UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics of Data

Final Dissertation

The effect of multi-nucleotide mutations on adaptive evolution

Internal supervisor

Sandro Azaele

External supervisor

Alexander Klug

Candidate

Michele Avella

Academic Year 2021/2022

Abstract

Multi-nucleotide mutations (MNMs) describe instances in which a single mutation event simultaneously alters more than one nucleotide site. Typically, these sites are close to each other. Studies estimate that about one in every 50-350 mutations is a MNM. Although these occur much less frequently, they offer the advantage that substantially more genotypes can be reached through one mutation event. Our study seeks to understand when and to which extent MNM mutations affect adaptive evolution. To pursue this research question, we employ Wright-Fisher simulations and use random and real fitness landscapes. Building on this, we derive analytical results that illustrate how relevant MNM are as a function of the mutation supply.

Summary

Multi-nucleotide mutations are not taken into account in adaptive evolution since they are considered to occur at a negligible rate. Despite this assumption, sequence comparisons and mutational events in laboratories have suggested that multi nucleotide events occur in nature with a higher rate than expected. In this thesis we are going to study the influence of multi nucleotide mutations on adaptive evolution. In particular we will focus on adjacent double nucleotide mutations.

In the first chapter we present the main theoretical and biological concepts. In section 1.1 we introduce the concept of fitness landscape. We quantitatively describe how populations evolve on the landscape and how the trajectories depend on the mutation supply which is the number of mutations per time step. In section 1.2 we introduce multi nucleotide mutations and adjacent double nucleotide mutations. We discuss about their rate and we briefly describe the two main advantages they have over single nucleotide mutation: a larger variability and the possibility of crossing fitness valleys. The project will focus on the first feature. We end the first chapter with section 1.3 where we describe the setup of the project i.e. the way we study the relevance of double mutations.

In the second chapter we introduce the *simple model*; a simplified model that enables us to gain a qualitative understanding of the dynamics and relevant details. We describe how to simulate it in section 1.2 and the analytical description in section 1.3. Then we show the results of this model assuming the same DFE for singles and doubles. We end the chapter by discussing how the number of beneficial mutants and the DFEs influence the dynamic.

In the third chapter we describe the Wright Fisher model. In section 3.1 we define the model and we discuss about the values of the variables of the system. At the end of the section we show some results. In section 3.2 we present two analytic approximation of the Wright Fisher model and we test their accuracy. We end the chapter with section 3.3 where we introduce the mean fitness jump and we study the influence of doubles on the total fitness. In the last chapter we simulate the Wright Fisher model on an empirical landscape. The landscape is of the TEM-1 gene. We show results and we discuss about the relevance of doubles.

Contents

Chapter 1

Intro

1.1 Fitness landscape

Determining the relationship between genotype and fitness is a fundamental question in evolutionary biology [1, 2, 3, 4]. Genotype is the the genetic constitution of an individual organism. Fitness is the reproductive success and reflects how well an organism is adapted to its environment. It is represented with a real positive number. The genotype-fitness map is a map from the discrete space of genotypes to the fitness. The shape of the genotype-fitness map has significant implications for how evolution proceeds. In order to study this implications Sewall Wright [5] introduced the concept of the fitness landscape (also known as the adaptive landscape). In this view, populations evolve by fitness increasing steps until they reach a fitness local peak. In fig 1.1 (left) there is a visualization of a fitness landscape where the genotypes are in the $x - y$ plane and the fitness on the z axis [6]. In this three-dimensional setting, we may picture the process of evolution as a form of hill climbing. Despite this, the space occupied by genotypes is both high-dimensional and discrete, and the representation of this space in three dimensions can be misleading. The use of graphs to represent genotypes is one approach that may be used to circumvent this problem. In fig 1.1 (right) we show a graph representation of a fitness landscape of 4 alleles. In this type of directed network, each node represents a genotype, and each edge connects two genotypes that are different from each other by a single mutation. Additionally, the edges lead in the direction of genotypes that have a higher fitness. Even this method of visually representing the genotype-fitness map has limitations and there are other methods which are more complex and more powerful [7].

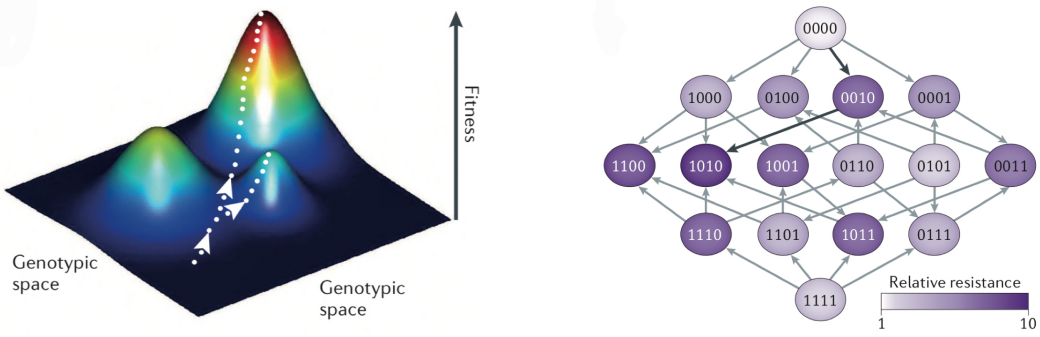

Figure 1.1: Fitness landscape visualization [6]. On the left a three dimensional visualization, on the $x - y$ plane the space of genotypes and on the z plane the fitness. The white dots represents two possible path starting from the same initial genotype. The evolution can be seen as uphill climbing. On the right a visualization on a graph. Nodes represent the $2^4 = 16$ genotypes; 0 and 1 indicate wild-type and mutant amino acids. Edges connect genotypes that differ by a single mutation and point towards genotypes with higher resistance. Bold black edge indicate the greedy walk from wild-type (0000) to the global maximum (1010).

1.1.1 Generating landscape

There are a variety of models that may be used to build a fitness landscape. In this project, we are going to make use of the House of Cards (HoC) model [8], which is characterized by the fact that all of the fitness values are picked at random from a distribution. This distribution is usually called *distribution of fitness effect* (DFE). More sophisticated models take also into account epistasis which is the interaction between two or more genes that has influence on the fitness of a genotype [9, 10, 11, 12, 13].

1.1.2 Weak mutation regime and clonal interference regime

Now we are going to describe how the population evolves on the fitness landscape and how the trajectory of evolution can depend on some parameters of the system. We define N as the population size. We define $\mu_{i,j}$ as the mutation rate from the genotype i to the genotype j. $U_i = \sum_i \mu_{i,j}$ is the rate of mutation of the genotype i. If we consider an homogeneous population, i.e. all the individuals have the same genotype i, we can define the mutation supply as $N \cdot U_i$ which is the expected number of mutations per generation.

Based on the value of the mutation supply we can define two different regimes:

- $N \cdot U_i \ll 1$ weak mutation regime [14, 15]. In this regime the population evolves through a random walk restricted on fitness-increasing steps. The probability of a path is proportional to the rate of the mutation times a monotonic function of the fitness of the mutant.

- $N \cdot U_i \sim 1$ clonal interference regime [16, 17]. In this regime we observe multiple mutants together adding competition to the dynamic. Competition favors mutants with a higher fitness. The probability of each path is again proportional to the rate of the mutation but more strongly dependent on the fitness of the mutant. If the mutation supply is such that we observe all the possible beneficial mutants the population evolves through a greedy walk. The evolution becomes deterministic and at each step evolution chooses the fittest mutant.

In figure 1.2 we show the evolution for three different values of the mutation supply assuming all the rates equal.

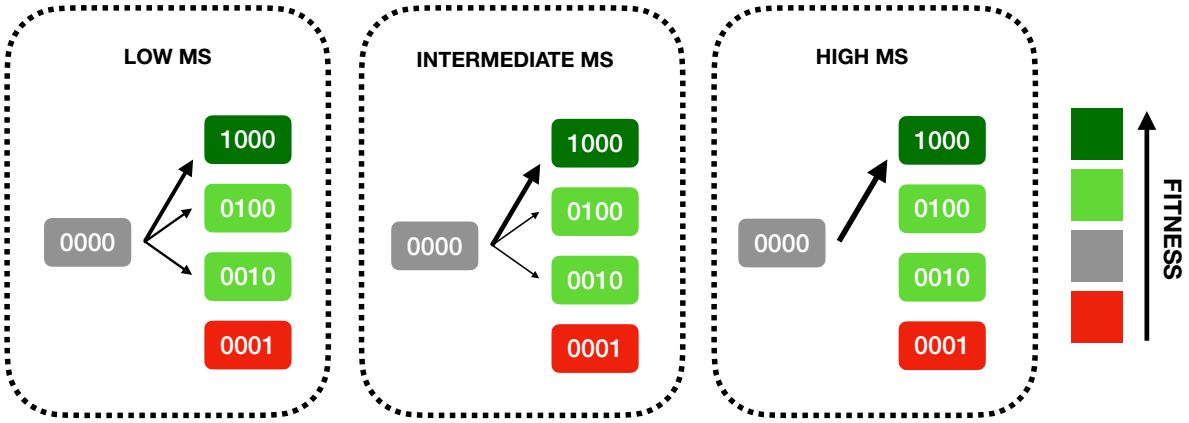

Figure 1.2: Evolution in different regimes. Each node is a genotype. On the right the legend with fitness ranking. The node on the bottom is deleterious. The thickness of the edge represent the probability of the path. The evolution can only increase fitness and so the bottom node is never reached. We are considering an homogeneous mutation rate. In the left panel we consider the week mutation regime. The path is a random walk biased to fitter genotypes. In the middle panel we are in the clonal interference regime. It is the same as before but now the bias is stronger. Right panel the mutation supply is such the dynamic becomes deterministic to the fittest genotype.

1.2 Multi nucleotide mutations

Multi-nucleotide mutations are not taken into account in adaptive evolution since they are considered to occur at a negligible rate. For instance, if we think of double nucleotide mutations, the rate of those mutations should be equal the square of the rate of single nucleotide mutations. In e-coli the single nucleotide mutation rate per base is $\mu = 10^{-10}$ [18, 19, 20]. Consequently, the probability of two mutations occurring by accident in the same generation is extremely low. Despite this assumption, sequence comparisons and mutational events in laboratories have suggested that multi nucleotide events occur in nature. In this paper [21], for example, the authors studied 283 parent-offspring trios. They estimated that the 3% of the mutations are part of a multi-nucleotide mutation. This MNM are more likely to happened in small cluster closer to each other. Other studies [22, 23, 24, 25] estimated this frequency and the results are always on the order of 1%. The cause of this kind of mutation is usually addressed to DNA polymerise zeta (Pol ζ) [26]. In this project we will focus on a subclass of multi nucleotide mutations which is the adjacent double nucleotide mutations. This kind of mutations involve two neighboring nucleotides. A few (old) papers [27, 28] estimated the rate of adjacent double mutations. We can also estimate the rate of those mutation from the results of multi nucleotide mutation studies. Based on all this results we are confident to consider adjacent double nucleotide mutations on the order of $0.1\% - 1\%$ of the total number of mutations.

Lets now discuss about the potential implications of adjacent double nucleotide (double) mutations on adaptive evolution. There are two main advantage of double mutations: larger variability of mutants, valley crossing.

1.2.1 Larger variability

If we consider a gene of length L, the number of possible single nucleotide substitutions is $L \cdot 3$ and the number of possible adjacent double nucleotide substitutions is $(L-1) \cdot 3 \cdot 3$. Therefore substantially more genotypes can be reached via adjacent double nucleotide substitutions.

Consider now a gene that encodes for an amino acid. The mutations that cause a change in an aminoacid's sequence are the ones that can have an impact on the fitness of a gene. Synonymous mutations are mutations in a nucleotide sequence that do not alter the encoded amino acid. For example a mutation from the codon ATT to the codon ATC does not change the fitness of the genotype since both codons encode the same aminoacid. This kind of mutation do not alter the fitness of a gene. Single

nucleotide mutation changes only a base of a codon. Because of that the average number of accessible amino-acids per codon through single nucleotide mutations is 6. Double adjacent nucleotide mutations change two neighboring nucleotides that can be inside a codon or across to neighboring codons. This can lead to a single amino-acid substitution but also to a double amino-acid substitution. The average number of accessible amino-acids per codon through double nucleotide mutations is 16. This number includes double amino-acid substitutions. As said in the previous part the rate of double mutations is 10^{-3} to 10^{-2} times the rate of single mutations. As a result, a genotype that can be reached by both types of mutations is more likely to arise as a result of a single nucleotide mutation. For this reason we will not consider this mutants as double mutants. Thus the average number of accessible amino-acids per codon through double nucleotide mutations becomes $16 - 6 = 10$. To clarify this part we have made a schematic representation (fig 1.3) of the accessible amino-acids through single and double mutations for a given codon.

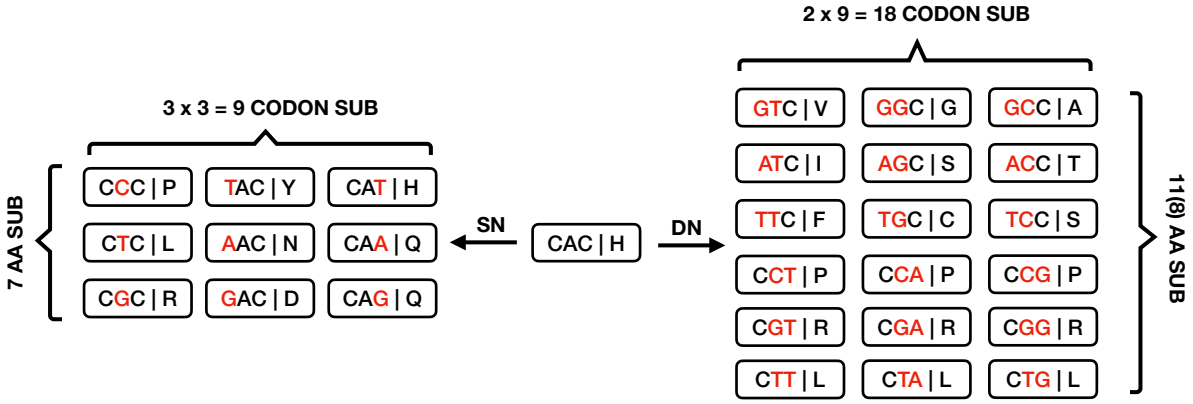

Figure 1.3: Acessible aminoacid. We compute all the 9 possible single nucleotide mutations and all the 18possible double nucleotide mutations of a codon. The number of accessible aminoacids is 7 for single and 11 for double. If we consider the number of aminoacids uniquely accessible via double the number becomes 8.

Let's now focus on a one-step dynamic. Consider a scenario in which we have a monomorphic population evolving in a House of Cards landscape (all fitness values are random). We want to investigate the next step of evolution. Because of what we said before, the average number of unique non-synonymous double mutants is 5/3 the number of unique non-synonymous single mutants. In a HoC landscape, the fact that there are more possible double mutants than there are single mutants implies that it is more likely that the fittest mutant is a double. This is one of the most crucial aspects of this project. For example, consider the evolution of a population on a fitness landscape under clonal interference. Before we have qualitatively shown that the probability of

fixation of a mutant is proportional to his rate times a term that is a monotonic function of the fitness. The higher is the mutation supply the larger is the value of the second term. As a consequence evolution drives the population to the fittest genotype. Double mutants have a lower rate but, since the number of doubles is larger than the number of singles, the fittest genotype is more likely to be a double. This trade-off between mutation rate and fitness in doubles is what we are going to study in the next chapters. What we expect is that in the week mutation regime the probability of fixation of the doubles will be on the order of μ_D/μ_S while in the clonal interference the doubles will become more and more relevant.

1.2.2 Valley crossing

We briefly comment the effect double mutations on valley crossing. This is not going to be a detailed examination; however, we want to give this argument that supports our hypothesis in a qualitative manner. With double mutations evolution can follow path that are inaccessible considering only single nucleotide (single) mutations [29]. Consider the case in which the population is in a local peak of single mutation, i.e. all single mutations decrease the fitness. In this scenario there could be double mutations that increase fitness and so the population can evolve to a new peak. In general we can say that double mutations reshape the fitness landscape connecting more genotypes and allowing the population to evolve to higher peaks. In fig 1.4 an example of what we just described. We can also quantify this effect by computing the average number

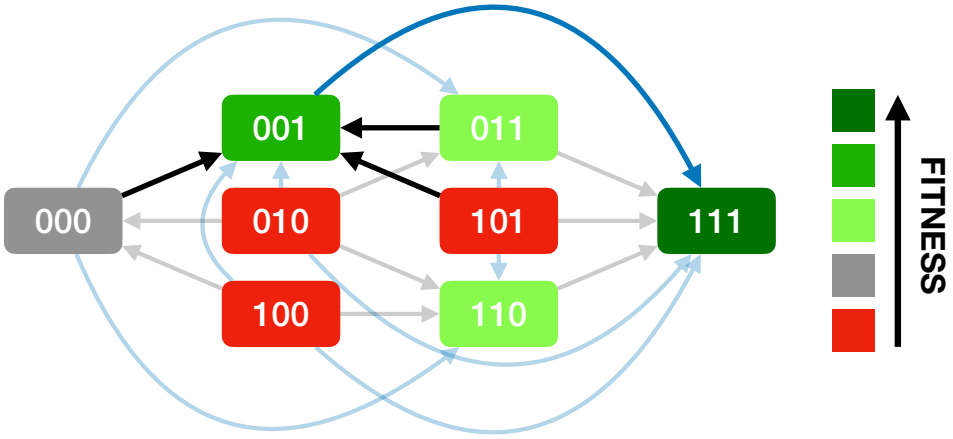

Figure 1.4: Valley crossing. Each node is a genotype. On the right the legend with fitness ranking. Genotype 001 is a local peak since all the single nucleotide mutations (black edges) lead to a lower fitness. If all the population is 001 it can not evolve via single mutations. Double nucleotide mutations (blue edge) allow genotype 001 to evolve to the fittest genotype 111.

of beneficial double mutations on a local peak. We generate a HoC landscape and we evolve the population through single nucleotide mutations until a local peak and then we compute the number of beneficial double in the peak. If $n_D/n_S = 5/3$ the number of double beneficial in the peak is around 1. Because of this, double mutations can become highly relevant when a population is stuck on a local peak because they are the only choice that is accessible. In figure 1.5 we show the distribution of the number of beneficial doubles in a local peak. This distribution is computed by simulating a population evolving on a random landscape. At each step the population evolves to a random beneficial single mutant until the peak is reached (all the mutants are deleterious). Once the population is on a local peak we compute the number of beneficial doubles fitter that the peak. As we can see, half of the time the population can escape from the local peak via adjacent double nucleotide mutations.

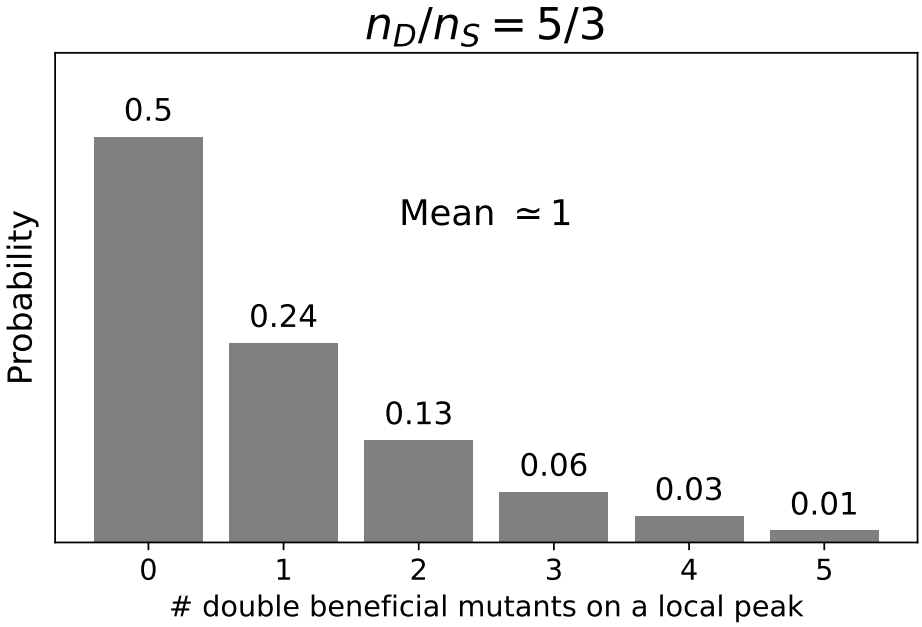

Figure 1.5: Distribution of number of double beneficial mutants on a fitness peak. We compute the distribution by simulating a population evolving on a random landscape. Once the population reaches a local peak we compute the number of beneficial double mutants.

1.3 Setup

We can now discuss about the setting of our study. We limit our analysis to a one-step dynamic, which means that we are only interested in tracing the evolutionary path

from wild-type (initial genotype of the population) to the next genotype. We impose this constraint for two reasons: the first is to simplify the dynamic, which will allow us to acquire better analytical conclusions; the second is to facilitate the use of empirical fitness data.

We assume a monomorphic initial population. For each genotype reachable through single and double nucleotide mutations we calculate its the probability of fixation. Our goal is to establish how significant double-nucleotide mutations are to the evolutionary process. In order to accomplish this, we define the following quantities:

- P_i = probability that the mutant i fixates first.
- $F_{DN} = \sum_{DN} P_i$ = probability that a double mutant fixates first. This is the probability that evolution goes through double mutations.
- f = fraction of double mutation. The number of double mutations over the total number of mutations.

The goal of this project is quantify F_{DN} and its dependence on the mutation supply. As said before we expect that double mutations, and thus F_{DN}, become relevant in the clonal interference regime where evolution is strongly biased to the fittest genotypes. We will first present a simple model to introduce the main features of this problem. Later we will present the Wright Fisher model which is a common population genetic model broadly used to study evolution. Lastly, we will present an empirical fitness landscape.

Chapter 2

Simple model

Before employing a population genetic model, we take a step back and consider this problem using a simplified model. This *simple model* enables us to gain a qualitative understanding of the dynamics and relevant details. It is mathematical much more tractable and shows what could potentially be expected in the population genetic model. The main limitations of the simple model are the following:

- Genetic drift is neglected. We discuss this limitation at the end of this chapter.
- There is no explicit dependence on the population size N which is related to the absence of genetic drift.
- Each generation is considered independently from previous ones.

2.1 Definition

In the following the core idea of the simple model is described. We assume an initially homogeneous population and observe all the mutations arising in a given time interval t. Of those mutants, the fittest one wins. This can be simulated with the steps listed below:

1. We draw a single fitness landscape $[\omega_1^S, ..., \omega_{n_S}^S] \sim f_S(\omega)$ and a double fitness landscape $[\omega_1^D, ..., \omega_{n_D}^D] \sim f_D(\omega)$ from two distributions $f_S(\omega)$, $f_D(\omega)$.
2. For a given time interval t the following is repeated for several runs:
   (a) A random number of single mutations $m_S \sim \text{Poisson}(\lambda_S = t \cdot N \cdot n_S \cdot \mu_S)$ and double mutations $m_D \sim \text{Poisson}(\lambda_D = t \cdot N \cdot n_D \cdot \mu_D)$ is drawn.

(b) We sample with replacement m_S mutations from the list $[\omega_1^S, ..., \omega_{n_S}^S]$ and m_D mutations from $[\omega_1^D, ..., \omega_{n_D}^D]$.

(c) The fittest mutation wins.

3. We compute $F_{DN} = \frac{\#\text{DN wins}}{\#runs}$.

Figure 2.1 also shows a schematic representation of this process.

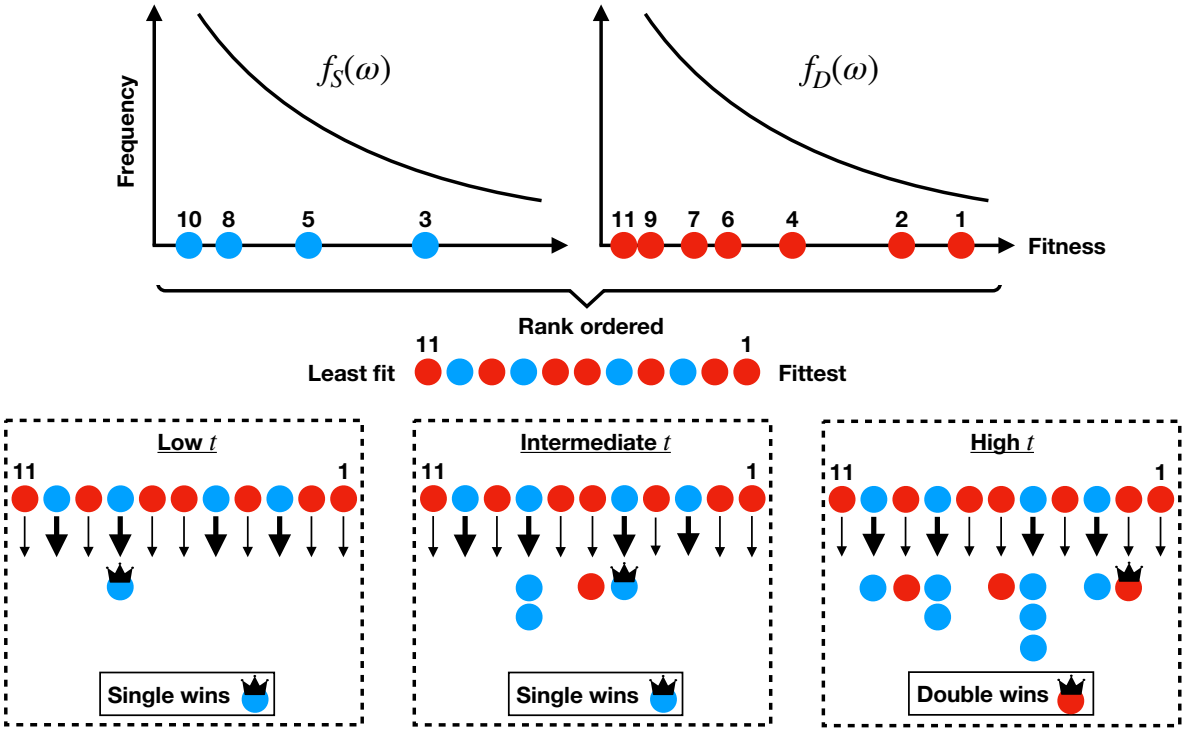

Figure 2.1: Simple model schematic. We draw single fitness landscape from distribution $f_S(\omega)$ and double fitness landscape from $f_D(\omega)$. We rank the fitness values from the highest to the lowest. Depending on the time interval t, the population N and the mutation rates μ_S, μ_D we draw with repetition singles and doubles. The fittest wins. We repeat this for multiple runs and we compute compute $F_{DN} = \frac{\#\text{DN wins}}{\#runs}$.

2.2 Analytical description

The model can also be expressed analytically. For this, we rank $[\omega_1^S, ..., \omega_{n_S}^S]$ and $[\omega_1^D, ..., \omega_{n_D}^D]$ from fittest to least fit and define:

$$P_i^S(Nt) = (1 - e^{-t \cdot N\mu_S}) \cdot e^{-t \cdot N\mu_S \cdot (i-1)}, \tag{2.1}$$

$$P_i^D(Nt) = (1 - e^{-t \cdot N\mu_D}) \cdot e^{-t \cdot N\mu_D \cdot (i-1)}. \tag{2.2}$$

These define the probability of observing at least once the i'th fittest mutant but zero times all fitter mutants of the same class (SN/DN). Using $P_i^S(Nt)$ and $P_i^D(Nt)$ we define the vectors $\vec{p}_S(Nt)$ and $\vec{p}_D(Nt)$:

$$\vec{p}_S(Nt) = [P_1^S, P_2^S, ..., P_{n_S}^S] \tag{2.3}$$

$$\vec{p}_D(Nt) = [P_1^D, P_2^D, ..., P_{n_D}^D] \tag{2.4}$$

Then F_{DN} can be computed by summing the probability of winning for all double mutants. The probability that double mutant j wins is the probability P_j^D that j is the fittest drawn double times the sum of the probabilities that the fittest single is i but considering only $i : \omega_i^S < \omega_j^D$:

$$P_j^D(\text{wins}) = P_j^D \cdot \sum_{\omega_i^S < \omega_j^D} P_i^S \tag{2.5}$$

$$F_{DN}(Nt) = \sum_j P_j^D(\text{wins}) \tag{2.6}$$

Additionally, we can represent this by using matrix notation:

$$\mathbb{H} \to H_{i,j} = \begin{cases} 1 & \text{if} \quad \omega_i^S < \omega_j^D, \\ 0 & \text{if} \quad \omega_i^S > \omega_j^D. \end{cases} \tag{2.7}$$

Consequently, F_{DN} may be computed as:

$$F_{DN} = \vec{p}_S^T \cdot \mathbb{H} \cdot \vec{p}_D + (1 - e^{-t \cdot N n_D \mu_D}) \cdot e^{-t \cdot N n_S \mu_S}, \tag{2.8}$$

where the last term is the probability of observing at least one DN mutant but zero SN mutants. In figure 2.2 we compare the simulation with the analytic prediction; the two lines overlap perfectly. Notice that $\vec{p}_S$ and $\vec{p}_D$ are independent of the fitness landscape. In fact, information about the fitness landscape is contained in the matrix $\mathbb{H}$.

2.3 General results

We start with a simple case, assuming that singles and doubles have the same DFE $f_S(\omega) = f_D(\omega)$. For 1000 different random landscape we compute $F_{DN}(Nt)$ and their average $\langle F_{DN}(Nt) \rangle$, see fig 2.3. There are three interesting points that we can observe in the panels:

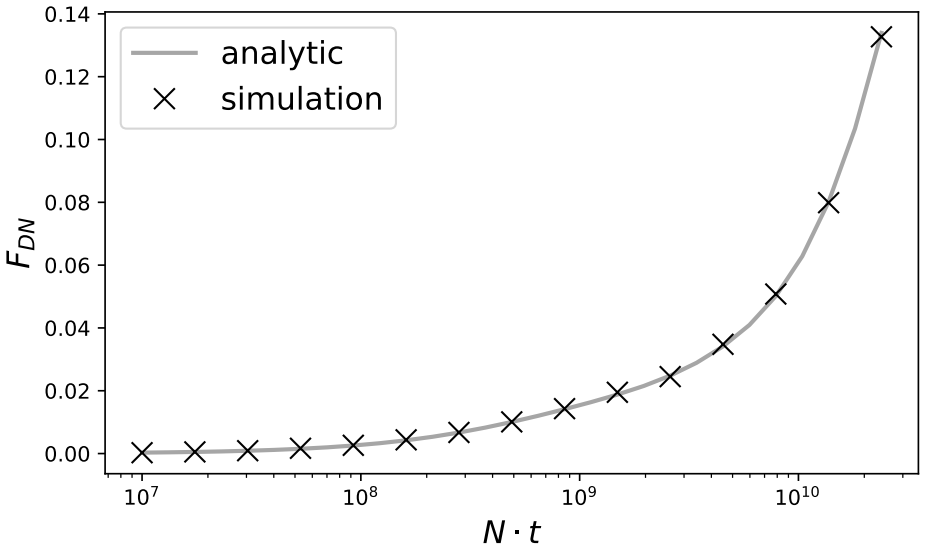

Figure 2.2: Comparison between simple model simulation and analytic prediction. As we can see the analytic prediction matches perfectly with the simulation.

a) The average behavior of F_{DN} (black line) is monotonically increasing.

b) The lines converges into different classes of convergence.

c) Some of the lines show a non-monotonic behavior.

Point a)

This is the most relevant aspect of our study. In grey we can see different landscapes and in black the average over 1000 random landscapes. Starting from the left, for small values of Nt (small mutation supply) all lines overlap. This is because in this regime we observe at most one mutation and so it is unlikely to have competition between doubles and singles. For this reason F_{DN} does not depend on the details of the landscape. This regime ends when Nt is such that at least one single mutant is always observed:

$$1 = n_S(1 - e^{-Nt\mu_S}) \tag{2.9}$$

$$Nt = \mu_S^{-1} \cdot \log(\frac{n_S}{n_S - 1}). \tag{2.10}$$

As a result from that point on competition is relevant for the dynamic, and so is the specific landscape.
For large values of Nt all the lines collapse into main convergence lines (see point b for details) and, more important, the average $\langle F_{DN} \rangle$ increases monotonically. This regime

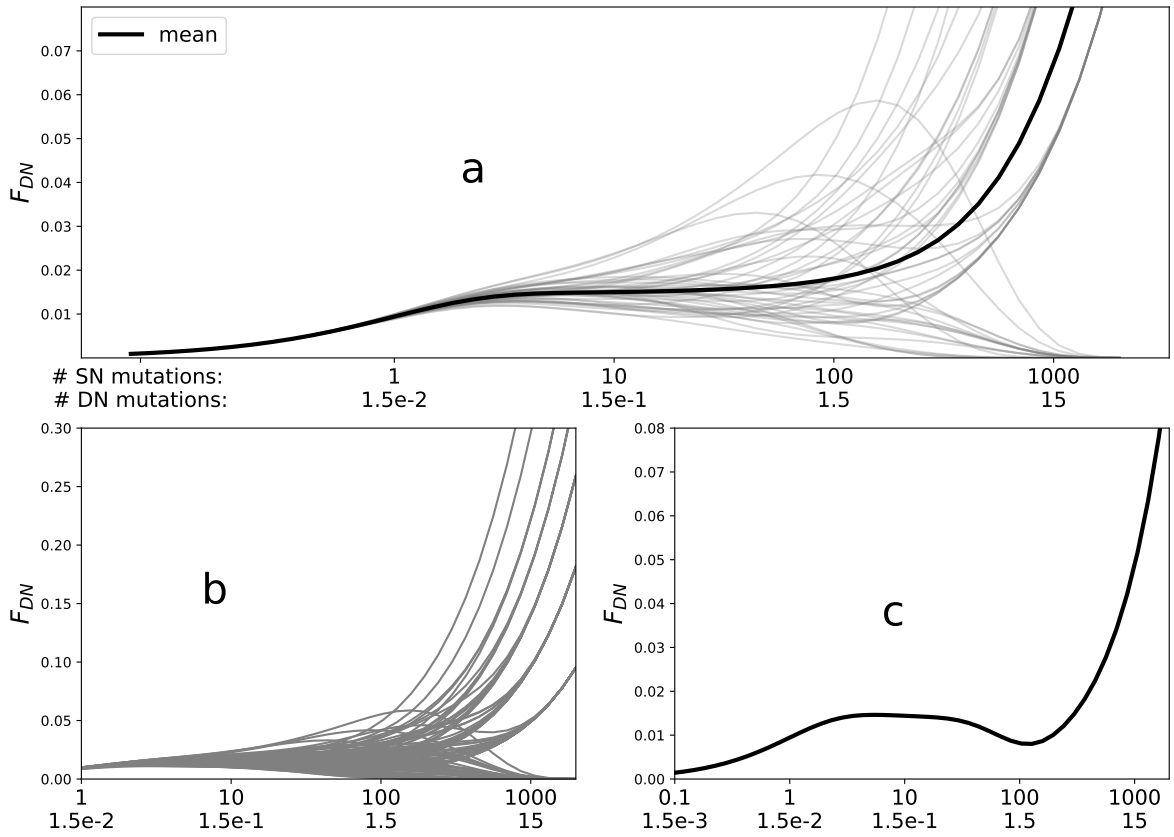

Figure 2.3: Probability of fixation of a double mutant (F_{DN}). Panel a: F_{DN} for several random landscape and their mean (solid black line). Panel b: F_{DN} for several random landscape with a wider y range, all the lines converge to different clusters for large values of t. Panel c: example of non-monotonicity.

begins when the singles start to saturate, meaning that we almost always observe all single mutants. Because of that the competition is only between the fittest single and the fittest double. If the fittest double is larger than the fittest single F_{DN} goes to 1, if the opposite it goes to 0. To compute the value of the average $\langle F_{DN} \rangle$ for large values of Nt we introduce the concept of order statistics [30]. Let $X_1, X_2, ..., X_n$ be iid random variables with a distribution f. We can relabel these Xs such that their labels correspond to their ranking from lowest 1 to highest n: $X_1 < X_2 < ... < X_n$. We define $f_{(k)}(x)$ as the distribution of the kth order statistic of a sample of n elements drawn from the distribution f:

$$f_{(k)}(x) = \frac{n!}{(k-1)!\,(n-k)!} f(x) F(x)^{k-1} \left(1 - F(x)\right)^{n-k} \tag{2.11}$$

So for example $f_{(1)}(x)$ is the distribution of the lowest element of a sample of size n and $f_{(n)}(x) = nf(x)F(x)^{n-1}$ is the distribution of the highest element of a sample of size n. For large value of t the competition is between the fittest single and the fittest double. We can predict the behavior of the average $\langle F_{DN} \rangle$ by studying the distribution of the fittest single $f^S_{(n_S)}(\omega)$ and of the fittest double $f^D_{(n_D)}(\omega)$:

$$\lim_{t\to\infty} \langle F_{DN} \rangle = P(D > S) = \int_0^\infty \int_y^\infty f^S_{(n_S)}(y) \cdot f^D_{(n_D)}(x)\, dy\, dx \tag{2.12}$$

$$\text{with} \quad S \sim f^S_{(n_S)}(\omega), D \sim f^D_{(n_D)}(\omega). \tag{2.13}$$

The easiest case is $f_S = f_D$ (the two DFEs are equal), if we compute eq 2.12 we obtain:

$$\lim_{t\to\infty} \langle F_{DN} \rangle = P(D > S) = \frac{n_D}{n_D + n_S}. \tag{2.14}$$

Since $n_D > n_S \Rightarrow \langle F_{DN} \rangle > 0.5$ and so it is more likely that the fittest double is fitter than the fittest single. If $f_S \neq f_D$ the problem may get more tricky, especially if the two distributions belong to different classes. However, it is always possible to find a numerical solution of equation 2.12.

So far we have given an analytic description of the behavior of F_{DN} and $\langle F_{DN} \rangle$ for small and large values of t. Having a rigorous description of the range in between is non trivial. The random landscapes have different behaviors while the average $\langle F_{DN} \rangle$ is flat in the first part and that increases. We can explain the flatness of $\langle F_{DN} \rangle$ with a qualitative argument. In this regime singles are not saturated yet and so the competition is between average numbers of unique doubles u_D doubles and average numbers of unique singles u_S:

$$u_S = n_S(1 - e^{-Nt\mu_S}); \quad u_D = n_D(1 - e^{-Nt\mu_D}). \tag{2.15}$$

If we assume that equation 2.14 holds also for real values of n_D, n_S we can compute $\langle F_{DN} \rangle$ as:

$$\langle F_{DN} \rangle = \frac{u_D}{u_D + u_S} = \frac{n_D(1 - e^{-Nt\mu_D})}{n_D(1 - e^{-Nt\mu_D}) + n_S(1 - e^{-Nt\mu_S})}. \tag{2.16}$$

In this regime Nt is still small and we can expand the previous equation at the first order:

$$\langle F_{DN} \rangle \simeq \frac{n_D \cdot Nt \cdot \mu_D}{n_D \cdot Nt \cdot \mu_D + n_S \cdot Nt \cdot \mu_S} = \frac{n_D\mu_D}{n_D\mu_D + n_S\mu_S}. \tag{2.17}$$

So both u_S and u_D grow linearly with Nt and so their ratio is constant. As soon as u_S starts growing sublinearly the slope of $\langle F_{DN} \rangle$ becomes positive.

Since this model is an oversimplification we can not derive quantitative results from this

analysis but still it suggests that the frequency of fixation of double adjacent nucleotide mutations can be non-negligible for large value of Nt.

Point b)

In this point we want to explain why all lines collapse to different convergence lines. The behavior of F_{DN} does not depend on the absolute values of $[\omega_1^S, ..., \omega_{n_S}^S]$ and $[\omega_1^D, ..., \omega_{n_D}^D]$ but only on the total rank. If we have n_S and n_D the number of possible ranks is:

$$\#\text{fitness landscapes} = \frac{(n_S + n_D)!}{n_S! \cdot n_D!} = \binom{n_D + n_S}{n_S}. \tag{2.18}$$

If $f_S(\omega) = f_D(\omega)$ each landscape is equiprobable. Since we only care about the rank we can collect fitness values in clusters as shown in figure 2.4. The rate of each cluster

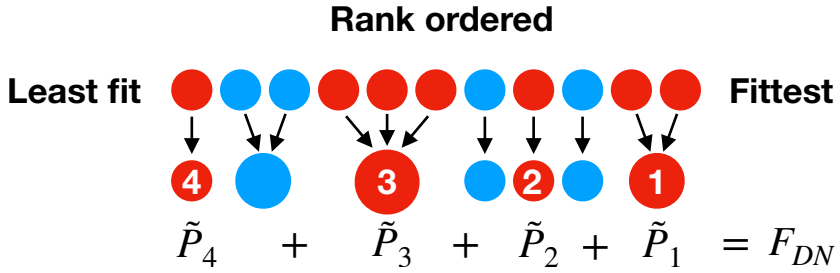

Figure 2.4: Clustering. Since in the simple model only the ranking matter, we rank the genotypes from the fittest to the least fit. Then we collect the neighbor genotypes of the same class (single/double) in clusters. We consider a cluster as a single genotype with rate proportional to the cluster size. If $\tilde{P}_i$ is the probability of the ith cluster to win, we can write F_{DN} as the sum of all the $\tilde{P}_i$.

is proportional to his size. If $\tilde{P}_i(Nt)$ is the probability of the cluster i to win, we can write F_{DN} as the sum of $\tilde{P}_i(Nt)$ for each i. For large values of Nt we almost always draw the fittest single. For this reason $\tilde{P}_1(Nt)$ becomes the most relevant term in F_{DN} since the cluster 1 is the only one that can win against the fittest single. This explains why we observe the convergence of all the lines in different fixed asymptotic behaviors. Consequently, if there is no double fitter than the fittest single the lines go to zero. In detail we can have $n_D + 1$ possible asymptotic behaviors which corresponds the possible sizes of the cluster 1. The asymptotic behaviors follow this equation:

$$F_{DN}^c = P_1^c = 1 - e^{-c \cdot t \cdot N \mu_d}, \quad c \in \{0, 1, ..., n_D\}, \tag{2.19}$$

where c is the size of the fittest cluster. For $c = 0$ it is just 0. We can also compute the probability of each cluster as follow. As said before the total number of landscapes is

$\# = \binom{n_D+n_S}{n_S}$. The number of landscapes belonging to cluster c is:

$$\#c = \binom{n_D + n_S - c - 1}{n_S - 1}$$
$$P(c) = \binom{n_D + n_S - c - 1}{n_S - 1} / \binom{n_D + n_S}{n_S}.$$

In figure 2.5 we plotted all the possible landscape for $n_S = 4, n_D = 6$. We can clearly see all the lines converging to the respective cluster for large Nt.

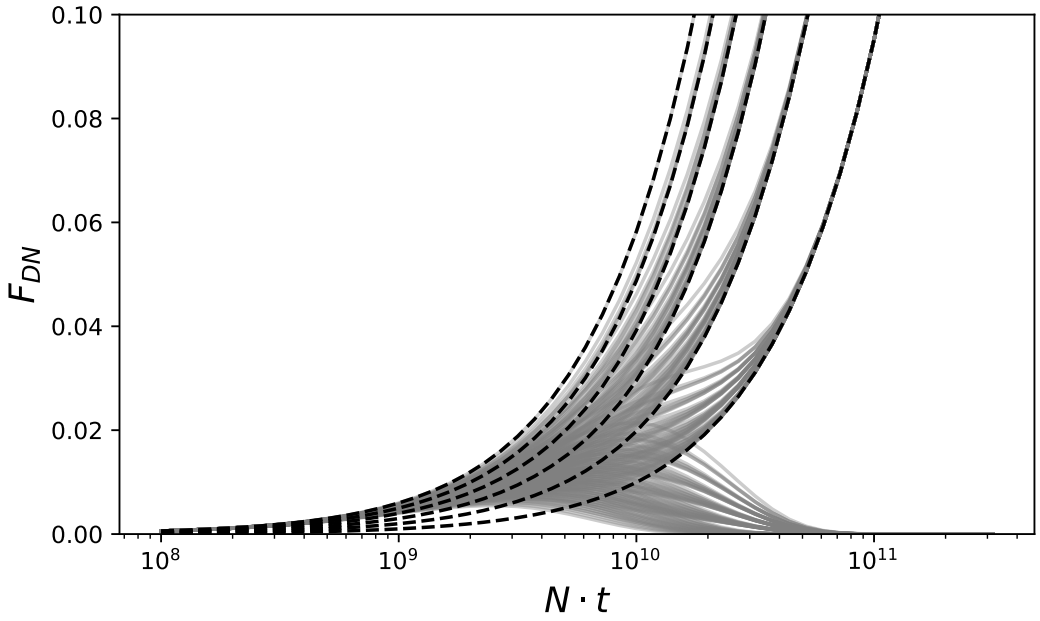

Figure 2.5: Probability of fixation of a double mutant (F_{DN}) for each possible landscape. We can see how each line converges to the probability the the fittest cluster wins (dashed black lines). Computed with $n_S = 4, n_D = 6, \mu_D/\mu_S = 10^{-2}$.

Point c)

Sometimes F_{DN} displays a non-monotonic behavior. As described in the previous points the behavior of F_{DN} can be explained quantitatively for large and small value of Nt. In the middle regime we observe many different behaviors; sometimes F_{DN} is monotonic and sometimes it is not. In this regime, the behavior of F_{DN} is very sensitive to the underlying landscape and thus it is hard to further categorize landscapes.

2.3.1 Dependence on n_S, n_D

The ratio between n_S and n_D depends on the class of mutations we are studying. In the case of double adjacent nucleotide mutations it is around 1.5 (see next chapter for details). However, the value of n_S is not fixed and relevant for the resulting dynamics. In figure 2.6 we show F_{DN} for three different values of $n_S \in \{20, 60, 180\}$. On the y axis there is the expected number of drawn single and double mutants. As we can see the three behaviors are slightly different. In particular the flat part is wider for larger values of the n_S. In point a) of the former subsection, we explained that the flat part begins when we start observing always a single mutant and it ends when the singles saturate. If n_S is smaller, the flat part gets shorter because we saturate immediately the singles. On the contrary, if n_S is larger also the time needed to saturate singles is larger and so the flat part is wider.

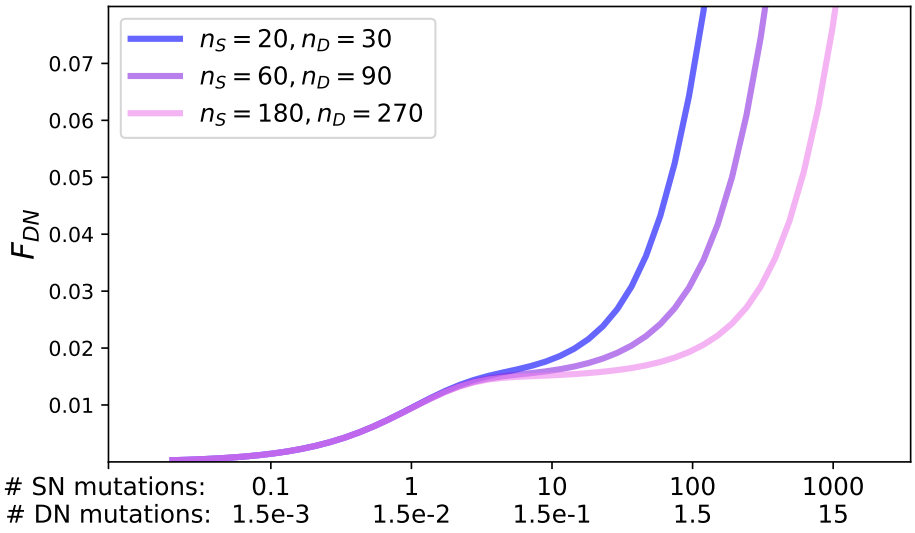

Figure 2.6: Probability of fixation of a double mutant (F_{DN}) for different values of n_S and n_D. Each F_{DN} is the mean over 1000 landscapes. The y-axis shows the expected number of single and double mutations.

2.3.2 Different DFE

Changing the DFE effects the probability of each landscape to be drawn. In our case we draw the single and double landscapes from two exponential distributions with mean λ_S and λ_D. In fig 2.7 we show the average $\langle F_{DN} \rangle$ for different value of λ_D which is the mean of the exponential distribution from which the doubles landscapes are drawn. As we did in the point a) of the first section of this chapter we are going to describe the

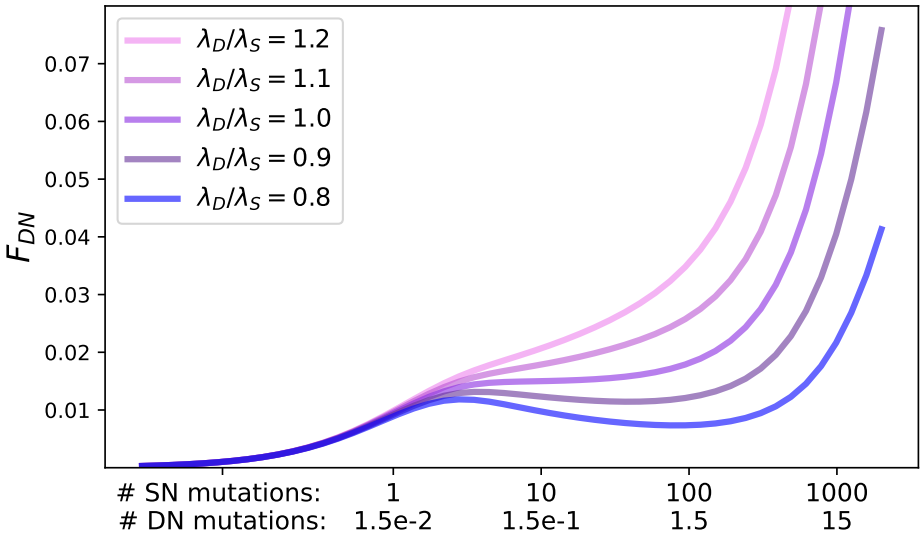

Figure 2.7: Probability of fixation of a double mutant (F_{DN}) for different distributions. Each F_{DN} is the mean over 1000 landscapes. We are using two exponential distributions as DFE. Each line is obtained by changing λ_D.

small, middle and large Nt regimes.

For small value of Nt the value of F_{DN} (and of its average) does not depend on the landscape. Therefore all the lines overlap. For large value of Nt the value of $\langle F_{DN} \rangle = P(D > S)$ increases with λ_D. This is what we expected since increasing λ_D means increasing the mean and the variance of the double DFE. We can compute the value of $\langle F_{DN} \rangle$ with equation 2.12.

For intermediate values of the mutation supply, we see distinct behaviors in what we previously referred to as the flat part. If $\lambda_D > \lambda_S$ there is a positive slope; on the contrary if $\lambda_D < \lambda_S$ there is a negative slope. This may be quantitatively explained in the same manner as in part a). In this case, we will employ uniform distributions as DFEs since they allow us to solve the problem analytically. If $f_D(\omega) = f_S(\omega) = U_{[0,1]}$ and if we observe u_S unique single mutants and u_D unique double mutants. The two largest order statistics distributions are:

$$f_{(u_S)}(x) = u_S \cdot x^{u_S - 1}, \quad f_{(u_D)}(x) = u_D \cdot x^{u_D - 1}. \tag{2.20}$$

As described before $\langle F_{DN} \rangle$ can be computed as:

$$\langle F_{DN} \rangle = \int_0^\infty \int_y^\infty f_{(u_S)}(y) \cdot f_{(u_D)}(x)\, dy\, dx = \frac{u_D}{u_D + u_S} \tag{2.21}$$

$$\simeq \frac{n_D \cdot Nt \cdot \mu_D}{n_D \cdot Nt \cdot \mu_D + n_S \cdot Nt \cdot \mu_S} = \frac{n_D \mu_D}{n_D \mu_D + n_S \mu_S} = f. \tag{2.22}$$

Now lets use the same uniform distribution for the singles $f_S(\omega) = U_{[0,1]}$ and a uniform distribution for double with a slightly different domain $f_D(\omega) = U_{[0,1+\epsilon]}$ with $|\epsilon| \ll 1$. The two largest order statistics distributions are:

$$f_{(u_S)}(x) = u_S \cdot x^{u_S - 1}, \quad f_{(u_D)}(x) = u_D \cdot \frac{1}{1+\epsilon} \cdot \left(\frac{x}{1+\epsilon}\right)^{u_D - 1} \tag{2.23}$$

We compute $\langle F_{DN} \rangle$ as before:

$$\langle F_{DN} \rangle = ... = \frac{u_D}{u_D + u_S} + \frac{u_D \cdot u_S}{u_D + u_S} \cdot \epsilon \tag{2.24}$$

$$\simeq \frac{n_D \mu_D}{n_D \mu_D + n_S \mu_S} + \frac{n_D \mu_D \cdot n_S \mu_S \cdot Nt}{n_D \mu_D + n_S \mu_S} \cdot \epsilon = f(1 + \epsilon \cdot n_S \mu_S \cdot Nt). \tag{2.25}$$

In this case $\langle F_{DN} \rangle$ depends on Nt because we have a quadratic term on Nt that does not cancel out. Because of that if $\epsilon > 0$, which means larger mean and larger variance, $\langle F_{DN} \rangle$ increases linearly with Nt. On the opposite if $\epsilon < 0$, which means smaller mean and smaller variance, $\langle F_{DN} \rangle$ decreases linearly with Nt. As we increase Nt, u_S starts growing sublinearly and so $\langle F_{DN} \rangle$ starts increasing independently on ϵ.

2.4 Genetic drift

One of the most important features that is missing in the simple model is the genetic drift (see next chapter for details). When we have a new mutant there is a non-zero probability of loosing it due to the stochasticity of the dynamic, this probability depends on the fitness of the mutant. We can easily add it to the simulation by introducing a probability $1 - \pi(\omega) = e^{-2(\omega - 1)}$ of throwing away each mutant before picking the fittest one. ω is the fitness.

The following method can be used if we want to include it to the analytical analysis. We define $g_i^S(t) = 1 - e^{-t \cdot N \mu_S \cdot \pi(\omega_i^S)}$ as the probability of observing at least once the single mutant i in a given time interval t. We do the same for doubles. We are reducing the rate of mutation by a factor $\pi(\omega)$ which is the probability surviving generic drift. Now we can define as the vectors:

$$\vec{p}_S : (\vec{p}_S)_i = g_i^S \cdot \prod_{j=1}^{i-1}(1 - g_j^S) \tag{2.26}$$

$$\vec{p}_D : (\vec{p}_S)_i = g_i^D \cdot \prod_{j=1}^{i-1}(1 - g_j^D) \tag{2.27}$$

$$F_{DN} = \vec{p}_S^T \cdot \mathbb{H} \cdot \vec{p}_D + \left(1 - \prod_{j=1}^{n_D} e^{-t \cdot N\mu_D \cdot \pi(\omega_j^D)}\right) \cdot \prod_{j=1}^{n_S} e^{-t \cdot N\mu_S \cdot \pi(\omega_j^S)} \tag{2.28}$$

where the last term is the probability of observing at least one DN mutant but zero DN mutants. Note that with this setting F_{DN} does not depend anymore only on the rank but also and the fitness values of the mutants.

Chapter 3

Wright Fisher model

3.1 Description

In order to simulate the evolution of a population on the fitness landscape we use the so called Wright fisher model [31]. This model is commonly utilized in population genetics since it captures several biological scenarios and is computationally more efficient than alternative methods.

The Wright Fisher (WF) model describes a population with discrete, non-overlapping generations and fixed population N. In each generation the entire population is replaced by the offspring from the previous generation. The offspring are generated considering mutation, selection and genetic drift. We assume the population to be asexual and haploid, which means that recombination does not occur. We define $p_i(t)$ as the population of the genotype i at step t. The wild-type is $i = 1$. Since the population is fixed $\sum_i p_i(t) = N$. Each individual has a probability per generation (rate) of mutating its genotype. We can define a mutation matrix $\mathbb{M}$ such that each element $m_{i,j}$ represents the mutation rate from genotype i to genotype j. The frequency (population) of genotype i after mutation is:

$$p_i^m(t+1) = \sum_j m_{j,i} \cdot p_j(t). \tag{3.1}$$

The fitness landscape associates a fitness to each genotype, ω_i is the fitness of the genotype i. Selection is the process whereby individuals with a fitter genotype tend to produce more offspring. The frequency of genotype i after mutation and selection is:

$$p_i^{m,s}(t+1) = \frac{\omega_i}{\bar{\omega}} \cdot p_i^m(t+1), \tag{3.2}$$

where $\bar{\omega} = \sum_i \omega_i \cdot p_i^m(t+1)$ is the mean fitness after mutation. Genetic drift is the randomness of the system. In each generation the offspring population is drawn from the parents population after mutation and selection making the dynamics nondeterministic. Finally, we can compute the effect of genetic drift and compute the population at step $t+1$:

$$[p_1(t+1), p_2(t+1), ...] \sim \text{Multinomial}(N, [p_1^{m,s}(t+1), p_2^{m,s}(t+1), ...]). \tag{3.3}$$

Genetic drift is very relevant in the early stage of a mutant because the population is small and fluctuations can drive it to extinction even if it is a beneficial mutant. Computationally, selection and mutation can be implemented with matrix multiplication followed by a normalization. Drift can be implemented by drawing from a multinomial random number generator. In figure 3.1 an oversimplified scheme of the Wright–Fisher model.

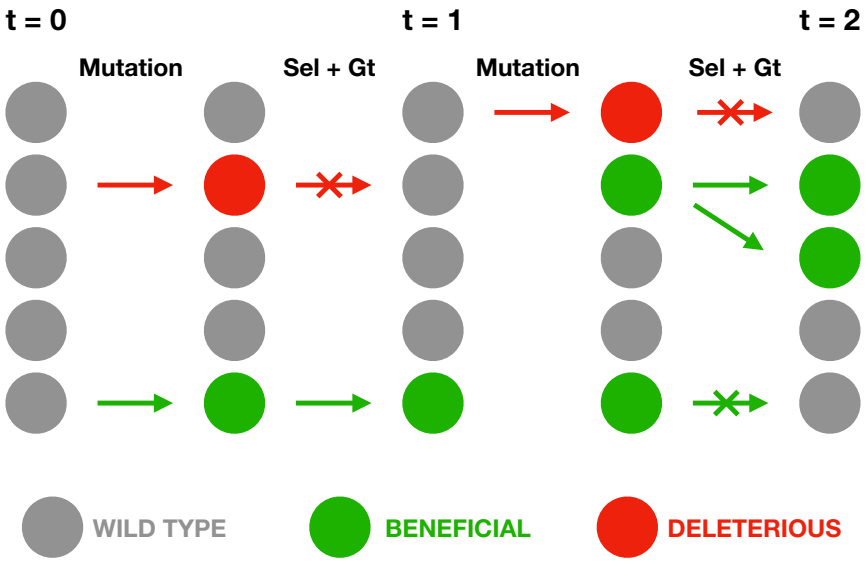

Figure 3.1: WFM schematic. We consider an oversimplified dynamic in which there are only 3 different genotypes: grey is the wild type, green is beneficial mutant, red is deleterious mutant. In each time step mutation acts first and then selection and genetic drift together. Mutation acts only on the wild type that can randomly change into beneficial or deleterious. Selection and genetic drift are performed by a biased sampling of the population after mutation. The weights of the sampling is the fitnesses of the genotypes. The deleterious mutations are unlikely to survive while beneficial mutant are likely to survive and spread. The dynamic is stochastic therefore a beneficial can go extinct like the lower right one.

The specific model we implement have two restrictions:

- The initial population is homogeneous, meaning that all the individuals have the wild type genotype: $p_1(t=0) = N$.
- We consider a one step dynamic. This has two main implications. We stop the simulation once a mutant reaches fixation. We allow only mutations from the wild type to mutant genotypes. This restriction make the dynamic easier to analyze and in the last chapter will allow us to use an empirical landscape.

In the panel 1 we show the pseudo-code of the model.

Algorithm 1 Pseudo code of the implementation of the Wright Fisher model.

Require: N ▷ population size
Require: $\vec{\omega}$ ▷ fitness landscape
Require: $\mathbb{M}$ ▷ mutation matrix

$\vec{n} \leftarrow [N, 0, 0, ..., 0]$ ▷ initial population
while 1 **do**
  $\vec{n} \leftarrow \mathbb{M} \cdot \vec{n}$ ▷ mutation
  $\vec{s} \leftarrow [n_0\omega_0, n_1\omega_1, ..., n_i\omega_i]$ ▷ selection vector
  $\vec{p} \leftarrow \mathcal{N}(\vec{s})$ ▷ normalization
  $\vec{n} \leftarrow \text{multinomial}(N, \vec{p})$ ▷ genetic drift
  if $\max(\vec{n}) \geq N - 1$ **and** $\text{argmax}(\vec{n}) \neq 0$ **then**
    break
  end if
end while
return $\text{argmax}(\vec{n})$

Before going to the results section we define the mutation matrix $\mathbb{M}$. Each term m_{ij} of the mutation matrix $\mathbb{M}$ represents the mutation rate from the genotype i to the genotype j. If we define μ_S as the mutation rate from the wild type to a single mutant and μ_D as the mutation rate from the wild type to a double mutant we can write the matrix as:

$$\mathbb{M}_{i \leftarrow j} = \begin{cases} 0\,; & \text{if} \quad j \neq 0 \text{ and } i \neq j \\ 1\,; & \text{if} \quad j \neq 0 \text{ and } i = j \\ \mu_S\,; & \text{if} \quad j = 0 \text{ and } j \to i \text{ SN} \\ \mu_D\,; & \text{if} \quad j = 0 \text{ and } j \to i \text{ DN} \\ 1 - (n_S\mu_S + n_D\mu_D)\,; & \text{if} \quad j = i = 0 \end{cases} \tag{3.4}$$

note that the mutant genotypes can not mutate. If we define the total mutation rate $U = n_S \cdot \mu_S + n_D \cdot \mu_D$ and the fraction of double de novo mutations $f = n_D \cdot \mu_D / U$ we can rewrite the mutation matrix as follow:

$$\mathbb{M}_{i \leftarrow j} = \begin{cases} 0\,; & \text{if} \quad j \neq 0 \text{ and } i \neq j \\ 1\,; & \text{if} \quad j \neq 0 \text{ and } i = j \\ \frac{U}{n_S}(1-f)\,; & \text{if} \quad j = 0 \text{ and } j \to i \text{ SN} \\ \frac{U}{n_D} f\,; & \text{if} \quad j = 0 \text{ and } j \to i \text{ DN} \\ 1 - U\,; & \text{if} \quad j = i = 0 \end{cases} \tag{3.5}$$

this second version is useful for two reasons: in the analytic part we can write the equations in a more compact way; in the simulation we can change the ratio μ_1/μ_2 by changing f without changing the total mutation rate U and vice versa.
In the figure 3.2 we show an example of a single Wright Fisher model simulation. On the left we can appreciate the effect of genetic drift. The first two mutants are beneficial, however they both go to extinction. On the left we can observe the competition between two genotypes. The darkest is the fittest, because of that it has a selective advantage on the other genotype.

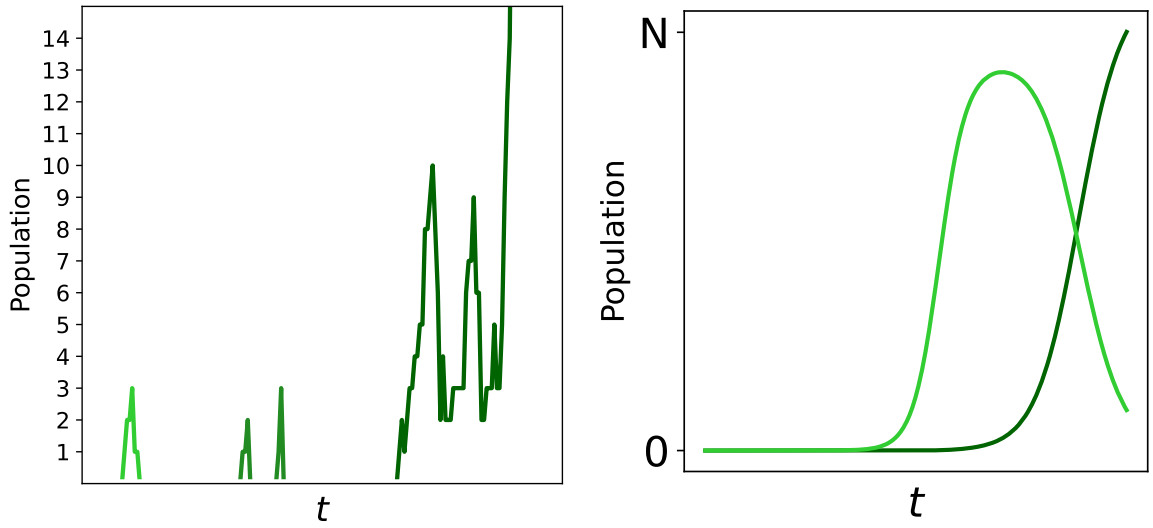

Figure 3.2: Wright Fisher model simulations. On the left we can see that the first two beneficial mutant do not survive genetic drift but the third does. On the left the competition between two genotypes.

3.1.1 Results

Now that we have defined the population genetic model we can simulate adaptive evolution. Before showing the results we briefly discuss about the values of the variables $(\mu_s, n_s, ...)$ that we are going to use in the model. We want them to be biologically realistic. We consider the genotype at amino-acid level. We decided to consider only beneficial mutant because are the only relevant mutants to the dynamic. Below the list of all variables of the systems and their values:

- n_s and n_d: the first is the number of beneficial genotypes that can be reached with a single nucleotide mutation, the second is the number of beneficial genotype that can be reached with a adjacent double nucleotide mutation. They were computed in the following way. If we consider a gene of length L (number of nucleotides), there are $L/3$ codons. For each codon, the average number of accessible amino-acid with a SN mutation is 6 and is 16 with an adjacent double nucleotide mutation. Some of the amino-acids that are accessible through a single mutation are also accessible through a double mutation and so most of the time they will be reached with a single mutation. For this reason we will consider $16-6 = 10$ as the number of accessible amino-acid with a double mutation. We assume that only the 1% of non synonymous mutations are beneficial so $n_s = \frac{L}{3} \cdot 6 \cdot 0.01 = 2L/100$ and $n_d = L/30$. In our simulation we use L between 600 and 6000 which correspond to n_S between 12 and 120 and n_D between 20 and 200.

- μ_s: the single point mutation rate is around $\mu_S^p = 10^{-10}$. We need the codon mutation rate therefore we multiply the single point mutation rate by 3. Since we are interested in the beneficial mutation rate we multiply the value by the average probability of having a non-synonymous mutation with a SN mutation which is 0.75. So the single beneficial mutation rate becomes: $\mu_s = 3 \cdot \mu_S^p \cdot P_S(\text{non syn}) = 3 \cdot 10^{-10} \cdot 0.75 \simeq 2.25 \cdot 10^{-10}$.

- μ_d: the ratio between the single and the double point mutation rate $\alpha = \mu_D^p/\mu_S^p$ is between 10^{-2} and 10^{-3}. For the same reasons described in the previous point: $\mu_d = 3 \cdot \mu_D^p \cdot P_D(\text{non syn}) = 3 \cdot \alpha \cdot 10^{-10} \cdot 0.99$; where 0.99 is the probability of having a non-synonymous substitution with double nucleotide mutation.

- Beneficial mutation supply and N: we chose the beneficial mutation supply to be between 10^{-3} and 10. The beneficial mutation supply is defined as $U \cdot N$. U is the total mutation rate: $U = n_S \cdot \mu_S + n_D \cdot \mu_D$. In the simulation we increment the beneficial mutation supply incrementing only the population size N.

- Random fitness landscape: as we did for the simple model we draw the SN fitness landscape $[\omega_1^S, ..., \omega_{n_S}^S] \sim f_S(\omega)$ and a DN fitness landscape $[\omega_1^D, ..., \omega_{n_D}^D] \sim f_D(\omega)$ from two distributions $f_S(\omega)$, $f_D(\omega)$.

Now that we have defined all the variables we can move to the simulation. For a given fitness landscape and a given value of the mutation supply we run the WFM simulation several times and we compute the probability of a double to fixate first F_{DN}. In figure 3.3 we show F_{DN} for different landscapes (grey) and their average (black) $\langle F_{DN} \rangle$. We computed $\langle F_{DN} \rangle$ for two values of α. In this case we are drawing the fitness values for singles and doubles from the same exponential distribution with mean $\lambda = 0.001$.

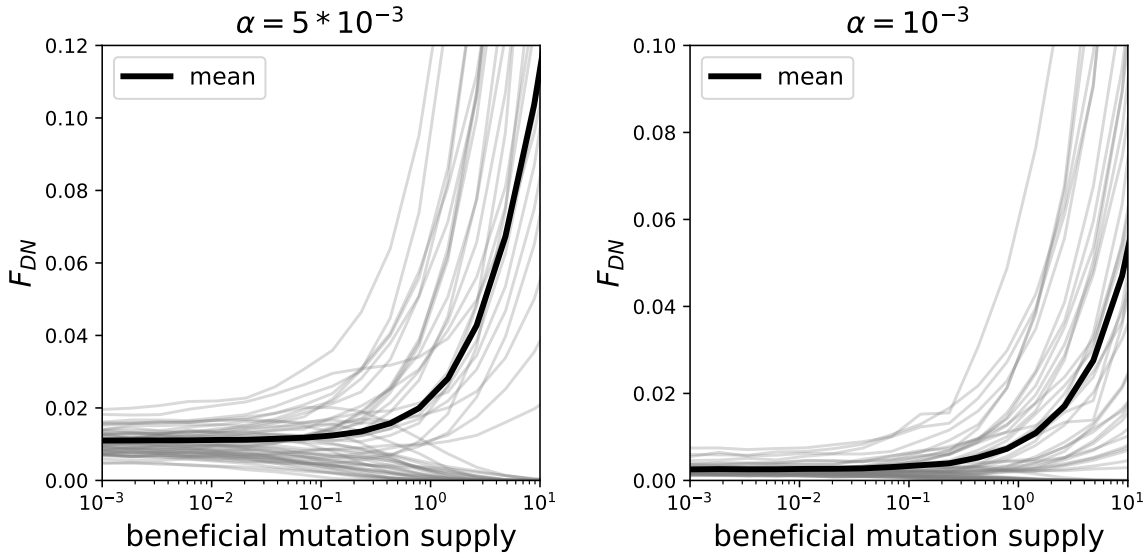

Figure 3.3: Probability of fixation of a double mutant (F_{DN}) for two different values of $\alpha = \mu_D/\mu_S$. In grey different random landscapes. In black the mean over 100 landscapes. The mean is monotonically increasing. For this two figures we used $n_S = 12, n_D = 20$.

The two curves have the same shape and the value of F_{DN} ranges between 0.2% and 5.5% for $\alpha = 10^{-3}$ and between 1% and 12% for $\alpha = 5 \cdot 10^{-3}$. As mentioned during the introduction the value of α depends on the species we are considering and it is a very important value for determining the relevance of double double nucleotide mutations. In general we can say that if F_{DN} is of the order of $1 - 10\%$ double can be non-negligible and can effect the trajectory of adaptive evolution. This result is sensible to the parameters we chose. In the next table we summarize the value of $\langle F_{DN} \rangle$ for the two values of α:

Parameters			Mutation supply			
α	n_S	n_D	10^{-2}	10^{-1}	10^{-0}	10^{1}
$1 \cdot 10^{-3}$	12	20	0.2%	0.3%	0.9%	5.5%
$5 \cdot 10^{-3}$	12	20	1%	1%	3%	12%

3.2 Analytical approximations

3.2.1 Weak mutation regime

We first want to analytically compute the probabilities of fixation in the weak mutation regime $\Rightarrow N \cdot U \ll 1$. We use a result from this papers [32, 33]. Consider a homogeneous population with fitness $\omega_{wt} = 1$, if a mutation occurs, the mutant can be beneficial $\omega_i > 1$ or deleterious $\omega_i < 1$. If it is deleterious it goes extinct after a few steps due to natural selection, if it is beneficial it can earthier go extinct or survive genetic drift. As explained before genetic drift is very important in the first few generations of a mutant because its population is small and the stochasticity can drive it to extinction (see left panel fig 3.2 left). Surviving genetic drift means to reach a population size such that the fluctuations of the dynamics become negligible. We define the selection coefficient $s_i = (\omega_i - \omega_{wt})/\omega_{wt} = \omega_i - 1$. The probability of surviving genetic drift is:

$$\pi_i = \begin{cases} 1 - e^{-2s_i} & s > 0 \\ 0 & s \leq 0 \end{cases} \tag{3.6}$$

the probability of surviving genetic drift is the probability of not going extinct due to genetic drift. In the weak mutation regime it is unlikely to have two or more mutants competing in the same time (fig 3.2) and so if a mutant survives genetic drift it reaches fixation: $P_i(\text{fix}|\text{de novo}) = \pi_i$. Finally we can compute the probability that i fixates first multiplying the last term by the mutation rate of the mutant and normalizing:

$$P_i = \frac{\mu_i \cdot \pi_i}{\mu_S \sum_{SN} \pi_i + \mu_D \sum_{DN} \pi_i} \tag{3.7}$$

$$F_{DN} = \frac{\mu_D \sum_{DN} \pi_i}{\mu_S \sum_{SN} \pi_i + \mu_D \sum_{DN} \pi_i} = \frac{f \cdot \langle \pi_i \rangle_{DN}}{f \cdot \langle \pi_i \rangle_{DN} + (1-f) \cdot \langle \pi_i \rangle_{SN}}. \tag{3.8}$$

Notice that if we draw the single and double landscapes from the same distribution we expect $\langle \pi_i \rangle_{SN} = \langle \pi_i \rangle_{DN}$ leading to $F_{DN} = f$. Remember that π_i is 0 for deleterious

mutations. If we consider $s_i \ll 1$ then $\pi_i = 2s_i$. Thus we can rewrite equation 3.8 as:

$$F_{DN} = \frac{f \cdot \rho_D \cdot \langle s_i \rangle_{DN}}{f \cdot \rho_D \cdot \langle s_i \rangle_{DN} + (1-f) \cdot \rho_S \cdot \langle s_i \rangle_{SN}}, \tag{3.9}$$

where ρ_S, ρ_D are the fraction of beneficial singles and double. In this way the value of F_{DN} in the weak mutation regime depends on the means of the beneficial distributions of fitness effect.

3.2.2 Clonal interference regime

We aim to incorporate the clonal interference regime into our analytical model [17]. In this regime, there is a non-negligible probability that two mutants will be in competition with each other (fig 3.2). We take into account the amount of interfering mutants in our model to add competition. Consider a mutant with selection coefficient s_i the time (number of generations) until fixation is on average:

$$\tau_i = \frac{N \cdot \log N}{\log(1 + s_i)}. \tag{3.10}$$

During this time interfering mutants can occur. The expected number of mutants with genotype j with $s_j > s_i$ that arise during τ_i and survive genetic drift is:

$$\lambda_{i,j} = \pi_j \cdot \mu_j \cdot \tau_i. \tag{3.11}$$

The overall number of interfering mutations is thus:

$$\Lambda_i = \sum_{\{j | s_j > s_i\}} \lambda_{i,j} \tag{3.12}$$

The probability that the mutant i reaches fixation is the probability that it survives genetic drift times the probability that no better mutants arise. If we assume that number of interfering mutants is Poisson distributed with mean Λ_i we can write:

$$P_i(\text{fix}|\text{de novo}) = P_i(\text{surv. gen. drift}) \cdot P_i(\text{no better mutants arise}) \tag{3.13}$$

$$= \pi_i \cdot e^{-\Lambda_i} \tag{3.14}$$

As we did with for the WM regime, we can now define P_i as the probability that the mutant i fixes first.

$$P_i = \frac{\mu_i \cdot \pi_i \cdot e^{-\Lambda_i}}{\sum_k \mu_k \cdot \pi_k \cdot e^{-\Lambda_k}}; \tag{3.15}$$

and:

$$F_{DN} = \sum_{DN} P_i = \frac{f \cdot \langle \pi_i \cdot e^{-\Lambda_i} \rangle_{DN}}{f \cdot \langle \pi_i \cdot e^{-\Lambda_i} \rangle_{DN} + (1-f) \cdot \langle \pi_i \cdot e^{-\Lambda_i} \rangle_{SN}}. \tag{3.16}$$

If $\mu_j \cdot N \ll 1 \Rightarrow \lambda_i \propto \mu_i N \to 0$, the equations reduce to the one previously found for the WM regime. As we increase the mutation supply we increase the number of interfering mutations $\Lambda_i(s_i)$ and the term $e^{-\Lambda_i(s_i)}$ decreases. If we treat it in this fashion, equation 3.8 is the same as equation 3.16 with discount factor that depends on the selection coefficient of the mutant. In the figure 3.4 we can see how the discount factor $e^{-\Lambda_i(s_i)}$ behaves for different regimes of the mutation supply.

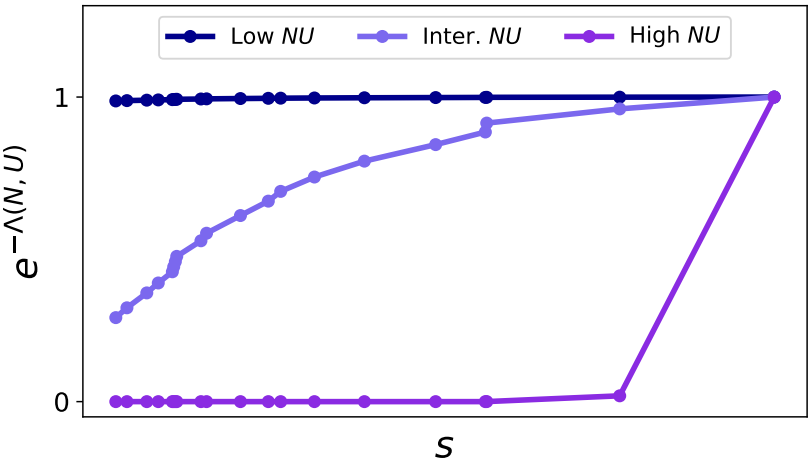

Figure 3.4: $e^{-\Lambda_i(s_i)}$ for different regimes of the mutation supply. For small values of the mutation supply the number of interfering mutation is almost zero. Therefore $e^{-\Lambda_i(s_i)} = 1$ independently on the selection coefficient. For intermediate values competition is on and so the number of interfering mutation larger for smaller values of the selection coefficient. For large values of the mutation supply all the mutants are always observed and so the discount factor is 0 for all the mutants and 1 for the fittest mutant.

This model is an extension of the weak mutation regime one and does not work for high values of the mutation supply. The biggest approximation is that the selection coefficient s_i is computed with respect to the wild type $\omega_{wt} = 1$. If a mutation occurs on top of another the value of s_i is not accurate anymore. This has consequences on the the probability of surviving genetic drift $\pi(s_i)$ and the average time until fixation τ_i which will be both inaccurate. As a result, the number of steps needed for a genotype to reach fixation will increase, leading to an larger number of interfering mutants that will favor genotypes with higher fitness. The model fails for large value of the mutation supply because too many mutations arise on average at each step. The accuracy of the model is also related to N since increasing N will make τ_i increase. Because of that more mutations then expected will occur before fixation making the model less

accurate.

3.2.3 Results

We want to test how accurate our models are. First of all, with a given fitness landscape, we run the Wright Fisher model for several times and we compute P_i for each genotype. We compute P_i also using the weak mutation approximation and the clonal interference approximation. We do it for three values of the mutation supply to highlight the difference of the two models and their accuracy. In figure 3.5 the comparison between the model predictions and the simulation predictions in three different regimes.
As we expected in the weak mutation regime (first panel) both models are accurate. In the clonal interference regime (central panel) we add competition to the dynamic and the WM model fails to predict the values overestimating the probabilities for small value of the selection coefficient and vice versa. For high values of the mutation supply both models fail. Again we want to stress the fact that the accuracy does not depends only on the mutation supply but also on N and s_i.
In figure 3.6 we plot the probability that a double fixates first F_{DN} as a function of the beneficial mutation supply $N \cdot U$ computed with simulation and with the two models.

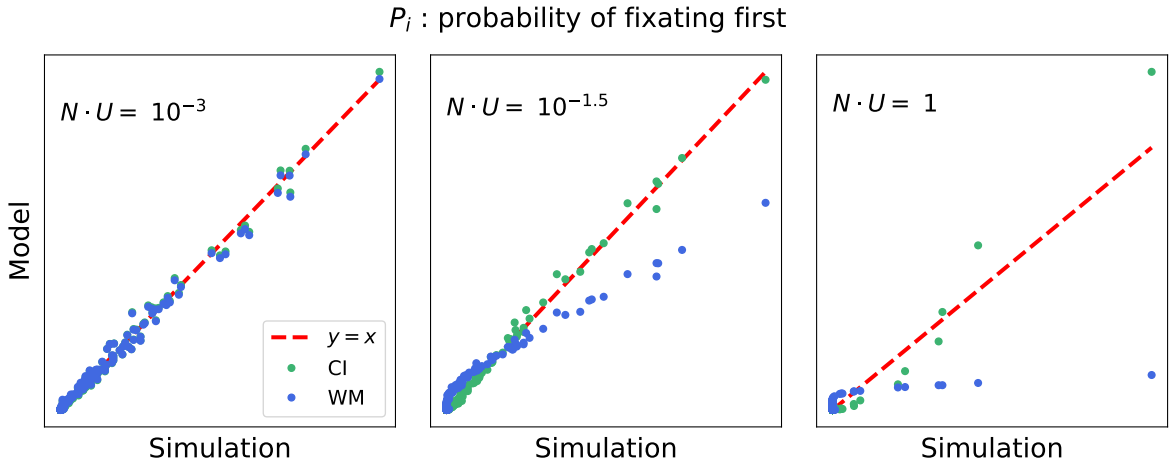

Figure 3.5: Comparison between models and simulation. Each dot is a genotype. On the x axis the probability of fixating first P_i computed with the Wright Fisher Model. On the y axis computed with the WM model (blue) and CI model (green). In each panel the values were computed using different values of the mutation supply. For the simulation we used $n_S = 120, n_D = 180, \lambda_S = \lambda_D = 0.1$.

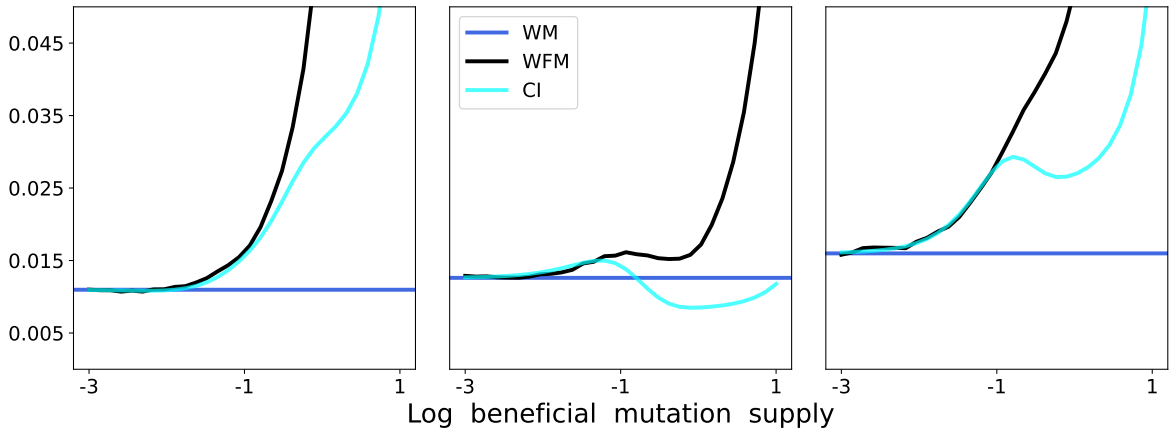

Figure 3.6: Probability that a double fixates first F_{DN} as a function of the mutation supply. In black computed with the WF model, in blue with the WM model and in cyan with the CI mode. For the simulation we used $n_S = 20, n_D = 30, \lambda_S = \lambda_D = 0.001$.

As before both models are accurate in the weak mutation regime and the CI model, in this case, is accurate until $N \cdot U = 0.1$.

3.3 Mean fitness jump

As last analysis we want to compute the mean fitness jump increment. The goal is to quantify the fitness advantage of allowing adjacent double nucleotide mutation. We

define mean fitness jump as:

$$\Delta W = \sum_i P_i \cdot \omega_i. \tag{3.17}$$

In order to quantify the advantage of having doubles we compute the mean fitness jump considering only single nucleotide mutations ΔW_S and considering both single and double mutations ΔW_{SD}. At this point we can compute the relative mean fitness increment as:

$$\Delta = \frac{\Delta W_{SD} - \Delta W_S}{\Delta W_S}. \tag{3.18}$$

In figure 3.7 on the top panel is reported the average $\langle \Delta \rangle$ (over 100 landscapes) as a function of the mutation supply for two values of α. In the weak mutation regime $N \cdot U \ll 1$ and $\langle F_{DN} \rangle \simeq 0.2\% - 1\%$ therefore the contribution of doubles is small. Furthermore, all the beneficial doubles have a non zero probability of fixation. Because of this two reasons and of the fact that $f_S(\omega) = f_D(\omega)$, the average of the relative mean fitness jump is almost zero $\langle \Delta \rangle \simeq 0\%$. The advantage of allowing adjacent double nucleotide mutations with $NU = 10$ is $\simeq 8\%$ for $\alpha = 5 \cdot 10^{-3}$ and $\simeq 4.5\%$ for $\alpha = 10^{-3}$. This values are comparable with the the values of $\langle F_{DN} \rangle$ and depending the the mutation supply and on α the doubles can can also influence the overall fitness of the population.

The next step is to investigate how the relative mean fitness jump changes for different values of n_S, n_D. In order to compare dynamics with different values of n_S, n_D we decided to compute $\langle F_{DN} \rangle$ and $\langle \Delta \rangle$ as a function of the population N instead of the mutation supply $N \cdot U$. We have to do it in this way since increasing n_S, n_D increase the total mutation rate U. In figure 3.7 on the bottom we plotted $\langle F_{DN} \rangle$ and $\langle \Delta \rangle$ as a function of the population N, for different values of the beneficial mutants. We can summarize the results for large population $N = 10^{12}$ in the following table:

n_S, n_D	$\langle F_{DN} \rangle$	$\langle \Delta \rangle$
$12, 20$	12%	14%
$120, 200$	8%	4%

We can see that increasing the number of beneficial slightly decreases $\langle F_{DN} \rangle$ and substantially decreases $\langle \Delta \rangle$. This is due to the finiteness of two random landscapes. Since $n_S < n_D$ and $f_S(\omega) = f_D(\omega) =$ Exponential, the fittest double is on average fitter than the fittest single. For large value of N the competition is between the fittest single $\tilde{\omega}_S$

and the fittest double $\tilde{\omega}_D$. We can rewrite F_{DN} and Δ as:

$$F_{DN} \simeq P(\tilde{\omega}_D) \tag{3.19}$$

$$\Delta \simeq F_{DN} \cdot \frac{\tilde{\omega}_D - \tilde{\omega}_S}{\tilde{\omega}_S} \tag{3.20}$$

For small number of beneficial mutants it is more likely that the singles's fitness values are not drawn from the tail of the distribution but some doubles's fitness values are. This leads (on average) to a large relative difference $\frac{\tilde{\omega}_D - \tilde{\omega}_S}{\tilde{\omega}_S}$ between the fittest double and the fittest single. Instead for larger values of beneficial also some singles's fitness values are drawn from the tail. This leads (on average) to a smaller relative difference between the fittest double and the fittest single. Because of that the value of $\langle F_{DN} \rangle$ is smaller for larger values of n_S, n_D. This effect is even stronger for $\langle \Delta \rangle$ because both terms $\langle F_{DN} \rangle$ and $\frac{\tilde{\omega}_D - \tilde{\omega}_S}{\tilde{\omega}_S}$ are smaller.
This explain what we observe in the figure. We can conclude by stating that the number of beneficial mutants is a relevant parameter of the dynamic. Note that what we have described is true for exponential distribution but not for heavy tail distributions.

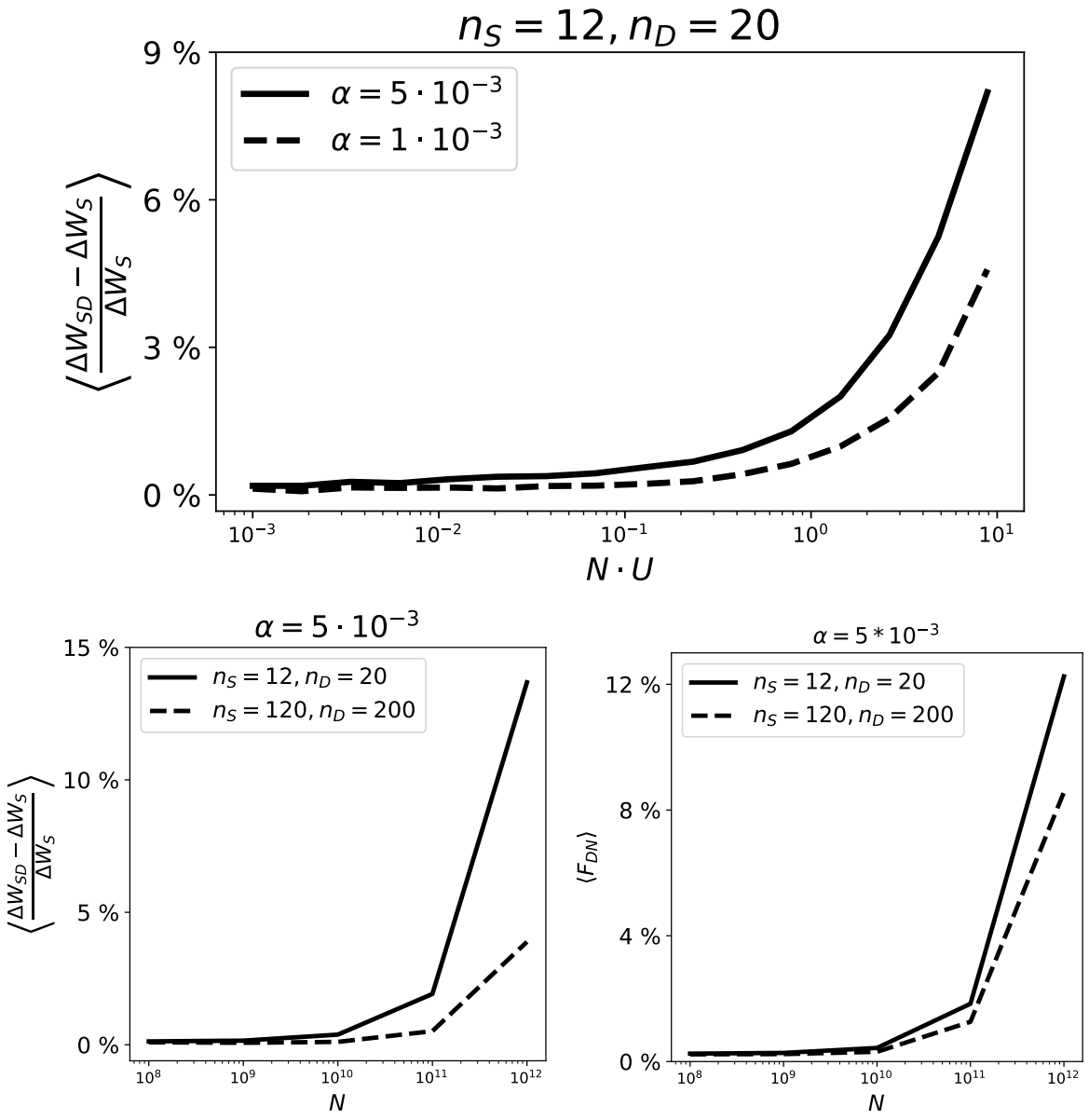

Figure 3.7: Mean fitness jump increment. The mean fitness jump increment is defined as $\Delta = \frac{\Delta W_{SD} - \Delta W_S}{\Delta W_S}$. We compute the average $\langle \Delta \rangle$ over 100 landscapes. On the top panel we compute $\langle \Delta \rangle$ as a function of the mutation supply for different values of α. Increasing the value of α increases the value of $\langle \Delta \rangle$ .On the two bottom panel we show $\langle \Delta \rangle$ (left) and $\langle F_{DN} \rangle$ (right) as a function of N for different values of n_S, n_D. As we can see increasing the values of n_S, n_D decreases the value of $\langle \Delta \rangle$ and $\langle F_{DN} \rangle$.

Chapter 4

Empirical landscape

As last analysis we want to simulate adaptive evolution on a empirical fitness landscape. In order to build this landscape we use the results of two papers [34, 35]. The subject of both of this studies is the TEM-1 β-lactamase gene that is a useful model for investigating evolution and the fitness effects of mutations. TEM-1 is a gene that confers high resistance to penicillin antibiotics such as ampicillin (Amp). Thus, if we put E. coli cells carrying TEM-1 in an environment with Amp, alleles conferring a higher ability to degrade the antibiotic will be advantaged . Consequently, Amp resistance is a key determinant of the fitness of an organism in the presence of Amp, although assessing TEM-1 fitness by measuring Amp resistance does not account for fitness variations unrelated to antibiotic resistance. The quantities used to estimate the fitness of a given allele is the minimum inhibitory concentration MIC. MIC is defined as the lowest concentration of a chemical, usually a drug, which prevents visible growth of bacteria. In this case the drug is the Amp.
These are the main results of the two studies:

- In the first paper they measured the fitness of point mutations and codon substitution. They were able to measure 98.2% (2536/2583) of all single nucleotide mutations and 83.9% (15167/18081) of all codon substitutions in the TEM-1 gene.
- In the second paper they measured the fitness of adjacent double amino-acid substitution. They were able to measure the 12.0% (12374/102855) of all possible adjacent double amino-acid substitutions.

Both paper are from the same lab and they took care of calibrating the results (fitness values) of the second paper on the results of the fist one. For this reason we can combine them together to obtain an empirical fitness landscape which involves single nucleotide mutations and adjacent double nucleotide mutations.

4.1 Dataset analysis

This is how the first dataset looks like:

Pos	WT codon	WT AA	Mut Codon	Mut AA	Fitness
0	ATG	M	GCA	A	0.002
0	ATG	M	TGC	C	0.009
...	...	...	...	...	...
287	TGG	W	TAT	Y	0.388

where:

- **Pos** is the position of the codon involved in the mutation.
- **WT codon** is the wild type codon.
- **WT AA** is the wild type amino-acid.
- **Mut Codon** is the mutant codon.
- **Mut AA** is the mutant amino-acid.
- **Fitness** is the fitness value of the mutant, $\omega = 1$ is the wild type fitness.

From this dataset we can also obtain the genotype of the wild type which is composed of 288 codons. If we look into the dataset we notice that mutations with different codon substitution but with the same amino-acid substitution have almost the same fitness value. Because of that we can reduce the dataset by looking at the fitness values at amino-acid level. The new fitness values are computed as the mean of all the mutations with same amino-acid mutant in the same position. Doing this we increase the number of fitness values from 85% (codon substitutions) to 98% (amino-acid substitutions). This is how the new dataset looks like:

Pos	WT AA	Mut AA	Fitness
1	M	A	0.002
1	M	C	0.009
...	...	...	...
288	W	Y	0.388

This the second dataset:

Pos	WT AA 1	WT AA 2	Mut AA 1	Mut AA 2	Fitness
1	M	S	R	C	0.003
...	...	...	...	...	...
287	H	W	L	A	0.008

It is very similar to the first one but with two amino-acid substitutions for each row. Pos is the position of the first aminoacid.

We now describe how we create the empirical landscape. Starting from the single nucleotide mutations. We compute all the possible SN substitutions $288 \cdot 3 \cdot 3 = 2592$. For each substitution we compute the mutant amino acid (Mut AA in the table). If the mutation is synonymous (WT AA = Mut AA) we assign fitness 1 to the mutant. If the mutation is non-synonymous we check in the fist dataset if there is the fitness value of the mutant. If the value is not the dataset we assign fitness 0 to the mutant. If the new codon is a stopping codon we assign fitness 0 to the mutant.

For DN mutations there are a total of $(288 \cdot 3 - 1) \cdot 9 = 7767$ possible adjacent double nucleotide substitutions. As we did for the single, we compute all the possible substitutions. For each substitution we check whether it is synonymous. If yes we assign fitness 1 to the mutant. If not we check if it is a single or double amino-acid substitution. In the first case we check in the fist dataset if there is the fitness value of the mutant. In the second case we check in the second dataset if there is the fitness value of the mutant. In figure 4.1 a schematic of this process.

4.2 Results

We summarize the results in the following table:

	Total	Data	$\langle\omega\rangle$	Beneficial	$\langle\omega_B\rangle$
SN	2592	2562 (99%)	0.67	473 (18%)	1.15
DN	7767	6960 (90%)	0.47	1421 (18%)	1.17

There are a total of $288 \cdot 3 \cdot 3 = 2592$ possible single substitutions. We have the fitness of 2562 of them which correspond to the 99%. The 24% of the single mutants are synonymous. This is due to the fact that the mutation on the third nucleotide of a codon are often synonymous. The mean fitness of the single mutants $\langle\omega\rangle = 0.67$. The mean fitness of the single mutants without considering the missing values is 0.70. If we consider only beneficial mutants it becomes $\langle\omega_B\rangle = 1.15$. Beneficial mutant are the 18% of the total mutations.

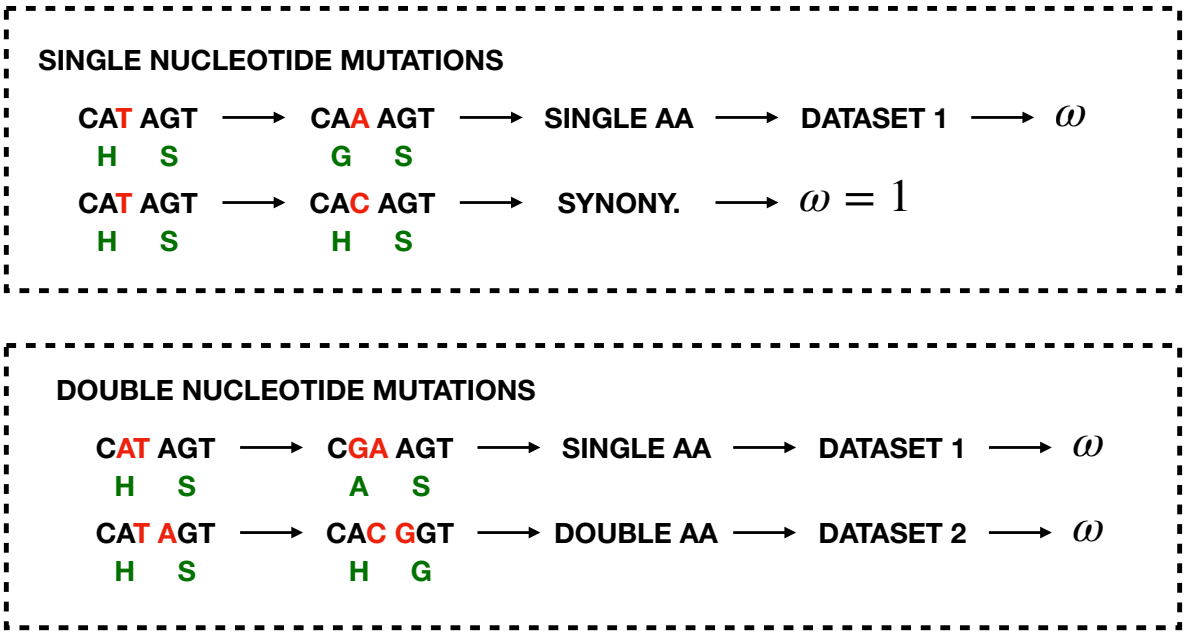

Figure 4.1: Schematic of the of empirical landscape building process. We calculate all the possible single and double nucleotides substitution. If a substitution leads to a synonymous substitution we assign fitness equal to 1 to the mutant. If a substitution leads to a single aminoacid substitution we check in the first dataset the fitness value of the mutant. If a substitution leads to a double aminoacid substitution we check in the second dataset the fitness value of the mutant. If in the first and in the second dataset there is not the fitness of the mutant we assign fitness equal to 0. At the end of this process we have the list of all the possible single mutants and double mutants and their fitness values.

The number of possible adjacent double nucleotide mutation is $(288 \cdot 3 - 1) \cdot 9 = 7767$. We have the fitness values of 6968 of them which correspond to the 90%. Most of the missing values are double mutations across two codons that lead to double amino-acid mutations. In details there are 691 double amino-acid mutations and we have data of only 107 of them. The mean fitness of the double mutants 0.47. The mean fitness of the double mutants without considering the missing values is 0.55. If we consider only beneficial mutants it becomes 1.17. Beneficial mutant are the 18% of the total mutations. It is interesting to observe that even if the mean fitness of the doubles is smaller than the mean fitness of the singles, if we restrict this analysis to the beneficial the result is the opposite. This is relevant for the Wright fisher dynamic where only the beneficial mutants matter.
We visualize the single landscape and the double landscape in figure 4.2.

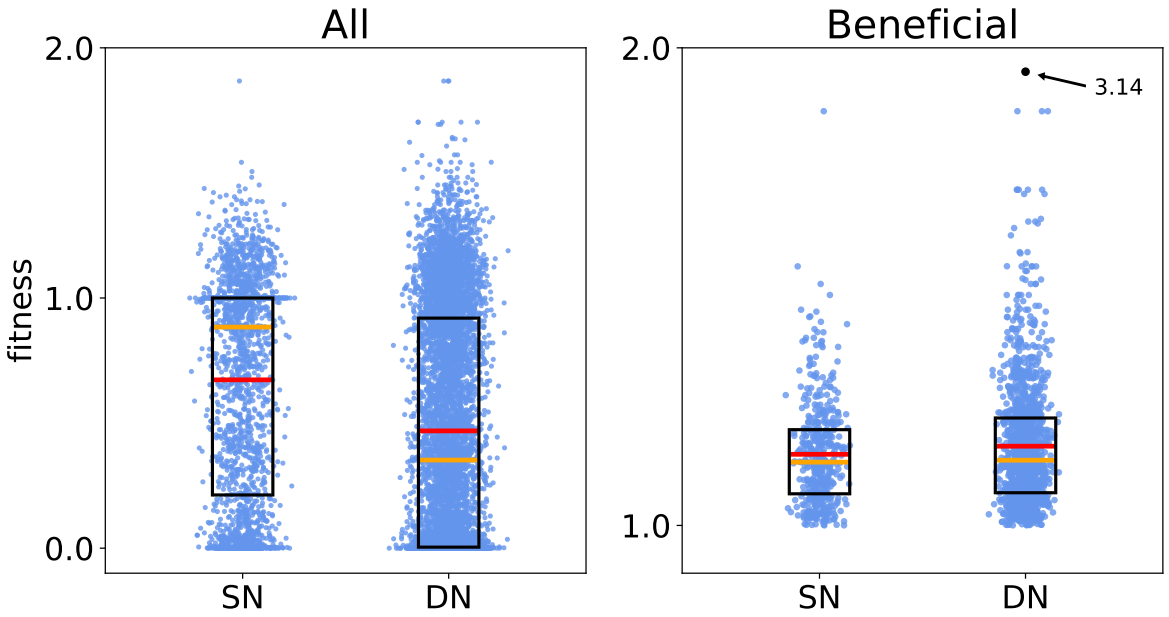

Figure 4.2: Empirical fitness landscapes. On the left we visualize the whole fitness landscapes. On the right we visualize only the beneficial fitness landscape. In red the mean of the distribution and in orange the median. On both figures there is a double mutant that is outside the range; we plotted it only in the beneficial distribution and it is the black dot in the top right corner.

4.2.1 Simulation on empirical landscape

We can now run the Wright Fisher model simulation on the empirical landscape. The parameters used in the simulation are the following:

$$\frac{\mu_D}{\mu_S} = \alpha = \{10^{-3}, 5 \cdot 10^{-3}\},\ \mu_S = 10^{-10},\ \mu_D = \alpha \cdot 10^{-10},\ n_S = 2592,\ n_D = 7767.$$

We compute F_{DN} as a function of the mutation supply for both values of α. In the figure 4.3 there are both plots side by side.
We can see that the two curves (black) have the same shapes, the only difference is their absolute value which is different by a factor of 5. As we look closely to the comparison of the two landscapes 4.2 we notice that the fittest mutant is a single with fitness value 3.14 and the second fittest is a single with fitness value 1.18. There are also three double mutants with fitness 1.18 belonging to the same amino-acid substitution of the best single. For large value of the mutation supply the fittest double is the only one contributing to F_{DN} as described in the simple model part.
For $\alpha = 10^{-3}$ the probability of fixation of a double mutant F_{DN} is 0.32% in the week mutation regime. Note that the fraction of double de novo mutations $f = \frac{\mu_D \cdot n_D}{\mu_D \cdot n_D + \mu_S \cdot n_S} \simeq$

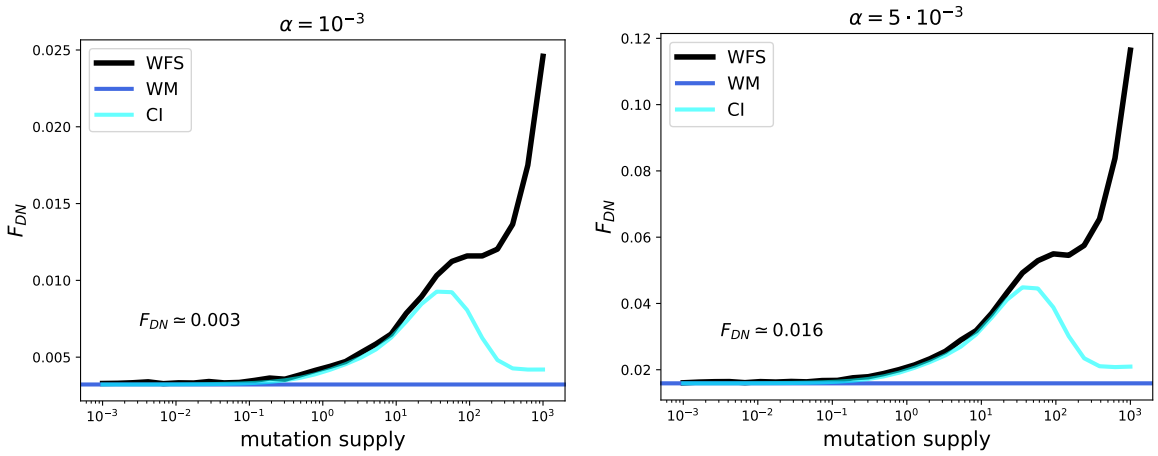

Figure 4.3: Probability of fixation of a double mutant (F_{DN}) for two different values of $\alpha = \mu_D/\mu_S$. In blue the weak mutation model and in cyan the clonal interference model. The shapes are identical and the absolute values is just multiplied by 5.

0.29% which is slightly smaller than F_{DN}. This is due to genetic drift, indeed the average fitness of beneficial is larger in doubles than singles and the fraction of beneficial mutant is the same. In the clonal interference regime the value of F_{DN} is between the 1% and 2%.

For $\alpha = 10^{-3}$ the probability of fixation of a double mutant F_{DN} is 1.59% in the week mutation regime. The fraction of double de novo mutations $f \simeq 1.48\%$. In the clonal interference regime the value of F_{DN} is between the 5% and 10%.

As in the previous chapter, the relevance of adjacent nucleotide mutations is strictly related to the value of α and to the regime (value of the mutation supply).

As last analysis we compared the mean fitness jump computed considering only single and considering single and doubles. We plot the results for two different values of α in figure 4.4. As said before the fittest single's fitness is 1.80 while the fittest double's fitness is 3.14. This is the reason why the mean fitness jump percentage increment (y axis) grows substantially as the mutation supply increases.

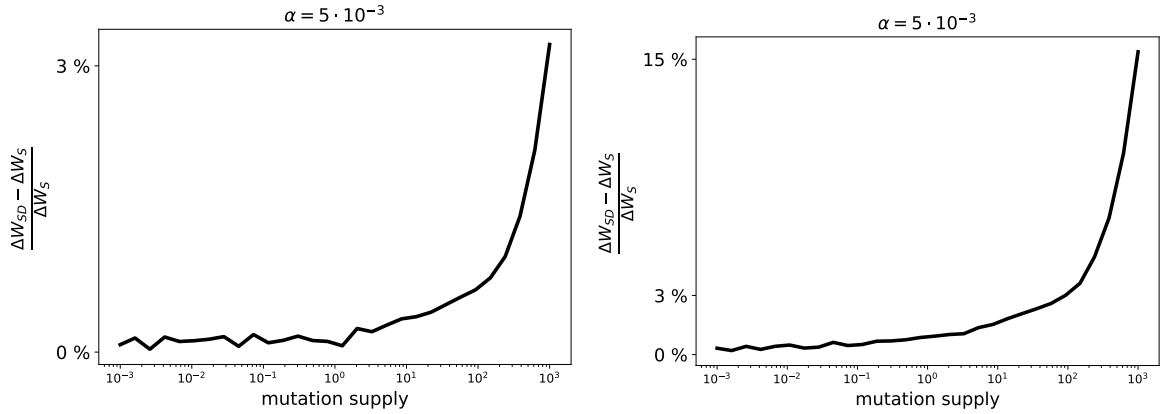

Figure 4.4: Mean fitness jump percentage increment. We computed the relative mean fitness jump: $\Delta = \frac{\Delta W_D - \Delta W_S}{\Delta W_S}$ as a function of the mutation supply.

Chapter 5

Conclusion

The main goal of this project is to study the effect of multi nucleotide mutations on adaptive evolution. Multi nucleotide mutations have a rate that is $10^{-3} - 10^{-2}$ times the rate of single nucleotide mutations. Although MNMs occur much less frequently, they offer the advantage that substantially more genotypes can be reached through one mutation event. The project focuses on a subclass of multi nucleotide mutations which is adjacent double nucleotide mutations (doubles). This kind of mutations involves two neighboring nucleotides. The number of beneficial doubles is 5/3 the number of beneficial singles. Because of that, if we assume a random landscape with the same DFE for singles and doubles, it is more likely that the fittest genotype is a double. The probability of fixation of a genotype depends on its rate of mutation and its fitness; as the mutation supply increases the fitness becomes more and more important. The trade off between mutation rate and fitness is the crucial point of this study. One important quantity that allow us to estimate the relevance of doubles is the probability that a double fixates first F_{DN}. Estimating F_{DN} means estimating how frequently evolution acts via adjacent double nucleotide mutations.

Before employing a population genetic model we introduce the simple model. The latter enables us to gain a qualitative understanding of the dynamics and relevant details. Thanks to it we can study the behavior of F_{DN} and its average $\langle F_{DN} \rangle$ at different mutation supply regimes (assuming $f_S(\omega) = f_D(\omega)$). For small value of the mutation supply the value of F_{DN} does not depend on the landscape but only on n_S, n_D and their rates μ_S, μ_D. For large values of the mutation supply competition is only between the fittest single and the fittest doubles. Since $n_D > n_S$ it is more likely that the fittest double is larger than the fittest single. For this reason $\langle F_{DN} \rangle$ increases monotonically with the mutation supply. For intermediate values of the mutation supply $\langle F_{DN} \rangle$ is constant. In this regime F_{DN} has a behavior that depends on the details of

the landscape and can be also non monotonic.
After that we introduce and implement the Wright Fisher model. We study the value F_{DN} and of its landscapes average $\langle F_{DN} \rangle$ as a function of the mutation supply $N \cdot U$. The results for a random landscape with the same DFE for singles and doubles are summarized in the following table:

Parameters			**Mutation supply**			
α	n_S	n_D	10^{-2}	10^{-1}	10^{-0}	10^{1}
$1 \cdot 10^{-3}$	12	20	0.2%	0.3%	0.9%	5.5%
$5 \cdot 10^{-3}$	12	20	1%	1%	3%	12%

This results suggest that the frequency of fixation of double adjacent nucleotide mutations can be non-negligible for large value of Nt. Afterwards, we analyze the effect of doubles on the mean fitness jump. We compute the relative difference between the mean fitness jump computed considering only singles and considering doubles and singles. The results for a random landscape with the same DFE for singles and doubles are summarized in the following table ($NU = 10, n_S = 12, n_D = 20$):

α	$\langle \Delta \rangle$
$1 \cdot 10^{-3}$	4.5%
$5 \cdot 10^{-3}$	8%

Also in this case the results suggest that for large values of the mutation supply NU the doubles can be non-negligible. To conclude this part we study the how of the number of beneficial mutants effects the relative mean fitness jump increment $\langle \Delta \rangle$. The results show that, for exponential DFEs, $\langle \Delta \rangle$ decreases as the number of beneficial mutants increases. A possible extension of this part should be to investigate how $\langle F_{DN} \rangle$ and $\langle \Delta \rangle$ changes if we use a heavy tail distribution as DFE.
Finally, we study evolution on an empirical fitness landscape. The landscape is of the TEM-1 gene. We compare the single's fitness landscape and the double's fitness landscape. The double mutants are on average less fit than the single mutants. However, if we restrict the analysis on the beneficial mutants the result is the opposite. If we compare the single's beneficial fitness landscape and the double's beneficial fitness landscape, the latter has a fatter tail and the fittest genotype is a double. By simulating the Wright Fisher model on the empirical landscape we obtain the following results:

α	$NU = 10^{-3}$	$NU = 10^{-}3$
$1 \cdot 10^{-3}$	0.3%	2.5%
$5 \cdot 10^{-3}$	1.5%	12%

$NU = 10^3$	
α	$\langle\Delta\rangle$
$1 \cdot 10^{-3}$	3%
$5 \cdot 10^{-3}$	15%

Even empirical landscape results suggest that double adjacent nucleotide mutations can be non-negligible for large value of Nt. Could be interesting to study other empirical landscapes to see how general are this results.

Eventually one possible way to take after this project could be to study the influence of adjacent nucleotide mutations on evolution on a whole random landscapes, without the one step restriction. Another way of extend this work could be to quantify the effect of valley crossing via double mutations.

Bibliography

[1] Günter P. Wagner and Jianzhi Zhang. "The pleiotropic structure of the genotype–phenotype map: The Evolvability of complex organisms". In: *Nature Reviews Genetics* 12.3 (2011), pp. 204–213. DOI: 10.1038/nrg2949.

[2] Patrick C. Phillips. "Epistasis — the essential role of gene interactions in the structure and evolution of Genetic Systems". In: *Nature Reviews Genetics* 9.11 (2008), pp. 855–867. DOI: 10.1038/nrg2452.

[3] Ben Lehner. "Genotype to phenotype: Lessons from model organisms for human genetics". In: *Nature Reviews Genetics* 14.3 (2013), pp. 168–178. DOI: 10.1038/nrg3404.

[4] Carl T. Bergstrom and Lee Alan Dugatkin. *Evolution*. Norton amp; Co., 2012.

[5] Sewall Wright et al. "The roles of mutation, inbreeding, crossbreeding, and selection in evolution". In: (1932).

[6] J. Arjan de Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution". In: *Nature Reviews Genetics* 15.7 (2014). DOI: 10.1038/nrg3744.

[7] David M. McCandlish. "Visualizing Fitness Landscapes". In: *Evolution* 65.6 (2011), pp. 1544–1558. DOI: 10.1111/j.1558-5646.2011.01236.x.

[8] J. F. Kingman. "On the properties of bilinear models for the balance between genetic mutation and selection". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 81.3 (1977), pp. 443–453. DOI: 10.1017/s0305004100053512.

[9] Uri Obolski, Yoav Ram, and Lilach Hadany. "Key issues review: Evolution on rugged adaptive landscapes". In: (2017). DOI: 10.1101/112177.

[10] Stuart Kauffman and Simon Levin. "Towards a general theory of adaptive walks on rugged landscapes". In: *Journal of Theoretical Biology* 128.1 (1987), pp. 11–45. DOI: 10.1016/s0022-5193(87)80029-2.

[11] Stuart A. Kauffman and Edward D. Weinberger. "The NK model of rugged fitness landscapes and its application to maturation of the immune response". In: *Journal of Theoretical Biology* 141.2 (1989), pp. 211–245. DOI: 10.1016/s0022-5193(89)80019-0.

[12] Johannes Neidhart, Ivan G Szendro, and Joachim Krug. "Adaptation in tunably rugged fitness landscapes: The rough mount fuji model". In: *Genetics* 198.2 (2014), pp. 699–721. DOI: 10.1534/genetics.114.167668.

[13] Takuyo Aita et al. "Analysis of a local fitness landscape with a model of the Rough Mt. fuji-type landscape: Application to prolyl endopeptidase and Thermolysin". In: *Biopolymers* 54.1 (2000), pp. 64–79. DOI: 10.1002/(sici)1097-0282(200007)54:1<64::aid-bip70>3.0.co;2-r.

[14] John H. Gillespie. "Some properties of finite populations experiencing strong selection and weak mutation". In: *The American Naturalist* 121.5 (1983), pp. 691–708. DOI: 10.1086/284095.

[15] H. Allen Orr. "The genetic theory of adaptation: A brief history". In: *Nature Reviews Genetics* 6.2 (2005), pp. 119–127. DOI: 10.1038/nrg1523.

[16] Scott William Roy. "Probing evolutionary repeatability: Neutral and double changes and the predictability of evolutionary adaptation". In: *PLoS ONE* 4.2 (2009). DOI: 10.1371/journal.pone.0004500.

[17] Philip J. Gerrish and Richard E. Lenski. "The fate of competing beneficial mutations in an asexual population". In: *Mutation and Evolution* (1998), pp. 127–144. DOI: 10.1007/978-94-011-5210-5_12.

[18] John W Drake. "Contrasting mutation rates from specific-locus and long-term mutation-accumulation procedures". In: *G3 Genes—Genomes—Genetics* 2.4 (2012), pp. 483–485. DOI: 10.1534/g3.111.001842.

[19] Martin Marinus. "Faculty opinions recommendation of rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing." In: *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2016). DOI: 10.3410/f.726446730.793521043.

[20] Sébastien Wielgoss et al. "Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with escherichia coli". In: *G3 Genes—Genomes—Genetics* 1.3 (2011), pp. 183–186. DOI: 10.1534/g3.111.000406.

[21] Søren Besenbacher et al. "Multi-nucleotide de novo mutations in humans". In: *PLOS Genetics* 12.11 (2016). DOI: 10.1371/journal.pgen.1006315.

[22] Danielnbsp;R. Schrider, Jonathannbsp;N. Hourmozdi, and Matthewnbsp;W. Hahn. "Pervasive multinucleotide mutational events in eukaryotes". In: *Current Biology* 21.12 (2011), pp. 1051–1054. DOI: 10.1016/j.cub.2011.05.013.

[23] Nadezhda V. Terekhanova et al. "Prevalence of multinucleotide replacements in evolution of primates and drosophila". In: *Molecular Biology and Evolution* 30.6 (2013), pp. 1315–1325. DOI: 10.1093/molbev/mst036.

[24] J.M. Goldmann, J.A. Veltman, and C. Gilissen. "De novo mutations reflect development and aging of the human germline". In: *Trends in Genetics* 35.11 (2019), pp. 828–839. DOI: 10.1016/j.tig.2019.08.005.

[25] Carla L. Dinardo et al. "Diversity of variant alleles encoding kidd, Duffy, and kell antigens in individuals with sickle cell disease using whole genome sequencing data from the nhlbi topmed program". In: *Transfusion* 61.2 (2020), pp. 603–616. DOI: 10.1111/trf.16204.

[26] Kelley Harris and Rasmus Nielsen. "Error-prone polymerase activity causes multinucleotide mutations in humans". In: *Genome Research* 24.9 (2014), pp. 1445–1454. DOI: 10.1101/gr.170696.113.

[27] Nick G. Smith, Matthew T. Webster, and Hans Ellegren. "A low rate of simultaneous double-nucleotide mutations in primates". In: *Molecular Biology and Evolution* 20.1 (2003), pp. 47–53. DOI: 10.1093/molbev/msg003.

[28] Michalis Averof et al. "Evidence for a high frequency of simultaneous double-nucleotide substitutions". In: *Science* 287.5456 (2000), pp. 1283–1286. DOI: 10.1126/science.287.5456.1283.

[29] Frida Belinky et al. "Crossing fitness valleys via double substitutions within codons". In: *BMC Biology* 17.1 (2019). DOI: 10.1186/s12915-019-0727-4.

[30] H.A David and H.N Nagaraja. *Order statistics*. Wiley-Interscience, 2003.

[31] Sarah P. Otto and Troy Day. *A biologist's Guide to Mathematical Modeling in Ecology and evolution*. Princeton university press, 2007.

[32] Alex S. Fraser. "An introduction to population genetic theory. by J. F. Crow and M. Kimura. Harper and row, New York. 656 pp. 1970". In: *Teratology* 5.3 (1972), pp. 386–387. DOI: 10.1002/tera.1420050318.

[33] David M. McCandlish and Arlin Stoltzfus. "Modeling evolution using the probability of fixation: History and implications". In: *The Quarterly Review of Biology* 89.3 (2014), pp. 225–252. DOI: 10.1086/677571.

[34] Elad Firnberg et al. “A comprehensive, high-resolution map of a gene’s fitness landscape”. In: *Molecular Biology and Evolution* 31.6 (2014), pp. 1581–1592. DOI: 10.1093/molbev/msu081.

[35] Courtney E. Gonzalez and Marc Ostermeier. “Pervasive pairwise intragenic epistasis among sequential mutations in TEM-1 -lactamase”. In: *Journal of Molecular Biology* 431.10 (2019), pp. 1981–1992. DOI: 10.1016/j.jmb.2019.03.020.