An Agent-Based Numerical Approach to Lenski's Long Term Experiment

An Agent-Based Numerical Approach to Lenski's Long Term Experiment

A thesis presented

by

James F. Almgren-Bell

To

Applied Mathematics

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 29, 2019

Abstract

In this thesis, we introduce and discuss two models of bacterial evolution within the framework of Lenski's experiment, which studies the evolution of bacterial colonies in a fixed environment over many generations. In the first model, we simulate the infinite-sites model, in which bacteria are characterized by their fitness. We consider exponential and Gaussian distributions for fitness draws and compare the results to analytical predictions. In this case, we primarily look at long-term fitness trajectories in both the strong selection weak mutation and clonal interference regimes. In the second model, we begin to explore the effects of bacterial competence, the intra-generation sharing of DNA, within the context of the spin glass model both with and without epistasis. We seek to understand the effects of bacterial competence as a function of the parameters chosen.

Acknowledgements

I would like to thank my thesis advisor, Chris Rycroft, for his mentorship throughout the project and his extensive help with the writeup. I would also like to thank Cengiz Pehlevan for serving as my second thesis reader. I am also grateful to Ariel Amir for his leadership of the overall project and his technical guidance, and to Yipei Guo for repeatedly providing mathematical assistance. Finally, I would like to express my particular thanks to Nick Boffi for working with me on a weekly basis and guiding me throughout this entire project. Without him, this thesis would not have been able to happen.

Contents

1 Introduction

Traditional work on evolutionary dynamics looks at how different species respond to a changing environment, wherein they must adapt and evolve in order to survive. This adaptation occurs through random genetic mutations within individuals in a population that are beneficial to the individual's aptitude, or fitness, to its environment. This increased fitness corresponds to an increased rate of reproduction. Because genetic material is passed down from parent to child, at long time scales, traits corresponding to higher fitness will be selected for and their prevalence within the population will rise, a process known as natural selection. While this process has been commonly accepted for many years, an important follow-on question to ask was if these same processes will continue to occur in a constant environment in which a species faces little to no threat or competition. In 1988, a team led by Richard Lenski began the *Escherichia coli* long term evolution experiment (LTEE), in which populations of *E. coli* bacteria are grown in a constant environment and are given time to evolve over thousands of generations [15]. In the experiment, which is still ongoing, Lenski separately grew 12 populations of the bacteria in separate (but the same) lab environments, which can be seen in Figure 1. Over the course of a day, each population, which may contain a number of different strains, grows and mutates via asexual reproduction. At the end of each day, approximately 1% of each population is separated, or diluted, and begins growth again.

Figure 1: Lenski's long-term evolution experiment; shown are the 12 populations of bacteria growing in a lab environment [19]

Lenski and his team found that despite no change in the environment, the populations

continued to evolve, even after 20,000 generations (albeit at a much decreased rate) [7]. After 50,000 generations, they showed that, rather than stopping, the fitness of the population was still growing according to a derived power law [24]. Lenski and his collaborators have made a number of different measurements and been able to infer several characteristics of this evolutionary process. For example, they were able to estimate the point-mutation rate of the strain of *E. coli* used to higher accuracy than ever before [23]. A second quality they observed was the presence of clonal interference, in which rather than one genotype dominating the population, there are several [5]. A third feature studied in the context of the LTEE was epistasis, the interaction between two different genes. Simulations done using "digital organisms", computer programs built to represent individuals that reproduce and mutate, support the hypothesis that epistasis is a feature of most genetic systems [14]. More recently, additional research has been done to better understand fitness landscapes, which describe the connections between different genotypes as well as their corresponding fitnesses [11, 13, 18, 17].

Here, our interest lies primarily in the infinite-sites model [18], in which an individual is represented purely by its fitness value, rather than a genome. In particular, we look at uncorrelated fitness landscapes, in which if a child mutates its fitness value has no correlation to its parent fitness value. The goal of much of this work is to understand the distribution of fitness effects (DFE) of new mutations. An understanding of the DFE can provide insight pertaining to the expected rate of evolution, applications of which include the study of complex human disease [8]. An early approach to this problem came from Gillespie in 1983, who used extreme value statistics to study the process of natural selection [10]. In 2006, Orr used Fisher's geometric model of mutation to expand on Gillespie's work and to analytically and computationally show that the DFE of beneficial mutations belongs to the Gumbel class [17]. This body of work is generally referred to as Orr–Gillespie theory. In particular, Orr showed that the distribution of fitness effects for beneficial mutations with low probability of occurrence is asymptotically exponential [17]. This idea was first introduced in 1978 by Kingman, who termed the use of the exponential distribution the "house of cards" model [12]. This has been further studied recently both analytically and numerically by multiple authors [13, 18]. With some limiting assumptions, significant progress has been made in the study of such distributions [9]. In the majority of simulations done, the Wright–Fisher model has been used,

which uses a fixed population size with no generational overlap.

Our work seeks to test the house of cards model using an agent-based simulation based on the dilution cycles in Lenski's original experiment. We are also interested in looking at the Gaussian distribution in the same context as the exponential distribution. Notable points of interest include fitness trajectories, measures of clonal interference, and fitness increments. Further, many of the results presented in Kryazhimskiy *et al.* were derived in the "strong selection weak mutation" (SSWM) regime, in which there is assumed to be only one genotype dominating the population at a given time [13]. As opposed to the SSWM regime, the clonal interference regime is known to cause differences, as multiple genes fix simultaneously [4]. We would like to compare the SSWM regime to the clonal interference regime in terms of the analytical formulations presented and the measurements mentioned above.

The second topic of interest is the process of bacterial competence, in which individuals within a given population may share their DNA [6]. This process begins when an individual drops a copy of a strand of its DNA into the environment. Another individual then takes up this free DNA and incorporates it into its genome. Because competence is a method of transfer of genetic information within a generation rather than from parent to child, it is also referred to as horizontal gene transfer [21]. Currently, most of the research done on bacterial competence is focused on the biological mechanisms of this process, in particular how to induce competence in a laboratory setting [16, 6]. Little work has been published on analysis or numerical simulation of bacterial competence. It was suggested to use a model of the genome based upon spin-glass models from magnetism that was previously developed by my collaborators in order to study bacterial competence, again using agent-based Lenski framework simulations [2]. An advantage of this model is that we have specific control over whether epistasis occurs or not, and thus can study the effects of competence both with and without epistasis.

2 The Two Models

In the work presented in this thesis, two different models of populations are considered, both within the Lenski framework. The two models considered are the infinite and finite sites

models, which will be explained in detail below. These models primarily differ in how they calculate fitness, which is quantitatively represented as the rate of reproduction. In the real world, the fitness of a certain organism depends on both the characteristics of the organism as well as its surrounding environment. In the case of the Lenski framework, a constant environment allows for tracking evolution using the measurement of fitness over time without needing to take changing external factors into consideration.

2.1 Infinite-Sites Model

The first model of interest, the infinite-sites model, has been studied by multiple authors [18, 13], but is given particular attention in a previously mentioned paper by Kryazhimskiy *et al.* [13]. Within this model, there is no explicit genome, and bacteria are thus distinguished purely by their fitness value. When a mutation occurs, the fitness of the new strain is equal to a value drawn randomly from a probability distribution. In the cases considered in this thesis, this probability distribution is independent of the parental fitness. The independence of this distribution means that the fitness landscape is referred to as an "uncorrelated landscape", where there is no relationship between the fitnesses of a parent and a mutant child. In their paper, Kryazhimskiy *et al.* use the Wright–Fisher framework for evolution to study these dynamics. This model holds the number of individuals N fixed, and in each time step an individual mutates with probability μ. This paper also primarily focuses on the "Strong Selection Weak Mutation" (SSWM) parameter regime, in which it is assumed for analysis that only one strain exists in the population at a time. Mathematically, this represents the regime in which the expected fixation time of a mutation is much less than the expected time for a beneficial mutation to occur. The expected fixation time of a mutation is the expected time for a mutation to fix, or take over the population, given that it will do so. The authors develop explicit formulas for a number of quantities, most interestingly the fitness trajectory (fitness measured as a function of time) [13]. They show their results provide an accurate model at large time scales within the SSWM regime and hypothesize that they hold in the non-SSWM regime, which occurs when the mutation rate is sufficiently high. As previously mentioned, this non-SSWM regime is referred to as the clonal interference regime, in which multiple strains coexist. In this work, we explore how these results hold up in the Lenski

framework in both regimes. We also analyze some measurements and discuss a probability distribution not discussed in the paper.

2.2 Spin Glass Model - Bacterial Competence

The second structure used to model the experiment is based upon spin glass models from condensed matter physics. The model was originally developed to model different spins of component atoms that make up an irregularly structured magnet. The Sherrington and Kirkpatrick spin glass is used here, which allows for infinite-range interactions between components, rather than just interactions of neighbors [20]. This makes biological sense, as the effect of a gene is measured by the proteins it produces, which can interact with other proteins created by genes anywhere on the genome. The set of non-zero interactions is sparse, such that not all genes will interact because they may produce unrelated proteins. This model uses a sequence of $+1$ and -1 values, which can be easily adapted to a genome; instead of spins, these represent alleles. When a mutation occurs, one of these allele values flips sign. An additional motivation for this type of model is that spin glass models tend to have logarithmic "relaxations" [3], which in magnetism refer to the drift of the magnetic structure toward desirable, low-energy states. In the context of evolutionary dynamics, a low-energy state corresponds to a high-fitness state, which is desirable. This logarithmic relaxation makes the spin-glass model attractive because it shows very slow long-term growth, similar to the the fitness growth seen in Lenski's experiment [24]. For this thesis, the primary interest in this spin glass model is to look at the effects of introducing bacterial competence, the process by which an individual can make a copy of a given fragment of its genome to drop into the surrounding environment. The fragment can then be taken up by another individual and swapped into its genome, replacing existing DNA. In some cases, bacteria must undergo a certain transitional process or stress in order to transform into a state in which competence is possible, whereas in other cases the process is natural. A diagram of the competence process is shown in Figure 2.

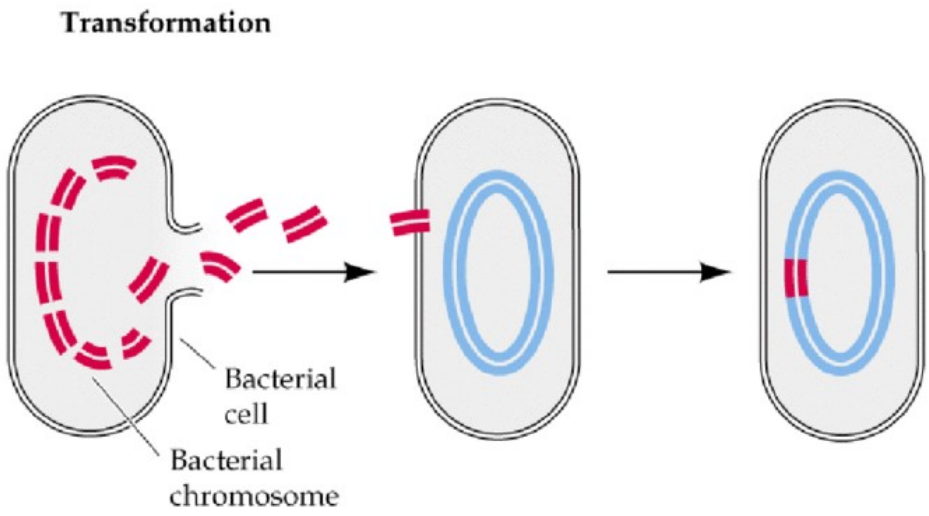

Figure 2: Diagram of bacterial competence; one individual makes a copy of its DNA and releases it into the environment, where the strand is absorbed and included into a second individual's DNA [1]

.

Our work on this subject seeks to determine if bacterial competence will increase the rate of evolution, determined by looking at the fitness trajectory. The expectation is to see qualitatively similar behavior both with and without epistasis. The hypothesis is that, because evolution favors the fittest individuals, the fittest individuals will be around longer and thus will be more likely to drop fragments of their genome. These fragments will likely lead to beneficial mutations for other strains, as they are from the fitter strains. Hence, the rate of increase of the average fitness across the population would increase.

3 Methodology

We explore two different agent-based models of bacteria, both of which are implemented within the Lenski framework, which involves tracking each strain of bacteria in a given population. In the infinite-sides model, individuals are solely represented by their fitness value, as they are modeled without a genome. In the spin glass model, individuals have a finite number of genes which can mutate. Before discussing the details of these representations of individuals, it is important to present the Lenski framework which they both share.

Each day begins with N_0 total individuals summed over all the strains in the population. In each time step, growth of each strain occurs. Writing $N_{s,0}$ as the initial number of individuals in a strain s with fitness F_s, the number of individuals in strain s, N_s, grows exponentially according to the equations

$$\frac{\partial N_s}{\partial t} = F_s N_s, \quad \text{where } N_s(0) = N_{s,0}. \tag{1}$$

After a time step Δt, the number of individuals in strain s with fitness F_s beginning with N_s individuals is

$$N_s(t + \Delta t) = N_s(t) e^{F_s \Delta t}. \tag{2}$$

Hence, the number of new individuals is

$$dN_s = N_s(t)(e^{F_s \Delta t} - 1). \tag{3}$$

From these new individuals, the number of mutants is drawn from a Poisson distribution with mean $\mu\, dN_s$, where μ is the mutation rate. When an individual mutates, it may either become a member of another existing strain or be the first member of a new strain. The rest of the offspring, which do not mutate, are exactly the same as their parents. Time steps are taken until the number of bacteria summed over all strains exceeds N_F. Then, growth is stopped, and the population is diluted down to N_0 individuals. This process is simply done by randomly selecting N_0 of the bacteria to keep and removing the rest. Note that this dilution process does not directly favor strains with higher fitness. Rather, it does so indirectly, because strains with higher fitness will undergo more growth and thus have more individuals to possibly be selected. Note that Lenski's original experiment allows the bacteria to grow for a fixed amount of time rather than up to a certain size before the dilution process. We chose the latter for the purpose of comparison between the numerical results and analytical predictions. The analytical predictions are formulated as functions of the number of mutations that have occurred, whereas the numerical results are a function of dilution cycles. With this model, the expected number of total mutations per day among over all strains is approximately $\mu(N_F - N_0)$. Thus, the analytical form can be linearly transformed for comparison against

the numerical results. If time was used instead to signal the end of a dilution cycle, this would not be possible, as the number of expected mutations per day would not be constant.

The primary data points of interest are:

- Fitness Trajectories - average fitness in the population as a function of time
- Number of strains in the population as a function of time

With this framework, we explore the two different models mentioned above. Below are more detailed explanations of each of these models.

3.1 Infinite-Sites Model

As previously mentioned, in the infinite-sites, or probabilistic, model, an individual is simply identified by its fitness value, meaning that it has no explicit model for its genetic code. Thus, two individuals are of the same strain if and only if they have the exact same fitness. When a mutation occurs, the fitness of the child mutant is drawn from a continuous probability distribution $\phi_F(y)$, where F represents the current fitness of the strain. $\phi_F(y)$ is referred to as the "neighbor fitness distribution" (NFD). For such a distribution, the probability of drawing the exact same number twice is zero. This means that mutation cannot occur between two currently existing strains. For this thesis, the focus will be placed on models such that $\phi_F(y)$ is independent of the current fitness value F, *i.e.* $\phi_F(y) = \phi(y)$ for all F.

3.2 Spin Glass Model

In the spin glass model, each individual's genetic code is a sequence of length L of $\alpha_i's$, corresponding to the individuals genetic code. These values $\alpha_1, \alpha_2, \ldots \alpha_L$ are either 1 or -1 and are used to determine an individual's fitness. Thus, two individuals are of the same strain if and only if their sequence of α_i's is identical. When a mutation occurs, a random site i is chosen and the sign of α_i is flipped. In this way, a strain's genome can be more efficiently stored by simply saving the set of locations j such that $\alpha_j = -1$. At the beginning of an experiment, random values h_i are drawn from a normal distribution with mean 0 to represent the fitness effect of each gene. Note that these h_i values are the same for all strains. An

individual's fitness is written as

$$F = \sum_{i=1}^{L} h_i \alpha_i. \tag{4}$$

When a mutation occurs in which we have an α_i flip sign, rather than recalculating the fitness F_c of the child strain from scratch, we can compute it from the parental fitness F_p. In this way, if we have a mutation at site k, the fitness of the child is written as

$$F_c = F_p + h_k \left(\alpha_k^{\text{child}} - \alpha_k^{\text{parent}} \right) = F_p + 2h_k \alpha_k^{\text{child}}, \tag{5}$$

because $\alpha_i^{\text{child}} = -\alpha_i^{\text{parent}}$. An equivalent calculation is used when multiple genes can change at once, which is only used when bacterial competence for strand length greater than 1 is implemented. An extension to this model considers interaction effects that can occur between different genes, a phenomenon called epistasis. This interaction is represented in the spin glass model by drawing random values J_{ij}. The matrix J of J_{ij}'s is sparse and thus many interactions have no effect; this sparsity is determined by a density parameter we label ρ (in all simulations done, $\rho = 0.05$). The non-zero values are drawn from a normal distribution with mean zero. Also note that a gene cannot interact with itself, and therefore $J_{ii} = 0$ for all i. As before, these J_{ij} values are the same for all strains. An interaction term is thus added, and fitness is written as

$$F = \sum_{i=1}^{L} h_i \alpha_i + \sum_{i=1}^{L} \sum_{j=1}^{L} J_{ij} \alpha_i \alpha_j. \tag{6}$$

We note that because of how the interaction parameters are defined, $J_{ij} = J_{ji}$. Thus, for computational efficiency, the form used is

$$F = \sum_{i=1}^{L} h_i \alpha_i + 2 \sum_{i=1}^{L} \sum_{j>i}^{L} J_{ij} \alpha_i \alpha_j. \tag{7}$$

A calculation similar to the one done above can be performed to quickly calculate the fitness of a child based upon the fitness of the parent when the interaction terms are present. With epistasis, the fitness of the child with a mutation at site k can be written as

$$F_c = F_p + 2h_k \alpha_k^{\text{child}} + 4h_k \alpha_k^{\text{child}} \sum_{i=1}^{L} J_{ik}. \tag{8}$$

This calculation is done slightly differently in practice because the genome is stored as a set of mutations rather than a full genome, but this is the basis for the calculation.

3.3 Bacterial Competence

A second extension to the spin glass model is to incorporate bacterial competence, which can be done with or without epistasis. Here, a set of genetic sequences of length l_{comp} in the surrounding environment is stored as free DNA. Also stored are the indices of their corresponding locations in the genome, because the sequence may not move to a different location within the genome. In each time step, decay of strands in the environment may occur. Biologically, this means that the strains will degrade and fall apart, thus no longer representing useful DNA sequences. Suppose there are S_{env} strands in the environment. The number of decaying strands is randomly chosen from a Poisson distribution with mean $p_{\text{decay}}S_{\text{env}}$, where p_{decay} represents the probability of decay. These strains are thus removed from the system. Next, each strain of bacteria can make a copy of a length l_{comp} sequence to drop into the environment. The number of strands dropped by each strain is chosen from a Poisson distribution with mean $p_{\text{drop}}N_s$, where p_{drop} represents the probability of dropping and as before N_s is the number of individuals of strain s. The starting location for each strand to be dropped is chosen randomly from the $L - l_{\text{comp}} + 1$ possible sites. Similarly, each strain can absorb and utilize a strand from the environment, thus destroying its current l_{comp}-length sequence. The number of strands absorbed by a strain s is chosen from a Poisson distribution with mean $p_{\text{take}}S_{\text{env}}N_s$, where p_{take} is the probability of absorbing a strand and N_s represents the population of strain s. If multiple strands are absorbed that overlap at a number of sites, the strand that is ultimately used is chosen randomly. After this, the usual stepping process continues. During the dilution phase, the strands in the environment are also diluted down by selecting a random fraction of them corresponding to the dilution factor.

3.4 Implementation

Both models are built from the same code base in C++, originally developed by Nick Boffi for the spin glass model without competence. Two extensions have been made, one for the probabilistic model and one for the spin glass model with competence. Simulations are run in

parallel using OpenMP, a shared memory parallel programming model. Because we need to perform an ensemble of independent simulations to gather statistics, the code is straightforward to parallelize. At the end of each dilution cycle, before growth begins again for the next cycle, data is written to binary files that are used for post-processing in a separate Python code. Here, the data is averaged across experiments before being analyzed. In summary, each simulation has the following parameters:

- N_0: Population at which each dilution cycle begins
- N_F: Population at which, when passed in a time step, growth is stopped and dilution occurs
- N_{cycles}: Number of dilution cycles
- N_{exps}: Number of experiments over which data is averaged
- Δt: Time step used for bacterial growth
- μ: Mutation rate
- F_0: Fitness of the initial strain

Empirically, it is found that a value of $N_{\text{exps}} = 100$ is more than sufficient to produce reproducible results. A value of $\Delta t = 0.1$ is chosen to be small enough that there is not too much overshoot in reaching N_F individuals but large enough for run-time purposes. Note that results here are presented in terms of dilution cycles. There is some variability in the physical time per dilution cycle from the dependence of the evolution on the fitnesses of the population. For simplicity, a value of $F_0 = 1$ is used for all experiments. A dilution factor of 1% is typically used to accurately model Lenski's original experiment [7]. Additionally, it was found that with this constant dilution factor, values of $N_0 = 100$, $N_0 = 1,000$, and $N_0 = 10,000$ produced behavior and results that were indistinguishable when time scaling was taken into account. Thus, for simplicity, most runs are done with $N_0 = 100$ and $N_F = 10,000$.

4 Results: Exponential Distribution

The first neighbor fitness distribution we consider is what Kryazhimskiy *et al.* refer to as the "house of cards" model, an example of an uncorrelated fitness landscape [13]. The house of cards model represents an exponential distribution, where

$$\phi(y) = \lambda \exp(-\lambda y) \tag{9}$$

for some parameter λ. The paper calculates an analytical form for the fitness trajectory based upon the equation

$$\frac{\partial F}{\partial m} = r(F), \tag{10}$$

where m is the number of mutations that have occurred. Here, $r(F)$ is the expected fitness increment, the expected increase in fitness when a mutation occurs for a starting fitness F. In the SSWM limit, $r(F)$ is calculated and is shown to be

$$r(F) = \frac{4(\lambda F + 1)}{\lambda(\lambda F + 2)^2} \exp(-\lambda F). \tag{11}$$

The authors then use an approximation appropriate for large F to write this quantity as

$$r_1(F) \approx \frac{4}{\lambda^2} \exp(-\lambda F). \tag{12}$$

The advantage of this approximation is that it can be integrated to obtain

$$F(m) = \frac{1}{\lambda} \ln\left(\exp(\lambda F_0) + \frac{4}{\lambda} m\right). \tag{13}$$

It is important to note, however, that this formula is based on the aforementioned approximation that only holds for large F. Based upon numerical simulations, it is observed that the fitness does not reach that limit until very late time. Thus, we consider the second form in Eq. 14,

$$\frac{\partial F}{\partial m} = \frac{4(\lambda F + 1)}{\lambda(\lambda F + 2)^2} \exp(-\lambda F). \tag{14}$$

While this cannot be solved analytically, it can easily be numerically calculated using Python's numerical solvers. As previously mentioned, it is important to note that, in this analytical formulation, time is measured as the number of mutations that have occurred. By contrast, the Lenski simulations represent time in the number of dilution cycles c. In order to compare trajectories from Lenski simulations and this analytical form, one must rescale by the expected number of mutations per dilution cycle, $\mu(N_F - N_0)$. Thus, to compare a fitness trajectory from a numerical simulation $F_L(c)$ and a fitness trajectory from the paper $F_K(m)$, the quantity that should be considered is $F_L(\mu \Delta N m)$.

Various simulations are done using the following parameters:

- $N_0 = 100$
- $N_F = 10,000$
- $N_{\text{cycles}} = 100,000$
- $N_{\text{exps}} = 400$

The following polots of fitness trajectories are produced and shown in Figure 3.

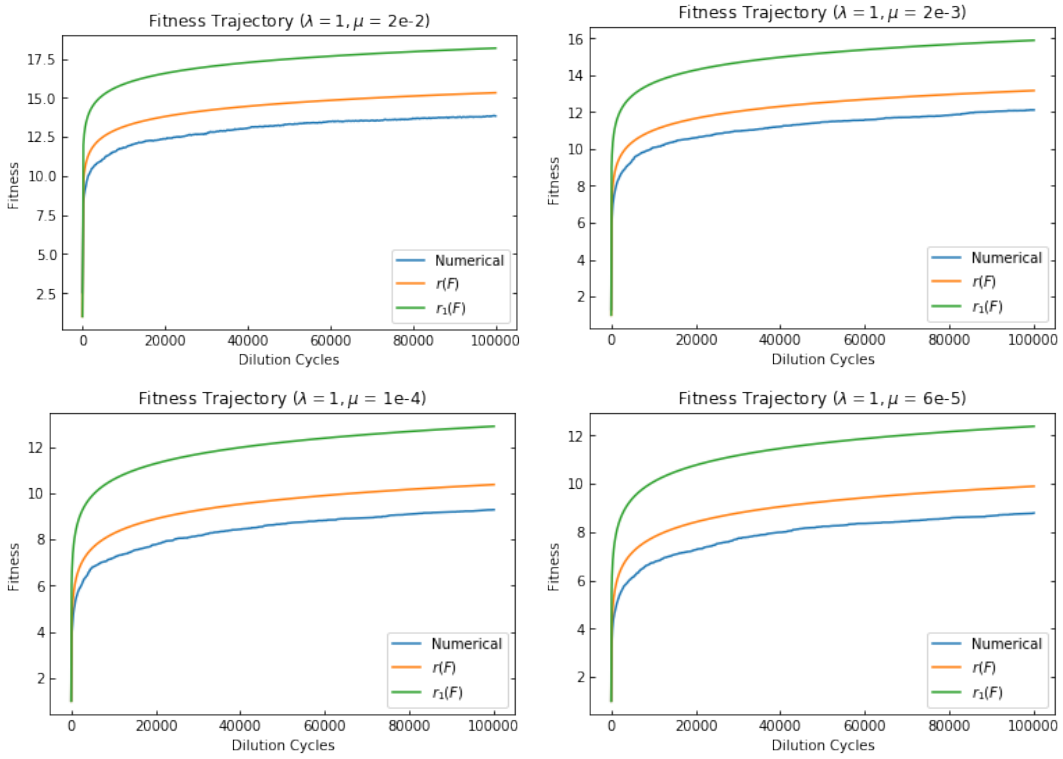

Figure 3: Fitness trajectories for exponential NFD; $\lambda = 1$; $\mu = 2 \times 10^{-2}, 2 \times 10^{-3}, 1 \times 10^{-4}, 6 \times 10^{-5}$

As can be seen from Figure 3, we see similar growth over time, but there appears to be a shift in the graphs. This disparity happens at very early time and remains fairly constant afterward. This is somewhat expected, as the analytical formulas were designed for large F, which occurs in long time, as well as in the SSWM regime. The simulations here satisfy neither of these assumptions. As an example, the simulation with $\mu = 2 \times 10^{-2}$ has approximately 3 coexisting strains (this will be discussed later in further detail). We now seek to determine if the trajectory is accurate, as this is a far more interesting quantity than the absolute fitness value. To do this, we will take a finite difference approximation of the derivative of the observed $F(c)$, writing

$$\frac{\partial F}{\partial c}(c_0) \approx \frac{F(c_0 + \Delta c) - F(c_0 - \Delta c)}{2\Delta c}. \tag{15}$$

This can also be done for the analytical formulations in the same way. A value of $\Delta c = 6,000$ dilution cycles is used in order to smooth the results, which are very noisy. The results are shown below in Figure 4.

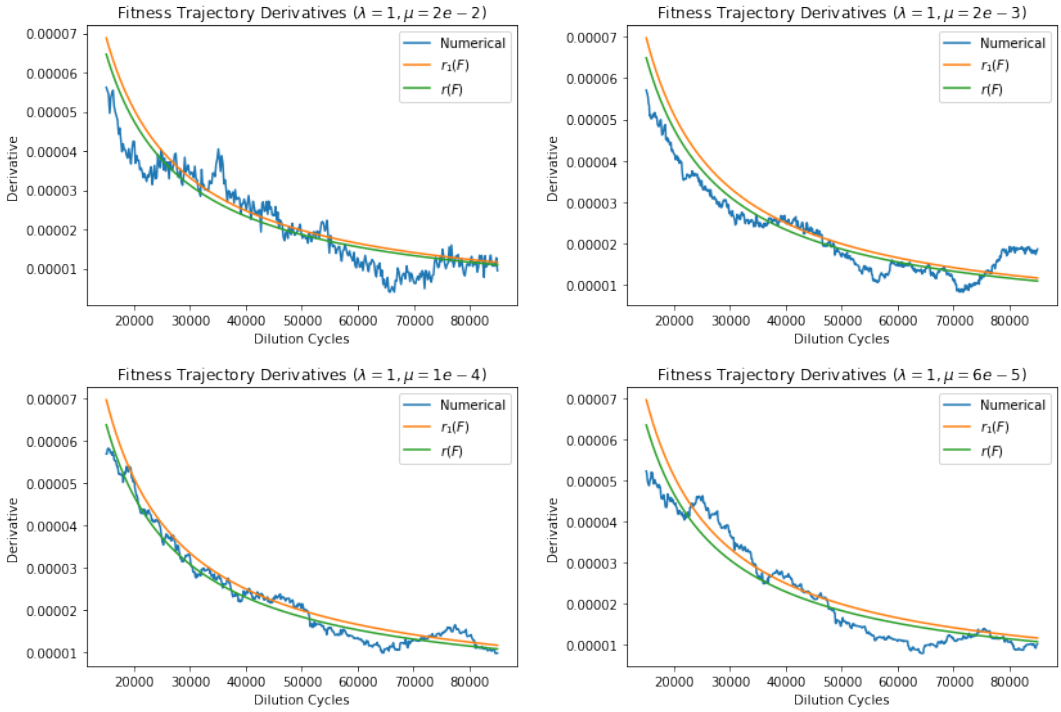

Figure 4: Fitness trajectory derivatives for exponential NFD; $\lambda = 1$

As can be seen, despite the noise, these analytical formulations appear to be good predictors of the long-term growth. These results are confirmed for other values of λ. Plots are shown below in Figure 5 and Figure 6 of similar results to those mentioned above for $\lambda = 2$ and $\lambda = 3$, both with $\mu = 3 \times 10^{-4}$.

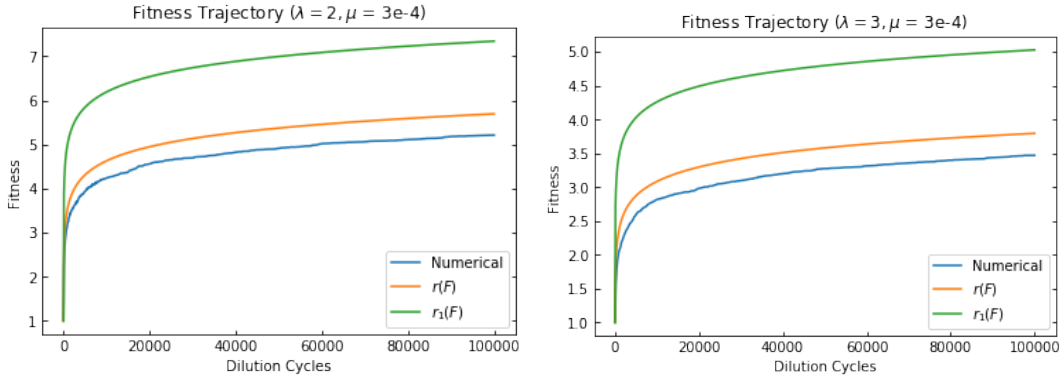

Figure 5: Fitness trajectories for exponential NFD; $\lambda = 2$, $\lambda = 3$

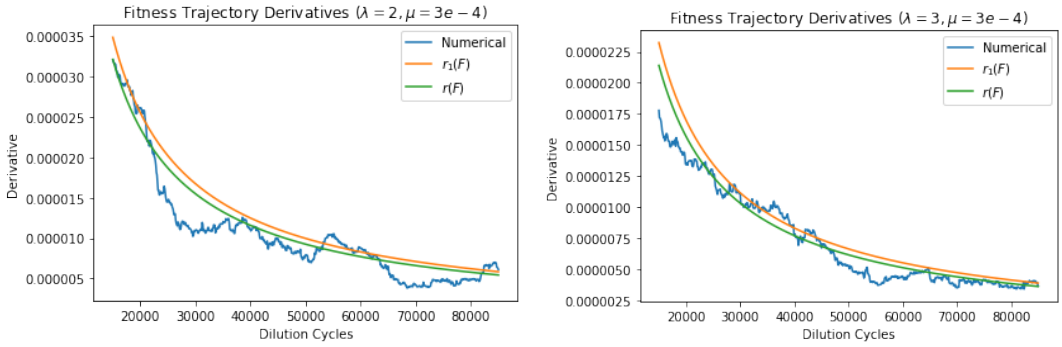

Figure 6: Fitness trajectory derivatives for exponential NFD; $\lambda = 2$, $\lambda = 3$

A final point of interest is the shift seen in the graphs. Empirically, the shift between the numerical results and $r(F)$ seems to be of magnitude $\frac{1}{\lambda}$. This shift appears to hold up regardless of the value of λ or μ. Some examples are shown below in Figure 7:

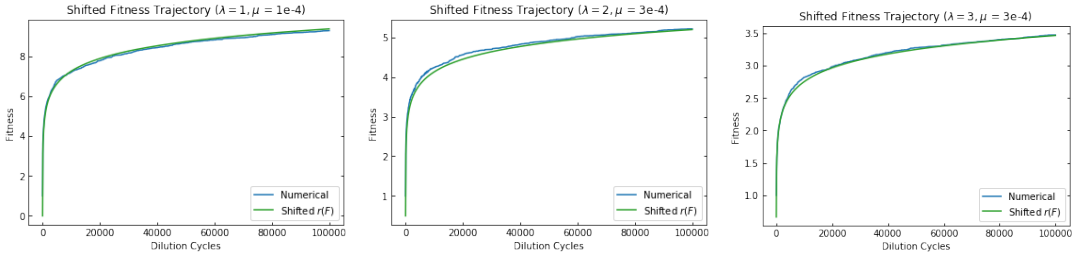

Figure 7: Shifted fitness trajectories for exponential NFD; $\lambda = 1, 2, 3$

Overall, we are able to show that the analytical formulation presented in the paper from Kryazhimskiy *et al.* is a good predictor of the fitness trajectory in the case of the Lenski framework. Additionally, we show that their result holds in the clonal interference regime, when mutation rates are high.

5 Results: Normal Distribution

The second distribution of consideration is the normal distribution, $\phi_F(y) \sim N(\eta, \sigma)$. A clear choice for the mean is to match the initial fitness in the experiment, *i.e.*, to set $\eta = F_0$. If η is chosen to be a greater value, the fitness value of the experiment will simply jump to around the mean, because all values drawn for mutations will have fitnesses near that value. This is a trivial shift of the trajectory upwards, and thus it does not produce interesting behavior.

The more interesting parameter worth considering is σ, which will determine the shape of the curve.

5.1 Analytical Form

To look for an analytical solution in the case of the normal distribution, we generalized a result from Vivo [22]. The relevant part of this paper focuses on describing a set of N identically normally distributed random variables. In the paper, Vivo derives an analytical form for a_N, the expectation of the maximum of the set after N numbers are drawn. This is an appropriate quantity to look at, particularly in the SSWM regime, because in the Lenski model the fitness will tend to follow the maximum of the previous fitnesses drawn. Strains with higher fitness values will grow faster and thus be more likely selected to be kept in the dilution stage. Although some mutants with higher fitness will die off after being drawn because they do not make it through the dilution process, the expectation is that $F(t)$ should match up closely to a_N when rescaling is done. At the minimum, the a_N curve should be an upper bound with the same trajectory, because the fitness cannot be higher than the maximum fitness value drawn. Further, Vivo's work suggests a slow convergence to this analytical form. Vivo's derivation was done with a mean $\eta = 0$ and a standard deviation $\sigma = 1$. To shift to using a non-zero mean, η can simply be added as a constant. To find the result for a different σ, the derivation from the paper is recreated below with an arbitrary σ value.

First, consider the cumulative distribution function after N draws, $Q_N(x)$, which in the general case is

$$Q_N(x) = \left[1 - \int_x^\infty dy\, p(y)\right]^N. \tag{16}$$

For the normal distribution, this becomes

$$Q_N(x) = \left[1 - \int_x^\infty dy \frac{\exp\left(\frac{-y^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\right]^N. \tag{17}$$

Using the change of variables $z = \frac{y}{\sigma}$, write

$$Q_N(x) = \left[1 - \int_{\frac{x}{\sigma}}^{\infty} dz \frac{\exp\left(\frac{-z^2}{2}\right)}{\sqrt{2\pi}}\right]^N. \tag{18}$$

As x becomes large, the integral becomes small, and the approximation $(1-\epsilon)^N \approx \exp(-N\epsilon)$ gives

$$Q_N(x) \approx \exp\left[-N \int_{\frac{x}{\sigma}}^{\infty} dz \frac{\exp\left(\frac{-z^2}{2}\right)}{\sqrt{2\pi}}\right]. \tag{19}$$

Using another change of variables $\tau = \frac{\sigma z}{x}$, the integral thus becomes

$$\frac{x}{\sigma} \int_1^{\infty} d\tau \frac{\exp\left(\frac{-x^2\tau^2}{2\sigma^2}\right)}{\sqrt{2\pi}}. \tag{20}$$

Using Taylor Series, expand the quantity $\frac{\tau^2}{2\sigma^2}$ around 1 as

$$\frac{\tau^2}{2\sigma^2} \approx \frac{1}{2\sigma^2} + \frac{\tau - 1}{\sigma^2}. \tag{21}$$

Then, approximate the integral as

$$\frac{x}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \int_1^{\infty} d\tau \exp\left(\frac{-x^2(\tau - 1)}{\sigma^2}\right), \tag{22}$$

which can be evaluated using Mathematica to obtain

$$\frac{\sigma}{\sqrt{2\pi}x} \exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{23}$$

Substituting this quantity back into the equation for Q_N gives

$$Q_N \approx \exp\left[\frac{-N\sigma}{x\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right]. \tag{24}$$

Following Vivo and rewriting the right hand side as $\exp[-\exp(-\gamma_N(x))]$ gives

$$\gamma_N(x) = -\log N - \log\sigma + \frac{1}{2}\log 2\pi + \log x + \frac{x^2}{2\sigma^2}. \tag{25}$$

As discussed in Vivo, the goal of extreme value statistics here is to write

$$\lim_{N\to\infty} Q_N(a_N + b_N z) \tag{26}$$

as a function of z. Thus we let $x = a_N + b_N z$ and let $x \to \infty$, where a_N is as described above, and b_N is considered a "scaling". Again following Vivo, this is done so that $\gamma_N(z) \sim z$ for large N, and hence $b_N = \frac{\sigma^2}{a_N}$. Substituting this into Eqn. (25) gives

$$\gamma_N = \log\left(\sqrt{2\pi}\right) - \log\left(N\sigma\right) + \log\left(a_N + \frac{\sigma^2 z}{a_N}\right) + \frac{a_N^2}{2\sigma^2} + \frac{z^2}{2a_N^2} + z. \tag{27}$$

We introduce an ansatz for a_N to cancel the $\log(N\sigma)$ term. Specifically, we write

$$a_N = \sqrt{\sigma^2(2\log N\sigma + c_N)}. \tag{28}$$

The $\frac{z^2}{2a_N^2}$ term can be ignored because in the large N limit it becomes very small. The equation can thus be simplified to

$$\gamma_N \approx z + \frac{c_N}{2} + \frac{\log 2\pi}{2} + \log\left(a_N + \frac{\sigma^2 z}{a_N}\right). \tag{29}$$

The $\frac{\sigma^2 z}{a_N}$ term can be ignored because it is small, and thus expanding the log term using Taylor series gives

$$\gamma_N \approx z + \frac{c_N}{2} + \frac{\log\left(2\pi\right)}{2} + \frac{\log\left(2\log\left(N\sigma\right)\right)}{2} + \frac{c_N}{4\log\left(N\sigma\right)} + \log\left(\sigma\right). \tag{30}$$

Because the goal is to get γ_N to be $\sim z$, set the sum of the rest of the terms to zero. Solving for c_N and then plugging back into a_N gives

$$\boxed{a_N = \sigma\sqrt{\left(2\log\left(N\sigma\right) - \log\left(4\pi\sigma^2\log\left(N\sigma\right)\right)\right)}}. \tag{31}$$

5.2 Numerical Simulations

As in the exponential case, we now perform simulations to test the accuracy of a_N as an estimate of the fitness as a function of number of mutations in both in the short and long

term. Various simulations are done using the following parameters:

- $N_0 = 100$
- $N_F = 10,000$
- $N_{\text{cycles}} = 100,000$
- $N_{\text{exps}} = 400$

Figures 8 and 9 show fitness trajectories that are produced by simulations done with $\sigma = 0.05$ and $\sigma = 0.1$:

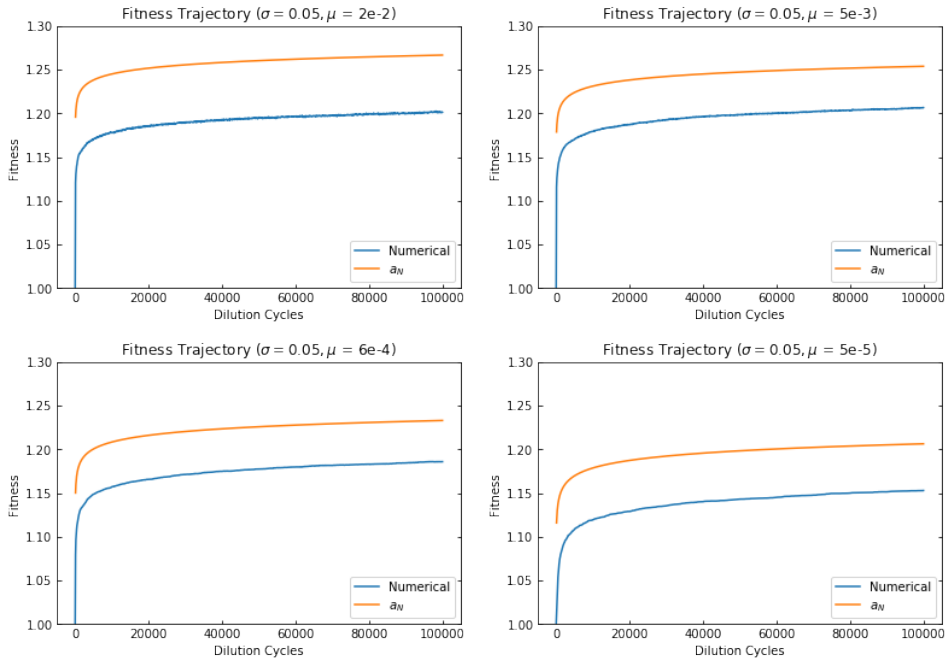

Figure 8: Fitness trajectories for Gaussian NFD; $\sigma = 0.05$

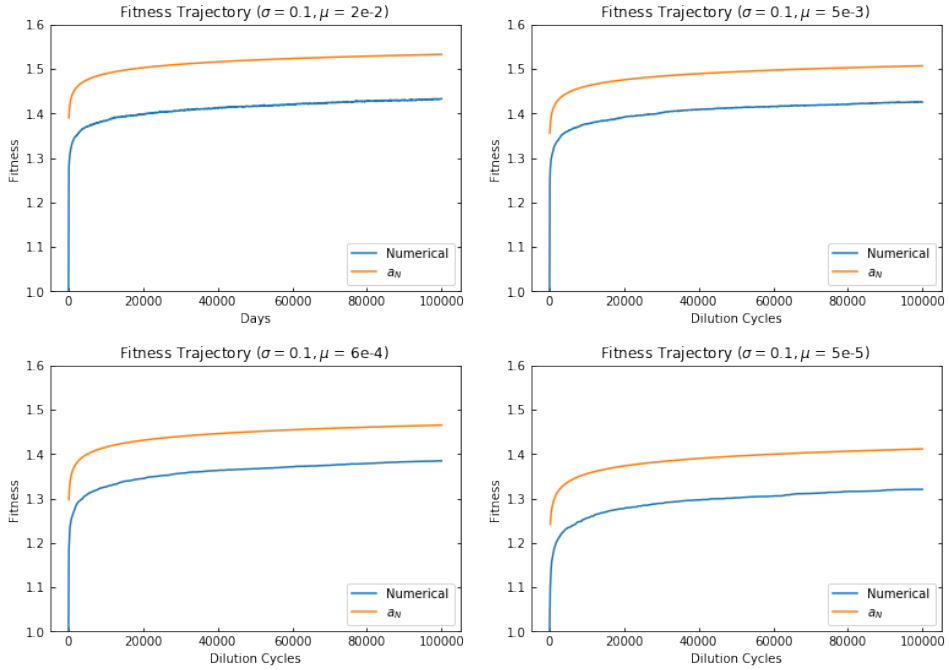

Figure 9: Fitness trajectories for Gaussian NFD; $\sigma = 0.1$

As can be seen in Figures 8 and 9, while the analytical form shows to be a good fit with regard to the slope, it is consistently too high. As mentioned in the derivation, the analytical form represents the maximum of N draws from the normal distribution. In the case of the Lenski framework, it is very possible for a mutation that corresponds to the highest fitness to not fix in the population, because it may be diluted out. Thus, it is expected that the numerical fitness trajectory has some lag behind the analytical trajectory. Vivo also mentions that the approximation is most appropriate in the long-term, which also suggests this may not me an appropriate fit for early time, which is where the gap between the two plots originates. Again, since we are focused on the rate of growth of fitness as opposed to the absolute fitness value, we look to the derivatives, which we calculate with finite differences exactly as done in the exponential case. This produces the following results for $\sigma = 0.05$ (results for $\sigma = 0.1$ are equivalent), which can be seen in Figure 10.

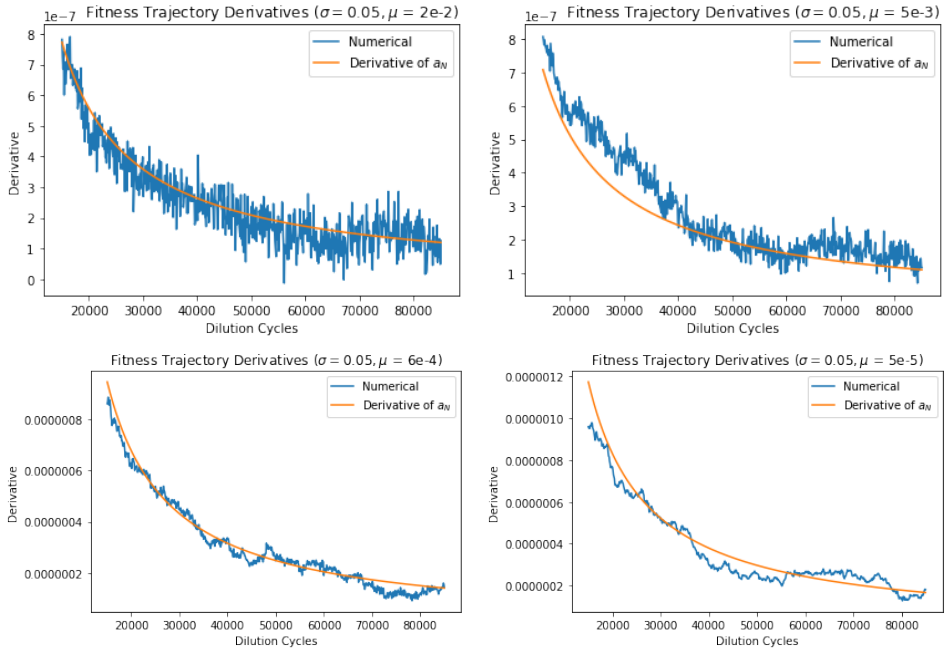

Figure 10: Fitness trajectory derivatives for Gaussian NFD; $\sigma = 0.05$

The plots in Figure 10 demonstrate that the analytical solution derived is a good fit for the long-term fitness trajectory rate of change in the case of the normal distribution. In particular, we see that the fit is accurate for all mutation values considered, from the SSWM range, when $\mu = 5 \times 10^{-5}$, to the clonal interference regime when $\mu = 2 \times 10^{-2}$.

6 Results: Number of Active Strains

A second data point considered is the number of strains present in the population at a given time when looking at simulations in the clonal interference regime. Finding a measure of the number of strains present for large time as a function of the parameters chosen can allow for calculating certain quantities in the clonal interference regime, in particular the expected fitness increment. In all simulations done with both exponential and Gaussian distributions, the plot of the number of strains as a function of time looks very similar qualitatively to the Figure 11. This image was produced from from a simulation in which an an exponential NFD was used, with $\lambda = 1, N_0 = 1,000, N_F = 100,000, F_0 = 1$, and $\mu = 1 \times 10^{-3}$.

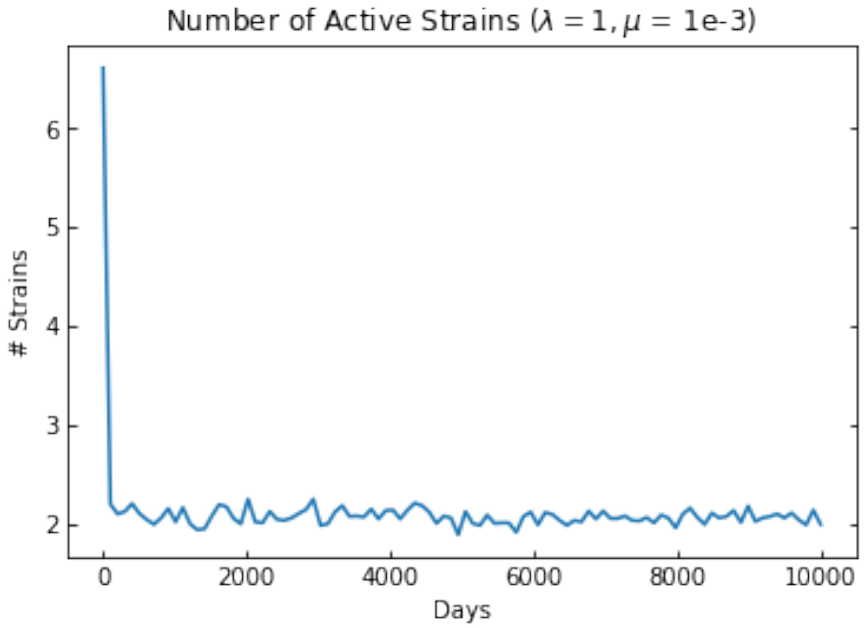

Figure 11: Num. strains for exponential NFD; $\lambda = 1, F_0 = 1, \mu = 1 \times 10^{-3}$

As can be seen in Figure 11, the number of strains has a brief peak at the very beginning, in which many beneficial mutations can occur and survive in the population. In the long term, however, it becomes much more difficult for this to occur, resulting in the dramatic drop. We note that this peak is simply a feature of the starting fitness F_0; if this value is increased, the simulation does not reach a peak number of strains, instead going straight to its steady state. This can be seen below in Figure 12, which was produced from a simulations in which an exponential NFD was used, with $\lambda = 1, N_0 = 100, N_F = 10,000, F_0 = 7$, and $\mu = 1 \times 10^{-4}$.

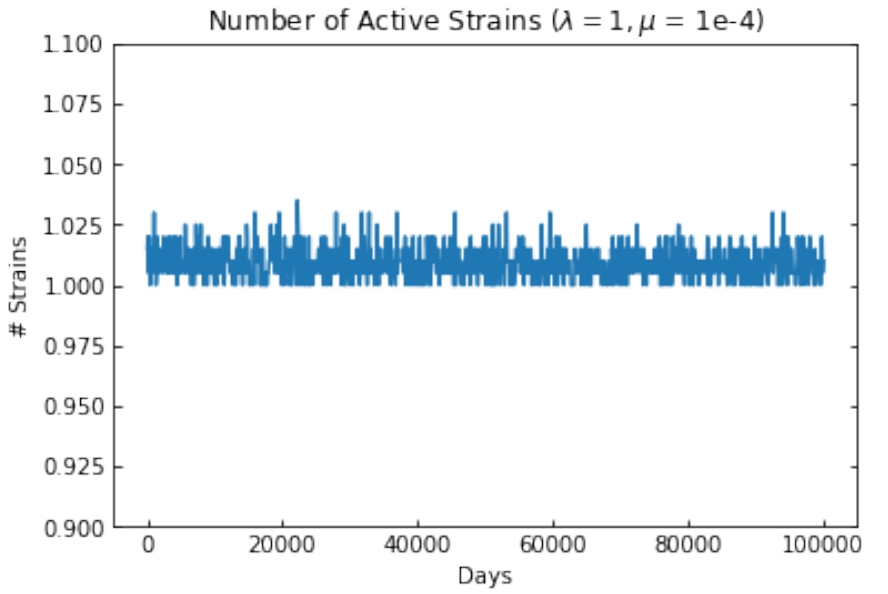

Figure 12: Clonal interference for exponential NFD; $\lambda = 1, F_0 = 7, \mu = 1 \times 10^{-4}$

Now we look to develop a relationship between the number of steady state strains as a function of mutation rate, size of the experiment, and choice of NFD. Data is gathered for a number of simulations using $\lambda = 1, N_0 = 100, N_F = 10,000$, and $F_0 = 1$ while varying the mutation rate μ. The asymptotic behavior can be seen below:

μ	Num. strains
4×10^{-4}	1.04
8×10^{-4}	1.085
2×10^{-3}	1.21
6×10^{-3}	1.63
1×10^{-2}	2.05
2×10^{-2}	3.1
3×10^{-2}	4.1
4×10^{-2}	5.2
5×10^{-2}	6.2

Table 1: Clonal interference measurement for exponential NFD

Putting this into a graph in Figure 13, the data seems to be linear with a slope of approx-

imately 100.

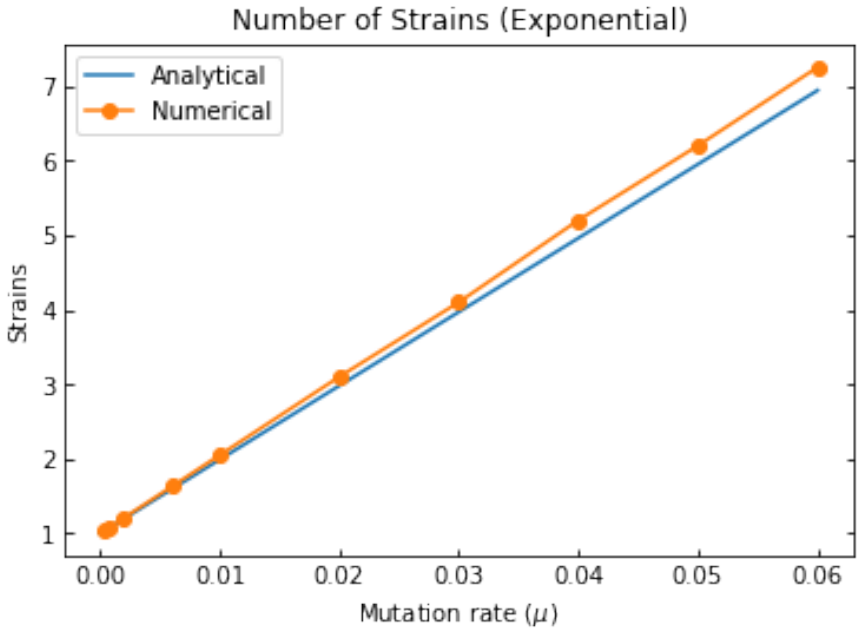

Figure 13: Clonal interference ($\lambda = 1, N_0 = 100, N_F = 10,000$)

A number of different simulations are done with exponential NFD's varying the exponential parameter λ, N_0, and N_F. The first test is done with different values of λ in the exponential distribution. As can be seen in Figure 14, the value of λ does not seem to affect the number of strains present.

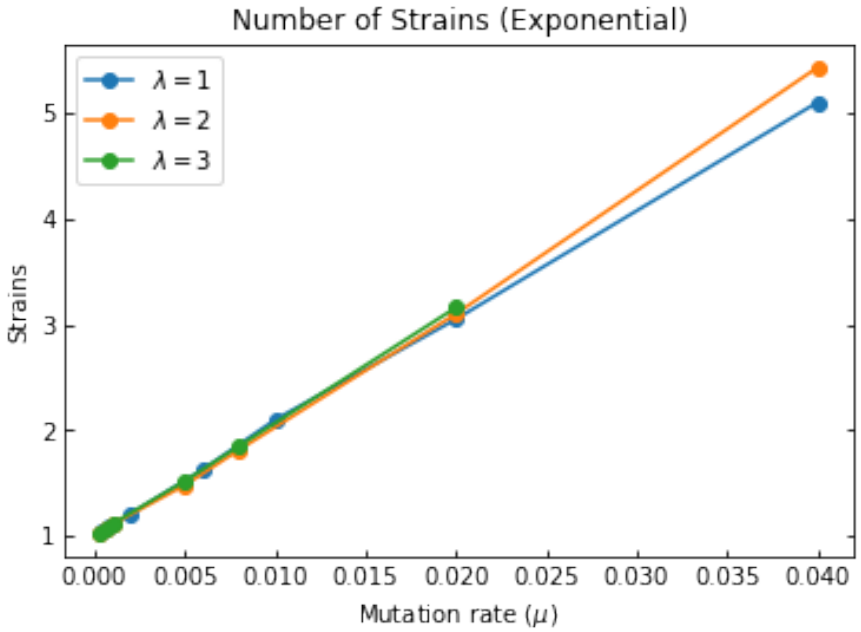

Figure 14: Clonal interference for exponential NFD; $\lambda = 1, 2, 3$

This confirms that the slope of this graph is independent of λ of this relationship. As expected, this slope changes proportionally with the size of the experiment. This can be seen by taking the above experiment and increasing N_0 and N_F by a factor of 4. The resulting plot in Figure 15 has a slope of approximately 400, 4 times greater than before.

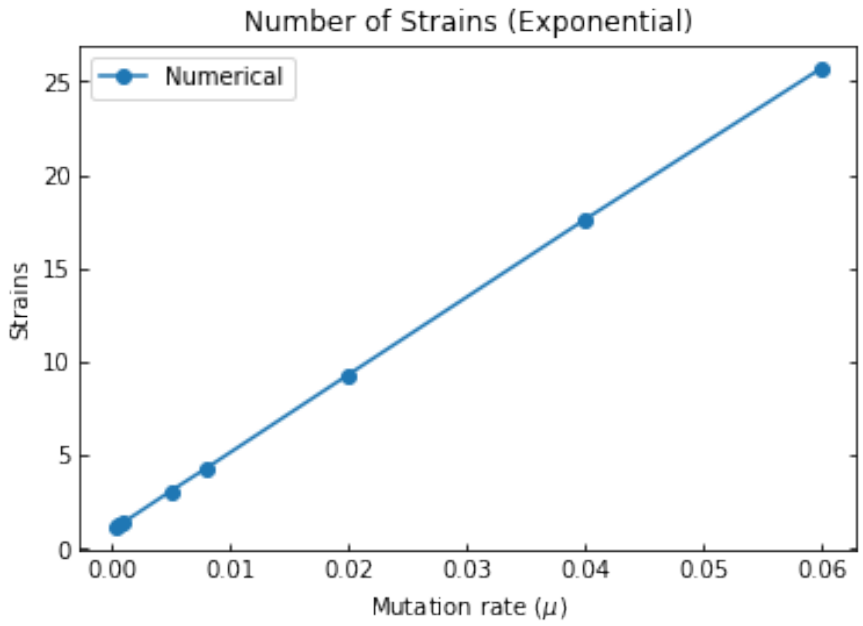

Figure 15: Clonal interference for exponential NFD; $\lambda = 1, N_0 = 400, N_F = 40,000$

It is also worth looking at this form in the case of a different NFD, namely the normal distribution. Data is gathered for a number of simulations using $\sigma = 0.1$, $N_0 = 100$, and $N_F = 10,000$ while varying the mutation rate μ, and can be seen in Figure 16:

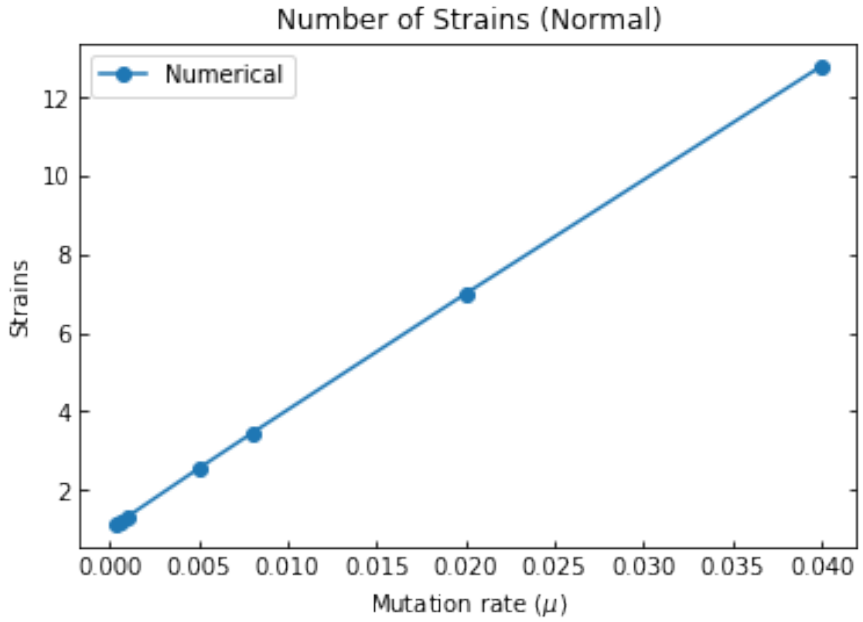

Figure 16: Clonal interference for Gaussian NFD; $\sigma = 0.1, N_0 = 100, N_F = 10,000$

Interestingly, despite the size of the experiment being the same as the exponential case mentioned above, the slope of the graph is clearly not the same. While the slope did not vary in the exponential case when λ was change, it did vary when the distribution was changed.

7 Results: Bacterial Competence

As mentioned previously, the expectation is that bacterial competence will improve the spread rate of beneficial mutations by allowing them to be shared within a generation, rather then only between generations. Thus, there should be a speedup in the fitness trajectory for a population. The hypothesis also suggests that the longer the strain allowed to be transferred, the faster the speedup. A longer fractional strain allows for a faster search of the possible genomes and thus can reach high fitness values faster. The goal of this work is to examine some of the relevant parameters in looking at the effects of competence. We will first consider the case without epistasis and then discuss the effects of epistasis.

It is worth discussing the relevant competence parameters in terms of how they relate to the real world and current scientific understanding. Values of the probabilities of decay, dropping, and uptake of strands are not generally well understood. Thus, for the purposes of this numerical simulation, those parameters are chosen to be sufficiently high such that the effects of competence are noticeable. The competence strand length l_{comp}, is similarly not well understood; for simulations, we choose to use on the order of approximately 1% of the entire genome length.

We begin by looking at a simulation done with no competence, *i.e.* $p_{\text{drop}} = p_{\text{take}} = 0$. We do a simulation with the following parameters:

- $L = 1,000$
- $N_0 = 100, N_F = 1,000$
- $\mu = 2.5 \times 10^{-6}$

The resulting fitness trajectory is plotted below in Figure 17:

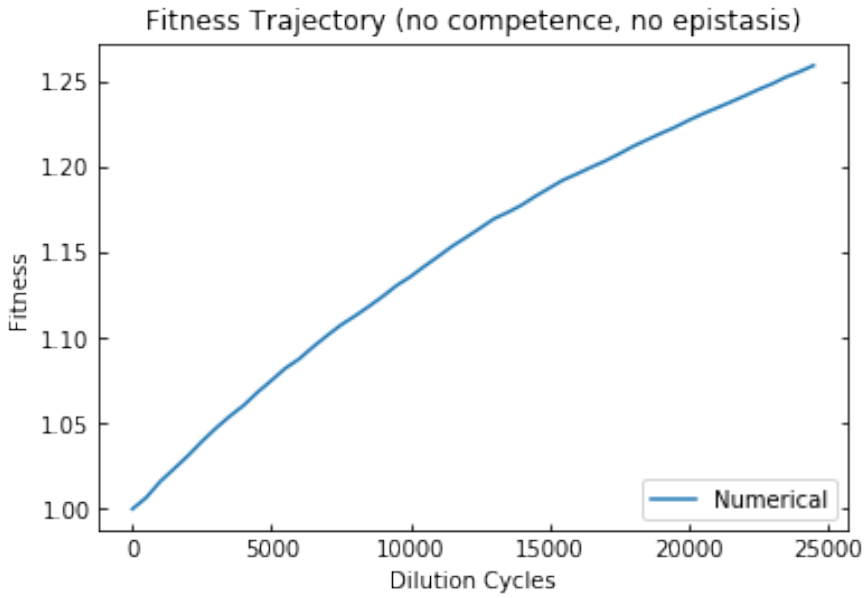

Figure 17: Fitness trajectory; no competence, no epistasis; $\mu = 2.5 \times 10^{-6}$

Competence is now introduced into the model, and the following parameters are set:

- $p_{\text{drop}} = 2.5 \times 10^{-2}$
- $p_{\text{take}} = 2.5 \times 10^{-2}$
- $p_{\text{decay}} = 1 \times 10^{-3}$

Here, we allow l_{comp} to vary. Running simulations with $L = 1,000, N_0 = 100, N_F = 1,000$, and $\mu = 2.5 \times 10^{-6}$ produces the plots shown in Figure 18.

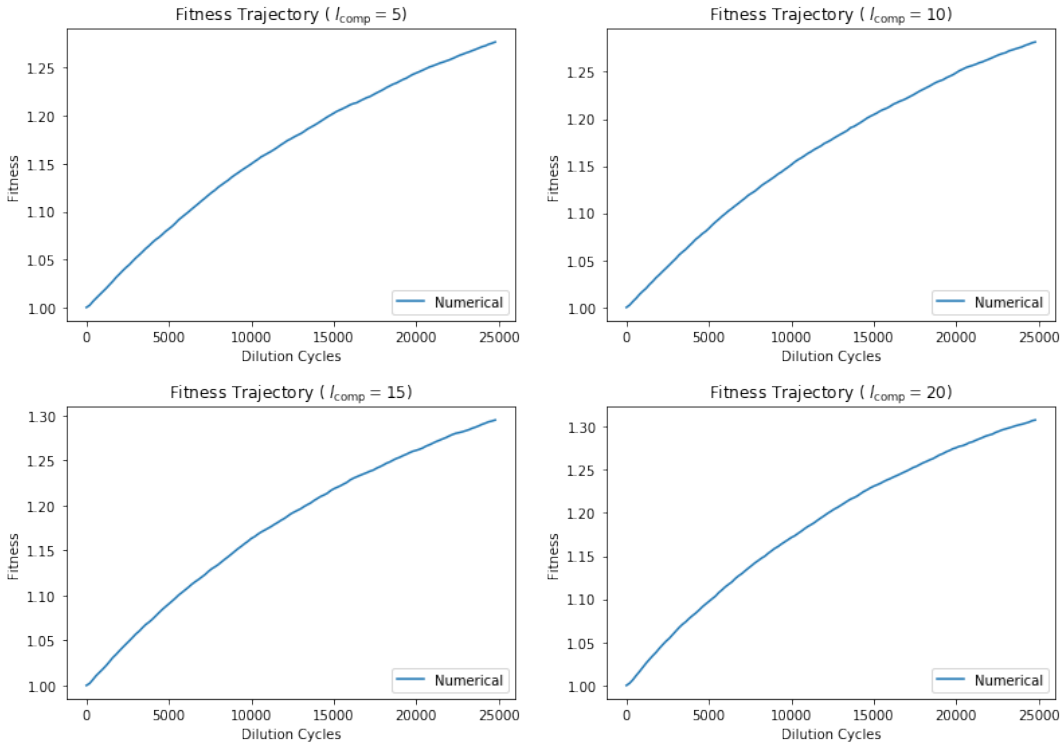

Figure 18: Fitness trajectories; competence on; epistasis off; $\mu = 2.5 \times 10^{-6}$

Figure 18 shows an speedup in fitness growth, which can be seen more clearly in Figure 19.

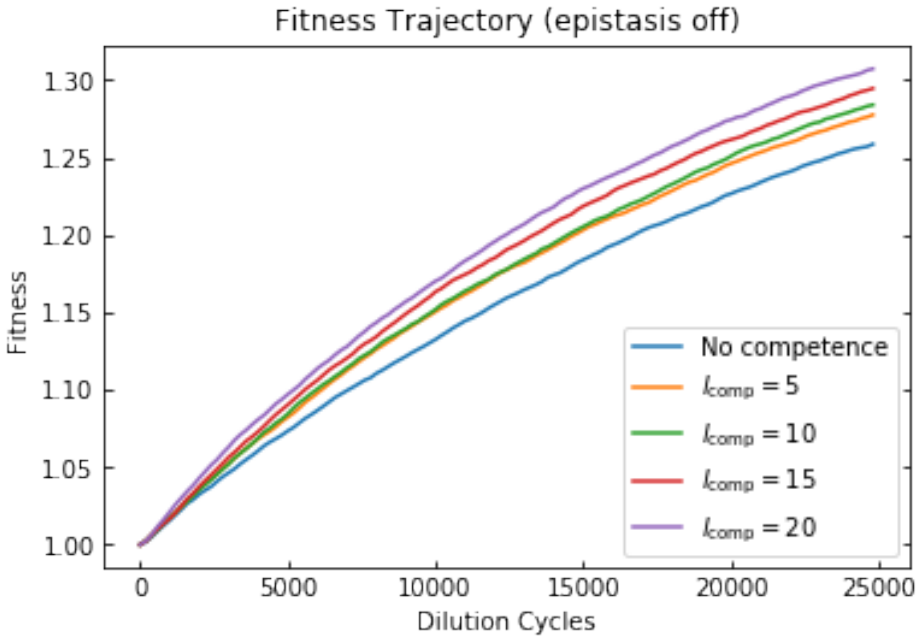

Figure 19: Fitness trajectories; competence on; epistasis off; $\mu = 2.5 \times 10^{-6}$

Figure 19 shows an increase in fitness after 25,000 dilution cycles for a greater strand length. Below is a table of this data:

l_{comp}	$F(25,000)$	Percent Increase
0 (no competence)	1.259	
5	1.277	1.43%
10	1.282	2.82%
15	1.295	2.95%
20	1.308	3.08%

Table 2: Fitness values after 25,000 days; epistasis off

The increases in fitness over time can be seen in Figure 20.

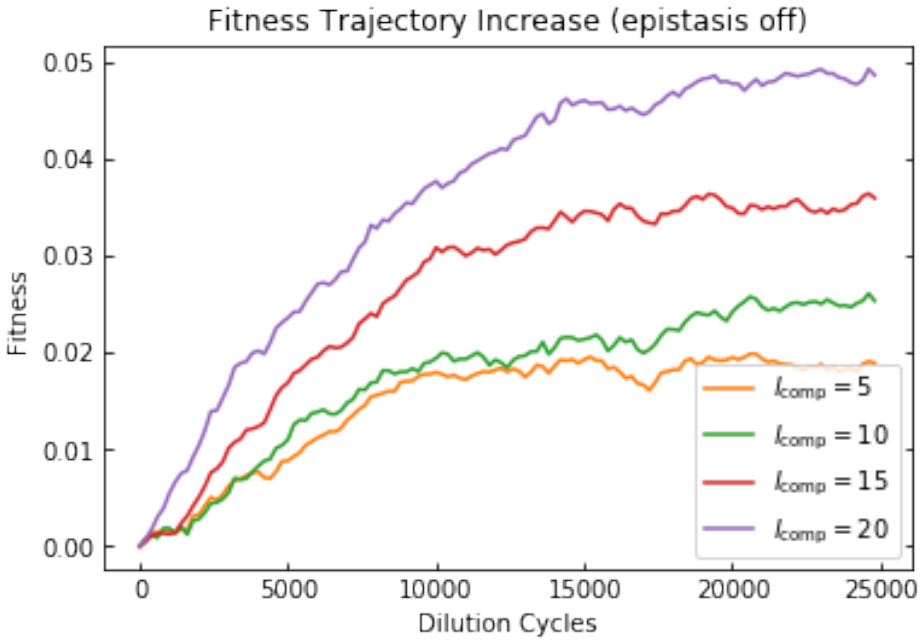

Figure 20: Fitness trajectory speedup; competence on; epistasis off; $\mu = 2.5 \times 10^{-6}$

These simulations show that, as expected, bacterial competence improves the rate of increase of fitness over time. Additionally, a greater strand length causes a greater increase, because more beneficial mutations can be transmitted at once.

Next, competence can be studied while epistasis is turned on within the model. First, we find our baseline trajectory with competence off. Figure 21 shows the fitness trajectory result

for a simulation done with the following parameters:

- $L = 1,000$
- $N_0 = 100, N_F = 10,000$
- $\mu = 2.5 \times 10^{-6}$

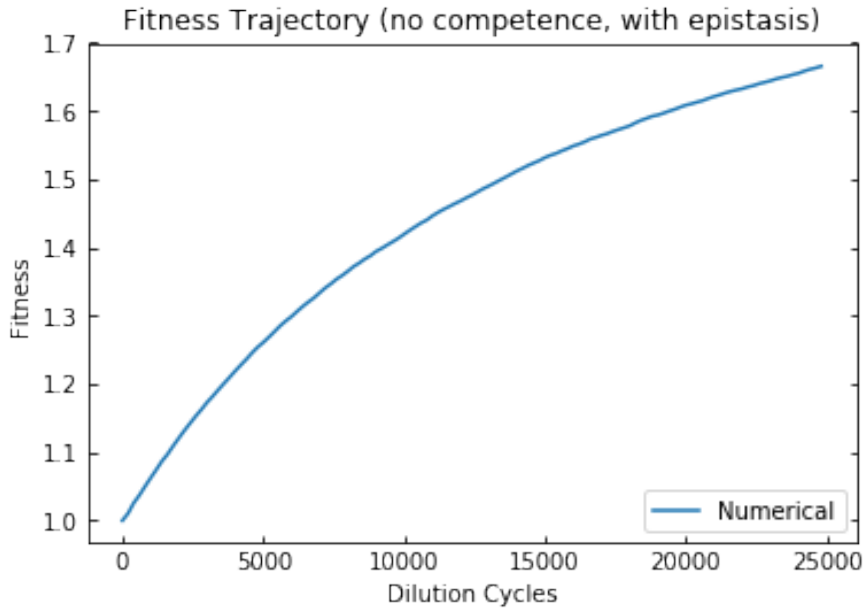

Figure 21: Fitness trajectory; competence off, epistasis on; $\mu = 2.5 \times 10^{-6}$

As before, competence is introduced with the same parameters, producing the plots in Figure 22.

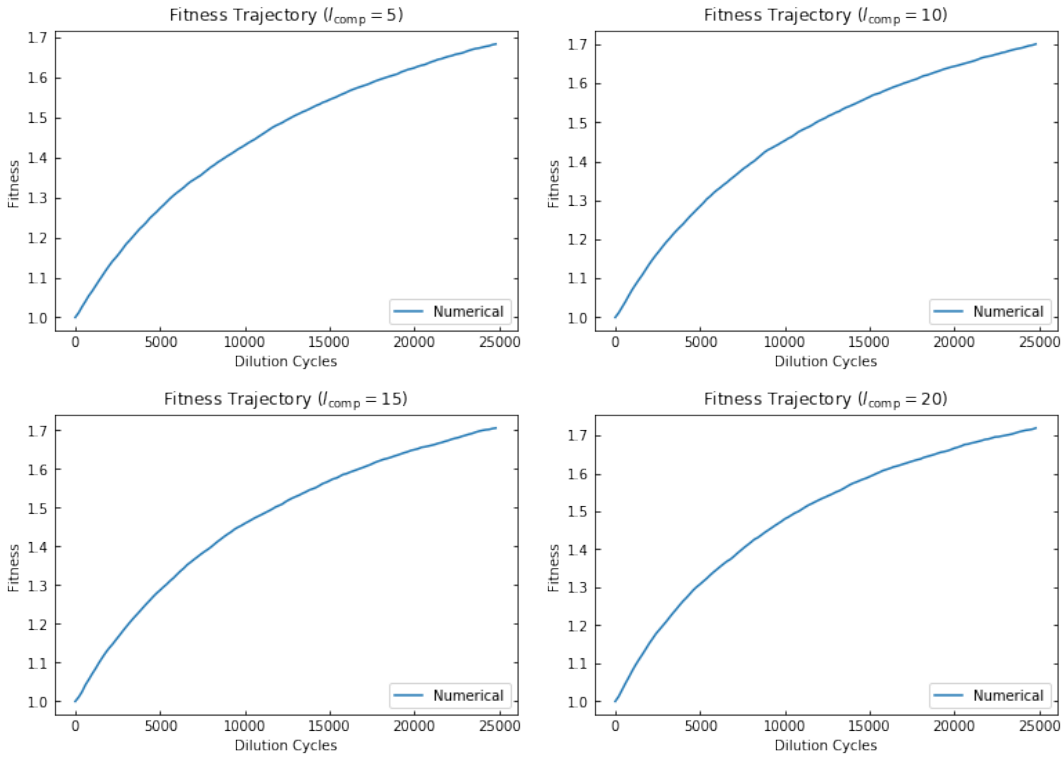

Figure 22: Fitness trajectories; competence on; epistasis on; $\mu = 2.5 \times 10^{-6}$

As before, we overlay the different trajectories to produce Figure 23.

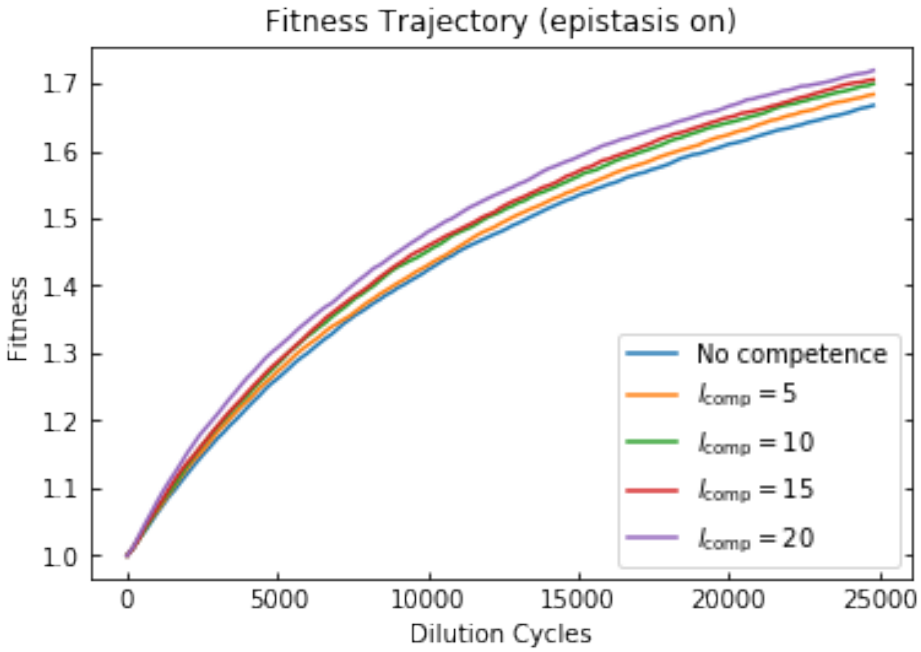

Figure 23: Fitness trajectories; competence on; epistasis on; $\mu = 2.5 \times 10^{-6}$

We can again create a table of the fitness values after $25,000$ days.

l_{comp}	$F(25,000)$	Percent Increase
0 (no competence)	1.665	
5	1.684	1.14%
10	1.701	2.16%
15	1.706	2.46%
20	1.719	3.24%

Table 3: Fitness values after $25,000$ days; epistasis on

As before, these fitness increases can be seen in Figure 24

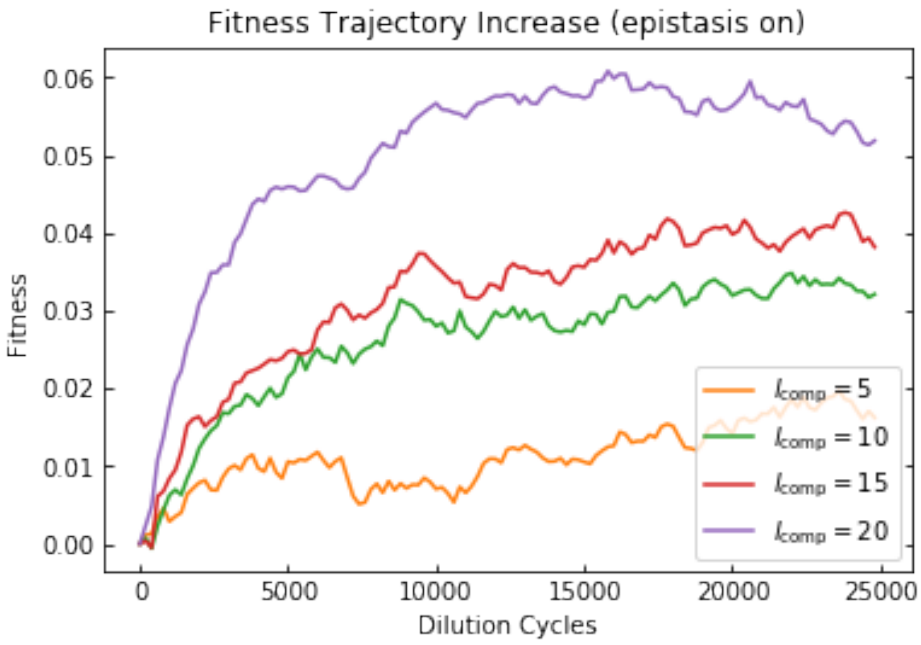

Figure 24: Fitness trajectory speedup; competence on; epistasis on; $\mu = 2.5 \times 10^{-6}$

We thus conclude that allowing for epistasis when competence is turned on leads to a similar increase in fitness as without epistasis. In both cases, we see noticeable increases in fitness values when competence is turned on that is proportional to the length of the competence fragment.

8 Conclusion

In this thesis, we use agent-based models of the Lenski framework to evaluate two different models of a genome. In the infinite-sites model, we present analytical formulas for fitness trajectories for exponential and Gaussian NFDs and show that for large time these forms are a good fit to data from simulations in measuring rate of increase of fitness. In future work, we would like to gain a better understanding of what happens at early time and determine how this can be expressed analytically. We would also like to look in more detail at the substitution trajectory, *i.e.* the number of mutations in the population as a function of time. Results for both long term fitness increase and number of active strains can also be extended to other probabilistic distributions, in particular distributions from the Gumbel class. In regard to

our measurement of clonal interference, we are interested in the development of an analytical form for the seemingly linear relationship between mutation rate and number of steady state strains.

In the spin glass model, we show that allowing for bacterial competence leads to a significant increase in fitness trajectories. Interestingly, we see relatively similar increases both with and without epistasis. We are interested in further experimentation with the parameters pertaining to competence. We would like to understand how these parameters control the effects of bacterial competence on fitness trajectories. This includes competence rates, the mutation rate, and the length of the shared fragments. We would also like to run larger simulations to further quantify the similarities and differences between competence with and without epistasis, particularly with higher mutation rates.

References

[1] Bacterial transformation. https://bio.libretexts.org/Bookshelves/Microbiology/Book%3A_Microbiology_(Boundless)/7%3A_Microbial_Genetics/7.11%3A_Genetic_Transfer_in_Prokaryotes/7.11B%3A_Bacterial_Transformation, 2017.

[2] Ariel Amir. Private communication.

[3] Ariel Amir, Yuval Oreg, and Yoseph Imry. On relaxations and aging of various glasses. *Proceedings of the National Academy of Sciences*, 109(6):1850–1855, 2012.

[4] Rowan D. H. Barrett, Leithen K. M'Gonigle, and Sarah P. Otto. The distribution of beneficial mutant effects under strong selection. *Genetics*, 174(4):2071–2079, 2006.

[5] Jeffrey E. Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, and Jihyun F. Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461:1243–1247, 2009.

[6] Jean-Pierre Claverys and Bernard Martin. Bacterial 'competence' genes: signatures of active transformation, or only remnants? *Trends in Microbiology*, 11(4):161–165, 2003.

[7] Vaughn S. Cooper and Richard E. Lenski. The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature*, 407:736–739, 2000.

[8] Adam Eyre-Walker and Peter D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8:610–618, 2007.

[9] Philip J. Gerrish and Richard E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102(0):127, March 1998.

[10] John H. Gillespie. A simple stochastic gene substitution model. *Theoretical Population Biology*, 23(2):202 – 215, 1983.

[11] Ryan T. Hietpas, Jeffrey D. Jensen, and Daniel N. A. Bolon. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901, 2011.

[12] J. F. C. Kingman. A simple model for the balance between selection and mutation. *Journal of Applied Probability*, 15(1):1–12, 1978.

[13] Sergey Kryazhimskiy, Gašper Tkačik, and Joshua B. Plotkin. The dynamics of adaptation on correlated fitness landscapes. *PNAS*, 106(44):18638–18643, 2009.

[14] Richard E. Lenski, Charles Ofria, Travis C. Collier, and Christoph Adami. Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400:661–664, 1999.

[15] 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.

[16] Stefany Moreno-Gámez, Robin A. Sorg, Arnau Domenech, Morten Kjos, Franz J. Weissing, G. Sander van Doorn, and Jan-Willem Veening. Quorum sensing integrates environmental cues, cell density and cell history to control bacterial competence. *Nature Communications*, 8:854, 2017.

[17] H. Allen Orr. The distribution of fitness effects among beneficial mutations in fisher's geometric model of adaptation. *Journal of Theoretical Biology*, 238(2):279 – 285, 2006.

[18] Su-Chan Park and Joachim Krug. Evolution in random fitness landscapes: the infinite sites model. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(04):P04014, apr 2008.

[19] Elizabeth Pennisi. The man who bottled evolution. *Science*, 342(6160):790–793, 2013.

[20] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, Dec 1975.

[21] Ane L. G. Utnes, Vidar Sørum, Nils Hülter, Raul Primicerio, Joachim Hegstad, Julia Kloos, Kaare M. Nielsen, and Pål J. Johnsen. Growth phase-specific evolutionary benefits of natural transformation in Acinetobacter baylyi. *The ISME Journal*, 9(10):2221–2231, 2015.

[22] Pierpaolo Vivo. Large deviations of the maximum of independent and identically distributed random variables. *European Journal of Physics*, 36(5):055037, 2015.

[23] Sébastien Wielgoss, Jeffrey E. Barrick, Olivier Tenaillon, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E. Lenski, and Dominique Schneider. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli. *G3: Genes, Genomes, Genetics*, 1(3):183–186, 2011.

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