion, for instance see Section 3.3.2, to not only ask which path to the target is fastest, but is one set of paths faster than another. Therefore we further introduce, for non-empty $\mathcal{J} \subset \{1, \ldots, n\}$, the notation

$$T^{\mathcal{J}} = \min_{j \in \mathcal{J}} \{T^{(j)}\}.$$

Again the final transition rates along each path will be assumed small and so to single these out, for path i between the root and the target, we suppose

$$b(p^{(i)}) = b_1(p^{(i)})u^{(i)}.$$

where again $b(p^{(i)})$ is the path benefit of path i. Observe that $b_1(p^{(i)})$ is simply the product of all but the final transition rates along path i. We will be taking the limit of small ν^* and in order to compare the final transitions we henceforth assume there exists $(\kappa^{(i)})_{i=1}^n$ such that

$$\lim_{\nu^* \to 0} \frac{u^{(i)}}{u^{(1)}} = \kappa^{(i)} > 0.$$

Then the following theorem, which concerns one set of paths populating the target x time units before another set of paths holds. Note that which set of paths comes first is dealt with by $x = 0$.

Theorem 3.3. *For non-empty disjoint $\mathcal{J}_1, \mathcal{J}_2 \subset \{1, \ldots, n\}$ and $x \in \mathbb{R}$ we have*

$$\lim_{\nu^* \to 0} \mathbb{P}(T^{\mathcal{J}_1} + x < T^{\mathcal{J}_2} | S_1) = \frac{\sum_{j \in \mathcal{J}_1} \frac{\kappa^{(j)} b_1(p^{(j)})}{c(p^{(j)})}}{\sum_{j \in \mathcal{J}_1} \frac{\kappa^{(j)} b_1(p^{(j)})}{c(p^{(j)})} + e^{\lambda x} \sum_{j \in \mathcal{J}_2} \frac{\kappa^{(j)} b_1(p^{(j)})}{c(p^{(j)})}}. \tag{3.22}$$

Proof. Consider initially the case when $n = 2$, that is there are only two paths from vertex 1 to vertex N. These two paths must diverge before N. If the population growth along these paths until the target vertex is given, then $T^{(1)}$ and $T^{(2)}$ are the first arrival times in two independent non-homogeneous Poisson processes. Returning to arbitrary n, to use this fact we define the population numbers along all the paths up until the target vertex as

$$\mathcal{X}(t) = (X(p_{1:j}^{(i)}, t))_{i=1,j=1}^{n,l^{(i)}}.$$

So that we can use this conditional independence, and prior results, we note

$$\mathbb{P}(T^{\vartheta_1} + x < T^{\vartheta_2} | S_1) = \frac{\mathbb{E}[\mathbb{I}_{S_1} \mathbb{P}(T^{\vartheta_1} + x < T^{\vartheta_2} | (\mathcal{X}(t))_{t \geq 0}]}{\mathbb{P}(S_1)}. \tag{3.23}$$

We now focus on the argument of the above expectation.

We also know that each $T^{(i)}$ appropriately centred converges. If we centre each $T^{(i)}$ by the same quantity, then the ordering is preserved. That is, clearly $T^{(1)} < T^{(2)} \iff T^{(1)} - \mu^{(1)} < T^{(2)} - \mu^{(1)}$ for any $\mu^{(1)}$. Let us use the shorthand notation $\eta^{(i)} = b_1(p^{(i)})/c(p^{(i)})$. As $T^{(i)}$ is the target time along the path graph $p^{(i)}$, Corollary 3.4 gives the appropriate centring. In particular with

$$\mu^{(i)} = \lambda^{-1} \log(\lambda^2 [\alpha \eta^{(i)} u^{(i)}]^{-1})$$

we have

$$\lim_{\nu^* \to 0} \mathbb{P}(\lambda(T^{(i)} - \mu^{(i)}) > t | (\mathcal{X}(t))_{t \geq 0}) = \exp(-e^t W \lambda / \alpha) \, a.s.$$

For order preservation (of the $T^{(i)}$s) we take $\mu^{(1)}$ as a common centring of all $T^{(i)}$, which is possible as $T^{(i)} - \mu^{(1)} = T^{(i)} - \mu^{(i)} + \mu^{(i)} - \mu^{(1)}$ and

$$\lim_{\nu^* \to 0} \mu^{(i)} - \mu^{(1)} = \lambda^{-1} \log \left(\frac{\eta^{(1)}}{\eta^{(i)} \kappa^{(i)}} \right).$$

Using this, and exponentiating (which again preserves order), leads to

$$\lim_{\nu^* \to 0} \mathbb{P}(\exp[\lambda(T^{(i)} - \mu^{(1)})] > t | (\mathcal{X}(t))_{t \geq 0}) = \exp \left(- \frac{t W \lambda \eta^{(i)} \kappa^{(i)}}{\alpha \eta^{(1)}} \right) \, a.s.$$

Recall the $T^{(i)}$ are conditionally independent, and hence convergence of the marginal distributions implies joint convergence. Therefore for $(t_j)_{j=1}^n$ with each $t_j \geq 0$, almost surely we have

$$\lim_{\nu^* \to 0} \mathbb{P} \left[\left(\exp[\lambda(T^{(j)} - \mu^{(1)})] > t_j \right)_{j=1}^n | (\mathcal{X}(t))_{t \geq 0} \right] = \prod_{j=1}^n \exp \left(- \frac{t_j W \kappa^{(j)} \eta^{(j)} \lambda}{\eta^{(1)} \alpha} \right).$$

Thus for fixed $(\mathcal{X}(t))_{t \geq 0}$ the random vector $(\exp[\lambda(T^{(j)} - b_1) + \lambda/\alpha])_{j=1}^n$ jointly tends in distribution to a mixture of independent exponential random variables and a point mass at infinity (when $W = 0$). The shift (the x term) is dealt with

by noting that

$$\lim_{\nu^* \to 0} \mathbb{P}(e^{\lambda x} \exp[\lambda(T^{(j)} - \mu^{(1)})] > t | (\mathcal{X}(t))_{t \geq 0}) = \exp\left(-\frac{tW e^{-\lambda x} \kappa^{(j)} \eta^{(j)} \lambda}{\alpha \eta^{(1)}} \right).$$

We see that so long as $W > 0$, we may use standard results for the exponential distribution to deduce the right hand side of (3.22). We know, up to null events, $\{W > 0\} = S_1$ and hence almost surely

$$\lim_{\nu^* \to 0} \mathbb{I}_{S_1} \mathbb{P}(T^{\mathcal{J}_1} + x < T^{\mathcal{J}_2} | (\mathcal{X}(t))_{t \geq 0}) = \mathbb{I}_{S_1} \frac{\sum_{j \in \mathcal{J}_1} \kappa^{(j)} \eta^{(j)}}{\sum_{j \in \mathcal{J}_1} \kappa^{(j)} \eta^{(j)} + e^{\lambda x} \sum_{j \in \mathcal{J}_2} \kappa^{(j)} \eta^{(j)}}.$$

Coupling the above with (3.23) gives the result. □

Thus an approximate distribution appears over the possible paths from the root to the target, where the probability of a given path is simply proportional to its benefit to cost. That is, using $\kappa^{(i)} \approx \kappa^{(1)} u^{(i)}$,

$$\mathbb{P}(T^{(i)} = T | S_1) \approx \frac{\frac{b(p^{(i)})}{c(p^{(i)})}}{\sum_{p \in \mathcal{P}_{1,N}} \frac{b(p)}{c(p)}} \tag{3.24}$$

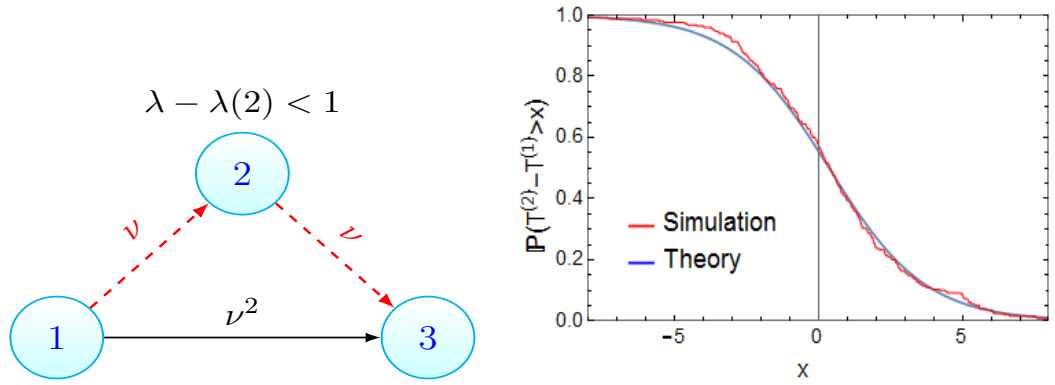
In particular we can examine two paths, i and j, and ask along which path is the target reached faster. The above result then yields the

$$\mathbb{P}(T^{(i)} < T^{(j)} | S_1) \approx \frac{\frac{b(p^{(i)})}{c(p^{(i)})}}{\frac{b(p^{(i)})}{c(p^{(i)})} + \frac{b(p^{(j)})}{c(p^{(j)})}}.$$

This expression is intuitively appealing, clearly showing the trade-off between the relative path benefits and costs.

A perhaps surprising consequence is the impact of paths with costs less than 1. To illustrate this, let us consider the case depicted in Figure 3.5a with $N = 3$, and paths $p^{(1)} = (1, 2, 3)$, $p^{(2)} = (1, 3)$ and $\nu = \nu(1, 2) = \nu(2, 3)$, $\nu(1, 3) = \nu^2$. Naively it may be expected that as the product of transition rates along both these paths are equal, and the first path contains a vertex with a fitness cost, the second path is more likely. However the result above yields

$$\mathbb{P}(T^{(2)} - T^{(1)} > x) \approx \frac{\frac{\nu}{\lambda - \lambda(2)}}{\frac{\nu}{\lambda - \lambda(2)} + e^{\lambda x} \nu} = \frac{1}{1 + (\lambda - \lambda(2)) e^{\lambda x}}. \tag{3.25}$$



(a) *Despite the red-dashed path containing a fitness valley, the type 3 population is more likely to arise via this path. See* (3.25).

(b) *Monte Carlo simulations of system in (a) vs theory* (3.25). *Parameters* $\alpha = 0.9$, $\beta = 0.3$, $\lambda = 0.6$, $\alpha(2) = 0.2$, $\beta(2) = 0.4$, $\lambda(2) = -0.2$, $\nu = 0.1$, *runs* $= 250$.

Figure 3.5 *Crossing valleys can be faster.*

for small ν. Thus the first path is more probable, i.e. $\frac{\mathbb{P}(T^{(2)}>T^{(1)})}{\mathbb{P}(T^{(1)}>T^{(2)})} > 1$, if $\lambda - \lambda(2) < 1$. That crossing a fitness valley can be faster is also confirmed by simulation, as is shown in Figure 3.5b. Note that for the simulation $\nu > 0.1\alpha$, demonstrating that for certain situations ν need not be very small for the above to be a good approximation.

3.2.5 Cyclic graphs and extensions

In Sections 3.2.3 and 3.2.4 it was demonstrated that the time until the target is populated and the path distribution have explicit representations in terms of the path benefits and costs. In this section, we allow for cycles in G and demonstrate that Theorem 3.3 holds for finite length walks. Further when transition rates are small over all edges, the target vertex is populated via a path. We follow this by two remarks on extensions of the model for which our results hold.

Cyclic graphs

Suppose G is as before, but without the acyclic assumption. For cyclic graphs in place of $\mathcal{P}_{1,N}$ we must consider $\mathcal{W}_{1,N}$, which we define to be the set of walks between the root and vertex N. The definition of vertex lineages holds identically

for walks. Therefore for $w \in \mathcal{W}_{1,N}$ we let

$$T_w = \inf\{t \geq 0 : X(w,t) > 0\} \tag{3.26}$$

where $X(w,t)$ is as given in Section 3.2.3. Here we show that so long as the number of transitions between vertices is finite, which corresponds to finitely many back transitions, we may map a graph containing cycles to an acyclic one and then use the results given above.

The set of walks between the root and vertex N of length at most $i-1$ will be denoted

$$\mathcal{W}_{1,N}^{(i)} = \{w \in \mathcal{W}_{1,N} : |w| \leq i\}.$$

Further for a graph $G = (V, E)$, let the vertex parameters be the birth and death rates associated to the vertices of G and the edge parameters the transition rates associated with the edges.

Proposition 3.2. *For a given graph $G = (V, E)$ with associated vertex and edge parameters and fixed $i \geq 1$, we can construct an acyclic graph $G' = (V', E')$ (dependent on i), which coupled with appropriate vertex and edge parameters, permits a bijection $g : \mathcal{W}_{1,N}^{(i)} \mapsto \mathcal{P}'_{1,|V'|}$. This bijection possesses the property that for $w \in \mathcal{W}_{1,N}^{(i)}$*

$$T_w \stackrel{d}{=} T'_{g(w)}$$

where T' is defined as in (3.26) *but with the process on G' with its associated parameters.*

Before giving the proof we briefly demonstrate this on the graphs shown in Figure 3.6. This elucidates the general idea.

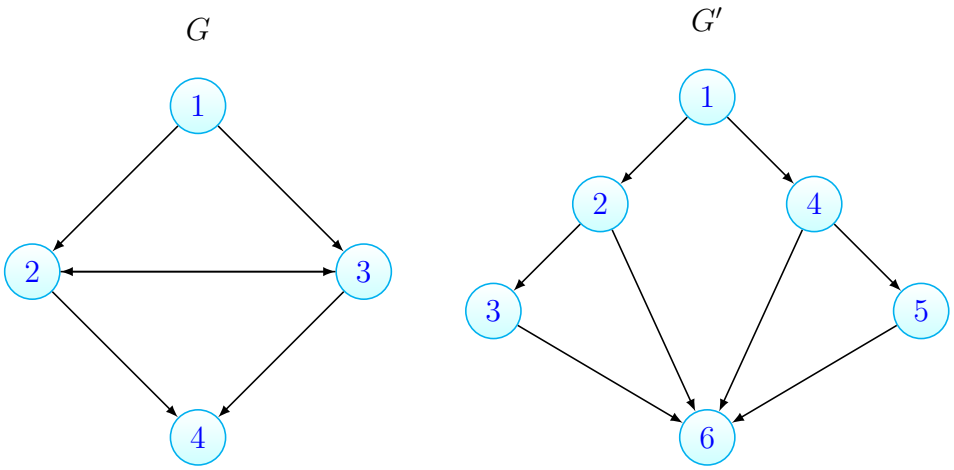


Figure 3.6 *Mapping cyclic graphs to acyclic graphs*

Here

$$\mathcal{W}^{(4)}_{1,N} = \{(1,2,4),(1,2,3,4),(1,3,4),(1,3,2,4)\}.$$

It is clear these may be mapped to the paths in G'

$$\mathcal{P}'_{1,|V'|} = \{(1,2,6),(1,2,3,6),(1,4,6),(1,4,5,6)\}.$$

Further $T_{(1,2,4)} \stackrel{d}{=} T'_{(1,2,6)}$ so long as the transition rates associated with edges $(1,2)$ and $(2,4)$ on G are equal to the rates with $(1,2)$ and $(2,6)$ on G' and that the birth and death rates at vertices 1 and 2 are equal in both cases.

Proof of Proposition 3.2. To avoid cumbersome notation and as the idea is illustrated in Figure 3.6 we sketch an algorithm. We enumerate the set $\mathcal{W}^{(i)}_{1,N} = \{w^{(1)}, w^{(2)}, \dots w^{(\mathcal{M})}\}$, where $\mathcal{M} = |\mathcal{W}^{(i)}_{1,N}|$. Again we use the notation $l^{(j)}$ for the length of the jth walk, that is $l^{(j)} = |w^{(j)}| - 1$. The graph G' is constructed iteratively. Let $V'_1 = \{1, 2, \dots, l^{(1)}, N'\}$, where N' is a placeholder and will ultimately be $|V'|$. Similarly $E'_1 = \{(1,2), \dots, (l^{(1)}, N')\}$ and $p^{(1)} = (1, 2, \dots, l^{(1)}, N')$.

Then for each further $w^{(j)}$, $2 \leq j \leq \mathcal{M}$, from the previously considered walks select the walk that $w^{(j)}$ 'agrees with the longest', say $w^{(k)}$, $1 \leq k < j$. If $w^{(j)}$ and $w^{(k)}$ agree up until their mth element, and $l^{(j)} - m > 0$ then add $l^{(j)} - m$ elements to V'_{j-1} (new vertices starting with $|V'_{j-1}|$). This creates V'_j. Denote the new vertices as $(v^{(j)}_x)^{l^{(j)}-m}_{x=1}$. To E_{j-1} add the edges $(p^{(k)}_m, v^{(j)}_1)$, $\{(v^{(j)}_x, v^{(j)}_{x+1})\}^{l^{(j)}-m-1}_{x=1}$ and $(v^{(j)}_{l^{(j)}-m}, N')$. This creates E_j. For $p^{(j)}$ we let $p^{(j)}_{1:m} = p^{(k)}_{1:m}$ and then add the new vertices of V'_j, and terminate the path at N'. If $l^{(j)} = m$ (and so the walk $w^{(j)}$

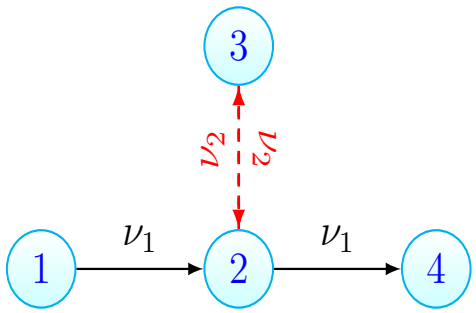


Figure 3.7 *Walks containing cycles can be faster. The walk of length 4 that traverses the red edges with transition rate ν_2 can still be faster than the length 2 path. See* (3.27).

agrees with $w^{(k)}$ until its final element) let $V'_j = V'_{j-1}$ and add the edge $(p^{(k)}_m, N')$ to E'_{j-1}. Then $p^{(j)}$ is equal to $p^{(k)}_{1:m}$ concatenated with N'. Finally our constructed graph is $G' = (V', E') = (V'_{\mathcal{M}}, E'_{\mathcal{M}})$ with $N' = |V'|$.

Identifying $g(w^{(j)}) = p^{(j)}$, we see that if $w^{(j)}$, $w^{(k)}$ agree until their mth element then so also will $g(w^{(j)})$, $g(w^{(k)})$, but after this the paths diverge. We further require that the birth and death rates for the population type represented by the kth element of w should be the same as the kth element of $g(w)$. Similarly for the transition rates between the kth and $k+1$th element of w and $g(w)$.

□

Coupling the above proposition with the results of Section 3.2.3 allows one to characterise

$$\min_{w \in \mathcal{W}^{(k)}_{1,N}} T_w.$$

Further for a set of walks $\mathcal{W}^{(i)}_{1,N}$, Theorem 3.3 holds but now with the walk benefit and costs, defined analogously to the definition given in (3.19).

We immediately use this to explore the concept of walks (with cycles) or paths seeding the target. Consider the graph and edge parameters displayed in Figure 3.7. Let $p = (1, 2, 4)$ and $w = (1, 2, 3, 2, 4)$. Despite the extra length, by using the above proposition with Theorem 3.3, for small ν_1 we see

$$\mathbb{P}(T_w < T_p | S_1) \approx \left(\frac{\frac{\nu_1^2 \nu_2^2}{(\lambda - \lambda(2))^2 (\lambda - \lambda(3))}}{\frac{\nu_1^2 \nu_2^2}{(\lambda - \lambda(2))^2 (\lambda - \lambda(3))} + \frac{\nu_1^2}{\lambda - \lambda(2)}} \right) = \left(1 + \frac{(\lambda - \lambda(2))(\lambda - \lambda(3))}{\nu_2^2} \right)^{-1}. \tag{3.27}$$

Thus the walk may be faster.

However, the following result demonstrates that when the transition rates between vertices are small, the first initiation of the target is dominated by the paths, as opposed to walks containing cycles, which offers a secondary justification for mainly focusing on acyclic graphs. To state the result let us define $q : \mathcal{W}_{1,N} \mapsto \mathcal{P}_{1,N}$ as the operation reducing walks to their respective paths. This may be accomplished recursively by searching for the first cycle in w and replacing this with the first element in the cycle. Then repeating the same procedure on the reduced walk until it is a path. This operation is the same as that which maps trajectories to their corresponding trajectory class in [80]. Further, for $p \in \mathcal{P}_{1,N}$ let

$$\mathcal{W}^{(i)}[p] = \{w \in \mathcal{W}^{(i)}_{1,N} : q[w] = p, w \neq p\}$$

that is the walks of length at most $i-1$ which when reduced to paths are p, but excluding p itself. Then with

$$\nu_{\max} = \max\{\nu(i,j) : (i,j) \in E\}$$

we have the following.

Proposition 3.3. *For $i \geq 1$, $p \in \mathcal{P}_{1,N}$ such that $\mathcal{W}^{(i)}[p]$ is non-empty, and all $w \in \mathcal{W}^{(i)}[p]$*

$$\lim_{\nu_{\max}\to 0} \lim_{\nu^*\to 0} \mathbb{P}(T_p < T_w | S_1) = 1.$$

Proof. We construct an acyclic graph with appropriate vertex and edge parameters as in Proposition 3.2 and use the mapping g from there with domain $\mathcal{W}^{(i)}_{1,N}$. As before let $\mathcal{P}'_{1,|V'|}$ be the codomain of g. Then for path $p \in \mathcal{W}^{(i)}_{1,N}$ and walk $w \in \mathcal{W}^{(i)}[p]$, $g(p)$, $g(w) \in \mathcal{P}'_{1,|V'|}$. Note due to the manner in which the edge parameters are chosen $\{\nu(i,N) : i \in \mathcal{N}^-(N)\} = \{\nu_{i,|V'|} : i \in \mathcal{N}^-(|V'|)\}$ and hence ν^* remains unchanged. Similarly $\nu_{\max}$ is unchanged. Therefore,

$$\lim_{\nu^*\to 0} \mathbb{P}(T_p < T_w | S_1) = \lim_{\nu^*\to 0} \mathbb{P}(T'_{g(p)} < T'_{g(w)} | S_1).$$

Furthermore each vertex and edge in p is also present in w. Hence the benefit of $g(w)$ satisfies $b(g(w)) = b(g(p))\tilde{b}(g(w))$ where $\tilde{b}(g(w))$ is the product of transition rates along the edges existing in cycles in w. Define $\tilde{c}(g(w))$ similarly for the costs pertaining to vertices in cycles in w, and $\Xi = \tilde{b}(g(w))/\tilde{c}(g(w))$. Then by Theorem 3.3

$$\lim_{\nu^*\to 0} \mathbb{P}(T'_{g(p)} < T'_{g(w)} | S_1) = \frac{1}{1+\Xi}.$$

As Ξ contains transition rates (from $\tilde{b}(g(w))$), this leads to the stated result upon

taking $\nu_{\max} \to 0$. □

Extensions

We remark also that Theorem 3.3 would also hold for a slightly more general branching process. Keeping the transition process as it is, we can have a type x cell giving birth to j type x cells at a rate $\alpha(x)\sigma_{x,j}$ where for each x, $(\sigma_{x,j})_{j\geq 0}$ is a probability distribution with finite second moment and mean $\bar{\sigma}_x$. We now have no need for $\beta(x)$. In this setting if we redefine $\lambda(x) = \alpha(x)(\bar{\sigma}_x - 1)$ then the path distribution will be unchanged. This is due to the fact that the key result of [47] used in the proof of Theorem 3.1 holds then also. However for the proofs, the definition of the vector $(\psi_x)_{x=1}^{N-1}$ in (3.4) should be amended to $\psi_x = \alpha(x) + \sum_{y\in\mathcal{N}^+(x)} \nu(x,y)$. This may be relevant to the case of viral dynamics, where an infected cell can release multiple virions upon bursting [70].

Throughout we have assumed the population is initiated at the root vertex and we are interested in the time for a single target population to arise. It may be of interest to consider several target populations, and then ask which of these populations arises first. Generally these populations can be grouped together and the above results will still apply. However if, as in the discussion directly preceding Section 3.2, we wait for the first individual to arise whose progeny does not go extinct, a minor modification occurs. For example if we have two target vertices labelled N_1, N_2 we should map $\nu(x, N_1) \mapsto \nu(x, N_1)\lambda(N_1)/\alpha(N_1)$ for x a neighbour of N_1 and $\nu(y, N_2) \mapsto \nu(y, N_2)\lambda(N_2)/\alpha(N_2)$ for y a neighbour of N_2. Other numbers of target vertices are treated analogously.

3.3 Applications

We now consider some applications of the preceding section. Firstly we consider the impact of imperfect drug penetration on resistance. Recently, two theoretical works have explored how resistance spreads in this setting, and have shown that poor penetration can accelerate resistance [33, 65]. We are able to recover and extend the findings of [33, 65]. Next the ordering by which resistance-causing mutations accrue is explored, firstly in the setting of cancer and then in bacterial infections. In the case of bacterial infections we explore how the risk of multidrug resistance depends on mutation rates, recovering simulation results

in [32]. Underlying the model in each of these scenarios is an acyclic graph, and therefore we will be investigating these applications with the approximate formulas implied by the results of Section 3.2.3. These are given in Table 3.1.

Full statement	Implied approximate formula
Theorem 3.2 (time to target)	$\mathbb{P}(T > t) \approx \left(\frac{\lambda/\alpha}{1+e^{\lambda(t-\mu)}} + \beta/\alpha\right)^z$
Corollary 3.3 (median time to target)	$t_{1/2} \approx \mu + h(z)$
Theorem 3.3 (path distribution)	$\mathbb{P}(T^{(i)} = T \mid S_1) \approx \frac{\frac{b(p^{(i)})}{c(p^{(i)})}}{\sum_{p \in \mathcal{P}_{1,N}} \frac{b(p)}{c(p)}}$

Table 3.1 *Approximate formulas implied by the results of Section 3.2 when the final transition rates are small. $b(p)$, $c(p)$ are again the path benefits and costs, μ is a simple function of these (see Corollary 3.7) and $h(z)$ is as given in Corollary 3.3*

.

Before continuing to the applications we make a remark on the interpretation of the approximate formulas. The results of Section 3.2 hold when the final transition rates in each path tends to 0. Therefore in the present section all transition rates, in particular the mutation and migration rates discussed below, will be taken to be small and statements are to be interpreted as approximations that hold true in this limiting regime. Furthermore, note that the results of Section 3.2 are given for the case when all transition rates between vertices are distinct (the limits are taken only on the final transition rates and not on all rates simultaneously). However below, and in several of the approximate formulas of Section 3.2, we will take the intermediate rates equal to the final transition rate. The results of this section could be presented with all transition rates distinct, and approximations given for when the final transition rates are small but non-zero. Then the numerical value for intermediate transition rates could be chosen to be equal to the final transition rates, where for the sake of interpretation they should be the same. We have chosen to avoid such a presentation.

3.3.1 Imperfect drug penetration: monotherapy

Our first two applications concern imperfect drug penetration and so the language introduced here shall hold in the next section also.

We consider a pathogenic population, with cancerous cells or bacteria as example

populations, being treated with one or more drugs with imperfect penetration profiles. The imperfect penetration results in low drug concentration in spatial locations, such that cells in these locations may still proliferate. We term the cumulation of these spatial regions the *sanctuary*. Sanctuaries have been observed in both bacterial infections [21, 88] and tumours [71]. We consider resistance to have arisen when cells with sufficiently many mutations such that all therapies are impotent come to exist in areas where the drugs have penetrated. The possibility of resistance occurring prior to treatment is excluded and we consider only the case when resistance arises from the sanctuary. Further, we do not consider cells acquiring mutations via gene transfer.

Throughout this section and the next we shall be comparing paths comprised of mutation and migration events. Growing cells in the sanctuary will be the type 1 cells from Section 3.1.2 and thus have birth and death rates α, β, with growth rate λ. The lack of exposure to drug motivates $\lambda > 0$. We fix the initial number of cells in the sanctuary to be z. We shall assume all cell types have birth rate α. Across all cell types, i.e. regardless of location or drug presence, the per cell mutation and migration rates will be denoted ν and m respectively. We suppose each mutation carries a fitness cost which will be denoted s. Resistance costs may not hold always but are well documented and often assumed in viruses, bacteria and cancer [2, 29, 44]. Migration events bring cells into contact with drugs and each drug invokes an additive fitness cost d, unless resistance to the drug has already been acquired. Drug resistant cells outside the sanctuary are the target type cells of Section 3.1.2 and we define resistance to have arisen when cells of this type exist. Thus, as there are 4 vertices in this system (i.e. $N = 4$ see Figure 3.8a) here escaping resistance is defined as $\{Z_4(t) = 0 \text{ for all } t \geq 0\}$.

We introduce the resistance paths and demonstrate the precise effect of the drug/resistance-mutation costs in tandem. There are two paths from which resistance might arise, see Figure 3.8a. Firstly, upon replication a cell in the sanctuary may produce a mutant able to grow in the presence of the drug. Then a mutated cell in the sanctuary can migrate to the drug compartment. We term this the mutation-migration path. Analogously we have the migration-mutation path which involves a susceptible cell migrating from the sanctuary, from whose progeny a mutant emerges. Due to the fitness cost of resistance, the death rate of mutated cells in the sanctuary is increased to $\beta + s$, while the death rate of migrated susceptible cells in the drug compartment is $\beta + d$, with $s, d > 0$.

The first quantity to determine is the probability of escaping resistance (drug

resistant cells outside the sanctuary never exist), and its dependence on the initial number of cells in the sanctuary. The approximation of (3.17) shows that the probability of resistance never occurring decreases exponentially with the initial number of cells in the sanctuary as $\mathbb{P}(\text{resistance never occurs}) \approx (\beta/\alpha)^z$. For the rest of the discussion we will assume resistance does occur. Further we suppose $z \gg 1$, so that we may use the approximate form for $h(z)$ given in Corollary 3.3. The case when z is small could be treated similarly using the exact expression for $h(z)$.

Let $t_{1/2}$ be the median time until resistance occurs. Then from Corollary 3.3 we immediately have

$$t_{1/2} \approx \lambda^{-1} \log(\frac{\lambda^2}{\alpha}) - \lambda^{-1} \log\left(\frac{\nu m}{s} + \frac{\nu m}{d}\right) + \lambda^{-1} \log\left(\frac{\alpha}{z\lambda}\right). \tag{3.28}$$

As expected increasing the resistance cost or drug efficacy increases the resistance time, while increasing the initial sanctuary speeds up resistance. This simple result can now be explored under different biological scenarios.

For example, suppose we are currently using drug A with fitness cost $d^{(A)}$ and penetration profile leaving $z^{(A)}$ initial cells in the sanctuary. We consider instead using a second drug which results in sanctuary size $z^{(B)}$ and has fitness cost $d^{(B)}$. Assume that

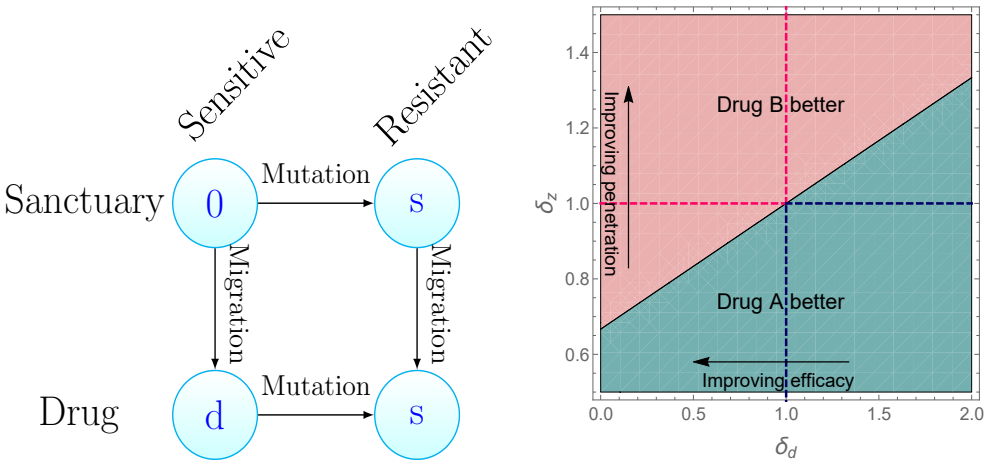
$$\begin{aligned} z^{(A)} &= \delta_z z^{(B)} \\ d^{(A)} &= \delta_d d^{(B)}, \quad \delta_z, \delta_d > 0. \end{aligned}$$

The median resistance times under the first and second drug shall be denoted $t_{1/2}^{(A)}$ and $t_{1/2}^{(B)}$ respectively. Using (3.28) we see $t_{1/2}^{(B)} > t_{1/2}^{(A)}$, and hence we should switch to drug B, if

$$\delta_z > \frac{1 + \delta_d s/d^{(A)}}{1 + s/d^{(A)}}. \tag{3.29}$$

This inequality is illustrated in Figure 3.8b.

Note as $\delta_d \to 0$, implying drug B is far more effective, the condition tends to $\delta_z > (1 + s/d^{(A)})^{-1}$. This is due to the fact that when $d^{(B)}$ is large resistance will always arise via the mutation-migration path (which we show below), so this provides an upper bound to the gain that can be obtained from increasing the drug efficacy. The above is under the assumption resistance will occur but note the probability of no resistance would decrease exponentially with larger



(a) *In the case of monotherapy, there are two paths by which drug resistance cells can arise in the region containing the drug. Horizontal edges represent changes in genotype (gaining resistance) whereas vertical edges represent changes in spatial location. The values of vertices are the fitness cost cells experience in that state.*

(b) *Comparison of two drugs with differing efficacies and penetration profiles using* (3.29)*. Arrows pertain to characteristics of drug B relative to drug A. Top left and bottom right quadrants, inside dashed lines, illustrate region where drugs are trivially better, i.e. have superior penetration and efficacy. Parameters:* $d^{(A)} = 10^{-2}$, $s = 5 \times 10^{-3}$.

Figure 3.8

sanctuary size.

Having discussed the time for resistance to occur we now ask by which of the two paths it will occur. Let $T^{\nu m}$ be the first time resistance arises via the mutation-migration path and $T^{m\nu}$ be defined analogously for the migration-mutation path. Then Theorem 3.3 gives us

$$\mathbb{P}(T^{\nu m} < T^{m\nu}) \approx (1 + s/d)^{-1}.$$

It is reassuring that as $s \to 0$, and therefore no cost is associated with the mutation-migration path, this path is taken with probability 1. Generally resistance arising via the mutation-migration path is more probable, i.e. the above probability is greater than $1/2$, if $d > s$. We naively expect this condition to hold for most drugs.

A similar question was asked in [33]. There the authors considered an approximate stochastic process and sought the parameter regime such that $T^{m\nu}$ stochastically dominated $T^{\nu m}$ under the approximate process. For the full process considered here, we can ask the same question but on the limit of the centred, in the sense of Theorem 3.2, resistance times. The resulting condition for stochastic

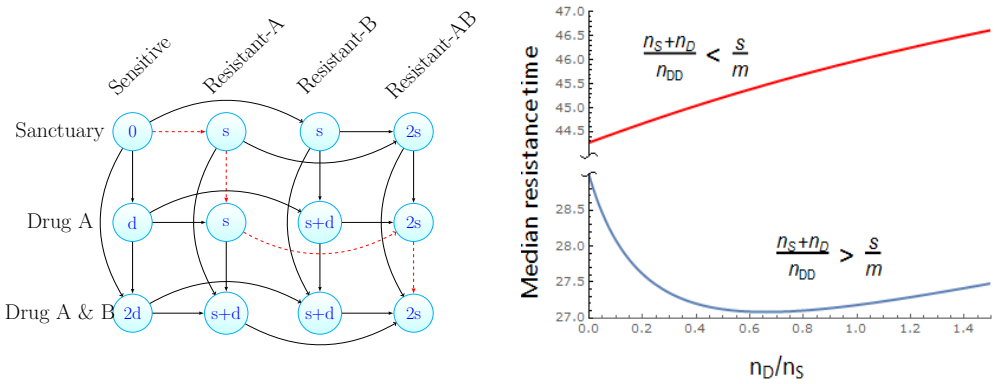
dominance is again $d > s$. The discrepancy in the condition given here and that stated in equation (28) in [33] exists as there ν is defined as the probability of mutation per birth event. Accounting for this leads to the same condition.

3.3.2 Imperfect penetration: combination therapy

There are at least two reasons for extending our analysis to include two drugs, say drugs A and B, being used in tandem. Firstly, combination therapies are widely implemented in both cancer and bacterial infections. Secondly, interesting new phenomena emerges, especially when a region still exists with only one drug present. This represents the scenario where one of the drug's penetration profile is a strict subset of the other.

We will only consider the case where the penetration profile for drug B is a subset of drug A, and thus a single drug region exists only for drug A. This allows us to clearly demonstrate the impact of unequal penetration profiles. In this setting there are 12 possible states that a cell may inhabit, defined by which of the drugs the cell is exposed to and the resistance it has obtained. These states are illustrated in Figure 3.9a. A more general case can be treated similarly. We further assume that the mutation rate bringing resistance to either drug is ν. This is for simplicity only, and having differing mutation rates is straightforward. Drug interactions and epistasis are also neglected but can be included.

To investigate the effect of the single drug region, we introduce further notation. Let n_{Tot} be the number of cells that may reside in our system of interest (e.g. the entire patient or tumour) in the absence of drug. Further, let the regions where both drugs act, the single drug (drug A) acts and the sanctuary have their respective cell capacities denoted by n_{DD}, n_{D}, and n_{S}. We assume that all these capacities are large, and hence we may still investigate the system with our branching process model. These capacities enter the dynamics of the process in the following manner. When a cell leaves the region it is in, at rate m, it now may migrate to one of two regions. Of these two regions, we specify that it migrates to a particular region at a frequency proportional to the region's capacity. For example a cell in the sanctuary migrates to the double drug region at rate $mn_{\text{DD}}/(n_{\text{D}} + n_{\text{DD}})$. The form of this migration rate can be motivated by assuming cells migrate via the vascular system, and that the interface size between particular regions and blood vessels is proportional to the capacity of those regions. It is also the form assumed in [65] which allows comparison with



(a) *Illustration of the paths to multidrug resistant cells arising in the region containing both drugs A and B, as in Figure 3.8a. Note a spatial location exists containing only drug A. The path benefits and costs for the red-dashed path are given in* (3.30).

(b) *Exploration of the median time to resistance as the relative size of the region containing only drug A to the sanctuary increases. A single drug region can speed up resistance when* (3.31) *holds. When true, resistance is fastest at* $n_D/n_S \approx 0.6$*, agreeing with* (3.32)*. Parameters; (top)* $\nu = 10^{-6}, m = 0.05, d = 0.9, s = 10^{-3}, \lambda = 0.4, \alpha = 0.3, n_{Tot} = 10^7, n_{DD} = 0.9n_{Tot}, \gamma = 10^{-3}$*; (bottom) all same except* $s = 10^{-2}$*.*

Figure 3.9

their results. If, instead, one wished to replace the regions' capacities with the volume or surface area, the results below would not differ, only the interpretation of the initial sanctuary size. Resistance mutations and the presence of the drug has the same additive effect as in Section 3.3.1.

Observe that, when all compartments have positive size, there are 18 possible paths. To see this, it is convenient to separate the paths into those that pass through the single drug compartment, which following [65] we term stepwise evolution, and those that involve migration from the sanctuary directly to the double drug compartment, which we collectively call direct evolution. For the direct evolution paths there are three transitions needed, mutations to drug A and B and then the migration to the double drug region. From the number of permutations of these events, this yields 6 direct evolution paths. In the stepwise evolution case four transitions are needed, both mutations and then the two requisite migration events. Notice that the migration from the sanctuary to the single drug region must precede the migration from the single drug region to the double drug region. This leads to 12 stepwise evolution paths.

An exemplar stepwise evolution path is a mutation causing resistance to drug A, followed by migration to the single drug region, then a mutation causing resistance

to drug B, and finally a migration to the double drug region. The edges of this path are red and dashed in Figure 3.9a. As before, the path benefits and costs are important. If we denote our exemplar stepwise evolution path, path 1, then its benefit and cost are

$$b(p^{(1)}) = \nu^2 m^2 \frac{n_{\mathrm{D}}}{n_{\mathrm{D}} + n_{\mathrm{DD}}} \frac{n_{\mathrm{DD}}}{n_{\mathrm{S}} + n_{\mathrm{DD}}}, \quad c(p^{(1)}) = 2s^3. \tag{3.30}$$

Note that d does not feature in the cost as resistance is always acquired prior to migration. The ordering of mutations matters for stepwise evolution paths, as acquiring resistance to drug B prior to migration does not negate the cost of exposure to drug A. We also observe that the path benefit given above will be the same for each stepwise evolution path.

Next we consider the biologically motivated scenario in which the size of the single drug region comes at the expense of the sanctuary. A trade-off may exist between reducing the initial number of proliferating sanctuary cells and increasing the number of paths to resistance. To study this we let $r = n_{\mathrm{D}}/n_{\mathrm{S}}$, fix $n_{\mathrm{Tot}} = n_{\mathrm{S}} + n_{\mathrm{D}} + n_{\mathrm{DD}}$, and specify that the initial number of cells in the sanctuary is related to its size via $z = \gamma n_{\mathrm{S}}$ with $0 < \gamma \ll 1$. Further let $t_{1/2}(r)$ be the median time to resistance with fixed r. Using the approximation provided by Corollary 3.3 we can explore the behaviour of $t_{1/2}(r)$. As can be seen in Figure 3.9b it is possible that the existence of the single drug compartment speeds up resistance. We now seek the conditions under which this is the case.

It is reasonable to expect that the efficacy of either drug is greater than the cost to resistance. Also we may hope that both drugs penetrate the majority of the target system. This motivates taking $s \ll d$ and $(n_{\mathrm{S}} + n_{\mathrm{D}}) \ll n_{\mathrm{Tot}}$. Under this limit, by examining the sign of $\frac{d}{dr} t_{1/2}(0)$, we find a simple condition for when increasing the single drug compartment speeds up resistance, namely

$$\frac{n_{\mathrm{S}} + n_{\mathrm{D}}}{n_{\mathrm{Tot}}} > \frac{s}{m}. \tag{3.31}$$

However, when (3.31) holds, due to depletions in the sanctuary size, as $n_{\mathrm{D}}/n_{\mathrm{S}}$ increases eventually a minimal resistance time is achieved. This represents a worse case scenario yielding fastest possible resistance. For $s \ll d$, this minimal time occurs when

$$\frac{n_{\mathrm{D}}}{n_{\mathrm{S}}} \approx 1 - \frac{2sn_{\mathrm{Tot}}}{m(n_{\mathrm{S}} + n_{\mathrm{D}})}. \tag{3.32}$$

We now examine whether resistance arises via stepwise evolution or direct

evolution. Let T_{SE} be the first time resistance occurs via stepwise evolution, and T_{DE} defined analogously for direct evolution. Then using Theorem 3.3, again with $s \ll d$, we find

$$\mathbb{P}(T_{\mathrm{SE}} < T_{\mathrm{DE}}) \approx \left[1 + \frac{s(n_{\mathrm{DD}} + n_{\mathrm{S}})}{m n_{\mathrm{D}}}\right]^{-1}. \tag{3.33}$$

Therefore the larger the proportional size of the single drug region, the more likely resistance will arise via stepwise evolution. A qualitatively similar result was derived in [65] (e-page 2878) by comparing the most likely stepwise evolution path with the most likely directed evolution path.

Furthermore, in [65] a minimal resistance time, in the sense of (3.32), was observed via simulations for the mutation$_A$-migration-mutation$_B$-migration path (path 1 from (3.30)). Using Theorem 3.3 we can show that this is the most probable (in the sense of the mode of the path distribution, see (3.24)) stepwise evolution path whenever $d > s$. Further, by differentiating the median time along this path, we see a minimal time occurs at $n_{\mathrm{D}} \approx n_{\mathrm{S}}$, which recovers the simulation result in [65] (Figure 3). That this path is the most probable stepwise evolution path also gives insight into the recurring s/m term in (3.31) and (3.33). Relative to direct evolution this path has an extra migration step, so higher migration should be beneficial. However fitness costs along this paths are incurred only via the cost of resistance, and a smaller s diminishes these fitness costs.

3.3.3 Path to resistance in chronic myeloid leukemia

We now consider the example given in [57], concerning the mechanism by which resistance to the targeted therapy imatinib occurs in chronic myeloid leukemia (CML). The two main resistance mechanisms are via gene amplification of the BCR-ABL fusion gene, which drives the CML, or a point mutation resulting in a modification to the target protein of imatinib. Approximately 100 point mutations have been identified that confer resistance, and a rough estimate for the probability of each of these occurring during a cell division is 10^{-7} [61]. Thus the probability of a point mutation conferring resistance is approximately 10^{-5} per division. The analogous quantity for resistance-causing gene amplifications, in a similar system, was found to be approximately 10^{-4} [83]. With the birth rate of leukemic stem cell as $\alpha = 0.008$ day^{-1} [63] we have the rate of point mutations as $\nu_{pm} = 8 \times 10^{-8}$ days^{-1}, while the rate of resistance due to gene amplifications

is $\nu_{ga} = 8 \times 10^{-7}$ days^{-1}. Therefore, using this information alone, the probability of resistance arising from a point mutation first is estimated to be 1/11. However, in the majority of cases, the primary mechanism of resistance is found to be point mutations [11].

Within our framework there are two possible explanations we can consider, both related to the reproductive capabilities of resistant cells. The first is simply that more lineages of amplified cells go extinct. The comments following (3.1) indicate how to treat this case. Indeed, if ρ_{pm} is the average survival probability for a population initiated by a single cell with a point mutation and ρ_{ga} the analogous quantity for gene amplification, then we find resistance arising first from a point mutation is more probable if $\rho_{pm} > 10\rho_{ga}$. This is possible as point mutations may be deleterious or advantageous [61] whereas gene amplifications appear deleterious [82].

The second explanation, that suggested in [57], is that several gene amplifications are required to attain resistance. Suppose a gene amplification increases the death rate of cells by s per day. Then if two gene amplifications are required, ignoring the survival probability from before, the probability that resistance arises via point mutation is now

$$\frac{\nu_{pm}}{\nu_{pm} + \nu_{ga}^2/s} = (1 + 8 \times 10^{-6}/s)^{-1}.$$

Resistance via point mutations is now more likely if $s > 10^{-6}$. Further explanations for the primacy of point mutation mediated resistance exist and the assumptions above may not hold. But when they do hold, our results offer a clear framework to investigate such questions.

3.3.4 Antibacterial multidrug resistance

We now move to consider multidrug resistant bacteria. As before we assume resistance incurs a fitness cost and thus the (unmutated) wild-type has the highest growth rate. Using Theorem 3.3 and measured fitness costs and mutation rates we can predict the most likely path to multidrug resistance.

We consider the acquisition of resistance to both rifampicin and streptomycin in *P. aeruginosa*. Recently the fitness costs associated with resistance to rifampicin [39] and streptomycin [87] have been reported. Fitness costs were determined

by the maximum growth rate of the bacteria undergoing exponential growth on media-rich plates. From [87] the ratio of streptomycin resistant bacteria to the wild-type's growth rate was 0.71 (here the wild-type's growth rate is estimated to be 5.1 mOD min^{-1} from Figure 2.a in [87]). The same value for resistance to rifampicin, in [39], was 0.88. Let T^{RS} be the first time of multidrug resistance when rifampicin resistance is acquired first, and T^{SR} defined analogously for streptomycin resistance first. Then from Theorem 3.3, we have that

$$\mathbb{P}(T^{RS} < T^{SR}) \approx \frac{1 - 0.71}{(1 - 0.88) + (1 - 0.71)} = 0.7.$$

We note that above we have taken the average fitness costs associated with the differing antibiotics. Fluctuations have not been taken into account.

Instead of considering the differing mutation rates to different drugs within a bacterial strain, we can also examine the impact of differing mutation rates across strains. In [32] the heightened occurrence of multidrug resistance in lineage 2 vs lineage 4 from *M. tuberculosis* was considered. Considering in particular resistance to rifampicin and isoniazid, via simulations the authors deduced that the heightened occurrence of resistance was due to elevated mutation rates in lineage 2 strains. The model simulated can be considered a discrete time counterpart to ours in the case of two paths of length 2. One of the main quantities of interest was the relative probability of resistance, when the only difference between the strains was greater mutation rates. For a particular strain, initiated with one wild type cell with wild-type birth and growth rates, α, λ, suppose we have mutation rate $\nu_{\text{R}}, \nu_{\text{I}}$ to rifampicin and isoniazid, each incurring the same fitness cost s. Then from Theorem 3.2, we have

$$\mathbb{P}(T \leq t) \approx 1 - \left(\frac{\lambda/\alpha}{1 + e^{\lambda(t-\mu)}} + \beta/\alpha \right), \quad \mu = \lambda^{-1} \log(\frac{\lambda^2}{\alpha}) + \lambda^{-1} \log \left(\frac{s}{2\nu_{\text{R}}\nu_{\text{I}}} \right). \tag{3.34}$$

For lineage-i ($i = 2, 4$) let $\nu_{\text{R}}^{(i)}$ and $\nu_{\text{I}}^{(i)}$ be the mutation rates to rifampicin and isoniazid. As in [32], keeping all other parameters equal between the lineages we have the relative probability of resistance

$$\frac{\mathbb{P}(T^{(2)} \leq t)}{\mathbb{P}(T^{(4)} \leq t)} \approx \frac{\nu_{\text{R}}^{(2)}\nu_{\text{I}}^{(2)}}{\nu_{\text{R}}^{(4)}\nu_{\text{I}}^{(4)}} \frac{(1 + 2\nu_{\text{R}}^{(4)}\nu_{\text{I}}^{(4)}e^{\lambda t}\alpha/\lambda^2 s)}{(1 + 2\nu_{\text{R}}^{(2)}\nu_{\text{I}}^{(2)}e^{\lambda t}\alpha/\lambda^2 s)}.$$

For times such that $\nu_{\text{R}}^{(i)}\nu_{\text{I}}^{(i)}e^{\lambda t} \ll 1$, $i = 2, 4$, we see the relative probability is simply $\nu_{\text{R}}^{(2)}\nu_{\text{I}}^{(2)}/\nu_{\text{R}}^{(4)}\nu_{\text{I}}^{(4)}$. Taking the mutation rates measured in [32], this gives

the relative probability of 22.06, which agrees with the simulation results in [32] (an approximately 22 fold increase is reported on page 788, and indicated on Figure 2c in the supplementary information). If in (3.34), we replace the λ^2/α with λ and $e^{\lambda t}$ with M, our expression is the same, to leading order in the mutation rates, as that given in [18] (equation 2) for the probability of resistance in a population of size M. This difference between fixed population size and fixed time results is the same as given in [16, 27].

3.4 Discussion

3.4.1 Shortcomings and future work

There are several shortcomings with the work presented in this chapter which we now discuss. We first consider those shortcomings arising from our reliance on the almost sure convergence of the growth of intermediate populations detailed in Theorem 3.1.

Firstly, our approach of always taking the final transition rates in the path to zero is rather unwieldy. As noted at the beginning of Section 3.3, we cannot rigorously cover the setting when the intermediate transition rates are equal to the final transition rates. Secondly, for a similar reason we cannot treat transitions that do not occur at replications, i.e. leave behind a cell of the original type. If this were to be the case then altering the final transition rates would affect the growth of the penultimate population type. This criticism is relevant in the scenario of migration treated in Sections 3.3.1, 3.3.2, as it is unlikely a migrating cell would do so upon replication. However if the migration rate is small, the dynamics of the penultimate population is unlikely to differ greatly. Hence we expect the above conclusions to broadly hold. Finally, while the assumption that the type 1 population is the most fit is often biologically motivated, it cannot easily be relaxed as then λ would cease to be the dominant eigenvalue of the matrix A from Section 3.2.1. The case where each transition leads to an increase in fitness was treated in [27], and a merging of the results given there with those presented here should be possible. This is left for future work.

In this chapter we considered the first time the target population arose and possibly the most straightforward extension would be to characterise the sequence of initiation times, as opposed to only the first, that is the Cox process of

immigrations to the target. Under the same centring as Theorem 3.2, we would expect the entire process to converge. For the case of $N = 2$, i.e. a two-type process, this was carried out in [16] and the same method should apply here.

As in Chapter 2 a possible extension is to consider the case where the fitnesses of types are themselves random variables. This is due to the observed fluctuations in fitness in both cancer [61] and bacteria [39]. So long as the type 1 is the most fit with probability one, the formulas presented in this chapter would then be the conditional results of the more general model.

In terms of biologically applicability, as with all branching process models with constant parameters, the exponential growth of the population is ultimately unrealistic. It would be interesting to know to what extent the features discussed above hold for populations with a large carrying capacity. The time scale to cross a fitness valley is given by the variable μ but this differs from the time scale observed in [36]. However there a carrying capacity of 100 was used. Carrying capacities were also used in [65] of the orders $10^5 - 10^7$ and we have been able to derive analytic expressions matching their simulation results, see Section 3.3.2. We would speculate that the carrying capacity should be large enough such that the target vertex is populated before the type 1 population saturates. This also implies the transition rates must be large enough also. A full investigation of this is also left for future work.

3.4.2 Concluding remarks

In this chapter we have considered a continuous time birth-death process with transitions. The time until the population of a particular target type arises was studied. This population type is accessible via one or more transitions from the initial population type. Which sequence of transitions leads to the target population arising was explored. This gave rise to a probability distribution over the paths composed of transitions between differing types. Motivated by applications, the setting when the initial population was most fit was focused on.

The important factor in determining whether the target was reached via a specific path was seen to be the path's benefit and cost. These terms are comprised of the transition rates and fitness costs associated with the intermediate types along the considered path. Both the time and path distribution have simple, explicit formulas. The utility of these formulas was demonstrated on a variety of

applications. The biological relations revealed might have been difficult to deduce by a simulation based approach.

Justified criticisms have been made on the use of birth-death processes as a modelling framework. In particular the potential for unlimited growth and lack of interactions between cells is unrealistic. However in systems with large carrying capacities and abundant resources, they provide a convenient approximation and are widely used. If this framework is deemed appropriate, this chapter provides quick and accessible results which may be used to guide and develop biological insight.

Chapter 4

Conclusions

At the outset to to this thesis, we outlined the importance of mathematical models concerning populations seeding further populations. These models have a diverse range of applications [1, 56, 62, 66] and are generally simple to state and intuitively motivate. However despite their apparent simplicity and impact, many aspects of these fundamental models remain unexplored. While several applications have been considered in the previous chapters, a fuller understanding of the models themselves was the driving motivation for this work.

Inspired by the success of Lea and Coulson [60], Chapter 2 explored a generalisation of their model, where stochastically growing mutant clones are initiated by a deterministically growing wild type population. It should be kept in mind that if stochastic wild type growth is desired, then Chapter 2 concerns clones conditioned on wild type growth. How differing wild type growth manifests itself in the clone size distribution was investigated. In some sense the intuition gleaned by considering deterministic mutant growth (Section 2.1.4) provided the take home message. Clones seeded from faster growing populations are more likely to arrive later in the process and so, on average, will be smaller. However the dichotomy that was demonstrated regarding clone sizes seeded from subexponential versus exponential-type wild type growth would have been difficult to predict a priori.

Our application of choice was the scenario of a primary tumour seeding metastases. From our theoretical results it was predicted that regardless of the primary tumour growth, a power law tail was expected in the metastasis size distribution. As a first step to explore this, likelihood analysis was applied to a dataset of metastasis sizes. The power law hypothesis was found to be plausible.

Moreover the estimated exponents of how the power law decayed in the empirical metastasis distributions gave a point estimate of the relative growth rate of the metastases to the primary tumour. This point estimate suggested the metastases grew faster than the primary, which is consistent with previous findings.

We next turned to the setting of a target population arising which might require multiple transitions from the original population. The questions of how long we wait until a target cell type exists and which sequence of intermediate types initiated the target population were considered. Our primary contribution here was to provide relatively simple, malleable formulas in terms of the path benefit and costs as answers to these questions. The formulas are intuitively appealing and amenable to further analysis. The applicability of these formulas was demonstrated in a variety of scenarios including exploring the impact of imperfect drug penetration, and how the ordering of mutations depends on fitness costs and mutation rates.

It may be argued the results of this thesis are special cases of more general results found in the branching process literature. However our aim here is to explicitly focus on those special cases which have biological significance, examining the biologically pertinent aspects and then reinterpreting our findings in light of the application. The transition of the statement of Theorem 3.2 in Chapter 3 to its application in Section 3.3 is a prime example. The original statement contains the entries of a given eigenvector of a given matrix. While for any set of parameter values this eigenvector may be obtained numerically, it is by considering their representation in terms of path benefits and costs that leads to a greater understanding of the process, and in turn the biology.

This last example illustrates what we hope this thesis contributes. If biological scientists wish to model a particular system, and if their model of choice is related to the models considered above, then the work presented here might be used to enhance their toolkit and give a firmer grounding to their intuition.

Appendix A

Appendix to Chapter 2

A.1 Special functions and definitions

Required definitions and identities taken from [26] unless otherwise stated.

With $s, z \in \mathbb{C}$ the polylogarithm of order s is defined as

$$\mathrm{Li}_s(z) = \sum_{k \geq 1} \frac{z^k}{k^s}.$$

Note that $\mathrm{Li}_1(z) = -\log(1-z)$. Gauss's hypergeometric function also appears and for complex a, b, c, z is defined by the power series

$$F\left(\begin{matrix} a, b \\ c \end{matrix}; z\right) = \sum_{k \geq 0} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} \quad \text{for} \quad |z| < 1,$$

and by analytic continuation elsewhere. Here $(a)_k$ denotes the Pochhammer symbol or rising factorial, that is

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} = a(a+1)(a+2)\cdots(a+k-1).$$

Some required identities for the hypergeometric function are:

$$F\left(\begin{matrix} a, b \\ c \end{matrix}; z\right) = (1-z)^{-b} F\left(\begin{matrix} c-a, b \\ c \end{matrix}; \frac{z}{z-1}\right), \tag{A.1}$$

$$F\left(\begin{matrix}1,b\\c\end{matrix};z\right)=1+\frac{b}{c}z\,F\left(\begin{matrix}1,b+1\\c+1\end{matrix};z\right), \tag{A.2}$$

$$F\left(\begin{matrix}1,1\\2\end{matrix};z\right)=-\frac{\log(1-z)}{z}, \tag{A.3}$$

and the following connection can be made to the incomplete beta-function

$$\frac{z^a}{a}F\left(\begin{matrix}a,1-b\\a+1\end{matrix};z\right)=B_z(a,b)=\int_0^x t^{a-1}(1-t)^{b-1}\,dt. \tag{A.4}$$

A.2 Exact time dependent clone size distributions

Here we derive the generating functions for the clone sizes given in Table 2.1 and the notation here is as given in Chapter 2. Where possible we also give expressions for the probability mass functions.

A.2.1 Exponential wild-type growth

We start with the case of exponential growth, $n_\tau = e^{\delta\tau}$ with $\delta > 0$. The distribution for the total number of mutants, B_t, was treated exhaustively in [51] and we follow their notation by letting $\gamma = \delta/\lambda$. Using (2.5) and the results found in section 3 of [51], the generating function is

$$\mathcal{Y}_t(s)=1+\frac{\lambda}{1-n_t^{-1}}\left[n_t^{-1}\,F\left(\begin{matrix}1,\gamma\\1+\gamma\end{matrix};\xi n_t^{-1/\gamma}\right)-F\left(\begin{matrix}1,\gamma\\1+\gamma\end{matrix};\xi\right)\right]. \tag{A.5}$$

Similarly the mass function is

$$\mathbb{P}(Y_t=0)=1+\frac{\lambda}{1-n_t^{-1}}\left[n_t^{-1}\,F\left(\begin{matrix}1,\gamma\\1+\gamma\end{matrix};\beta n_t^{-1/\gamma}\right)-F\left(\begin{matrix}1,\gamma\\1+\gamma\end{matrix};\beta\right)\right]$$

and for $k \geq 1$

$$\begin{aligned}\mathbb{P}(Y_t=k)=&\frac{\delta}{\alpha(n_t-1)}\sum_{j=1}^{k}\binom{k-1}{j-1}\frac{1}{j+\gamma}\left(\frac{\lambda}{\beta-n_t^{1/\gamma}}\right)^j F\left(\begin{matrix}1,\gamma\\1+\gamma+j\end{matrix};\beta n_t^{-1/\gamma}\right)\\&+\frac{\delta}{\alpha(1-n_t^{-1})}\frac{(k-1)!}{(\gamma+1)_k}F\left(\begin{matrix}k,\gamma\\1+\gamma+k\end{matrix};\beta\right).\end{aligned}$$

again $F\left(\begin{matrix}a,b\\c\end{matrix};z\right)$ is Gauss's hypergeometric function and $(a)_k$ is the Pochhammer symbol defined in Appendix A.1. The above expressions are given in terms of n_t to allow easy comparison to the formulas in [51]. For these exact time-dependent formulas, their form is somewhat cumbersome, however simpler long-time limit expressions are given in Section 2.3.1. A reduction in complexity is also obtained for the special case of neutral mutants ($\delta = \lambda$) where, by using (A.3), the generating function in (A.5) simplifies to

$$\mathcal{Y}_t(s) = 1 + \frac{\lambda}{\xi(1-e^{-\delta t})} \log \frac{1-\xi}{1-\xi e^{-\delta t}}.$$

If additionally the neutral mutants are immortal, the above expression further simplifies to

$$\mathcal{Y}_t(s) = 1 + \frac{1-s}{s\phi} \log(1-s\phi) \text{ where } \phi = 1 - e^{-\delta t}.$$

The probabilities are then concisely given by

$$\mathbb{P}(Y_t = k) = \frac{\phi^{k-1}}{k} - \frac{\phi^k}{k+1} \text{ or } \mathbb{P}(Y_t > k) = \frac{\phi^k}{k+1}$$

which corresponds to the classical Lea-Coulson result [60]

$$\mathcal{B}_t(s) = (1-s\phi)^{\theta(1-s)/s}$$

with $\theta = \nu e^{\delta t}$.

A.2.2 Power-law wild-type growth

Now we assume that the wild-type population grows according to a general power-law $n_\tau = \tau^\rho$, for some non-negative integer ρ, and therefore $a_t = \frac{t^{\rho+1}}{\rho+1}$. To obtain the generating function we start from (2.4), which gives

$$\mathcal{Y}_t(s) = \frac{\rho+1}{t^{\rho+1}} \int_0^t \tau^\rho \left(1 - \frac{\lambda}{1-\xi e^{-\lambda(t-\tau)}}\right) d\tau = 1 - \frac{(\rho+1)\lambda}{t^{\rho+1}} \int_0^t \frac{\tau^\rho}{1-\xi e^{-\lambda(t-\tau)}}\, d\tau.$$

We claim that

$$\int \frac{\tau^\rho}{1-\xi e^{-\lambda(t-\tau)}}\,d\tau = \frac{\tau^{\rho+1}}{\rho+1} + \rho! \sum_{i=0}^{\rho} \frac{(-1)^i}{(\rho-i)!\lambda^{i+1}} \tau^{\rho-i} \mathrm{Li}_{i+1}(\xi e^{-\lambda(t-\tau)}) \quad \text{(A.6)}$$

where $\mathrm{Li}_i(s)$ is the polylogarithm of order i defined in Appendix A.1. This may be derived by a binomial expansion of the denominator and an identity for the incomplete gamma function, but for succinctness we differentiate both sides with respect to τ. First we note that

$$z\partial_z \mathrm{Li}_i(z) = \mathrm{Li}_{i-1}(z).$$

Now differentiating the right hand side of (A.6) yields

$$\tau^\rho + \frac{\tau^\rho \lambda \mathrm{Li}_0(\xi e^{-\lambda(t-\tau)})}{\lambda} + \rho! \sum_{j=0}^{\rho-1} \frac{(-1)^j(\rho-j)\tau^{\rho-j-1}\mathrm{Li}_{j+1}(\xi e^{-\lambda(t-\tau)})}{(\rho-j)!\lambda^{j+1}} +$$
$$\rho! \sum_{i=1}^{\rho} \frac{(-1)^i \tau^{\rho-i}\lambda \mathrm{Li}_i(\xi e^{-\lambda(t-\tau)})}{(\rho-i)!\lambda^{i+1}} = \tau^\rho(1+\mathrm{Li}_0(\xi e^{-\lambda(t-\tau)}))$$

where the equality follows by the telescoping nature of the sums. Noting that $(1-\xi e^{-\lambda(t-\tau)})^{-1} = \mathrm{Li}_0\left(\xi e^{-\lambda(t-\tau)}\right) + 1$ and applying the limits of the integral reveals the generating function to be

$$\mathcal{Y}_t(s) = \beta + \lambda(\rho+1)!\left[\frac{(-1)^\rho \mathrm{Li}_{\rho+1}(\xi e^{-\lambda t})}{(t\lambda)^{\rho+1}} + \sum_{i=0}^{\rho} \frac{(-1)^{i+1}\mathrm{Li}_{i+1}(\xi)}{(\rho-i)!(t\lambda)^{i+1}}\right]. \quad \text{(A.7)}$$

To determine the mass function, we seek a power series representation of the generating function. We focus on the $\beta = 0$ case and thus $\xi = \frac{s}{s-1}$. By the definition of the polylogarithm and the binomial theorem

$$\mathrm{Li}_i\left(\frac{s}{s-1}\right) = \sum_{k\geq 1}\sum_{j\geq 0} \binom{k+j-1}{j}(-1)^k \frac{s^{j+k}}{k^i}$$

Reindexing the sum we obtain

$$\mathrm{Li}_i\left(\frac{s}{s-1}\right) = \sum_{m\geq 1} s^m \sum_{k=1}^{m} \binom{m-1}{m-k}\frac{(-1)^k}{k^i}$$
$$\mathrm{Li}_i\left(\frac{se^{-\alpha t}}{s-1}\right) = \sum_{m\geq 1} s^m \sum_{k=1}^{m} \binom{m-1}{m-k}\frac{(-e^{-\alpha t})^k}{k^i}.$$

Applying this to the polylogarithmic terms in $\mathcal{Y}_t(s)$, and noting

$$\sum_{k=1}^{m} \binom{m-1}{m-k} \frac{(-1)^k}{k^i} = \frac{1}{m} \sum_{k=1}^{m} \binom{m}{k} \frac{(-1)^k}{k^{i-1}} \quad \text{and} \quad \sum_{k=1}^{m} \binom{m}{k} (-1)^k = -1,$$

yields the mass function

$$\mathbb{P}(Y_t = m) = \frac{(\rho+1)}{mt\alpha} + \frac{(\rho+1)!}{mt\alpha} \left[\frac{(-1)^\rho}{(t\alpha)^\rho} \sum_{k=1}^{m} \binom{m}{k} \frac{(-e^{-\alpha t})^k}{k^\rho} + \sum_{i=1}^{\rho} \frac{(-1)^{i+1}}{(t\alpha)^i (\rho - i)!} \sum_{k=1}^{m} \binom{m}{k} \frac{(-1)^k}{k^i} \right].$$

If mutants may die, the exact mass function is most easily obtained via Cauchy's integral formula which may be efficiently computed using the fast Fourier transform. For a brief discussion on implementation see [4] and references therein.

Note for $\rho \geq 1$, $n_0 = 0$ which, while useful for analytic tractability, is unrealistic. This can be overcome by letting $n_\tau = n_0 + \tau^\rho$. Then, by splitting the integral in the generating function (2.4) and using the above analysis, one can obtain the mass function for any n_0. However for practical purposes the contribution of n_0 is negligible.

A.2.3 Constant size wild-type

For the specific power-law growth when $\rho = 0$, i.e. $n_\tau = 1$ (recall that this is equal to the general case when $n_\tau = n_0$), we recover some classical results for constant immigration [52]. We note that the distribution of the ordered clone size, depending on initiation time, was discussed in [48]. From (A.7) with $\rho = 0$, the generating function is

$$\mathcal{Y}_t(s) = 1 - \frac{1}{t\alpha} \log \left(\frac{1 - s\mathcal{S}_t^{-1}}{1 - \mathcal{S}_t^{-1}} \right). \tag{A.8}$$

with $\mathcal{S}_t$ as given in (1.6). By expanding this generating function in terms of s we obtain the probabilities

$$\mathbb{P}(Y_t = k) = \begin{cases} 1 + t^{-1} \log(1 - \mathcal{S}_t^{-1}) & k = 0 \\ \frac{1}{t\alpha k} \mathcal{S}_t^{-k} & k \geq 1. \end{cases}$$

Then using (2.5) with the clone sizes (A.8) we obtain the generating function of the total number of mutants

$$\mathcal{B}_t(s) = \left[\frac{1-\mathcal{S}_t^{-1}}{1-s\mathcal{S}_t^{-1}}\right]^{\nu},$$

and from the binomial theorem we also get the probabilities

$$\mathbb{P}(B_t = m) = \binom{m+\nu\alpha-1}{m}\left(1-\mathcal{S}_t^{-1}\right)^{\nu\alpha}\mathcal{S}_t^{-m}.$$

We recognise this as a negative binomial distribution under the interpretation that B_t is the number of failures until ν successes, with failure probability $\mathcal{S}_t^{-1}$. This result for B_t was first derived by Kendall [52] who was attempting to explain the appearance of the logarithmic distribution for species number when randomly sampling heterogeneous populations, conjectured by R.A. Fisher. From the distribution of B_t, by an argument which may be considered a special case of Proposition 2.1, he derived that for constant rate initiation, the clone size conditioned on non-extinction is logarithmically distributed again with parameter $\mathcal{S}_t^{-1}$, which can be obtained via (2.7).

Constant immigration may imply a constant size source, hence mutants with equal birth and death rates (i.e. evolving as a critical branching process) are particularly interesting. This case yields analogous formulas to those above but $\mathcal{S}_t$ is replaced with the expression given in (1.7).

A.2.4 Logistic wild-type growth

Starting from a population of one and having a carrying capacity K, logistic growth is given by $n_\tau = \frac{Ke^{\lambda\tau}}{K+e^{\lambda\tau}-1}$. We assume neutral mutations, i.e. λ is also the wild-type growth rate. Integrating the growth function gives $a_t = \frac{K}{\lambda}\log(\frac{e^{\lambda t}}{n_t})$.

We aim to calculate the generating function using (2.4). Recalling the definition of $\mathcal{Z}_{t-\tau}(s)$ we have the following antiderivative

$$\int \frac{1}{1-\xi e^{-\lambda(t-\tau)}} n_\tau \, d\tau = \frac{K}{\lambda[(K-1)\xi e^{-\lambda t}+1]} \log\left(\frac{1-e^{\lambda\tau}-K}{e^{\lambda t}-\xi e^{\lambda\tau}}\right).$$

Therefore the generating function is

$$\mathcal{Y}_t(s) = 1 + \frac{\lambda e^{\lambda t}}{[e^{\lambda t} + (K-1)\xi] \log(\frac{e^{\lambda t}}{n_t})} \log\left(\frac{n_t(1-\xi)}{e^{\lambda t}(1-\xi e^{-\lambda t})}\right).$$

Agreeing with intuition for $K = 1$ we recover the generating function of the constant case, and $\lim_{K\to\infty} \mathcal{Y}_t(s)$ gives the generating function for exponential wild-type growth. Therefore the logistic case interpolates between the constant and exponential growth cases. The mass function can be obtained by expanding the non-logarithmic and logarithmic function in $\mathcal{Y}_t(s)$ and using the Cauchy product formula. However, this method provides little insight and numerically it is simpler to use the fast Fourier transform.

Appendix B

Appendix to Chapter 3

B.1 Topological sorting of a directed, acyclic graph

In this short appendix we show that for a directed acyclic graph, a relabelling of the vertices exist such that the adjacency matrix of the relabelled graph is upper triangular. This labeling is often called the topological sorting [19] (chapter 22.4).

Lemma B.1. *If $G = (V, E)$ is a simple, finite, directed acyclic graph (DAG) then G contains an element with in-degree 0.*

Proof. Suppose not. Then take any $v \in V$. There exists $(u, v) \in E$ for some $u \in V$. This may repeated indefinitely. Hence due to finiteness we have constructed a cycle, yielding a contradiction. □

Lemma B.2. *For a DAG, we may label the vertices $v_1, \ldots, v_n$, with $n = |V|$, such that for all $(v_i, v_j) \in E$ we have $i < j$.*

Proof. Take a vertex with in-degree 0, which exists by Lemma B.1, and label this v_1. Let $G_1 = G$. For $k = 2, \ldots, n$, consider the subgraphs $G_k = G_{k-1} - \{v_k\}$, where we choose v_{k-1} to be a vertex with indegree 0 from G_{k-1}. Such a vertex will exist as all G_k are still DAGs. Under this labeling, suppose $(v_i, v_j) \in E$ with $j < i$. This implies at step $k = j$ in the above construction, v_j has positive indegree, which contradicts the above construction. Hence the result. □

Bibliography

[1] Altrock, P. M., L. L. Liu, and F. Michor. "The mathematics of cancer: Integrating quantitative models." *Nature Reviews Cancer* 15, 12: (2015) 730–745.

[2] Andersson, D. I., and D. Hughes. "Antibiotic resistance and its cost: is it possible to reverse resistance?" *Nature Reviews Microbiology* `http://www.nature.com/doifinder/10.1038/nrmicro2319`.

[3] Angerer, W. P. "An explicit representation of the Luria-Delbrück Distribution." *Journal of Mathematical Biology* 42, 2: (2001) 145–174.

[4] Antal, T., and P. L. Krapivsky. "Exact solution of a two-type branching process: clone size distribution in cell division kinetics." *Journal of Statistical Mechanics: Theory and Experiment* 2010, 07: (2010) P07,028. `http://stacks.iop.org/1742-5468/2010/i=07/a=P07028`.

[5] ———. "Exact solution of a two-type branching process: models of tumor progression." *Journal of Statistical Mechanics: Theory and Experiment* 2011, 08: (2011) P08,018. `http://stacks.iop.org/1742-5468/2011/i=08/a=P08018`.

[6] Athreya, K. B., and P. E. Ney. *Branching Processes*. Dover Publications, 2004.

[7] Bartlett, M. *An Introduction to Stochastic Processes*. Cambridge University Press, 1955, 3rd edition.

[8] Benzekry, S., C. Lamont, A. Beheshti, A. Tracz, J. M. L. Ebos, L. Hlatky, and P. Hahnfeldt. "Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth." *PLOS Computational Biology* 10, 8: (2014) 1–19. `https://doi.org/10.1371/journal.pcbi.1003800`.

[9] Billingsley, P. *Convergence of Probability Measures*. John Wiley and Sons, Inc, 1999, 2nd edition.

[10] Bingham, N., C. Goldie, and J. Teugels. *Regular Variation*. Cambridge University Press, 1987.

[11] Bixby, D., and M. Talpaz. "Seeking the causes and solutions to imatinib-resistance in chronic myeloid leukemia." *Leukemia* 25: (2010) 7. http://dx.doi.org/10.1038/leu.2010.23810.1038/leu.2010.238.

[12] Bozic, I., and M. A. Nowak. "Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers." *Proceedings of the National Academy of Sciences* 111, 45: (2014) 15,964–15,968. http://www.pnas.org/content/111/45/15964.

[13] ———. "Resisting Resistance." *Annual Review of Cancer Biology* 1, 1: (2017) 203–221. https://doi.org/10.1146/annurev-cancerbio-042716-094839.

[14] Bozic, I., J. G. Reiter, B. Allen, T. Antal, K. Chatterjee, P. Shah, Y. S. Moon, A. Yaqubie, N. Kelly, D. T. Le, E. J. Lipson, P. B. Chapman, J. Diaz, Luis A, B. Vogelstein, and M. A. Nowak. "Evolutionary dynamics of cancer in response to targeted combination therapy." *eLife* 2: (2013) e00,747. https://doi.org/10.7554/eLife.00747.

[15] Carlson, J. A. "Tumor Doubling Time of Cutaneous Melanoma and its Metastasis." *The American Journal of Dermatopathology* 25, 4: (2003) 291–299. http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000372-200308000-00003.

[16] Cheek, D., and T. Antal. "Mutation frequencies in a birth-death branching process." *ArXiv e-prints* , 1710.09783. https://arxiv.org/abs/1710.09783.

[17] Clauset, A., C. R. Shalizi, and M. E. J. Newman. "Power-law distributions in empirical data." *SIAM Review* 51: (2009) 661–703.

[18] Colijn, C., T. Cohen, A. Ganesh, and M. Murray. "Spontaneous Emergence of Multiple Drug Resistance in Tuberculosis before and during Therapy." *PLOS ONE* 6, 3: (2011) 1–7. https://doi.org/10.1371/journal.pone.0018327.

[19] Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*, volume 7. The MIT Press, Cambridge, Massachusetts., 2009, third edition.

[20] Daley, D. J., and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer-Verlag, 1988.

[21] Deresinski, S. "Vancomycin in Combination with Other Antibiotics for the Treatment of Serious MethicillinResistant Staphylococcus aureus Infections." *Clinical Infectious Diseases* 49, 7: (2009) 1072–1079. http://cid.oxfordjournals.org/content/49/7/1072.long.

[22] Dewanji, A., E. G. Luebeck, and S. H. Moolgavkar. "A generalized Luria-Delbrück model." *Mathematical Biosciences* 197, 2: (2005) 140–152.

[23] Dewanji, A., J. Jeon, R. Meza, and E. G. Luebeck. "Number and Size Distribution of Colorectal Adenomas under the Multistage Clonal Expansion Model of Cancer." *PLOS Computational Biology* 7, 10: (2011) 1–10. https://doi.org/10.1371/journal.pcbi.1002213.

[24] Diaz Jr, L. A., R. T. Williams, J. Wu, I. Kinde, J. R. Hecht, J. Berlin, B. Allen, I. Bozic, J. G. Reiter, M. A. Nowak, K. W. Kinzler, K. S. Oliner, and B. Vogelstein. "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers." *Nature* 486: (2012) 537–540.

[25] Dingli, D., F. Michor, T. Antal, and J. M. Pacheco. "The emergence of tumor metastases." *Cancer Biology & Therapy* 6, 3: (2007) 383–390. https://doi.org/10.4161/cbt.6.3.3720. PMID: 17312385.

[26] DLMF. "*NIST Digital Library of Mathematical Functions*." http://dlmf.nist.gov/, Release 1.0.18 of 2018-03-27, . http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[27] Durrett, R., and S. Moseley. "Evolution of resistance and progression to disease during clonal expansion of cancer." *Theoretical Population Biology* 77, 1: (2010) 42–48. http://dx.doi.org/10.1016/j.tpb.2009.10.008.

[28] Durrett, R., J. Foo, K. Leder, J. Mayberry, and F. Michor. "Evolutionary dynamics of tumor progression with random fitness values." *Theoretical Population Biology* 78, 1: (2010) 54 – 66. http://www.sciencedirect.com/science/article/pii/S0040580910000444.

[29] Enriquez-Navas, P. M., Y. Kam, T. Das, S. Hassan, A. Silva, P. Foroutan, E. Ruiz, G. Martinez, S. Minton, R. J. Gillies, and R. A. Gatenby. "Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer." *Science Translational Medicine* 8, 327: (2016) 327ra24–327ra24. http://stm.sciencemag.org/content/8/327/327ra24.

[30] Flajolet, P., and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, 2009.

[31] Foo, J., and F. Michor. "Evolution of acquired resistance to anti-cancer therapy." *Journal of Theoretical Biology* 355: (2014) 10–20.

[32] Ford, C. B., R. R. Shah, M. K. Maeda, S. Gagneux, M. B. Murray, T. Cohen, J. C. Johnston, J. Gardy, M. Lipsitch, and S. M. Fortune. "Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis." *Nature Genetics* 45, 7: (2013) 784–790.

[33] Fu, F., M. A. Nowak, and S. Bonhoeffer. "Spatial Heterogeneity in Drug Concentrations Can Facilitate the Emergence of Resistance to Cancer Therapy." *PLoS Computational Biology* 11, 3.

[34] Gillespie, J. H. "A simple stochastic gene substitution model." *Theoretical Population Biology* 23, 2: (1983) 202 – 215. http://www.sciencedirect.com/science/article/pii/004058098390014X.

[35] Gokhale, C. S., Y. Iwasa, M. A. Nowak, and A. Traulsen. "The pace of evolution across fitness valleys." *Journal of Theoretical Biology* 259, 3: (2009) 613–620.

[36] Greulich, P., B. Waclaw, and R. J. Allen. "Mutational Pathway Determines Whether Drug Gradients Accelerate Evolution of Drug-Resistant Cells." *Phys. Rev. Lett.* 109: (2012) 088,101. https://link.aps.org/doi/10.1103/PhysRevLett.109.088101.

[37] Haeno, H., M. Gonen, M. Davis, J. Herman, C. Iacobuzio-Donahue, and F. Michor. "Computational Modeling of Pancreatic Cancer Reveals Kinetics of Metastasis Suggesting Optimum Treatment Strategies." *Cell* 148, 1: (2012) 362–375. http://dx.doi.org/10.1016/j.cell.2011.11.060.

[38] Haeno, H., Y. Iwasa, and F. Michor. "The Evolution of Two Mutations During Clonal Expansion." *Genetics* 177, 4: (2007) 2209–2221. http://www.genetics.org/content/177/4/2209.

[39] Hall, A. R., J. C. Iles, and R. C. MacLean. "The fitness cost of rifampicin resistance in Pseudomonas aeruginosa depends on demand for RNA polymerase." *Genetics* 187, 3: (2011) 817–822.

[40] Hanin, L., J. Rose, and M. Zaider. "A stochastic model for the sizes of detectable metastases." *Journal of Theoretical Biology* 243, 3: (2006) 407–417.

[41] Hermsen, R., and T. Hwa. "Sources and Sinks: A Stochastic Model of Evolution in Heterogeneous Environments." *Phys. Rev. Lett.* 105: (2010) 248,104. https://link.aps.org/doi/10.1103/PhysRevLett.105.248104.

[42] Houchmandzadeh, B. "General formulation of Luria-Delbrück distribution of the number of mutants." *Physical Review E, 012719* 92.

[43] Hudson, D. J. "Interval estimation from the likelihood function." *Journal of the Royal Statistical Society. Series B (Methodological)* 33, 2: (1971) 256–262.

[44] Hughes, D., and D. I. Andersson. "Evolutionary consequences of drug resistance: Shared principles across diverse targets and organisms." *Nature Reviews Genetics* 16, 8: (2015) 459–471.

[45] Iwasa, Y., F. Michor, and M. A. Nowak. "Evolutionary dynamics of invasion and escape." *Journal of Theoretical Biology* 226, 2: (2004) 205–214.

[46] Iwasa, Y., M. A. Nowak, and F. Michor. "Evolution of resistance during clonal expansion." *Genetics* 172, 4: (2006) 2557–2566.

[47] Janson, S. "Functional limit theorems for multitype branching processes and generalized Pólya urns." *Stochastic Processes and their Applications* 110, 2: (2004) 177–245.

[48] Jeon, J., R. Meza, S. H. Moolgavkar, and E. G. Luebeck. "Evaluation of screening strategies for pre-malignant lesions using a biomathematical approach." *Mathematical Biosciences* 213, 1: (2008) 56–70.

[49] Jones, S., W.-d. Chen, G. Parmigiani, F. Diehl, N. Beerenwinkel, T. Antal, A. Traulsen, M. A. Nowak, C. Siegel, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, J. Willis, and S. D. Markowitz. "Comparative lesion sequencing provides insights into tumor evolution." *Proceedings of the National Academy of Sciences* 105, 11: (2008) 4283–4288. `http://www.pnas.org/content/105/11/4283`.

[50] Karlin, S., and H. M. Taylor. *A Second Course in Stochastic Processes*. New York - San Francisco - London: Academic Press, Inc., a subsidiary of Harcourt Brace Jovanovich, Publishers. XVI, 1981.

[51] Keller, P., and T. Antal. "Mutant number distribution in an exponentially growing population." *Journal of Statistical Mechanics: Theory and Experiment* 2015, 1: (2015) P01,011. `http://stacks.iop.org/1742-5468/2015/i=1/a=P01011`.

[52] Kendall, D. G. "On some modes of population growth leading to R. A. Fisher's logarithmic series distribution." *Biometrika* 35, 1/2: (1948) 6–15.

[53] ———. "Birth-and-death processes, and the theory of carcinogenesis." *Biometrika* 47: (1960) 13–21.

[54] Kessler, D., and H. Levine. "Scaling solution in the large population limit of the general asymmetric stochastic LuriaDelbrück evolution process." *Journal of Statistical Physics* 158,, 4: (2015) 783–805.

[55] Kessler, D. A., R. H. Austin, and H. Levine. "Resistance to chemotherapy: patient variability and cellular heterogeneity." *Cancer research* 74, 17: (2014) 4663–70.

[56] Klein, A. M., D. E. Brash, P. H. Jones, and B. D. Simons. "Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia." *Proceedings of the National Academy of Sciences* 107, 1: (2010) 270–275. `http://www.pnas.org/content/107/1/270`.

[57] Komarova, N. L., and D. Wodarz. "Drug resistance in cancer: Principles of emergence and prevention." *Proceedings of the National Academy of Sciences* 102, 27: (2005) 9714–9719. `http://www.pnas.org/cgi/doi/10.1073/pnas.0501870102`.

[58] Komarova, N. L., L. Wu, and P. Baldi. "The fixed-size Luria-Delbrück model with nonzero death rate." *Mathematical Biosciences* 210: (2007) 253–290.

[59] Krapivsky, P. L., and S. Redner. "Organization of growing random networks." *Phys. Rev. E* 63: (2001) 066,123. https://link.aps.org/doi/10.1103/PhysRevE.63.066123.

[60] Lea, D. E., and C. A. Coulson. "The distribution of the numbers of mutants in bacterial populations." *Journal of Genetics* 49, 3: (1949) 264–285.

[61] Leder, K., J. Foo, B. Skaggs, M. Gorre, C. L. Sawyers, and F. Michor. "Fitness Conferred by BCR-ABL Kinase Domain Mutations Determines the Risk of Pre-Existing Resistance in Chronic Myeloid Leukemia." *PLOS ONE* 6, 11: (2011) 1–11. https://doi.org/10.1371/journal.pone.0027682.

[62] Luria, S. E., and M. Delbrück. "Mutations of bacteria from virus sensitivity to virus resistance." *Genetics* 48, 6: (1943) 491–511.

[63] Michor, F., T. P. Hughes, Y. Iwasa, S. Branford, N. P. Shah, C. L. Sawyers, and M. A. Nowak. "Dynamics of chronic myeloid leukaemia." *Nature* 435, 7046: (2005) 1267–1270. http://dx.doi.org/10.1038/nature03669.

[64] Möhle, M. "Convergence Results for Compound Poisson Distributions and Applications to the Standard Luria-Delbrück Distribution." *Journal of Applied Probability* 42, 3: (2005) pp. 620–631. http://www.jstor.org/stable/30040845.

[65] Moreno-Gamez, S., A. L. Hill, D. I. S. Rosenbloom, D. A. Petrov, M. A. Nowak, and P. S. Pennings. "Imperfect drug penetration leads to spatial monotherapy and rapid evolution of multidrug resistance." *Proceedings of the National Academy of Sciences* 112, 22: (2015) E2874–E2883. http://www.pnas.org/content/112/22/E2874.abstract.

[66] Murray, J. D. *Mathematical Biology I. An Introduction*, volume 17. Springer, 2002.

[67] Newman, M. "Power laws, Pareto distributions and Zipf's law." *Contemporary Physics* 46 (5): (2005) 323–351.

[68] Nicholson, M. D., and T. Antal. "Universal asymptotic clone size distribution for general population growth." *Bull Math Biol* 78: (2016) 22432276.

[69] Paterson, C., M. A. Nowak, and B. Waclaw. "An exactly solvable, spatial model of mutation accumulation in cancer." *Scientific Reports* 6: (2016) 39,511. http://dx.doi.org/10.1038/srep3951110.1038/srep39511https://www.nature.com/articles/srep39511#supplementary-information.

[70] Pearson, J. E., P. Krapivsky, and A. S. Perelson. "Stochastic Theory of Early Viral Infection: Continuous versus Burst Production of Virions." *PLOS Computational Biology* 7, 2: (2011) 1–17. https://doi.org/10.1371/journal.pcbi.1001058.

[71] Primeau, A. J., A. Rendon, D. Hedley, L. Lilge, and I. F. Tannock. "The Distribution of the Anticancer Drug Doxorubicin in Relation to Blood Vessels in Solid Tumors." *Clinical Cancer Research* 11, 24: (2005) 8782–8788. http://clincancerres.aacrjournals.org/content/11/24/8782.

[72] Robbins, R. J. "Electronic Scholarly Publishing project.", 2001. http://old.esp.org/foundations/genetics/classical/holdings/l/slmd-43.pdf.

[73] Rodriguez-Brenes, I. A., N. L. Komarova, and D. Wodarz. "Tumor growth dynamics: insights into evolutionary processes." *Trends in Ecology & Evolution* 28, 10: (2013) 597 – 604. http://www.sciencedirect.com/science/article/pii/S0169534713001420.

[74] Rosche, W. A., and P. L. Foster. "Determining mutation rates in bacterial populations." *Methods* 20, 1: (2000) 4 – 17. http://www.sciencedirect.com/science/article/pii/S1046202399909015.

[75] Schiff, J. *The Laplace Transform: Theory and Applications*, volume 85. Springer, 1999.

[76] Simon, H. A. "On a class of skew distribution functions." *Biometrika* 42 (3-4): (1955) 425–440.

[77] Sniegowski, P. D., and R. E. Lenski. "Mutation and Adaptation: The Directed Mutation Controversy in Evolutionary Perspective." *Annual Review of Ecology and Systematics* 26: (1995) 553–578. http://www.jstor.org/stable/2097219.

[78] Spratt, J. S., J. S. Meyer, and J. A. Spratt. "Rates of growth of human neoplasms: Part II." *Journal of Surgical Oncology* 61, 1: (1996) 68–83.

[79] Stein, E. M., and R. Shakarchi. *Complex Analysis*. Princeton University Press, 2003.

[80] Tadrowski, A. C., M. R. Evans, and B. Waclaw. "Phenotypic switching can speed up biological evolution of microbes." *ArXiv e-prints* , 1601.04600. https://arxiv.org/abs/1601.04600.

[81] Tavare, S. "The birth process with immigration, and the genealogical structure of large populations." *Journal of Mathematical Biology* 25: (1987) 161–168.

[82] Tipping, A. J., F. X. Manon, V. Lagarde, J. M. Goldman, and J. V. Melo. "Restoration of sensitivity to STI571 in STI571-resistant chronic myeloid leukemia cells." *Blood* 98, 13: (2001) 3864–3867.

[83] Tlsty, T. D., B. H. Margolin, and K. Lum. "Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbruck fluctuation analysis." *Proc Natl Acad Sci U S A* 86, 23: (1989) 9441–9445.

[84] Tomasetti, C. "On the probability of random genetic mutations for various types of tumor growth." *Bulletin of Mathematical Biology* 74, 6: (2012) 1379–1395.

[85] Turajlic, S., and C. Swanton. "Metastasis as an evolutionary process." *Science* 352, 6282: (2016) 169–175. `http://science.sciencemag.org/content/352/6282/169`.

[86] Varga, R. S. *Matrix Iterative Analysis*. Prentice-Hall International, 1962.

[87] Ward, H., G. G. Perron, and R. C. MacLean. "The cost of multiple drug resistance in Pseudomonas aeruginosa." *Journal of Evolutionary Biology* 22, 5: (2009) 997–1003.

[88] Warner, D. F., and V. Mizrahi. "Tuberculosis chemotherapy: The influence of bacillary stress and damage response pathways on drug efficacy.", 2006.

[89] Weissman, D. B., M. M. Desai, D. S. Fisher, and M. W. Feldman. "The rate at which asexual populations cross fitness valleys." *Theoretical Population Biology* 75, 4: (2009) 286 – 300. `http://www.sciencedirect.com/science/article/pii/S0040580909000264`. Sam Karlin: Special Issue.

[90] Williams, D. *Probability with Martingales*. Cambridge University Press, 1991.

[91] Williams, M. J., B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. "Identification of neutral tumor evolution across cancer types." *Nature Genetics* 48: (2016) 238–244.

[92] Yachida, S., S. Jones, I. Bozic, T. Antal, R. Leary, B. Fu, M. Kamiyama, R. H. Hruban, J. R. Eshleman, M. A. Nowak, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and C. A. Iacobuzio-Donahue. "Distant metastasis occurs late during the genetic evolution of pancreatic cancer." *Nature* 467, 7319: (2010) 1114–1117.

[93] Zheng, Q. "Progress of a half century in the study of the Luria-Delbrück distribution." *Mathematical Biosciences* 162, 1–2: (1999) 1–32.

[94] ———. "rSalvador: An R Package for the Fluctuation Experiment." *G3: Genes, Genomes, Genetics* 7, 12: (2017) 3849–3856. `http://www.g3journal.org/content/7/12/3849`.