UNIVERSITY OF CALIFORNIA

Los Angeles

A Search for Punctuated Patterns through Computational Modeling of Asexually Reproducing Unicellular Populations

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Geophysics and Space Physics

by

Mai Tran

2023

ABSTRACT OF THE DISSERTATION

A Search for Punctuated Patterns through Computational Modeling of Asexually
Reproducing Unicellular Populations

by

Mai Tran
Doctor of Philosophy in Geophysics and Space Physics
University of California, Los Angeles, 2023
Professor William I. Newman, Chair

In this dissertation, we investigate whether evolutionary paths for simple biological systems exhibiting punctuated or intermittent behavior through computational modeling of asexually reproducing unicellular populations. Punctuated equilibrium is a controversial theory proposed by Gould and Eldredge in 1972 [1] to explain observations of species experiencing long periods of stasis followed by bursts of change in the fossil record. The challenges in interpreting the fossil record and in the modeling of complex ecosystems has resulted in limited progress in settling the debate around this conjecture. In our work, we aim to build computational models of a simple biological system to study the behavior of fitness evolution at the population scale where fitness refers to an organism's reproductive success. We are motivated by the results from the ongoing Long-Term Evolution Experiment (LTEE) [2], which observes punctuated patterns in the fitness of *E. coli* populations grown in a laboratory setting. We construct two models of organismic evolution whose results are consistent with the fitness trends observed in the LTEE. First, we develop a Monte Carlo Wright-Fisher model that we adapt to model the LTEE populations. Our adapted Wright-Fisher model exhibits punctuated patterns at the bacteria's generational time scale. Second, we develop and introduce the smooth-colony dynamics model to study a population of colonies, which is more naturally realistic than the adapted Wright-Fisher model's assumptions and the LTEE's laboratory environment. Our main result shows that while this more biologically realistic model fits the LTEE data, it does not display punctuated patterns in absence of epistasis, or the interactions of genes. This indicates that punctuated behavior at the generation time scale may be limited to specific conditions that may not be naturally occurring. Further work aims to expand the complexity of the smooth-colony dynamics model to provide additional insights into the evolutionary trends of simple biological systems.

The dissertation of Mai Tran is approved.

David Clifford Jewitt

Aaron S. Meyer

James William Schopf

William I. Newman, Committee Chair

University of California, Los Angeles

2023

To my family

Table of Contents

List of Figures

Acknowledgments

I want to thank my parents and my siblings for being the perfect models of kindness and resilience, words cannot describe how indebted I am to you all.

It has been a humbling and rewarding experience. I am grateful for the opportunities I have been given here to learn and grow intellectually and personally. I want to thank my advisor, Professor William I. Newman, for his guidance and insight. He has been generous with his time and knowledge. My committee – Professor William Schopf, Professor David Jewitt, and Professor Aaron Meyer – have been encouraging, supportive, and thoughtful. I want to especially thank Professor David Jewitt for his guidance and support during challenging times. I want to thank Professor Jean-Luc Margot for taking a chance on me and Professor An Yin for always reaching out and giving me the necessary reassurance. I want to thank Professor David Paige for his support and providing a different perspective. I want to thank Professor Tina Treude for her support. I want to thank Dr. Sorel Fitz-Gibbon for the introduction of the Long-Term Evolution Experiment. I want to thank Dr. Anthony Friscia, for nurturing my teaching interests.

I am thankful to all my friends and colleagues, Fernando Becerra, Tina Diao, Kalpit Dixit, Charunethran Panchalam Govindarajan, Ariel Graykowki, Edward Jung, Dana Li, Isaac Matthews, Kiran Palla, Liutauras Rusaitis, Sanchita Sharma, Edward Tierra, Ashley Yu, who have accompanied me on this journey and made it all the richer. I am fortunate to have witnessed their grace, patience, humor, generosity, intellect, and hopefulness. I am thankful to Adam Lawrence for widening my intellectual interests. I am especially grateful to Kartik Ahuja, who has never doubted me nor failed to challenge me intellectually. I want to thank Alireza Sobrabi for our long chats during the COVID-19 pandemic, his curiosity and openness inspire me every day. I want to thank Debraj Bose for helping me edit this dissertation and picking up my numerous phone calls over the years.

Lastly, I want to thank the Gates Millennium Scholars Program for partially funding my graduate education.

Curriculum Vitae

2010 – 2014	B.S. in Managerial Economics, University of California, Davis
2015 – 2017	M.S. in Management Science and Engineering, Stanford University, California
2017 – 2021	M.S. in Geophysics and Space Physics, University of California, Los Angeles
2017 – present	Ph.D. student in Geophysics and Space Physics, University of California, Los Angeles
2019 – 2021	Teaching Assistant and Teaching Fellow, CLUSTER 70: Evolution of Cosmos and Life, Undergraduate Education Initiatives, UCLA

Publications

M. Tran, K. Ahuja, "Convolutional Neural Networks Based Random Projections for EEG Prediction Tasks", extended abstract in *IEEE Engineering in Medicine and Biology Society's Biomedical and Health Informatics Conference, 2019*

CHAPTER 1

Motivation

The observation of long periods of stasis with intermittent episodes of abrupt changes in the fossil record is well established in paleontology. The conjecture put forward by Gould and Eldredge in 1972 [1] proposes that this observation is a feature of the evolutionary process and not simply ad-hoc conditions that are case specific nor a result of sampling bias. It has been suggested since Gould and Eldredge's original 1972 publication that certain conditions can induce a punctuated behavior in evolution [4], though there has been limited evidence supporting this hypothesis. There is currently no consensus on the mechanisms resulting in punctuation nor total refutation of Gould and Eldredge's conjecture leaving this important and interesting observation in the fossil record unexplained.

The majority of the mathematical modeling efforts in the literature found, e.g. the Bak-Sneppen model [5], have focused on the species level, for the understandable reason that this is likely what the fossil record can capture. However, there are numerous challenges due to, but not limited to, the length of time a species survives. For example, individual mammal species exist for about 1 to 2 million years in changing environments, and experience complex interactions within their ecosystem. In this work, we focus on the mathematical and computational modeling of a simpler biological system at the level of organisms for the study of punctuated equilibrium. We utilize real laboratory data by the Lenski group [6] to constrain our models, which is seldom feasible in species modeling to study punctuated equilibrium. Although this does not necessarily provide evidence at the species level, which would be necessary to explain the fossil record, we believe the observation of punctuation at the organism level in our work, if observed, would demonstrate that identifiable biological mechanisms can be modeled to replicate the observed behaviors. These insights open a potential pathway forward to look at more complex systems in the study of punctuated equilibrium. Additionally, the models developed and utilized in this dissertation, which are constrained by real biological data, provide new tools to tease out conditions and mechanisms for further research.

The objective of this dissertation is to demonstrate that the modeling of fitness evolutionary paths for simple biological systems, where **fitness** is defined as reproductive success being represented by a population's growth rate, can provide insights into population evolutionary trends. We investigate whether punctuated patterns can emerge where asexually reproducing populations experience long periods of stasis which are interrupted by abrupt episodes of change. We want to highlight the contribution of our work in deploying two models for our investigation. We develop a Monte Carlo Wright-Fisher model that we have not seen utilized in the study of punctuated equilibrium. We construct and introduce the new smooth-colony dynamics model that combines elements of discrete populations and overlapping generational times. Additionally, we provide an approximation of the Gerrish-Lenski mathematical model that allows faster computational ability without significant reduction in accuracy.

In this chapter, we discuss the concept of punctuated equilibrium and provide an outline of the dissertation. Terms highlighted in **bold** are defined briefly in the text and accompanied by more extensive definitions provided in the glossary, or appendix B.

1.1 Punctuated equilibrium

The fossil record provides evidence of evolutionary processes over an extended geological period of time, albeit a very incomplete picture. Paleontologists have long observed that the fossil record seems to show that species experience morphological stasis for the majority of their lifespans, or time from origination to extinction, followed by periods of rapid changes relative to geological time. An extreme example, evoked by Bergstrom and Dugatkin (2016) [7], of rapid changes is the Cambrian explosion where a larger number of new body forms emerged approximately 543 to 490 million years ago. There have been attempts to explain these step-like observations in the fossil record through “data collection bias.” Indeed, the likelihood that any individual organism becomes fossilized is extremely small and is dependent on a numerous factors, including its composition, structure, behavior, and habitat. Members of long-living large populations have a higher, yet nevertheless low chance of leaving a discernable preserved record. The sedimentary rock cycle adds yet another layer of complexity, biasing in favor of younger fossils being preserved. Taken together, these factors lead to a very low sampling rate.

The data from the fossil record are not inconsistent with a step-like, rather than a gradual progression, of **morphological change**. In 1972, Gould and Eldredge [1] suggested that the unchanging intervals seen in the fossil species may not be due to a limitation of the data-collection methods but instead evidence of a real pattern in macro-evolution. They coined the term ***punctuated equilibrium*** (PE) to describe the conjecture that species spend most of their life spans without obvious evolutionary change interrupted by shorter periods of rapid morphological diversification, a result of **speciation** events, or creation of species, that occur rapidly relative to geological time. In other words, the history of life could be defined by species being evolutionarily inactive and relatively stable and then undergoing bursts of rapid changes or speciation. This contrasts with the view that species accumulate gradual morphological changes over time that may lead to new species. Figure 1.2 provides a schematic distinguishing between these two ideas. In this framework, a more well-sampled fossil record would show continuous changes over time with variable rates. An example is the changes in shell conicity of the planktonic foraminifer *Contusotruncana* lineage measured in the range of about 5 million years by Kucera and Malmgren (1998) [8] which shows gradual changes over time, see figure 1.1.

Since its introduction, the PE concept has been extensively debated in the evolutionary biology and paleontology communities. The current evidence suggests that species can exhibit both gradual changes and periods of rapid shifts (Erwin and Anstey 1995 [10]). The factors that influence this rate are complicated and difficult to disentangle. A punctuated

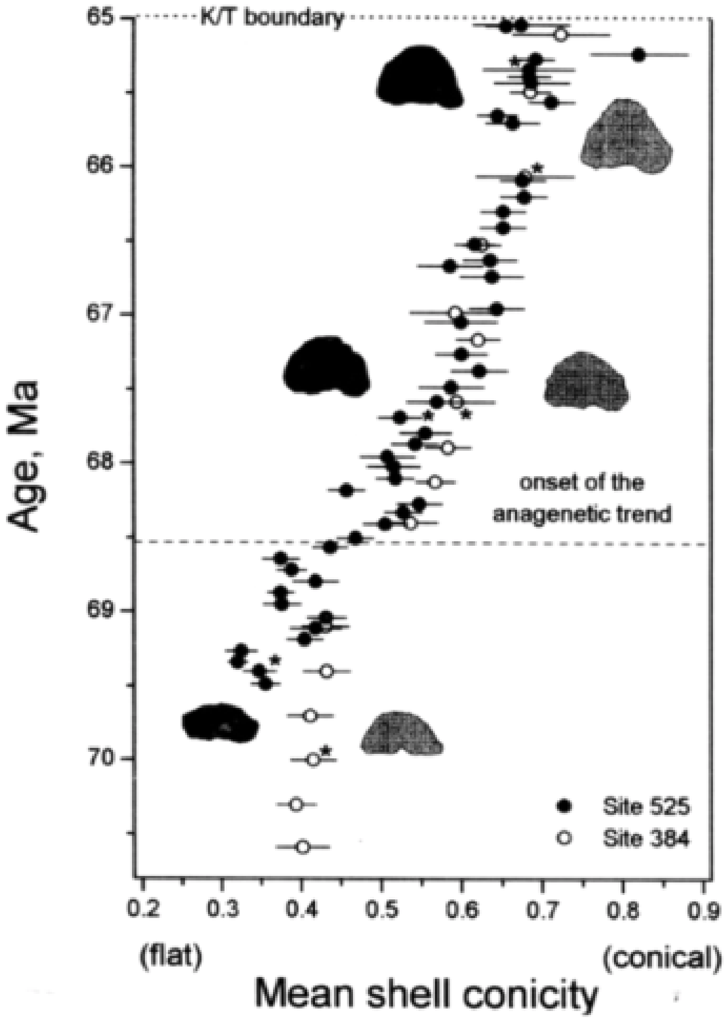


Figure 1.1: The mean shell conicity of the planktonic foraminifer *Contusotruncana* lineage measured from 70 million years ago to 65 million years ago shows gradual changes over time. This figure is extracted from Kucera and Malmgren (1998) [8].

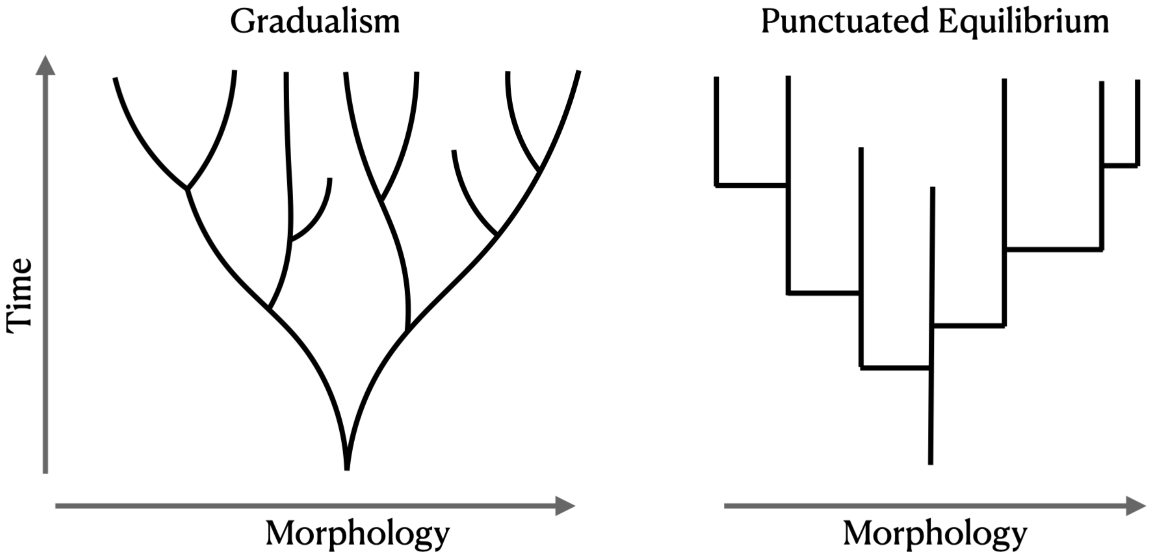


Figure 1.2: Punctuated equilibrium proposes that species experience little morphological change for an extended period of time followed by an abrupt speciation event. The concept of PE thus sharply contrasts with "gradualism" according to which changes are accumulated moderately over time. As seen on the left schematic demonstrating gradualism, species experience morphological changes gradually over time. On the right schematic demonstrating PE, species experience abrupt changes then stay in stasis. We adapt this figure from Benton and Pearson (2001) [9].

equilibrium-like trend has also been observed at the **genomic**, or DNA, level, where bursts of elevated **mutation** accumulation rate have been observed (Heasley *et al.* 2021 [11], Lenski 2017 [12]). How this observation shapes speciation events and the overall macro-evolutionary trend is not clear. Questions regarding the predominant pattern of evolutionary progress and under what conditions one might expect to observe one pattern rather than another are yet to be resolved (Bergstrom and Dugatkin 2016 [7]).

1.2 Long-Term Evolution Experiment (LTEE)

We aim to construct our theoretical model to be simple yet fully driven by our understanding of biology. We are motivated by the work of Richard Lenski and his collaborators on microorganisms through an ongoing project that he started in 1988 called the Long-Term Evolution Experiment or LTEE [6]. For over 30 years — and continuing — Lenski and his colleagues have been growing twelve populations of ***Escherichia coli (E. coli)*** cultivated from one ancestral strain under identical laboratory conditions (Lenski 2017 [12]). The ancestral strain does not have the capacity for **horizontal gene transfer**, or the non-reproductive transfer of genetic material from one organism to another (Lenski 2017 [12]). Thus all genotypic variations that emerge in these populations arise from genetic mutations. Selection then acts on these mutations depending on their impact on the organisms' fitness. In the context of the LTEE, fitness is a measure of a population's reproductive success, approximated by its growth rate relative to the ancestral population. A fitness of two means that a population grows twice as fast as its ancestral population during one growth period of 24 hours. The LTEE workers take fitness measurements of the populations and also freeze the samples periodically for future analysis. Each population undergoes a 100 fold growth, or roughly 6.64 generations, within a 24-hour period. They have published the fitness of the twelve populations through 50,000 generations and performed a range of genomic studies as technology has progressed (Lenski 2017 [12]). The fitness paths for all twelve populations display a non-decreasing trend with a declining rate of increase. We are interested in understanding these fitness trajectories that follow an aggregate **trait**, as fitness is determined by a number of different traits, providing a convenient starting point for modeling which can be built upon in future research.

1.3 Outline of dissertation

The observations highlighted above motivate the work outlined here. The focus of this dissertation is to establish a theoretical mathematical model built on a biological paradigm. In the simplest possible approach, we wish to explore the issues of evolutionary rate and pattern in a relevant biological system. Our model is guided by experimental data, allowing a more reliable interpretation of our results.

Genetic randomness is an essential feature of our model where we explore how a population can take different evolutionary paths in adapting to its environment. Central to the theory of evolution are the random processes that give rise to genetic variations. These can be advantageous, disadvantageous, or neutral. Natural selection would then weed out disadvantageous mutations, although under certain scenarios they may remain in the population. In looking at highly varied species, the associated **fitness landscape** is highly multi-dimensional, due to the presence of many different genetic attributes, complicating the study of related biological behaviors. From thermodynamic principles, we expect that there will be errors in genetic material **transcriptions**, **replications**, etc. during DNA related processes, due to the randomness present in the associated complex chemical kinetics. The different transcription and replication possibilities or routes are determined by reaction rates which are in turn controlled by chemical kinetics. Populations with different **genotypes**, or genetic makeup, undergo competition defined by these chemical reactions.

To model genetic randomness, we employ what is called a **mutation rate** μ, which is the probability that an organism experiences a mutation during its reproduction in one generation. A mutation altering an organism's genetic material may or may not impact the fitness of the organism. Mutation is important in providing new genetic material for natural selection to act upon. However, for this work, we are focusing on beneficial mutations and adaptation. We are using the mutation rate definition as the probability of an organism to obtain a beneficial mutation during one generation both for convenience and it being a reasonable metric. Although mutations can be neutral, deleterious, or beneficial, for large asexually reproducing populations, deleterious mutations have a relatively minor effect on fitness as they are eliminated from the population the same rate as their introduction [13]. Additionally, Wielgoss *et al.* (2013) [14] measures a relatively small fitness impact from deleterious mutations in the LTEE populations.

This dissertation is structured as follows. Chapter Two provides an overview of the Long-Term Evolution Experiment (LTEE) and statistical modeling of its fitness trend. Lenski *et al.* (1991) [2] shows step-like trend in fitness measurements of a single LTEE population. Wiser *et al.* (2013) [3] provides fitness measurements across the LTEE populations for 50,000 generations that exhibit a diminishing fitness growth rate over time. We discuss the Gerrish-Lenski model, introduced by Gerrish and Lenski (1998) [15], that is employed by the LTEE's experimenters in predicting the fitness trajectory of the LTEE populations. The Gerrish-Lenski model is the first comprehensive description at attempting to model the genotypic competition within an asexually reproducing populations. Wiser *et al.* (2013) [3] adapted the Gerrish-Lenski model and was able to predict the LTEE fitness data quantitatively well.

Chapter Three covers our reproduction and further computational adaptation of the Bak-Sneppen model (Bak and Sneppen 1993 [5]). We employ the hierarchical data structure developed by Newman and Phoenix [16][17] for a class of problems arising in materials science where there is, in effect, competition between nearest neighbors under stress. This algorithm provides the most efficient way, theoretically, to perform localized comparison and identify the weakest elements. This algorithm enhances simulation speed by a factor of $N/\log N$, offering dramatically improved performance by a factor of 10^5 or greater. Moreover, since the time of the Bak-Sneppen model [5], computer processor speeds have increased by a factor of 1,000. The combination of methodological and hardware advancement makes it possible for us to undertake an ambitious project that can provide a much more realistic model for genetic changes where we analyze the behavior of a system with genotypes competing with nearest neighbors.

In Chapter Four, we adapt the Wright-Fisher model in modeling the LTEE and analyzing fitness trends in different time scales. The Wright-Fisher model is a commonly used population genetic model that attempts to characterize genotype frequencies under genetic drift, mutation, and selection. We find that for simulations of small populations with low mutation rate leading to few mutations per generation, distinct step-like trends in fitness can be observed. However, for large populations with the same mutation rate, simulations of the Wright-Fisher model exhibit a more smooth curve with small "jumps" in fitness that are not distinct and occur in relatively few generations. Utilizing the parameters proposed by Wiser *et al.* (2013) [3], our simulation results of the Wright-Fisher model provide a quantitatively similar fitness trajectory for the LTEE.

In Chapter Five, we introduce our smooth-colony dynamics model for an asexually reproducing population under more realistic assumptions. This model is biologically more relevant with a character similar to the Bak-Sneppen model. We model **colonies** of organisms residing in a 1-D configuration where they evolve as discrete units. In our base model, the colonies evolve independently, like the populations in the LTEE, which we refer to as the non-interacting neighbors model. This is closely-related to the Wright-Fisher model where each colony is a population in the Wright-Fisher model. In this model, the colonies have overlapping generations with varying reproduction times. We extend this model into the interacting neighbors model, where the evolutionary path of each colony is impacted by their neighbors'. Our results show that the population – the collection of all colonies – experiences faster fitness improvements in the interacting neighbors model. We initially observe step-like trend in the early generations, possibly a transient effect due to our model construction. Unlike the Wright-Fisher model's results, observations in the longer term, after several thousands of generations, show exponential growth but no distinct jumps in fitness. Utilizing the parameters proposed by Wiser *et al.* (2013) [3], our simulation results of our smooth-colony dynamics model provide a quantitatively similar fitness trajectory for the LTEE. The fitness across a large number of colonies appears to eliminate the fitness jumps observed in a single population simulation in the Wright-Fisher model. This suggests the step-like fitness observed in the LTEE data is possibly due to a single population's data.

All three models, the Wright-Fisher, the non-interacting neighbors and interacting neighbors smooth-colony dynamics models, are able to fit the LTEE data quantitatively well. However, we observe "jumps" in fitness trajectories, or evolutionary paths, in the Wright-Fisher model. This is due to a single population simulation. Further research is needed to distinguish which model is more accurate in capturing the operations of a naturally occurring asexually reproducing population.

We end with a summary chapter where we highlight our results and the direction of future research efforts. After exploring a variety of models with varying degrees of sophistication, we are unable to verify concrete evidence confirming punctuated equilibrium at the scale of thousands of generations for asexual single organisms. Although we are able to observe step-like behavior in a single population in our Monte Carlo Wright-Fisher simulations, this also depends on the size of the population and the resolution of the number of generations. Averaging over a greater number of simulations reduces this effect. The smooth-colony dynamics model accounts for a large number of populations simultaneously and fails to observe any "punctuated" behavior. Thus, for large populations which operate as a collection of sub-populations, we are likely to observe smooth changes in the sampling of fitness growth over time. However, it might be possible to observe step-like behavior in a natural environment when sampling the fitness growth of a small population with no subdivisions. Further research is needed to tease out which regime naturally occurring populations operate in, and the conditions of the data sampled.

To assist the reading of this dissertation, we incorporate several appendixes, namely appendix A provides a list of acronyms and mathematical symbols; appendix B contains a glossary of important or relevant terms used in this dissertation that appear in **bold** text; and appendix C includes clarification of relevant mathematical concepts and algorithm procedures referenced in the main text.

CHAPTER 2

The Long-Term Evolution Experiment

In this chapter, we provide an overview of the on-going Long-Term Evolution Experiment, or LTEE, conducted by Lenski and his collaborators. The experiment aims to probe the nature of evolution in a tractable biological system. We highlight the results of particular interest for this work as they relate to our objective of investigating punctuated equilibrium. We then discuss the Gerrish-Lenski model of clonal interference in asexually reproducing organisms and the Wiser *et al.* (2013) [3]'s adaptation to predict the LTEE's data trajectory.

2.1 Introduction

Since 1988, Lenski and collaborators have been growing twelve populations of *Escherichia coli (E. coli)* cultivated from the same ancestral strain [12]. By keeping the environment constant – controlling for food source, temperature, and restricting horizontal gene transfer – the experiment allows differences in these populations to emerge through selection on random mutations and genetic drifts. Fitness changes are measured periodically as the evolved population grows alongside the ancestral strain. Fitness is a measurement of reproductive success. The experimenters estimate this through the relative growth ratio of the evolved population to the ancestral population, which will be elaborated momentarily. Lenski and collaborators have since expanded the original scope of the experiment by incorporating gene sequencing to study the genetic evolution of the populations across time.

The experiment starts by taking twelve sampled populations from one ancestral strain of *E. coli.* They are grown in a nutrient-limited medium and propagated daily by transferring 1% of the previous day's medium into a fresh one. The medium containing the bacteria is stored in a flask set in a shaking incubator at a constant temperature and rotation. During a 24-hour cycle, the bacterial population obtains a stationary phase density, where their population size remains constant. Thus, the *E. coli* population grows roughly 100 fold during each incubation period of a day, resulting in ≈ 6.64 generations of binary fission ($2^{6.64} = 100$). A more technical description of the experiment can be found in Lenski *et al.* (1991) [2].

Fitness is estimated as the relative growth rate of an evolved population to its ancestral population in the LTEE. This sometimes is referred to as the relative fitness. We will simply refer to this as a population's fitness for the rest of this work. For the first 2,000 generations, evolved populations are grown alongside the ancestral population, with fitness measurements obtained every 100 generations. This was later extended to 50,000 generations but for longer measurement intervals of 500 generations and 1,000 generations [3]. The fitness of population i, usually the derived population, relative to the population j, usually the ancestral population, is indicated as, W_{ij}. The relative fitness is estimated as the ratio of the number of doublings, D_i and D_j, of the two populations grown together in one 24-hour growth cycle:

$$W_j = D_i/D_j = \frac{\ln[N_i(1)/N_i(0)]/\ln(2)}{\ln[N_j(1)/N_j(0)]/\ln(2)} \tag{2.1}$$

where $N_i(0)$ and $N_i(1)$ indicate the initial density and the density after one day, respectively [2]. Densities are estimated by the counting of colonies. Additionally, at 100-generation intervals, samples are taken from the twelve evolving populations for future analysis, effectively keeping a snapshot of the evolutionary progress [2].

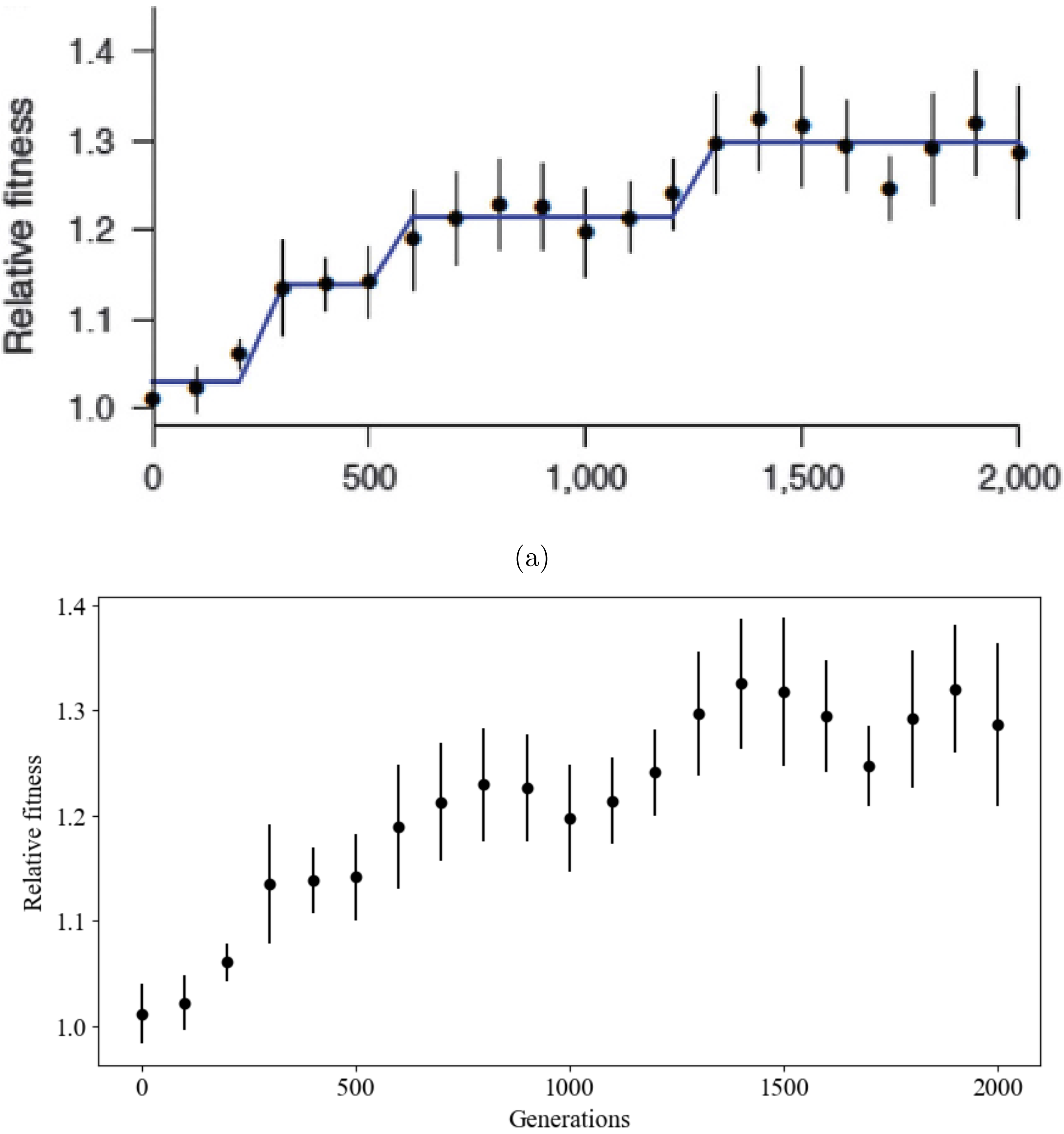
We recall that mutation rate refers to the probability of an organism in obtaining a mutation during one generation. Mutations are important in providing new genetic material for natural selection to act upon. Evolution of mutation rates in asexual populations is a complex process that involves the balancing of deleterious and beneficial mutations' effects [18] [19]. During the 50,000 generations, 6 of the 12 populations evolved **hypermutability** such that their mutation rate increased from the ancestral mutation rate [12]. Since we are interested in constructing the simplest of models to study punctuated equilibrium, we are keeping this parameter constant in our models. Thus, we are not including those hypermutating populations in forthcoming analyses as they have different fitness trajectories than the ones that maintain the ancestral mutation rate.

The simplicity of the setup and relatively inexpensive nature of running the experiment have allowed the researchers to keep the LTEE going for decades. They have utilized the experiment to approach a variety of evolutionary questions including but not limited to, the relationship between genotypic and phenotypic evolution in a population, and the patterns of the twelve populations' evolutionary paths. We are interested in the evolution of fitness across time observed within the population. Wiser *et al.* (2013) have measured relative fitness for most of the 12 populations through the first 50,000 generations [3]. We will be using this data set to analyze the evolutionary trajectory and validate our models in later chapters.

2.2 Fitness trend

The LTEE experimenters have observed that fitness relative to the ancestral strain increases over time, albeit at a decreasing rate [3]. The trend in fitness growth is particularly important to us as we would like to model the impact of mutations on fitness over time in an asexually reproducing population. Lenski *et al.* (1991) reported a step function-like trend in the mean fitness increase of the 12 populations in the first 2,000 generations [2]. Figure 2.1a shows the mean fitness data of one of the 12 populations measured every 100 generations with a fitting curve using an isotonic regression model with three steps as the final fit to the trajectory by Lenski and Travisano (1994) [20]. However, due to the discretized nature of the measurements and a single population's sample, we cannot definitely conclude that it exhibits a step-like trend. We have reproduced the same figure using Lenski and Travisano (1994) [20]'s data without a trend line in figure 2.1b. Importantly, this shows that we can be misled into saying "punctuation" if we fit straight line segments.

Wiser *et al.* (2013) [3] periodically measured the fitness of the LTEE populations over



(a)

(b)

Figure 2.1: (a) Fitness trajectory relative to the ancestral population for one of the 12 populations. Error bars are 95% confidence intervals based on 10 replicated assays. The line is fitted using a step model. This figure is extracted from Lenski (2017) [12]. (b) We reproduce figure 2.1a without the fitted step model trend line using data from the LTEE's archived website [6]. It is significant that given the prevailing uncertainty in measurement as indicated in the error bars that a step-wise progression is not necessarily a good representation for the raw data.

50,000 generations and found that the although the fitness continues to increase during this period, the rate of increase diminishes substantially over time. Figure 2.2 shows the mean relative fitness of 6 populations that retain their ancestral mutation rate over time (generations) obtained by the LTEE for 50,000 generations. As mentioned previously, we are not including the LTEE populations that have developed hypermutability, i.e. elevated mutation rate relative to the ancestral rate, in our analysis.

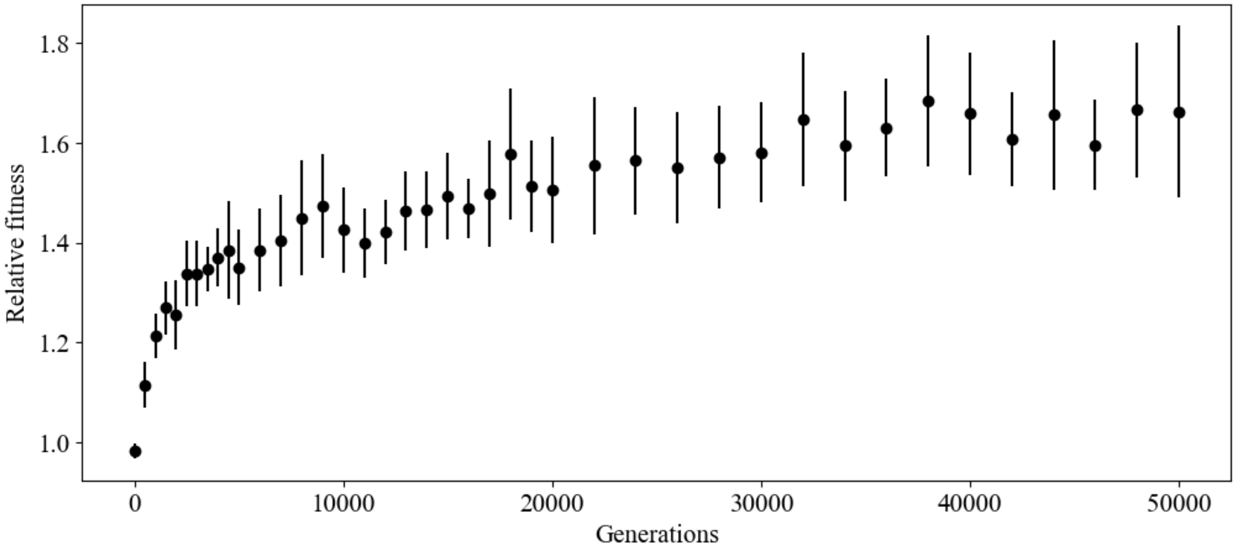


Figure 2.2: Mean fitness for the 6 populations that retained their ancestral mutation rate, error bars are 95% confidence limits. This figure is reproduced using data from Wiser *et al.* 2013 [3].

We now turn to the issue of the step-like trend observed in the fitness data during the first 2,000 generations. Lenski *et al.* (1991) attributes this fitness trend behavior to the notion that new beneficial mutations take a considerable number of generations, depending on their effects, to have appreciable impacts on the mean fitness of the population [2]. Their example works as follows. Let's look at a simple case where a mutation occurs in a purely asexual population, i.e. where an offspring is produced by a single parent and no horizontal gene transfer is present, with N members. Let $P(t)$ be the proportion of the population having this mutation at time t, starting at some arbitrary time, then $P(0)$ is $1/N$ when the mutation emerges in one member of the population. Assuming this mutation is beneficial, i.e. it increases fitness of the individual and enables it to reproduce at a faster rate than other members in the population, and the mutation is not lost due to random drift, then it spreads through the population as:

$$dP/dt = rP(1 - P) \tag{2.2}$$

where r is the selection-rate constant, the increase in reproduction rate characterized by the difference of the Malthusian parameters between the mutated organism and its ancestor [2]. This is of course, the derivative of the logistic function, see appendix C.1.2. Thus we have:

$$P(t) = \frac{1}{1 + (\frac{1}{P(0)} - 1)e^{-rt}} \tag{2.3}$$

Consequently, the mean fitness of the population, relative to the ancestral population, can be captured as a function of time:

$$\overline{W(t)} = 1 + \frac{r}{m_0}P(t) = 1 + sP(t) \tag{2.4}$$

where m_0 is the Malthusian parameter for the ancestor population. Let s be the selection coefficient, measuring the difference in relative fitness of the mutated genotype and the ancestral genotype or $s = W - 1$ where W is the relative fitness of the mutated genotype using the ancestral genotype as the reference point and it is estimated as $W = \frac{m_{\text{mutated}}}{m_0}$ then $s = \frac{r}{m_0}$. The derivation of equations (2.2) and (2.4) can be found in appendix C.2.

As shown in figure 2.3, which we reproduce from Lenski *et al.* (1991) [2], it can take many generations before the mutations' effects are reflected in the population's mean fitness. Note the selection rate constant used for the plots is per day, thus t in equation 2.4 is measured in days. We convert it to binary fission generations as the second axis which is more intuitive. For the LTEE experiment, there is 100-fold growth per day resulting in ≈ 6.64 generations per day. The increase takes place in relatively few number of generations before reaching a plateau, where the mutation is fixed in the population. **Fixation** of a mutation occurs when it becomes the most common, or **wild**, type in the population, i.e. the mutation is fixed. For two genotypes, it is easy to observe that when the genotype with a higher fitness takes over the population, a new mean fitness is obtained. However, for a population with many

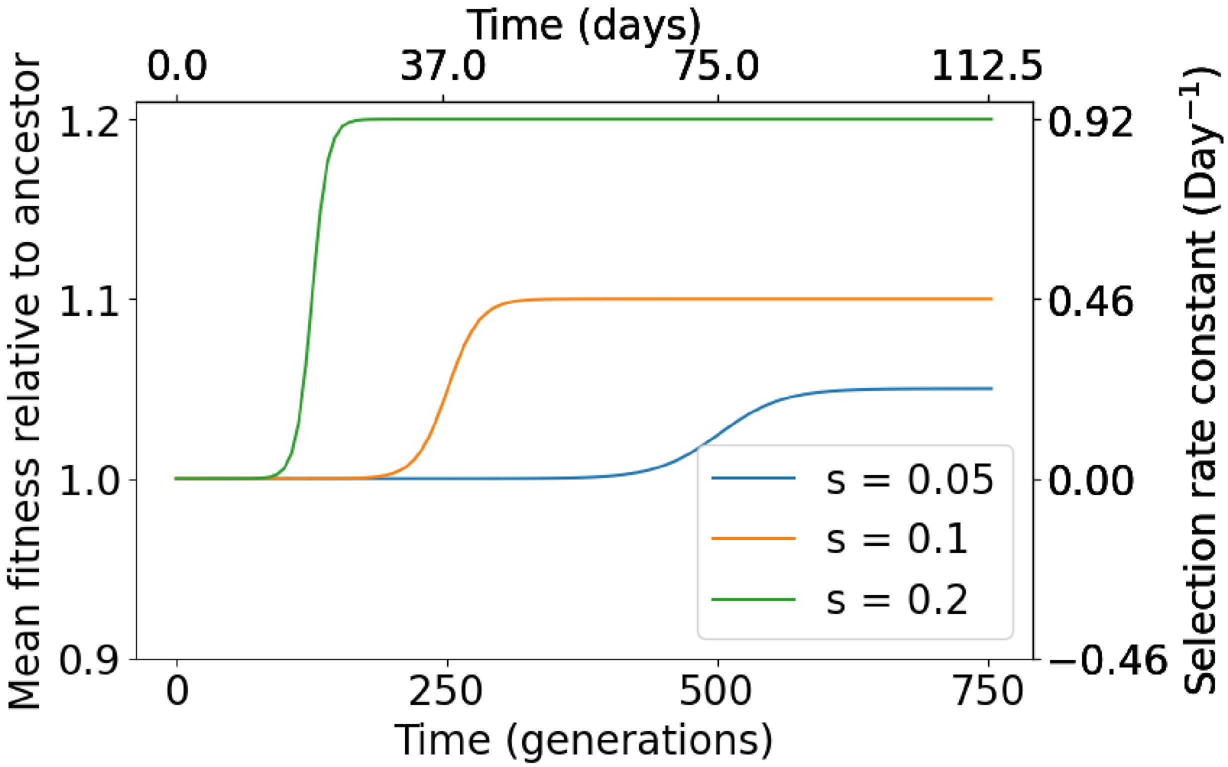
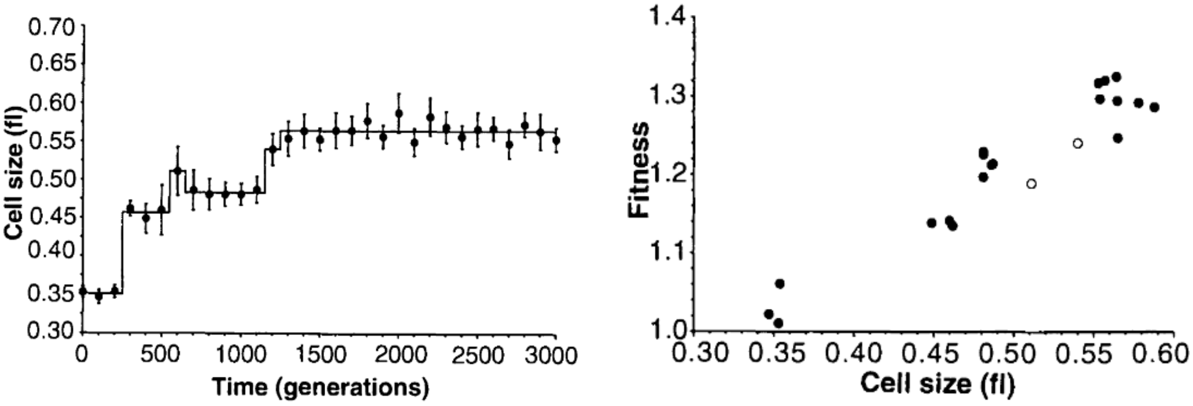


Figure 2.3: Plot for equation 2.4 with different s values: 0.05, 0.1, and 0.2. We reproduce this figure from Lenski *et al.* (1991) [2]. It takes a number of generations after the mutation is introduced into the population before it makes a measurable impact on the fitness of the population. Mutations with smaller selection coefficient s take longer than ones with larger values to be fixed in the population.

competing genotypes this new jump in fitness may not be observed so readily. We address this scenario in Chapter Four where we discuss the Wright-Fisher model.



(a)

(b)

Figure 2.4: (a) Average cell size over 3,000 generations for one population in the LTEE experiment [21]. Elena *et al.* (1996) fits a step function model to the data showing distinct steps in cell size growth. Error bars indicate 95% confidence intervals based on 10 replicate assays. (b) Mean relative fitness appears to correlate strongly with average cell size [21]. These figures are extracted from Elena *et al.* (1996), please refer to their paper [21].

Since the growth in fitness that is observed in figure 2.1a may not necessarily mean a morphological change, it is worth noting an observation by Elena *et al.* (1996) where the authors report parallel behavior in cell size growth over time during first 3,000 generations of the LTEE [21], figure 2.4a. They also show high correlation with the increasing-fitness trend where the increase in cell size corresponds to the increase in relative fitness, figure 2.4b. The authors did not pursue the question of whether cell size is selected for, or a byproduct of other phenotypes being targeted. For example, bigger cells may be able to commence growth faster than smaller cells. Another scenario is that faster growing cells tend to be larger so large cell phenotypes may be correlated with faster growth [21]. This provides a line of evidence that the fitness trend observed in the LTEE is a reasonable benchmark of morphological changes, although we must be cautious in our interpretation when using this data to inform our model.

2.3 Gerrish-Lenski model

A key feature for genetic evolution in asexual reproducing organism is their limited ability to combine different beneficial mutations originating from different organisms. In an asexual population, adaptation can potentially display an abrupt increase in fitness through clonal interference that occurs where different lineages are in competition, a concept introduced by Muller (1932) to assert the advantageous of sexual reproduction [22]. Since asexual populations cannot combine advantageous mutations through sexual reproduction, these beneficial mutations have to compete to be fixed in a population. As with the LTEE, horizontal gene transfer, i.e. the exchange of genetic material between bacteria without sexual reproduction, is absent in this model. Thus, they are likely to be acquired sequentially, unless in the event that later beneficial mutations occur in the mutated genotype before the original mutation is fixed in a population, creating a scenario where less beneficial mutations are hitchhiked through the superior mutations and multiple mutations are fixed at the same time. This competition between lineages is the effect of clonal interference [22].

The Gerrish-Lenski (1998) [15] model is the first mathematical model attempting to capture the phenomenon of clonal interference in an asexually reproducing population where genetic material cannot be exchanged or combined. It has been shown to be a robust estimation of this process [23]. In effect, the model aims to generalize the simple example we demonstrated in the section earlier to include additional mutations arising after the first beneficial mutation occurs. For a stable population of size N and mutation rate μ, the model assumes that each organism in the population has a likelihood μ of obtaining a beneficial mutation with some effect s each time it undergoes binary fission, i.e. generation.

Each binary fission an organism undergoes indicates one generation. The measure of relative improvement in fitness is often called the **selection coefficient**, denoted as s. This is commonly defined as $\frac{w_1}{w_0} = 1 + s$ where w_1 and w_0 indicate the fitness of the organism after the mutation and before the mutation respectively. Gerrish-Lenski assumes an exponential distribution of s with the form $\alpha e^{-\alpha s}$, where α is some scaling parameter. Once an organism is mutated, its newly improved genotype begins to increase in frequency. If no additional mutations occur that have higher selective coefficients, the original mutation fixes and becomes the only genotype in the population. However, if another mutation occurs with a higher s, the original mutation may be eliminated before being fixed in the population. This is the effect of clonal interference. Additionally, the new genotype can randomly or stochastically fluctuate in frequency when it is still few in number, via the process of genetic drift. It may also die out due to genetic drift even with a higher fitness genotype. Gerrish-Lenski assumes these two processes are independent with genetic drift being the dominant force when the genotype is low in frequency and clonal interference playing a larger role when the genotype obtains a greater frequency in the population. The probability of a genotype with a mutational effect s surviving genetic drift is often estimated as $\approx 4s$ [15]. With these assumptions and the accompanying model construction, Gerrish-Lenski calculates that the distribution of successfully fixed selection coefficient, labelled here as s_{f}, for a population of size N, mutation rate μ, and exponential distribution scaling parameter α is:

$$p(s_{\mathrm{f}}) = K4se^{-\lambda(s_{\mathrm{f}},\alpha,\mu,N)-\alpha s_{\mathrm{f}}} \tag{2.5}$$

where K is a normalizing constant such that $\int_0^\infty p(s_{\mathrm{f}})\,ds_{\mathrm{f}} = 1$ and

$$\lambda(s_{\mathrm{f}},\alpha,\mu,N) = \frac{\mu}{s_{\mathrm{f}}} N\ln(N)e^{-\alpha s_{\mathrm{f}}}4(s_{\mathrm{f}} + \frac{1}{\alpha}) \tag{2.6}$$

In other words, $p(s_{\mathrm{f}})$ indicates the probability that s_{f} is produced and fixed in the population given the parameters above. Then the expected value of s_{f} could be found by integrating $s_{\mathrm{f}}\,p(s_{\mathrm{f}})$ over all possible values of s_{f} or

$$\mathbb{E}\left(s_{\mathrm{f}}\right) = \int_0^\infty s_{\mathrm{f}}\,p\left(s_{\mathrm{f}}\right)ds_{\mathrm{f}} \tag{2.7}$$

The Gerrish-Lenski model estimates the expected substitution rate, or number of mutations, per generation for the population as:

$$\mathbb{E}\left(\text{substitution rate}\right) = \mu N\alpha \int_0^\infty 4s_{\mathrm{f}}\,e^{-\lambda(s_{\mathrm{f}},\alpha,\mu,N)-\alpha s_{\mathrm{f}}}\,ds_{\mathrm{f}} \tag{2.8}$$

The inverse of this gives the expected number of generations for a mutation to be fixed. Let $\mathbb{E}\left(t_{\text{fix}}\right)$ be this value, then we have $\mathbb{E}\left(t_{\text{fix}}\right) = 1/\mathbb{E}\left(\text{substitution rate}\right)$. Let W_0 and W_1 be the fitness of the population before and after a mutation is fixed and $\frac{W_1}{W_0} = 1 + s_{\text{f}}$. Then the Gerrish-Lenski model estimates that $d\ln(W) = \frac{\ln(1+\mathbb{E}(s_{\text{f}}))}{\mathbb{E}(t_{\text{fix}})}$ and $W(t) = (1 + \mathbb{E}(s_{\text{f}}))^{t/\mathbb{E}(t_{\text{fix}})}$ where $W(t)$ is the mean fitness of the population as a function of time (generation). To sample selection coefficient values from the distribution in equation 2.5, we need to numerically solve the equation, which may be computationally inefficient for some applications. Thus, there is potential utility to provide a simpler, yet quantitatively similar, probabilistic distribution. We present a further simplification later in Chapter 5 that remedies this situation.

2.3.1 Adaptation to LTEE

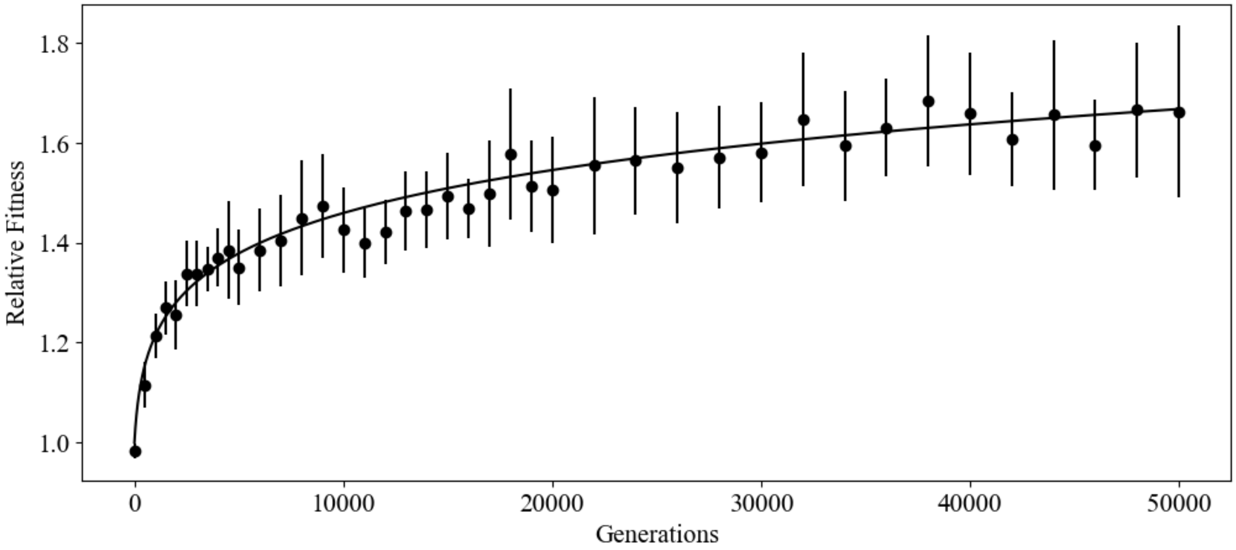


Figure 2.5: Black dots are the mean fitness for the 6 populations that retained their ancestral mutation rate, error bars are 95% confidence limits. The curve is the predicted trajectory equation 2.9. We reproduce this figure using data and model from Wiser *et al.* 2013 [3].

Wiser *et al.* (2013) [3] adapts the Gerrish-Lenski model to predict the LTEE's fitness trajectory over time. The authors assume a diminishing improvement from additional mutations by changing the beneficial mutation distribution as mutations become fixed in the population. The parameter α in the exponential distribution of beneficial mutation is updated as

$\alpha_{n+1} = \alpha_n(1+g*\mathbb{E}\left(s_{\mathrm{f},\,n+1}\right))$, where n is the number of mutations fixed in the population and g is the diminishing constant to be fitted by the experimental data. $\mathbb{E}\left(s_{\mathrm{f},\,n+1}\right)$ indicates the expected value for the $n+1^{\text{th}}$ selection coefficient s_{f} in the probability distribution captured in equation 2.5. Wiser *et al.* (2013) [3] derived the following equation to predict the fitness trajectory of the LTEE's populations:

$$W(t) \approx \left(2g\,\mathbb{E}\left(s_{\mathrm{f},\,1}\right)e^{g\,\mathbb{E}\left(s_{\mathrm{f},\,1}\right)}\frac{t}{\mathbb{E}\left(t_{\mathrm{fix},\,1}\right)}+1\right)^{1/2g} \tag{2.9}$$

where W is the mean fitness, $\mathbb{E}\left(s_{\mathrm{f},\,1}\right)$ is the expected effect of the first beneficial mutation fixed, and $\mathbb{E}\left(t_{\mathrm{fix},\,1}\right)$ is the expected time, in binary fission generations, it takes for $\mathbb{E}\left(s_{\mathrm{f},\,1}\right)$ to be fixed in the population of size N [3]. As we have seen in the the previous section, both $\mathbb{E}\left(s_{\mathrm{f},\,1}\right)$ and $\mathbb{E}\left(t_{\mathrm{fix},\,1}\right)$ depends on α, N, and mutation rate μ. For the LTEE, N is approximated to be 3.3×10^7. The authors estimate $\mathbb{E}(s_{\mathrm{f},\,1})$ and $\mathbb{E}\left(t_{\mathrm{fix},\,1}\right)$ to be ≈ 0.1 and ≈ 300 respectively from the LTEE data [3]. Figure 2.5 contains the same fitness data shown in figure 2.2 with a trajectory curve predicted equation 2.9.

The adapted Gerrish-Lenski model appears to do well in fitting the LTEE data. However, since it is approximated over the intervals of time it takes for the fixation of the mutations, i.e. the mutations make up the most common genotypes in the population, it appears as a smooth trend line. Additionally, the LTEE data was measured at relatively long intervals – ranging from 500 generations initially then 2,000 generations later on – thus lacking the granularity provided by figure 2.1a. We are interested in the utility of a model at the generational scale to see the trend in fitness. Thus, we turn to the Monte Carlo simulation of the Wright-Fisher model which will be covered in Chapter 4.

CHAPTER 3

The Bak-Sneppen model

At the species level, the Bak-Sneppen model introduced in 1993, is perhaps the simplest mathematical model of evolution that claims to exhibit punctuated equilibrium in biological evolution [5]. This highly cited work models a system of species arranged in one dimension with interacting fitness landscapes over time. At equilibrium, the system displays behaviors the authors argue demonstrate a highly-correlated system. In this chapter, we provide an overview of the Bak-Sneppen model. We introduce the Newman-Phoenix hierarchical data structure for efficient repeated selection of complex data. We then show that we can reproduction the Bak-Sneppen model's results using this methodology and provide further analysis of their results with a dramatically reduced computational cost. We end by contextualizing this model in relation to our work.

3.1 The Bak-Sneppen model summary

The Bak-Sneppen model constructs a geometry-independent, simplified, one-dimensional ecosystem and is claimed to simulate the interdependent evolutionary paths of its species. The model is derived through the lens of **self-organized criticality** which observes scale-invariance characteristics in a system which determines the critical threshold [5]. This concept was introduced by Bak *et al.* (1987) [24] and has been applied to different dynamical systems including the forest fire and the sandpile models [25]. Since its publication, the Bak-Sneppen model has been cited close to 2,000 times. This model of a biological system has demonstrated that complicated system behaviors can emerge even in an oversimplified model, including the appearance of punctuated equilibrium in species evolution. However, this intentional simplicity also greatly limits the ability of the model to be extensively applied in biology [26]. As a result, it has been predominantly explored in the physical and mathematical sphere — some examples are Grassberger (1995) [27], Meester and Znamenski (2004) [28], and Fontes *et al.* (2020) [29]. Another source for a thorough treatment of this model and its extension can be found in Sneppen and Zocchi (2005) [30].

The Bak-Sneppen model creates and initializes an ecosystem with N species. Each species carries a fitness-related quantity, labeled as B_i, that is drawn from a uniform distribution [0,1). The authors introduced this representation as "barrier heights," defined as the measurement of the amount of genetic code needed to be changed to move from one local fitness maxima to a higher local fitness maxima. These barrier heights and fitness are positively correlated, as species with high fitness also have high barrier values. The N species are arranged in a sequential circular array with "periodic boundary" conditions - the first barrier, B_1 and the last barrier, B_N are nearest neighbors. The subscript i indicates the position of the species in the array. The system is updated at each discrete time step t, an arbitrary counter to keep track of the number of cycles in a simulation. At each time step, the lowest barrier is updated, through genetic mutation or species replacement, with a new barrier value drawn from uniform [0,1). The two neighboring species are also impacted, not through mutation but by the changes in fitness landscape interactions with the mutated species, and are updated from the same distribution. The simulation will run until a predetermined number of time steps are reached. Figure 3.1 shows a schematic of one ecosystem with eight species and the accompanying updating process.

Initially, the positions of the species selected at each time step appear uncorrelated but as the average barrier height of the ecosystem increases, near neighbors of mutating species are more likely to be selected at the next time step to mutate, exhibiting self-organized behavior [5]. Figure 3.2a is a log-log plot that shows the distribution of distances between consecutive updates [5]. For example, in some time step t, if B_i is selected for update and in the next time step $t + 1$, B_j is selected, the distance will be the shortest difference between their locations in the array ($|j - i|$). Initially, selection of subsequent updates are randomly

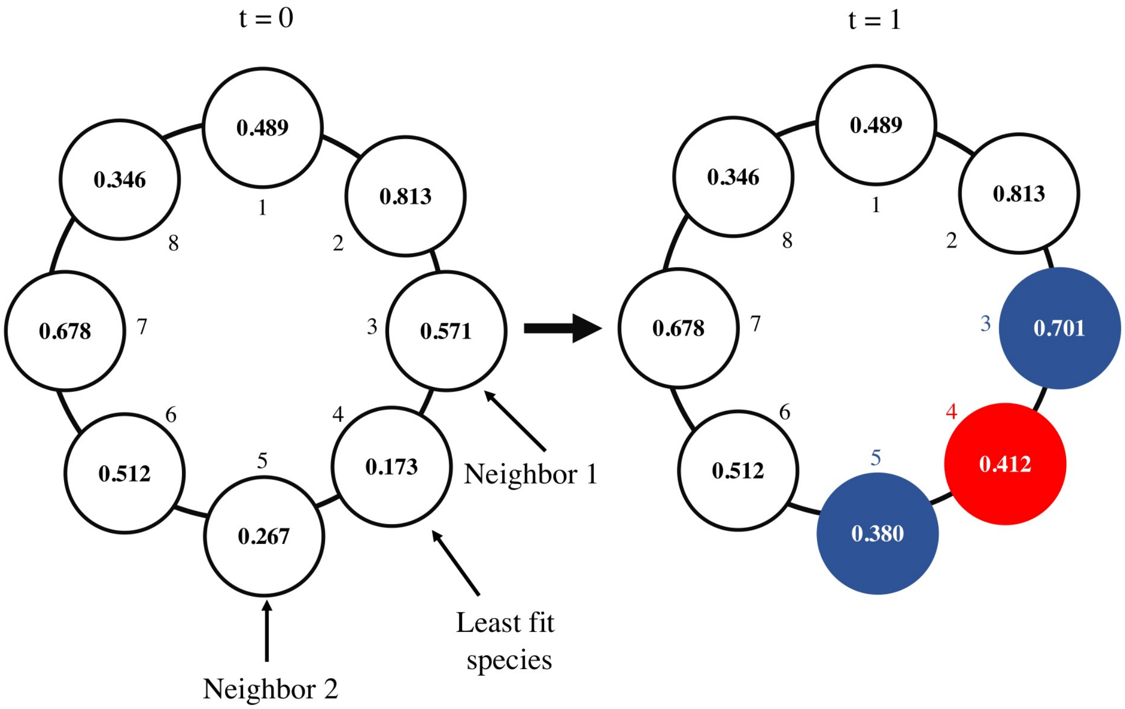
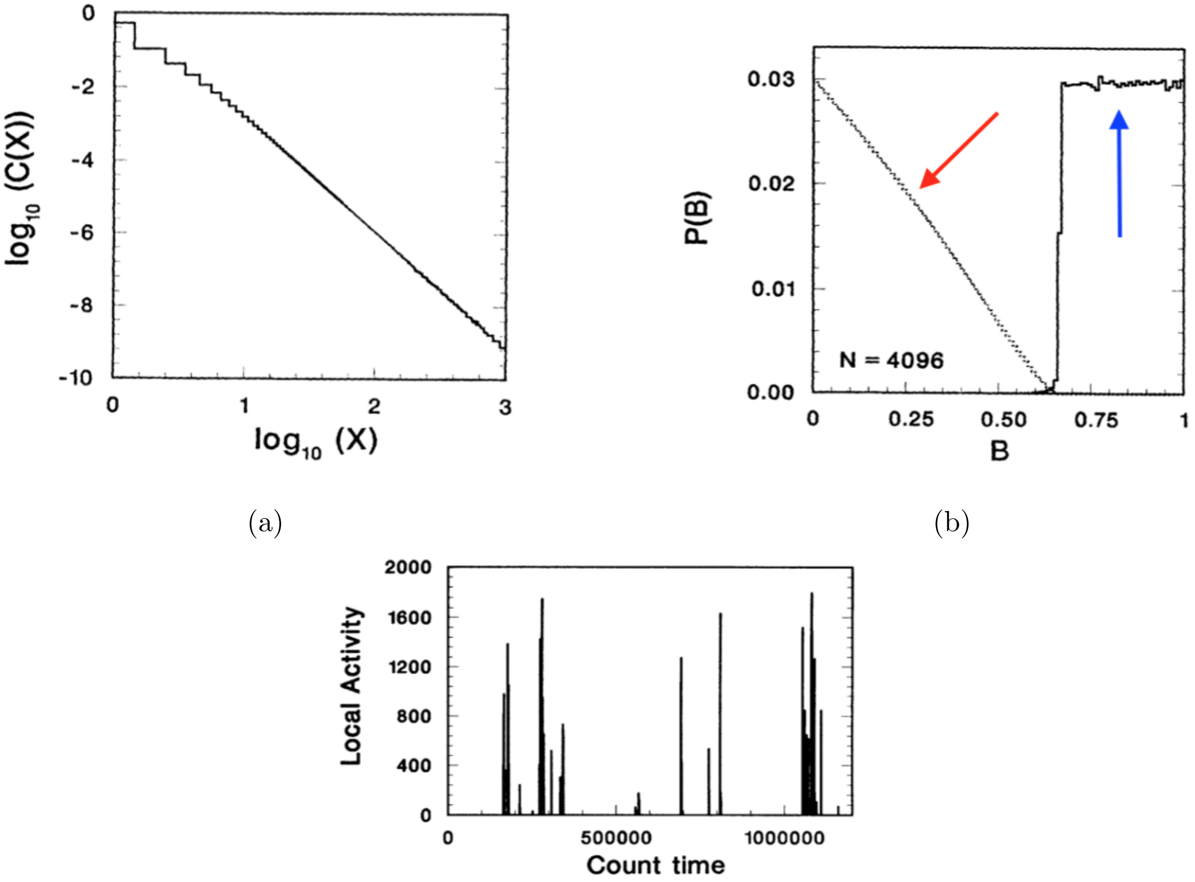


Figure 3.1: This ecosystem has eight independent species (N = 8). Time step is indicated by t. At $t = 0$, B_4 contains the lowest barrier value, thus it is updated. Its neighbors also experience a change in barrier values. The new barrier values for these three species are drawn uniformly over the range between 0 and 1.

determined. However, as time progresses, it is more likely that nearby neighbors are selected for update in the next time step. The distribution of these distances becomes stationary, i.e. they become statistically in equilibrium. The log-log plot exhibits a power law in the distance distribution. Figure 3.2b shows two distributions. The first, highlighted by the blue arrow, is the barrier's distribution at equilibrium. The majority of the barriers have values above $2/3(\approx 0.67)$ and are uniformly distributed. The second curve, highlighted by the red arrow, is the distribution of minimum barriers that are chosen to be mutated at each time step. Figure 3.2c shows that mutation events are sporadic for a species over time when inactive periods are interrupted by fragmented spikes of activity. Local activity indicates the number of selections that occurred in that location vicinity in the array. Count time keeps track of the number of mutations (selections) that have occurred, or the time steps the system undergoes. Although these results are quite intriguing, it is unclear that the power law observed in the Bak Sneppen model occurs in real biological systems.

3.2 Adapting the Newman-Phoenix hierarchical data structure to reproduce the Bak-Sneppen model

In order to reproduce the Bak-Sneppen model results, we have employed hierarchical optimization methodology proposed by Newman and Phoenix [16][17] for a class of problems arising in materials science where there is, in-effect, competition between nearest neighbors under stress. The Newman-Phoenix hierarchical algorithm, which we will refer to as Newman-Phoenix onward, optimizes repeated searches over complex data systems, a feature that is applicable today with data mining and machine learning. For example, decision trees build hierarchical structures to categorize samples based on their features. Their methodology identifies and eliminates redundant mathematical operations thereby reducing N^2 arithmetic operations to $N \log_2 N$, exploiting a hierarchical structure analogous to what is accomplished by the Fast Fourier Transform (FFT) in signal processing [16]. A relevant real life example is competition among fibers of composite materials under stress [16]. This



(a)

(b)

(c)

Figure 3.2: At equilibrium state, (a) the log-log plot of the distribution of distances [C(X), where X is the distance)]between consecutive updates [5]. The straight line indicates a power law distribution of distances where the majority of the distances are small. For any species being selected for update, the likelihood of the nearest neighbors being selected next for update is not random at equilibrium. (b) Distribution of barrier values at equilibrium highlighted by the blue arrow where B is the barrier and P(B) is the frequency of that barrier [5]. The majority of the barrier values are above a certain value (≈ 0.67). The dashed line, highlighted by the red arrow, shows the distribution of the minimum barrier values selected at each time step. (c) Local activity, selections, of 10 consecutive sites in a system over time, measured in units of selections in the system [5]. Collectively, these sites experience selections intermittently with long periods of inactivity. These figures were extracted from Bak and Sneppen (1993) [5].

algorithm is currently the most efficient way to perform localized comparisons and identify the smallest elements. It enhances simulation speed by a factor of $N/\log N$, offering dramatically improved performance by a factor of 10^5 for $N \geq 10^6$ or greater depending on N, compared to using algorithms to locate the minimum value in an unsorted array at each time step which has $\mathcal{O}(N)$ time complexity. Additionally, computational speed since the publication of the Bak-Sneppen model has increased by a factor of 1,000. The combination of methodological and hardware advancement allows us to not only reproduce the Bak-Sneppen model efficiently but also gives us the ability to expand the simulation parameters. This hierarchical data structure and the gain in computational performance will allow us to perform a larger number of simulations with a greater number of elements for longer times. The ability to perform a large number of independent simulations tractability with a large number of elements reduces sampling bias.

The most resource intensive operation in the Bak-Sneppen scheme is locating the minimum barrier value, often requiring traversing the whole array. The Newman-Phoenix hierarchical data structure reduces the need to traverse the whole array at each time step by maintaining a perfect binary tree structure with the root node holding the minimum value (see Figure 3.3). Although a one-time initialization requires $\mathcal{O}(N)$, each time step takes $\mathcal{O}(\log_2 N)$ to update the tree and locate the minimum value. A direct outcome of using this computational methodology is that the Monte Carlo simulations we will perform in Chapter 5 allow us to obtain our model results in a time scale measured in minutes instead of decades with a sample size hundreds of times larger than that in the LTEE.

Once the species in the array have been assigned their barrier values at the start of the simulation, the Newman-Phoenix hierarchical data structure is deployed to store the barrier values. The N species or “nodes” of the ecosystem are organized to form a binary tree with $\log_2 N$ layers, and if the number of species is not a power of two, we can “pad” the array with artificial members or nodes that have no influence or introduce any error with values outside the accepted range to make its length a power of 2. The base layer holds the indices of each species in array. For each successive layer above that, the nodes hold the indices of the smaller barrier values of their immediate children nodes. Hence, the root or lowest node holds the smallest value of the B_i array (see Figure 3.3). Thus, the initial tree takes $\sum_{i=1}^{\log_2(N)} \frac{N}{2^i} = N - 1$ operations or $\mathcal{O}(N)$ where N is the number of species or size of the array. However, we only have to do this once and locating the minimum barrier value is one operation, or $\mathcal{O}(1)$, since it is simply the root node. Changing the minimum value and updating the binary tree take $\mathcal{O}(\log_2 N)$ as we only have to make one comparison for each layer with every new single value update in the array. We can see that without this structure, each update will take $\mathcal{O}(N)$ as we need to visit each index to find the new minimum barrier value. A schematic of this structure is displayed in figure 3.3 with $N = 8$ and the minimum value is 0.163 at index 3. A more extensive description of the algorithm's procedure can be found in Appendix C.3. Importantly, the algorithmic developments we have made will provide an improvement in computational speed that will allow us to simulate much more realistic biological systems with large N and for a longer time.

Using the Newman-Phoenix hierarchical data structure, we have reproduced the Bak-Sneppen model's results and our work seems to be in agreement. The figures 3.2a, 3.2b, and 3.2c extracted from their paper are mirrored by our simulation, see figures 3.4a, 3.4b, and 3.4c. We set $N = 1024$ and a range of time steps from 10^6 to 10^9 for our simulations. The graphs from the Bak-Sneppen's paper have inconsistent parameters with some produced using $N = 4096$ and another having $N = 512$. However, our results seem to be in agreement. We are unsure how long their simulations ran for to reproduce figure 3.2a. Thus, we ran our simulations over $t > 10^9$ and included the distance distributions at different simulation points in figure 3.4a. We can see over time, the simulation results are reaching the distance distribution documented by the original paper of the Bak-Sneppen model [5]. The preceding validates not only our ability to replicate the Bak-Sneppen results but the potential to do so orders-of-magnitude more rapidly, thereby permitting us to pursue much more expansive models that are in close conformity with biological data.

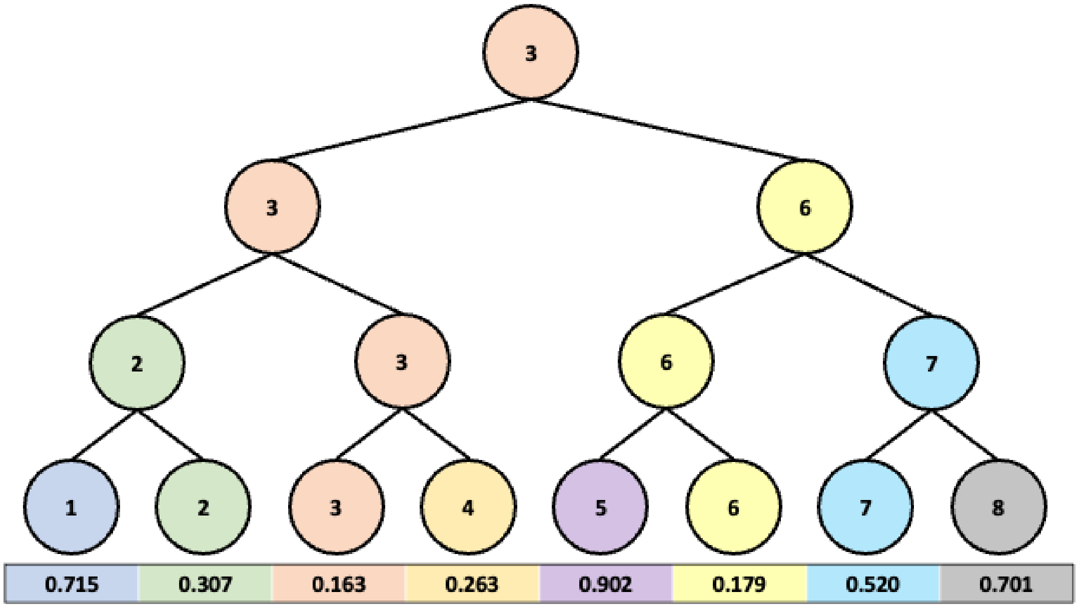
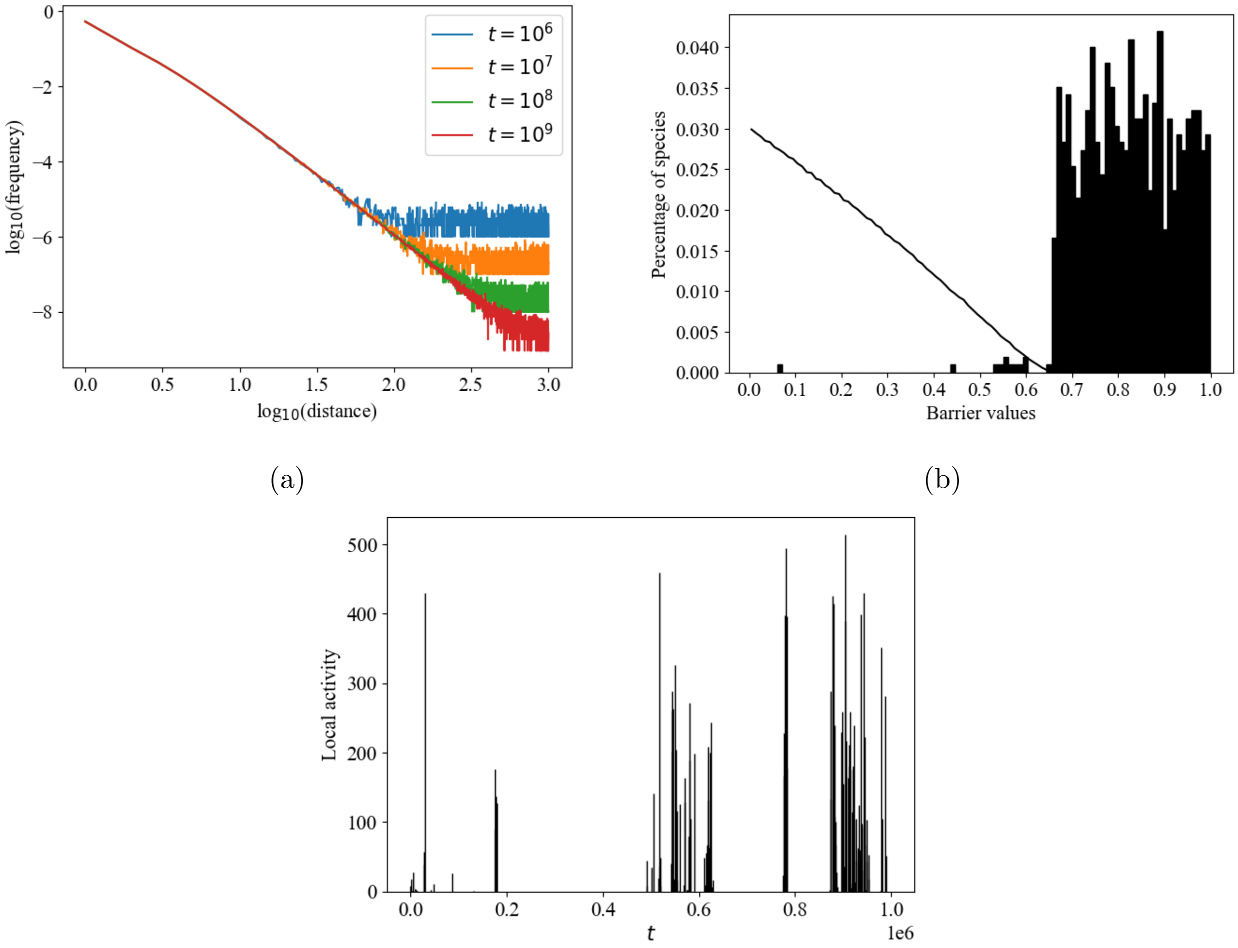


Figure 3.3: Hierarchical data structure based on Newman and Phoenix (2001, 2009) [16][17] with $N = 8$. A binary tree is constructed to store information of an array holding eight elements with varying values. The first layer compares adjacent elements and hold the index of the smaller one. The next layer compares the elements of the adjacent indices of the layer below it and hold the index of the smaller one. This occurs until the root node holds the index of the smallest element.



(a)

(b)

(c)

Figure 3.4: We reproduce these graphs from Bak and Sneppen (1993) [5] showing the behavior of the system as it approaches equilibrium state using results from our simulation. (a) Distribution of distances between consecutive updates. Note the system runs for t time steps for each curve. (b) Distribution of barrier values at $t = 10^6$. The black line indicates the distribution of the minimum barriers selected. (c) Local activity, or updates, of 10 consecutive sites (242, ..., 251) in a system over time, which is the total number of time steps. The system runs for $t = 10^6$.

3.3 Analysis

Since the time that the Bak-Sneppen model was published, it has received a lot of attention in the literature but there seems to be few rigorous analytical results for its statistical observations [29][27]. The simplicity of the model has allowed great flexibility in adaptation and invited substantial further mathematical analyses. Of a particular interest is the distribution of barrier values at equilibrium. There appears to be a critical threshold, above which the majority of the barrier values lie uniformly. Empirical evidence suggests critically at $p_c = 0.667$, i.e. 2/3 for the Bak-Sneppen model, where the marginal distribution is uniform on $(p_c, 1)$ as $N \longrightarrow \infty$ [28][5]. In this analysis, we want to dissect this observation further. Although there is currently no analytical solution to this result, we proceed in analyzing simpler updating schemes to provide insights to this somewhat surprising result. We look at two updating schemes of a one-dimensional array, initialized with elements drawn from a uniform distribution [0,1): a) minimum value update at each time step b) minimum value and one randomly chosen index are updated at each time step. For both cases, the selected elements are updated with values from a uniform distribution [0,1).

For the first case where only the minimum value is selected for update at each time step, we can see that over time this results in a monotonic increase in the critical value of the array, where all but one element is greater than this value. For each time step, the minimum value is selected and replaced with a value from a uniform distribution [0,1) . If the new value is less than the previous value, then this index is again selected for update in the next time step. This index is continuously selected until the new updated value is greater than the initial minimum value identified at some time step t. Once this is achieved, a new critical value is established. Thus we can see over time, the majority — all but one — of the elements in the array are greater than this critical value.

For the second case, the minimum value and a randomly chosen index are updated with values drawn from the uniform distribution [0,1). This is an extension of the first case with the random index update. If we simply perform a random index update at each time step, and not the minimum value update, we can see that the distribution of the array stays uniform between 0 and 1. Thus, combining these two updating schemes at each time step intuitively creates a situation where the distribution becomes increasingly narrower and approaches 1 by the minimum value update, and is pulled back to the uniform [0,1) distribution by the random index update. The distribution is stationary at equilibrium. Since at each time-step the minimum barrier value is eliminated, the system evolves to have a step function distribution with the majority of the values being uniformly distributed above a certain critical threshold B_c. Thus, the minimum value $B_{\min}$ selected is always less than B_c and values above B_c are only selected through the random index. For this distribution to stay stationary, the number of elements below B_c needs to stay stable. Thus, we can solve for the critical threshold through the following probability distribution for the number of additional/loss elements below B_c at each time step. Let X be the random variable that denotes the number of elements below B_c gained at each time step, we have:

$$P(X) = \begin{cases} (1 - B_c)^2, & \text{for } X = -1 \text{, both new } B_i\text{'s are greater than } B_c \\ 2B_c(1 - B_c), & \text{for } X = 0 \text{, one new } B_i \text{ is above and one below } B_c \\ B_c^2, & \text{for } X = 1 \text{, both new } B_i\text{'s are less than } B_c \end{cases} \tag{3.1}$$

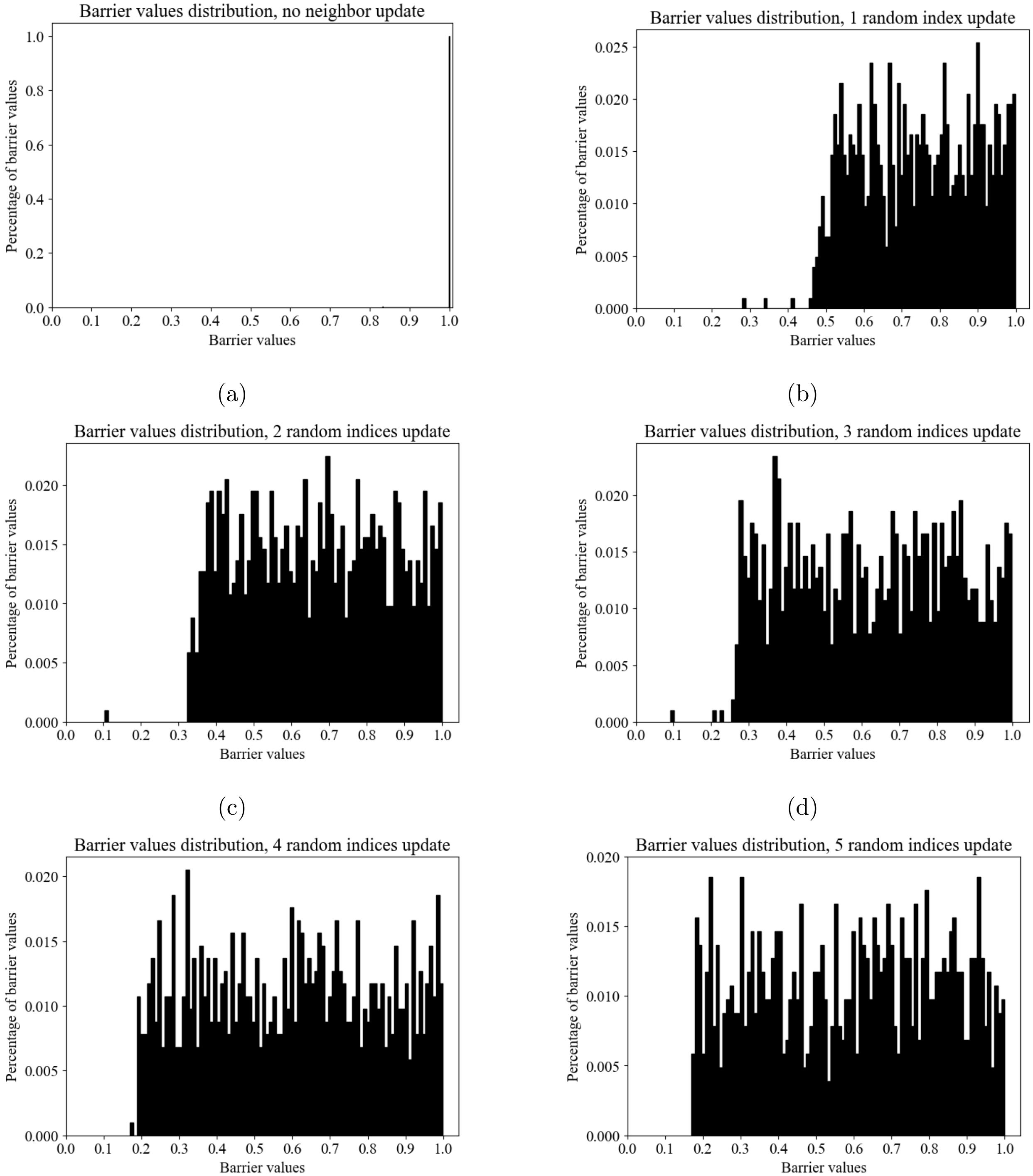
Solving the set of equations above gives a B_c value of 0.5, see Sneppen and Zocchi (2005) [30]. This is in contrast with the Bak-Sneppen model which has the critical barrier value at 2/3. As a reminder, the Bak-Sneppen updates the minimum barrier value and two nearest, i.e. non-random, neighbors at each time step. A more thorough mathematical treatment of this problem can be found in Flyvbjerg *et al.* (1993) [31]. Simulation has shown that at equilibrium, the distribution is uniform above 0.5. We can extend this to cases there multiple random indices are selected for update at each time step. We can see that these additional random selections will pull the minimum value update even further, resulting in a barrier threshold B_c that is increasingly approaching 0 as the number of random selections at each time step increases. The critical threshold can be found using the following equation: $1/(N + 1)$, where N is the number of random indices. As $N \longrightarrow \infty$, the critical threshold approaches zero. Distributions at equilibrium, or stationary distributions, where 1, 2, 3, 4, and 5 random indices are updated in addition to the minimum value update are included in figure 3.5. As we can see, the critical threshold decreases as the number of random indices increases.

We can extrapolate from the analysis of the two cases above to analyze the behavior of the system in the Bak-Sneppen model. The distinction – and complication – here is the non-random neighbor update. The neighbors are not chosen independent of the minimum value update thus the system behaves differently than the cases we have highlighted. Initially, the updated values are not correlated as the minimum values are small. However, as the number of updates completed increases and the minimum value selected increases, the neighbors are more likely to be updated with values smaller than the minimum values, causing them to be selected at the next time step to be updated. We can again see how the distribution is pushed upward toward 1 by the minimum value update but pulled downward towards 0 because of the neighbor updates. The impact of the coupled neighbor updates however, causes this balancing to be less pronounced since the values of the minimum value and the values of the neighbors are correlated, as opposed to case two that is highlighted earlier. We can observe this behavior played out in the simulations provided in figure 3.6 for the one nearest neighbor update and the two nearest neighbor updates. Their distributions at equilibrium are still uniform from a certain B_c value to 1. The B_c values are much smaller than in the cases where randomly selected neighbors are updated.

3.4 Summary

In addition to the mathematical analysis, other works have adapted the Bak-Sneppen model to analyze adjacent systems [32] [29]. For example, Guiol *et al.* (2010) proposed a stochastic random walk model of evolution showing similar long term behavior as Bak-Sneppen but with analytical results [32]. Unlike Bak-Sneppen model, however, the number of species in Guiol *et al.* (2010) is not static and the species do not interact, i.e. neighboring species are not affected as a species is introduced or become extinct [32].

Works like the Guiol *et al.* (2010)'s model have extended beyond the original work by



(a)

(b)

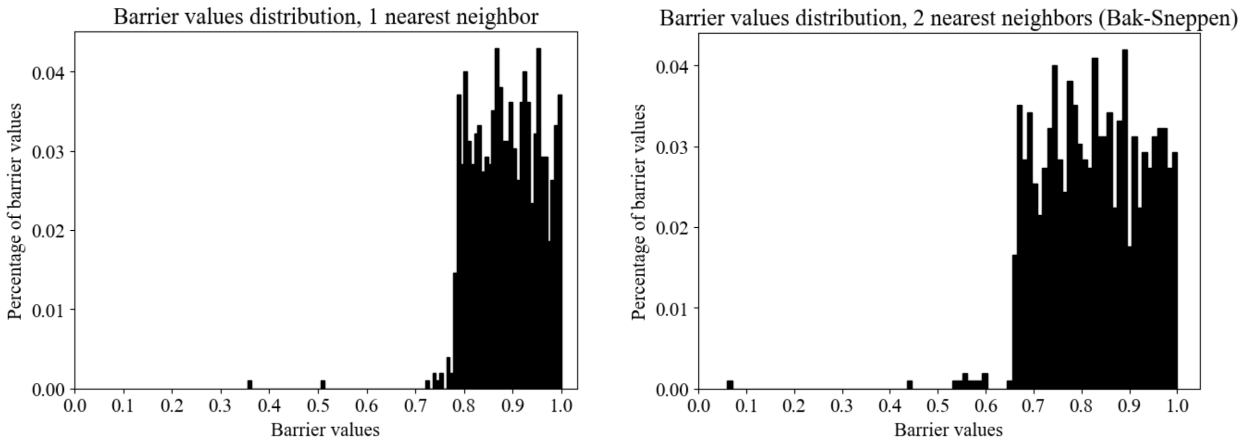
(c)

(d)

(e)

(f)

Figure 3.5: Distributions of barrier values at $t = 10^6$, $N = 1024$. For each figure, the minimum value and (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, (f) 5 random indices are selected for update for each time step.



(a) (b)

Figure 3.6: For each figure, the minimum value and 1 or two nearest neighbors are chosen for update. Distributions of barrier values at $t = 10^6$, $N = 1024$ for (a) one nearest neighbor update and (b) two nearest neighbors update (Bak-Sneppen model).

Bak and Sneppen and provided greater insights into the interesting results obtained by the original model. However, the extent they can provide insights on punctuated equilibrium behavior at the species scale remains an open question as argued in Gould 2009 [26]. Thus, we are interested in looking at constrained biological systems at population and organismic scale in order to explore biologically relevant issues including the question of evolutionary rate. Now that we have adapted the Newman-Phoenix hierarchical data structure and provided further analysis of the Bak-Sneppen model, we move forward in building computational models grounded on the works of the Long-Term Evolutionary Experiment [6].

CHAPTER 4

The Wright-Fisher model

In this chapter, we present an adaptation of the Wright-Fisher model using a Monte Carlo approach to study the fitness evolutionary path of asexually reproducing organisms at the unicellular level. We have not seen this model utilized in the literature to study punctuated equilibrium. This is a biologically simplest and least complicated situation. Using parameters estimated by Wiser *et al.* (2013) [3] for the Long-Term Evolutionary Experiment (LTEE), our Wright-Fisher simulations are able predict the fitness trajectory of the LTEE's populations well. We observe the clonal interference effects in our results, demonstrating the possibility for simple models of biological systems to exhibit step-like behavior in their evolutionary path at the generation time scale.

4.1 The Wright-Fisher model

The Wright-Fisher model, developed implicitly by Ronald Fisher (1923) [33] and later explicitly by Sewall Wright (1931) [34], attempts to describe the evolution of genotypic frequency in a population due to random chance. This forms the basis of modeling genetic drift in a finite population where there are no presence of beneficial/deleterious mutations thus no selection. In this model, there is no migration into the population. It assumes non-overlapping, discrete generations where the whole population is replaced at each generation by the offspring from the previous one with each parent equally likely to reproduce [35] [7]. This is effectively random sampling with replacement. A schematic of the Wright-Fisher model is shown in figure 4.1. The Wright-Fisher uses a highly simplified probabilistic expression arising from combinatorics to convey an idealized expectation for the relative frequency distribution for site selections according to the stochastic processes at hand. Starting with some arbitrary distribution of genotypes in a population, for each generation t, every organism in the population is replaced by a progeny of the previous generation. For example, at generation $t = 0$, there are two genotypes a and b with equal likelihood of reproduction, each making up 50% of the total population of size N. Since the progeny are equally likely to be descendant of any parent organism, in generation t = 1 the organisms of the new population are equally likely to carry either genotype a or b. But due to stochasticity, the frequency of genotypes a and b may not stay 50% in $t = 1$. In fact the probability distribution of genotype a at $t = 1$ follows the binomial distribution with the **probability mass function** (for discrete distributions) or PMF:

$$Pr(X = k) = \binom{N}{k} {p_a}^k {p_b}^{N-k} \tag{4.1}$$

where p_a and p_b are frequencies of a and b respectively in the parents' generation and k is the number of offspring having genotype a. This PMF gives us the probability of drawing k offspring with genotype a. For example, the probability of having 50% of the population carrying genotype a at $t = 1$ is the same as having $k = N/2$, or:

$$Pr(X = N/2) = \binom{N}{N/2} 0.5^{N/2} 0.5^{N/2} = \binom{N}{N/2} 0.5^{N} \tag{4.2}$$

For very large N, the binomial distribution is approximately normal (Feller 1968 [36]) and:

$$Pr(X = k) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{4.3}$$

where μ is the mean and σ is the standard deviation. The mean for the binomial distribution in this case is simply Np_a and standard deviation is $\sqrt{Np_a(1 - p_a)}$, Thus:

$$Pr(X = N/2) \approx \frac{1}{\sqrt{Np_a(1 - p_a)}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-Np_a}{\sqrt{Np_a(1-p_a)}})^2} \approx \frac{1}{\sqrt{0.5N\pi}} \tag{4.4}$$

Since $p_a = p_b$ in our example, genotype b has the same PMF as genotype a. Note that the PMF's of a and b change from one generation to another, depending on the frequencies of a and b in the previous generation.

The standard Wright-Fisher model can be extended by adding weights to the genotypes according to their fitness advantage, allowing us to introduce mutation and selection. This

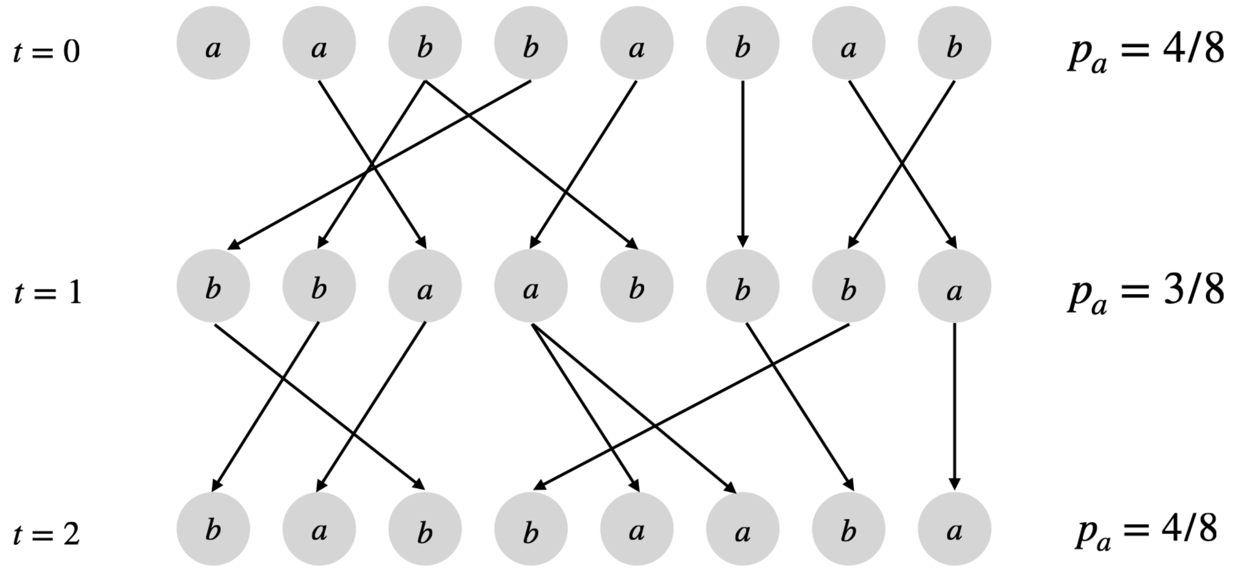


Figure 4.1: A schematic of the Wright-Fisher model for genetic drift. In this illustration, there are eight organisms in the population. Each organism has equal likelihood of reproducing in the next generation, indicating they have the same fitness. At generation $t = 0$, half of the organisms carries genotype a and the other half carries genotype b. At $t = 1$, the frequency of a drops from 50% to 37.5% due to random chance. This stochastic fluctuation in genotypic frequency is genetic drift [37]. This visual is provided to illustrate the possible pathways that is explained mathematically in the text.

forms a simplified basic mathematical model of asexually reproducing organisms. Thus, we can adapt this model to simulate the fitness evolutionary path of the populations in the LTEE. Wiser *et al.* (2013) has also included a simulation of Wright-Fisher to demonstrate the robustness of their adapted model [3]. This is expected as the original Gerrish-Lenski model, which Wiser *et al.* (2013) adapted, has been shown to approximate the Wright-Fisher simulations well [23]. We have outlined both of these models in Chapter 2. We have two objectives in utilizing Wright-Fisher for our investigation. First, we want to investigate whether a finer resolution in single generation scale exhibit distinct jumps in fitness, since the Gerrish-Lenski model and in effect, the Wiser *et al.* (2013) [3]'s adaptation, are more coarse-grained. Second, we want to analyze the possibility of observing the punctuated equilibrium behavior in an asexually reproducing population under different conditions using the Wright-Fisher model.

4.2 The Wright-Fisher model with mutation and natural selection

As seen in the description above, in the standard Wright-Fisher model, the genotypic frequency progresses stochastically over time from one generation to the next, forming the foundation of studying genetic drift. To model the LTEE experiment, we need to introduce mutations that can affect the likelihood of an organism to reproduce, or its fitness. This allows us to model mutation and natural selection in an asexually reproducing population using the Wright-Fisher approach. We use a Monte Carlo model, which is outlined as follows, to simulate the Wright-Fisher model.

4.2.1 Monte Carlo algorithm outline

In this section, we are outlining the algorithmic construction of the model. We start with a fixed population of size N, which stays constant from one generation to the next in the simulation. Each organism i in the population carries fitness $w_{i,t}$ at generation t. At $t = 0$, all organisms in the population is initialized with the same fitness of 1, or $w_{i,0} = 1$ for all i from 1 to N. In this dissertation, we retain the definition of fitness as measured in the LTEE. Fitness is estimated as relative binary fission rate between one population to another, usually the ancestral population. Since we are tracking the fitness evolution of individuals in the Wright-Fisher model, we define individual fitness $w_{i,t}$ as the relative binary fission rate between the i^{th} organism at generation t and its ancestor at $t = 0$. When the ancestral organism had undergone 1 doubling event, organism i has undergone $w_{i,t}$ doubling events with t being the current generation. The probability of an organism in the next generation having organism i as a parent, therefore obtaining its genotype and fitness, is $w_{i,t}/(\overline{W}(t)N)$ where $\overline{W}(t)$ is the mean fitness of the current generation t. Thus, an organism with relatively larger fitness $w_{i,t}$ within a population is more likely to reproduce at generation t. The population mean fitness $\overline{W}(t)N$ is calculated as $\sum_{i=1}^{N} w_{i,t}/N$. Since each fitness value is associated with a particular genotype, the genotype's impact on an organism's reproductive rate is measured using fitness. Thus, in our simulation, we only keep track of the number of organisms with each particular fitness value to approximate its genotypic frequency. We are using the multinomial distribution to reduce computational time as implemented by Park and Krug (2007) [23]. This will allow us to keep track of the number of organisms carrying a particular fitness value without having to store the values for each organism. For example, we only need to store one value of n_a as the number of organisms carrying genotype a without having to store n_a organisms' fitness values separately. Please refer to Park and Krug (2017)'s supplement for a full treatment of drawing from a multinomial distribution [23]. Feller (1968) also serves as a classic textbook on this subject [36].

Once the population has been repopulated with offspring from the previous generation, it then undergoes mutation. Let the mutation rate μ be the probability of an organism obtaining a beneficial mutation during each generation. Instead of randomly selecting the organisms to be mutated with probability μ, we draw a random variable Q from a binominal distribution with success μ as the number of organisms to be mutated. This would reduce computational resources. We then randomly select Q organisms in the population to undergo mutation. When an organism mutates, its new fitness is determined by a selection coefficient, $s_{i,t}$, and calculated as $w_{i,t}(1 + s_{i,t})$. The selection coefficient $s_{i,t}$ is drawn as a random variable from the exponential distribution $\alpha e^{-\alpha s}$. This is a standard distribution used to model beneficial selection coefficient in population genetics [23]. We continue to repopulate and mutate the organisms until a set simulation time is met. Please refer to figure 4.2 which we provide as a visualization of this algorithm.

4.2.2 Results

We have simulated the Wright-Fisher model using parameters from Park and Krug (2007) [23] to validate our code setting $\alpha = 50$, $\mu = 10^{-6}$, and the number of generations as 20,000. Park and Krug (2007) simulates their algorithm for population size N ranging from 10^3 to 10^9 [23]. The smallest population was run for 10^8 times or iterations and the largest was run for 32,000 iterations. Since we simply want to validate our code, we ran 1,000 iterations for population sizes $N = 10^3, 10^4, 10^5$, 100 iterations for population sizes $N = 10^6, 10^7, 10^8$, and 10 iterations for population size $N = 10^9$. The results of the validation check align well with the original model, allowing us to move forward and add the additional parameters to model the LTEE, see figure 4.3. The resulting model is outlined in the next section.

In figure 4.3, we show the rate of fitness improvement for each population size. We include our numerical estimates of the Gerrish-Lenski mathematical model and Park and Krug (2007) [23]'s estimates. Park and Krug (2007) [23]'s approximations of the Gerrish-Lenski model are lower than ours due to an adjustment they have proposed in dealing with the probability of surviving drift. The authors argue that the independence assumption the Gerrish-Lenski model makes regarding clonal interference and genetic drift is likely to cause the model to overestimate the survival of the beneficial mutations, thus reducing the rate of fitness improvement. Instead of utilizing $4s$, or some similar linear approximation, as the survival probability estimation of a mutation with selection coefficient s as in Gerrish-Lenski, they propose a correction in the form $1 - e^{-2s}$ [23]. Although this allows their estimations to be closer to the Wright-Fisher simulation results, we find that by adjusting the survival probability estimation to be $2s$ (another commonly used approximation), we can

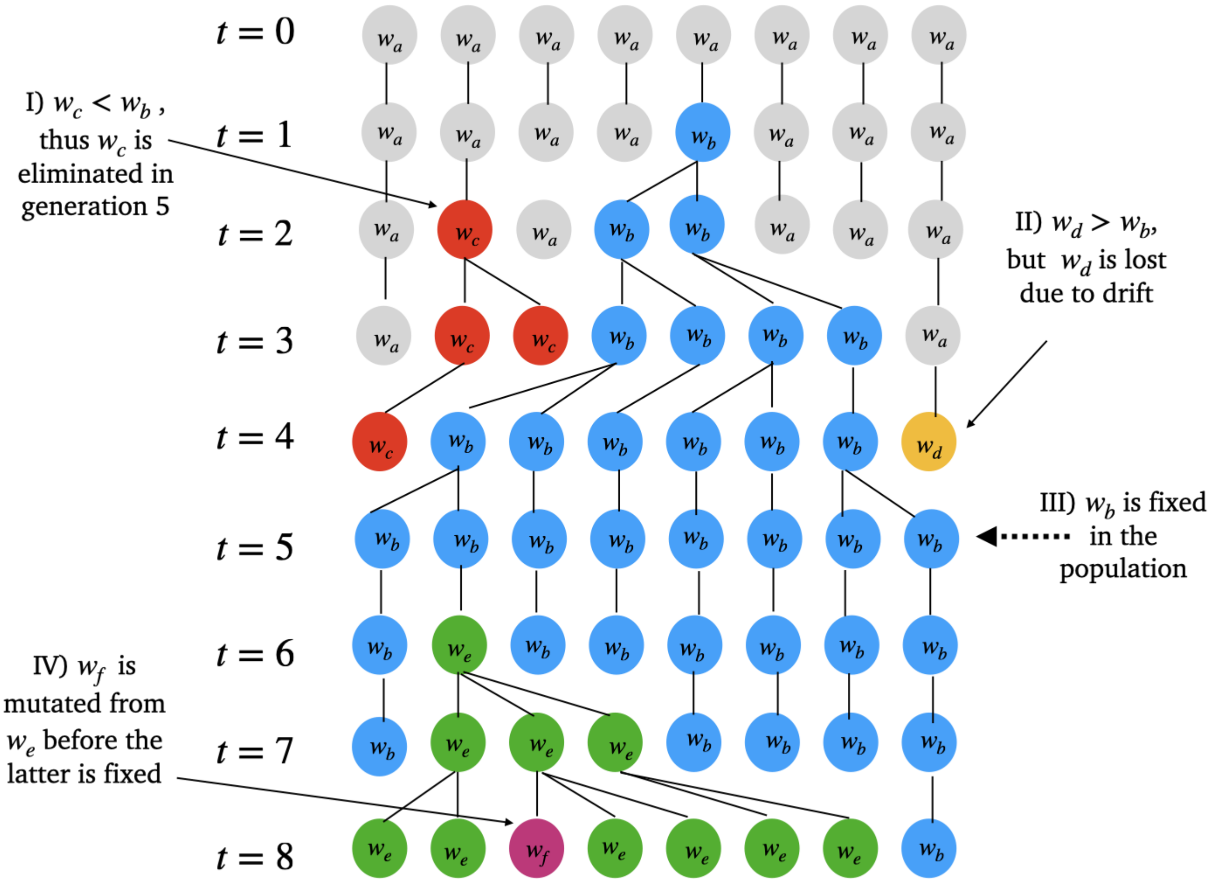

$t = 0$
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 5$
$t = 6$
$t = 7$
$t = 8$
w_a
w_b
w_c
w_d
w_e
w_f
I) $w_c < w_b$, thus w_c is eliminated in generation 5
II) $w_d > w_b$, but w_d is lost due to drift
III) w_b is fixed in the population
IV) w_f is mutated from w_e before the latter is fixed

Figure 4.2: A schematic of the Wright-Fisher model with mutation and selection. Three genotypes with fitness w_b, w_c, and w_d, are direct descendants of the original genotype w_a. Assume $w_a < w_c < w_b < w_d$, w_a and w_c genotypes are out-competed by the other two genotypes [event I)]. Although w_d has a higher fitness than w_b, it is lost due to random chance [event II)]. Genotype w_b is fixed in the population in generation 5, or become the single genotype [event III)]. A new genotype w_e emerges from w_b but before it is fixed in the population, it is mutated once again to form genotype w_f [event IV)]. If genotype w_f is fixed in the population in later generations, the mutation that creates genotype w_e will also be fixed (assuming the mutation forming w_f does not interfere with the mutation forming w_e).

get relatively close to their adjusted estimation for the simulated population range, as shown in figure 4.3. Since the updated approximation does not significantly improve the model's estimations, we will continue to use the $4s$ to remain consistent with Gerrish-Lenski's original model for further analysis in this chapter and the rest of the dissertation.

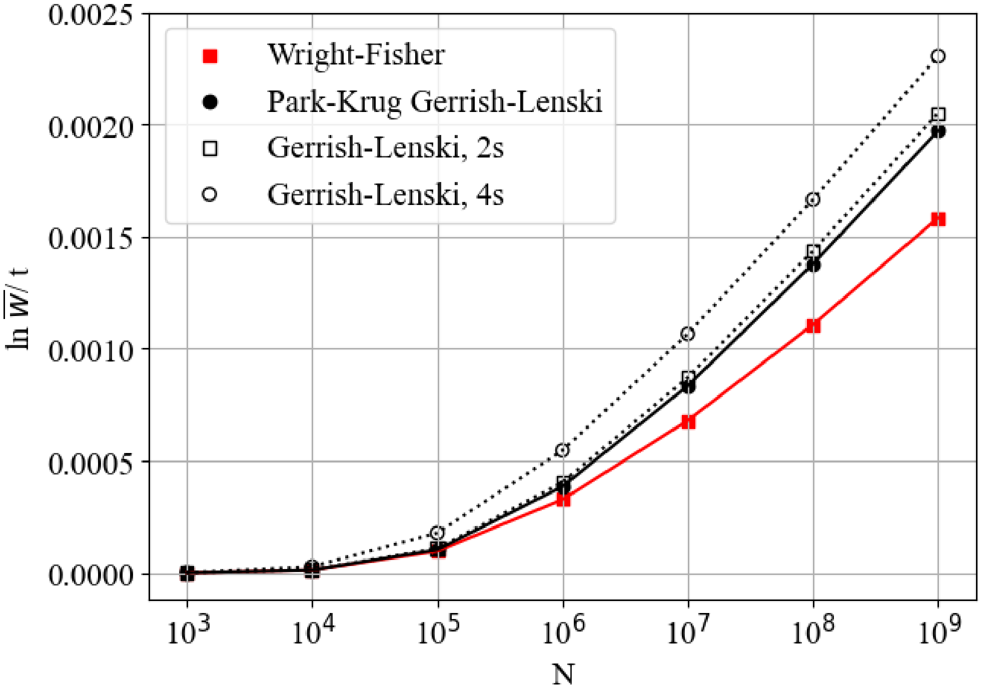


Figure 4.3: The Wright-Fisher model simulated with varying population sizes. The Gerrish-Lenski model using different survival probabilities are also included (Park and Krug's, $4s$, $2s$). This graph has been reproduced using the same parameters as Park and Krug (2007) [23], please refer to their paper for the original figure.

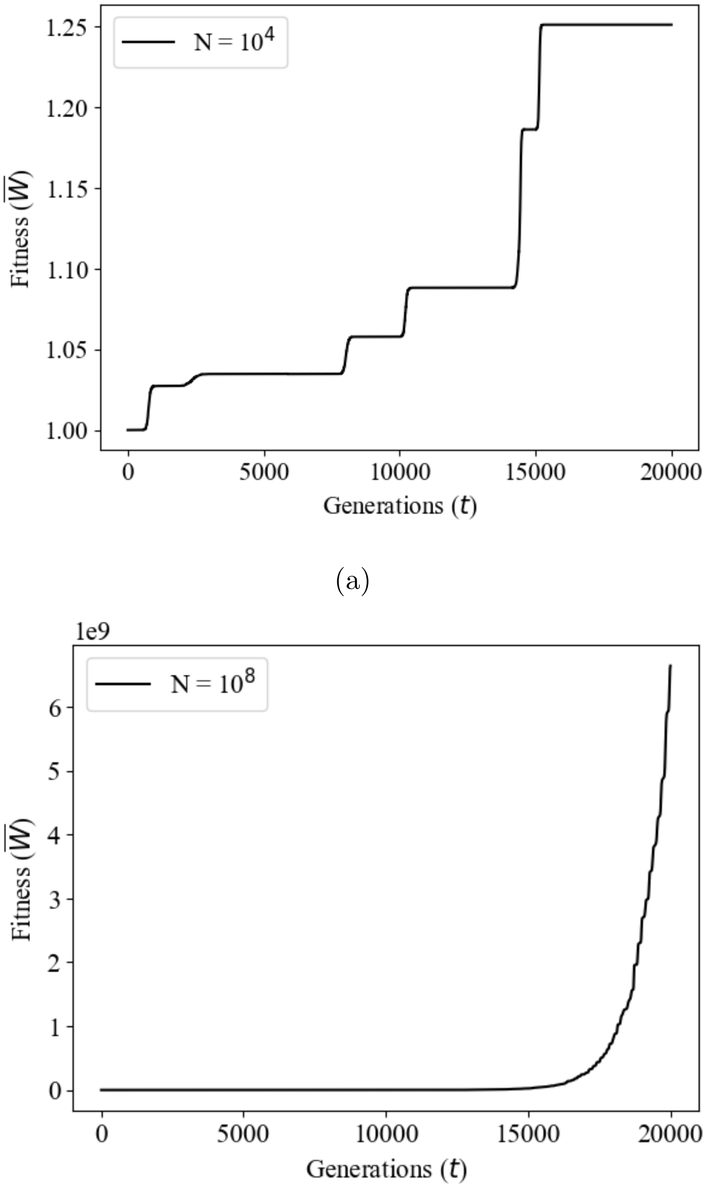
Two examples of our simulations are shown in figure 4.4 with varying population sizes. As expected, fitness increases exponentially with a natural log plot from figure 4.6 showing a linear trajectory for the range of population sizes simulated. Fitness increases much faster with larger population sizes since there are many more mutations per generation that can be selected for, given the same mutation rate and mutational effect distribution. All our simulations exhibit fitness jumps though they are more defined in smaller populations for the number of generations simulated, as seen in figure 4.4a for population size $N = 10^4$. In fact, depending on the population size N, fitness jumps can be observed given the appropriate generational time scale. Although the fitness jumps in figure 4.4b for $N = 10^8$ are less pronounced, zooming into the fitness trajectory from generation 5,000 to 5,500 shows more distinct steps in figure 4.5. This is likely an observation of clonal interference where the beneficial mutation does not measurably affect the population's fitness until it has spread to a significant portion of the population. As shown in figure 4.7, the "jumps" observed in fitness occur in close conjunction with generations in which the dominant genotypes reach their peak frequency, which can be 100%. Fitness stays relatively constant until the dominant genotype reaches a certain frequency to have measurable effect on the population fitness. It is important to note that the issue of smoothness of the curves depends substantially on the number of generations involved as well as the number of organisms in a population. Moreover, the relevance to the fossil record is complicated because of the incompleteness present over long time scales and the distribution geographically over the surface of the Earth.

4.3 The Wright-Fisher model with diminishing-returns epistasis

The Wright-Fisher model simulations in figure 4.3 does not account for **epistasis**, or the interaction between mutations. Khan *et al.* (2011) [38] show that there is a **diminishing-returns epistasis** in the LTEE populations where as fitness increases in a population, the marginal improvement from additional mutations decreases. In this section, we account for this diminishing fitness improvement effect in order to predict the fitness trajectory of the LTEE populations.

4.3.1 Algorithm outline

In order to decrease the fitness improvement rate over time, we can change the distribution of the selection coefficient s, which determines the fitness effects of the mutations. We can reduce the likelihood of obtaining large values of s by increasing the scaling parameter α



(a)

(b)

Figure 4.4: Two Wright-Fisher simulations running for 20,000 generations with $\alpha = 50$, $\mu = 10^{-6}$, and $N = 10^4$ in (a) and $N = 10^8$ in (b). The fitness growth varies greatly as population size increases. The exponential growth in fitness can be seen more effectively in (b) since it's likely to have experienced substantially more mutations during the same number of generations as (a). These simulation results simply demonstrate the behavior of the model for the given parameters, which need to be constrained for the appropriate system.

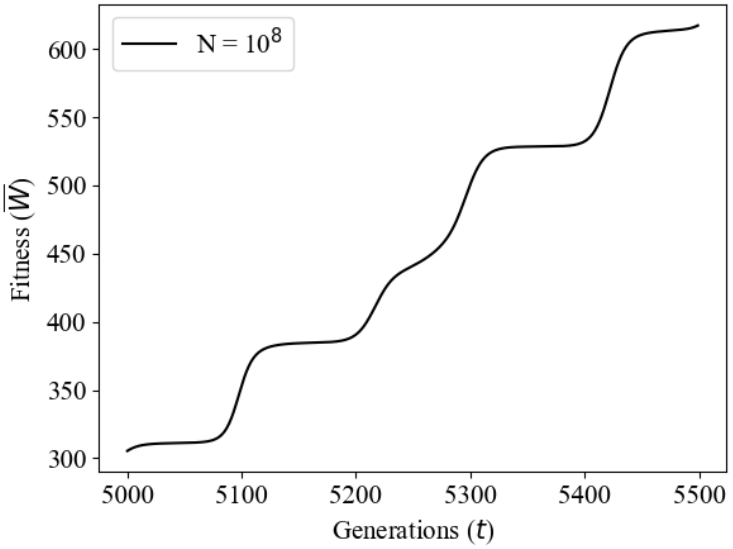


Figure 4.5: Fitness trajectory from figure 4.4b for population size $N = 10^8$ zooms into generation 5,000 to 5,500.

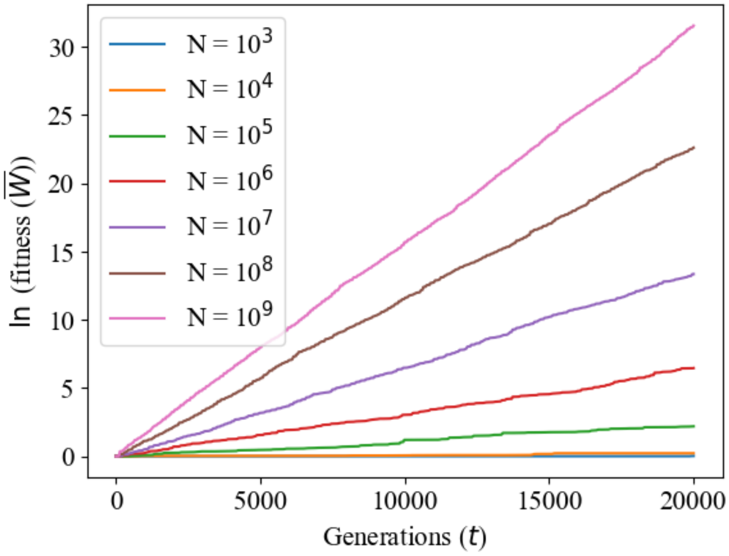
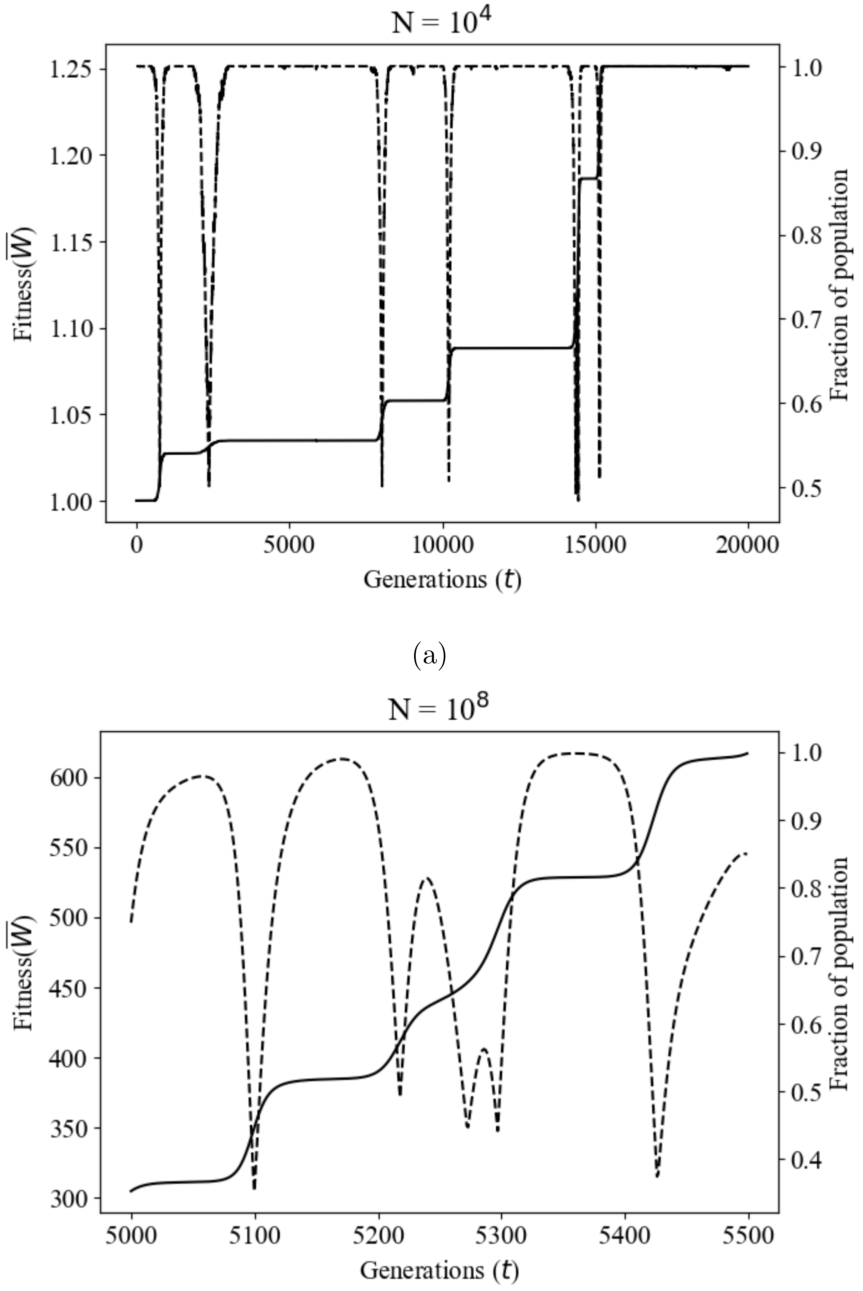


Figure 4.6: Natural log fitness trajectories for populations with varying sizes using the Wright-Fisher model. These simulations are ran for 20,000 generations with scaling parameter $\alpha = 50$ and mutation rate $\mu = 10^{-6}$. All populations' fitness experience exponential growth, as indicated by the straight lines in the natural log scale.



(a)

(b)

Figure 4.7: We superimpose on figures 4.4a (a) and 4.5 (b) the frequency of the dominant genotype at each generation as dashed curves (−−). The solid curves are the mean fitness of the populations. The "jumps" observed in fitness occur in close conjunction with the generations when the dominant genotypes reach their peak frequency, which can be 100%.

over time since the mean value of s with the exponential distribution $\alpha e^{-\alpha s}$ is $1/\alpha$. Wiser *et al.* (2013) proposes increasing α by setting $\alpha_{n+1} = \alpha_n(1 + g * \mathbb{E}(s_{\mathrm{f},\,n+1}))$ as mentioned in Chapter 2 with g being fitted from the LTEE data and found to be 6 [3]. This corresponds to updating α as $\alpha_{i,t} = \alpha_{i,t-1}(1 + g * s_{i,t})$ for each genotype in our Wright-Fisher model. Using the remaining parameters from Wiser *et al.* (2013) [3], we set the initial scaling parameter $\alpha_{i,0} = 85$ for all organisms, mutational rate $\mu = 1.7 \times 10^{-6}$, and population size $N = 3.3 \times 10^7$ (which is the effective population of the experiment). Since the Wiser *et al.* (2013) parameters are fitted for natural generations [3], we need to adjust our Wright-Fisher model for binary fission by applying a factor of $\ln(2)$.

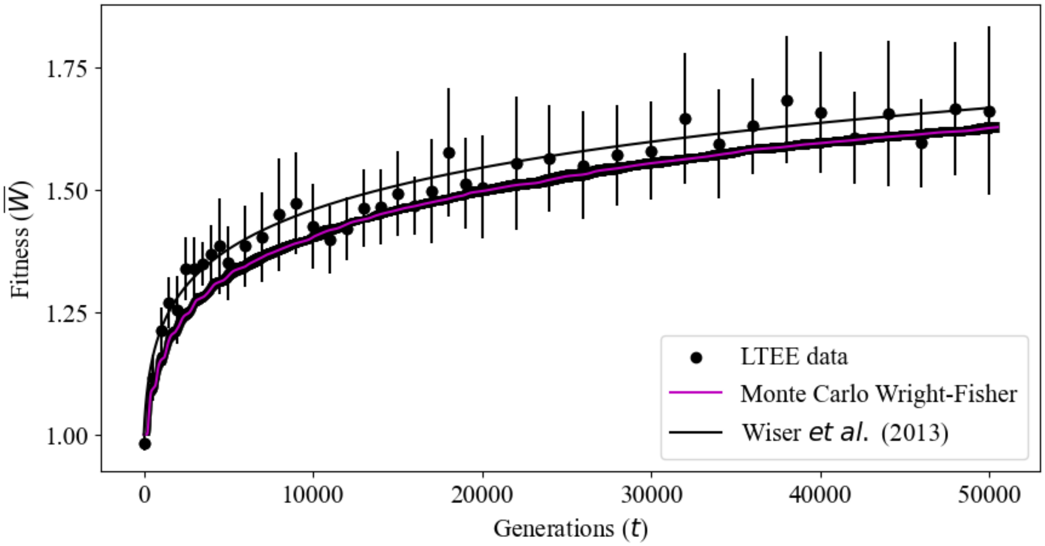
4.3.2 Results

The results of our simulations when accounting for epistasis is shown in figure 4.8a. Although it approximately follows the data trend from the LTEE experiment, utilizing Wiser *et al.* (2013) [3]'s parameter estimates appear to underestimate its prediction of the LTEE's fitness data. This is expected as we have shown empirically in the previous section that the Gerrish-Lenski model overestimates the fitness improvement of a Wright-Fisher model. Setting the population size and the diminishing constant g at the same values as the Wiser *et al.* (2013) [3]'s model, we can adjust either μ or the initial scaling factor α for the selection coefficient distribution to obtain a better fit to the LTEE data. We found the combinations of $\mu = 5 \times 10^{-6}$ and $\alpha_{i,0} = 85$ or $\mu = 1.7 \times 10^{-6}$ and $\alpha_{i,0} = 75$ appear to simulate the data well, see figure 4.9. Since the mutation rate for *E. coli* is normally set to be roughly 10^{-6}, we think the second combination that keeps the same mutation rate as Wiser *et al.* (2013) [3] may be a better estimate for α.

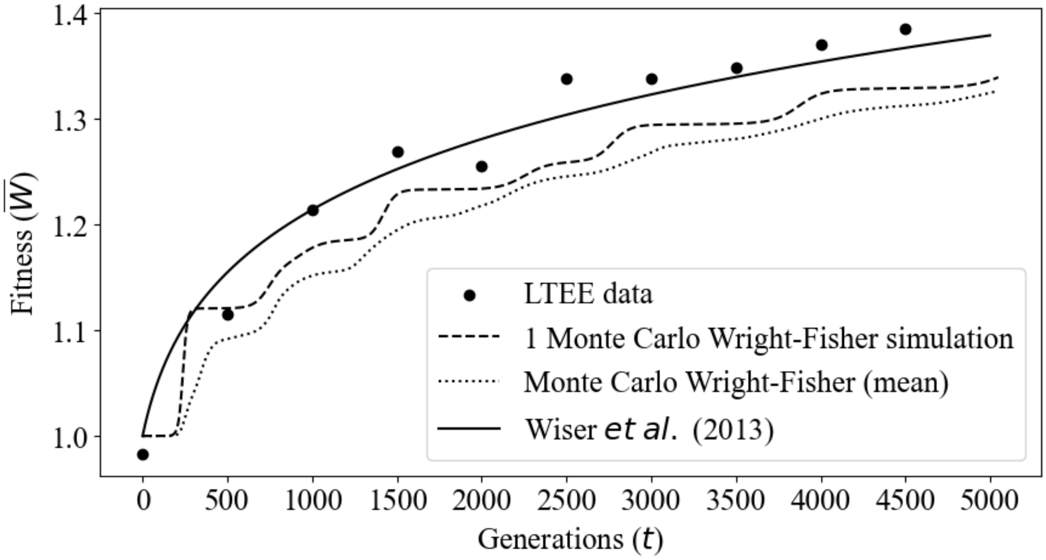
By adding a diminishing constant, the Wright-Fisher model is able to capture the declining fitness growth rate observed in the LTEE data. As mentioned previously, this diminishing constant has been fitted to the observed data. Without a diminishing constant, fitness follows an exponential growth in our Wright-Fisher model. As each genotype experiences additional mutations, the benefits from the new mutations are drawn from exponential distributions with a smaller means. This effectively reduces the collective fitness growth rate of the population over time. As shown, the trajectory of the LTEE fitness data can be fitted through different combinations of the variable inputs mutation rate μ and initial exponential distribution scaling parameter $\alpha_{i,0}$. Population size N can also be adjusted, although it is constrained by the LTEE so we chose not to vary it in this section.

4.4 Discussion

Wiser *et al.* (2013) [3]'s fit to the LTEE data exhibits a smooth line while our simulation results are more analogous to an irregular staircase. This is likely due to the fact that we only use the mean for six simulations to mimic the six populations in LTEE that did not undergo hypermutability. Although we expect a larger number of simulations will make our Wright-Fisher results appear more smooth, a single simulation is likely to exhibit rapid jumps in its fitness trajectory as we have seen in the case of no epistasis earlier due to clonal interference. In figure 4.8b, we zoom into a shorter time scale to highlight this behavior. We include both the mean values of six simulations and one sample simulation in addition to Wiser *et al.* (2013) [3]'s result. Due to the coarse-grained approach, their result exhibits a smooth trajectory of fitness. The Gerrish-Lenski model adapted by Wiser *et al.* (2013) evaluates the expected increase in fitness based on the fixation of one beneficial mutation at a time thus it does not track fitness values at every generation [3]. The fitness trend before fixation time is approximated in retrospect using the expected fitness increase once the mutation is fixed, or reaches 100% frequency in the population. The Wright-Fisher model is able to provide a higher trend resolution since it captures the population's fitness at each generation. For a single simulation of Wright-Fisher (4.8b), we can see the jumps in fitness more distinctly. The jumps get smaller and have longer time intervals in later generations due to a growing α which reduces the fitness improvement rate. Intuitively, increasing α values decreases the probability of producing mutations with large beneficial coefficients, leading to the declining effects of the mutations fixed. Mutations with smaller s values also take a longer time to circulate and become the standard type in the population.

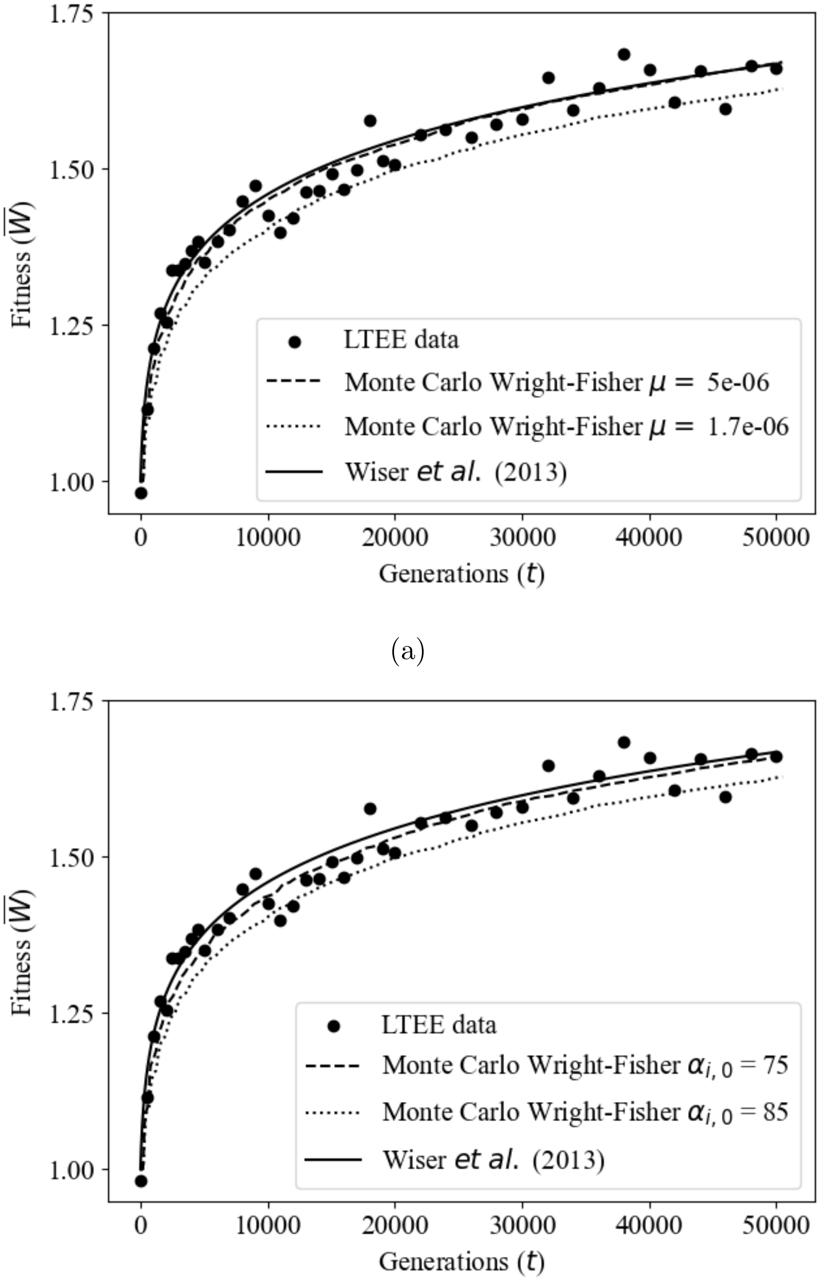


(a)



(b)

Figure 4.8: (a) Black circles are data from LTEE of the 6 populations that retained the ancestral mutation rate. Error bars are 95% confidence intervals. The black curve is the fitness trajectory predicted by Wiser *et al.* (2013) [3] and the magenta curve is the fitness trajectory from the Wright-Fisher simulations with error bars of 95% confidence intervals based on multiple simulations. (b) Same graph as a) up to 5,000 generations with an additional sample Wright-Fisher simulation and without error bars.



(a)

(b)

Figure 4.9: Wright-Fisher simulations with differing parameters. For both graphs, population size $N = 3.3 \times 10^7$ and diminishing constant $g = 6$ with (a) mutation rate $\mu = 5 \times 10^{-6}$ and scaling parameter $\alpha_{i,0} = 50$ and (b) mutation rate $\mu = 1.7 \times 10^{-6}$ and scaling parameter $\alpha_{i,0} = 75$. By adjusting the inputs mutation rate μ and/or scaling parameter $\alpha_{i,0}$, the Wright-Fisher model can robustly predict the LTEE trajectory.

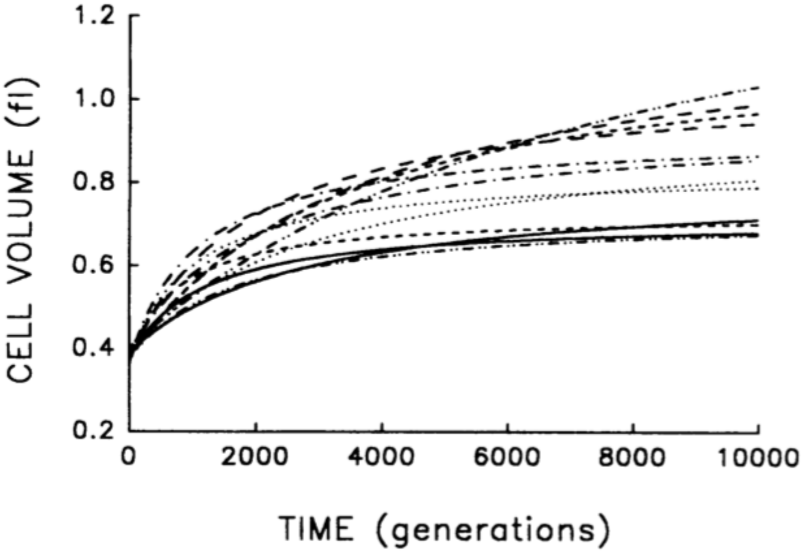


Figure 4.10: Trajectories of cell size growth for the twelve populations in the LTEE. Each curve is the best fit of a hyperbolic model for one population [20]. The cell growth trajectories of all populations show diminish rate in later generations. This figure is extracted from Lenski and Travisano 1994, please refer to their paper [20].

The jumps in fitness shown in figures 4.4 and 4.8b do not strictly demonstrate examples of punctuated equilibrium since they are not at the species scale, and occur in a relatively short time frame, both in the number of generations and in normal time. However, they show that the behavior of step-like changes can indeed occur in a biological system as observed in the simple laboratory experiment of the LTEE and even simpler mathematical model. Another possible interpretation of the LTEE data, and the Wright-Fisher simulation in conjunction, is that they seem to exhibit punctuated equilibrium at a higher scale than the granular generational level, at thousands of generations. Because of diminishing-returns epistasis in later fixed mutations, the fitness improvement is markedly smaller in latter generations. This was the same argument made by Gould (2009) when observing the trend of the cell size growth of LTEE [26]. Gould (2009) used the cell growth reported by Lenski and Travisano (1994) as an example of a punctuated pattern observed at a scale below species level [26][20]. As we have mentioned in Chapter 2, it exhibits a parallel trend to the fitness trajectory, showing rapid growth in earlier generations followed by little changes in the later generations (See figure 4.10). Since the *E. coli* populations in LTEE are cultivated in a different environment than what they were adapted to prior to the experiment, they may experience a rapid fitness increase which slows down as they adapt to the new environment. As the organisms are fully adapted to the environment and/or a balance is reached between beneficial mutations and deleterious mutations, no further fitness improvement will occur [3]. Our model and the LTEE data, of course, do not represent any remote interpretation of the fossil record since they are vastly simplified and at the organismic scale. However, it still demonstrates how a simple model like the Wright-Fisher can exhibit step-like trends in the relevant scales.

With one Wright-Fisher simulation, we observe "jumps" in fitness in both exponential fitness growth and the inclusion of a diminishing returns constant to model the LTEE observations. However, as shown in our prediction of the LTEE's fitness trajectory, averaging over a larger number of simulations reduces this effect. Increasing the number of the Monte Carlo simulations of the Wright-Fisher could reduce the uncertainties in the simulation results. As the LTEE's fitness data comprises of only twelve populations, and just six retain the ancestral mutation rate, we anticipate that a much larger sample size, e.g. 1,000 populations, will likely to eliminate any hint of intermittency. If a collection of populations are more likely to be naturally occurring, as the homogeneous conditions sustained in the LTEE and required in the Wright-Fisher model are difficult to be met in nature, this negates the potential for "punctuation" observed in this system. However, if a single population is likely to be sampled in naturally occurring conditions, we might be more likely to observe step-like behavior in the data. For an extreme example, consider an entire continent that is populated by one species of asexually reproducing organisms. If these organisms operate as discrete populations, samplings of their fitness growth over time will likely show a smooth trajectory. However, if they operate as one population, then we potentially could observe punctuated behavior in the sampling of their fitness growth over time.

CHAPTER 5

Smooth-colony dynamics

In this chapter, we develop and introduce our smooth-colony dynamics model to study the evolutionary progression of asexually reproducing populations at the sub-population scale. In this model, we look to analyze how population structure affects fitness evolution in search of punctuated equilibrium behavior. We are interested in building upon the principles derived from adapting the Wright-Fisher model and the conditions of the LTEE. The Wright-Fisher model is one of the most widely considered models in studying population evolutionary dynamics [39]. As we have described in the previous chapter, its most basic construction assumes a homogeneous population with discrete non-overlapping generations, no selection, no mutation, and no migration. By using an adapted model with selection and mutation through a Monte Carlo approach, we are able to incorporate the conditions of the LTEE more closely and this model has performed well in replicating the observed fitness trend. However, the laboratory environment of the LTEE and the subsequent assumptions built into the adapted Wright-Fisher model are still quite limited, and unlikely to be found in natural settings. In this chapter, we look to relax the homogeneity condition while introducing spatiality and local interactions in an evolving population. This provides a more accurate model of organisms' interactions in a natural environment where they are likely to be affected by local conditions. We focus on the one dimensional case in this chapter and look to extend the model into two dimensional in our future works.

5.1 Motivation

In contrast to the Wright-Fisher and Gerrish-Lenski models, we are interested in assessing the potential impacts of the spatial structure of a population. We use a bacteria colony, or a group of single cells grown in physical proximity, as the evolving unit that determines the mean fitness of the population. Bacteria often grow in clumps as the same strain tend to bind together. Although they are often easily detectable once grown in certain size, bacterial colonies are not so well-defined as highlighted in a review by Jeanson *et al.* 2015 [40]. A bacterial colony is most commonly viewed as a group of organisms that are grown into a cluster and usually share the same genotype as they are often derived from the same single parent cell [41]. This also has a physical manifestation where they form a grouping at the same location when grown on a surface medium in a laboratory, which is the definition we are employing in this work. A picture of bacteria colonies grown on a plate from the LTEE is shown in figure 5.1. We treat each cluster of organisms as a single independent entity where the colonies interact with each other as collective units. We track the evolution of these colonies in our system and how they impact the evolution of the population. The construction of our model is influenced by ideas underpinning the smoothed-particle hydrodynamics (SPH) model that's been widely used in physics and related fields. SPH is a complex system model that use distinct particles to simulate behavior in a continuum environment [42] [43]. Similar ideas have been found in the biological literature, although we have not explicitly incorporated them. For example, there have been numerous works attempting to address the issues of local interactions and collective motion in sub-population units for various biological systems [44][45].

The colonies in our simulation are situated next to each other in a line. We aim to extend this to a two dimensional surface in future work. Interaction between subpopulations with a linear spatial distribution are not uncommon. **Ring species**, though documented for multicellular organisms, is defined to be species connected geographically where each species is more likely to interact with the species that is closest in proximity [47]. Perhaps the most cited example is the *Ensatina eschscholtzii* species of salamanders with seven

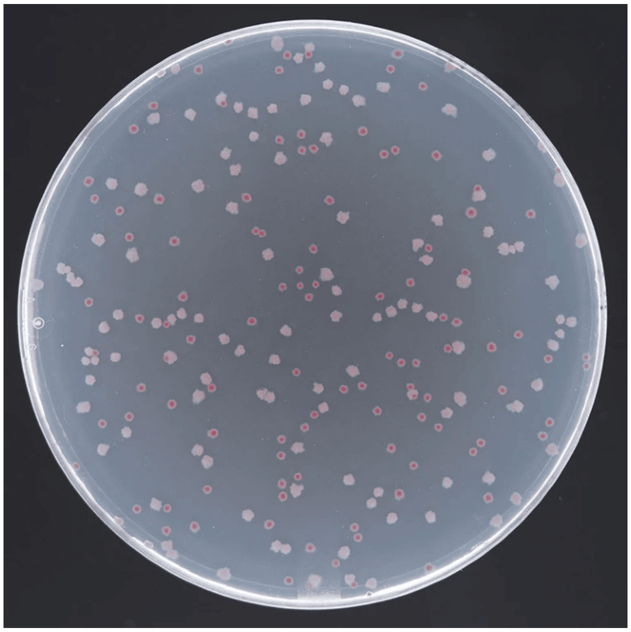


Figure 5.1: Bacteria colonies develop on a plate from the LTEE. They stay physically stationary and expand in size as more bacteria are grown. Red and white colonies mark strain Ara- and Ara+ respectively. This image is captured by Jeffrey Barrick [46].

subspecies being distributed around the Central Valley of California in a ring-like shape [48]. Computational and mathematical models at both the organismic and sub-population levels suggest that the interaction between drift and natural selection can be influenced by local spatial structure [44][49][45].

5.2 Model

We start with P colonies arranged in one-dimension, each with an accompanying index i ranging from 1 to P, making up the whole population. Each colony has N organisms and initially has the same genotype and thereby, the same fitness. Note the population in this model is made of colonies as opposed to organisms in the Wright-Fisher model in Chapter 4 and the Gerrish-Lenski model. However, as an individual organism multiples, it could ultimately be regarded as analogous to a colony.

In the new model considered here, our previous use of N in Wright-Fisher model has been replaced by P. N now describes the number of organisms in a given colony and P describes the number of colonies. We set this initial fitness as 1 as the ancestral reference fitness. Each of these colonies is treated as one homogeneous entity that carries similar assumptions to the Wright-Fisher model. We assume the time between binary fission for each colony follows a Rayleigh distribution. Each colony generation time is treated as a discrete event but the whole population has overlapping generational times, i.e. the colonies do not experience a generation at the same time. Generation time distributions for individual organisms have been a focus of study for different bacteria types and both symmetric and asymmetric distributions have been observed [50][51][52]. Thus time t takes on a continuous character. To eliminate any ambiguity, we will use τ for the random continuous time variable. A useful analogy for this is the human gestation period with a mean of 268 days and a standard deviation of 10 days. We assume a Rayleigh distribution for colony generation time for its non-negative values and being a common distribution. We have evaluated other distributions, which is discussed in a later section, and found the Rayleigh distribution to perform best as an approximation. After a certain number of generations, each colony undergoes a random improvement in fitness, or a fixation event. The fixation time, or the number of generations before a fixation event, depends on the size of the colony, the mutation rate, and the selection coefficient fitness distribution. Once a fixation event occurs, all members in the colony carry the same genotype with a fitness increase from the last fixation event. The clonal fitness improvement distribution is derived based on the evolutionary dynamic at the organism level. Although we are discretizing the fitness improvement into one single event, population genetic modeling based on clonal interference in asexually reproducing organisms suggests that beneficial mutations do not meaningfully affect the population's fitness until it becomes the dominant genotype [2]. In fact, the growth in the higher fitness genotype's frequency can be well approximated as logistic when there are only two genotypes as we have shown in Chapter 2. When multiple genotypes are allowed, we also observe this jump in population fitness as seen in the Wright-Fisher simulations in Chapter 4.

When a colony reaches its fixation time, its fitness improves by a value based on a distribution modeled for clonal interference. We want to make a distinction here between a mutation event and a fixation event. A mutation event occurs at the cellular level. Each time a bacterial cell undergoes binary fission, it can potentially be exposed to a mutation event. As we have discussed in Chapter 1, we assume a mutation–selection equilibrium has been reached for deleterious mutations for large populations and/or populations with high mutation rates thus adaptation rate is determined by beneficial mutations [23] [13]. A fixation event, which determines the colony fitness improvement event in our model, occurs when a mutation or multiple mutations become ubiquitous in the population [53]. In our model, we are interested in the fixation event of a colony. Thus, we will not distinguish between the number of mutations being fixed but only the accompanying genotype being dominant in the colony. Although we are treating a fixation event as the point when only one genotype remains in the colony, in a natural environment, large populations will almost always be polymorphic since they are likely to generate numerous mutations each generation.

5.2.1 Clonal distribution of selection coefficients

To determine the fitness evolution for each colony, we need a distribution of selection coefficient at the clonal unit. In the Wright-Fisher model, we use the exponential distribution to draw the selection coefficient effect for each mutation. Similarly, we need a distribution for the selection coefficient effect that is fixed in a colony. Intuitively, when an organism in the Wright-Fisher model experiences a mutation, its fitness improves by an exponential rate of beneficial effect s. Thus, when a colony in the smooth-colony dynamics model experiences a fixed mutation, its fitness improves by an exponential rate of beneficial effect s_{f} for the whole colony. In this chapter, we establish the distribution for the fixed selection coefficient for a colony.

We assume that within a colony, the asexually reproducing organisms experience mutations and selection with no horizontal gene transfer and its fitness evolutionary trend may be approximated by the Gerrish-Lenski model. This provides us with a continuous probability distribution of fixed selection coefficient effects and an expected rate of substitution. As outlined in Chapter 2, the Gerrish-Lenski model approximates the distribution of the fixed selection coefficient s_{f} for a population (colony in this smooth-colony dynamics model) of size N, mutation rate μ, and exponential distribution scaling parameter α as:

$$p(s_{\mathrm{f}}) = K4s_{\mathrm{f}}\, e^{-\lambda(s_{\mathrm{f}},\alpha,\mu,N)-\alpha s_{\mathrm{f}}} \tag{5.1}$$

where K is a normalizing constant such that $\int_0^\infty p(s_{\mathrm{f}})\, ds_{\mathrm{f}} = 1$ and

$$\lambda(s_{\mathrm{f}},\alpha,\mu,N) = \frac{\mu}{s_{\mathrm{f}}} N \ln(N) e^{-\alpha s_{\mathrm{f}}}\, 4(s_{\mathrm{f}} + \frac{1}{\alpha}) \tag{5.2}$$

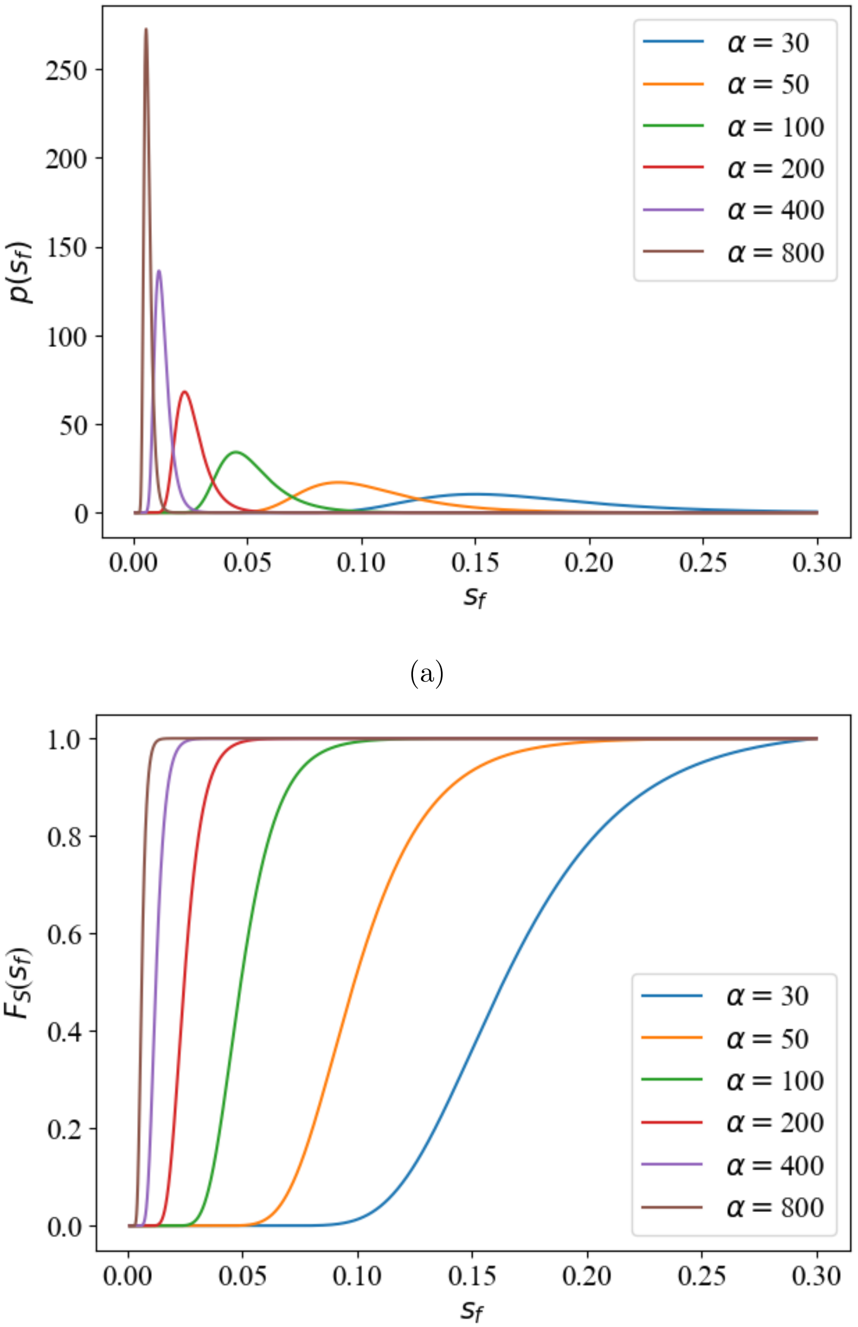
Then the expected value of s_{f} could be found by integrating $s_{\mathrm{f}}\, p(s_{\mathrm{f}})$ over all possible values of s_{f} or

$$\mathbb{E}(s_{\mathrm{f}}) = \int_0^\infty s_{\mathrm{f}}\, p(s_{\mathrm{f}}) ds_{\mathrm{f}} \tag{5.3}$$

and the expected substitution rate, or number of mutations, per generation for the population is:

$$\mathbb{E}(\text{ substitution rate}) = \mu N \alpha \int_0^\infty 4s_{\mathrm{f}} e^{-\lambda(s_{\mathrm{f}},\alpha,\mu,N)-\alpha s_{\mathrm{f}}} ds_{\mathrm{f}} \tag{5.4}$$

Since we are interested in the fitness evolution of each colony as a unit, we are using equation 5.1 to computationally sample fixed selection coefficient s_{f} for a colony of size N and mutation rate μ. We can utilize inverse transform sampling to computationally sample random values of s_{f} which convert random variables drawn from a uniform distribution $[0, 1)$ into any distribution. Appendix C.1.1 provides an overview of the inverse transform sampling method. In order to perform inverse transform sampling computationally, we need the **cumulative distribution function** or CDF of the distribution described by equation 5.1, which is the PDF or the **probability density function**. Let S be a random variable drawn from $p(s_{\mathrm{f}})$ and $F_S(s_{\mathrm{f}})$ be the CDF of $p(s_{\mathrm{f}})$. Then $F_S(s_{\mathrm{f}}) = P(S \leq s_{\mathrm{f}})$ where the right hand side represents the probability of the random variable S is less than s_{f}. Since $P(S \leq s_{\mathrm{f}}) = \int_0^{s_{\mathrm{f}}} p(s_{\mathrm{f}})\, ds_{\mathrm{f}}$, we have not been able to recover a closed form solution to $F_S(s)$. We can numerically calculate the CDF but we would like to pursue a closed form CDF that performs similarly to the CDF of equation 5.1. Holding mutation rate μ and colony size N constant, we plotted equation 5.1 for varying exponential distribution scaling parameter α values and the accompanying $F_S(s)$ in figures 5.2a and 5.2b respectively.



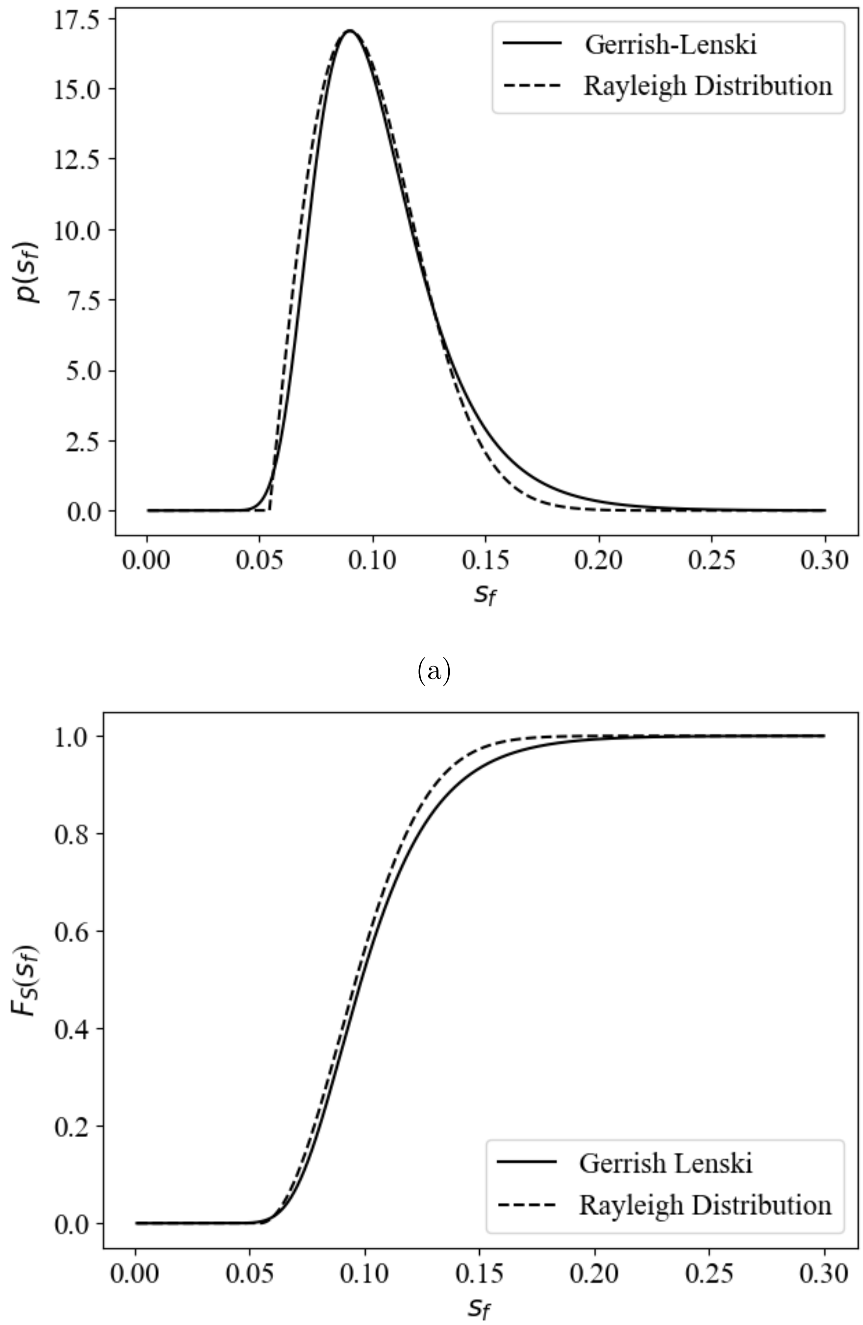
(a)

(b)

Figure 5.2: Equation 5.1 and its integral plotted for a range of exponential distribution scaling parameter α values in plot (a) and (b) respectively. We set colony size $N = 10^6$ and mutation rate $\mu = 10^{-6}$.

The distribution of fixed selection coefficient s_f resembles the Rayleigh distribution of the form $\frac{s_\text{f}-d}{\sigma^2}e^{-(s_\text{f}-d)^2/(2\sigma^2)}$ where σ and d are characteristic parameters. In figures 5.3a and 5.3b, we have overlaid the Rayleigh distribution over one of the curves in figure 5.2. The Rayleigh distribution's parameters σ and d are fitted so the modes of the two distributions overlap. As the Rayleigh distribution fits the Gerrish-Lenski model well, we have opted to utilize its CDF to numerically draw the random variable of fixed selection coefficient for a colony. This improves computational efficiency without sacrificing much accuracy. Given colony size N, mutation rate μ, and exponential distribution scaling parameter α, we can fit for the Rayleigh distribution parameters σ and d. For all our simulations, we hold the colony size N and the mutation rate μ constant throughout the simulation. By holding colony size N constant, we are establishing a multicellular unit there each colony's organisms are free to interact with each other internally but must interact with the external environment as a whole entity. This is an attempt to model the interaction between global and local effects. The size of a unit that can act cohesively is theoretical and may fluctuate in reality but it is biologically reasonable to model it as some fixed value that can be fitted with real life data. Evolutionary dynamics of subdivided populations have been studied in a number of different contexts with varying population structures [45][54][55].

Similar to the Wright-Fisher model, we are not addressing a varying mutation rate μ in this model and are focusing on the LTEE populations that have maintained an ancestral mutation rate. LTEE populations can experience different mutation rates over time [56]. Populations can carry the trait for many generations if it's accompanied by beneficial mutations. However, there is a trade-off between increasing mutation rate and the accumulations of deleterious mutations [19]. Higher mutation rates allow the populations to generate more mutations in a shorter period of time enabling faster evolution but at a cost of potentially introducing more deleterious mutations into the population. If the hypermutability trait is accompanied by beneficial mutations that outweigh the effects of the deleterious mutations, then it may be able to stick around. This may change when the push and pull between costs associated with deleterious mutations become too high to be balanced. For example, as the



(a)

(b)

Figure 5.3: (a) PDF's and (b) CDF's from the Gerrish-Lenski model and the Rayleigh distribution with parameters σ and d fitted for the modes of the two distributions to overlap. This provides evidence that our empirical selection of a Rayleigh distribution captures features modeled quantitatively by the Gerrish-Lenski model. We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.

populations in the LTEE become more adapted to their environment, there is potentially less room for improvement leading to a decline in fitness growth. In this regime, the hypermutability trait may be weeded out which has been observed in some hypermutating LTEE populations that later reverted to their original mutation rate [14]. Thus, more thought is required to tease out the relationship between mutation rate and the distribution of beneficial mutations which is outside the scope of this work. Additionally, changing the mutation rates would only change the magnitude of the trends observed in our simulations and not the qualitative trends.

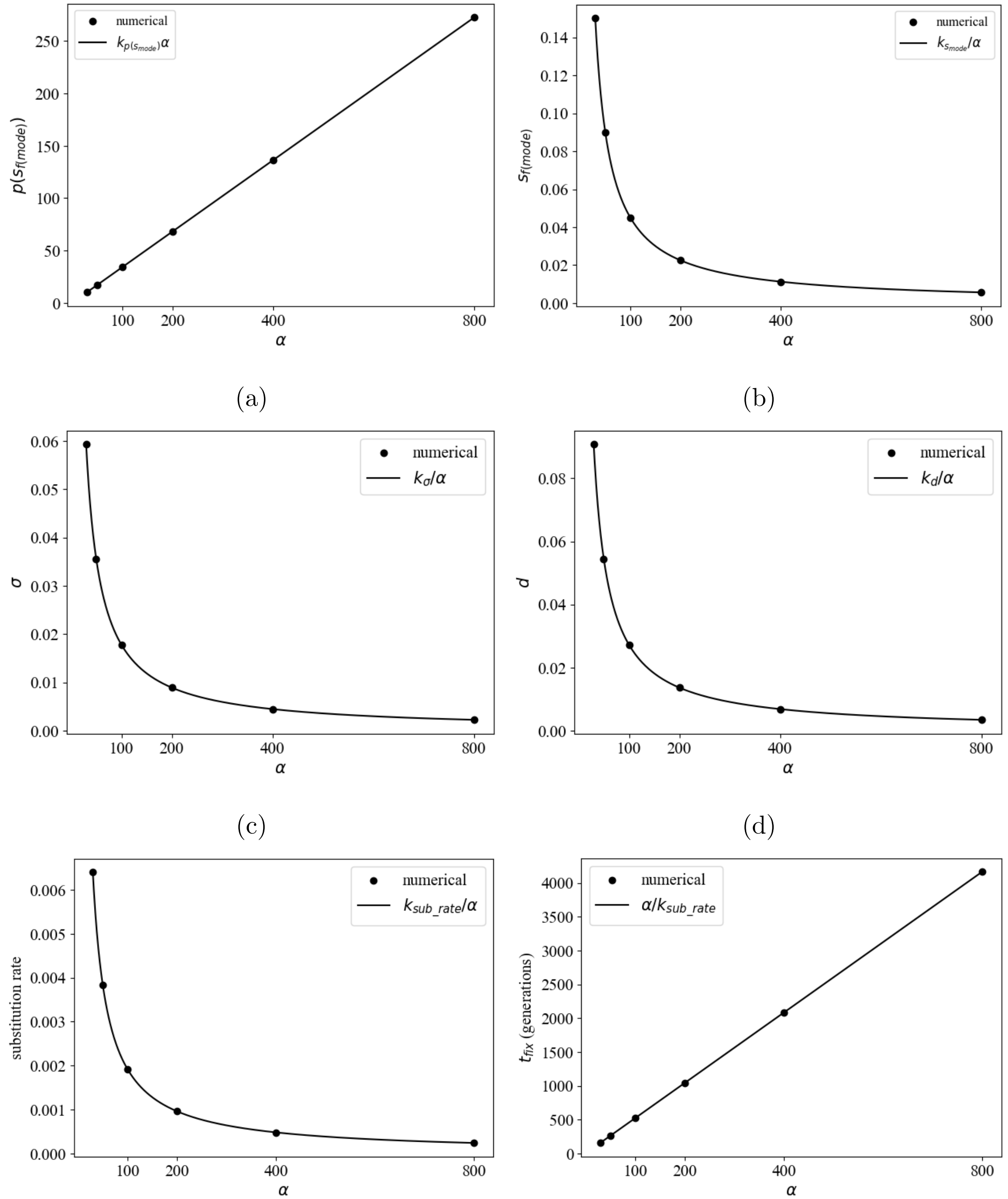
In order to model the LTEE's fitness observations, we are interested in the scenario where α varies due to diminishing returns in additionally fixed mutations as we have done with the Wright-Fisher model in Chapter 4. As α varies, the distribution of the fixed selection coefficient, equation 5.1, and the expected substitution rate, equation 5.4, will change. This in turn affects the σ and d parameters in the Rayleigh distribution, the approximation of equation 5.1 that we are utilizing to draw random variable S. We are interested in characterizing the relationship between these two parameters and the exponential distribution scaling parameter α to reduce the computational time needed to numerically calculate them as α changes. Since the Rayleigh distribution approximates the distribution of the fixed selection coefficient $p(s_{\mathrm{f}})$ well and its mode is σ, we conject that the parameters vary proportionally to α while colony size N and mutation rate μ stay constant. Indeed, we find this to be numerically the case. In fact, we find this to be approximately the case for varying μ as well, when holding colony size N and exponential distribution scaling parameter α constant. Due to the $\ln(N)$ term, the variables' relationship with colony size N is less linear while holding the other two inputs constant. The mode of the distribution $p(s_{\mathrm{f}})$ varies in direct proportional to α by some constant $k_{p(s_{\mathrm{f\ (mode)}})(\mu,N)}$ making the Rayleigh distribution parameters σ and d varying inversely proportional to α by constants $k_\sigma(\mu, N)$ and $k_d(\mu, N)$ respectively. The substitution rate has an inverse proportional relationship with α thus making t_{fix} varying in direct proportional to α by constant $1/k_{\mathrm{sub\_rate}}(\mu, N)$. Please refer to figure 5.4 where mutation rate μ and colony size N are kept constant while varying α. We used the same exponential distribution scaling parameter α values as figure 5.2. We find these trends to be true for a wide range of mutation rate μ and colony size N values. This in turn allows us to perform the simulation very rapidly without the need for extensive implicit calculations. By utilizing the Rayleigh distribution and recognizing the relationship between α and the distribution of fixed selection coefficient $p(s_{\mathrm{f}})$, we are able to provide a newly enhanced approximation of the Gerrish-Lenski mathematical model for faster computation without sacrificing significant accuracy.

5.2.2 Algorithm

A general conception for this algorithm proceeds as follows. For P colonies starting with the same initial fitness, we track the fitness evolution of each colony. Each colony carries two timers, one that determines its next binary fission event and the second to determine its next fixation event. During each selection, the colony with the smallest time to binary fission undergoes one doubling event. Once a colony's time to fixation event is reached, it experiences a fitness improvement event. The simulation is run until a predetermined number of M selections is reached. A computational description of our algorithm can be found in Appendix C.4 and a simple visualization can be seen in figure 5.5.

5.2.3 Two model versions: non-interacting neighbors and interacting neighbors

We create two versions of our model: the non-interacting neighbors (NIN) and the interaction neighbors (IN). The two versions follow the same algorithm described above with a difference in how the colonies experience fitness improvement during a fixation event. In the NIN version, the colonies evolve independently and experience fitness improvement through mutations that occur in each colony. In the IN version, the colonies' fitness impact their neighbors. During a fixation event, a colony can take on its neighbors' fitness if they are higher than its own. This allows an interaction between the colonies.

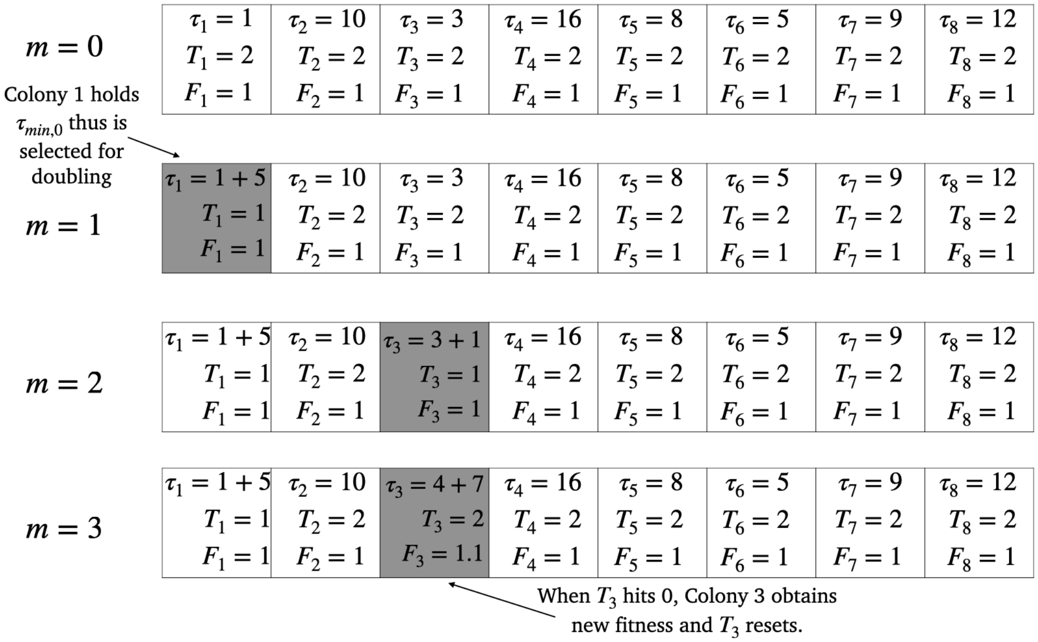


(a) (b)

(c) (d)

(e) (f)

Figure 5.4: Using the the Gerrish-Lenski model, exponential distribution scaling parameter α is plotted against: (a) fixed selection coefficient distribution $p(s_{\text{f(mode)}})$ where $s_{\text{f(mode)}}$ is the mode of the distribution, (b) mode of $p(s_{\text{f(mode)}})$, the Rayleigh distribution parameters (c) σ and (d) d versus α, (e) the substitution rate $\mathbb{E}(\text{substitution rate})$ and f) fixed time $\mathbb{E}(t_{\text{fix}})$.



$m = 0$	$\tau_1 = 1$ $T_1 = 2$ $F_1 = 1$	$\tau_2 = 10$ $T_2 = 2$ $F_2 = 1$	$\tau_3 = 3$ $T_3 = 2$ $F_3 = 1$	$\tau_4 = 16$ $T_4 = 2$ $F_4 = 1$	$\tau_5 = 8$ $T_5 = 2$ $F_5 = 1$	$\tau_6 = 5$ $T_6 = 2$ $F_6 = 1$	$\tau_7 = 9$ $T_7 = 2$ $F_7 = 1$	$\tau_8 = 12$ $T_8 = 2$ $F_8 = 1$
$m = 1$	$\tau_1 = 1 + 5$ $T_1 = 1$ $F_1 = 1$	$\tau_2 = 10$ $T_2 = 2$ $F_2 = 1$	$\tau_3 = 3$ $T_3 = 2$ $F_3 = 1$	$\tau_4 = 16$ $T_4 = 2$ $F_4 = 1$	$\tau_5 = 8$ $T_5 = 2$ $F_5 = 1$	$\tau_6 = 5$ $T_6 = 2$ $F_6 = 1$	$\tau_7 = 9$ $T_7 = 2$ $F_7 = 1$	$\tau_8 = 12$ $T_8 = 2$ $F_8 = 1$
$m = 2$	$\tau_1 = 1 + 5$ $T_1 = 1$ $F_1 = 1$	$\tau_2 = 10$ $T_2 = 2$ $F_2 = 1$	$\tau_3 = 3 + 1$ $T_3 = 1$ $F_3 = 1$	$\tau_4 = 16$ $T_4 = 2$ $F_4 = 1$	$\tau_5 = 8$ $T_5 = 2$ $F_5 = 1$	$\tau_6 = 5$ $T_6 = 2$ $F_6 = 1$	$\tau_7 = 9$ $T_7 = 2$ $F_7 = 1$	$\tau_8 = 12$ $T_8 = 2$ $F_8 = 1$
$m = 3$	$\tau_1 = 1 + 5$ $T_1 = 1$ $F_1 = 1$	$\tau_2 = 10$ $T_2 = 2$ $F_2 = 1$	$\tau_3 = 4 + 7$ $T_3 = 2$ $F_3 = 1.1$	$\tau_4 = 16$ $T_4 = 2$ $F_4 = 1$	$\tau_5 = 8$ $T_5 = 2$ $F_5 = 1$	$\tau_6 = 5$ $T_6 = 2$ $F_6 = 1$	$\tau_7 = 9$ $T_7 = 2$ $F_7 = 1$	$\tau_8 = 12$ $T_8 = 2$ $F_8 = 1$

Colony 1 holds $\tau_{min,0}$ thus is selected for doubling

When T_3 hits 0, Colony 3 obtains new fitness and T_3 resets.

Figure 5.5: A schematic of the smooth-colony dynamics algorithm. Eight colonies start with the same fitness and time to fixation event but varying time to doubling $\tau_{i,m}$'s. At each selection, the colony with the lowest $\tau_{i,m}$ value experiences a doubling event which is reflected in its reduction in the time to the next fixation event. When a colony reaches a fixation event, its fitness increases and the time to fixation event resets.

5.2.4 Doubling time ordering

Array τ determines which colony is to undergo doubling at the current selection number m. As mentioned in the algorithm outline above, at the current time step, the smallest value in τ reaches its generation time. We are using the Newman-Phoenix hierarchical data structure that was utilized to reproduce the Bak-Sneppen model. This allows accessing the minimum value of an array in $\mathcal{O}(1)$ time and updating the data structure in $\mathcal{O}(\log_2 N)$ time. The description of the implementation of the Newman-Phoenix hierarchical data structure can be found in Chapter 3.

Although we have chosen to use the Rayleigh distribution for the initial values of $\tau_{i,0}$ and their subsequent additives, the behavior of our simulations is surprisingly robust to the distribution used. We demonstrate this by comparing the number of selections while sampling values from the exponential and uniform distributions. The distribution of the number of selections for the colonies is approximately normal for when fitness stays constant over time, which indicates that the process of selection is sufficiently random such that all colonies are equally likely to double over time. In the non-interacting neighbors case (NIN), the colonies evolve independently. In the interacting neighbors (IN) case, the fitness evolution of the colonies is impacted by their neighbors. However, when there are no mutations and no selection, or the fitness stays constant, both cases are effectively the same. We define as a generation every P time steps since that is on average the number of selections necessary for all colonies to have an opportunity to double. Figure 5.6 is a normal probability plot of the number of selections for a number of cases where we utilize different distributions to draw $\tau_{i,m}$. These distributions include uniform, exponential, and a referenced distribution labelled binomial. We ran these simulations for 100,000 generations. A probability plot helps visualize and assess whether a certain data set follows some hypothesized distribution, in this case a normal distribution [57]. A linear fit to these simulation results indicates little deviation from the normal distribution with varying standard deviations. This indicates the majority of the colonies experience a similar number of selections. We use a reference distribution, labelled binomial, for comparison. In this reference case, all colonies are equally likely to be

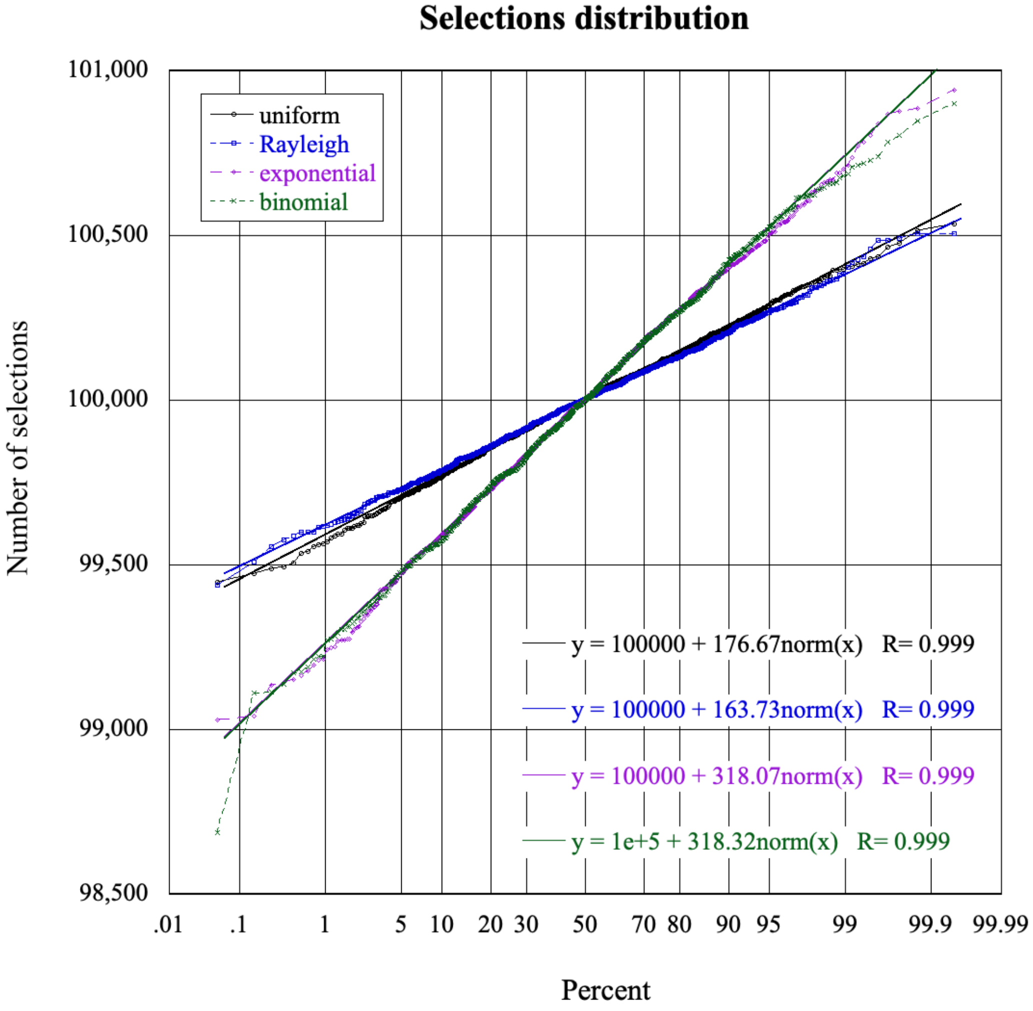


Figure 5.6: Normal probability plots for the number of selections using three different distributions to draw generation times: uniform, Rayleigh, and exponential. The binomial distribution is the reference case where all colonies are equally likely to be selected at each selection.

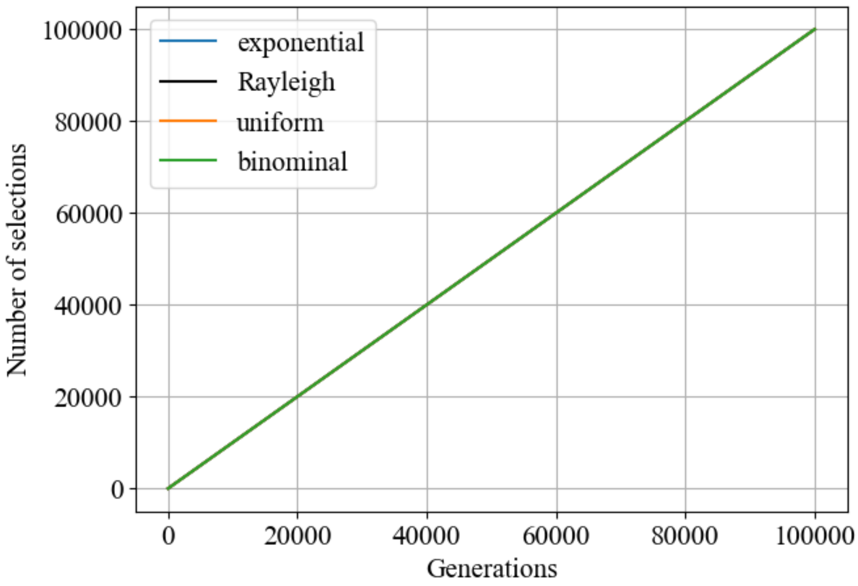


Figure 5.7: Mean of number selections of all P colonies for 100,000 generations. The mean of number selections stays the same for all distributions simulated. Note, the lines in the plot are nearly indistinguishable.

selected at each time step. Thus, the number of selections for all colonies follows a binomial distribution. The exponential distribution appears to mirror the reference distribution. The Rayleigh and the uniform distributions have smaller standard deviations than the reference distribution, indicating that the colonies' numbers of selections are closely aligned.

Figure 5.7 and figure 5.8 show the mean number of selections for the whole population, i.e. all colonies, and its variance for 100,000 generations respectively. All distributions produce a consistent mean number of selections over time, following closely the trajectory of the reference case. The variance for the reference case grows faster or at the same rate as the variance of other distributions, indicating that the colonies in all distributions studied have closely distributed number of selections relative to the reference case. Another useful statistic to investigate the performance of these distributions is the coefficient of variation which measures the dispersion of a data distribution. The coefficient of variation is the ratio between the standard deviation and the mean. Figure 5.9 is a log–log plot for the coefficient of variation for 100,000 generations for all distributions used. Regardless of distribution used for selection, the dispersion of the number of selections for all colonies declines over time and exhibits a power-law behavior. Since the results for different distribution functions display similar behavior, this demonstrates the robustness of our model.

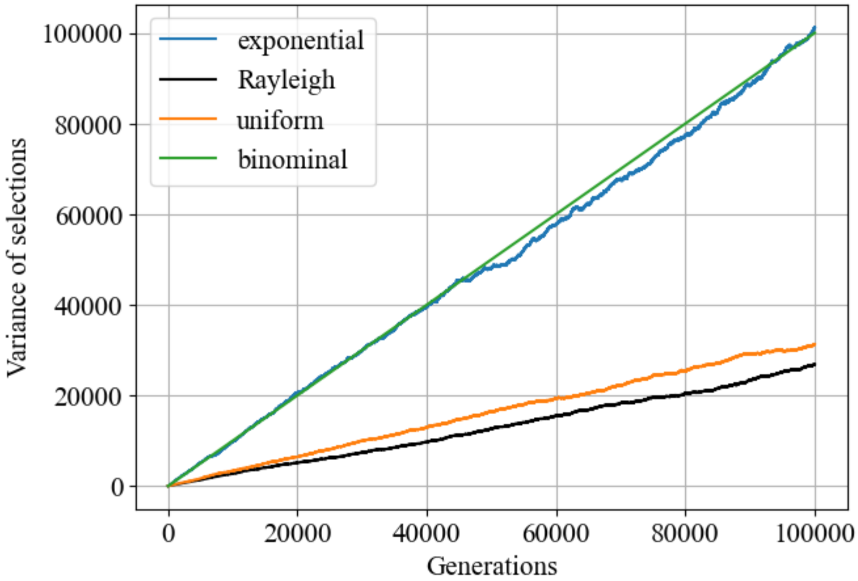


Figure 5.8: The variance of the number selections for 100,000 generations.

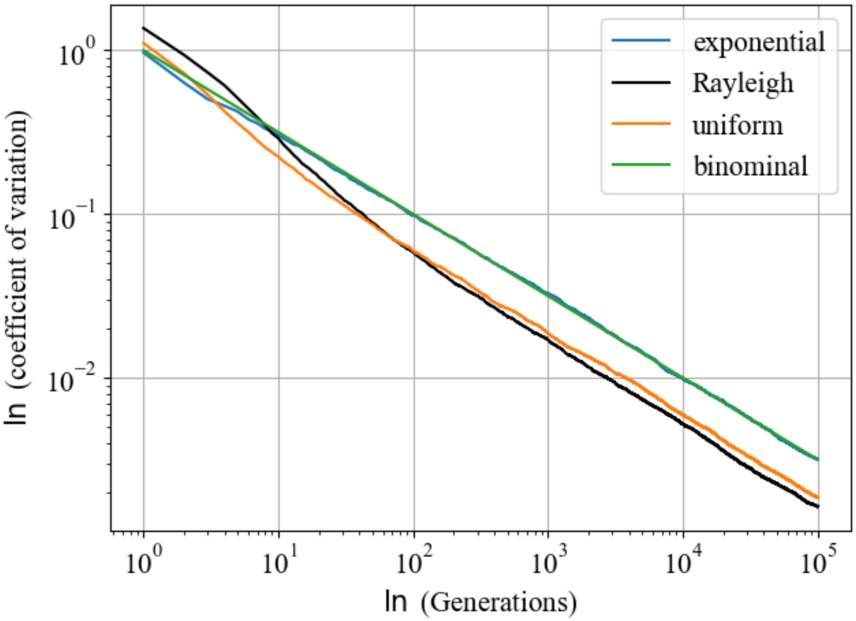
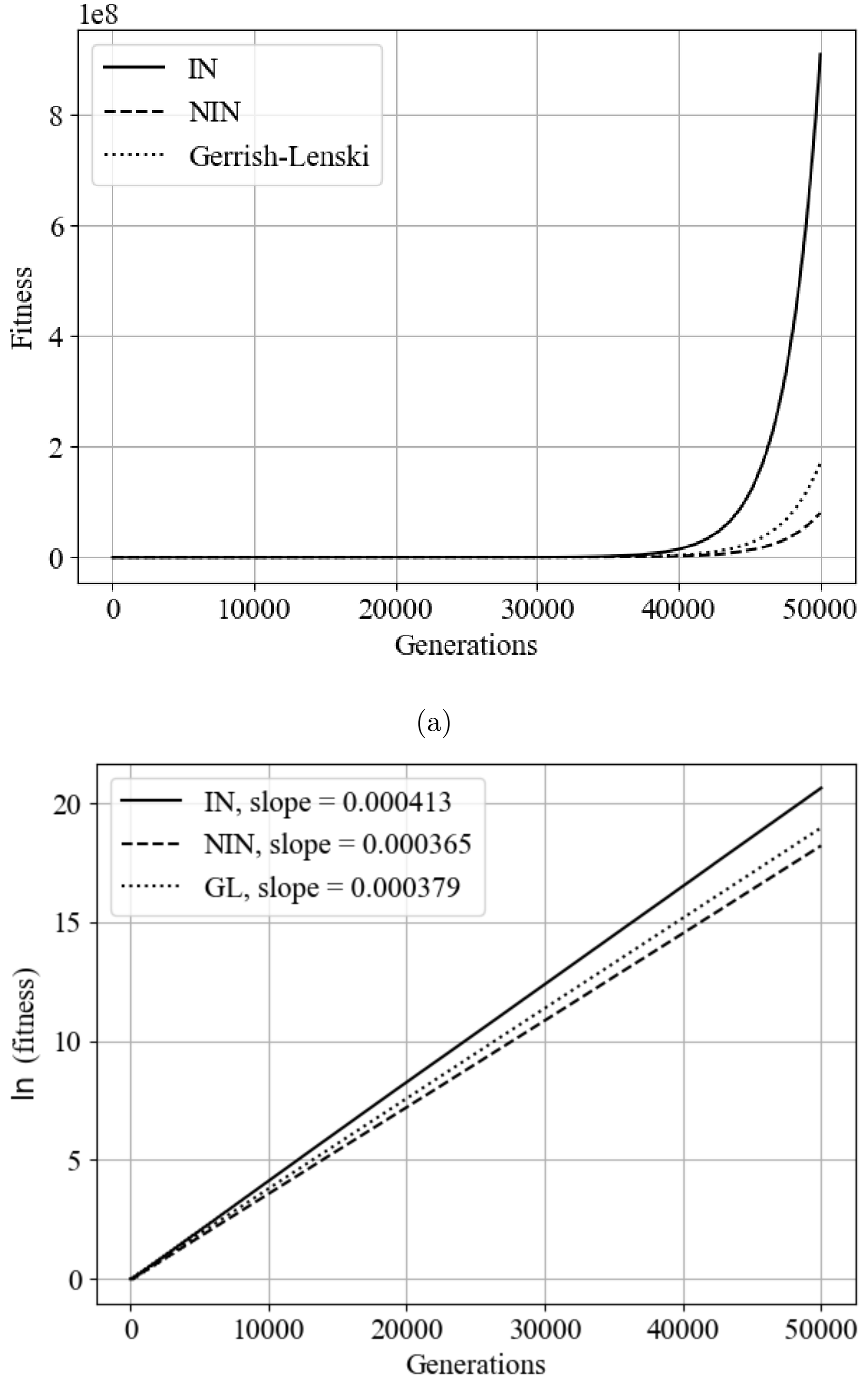


Figure 5.9: The log–log plots of the coefficient of variation for 100,000 generations.

5.3 Results

For the results in this section, the parameters are set as follows: number of colonies $P =$ 1,000, colony size $N = 1 \times 10^6$, mutation rate $\mu = 1 \times 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$. With the exception of the number of colonies, the values for the remaining parameters are reasonable for mathematical modeling of *E. coli*. The colony size N can range anywhere from 50 organisms to 10^9, depending on cell size and packing density of the colony [58][59][60]. We ran our simulations for $M = 5 \times 10^7$ selections, resulting in 50,000 generations. As expected, the fitness in the neighbor-interacting case grows faster than the non-neighbor-interacting case, see figure 5.10a. We also include the prediction from the Gerrish-Lenski model using the same parameters for colony size N, mutation rate μ, and exponential distribution scaling parameter α. All three curves demonstrate exponential growth, which is expected since fitness increases by a multiplicative function of the fixed selection coefficient. In figure 5.10b, we plot the fitness trajectories in log scale to show this trend more clearly over time. Using the natural log of mean fitness over time, we can calculate $\frac{\mathrm{d}\ln(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}}$ empirically then subsequently calculate $\frac{\mathrm{d}(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}}$ since $\frac{d(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}} = \frac{d\ln(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}} e^{\frac{\mathrm{d}\ln(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}} t_{\mathrm{generations}}}$. For figure 5.10, we calculate $\frac{\mathrm{d}\ln(f_{\mathrm{mean}})}{\mathrm{d}t_{\mathrm{generations}}}$ to be ≈0.000365, ≈0.000413 for the NIN and IN cases respectively. Compare this to the Gerrish-Lenski's model, which utilizes only the expected values, it is found to be ≈0.000379. This is reflected in figures 5.10a and 5.10b where we can see that the NIN curve grows slower than both the Gerrish-Lenski curve and the IN curve, though not by much. We suspect the stochastic nature of the NIN case slows down its fitness growth trend compared to the Gerrish-Lenski curve. For the IN case, the interacting feature allows the population to improve in fitness quicker than the Gerrish-Lenski model, even when utilizing their approximation for clonal selection coefficient distribution.

Figure 5.11 shows the fitness values of the P colonies after M selections for both IN and NIN models. The distribution for the natural log of fitness in the NIN case appears random. The fitness distribution for the interacting neighbors case shows patterns of interactions resembling diffusion through the colonies. The probability plot suggests that they both

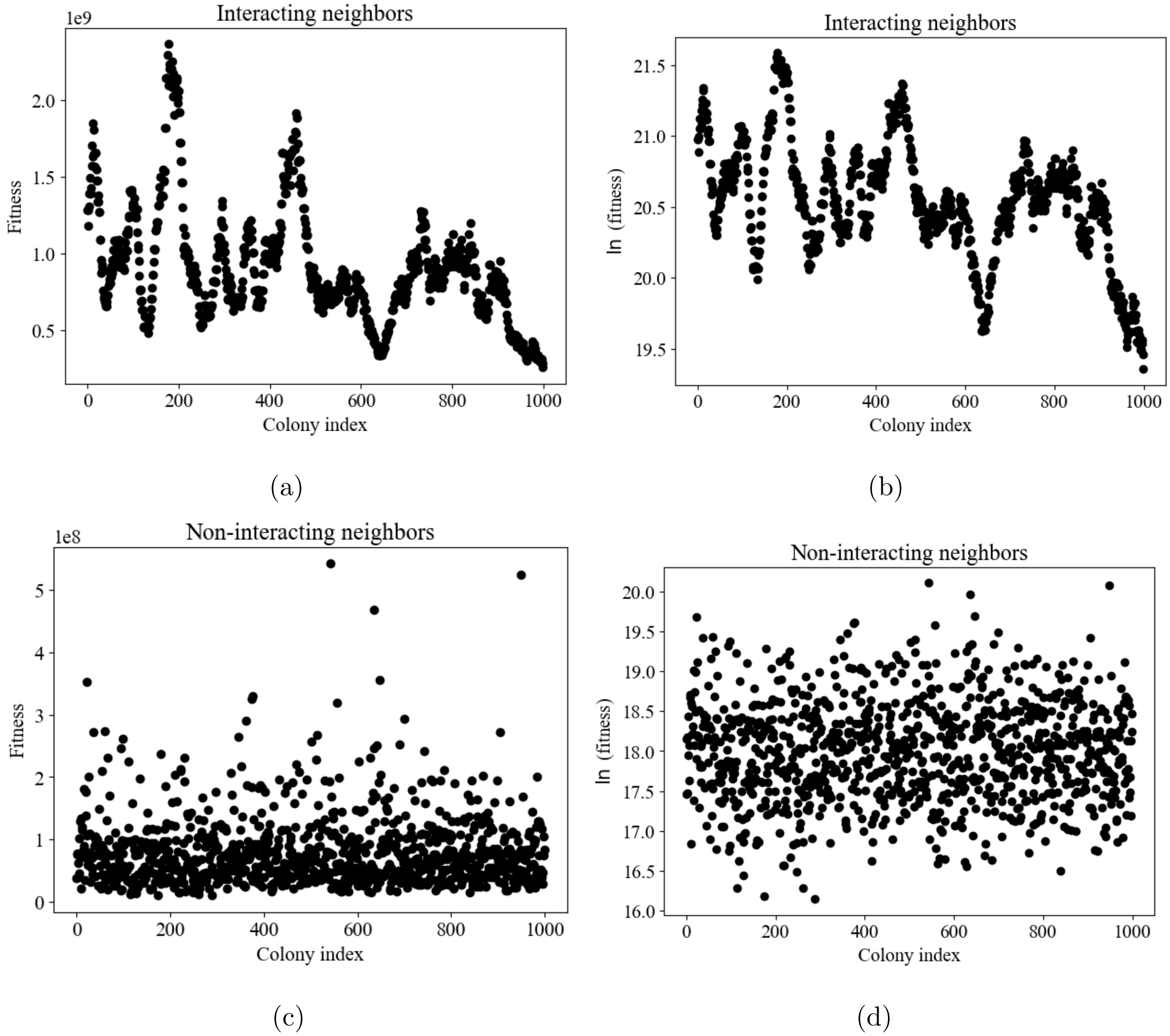


(a)

(b)

Figure 5.10: (a) Fitness (b) Natural log fitness for IN and NIN cases for 50,000 generations. We also include the Gerrish-Lenski prediction. We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.

follow a normal distribution with the NIN case being a better fit, see figure 5.12.



(a)

(b)

(c)

(d)

Figure 5.11: Fitness distributions for IN and NIN cases after 50,000 generations. (a) Fitness (IN) (b) Natural log fitness (IN) (c) Fitness (NIN) (d) Natural log fitness (NIN). We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.

Figure 5.13 displays the distribution of the number of selections for P colonies in each case after undergoing M selections. The histogram for the NIN case can be closely approximated with a normal distribution whereas the IN case has a more narrow and right skewed distribution. The variance in the NIN case is larger than the IN. This is reasonable as the in-

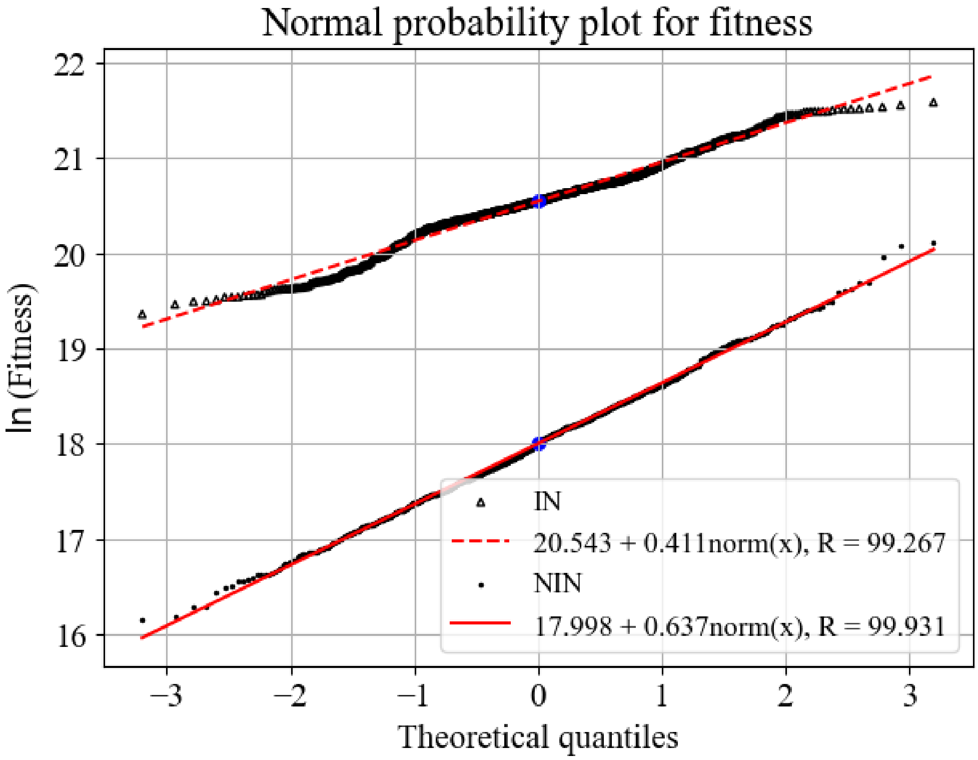
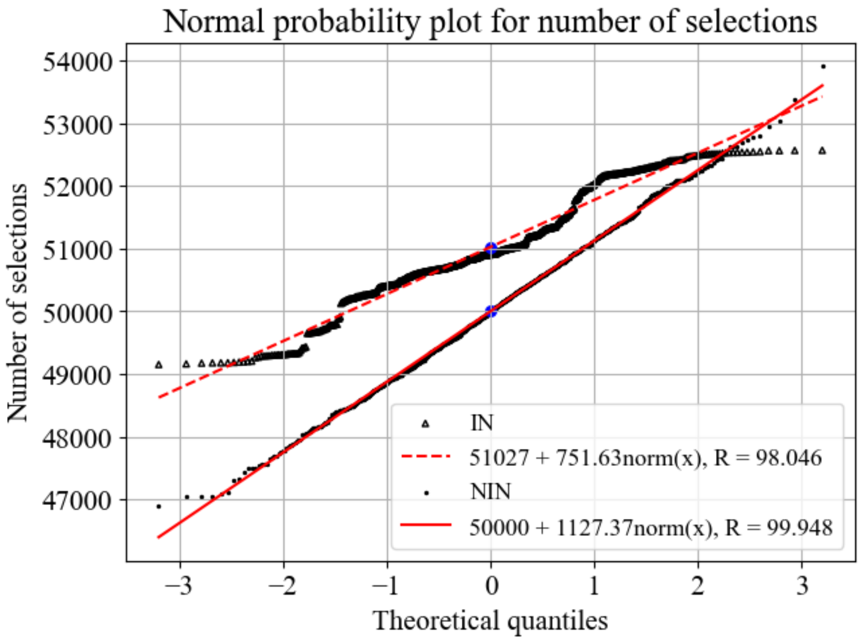


Figure 5.12: Normal probability plot for fitness distributions for IN and NIN cases after 50,000 generations. We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.

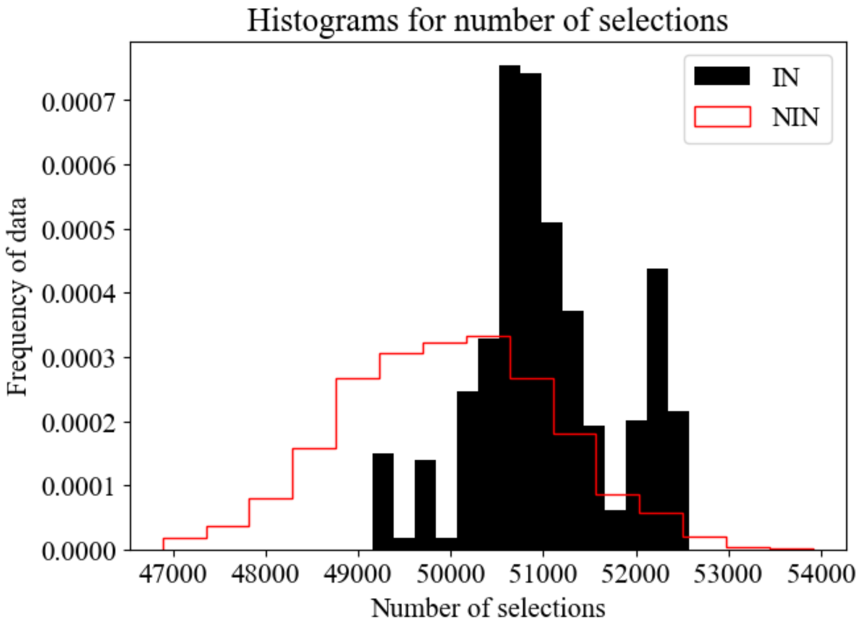
teracting neighbor features is keeping localized colonies related in characteristics. The mean number of selections is higher in the IN case as one would expect as the fitness among the colony grows faster. Figure 5.14 shows the distribution of number of fixations for P colonies. Variance in the NIN case is again larger than the IN case. However, the mean number of fixations in the IN case is slightly less than the NIN case. This is expected since the colony that takes on the fitness of its neighbors also takes on its number of selections and number of mutations, thus disconnecting the two variables. For example, if we assume colony i takes on the fitness value, the number of selections, and the number of fixations of colony $i+1$'s at selection number m when its number of selections equals the number of mutations multiplied by fixation time. However, when colony i takes on colony $i+1$'s values, the new number of selections is slightly more than the number of fixations multiplied by fixation time since colony $i+1$ is waiting to reach its next fixation event. Colony i now has the number of fixations multiplied by the fixation time plus some remainder as its number of selections. It then experiences a reset of its fixation time.

5.3.1 Sensitivity to simulation parameters

To assess the sensitivity of our simulation parameters, we vary each parameter number of colonies P, mutation rate μ, colony size N, and exponential distribution scaling parameter α independently while holding the other three constant. The standardized values for this analysis are number of colonies $P = 1{,}000$, mutation rate $\mu = 10^{-6}$, colony size $N = 10^6$, and exponential distribution scaling parameter $\alpha = 50$. When we vary any parameter, we set the other three parameters to the standardized values. The number of colonies is varied from 0 to 10,000. The mean fitness trend starts to converge, with increasing variance, once the number of colonies P reaches 100 and beyond, see figure 5.15. For colony size N and mutation rate μ, an increase in the number of organisms in a colony or the mutation rate results in a faster growth rate of mean fitness, see figures 5.16 and 5.17. This is expected as a larger colony size N or mutation rate μ provides more mutations per generation in each colony for natural selection. Reduced exponential distribution scaling parameter α values also result in faster

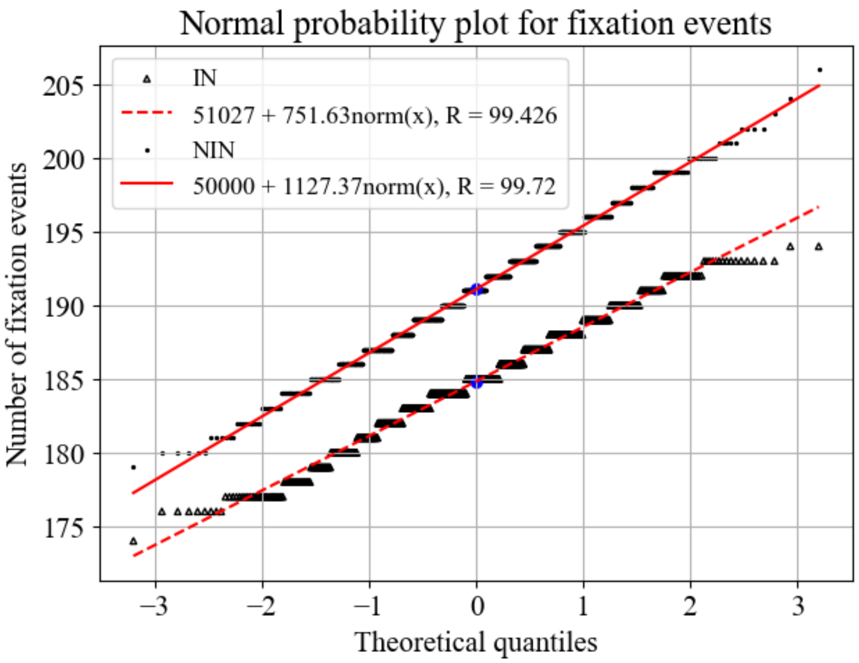


(a)

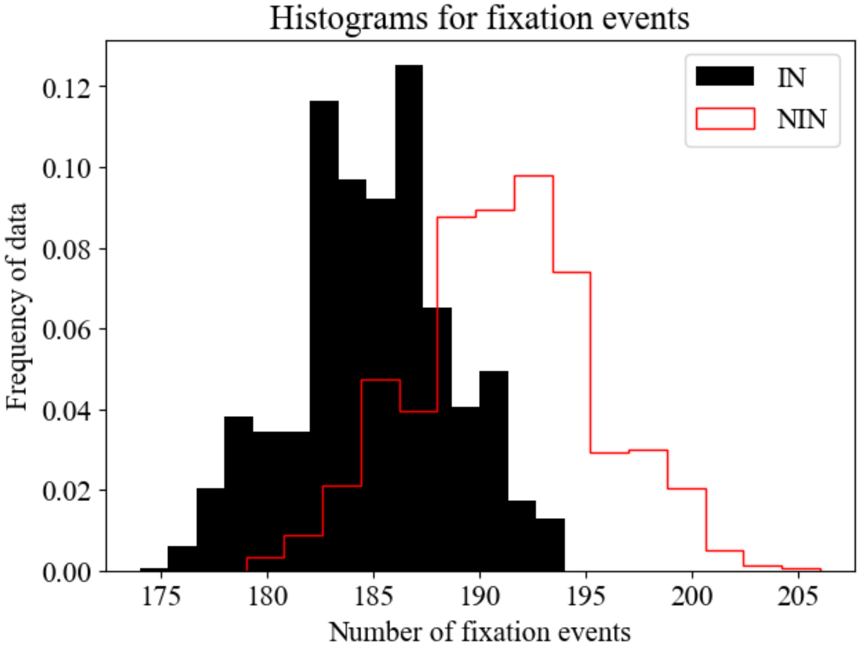


(b)

Figure 5.13: (a) Normal probability plot for the number of selections and (b) the accompanying histograms for both IN and NIN cases. We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.



(a)



(b)

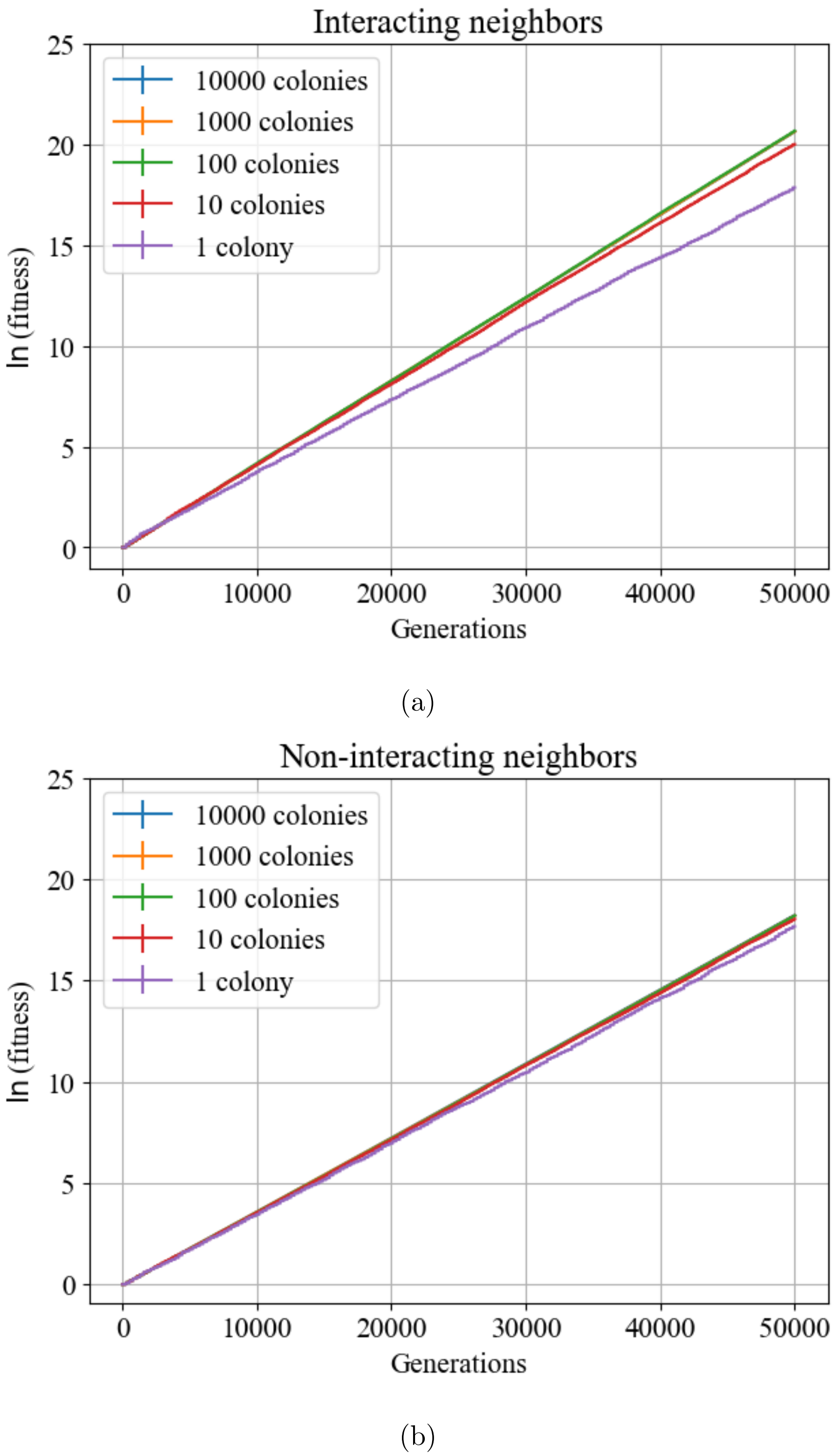
Figure 5.14: (a) Normal probability plot for the number of fixations and (b) the accompanying histograms for both IN and NIN cases. We set colony size $N = 10^6$, mutation rate $\mu = 10^{-6}$, and exponential distribution scaling parameter $\alpha = 50$.

mean fitness growth since the mutational effects are larger on average when α is small, see figure 5.18. In all cases, the behavior of the mean fitness trends are qualitatively similar, displaying exponential mean fitness growth with varying degrees. Moreover, the IN model experiences a higher mean fitness growth than the NIN model under all conditions tested.

5.3.2 LTEE: diminishing returns

As mentioned previously, the LTEE's fitness data for 50,000 generations display diminishing improvement in fitness over time. We are using Wiser *et al.* (2013) [3]'s estimation of $N = 3.3 \times 10^7$ and $\mu = 1.7 \times 10^{-6}$ and the initial exponential distribution scaling parameter $\alpha_{i,0} = 85$ values. Adopting their diminishing constant $g = 6$, we update $\alpha_{i,m} = \alpha_{i,m}(1+gs_{i,m})$ for colony i if selection m marks a fixation event. Since the colonies in our NIN model do not interact and they are treated as homogeneous populations, they are appropriate to model the LTEE experiment. We set the number of colonies P as 1,000. The result of our simulations follows the mean fitness prediction curve established in Wiser *et al.* (2013) [3] albeit at a slower rate, refer to figure 5.19 which shows the trend of fitness over 50,000 generations. Since we are using selection instead of generation time, we approximate that every P selections mark one generation where all colonies have had an opportunity to be selected to undergo one generation.

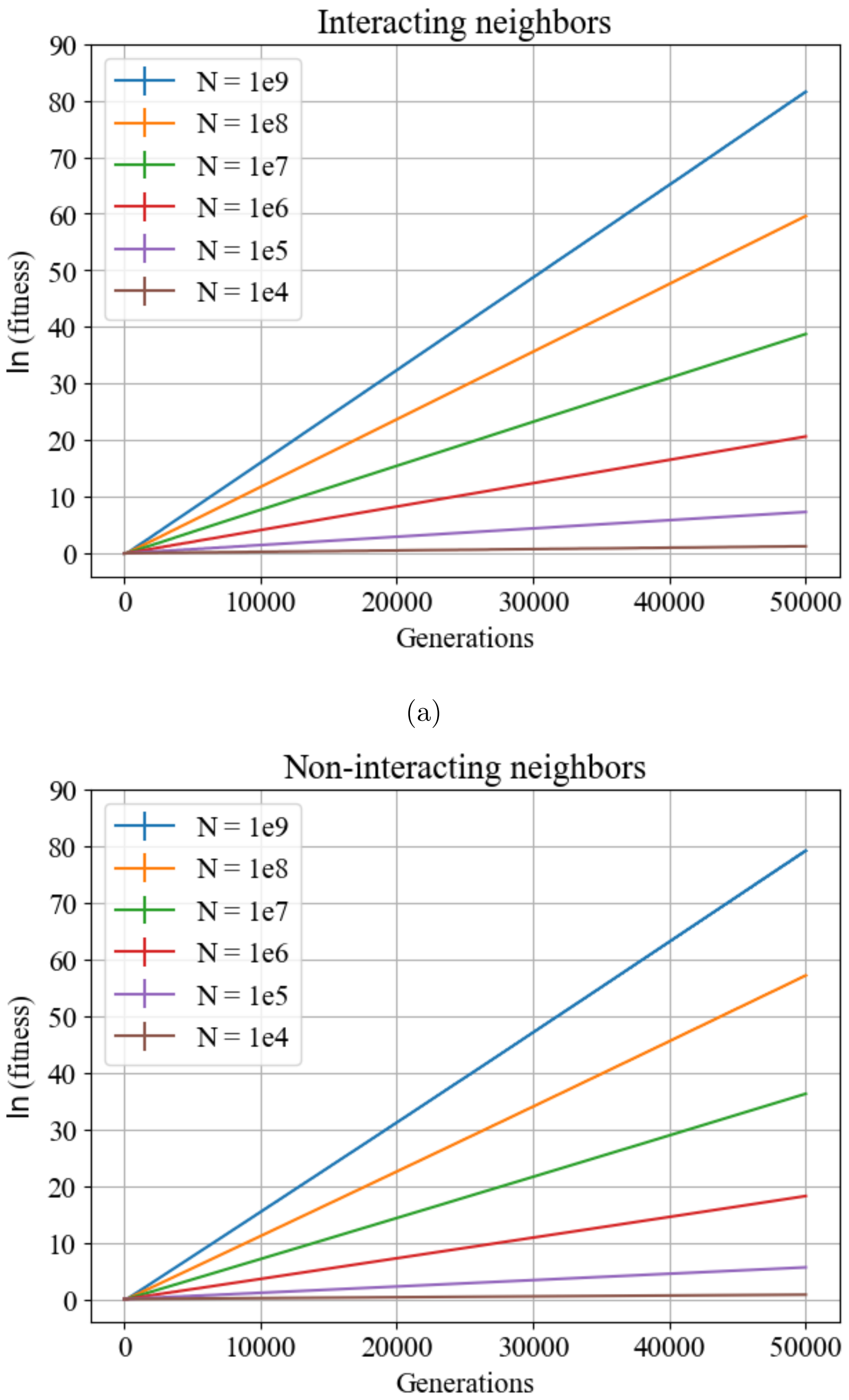
We perform the same simulation for the case where the colonies interact and found the trend to be qualitatively similar as the NIN case. We use the same parameters: number of colonies $P = 1{,}000$, colony size $N = 3.3 \times 10^7$, mutation rat $\mu = 1.7 \times 10^{-6}$, and initial exponential distribution scaling parameter $\alpha_{i,0} = 85$ with $g = 6$. The values of mean fitness for the IN case are slightly higher than the NIN case but lower than predicted quantities by Wiser *et al.* (2013) [3] at all generations, though they all follow the raw LTEE data well. The distributions of the fitness values for both cases after 50,000 generations are approximately normal, see figure 5.22. Figures 5.21a and 5.21b show the raw fitness distributions for the IN and the NIN cases respectively after 50,000 generations. The fitness distribution in the NIN case appears random while the fitness distribution in the IN case exhibits clumped peaks,



(a)

(b)

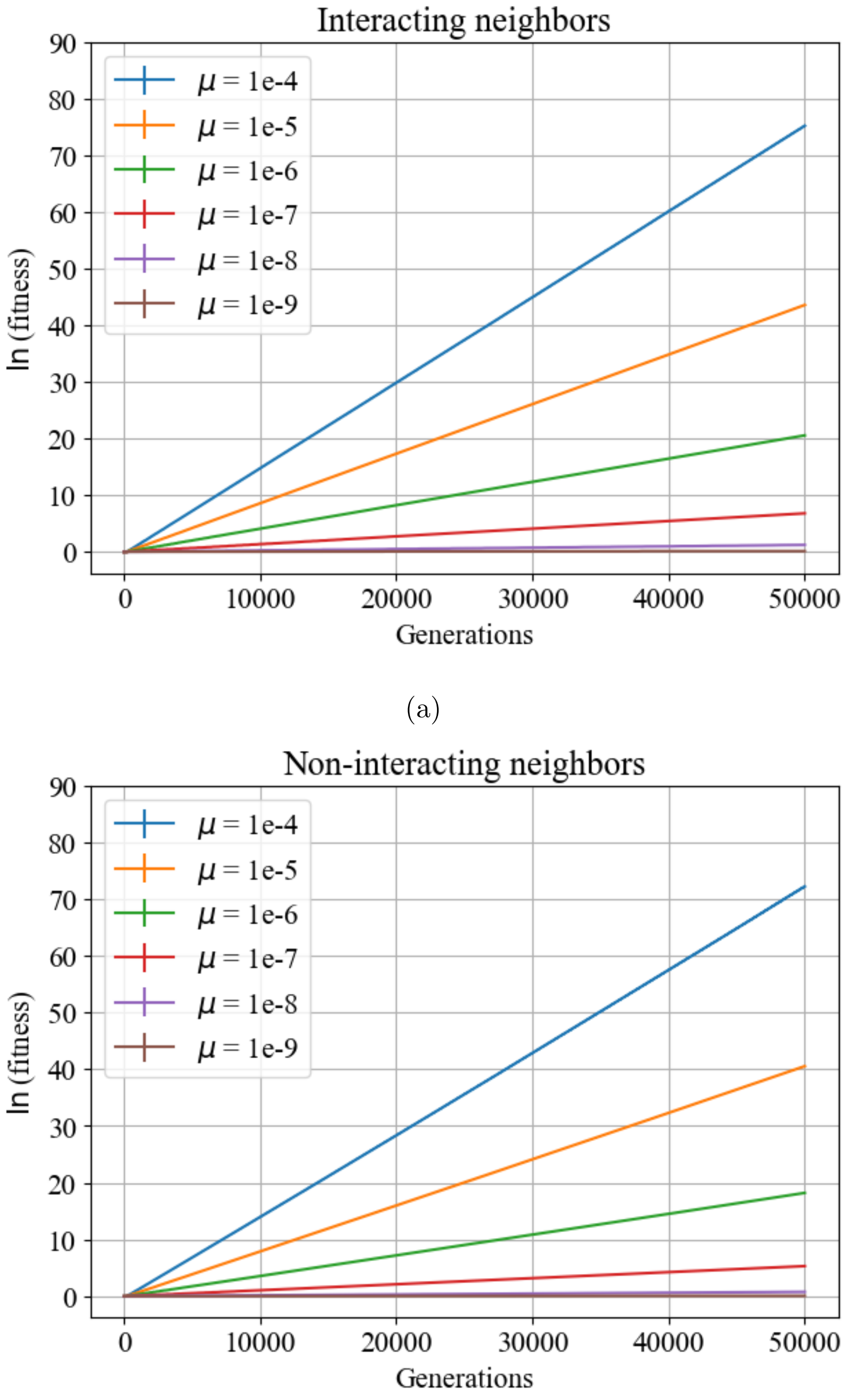
Figure 5.15: Mean natural log fitness for 50,000 generations with P = [1,10,10^2, 10^3, 10^4] for (a) IN case and (b) NIN case.



(a)

(b)

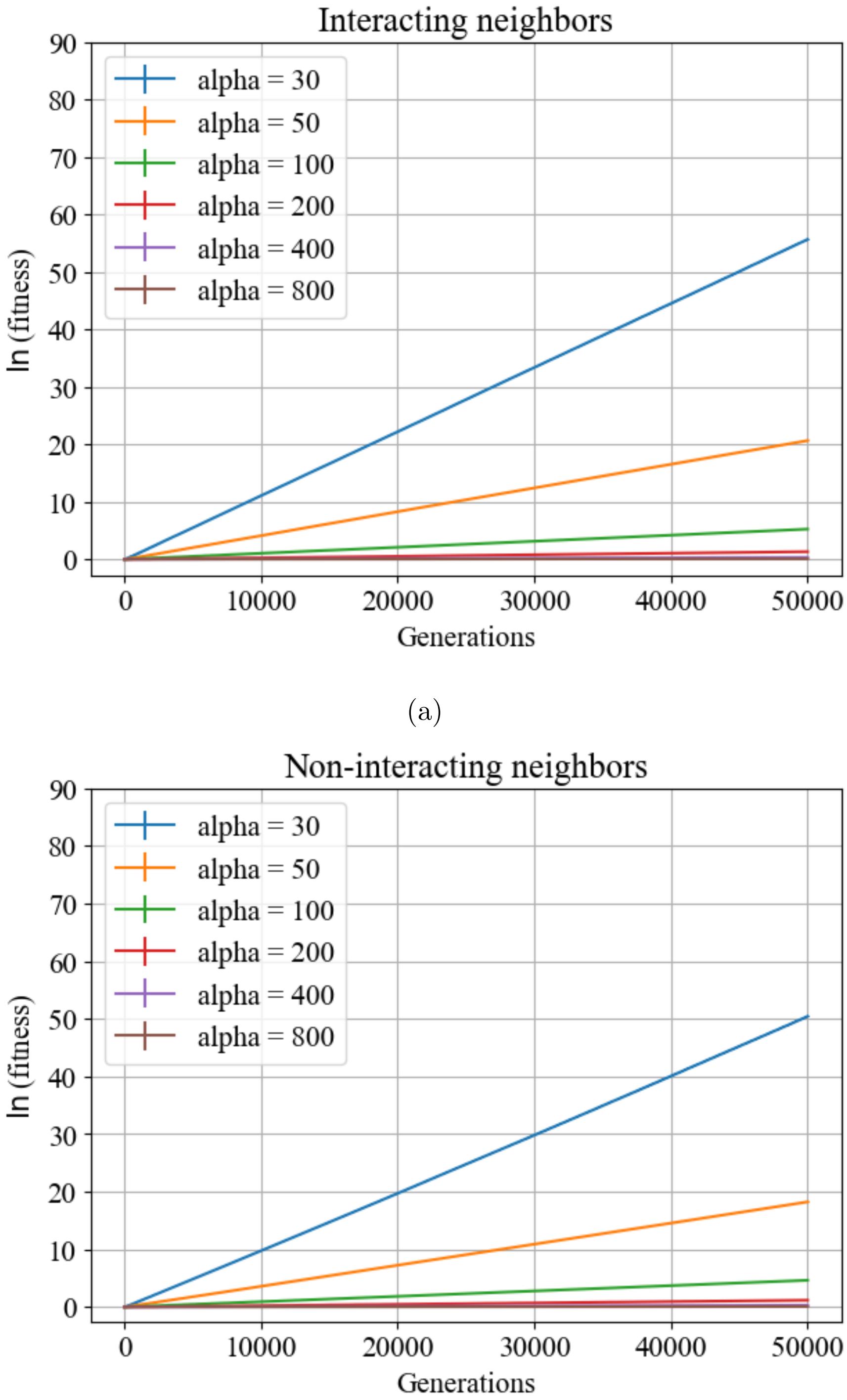
Figure 5.16: Mean natural log fitness for 50,000 generations with colony size $N = [10^4, 10^5, 10^6, 10^7, 10^8, 10^9]$ for (a) IN case and (b) NIN case.



(a)

(b)

Figure 5.17: Mean natural log fitness for 50,000 generations with mutation rate $\mu = [10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}]$ for (a) IN case and (b) NIN case.



(a)

(b)

Figure 5.18: Mean natural log fitness for 50,000 generations with exponential distribution scaling parameter $\alpha = [30, 50, 100, 200, 400, 800]$ for (a) IN case and (b) NIN case.

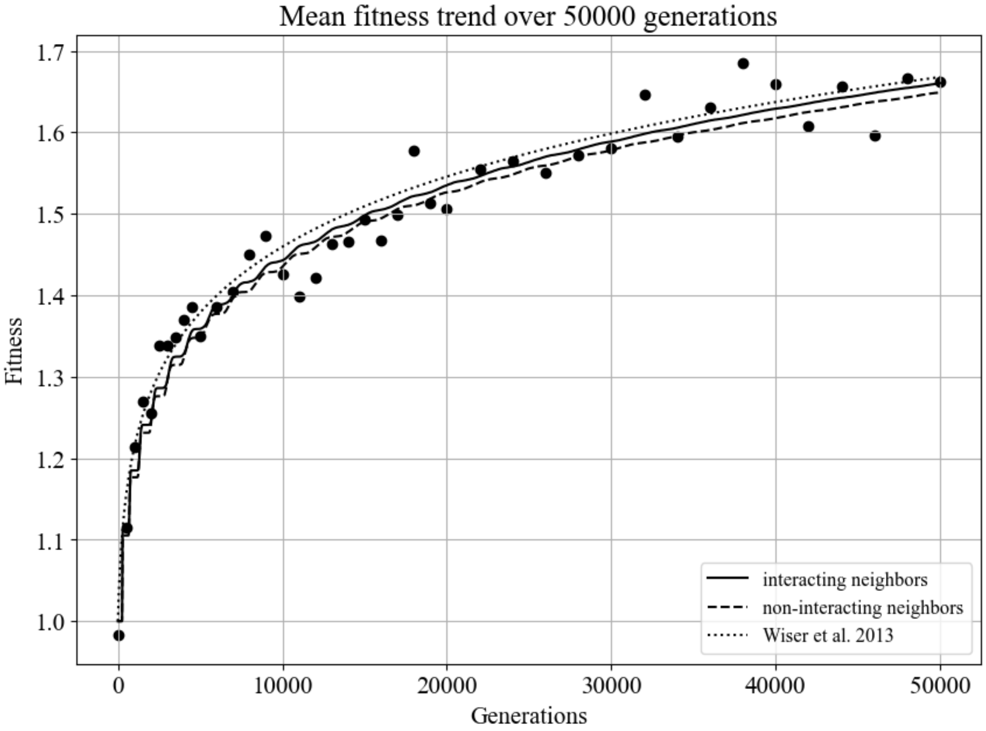


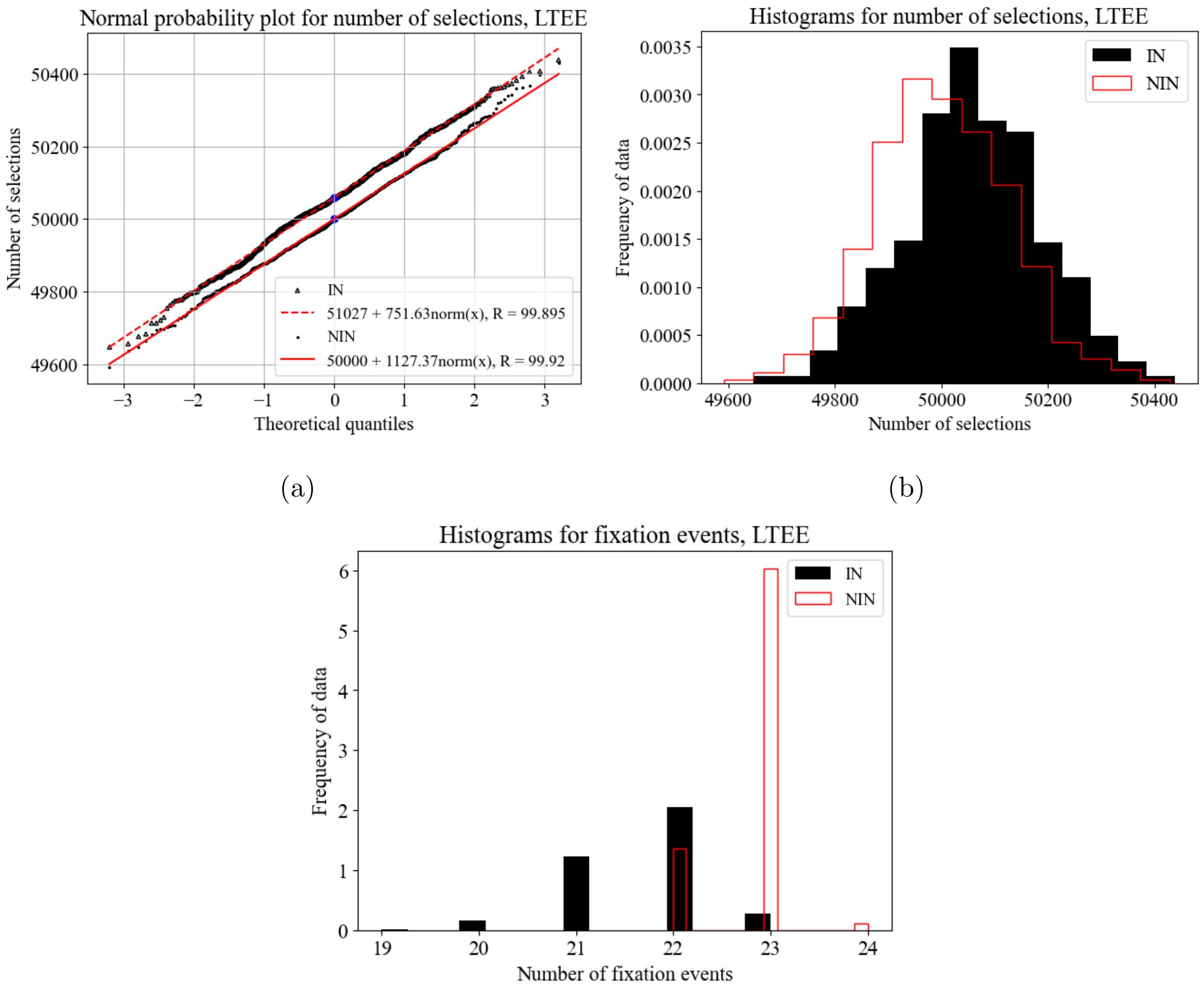
Figure 5.19: Fitness trajectory for mean fitness of 1,000 colonies in IN and NIN models. They both grow slightly slower than Wiser *et al.* (2013) [3] 's prediction but appear to follow the raw LTEE data trend well, which is shown as black dots.

demonstrating the effect of the interactive neighbors. The distributions for the number of selections are also approximately normal for both cases as well, with the IN case having a higher mean, see figure 5.20a. There were less than 25 mutations in both cases, see figure 5.20c. With a large initial exponential distribution scaling parameter $\alpha_{i,0}$ that also increases over time, it takes many generations for a mutation to be fixed in a colony in later generations and the effect of the fixed mutations also declines. Consequently, we see that the rate of fitness improvement decreases over time. The fitness jumps observed in both the NIN and IN cases during the early generations are likely artifacts from our algorithm given that the colonies only experience a fitness increase in discrete measures. In other words, the simulation requires a period of relaxation time to reach equilibrium, at which point the fitness curves appear smooth. We suspect that Wiser *et al.* (2013) [3]'s prediction is slighter higher than the IN case because the interacting neighbors force the colonies to experience diminishing returns in fitness more quickly. We speculate the scattering in the LTEE data points is due to a finite number of samples.

5.4 Discussion

For both NIN and IN models, varying parameters colony size N, mutation rate μ, and exponential distribution scaling parameter $\alpha_{i,m}$ appear to only change the rate of fitness increase and not the behavior of the mean fitness trajectory which follows an exponential curve. NIN experiences slower fitness growth rate than the Gerrish-Lenski model while the IN model has a faster fitness growth rate than both NIN and Gerrish-Lenski. When the diminishing constant g is greater than 0, the initial exponential distribution scaling parameter $\alpha_{i,0}$ decreases over time and fitness improvement rate also declines although fitness continues to increase. Both models are able to predict the trend in fitness well for the LTEE data with IN growing slightly faster than the NIN model, although the former model is not applicable in studying the LTEE populations since they do not interact. Our models do not display substantially different results than the Wiser *et al.* (2013) [3].

The behavior of fitness dynamics over time in our models do not indicate punctuated

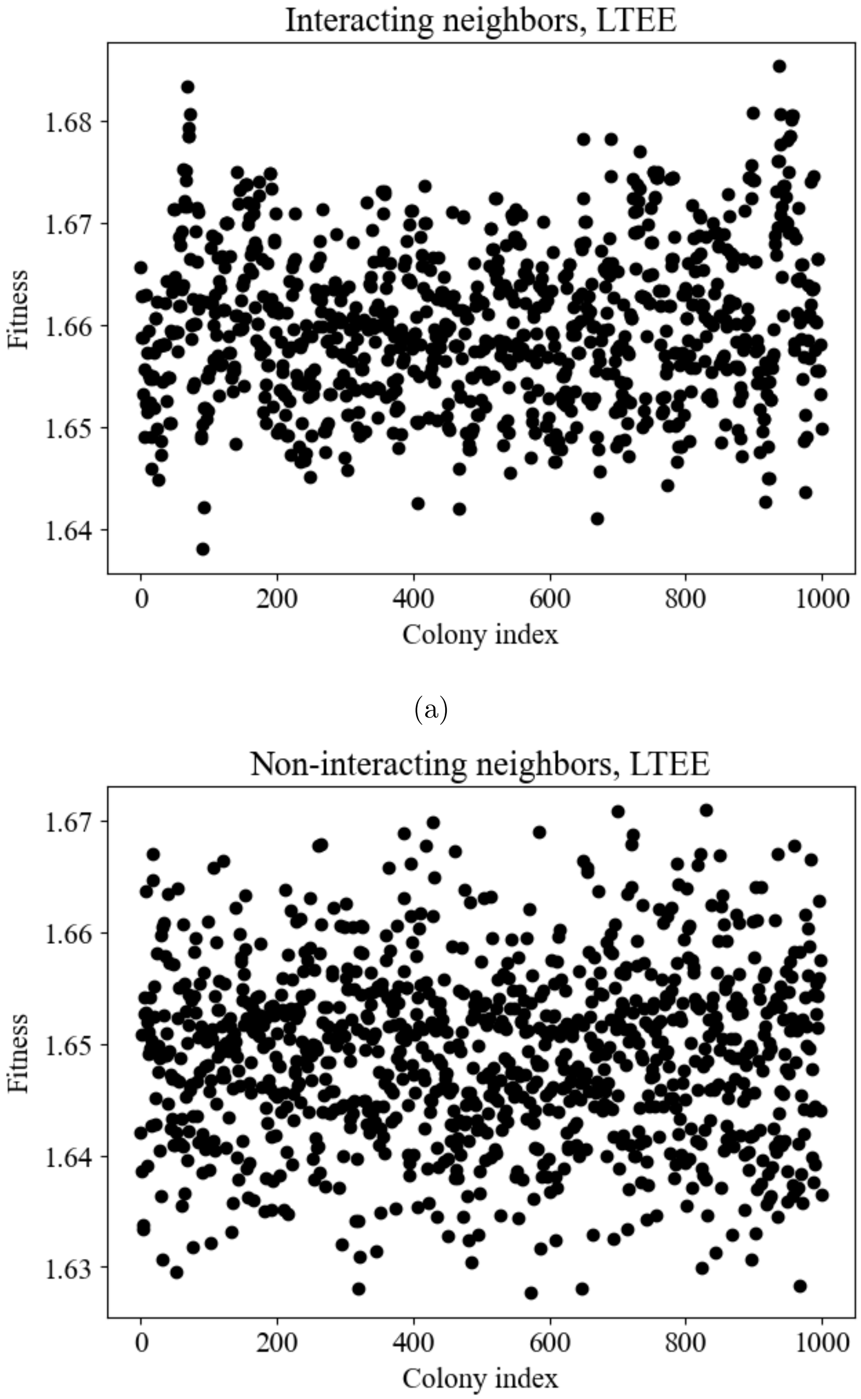


(a)

(b)

(c)

Figure 5.20: (a) Normal probability plots and (b) histograms for the number of selections that colonies experienced after 50,000 generations. (c) Histograms for the number of fixations that colonies accumulated after 50,000 generations. We set number of colonies $P = 1{,}000$. The remaining inputs colony size $N = 3.3 \times 10^7$, mutation rate $\mu = 1.7 \times 10^{-6}$, and initial exponential distribution scaling parameter $\alpha_{i,0} = 85$ with $g = 6$ are Wiser *et al.* (2013) [3]'s parameter approximations for LTEE.



(a)

(b)

Figure 5.21: Fitness distribution for (a) IN case and (b) NIN case after 50,000 generations.

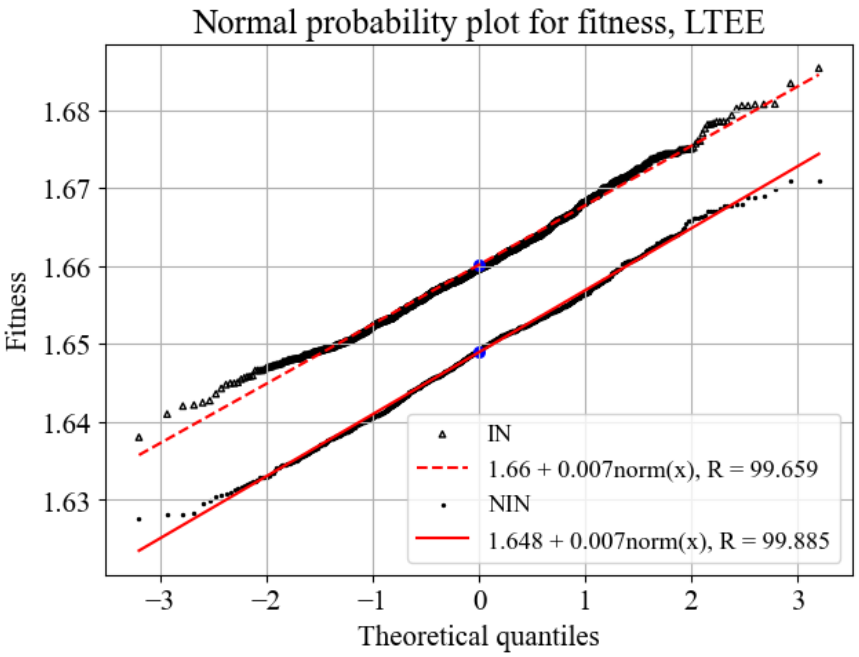


Figure 5.22: Normal probability plot for fitness distributions for IN and NIN cases modeling LTEE after 50,000 generations. The mean fitness in IN is systematically greater than NIN because the fitness of each colony is affected by its neighbors where a more fit neighbor will induce the colony to gain in fitness faster.

equilibrium behavior. In the case where there is no diminishing returns, the fitness trajectory grows exponentially into infinity given the parameters we have considered. When epistasis is present as in the case of modeling the LTEE, one possible interpretation as we have discussed in Chapter 4 in regards of the Wright-Fisher model is this declining fitness rate itself demonstrates punctuated behavior as argued by Gould (2009) regarding the LTEE's cell size data [26]. Although that cannot be excluded, it is inconclusive, and we think further extensions of our model are needed to provide stronger evidence. Two venues we look to build on our model are adapting it into 2-D and 3-D environments and adding additional traits beyond fitness.

In the previous chapter, we observe "jumps" in fitness in the Wright-Fisher model for one population. When considering a large number of colonies as we have done in the smooth-colony dynamics model, where each colony behaves as one population in the Wright-Fisher model, this observation is eliminated in the long run. Further increase in the number of Monte Carlo simulations of our smooth-colony dynamics model of both (IN) and (NIN) cases could reduce the uncertainties in the simulation results. Our model captures the structure of colonies grown on surfaces which can be hundreds on a plate [61]. Since a collection of colonies is more likely to mimic a natural occurring system, this negates the potential for observing punctuated patterns in evolution for similar biological structure.

CHAPTER 6

Summary and future work

In this dissertation, we have investigated the possible emergence of punctuated patterns through mathematical modelling of simple biological systems. We looked at mathematical and computational modelling using three different units: species, colonies, and organisms to assess the hierarchical level necessary to observe punctuated patterns if they exist. We want to highlight the contribution of our work in deploying two models for our investigation. We developed a Monte Carlo Wright-Fisher model that we have not seen utilized in the study of punctuated equilibrium. We constructed and introduced the new smooth-colony dynamics model that combines elements of discrete populations and overlapping generational times. Additionally, we provided an approximation of the Gerrish-Lenski mathematical model that allows for faster computation.

Punctuated equilibrium has been proposed to explain observations in the paleontological record wherein species stay in stasis for the majority of their lifetimes with sporadic bursts of change that mark speciation events [1]. Although the Bak-Sneppen model, proposed by Bak and Sneppen (1993) [5], shows the behavior of burst of activities follows by stasis at the species level, no relevant biological system has been identified to provide evidence to its applicability [26]. At the population level, we extend the Wright-Fisher model and develop the smooth-colony dynamics model (with two variations) based on the LTEE data in order to understand the simplest biological system at the generational scale. We find that although these models may provide some insights into the evolutionary dynamics, they are limited in their ability to resolve the question of punctuated patterns in a population due to their conflicting behavior. All three models are able to fit the LTEE data quantitatively well. However, the Wright-Fisher shows step-like behavior while the two versions of the smooth-colony dynamics model do not. We suggest that the Wright-Fisher results are more applicable in the laboratory setting while the smooth-colony dynamics model captures evolutionary dynamics of bacteria in a more natural environment. Another possibility is the discretization of generations in the Wright-Fisher may induce more step-like behavior. Further work is required to resolve the question of which regime populations of asexual single organisms operate under. We aim to extend this work in the future by building the smooth-colony dynamics model into a 3-D model and accounting for additional traits that will allow it to be more biologically complex.

At the species level, we adapt the highly-efficient Newman-Phoenix hierarchical data structure to the Bak-Sneppen model. We provide an overview and further mathematical analysis of their results. This model observes punctuated behavior at the species level where the species experiences fitness transitions intermittently with periods of inactivity interrupted by relatively rapid periods of changes. However, this model lacks any relevant data to support its claims. Thus, their results are limited for providing insight to the question of punctuated behavior observed in the fossil record [26].

The resolution of observations at the species level is often limited due to data sampling issues. Patterns at the species level are also complicated and impacted by a wide range of variables that operate on varying timescales. Thus, we pursue a simplified system that evolves in a shorter timescale within which data are accessible to constrain our model. The Long-Term Evolution Experiment (LTEE) on *E. coli* by Lenski and his colleagues serves as our motivation to study models in the organismic and clonal levels of asexually reproducing single cells. Although Lenski and Travisano (1994) suggests there are distinct jumps in fitness in the first 2,000 generations [20], we find this description inconclusive based on the discretized data. Another suggestion of punctuated patterns highlighted by Gould (2009) is the the diminishing growth rate in cellular size after the initial few thousand generations [26]. Again, we find this interpretation not conclusive since the populations are continuously experiencing changes genetically and morphologically, albeit at a slower rate.

We employ the Wright-Fisher model using a Monte Carlo approach to model the LTEE fitness data of 50,000 generations. The Wright-Fisher assumes homogeneous interaction in a population of size N and non-overlapping generations. We extend this model to include mutation and natural selection by introducing a mutation rate μ per generation and a distribution of mutational effects. We use a Monte Carlo approach to simulate our adapted Wright-Fisher model. Using Wiser *et al.* (2013) [3]'s estimations for our N and μ parameters and a commonly used exponential distribution for modeling bacteria's beneficial mutational effects, our Wright-Fisher model is able to reproduce the LTEE's data trend close in quantitative agreement over many thousand generations. We do observe jumps in fitness in the simulation results of our model. This is due clonal interference where beneficial mutations are fixed in a population sequentially since one advantageous genotype cannot exchange genetic material with another advantageous genotype [13]. The new beneficial genotype needs to reach a relatively high frequency for the fitness improvement to be observed. *E. coli* in favorable laboratory conditions can undergo binary fission every 20 minutes. In nature, many bacteria species have generation times ranging from under 30 minutes to over 100 hours [62]. Although generation times for bacteria are extremely short relative to geological time, these results demonstrate that punctuated patterns can be observed at the population level within the generational scale.

We next use a colony unit to model evolutionary dynamics of an asexually reproducing population in our smooth-colony dynamics model. By dividing the population into colonies, we assume homogeneous interactions in the sub-population and local interactions between the sub-populations as single units. This assumption is more realistic than the Wright-Fisher model's for naturally occurring populations. We estimate the mutational effects at the clonal level using the Gerrish-Lenski mathematical model of clonal interference [15]. We build two versions of this model, one without interacting neighbors and one with. In the first version without interacting neighbors, each colony evolves independently. In the second case, each colony's fitness is impacted by its neighbors wherein the less fit colony adopts the higher fitness of its neighbor's. Both of our smooth-colony dynamics models also robustly fit the LTEE data quantitatively. However, the population mean fitness in the interacting neighbors model grows faster than the non-interacting neighbors model. We cannot resolve the generational jumps observed in the Wright-Fisher model as our explicit assumption treats each fitness improvement event as discrete. Thus, the jumps in fitness observed in early generations in both models are likely artifacts from our modeling treatment. However, the trend for population fitness becomes smoother over time as colonies' fixation times overlap in our simulations. In these two models, we do not observe punctuated behavior under any combination of parameters used for our simulations in absence of epistasis. Given that the fitness growth is exponential, except when explicitly adding a diminishing growth term, it is unlikely this model will yield any behavior deviating from this growth behavior in a one trait model.

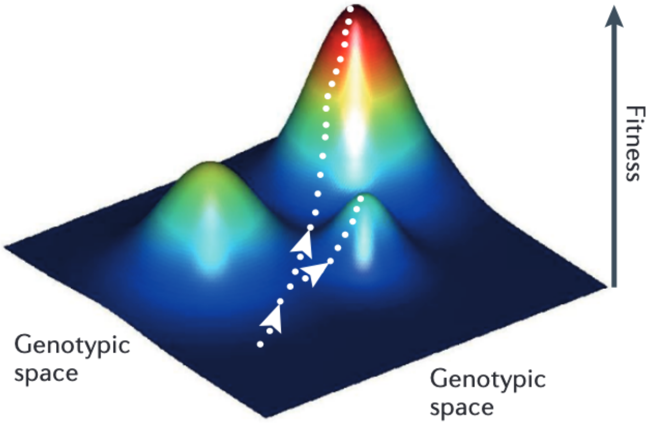


Figure 6.1: A graphical representation of fitness landscape with the horizontal axes representing the genotypic space and the vertical axis representing the organism's fitness. The dotted white paths show two potential evolutionary pathways. Note the limitation of this visualization for high dimensional system with more than two traits. This was taken from De Visser and Krug (2014) [63].

Although we have observed punctuated patterns using the Wright-Fisher model, the more naturally realistic smooth-colony dynamics model fails to detect this trend. For our future work, we will look to extend our smooth-colony dynamics model to 2-D and 3-D environment and include different traits that are less aggregated than the fitness measurement. We speculate that the observation of punctuated patterns, or episodic change, will more likely be exhibited in a scenario where multiple traits evolve as opposed to single trait evolution. In this scenario, we will assume two cases where these traits evolve independently, i.e. the evolution of one trait does not correlate in the change in another, and dependently. To visualize this relationship between genotypic makeup and fitness outcome, Sewall Wright developed the fitness landscape in 1932 [64], see figure 6.1 for a representation. Giorgio Parisi has recently received the 2021 Nobel Prize in Physics for his work in **complexity** theory that also extends to fitness landscapes, visualized through "corrugated landscapes" with valleys and mountains, exhibiting the relation between genetic makeup and fitness distribution (Giorgio Parisi – Nobel Prize lecture [65]). The trade-offs between different traits in a high dimensional fitness landscape may become more important than the evolution of any single trait. In terms of the fitness landscape's maxima and minima, when we go from one dimension to two or three, we can have isolated maxima and minima where more activity is generated in one trait at one time and in another at a different time. Although our simplified single-trait model demonstrates that a population residing in a natural environment is not likely to exhibit punctuated patterns, the modeling of evolution involving multiple traits can potentially show more complex behavior. This is one outcome we may observe in the extension of our work.

APPENDIX A

Acronyms and Mathematical symbols

All Chapters

α	scaling parameter for the exponential distribution
N	population size for the Wright-Fisher model and colony size for the smooth-colony dynamics model
μ	beneficial mutation rate
t	generation number
s	selection coefficient per beneficial mutation
$w_{i,t}$	fitness for i'th organism at generation t in the Wright-Fisher model
$\overline{W}(t)$	population mean fitness as a function of generation t
PE	punctuated equilibrium

Chapter 5

$s_{\mathbf{f}}$	fixed selection coefficient per colony or group of organisms
m	selection number in the smooth-colony dynamics model
M	total number of selections in the smooth-colony dynamics model
P	number of colonies in the smooth-colony dynamics model

$\tau_{i,m}$ a continuous random variable that indicates time to next doubling for i'th colony during selection m in the smooth-colony dynamics model

d shift parameter in the Rayleigh distribution

σ scaling parameter in the Rayleigh distribution

IN interacting neighbors

NIN non-interacting neighbors

APPENDIX B

Glossary

- **colony:** although bacterial colonies are not well defined [40], in this dissertation we define it as a group of cells having the same ancestral organism. They grow in one place physically, usually on a surface.

- **complexity:** systems with many interdependent or interacting components whose behavior is highly sensitive to their initial conditions.

- **cumulative distribution function (CDF):** a function that when evaluated at x gives the probability that a random variable can take on value less than or equal to x.

- **diminishing-returns epistasis:** diminishing improvement from additional mutations due to negative interaction between mutations.

- **epistasis:** the interaction of genes where the effects of the same mutation can vary depending on the genotype it occurs in [66].

- ***Escherichia coli (E. coli):*** a type of bacteria that has been used as a model organism in studying prokaryotic cells. They are practical to be handled in a laboratory setting and have rapid growth, thus making them ideally suitable for biological investigations.

- **fitness:** the expected number of offspring an individual will produce. For binary fission, the LTEE estimates this value by counting the number of times a population doubles relative to the ancestral population.

- **fitness landscape:** the mapping of an organism's genotype and its reproductive success, or *fitness*, which can display peaks and valleys. The horizontal axes represent the space of possible genotypes and the height represents the fitness of the organism.
- **fixation:** when a mutated genotype becomes the only genotype in the population
- **gene:** a unit of heredity that determines a physical characteristic of an organism. There are multiple definitions of genes and the employment of each depends on the topic of study. For example, morphology focuses on the study of physical appearances of an organism and may treat a feature like beak size as a targeted gene. We will use the definition that a gene is a sequence of genetic nucleotides in DNA that encodes for the production of a particular protein.
- **genome:** complete set of genetic material in an organism.
- **genotype:** the complete genetic makeup of an organism.
- **horizontal gene transfer:** the non-reproductive transfer of genetic material from one organism to another.
- **hypermutability:** an increase in the mutation rate relative to the ancestral mutation rate
- **morphological change:** changes to the outward appearance of an organism.
- **mutation:** changes that affect an organism's genetic material. In cellular organism, that usually refers to DNA and RNA.
- **mutation rate:** the probability that an organism experiences a mutation during its reproduction in one generation.
- **nucleotide:** a building block of nucleic acid (DNA and RNA) consisting of a sugar molecule, a phosphate group, and a nitrogen-containing base. The primary bases are adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U). DNA uses the

first four and RNA replaces thymine with uracil. A is normally paired with T — or U in RNA — and C is normally paired with G.

- **point mutation:** the smallest form of genetic mutation where a single nucleobase pair is altered.

- **probability density function (PDF)**: a function that maps the likelihood that a random variable will take on a value.

- **probability mass function (PMF)**: same as PDF but for discrete values.

- **punctuated equilibrium:** a theory proposed by Niles Eldredge and Stephen Gould that species experience bursts of morphological change in a short period of time then stay in stasis, with little to no change, for the majority of their existence.

- **replication:** [DNA replication] the process of making two identical copies of DNA from the original DNA molecule. This is critical for cell division. A mutation can occur during this process which can introduce new variants in the population, which can be advantageous, disadvantageous, or neutral. Mutations that occur during somatic cell (regular body cell) division do not get passed down to the next generation. Mutations that occur in germ cells can be passed down.

- **ring species:** an arrangement of spatially connected populations where neighboring populations, geographically close, are able to interbreed but at least two populations, occupying the boundary of the series, are too distant to interbreed.

- **selection coefficient:** the difference in relative fitness between two genotypes. In the context of our model, it's the advantage in fitness improvement of the mutation.

- **self-organized criticality:** the behavior of large dynamical systems that tends towards an organization of scale-invariant characteristic, either temporally or spatialy, without the need to tune the related parameters to precise values.

- **speciation:** formation of a new species when a population of organisms distinct itself from an existing species by developing different characteristics. A species is sometime defined as a group of similar organisms that can exchange genes. There are different definitions of species in the biology community which can often complicate the study of speciation.

- **trait:** characteristic of an organism that is controlled by its genetic makeup and may be influenced by the environment. The definition is further refined based on the topic of study. Traits can be controlled by one or more genes. Some examples are eye color, tree height, bird vocalization, etc.

- **transcription:** the process of copying a segment of DNA into RNA which may contain genetic sequence to make proteins. A mistake can occur during this process which can potentially cause a gene to be incorrectly expressed. For example, adenine (A) may be transcribed as guanine (G) instead of uracil (U).

- **wild type:** Although this is generally used to describe traits and or specific genes [67], in this work, we will refer to the wild type genotype as the most common genotype in the population, i.e. in the wild. Mutations can change this genotype into a mutant genotype.

APPENDIX C

Additional mathematical issues

C.1 Mathematical concepts

In this section, we are providing a brief outline of some mathematical concepts that are applied in the dissertation. This is intended as an overview, and a comprehensive treatment of these topics can be found in the appropriate mathematical and/or computer science textbooks. We also provide a more extensive description of the Newman-Phoenix hierarchical data structure and the smooth-colony dynamics model's algorithm.

C.1.1 Inverse transform sampling

Inverse transform sampling is a method to draw random numbers from any probability distribution. We utilize this method to draw random numbers for both the Monte Carlo Wright-Fisher model in Chapter 4 and both versions of the the smooth-colony dynamics model in Chapter 5. Please refer to Donald Knuth's classic book *Art of computer programming, volume 2: Seminumerical algorithms* (2014) as a reference [68].

Let $F(x)$ denotes some cumulative distribution function (CDF) and $F^{-1}(x)$ be its inverse. We can define a random variable X such that $X = F^{-1}(U)$ where U is uniformly distributed between the interval $[0, 1)$. Then X is distributed according to the CDF $F(x)$. We are not providing a proof here but an example. In Chapter Four, we want to draw random variables that follow the Rayleigh distribution with a CDF of the form $1 - e^{-x^2/2\sigma^2}$ for some constant $\sigma > 0$ and for $0 \leq x < \infty$. Let $F(x) = 1 - e^{-x^2/2\sigma^2}$ and $y = F(x)$, then $x = F^{-1}(y) = \sqrt{-2\sigma^2 \ln(1 - y)}$. Then random variable $X = \sqrt{-2\sigma^2 \ln(1 - U)}$, where $U \sim$

$uniform(0,1)$. Note $(1-U)$ is also uniform thus we can replace it by U. Computationally, we can draw a psuedo-random number U that's uniformly distributed from $(0,1)$ then compute $X = \sqrt{-2\sigma^2 \ln(U)}$.

C.1.2 Logistic function

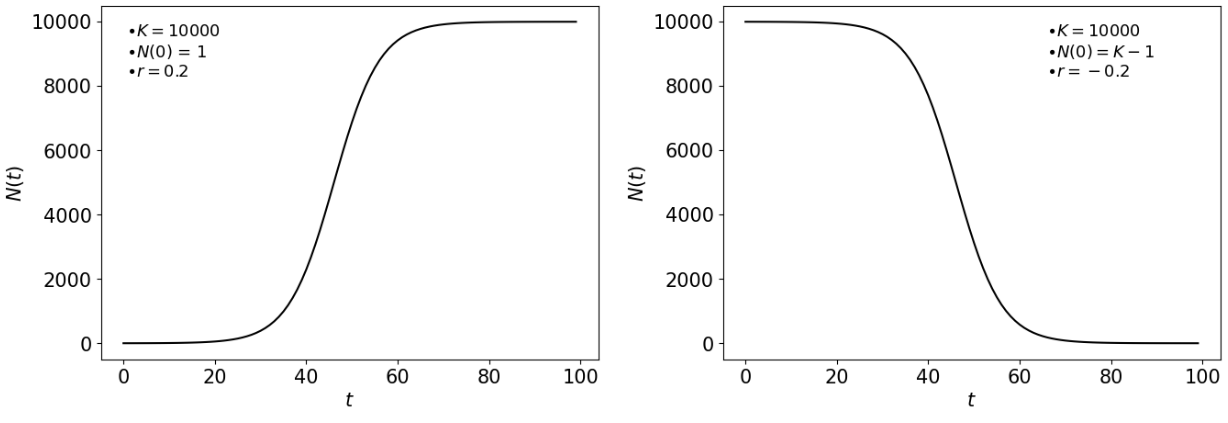
The logistic function shows initial exponential growth which slows down as the population reaches a saturation point. This is relevant for the discussion of clonal interference in Chapter 2. The logistic function is covered by many textbooks including James Steward's *Calculus: early transcendentals* (2008) edition [69] and William I. Newman's *Mathematical methods for geophysics and space physics* (2016) [70]. The logistic function is derived from a simple differential equation of the form:

$$\frac{dN(t)}{dt} = rN(1 - \frac{N(t)}{K}) \tag{C.1}$$

where K is some constant greater than $N(0)$ and r is some rate that determines if $N(t)$ is increasing or decreasing. The solution to the differential equation above is:

$$N(t) = \frac{K}{1 + (K/N(0) - 1)e^{-rt}} \tag{C.2}$$

K marks the saturation point of the equation, such that as $N(t)$ approaches K, its growth



(a)

(b)

Figure C.1: (a) Logistic function with $r > 0$. (b) Logistic function with $r < 0$.

(decline) rate reduces. For $r > 0$, the hallmark logistic curve is an S-shape with $N(t)$ initially experiencing exponential growth follows by diminishing growth. For $r < 0$, $N(t)$'s declining trend follows a similar pattern but in reverse. Figures C.1a and C.1b provide two examples of the logistic function.

C.1.3 Multinomial distribution

In this section, we discuss the multinomial distribution which is utilized in the adapted Wright-Fisher model in Chapter 4. The multinomial distribution is a generalized version of the binomial distribution where there are multiple outcomes instead of two. The following description has been adapted from Park and Krug 2007, please refer to the supplement of their paper for a more thorough treatment [23]. Feller's *An Introduction to Probability Theory and Its Applications* (1968) also covers this topic [36]. Let N be the number of times an experiment is performed and K be the number of possible outcomes each experiment can have. Let n_i denote the number of times outcome i is observed and p_i is the probability a given experiment results in outcome i, where $i \in [1, K]$. Then we have the following multinomial distribution:

$$P(n_1, n_2, ...n_K) = N! \prod_{i=1}^{K} \frac{p_i^{n_i}}{n_i!} \tag{C.3}$$

where $\sum_{i=1}^{K} n_i = N$.

C.2 Clonal interference

In this section, we derive the equations $dP/dt = rP(1 - P)$ and $\overline{W(t)} = 1 + sP(t)$. which is relevant for the discussion of clonal interference in Chapter 2.

Let the initial genotype be a and the mutated genotype be A. We start the clock when the mutated genotype first emerges thus the initial number of A copies is 1 and thus the number of a copies is $N - 1$. Let $P_a(t)$ and $P_A(t)$ be the numbers of a individuals and A individuals, respectively and $P(t)$ be total population at time t. Let m_a and m_A be the Malthusian parameters for each sub-population. Thus we have:

$$P_A(t) = (1)e^{m_A t}$$

$$P_a(t) = (N-1)e^{m_a t}$$

$$P = P_a(t) + P_A(t)$$

Let $p_a(t)$ and $p_A(t)$ be the percentage of a and A copies in the total population. Although we are allowing for both sub-populations to grow indefinitely, we can curtail the size of the population to any value given $p_a(t)$ and $p_A(t)$, which can be calculated as:

$$p_A(t) = \frac{P_A}{P_A+P_a} = \frac{e^{m_A t}}{e^{m_A t}+(N-1)e^{m_a t}}$$

$$p_a(t) = 1 - p_A(t) = \frac{(N-1)e^{m_a t}}{e^{m_A t}+(N-1)e^{m_a t}}$$

Then:

$$\frac{dp_A}{dt} = \frac{m_A e^{m_A t}}{e^{m_A t}+(N-1)e^{m_a t}} - e^{m_A t}\left[\frac{m_A e^{m_A t}+m_a(N-1)e^{m_a t}}{(e^{m_A t}+(N-1)e^{m_a t})^2}\right]$$

$$\frac{dp_A}{dt} = m_A p_A(t) - p_A(t)[m_A p_A(t) + m_a(1-p_A(t))]$$

$$\frac{dp_A}{dt} = p_A(t)[m_A - m_A p_A(t) - m_a + m_a p_A(t)]$$

$$\frac{dp_A}{dt} = p_A(t)[(m_A - m_a) - p_A(t)(m_A - m_a)]$$

$$\frac{dp_A}{dt} = (m_A - m_a)p_A(t)[1-p_A(t)]$$

Let $m_A - m_a$ be r, we recover equation (2.2): $dp_A/dt = rp_A(1-p_A)$. Since relative fitness is defined as $W_{Aa} = m_A/m_a$ with a be the ancestral genotype, we have the mean fitness of the population as:

$$\overline{W(t)} = \frac{m_a}{m_a}p_a(t) + \frac{m_A}{m_a}p_A(t) = (1-p_A(t)) + W_{Aa}p_A(t)$$

$$\overline{W(t)} = 1 + p_A(t)(W_{Aa}-1) = 1 + s_{Aa}p_A(t)$$

where s_{Aa} is the selection coefficient of the mutated genotype, recovering equation (2.4).

C.3 The Newman-Phoenix hierarchical data structure

In this section, we outline the Newman-Phoenix hierarchical data storage that is relevant to our smooth-colony dynamics model and in reproducing the Bak-Sneppen model.

Given an array of N elements, we first deal with the case where $\log_2 N$ is an integer. The first array (denote array 0), contains the indices of the given data with N elements, or simply $\{1, 2, ...N\}$. The next array, array 1, contains indices of the smaller elements after comparing pairs of consecutive elements with each element being compared once (i.e. compare element 1 with element 2, element 3 with element 4, etc. from layer 0). For example, if we compare the values the indices in element 1 and element 2 of layer 0 point to, the index of the smaller value is stored in element 1 of array 1. Thus, array 1 has $N/2$ elements. We do the same for array 2 by comparing the values the elements in array 1 point to. Thus, the last array, array $\log_2 N$, has one element, holding the index of the smallest value of the given data. There are $\log_2 N$ arrays in addition to array 0. We deal with the case where $\log_2 N$ is not an integer later on.

We now store the data in the set up we just discussed. Let the given array be called A. The method in storing A with N elements, where $\log_2 N$ is an integer, is as follows:

1. Create array 0 with indices of the values in A resulting in layer $0 = \{1, 2, ..., N\}$
2. Create $\log_2 N$ arrays with array k having $N/2^k$ elements, where $k = \{1, 2, ..., \log_2 N\}$
3. For element i in array k:
   - Compare the values the indices in elements $2i - 1$ and $2i$ in array $k - 1$ point to in A, store the index of the smaller value in element i of array k.
4. Repeat step 3 for all elements of all $\log_2 N$ arrays.

We are now left with $\log_2 N$ arrays with the last array containing the index of the smallest value in A. This storing algorithm takes $\sum_{i=1}^{\log_2 N} \frac{N}{2^i} = N-1$ operations or $\mathcal{O}(N)$. However, we only have to do this once and locating the minimum value is $\mathcal{O}(1)$. Changing the minimum value and update the storing configuration takes $\mathcal{O}(\log_2 N)$ as we only have to make one comparison for each array with every new single value update in A.

For cases where $\log_2 N$ is not an integer, we first pad array A as follows:

1. Let k be the smallest integer greater than $\log_2 N$
2. Create array A^* with 2^k elements, where the first N elements contains A and the remaining $2^k - N$ are values outside of range. For example, in the Bak-Sneppen model, the barriers range from 0 to 1 so we chose to pad the new array with value 10.

Now we will store the data for array A^* the same as in the algorithm above.

C.4 The smooth-colony dynamics algorithm

In this section, we outline the algorithm used to simulate our model. A general conception for this algorithm proceeds as follows. For P colonies starting with the same initial fitness, we track the fitness evolution of each colony. Each colony carries two timers, one that determines its next binary fission event and a second that determines its next fixation event. During each selection, the colony with the smallest time to binary fission undergoes one doubling event. Once a colony's time to fixation event is reached, it experiences a fitness improvement event through a mutation fixation. The simulation is run until a predetermined number of selections is reached. Below is an computational description of our algorithm.

Given mutation rate μ and colony size N, we numerically obtain $1/k_{\text{sub\_rate}}(\mu, N)$, $k_\sigma(\mu, P)$, and $k_d(\mu, P)$. Starting with P colonies arranged in one-dimension, each assigned an index i ranging from 1 to P, we create the following arrays:

- Array F with $F_{i,m}$ elements where i ranges from 1 to P, and m indicates the current selection number ranging from 1 to M. Each element $F_{i,m}$ holds the current fitness value for colony i.
- Array τ with $\tau_{i,m}$ elements where i ranges from 1 to P and m indicates the selection

number ranging from 1 to M. Each element $\tau_{i,m}$ holds the generation time for colony i when the colony will next experience a generation. This time is arbitrary in order to determine which colony is next to experience a generation and not related to physical time. We use the Rayleigh distribution to draw random $\tau_{i,0}$ variables. This can be accomplished computationally using inverse transform sampling. Please see the earlier section C.1.1 on inverse transform sampling.

- Array T with $T_{i,m}$ elements where i ranges from 1 to P, and m indicates the current selection number ranging from 1 to M. Each element $T_{i,m}$ is the number of generations until colony i's next fixation event. We use equation 2.8 from the Gerrish-Lenski model for the estimation of $T_{i,m}$, which depends on the size of the colony, the beneficial mutation rate for the organisms, and the scaling parameter for the distribution of the beneficial mutations. Since the substitution rate, equation 2.8, indicates the rate mutations are fixed in a colony, its inverse provides the expected number of generations for one mutation to be fixed.

- Array A with $A_{i,m}$ elements where i ranges from 1 to P, and m indicates the current selection number ranging from 1 to M. Each element $A_{i,m}$ is the current exponential distribution scaling parameter for colony i.

- Array D with $D_{i,m}$ elements where i ranges from 1 to P, and m indicates the current selection number ranging from 1 to M. Each element $D_{i,m}$ keeps track of the number of times colony i is selected for doubling.

- Array G with $G_{i,m}$ elements where i ranges from 1 to P, and m indicates the current selection number ranging from 1 to M. Each element $G_{i,m}$ keeps track of the number of fixations colony i has accrued.

Initialize the arrays as outlined below:

- Set $F_{i,0} = 1$, $D_{i,0} = 0$, $G_{i,0} = 0$ for i from 1 to P.

- Set $\tau_{i,0} = \sqrt{-2 * \ln(u_{i,0})}$ for i from 1 to P where $u_{i,0}$ is a random variable drawn from the uniform distribution $[0, 1)$.
- Set $A_{i,0} = \alpha_0$, for i from 1 to P where α_0 is a simulation input.
- Set $T_{i,0} = (1/k_{\text{sub\_rate}}(\mu, N))\alpha_{i,0}$, for i from 1 to P, where $T_{i,0}$ is the number of generations until colony i experiences a fitness improvement. Each element $T_{i,0}$ and the subsequent resets of $T_{i,m}$ are fixation times.

For M selections, during each selection m, we perform the following operations:

- Locate the colony with the smallest generation time. Let this colony's index be $i_{\min}$. This colony undergoes one generation during this selection. Set $D_{i_{\min},m} = D_{i_{\min},m-1}+1$.
- Set $T_{i_{\min},m} = T_{i_{\min},m-1} - 1$. If $T_{i_{\min},m} > 0$, continue to the next step. If $T_{i_{\min},m} \leq 0$, colony experiences a fixation event, set $G_{i_{\min},m} = G_{i_{\min},m-1} + 1$. Here we have two different variations of our algorithm. The first version is the non-interacting neighbors scheme (NIN) and the second version is the interaction neighbors scheme (IN) with the following treatment of fitness improvement:
  - NIN: $F_{i_{\min},m} = F_{i_{\min},m-1}(1 + s_{i_{f(\min,m)}}))$
  - IN: $F_{i_{\min},m} = max(F_{i_{\min}-1,m}, F_{i_{\min},m-1}(1 + s_{f(i_{\min},m)}), F_{i_{\min}+1,m})$

  where $s_{f(i_{\min},m)} = \sqrt{-2\sigma^2_{i_{\min},m} \ln(p_{i_{\min},m}) + d_{i_{\min},m}}$ and $p_{i_{\min},m}$ is a randomly drawn value from the uniform distribution [0,1). The other two variables are the Rayleigh distribution parameters $d_{i_{\min},m} = k_d(\mu, N)\alpha_{i_{\min},m}$, and $\sigma_{i_{\min},m} = k_\sigma(\mu, N)\alpha_{i_{\min},m}$. Reset the fixation time $T_{i_{\min},m} = (1/k_{\text{sub\_rate}}(\mu, N))\ \alpha_{i_{\min},m}$.
- Set $\tau_{i_{\min},m} = \frac{\sqrt{-2*\ln(u_{i_{\min},m})}}{\ln(F_{i_{\min},m})+2} + \tau_{i_{\min},m-1}$ where $u_{i_{\min},m}$ is a random variable drawn from the uniform distribution [0,1). Scaling $\tau_{i_{\min},m}$ with its fitness value $F_{i_{\min},m}$ effectively allows colonies with higher fitness to have relatively shorter generation times. We are using the natural log of $F_{i_{\min},m}$ instead of $F_{i_{\min},m}$ to prevent $\Delta\tau_{i_{\min},m}$ from diminishing too quickly as $F_{i_{\min},m}$ grows large since it experiences exponential growth. This time

update is valid up to a certain point until the term $\ln(F_{i_{\min},m})$ gets sufficiently large. However, it is amply sufficient for the time scale we are assessing biologically in this chapter.

The steps above are repeated until M selections is reached.

Bibliography

[1] Stephen Jay Gould and Niles Eldredge. Punctuated equilibria: an alternative to phyletic gradualism. *Models in paleobiology*, pages 82–115, 1972.

[2] Richard E Lenski, Michael R Rose, Suzanne C Simpson, and Scott C Tadler. Long-term experimental evolution in Escherichia coli. i. Adaptation and divergence during 2,000 generations. *The American Naturalist*, 138(6):1315–1341, 1991.

[3] Michael J Wiser, Noah Ribeck, and Richard E Lenski. Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367, 2013.

[4] Peter R Sheldon. Plus ça change—a model for stasis and evolution in different environments. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 127(1-4):209–227, 1996.

[5] Per Bak and Kim Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Physical review letters*, 71(24):4083, 1993.

[6] R. E. Lenski. The E. coli long-term experimental evolution project site. `http://myxo.css.msu.edu/ecoli`. Accessed: 2023-07.

[7] Carl T Bergstrom and Lee Alan Dugatkin. *Evolution*. WW Norton & Company, 2016.

[8] Michal Kucera and Björn A Malmgren. Differences between evolution of mean form and evolution of new morphotypes: an example from Late Cretaceous planktonic foraminifera. *Paleobiology*, 24(1):49–63, 1998.

[9] Michael J Benton and Paul N Pearson. Speciation in the fossil record. *Trends in Ecology & Evolution*, 16(7):405–411, 2001.

[10] Douglas H Erwin and Robert L Anstey. Speciation in the fossil record. *New approaches to speciation in the fossil record*, pages 11–38, 1995.

[11] Lydia R Heasley, Nadia Sampaio, and Juan Lucas Argueso. Systemic and rapid restructuring of the genome: a new perspective on punctuated equilibrium. *Current Genetics*, 67(1):57–63, 2021.

[12] Richard E Lenski. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME journal*, 11(10):2181–2194, 2017.

[13] John Haigh. The accumulation of deleterious genes in a population—Muller's ratchet. *Theoretical population biology*, 14(2):251–267, 1978.

[14] Sébastien Wielgoss, Jeffrey E Barrick, Olivier Tenaillon, Michael J Wiser, W James Dittmar, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E Lenski, and Dominique Schneider. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences*, 110(1):222–227, 2013.

[15] Philip J Gerrish and Richard E Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144, 1998.

[16] WI Newman and SL Phoenix. Time-dependent fiber bundles with local load sharing. *Physical Review E*, 63(2):021507, 2001.

[17] S Leigh Phoenix and William I Newman. Time-dependent fiber bundles with local load sharing. II. General Weibull fibers. *Physical Review E*, 80(6):066115, 2009.

[18] Paul D Sniegowski, Philip J Gerrish, Toby Johnson, and Aaron Shaver. The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066, 2000.

[19] Erick Denamur and Ivan Matic. Evolution of mutation rates in bacteria. *Molecular microbiology*, 60(4):820–827, 2006.

[20] Richard E Lenski and Michael Travisano. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15):6808–6814, 1994.

[21] Santiago F Elena, Vaughn S Cooper, and Richard E Lenski. Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272(5269):1802–1804, 1996.

[22] Hermann Joseph Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932.

[23] Su-Chan Park and Joachim Krug. Clonal interference in large populations. *Proceedings of the National Academy of Sciences*, 104(46):18135–18140, 2007.

[24] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381, 1987.

[25] Kan Chen, Per Bak, and Mogens H Jensen. A deterministic critical forest fire model. *Physics Letters A*, 149(4):207–210, 1990.

[26] Stephen Jay Gould. *Punctuated equilibrium*. Harvard University Press, 2009.

[27] Peter Grassberger. The Bak-Sneppen model for punctuated evolution. *physics Letters A*, 200(3-4):277–282, 1995.

[28] Ronald Meester and Dmitri Znamenski. Critical thresholds and the limit distribution in the Bak-Sneppen model. *Communications in mathematical physics*, 246(1):63–86, 2004.

[29] Luiz Renato Fontes, Carolina Grejo, and Fábio Sternieri Marques. An evolution model with event-based extinction. *Journal of Physics A: Mathematical and Theoretical*, 53(19):195601, 2020.

[30] Kim Sneppen and Giovanni Zocchi. *Physics in molecular biology*. Cambridge University Press, 2005.

[31] Henrik Flyvbjerg, Kim Sneppen, and Per Bak. Mean field theory for a simple model of evolution. *Physical review letters*, 71(24):4087, 1993.

[32] Herve Guiol, Fabio P. Machado, and Rinaldo B. Schinazi. A stochastic model of evolution. *Markov Processes and Related Fields*, 17:253–258, 2011.

[33] Ronald A Fisher. XXI.—On the dominance ratio. *Proceedings of the royal society of Edinburgh*, 42:321–341, 1923.

[34] Sewall Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97, 1931.

[35] PW Messer. Neutral models of genetic drift and mutation. *Encyclopedia of Evolutionary Biology*, pages 119–123, 2016.

[36] William Feller. *An Introduction to Probability Theory and its Applications, Volume I, 3rd edition.* John Wiley Sons, Inc., 1968.

[37] National Human Genome Research Institute. Genetic drift. `https://www.genome.gov/genetics-glossary/Genetic-Drift`, 2023. Accessed: 2023-07-31.

[38] Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196, 2011.

[39] Tat Dat Tran, Julian Hofrichter, and Jürgen Jost. An introduction to the mathematical structure of the Wright–Fisher model of population genetics. *Theory in Biosciences*, 132(2):73–82, 2013.

[40] Sophie Jeanson, Juliane Floury, Valérie Gagnaire, Sylvie Lortal, and Anne Thierry. Bacterial colonies in solid media and foods: a review on their growth and interactions with the micro-environment. *Frontiers in Microbiology*, 6:1284, 2015.

[41] James A Shapiro. The significances of bacterial colony patterns. *BioEssays*, 17(7):597–607, 1995.

[42] Robert A Gingold and Joseph J Monaghan. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly notices of the royal astronomical society*, 181(3):375–389, 1977.

[43] Joseph J Monaghan. Smoothed particle hydrodynamics and its diverse applications. *Annual Review of Fluid Mechanics*, 44:323–346, 2012.

[44] Erez Lieberman, Christoph Hauert, and Martin A Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005.

[45] Montgomery Slatkin. Fixation probabilities and fixation times in a subdivided population. *Evolution*, pages 477–488, 1981.

[46] Ewen Callaway. Legendary bacterial evolution experiment enters new era. *Nature*, 606(7915):634–635, 2022.

[47] Darren E Irwin, Jessica H Irwin, and Trevor D Price. Ring species as bridges between microevolution and speciation. *Microevolution rate, pattern, process*, pages 223–243, 2001.

[48] Shawn R Kuchta, Duncan S Parks, Rachel Lockridge Mueller, and David B Wake. Closing the ring: historical biogeography of the salamander ring species Ensatina eschscholtzii. *Journal of Biogeography*, 36(5):982–995, 2009.

[49] Nicholas H Barton. The probability of fixation of a favoured allele in a subdivided population. *Genetics Research*, 62(2):149–157, 1993.

[50] Hans Bremer. Variation of generation times in Escherichia coli populations: its cause and implications. *Microbiology*, 128(12):2865–2876, 1982.

[51] EO Powell. Growth rate and generation time of bacteria, with special reference to continuous culture. *Microbiology*, 15(3):492–511, 1956.

[52] LD Plank and JD Harvey. Generation time statistics of Escherichia coli B measured by synchronous culture techniques. *Microbiology*, 115(1):69–77, 1979.

[53] Craig A Fogle, James L Nagle, and Michael M Desai. Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics*, 180(4):2163–2173, 2008.

[54] Christopher J Skalnik, Eran Agmon, Ryan K Spangler, Lee Talman, Jerry H Morrison, Shayn M Peirce, and Markus W Covert. Whole-colony modeling of Escherichia coli. *bioRxiv*, pages 2021–04, 2021.

[55] Takeo Maruyama. Effective number of alleles in a subdivided population. *Theoretical population biology*, 1(3):273–306, 1970.

[56] Paul D Sniegowski, Philip J Gerrish, and Richard E Lenski. Evolution of high mutation rates in experimental populations of E. coli. *Nature*, 387(6634):703–705, 1997.

[57] J Ferré. Regression diagnostics. 2009.

[58] Maximilian Niyazi, Ismat Niyazi, and Claus Belka. Counting colonies of clonogenic assays by using densitometric software. *Radiation oncology*, 2(1):1–3, 2007.

[59] Kazumi Mashimo, Yuki Nagata, Masakado Kawata, Hiroshi Iwasaki, and Kazuo Yamamoto. Role of the RuvAB protein in avoiding spontaneous formation of deletion mutations in the Escherichia coli K-12 endogenous tonB gene. *Biochemical and biophysical research communications*, 323(1):197–203, 2004.

[60] Geoffrey M Cooper. *The Cell: A Molecular Approach. 2nd edition.*, chapter Cells As Experimental Models. Sunderland (MA): Sinauer Associates, 2000. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK9917/`.

[61] Robert S Breed and WD Dotterrer. The number of colonies allowable on satisfactory agar plates. *Journal of bacteriology*, 1(3):321–331, 1916.

[62] Beth Gibson, Daniel J Wilson, Edward Feil, and Adam Eyre-Walker. The distribution of bacterial doubling times in the wild. *Proceedings of the Royal Society B*, 285(1880):20180789, 2018.

[63] JAGM De Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, 2014.

[64] Sewall Wright et al. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. 1932.

[65] NobelPrize.org. Nobel Prize Outreach AB 2022. Giorgio Parisi – Nobel Prize lecture. `https://www.nobelprize.org/prizes/physics/2021/parisi/lecture/`, 2013. Accessed: 2022-04-01.

[66] Ilona Miko. Epistasis: gene interaction and phenotype effects. *Nature Education*, 1(1):197, 2008.

[67] Wild Type vs. Mutant Traits. `https://www.bio.miami.edu/dana/dox/wildtype.html`. Accessed: 2023-07.

[68] Donald E Knuth. *Art of computer programming, volume 2: Seminumerical algorithms.* Addison-Wesley Professional, 2014.

[69] James Stewart. Calculus: early transcendentals. *Thomson Brooks/Cole*, 2008.

[70] William I Newman. *Mathematical methods for geophysics and space physics.* Princeton University Press, 2016.