UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

OSMAR FREITAS DA SILVA JÚNIOR

THE INFLUENCE OF THE BOTTLENECK PROTOCOL ON THE ADAPTATION RATE AND PREDICTABILITY

Recife

2021

OSMAR FREITAS DA SILVA JÚNIOR

THE INFLUENCE OF THE BOTTLENECK PROTOCOL ON THE ADAPTATION RATE AND PREDICTABILITY

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Física.

Área de Concentração: Física Teórica e Computacional

Orientador: Paulo Roberto de Araújo Campos

Recife

2021

Catalogação na fonte
Bibliotecária Fernanda Bernardo Ferreira, CRB4-2165

S586i Silva Júnior, Osmar Freitas da
The influence of the bottleneck protocol on the adaptation rate and predictability / Osmar Freitas da Silva Júnior. – 2021.
99 f.: il., fig.

Orientador: Paulo Roberto de Araújo Campos.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Física, Recife, 2021.
Inclui referências e apêndices.

1. Física Teórica e Computacional. 2. Bottleneck. 3. Previsibilidade. I. Campos, Paulo Roberto de Araújo (orientador). II. Título.

530.1 CDD (23. ed.) UFPE- CCEN 2021 - 81

OSMAR FREITAS DA SILVA JÚNIOR

THE INFLUENCE OF THE BOTTLENECK PROTOCOL ON THE ADAPTATION RATE AND PREDICTABILITY

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Física.

Aprovada em: 16/04/2021.

BANCA EXAMINADORA

_____

Prof. Paulo Roberto de Araujo Campos
Orientador
Universidade Federal de Pernambuco

_____

Prof. Pedro Valadão Carelli
Examinador Interno
Universidade Federal de Pernambuco

_____

Profa. Sabrina Borges Lino Araújo
Examinadora Externa
Universidade Federal do Paraná

ACKNOWLEDGEMENTS

Agradeço à meu pai por ter me ensinado que pra tudo na vida se dá um jeito, à minha mãe por me mostrar que quanto mais se vive mais a gente aprende, e aos meus irmãos, sem os quais me faltariam cúmplices nesse dilema que é fazer parte da nossa família.

Agradeço ao Prof. Paulo Campos pela oportunidade que me foi dada e pela paciência ao me orientar em uma área na qual sempre tive muito fascínio, porém poquíssima experiência. Em tempos de distanciamento social, sua confiança e disponibilidade foram essenciais para desenvolver essa pesquisa.

Em particular, aos professores Leonardo Cabral, Leonardo Menezes e Alessandro Villar, que muito me ensinaram para além da sala de aula.

Agradeço também às instituições de fomento CNPq e CAPES, sem as quais esta pesquisa não se realizaria; à estrutura fornecida pela UFPE, pelo Departamento de Física e pelo laboratório de Dinâmica Evolucionária.

E, por fim, agradeço à estocaticidade dos eventos históricos que fizeram com que hoje eu tenha, sempre ao meu lado, minhas companheiras de vida: Ravena e Suricat.

ABSTRACT

Bottlenecks are evolutionary events that reduce the population size. In experimental evolution, the cultivation of microorganisms in laboratory conditions follows a serial passaging protocol in which periodic bottlenecks are an inherent aspect. In this work, we study a computational model of microbial evolution to understand the influence of the bottleneck protocol on the rate of adaptation and predictability of the population. To address these questions, evolutionary simulations are run using a standard Wright-Fisher model integrated with a bottleneck regime, and the implementation of a fitness landscape with varying ruggedness. The rate of adaptation is analyzed as a function of bottleneck and population sizes at different time units. It is found that the rate of adaptation depends monotonically on bottleneck size when time is expressed *per generation*, but is maximal at intermediate bottleneck size for times expressed *per bottleneck* and *per birth*. Additionally, fitness landscapes allow to register the change in the population's genetic composition as trajectories over the genotypic space. Under this path-dependent perspective, an ensemble of trajectories is generated in many independent runs, for which statistical and computational measurements of predictability are inferred. Irrespective of the timing of population bottlenecks, we find that predictability increases with population size. We also find that predictability of the adaptive pathways increases in increasingly rugged fitness landscapes.

Keywords: Bottleneck. Predictability. Adaptation rate. Population genetics.

RESUMO

Gargalos, ou *bottlenecks*, são eventos evolutivos que reduzem o tamanho de uma população. Em experimentos evolutivos, o protocolo para o cultivo de microrganismos em laboratório costuma impor estas amostragens de maneira regular e periódica. Neste trabalho, estudamos um modelo computacional de evolução microbiana para entender como este protcolo pode influenciar a taxa de adaptação de uma população e sua previsibilidade. Para abordar essas questões, simulamos uma população que evolui sob este protocolo usando o modelo de Wright-Fisher e um relevo de adaptação com epistasia regulável. A taxa de adaptação é medida em função da severidade do *bottleneck* em diferentes unidades de tempo. Verificamos que a taxa de adaptação depende monotonicamente do tamanho do *bottleneck* quando o tempo é expresso *por geração*, mas apresenta um máximo em tamanhos intermediários quando os tempos são expressos em unidades *por bottleneck* e *por nascimento*. A adoção de um relevo de adaptação permite registrar a mudança na composição genética da população, e obter um *ensemble* das trajetórias evolutivas no espaço de genótipos. Utilizamos medidas estatísticas e computacionais para inferir o grau de previsibilidade destas trajetórias. Destacamos que a previsibilidade aumenta com o tamanho da população, independentemente do tamanho da amostragem. Também observamos que a previsibilidade das trajetórias evolutivas aumenta em função da epistasia do relevo de adaptação.

Palavras-chaves: *Bottleneck*. Previsibilidade. Taxa de adaptação. Dinâmica populacional.

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

CDF	Cumulative Density Function
CI	Clonal Interference
GO	Global Optimum
LTEE	Long-Term Evolution Experiment
MDS	Multidimensional Scaling
PGF	Probability Generating Function
RAW	Random Adaptation Walks
SSWM	Strong-Selection Weak-Mutation
SWW	S-Weighted Walks

CONTENTS

1 INTRODUCTION

"Any observation of a living system
must ultimately be interpreted
in the context of its evolution." (1)

Evolutionary Theory is one of the most beautiful and successful theories in science. Through the set of its essential mechanisms, we came to understand the complex collective behavior of living organisms, to track the origin of their common ancestors, and to account for the abundant variety of life forms.

Over the years, formal models were developed to study biological systems whose fates are strongly driven by evolutionary processes, such as selection, mutation, migration, and genetic drift. Working on quantification of those processes, Evolution became a field grounded in well established mathematical foundations, in which hypotheses can be explored and confronted with empirical data. Its body of work ranges over different scales of interaction between living organisms - from Population Genetics to Ecology.

At the genetic level, problems revolve around the change of a population's genetic composition over time. While selection favors the most adapted individuals, the nature of the reproduction process itself affects the population's diversity through genetic drift and mutation. The non-trivial interaction between these factors over time may render the trajectory of the evolutionary process highly susceptible to chance events.

In the 80s, a warm debate has emerged about how the fate of an evolving population is affected by the particularities of its history - an aspect known as Contingency (2). Would such a fate be completely arbitrary and thus unpredictable? Are there enough constraints that effectively reduce the number of 'paths' an evolving population could take? If the latter is true, how reproducible an evolutionary process is? Could we "replay the tape of life"?

Although such ideas, at the time, were more heuristic than passive of observation, it was the seed - or Gedanken experiment, if you prefer - to consider the quantification of randomness alongside the evolutionary history: how stochastic is the dynamic of an evolving population? What kind of elements can affect the predictability of its composition over time? In recent years, parallel evolution experiments alongside exhaustive analyzes of genetic material from lab and natural populations have started to accumulate evidence that some 'paths' leading to the final composition have different probabilities of occurrence (3). Such a novelty has fundamental

implications in the understanding of mechanisms of adaptation, immune response, protein folding, as well as antibiotic resistance (4).

Under well-controlled environment conditions, one can cultivate simple organisms - such as bacteria, viruses, and yeast - to observe evolution acting upon them at shorter time scales, selecting those more adapted to the lab conditions. Through an Experimental Protocol developed throughout the years, these Evolutionary Experiments became a rich niche for testing theoretical models with empirical data (5). Inherent to this procedure, the adopted convention of a bottleneck regime in those protocols - severe and periodic reduction in population size, after a growth phase - is of interest to the theme of contingency, since it is well known that larger populations are less sensitive to randomness than smaller ones, besides its effect on the adaptation rate of the evolutionary process (6).

As such, the goal of the present work is to study a model - through computer simulations - of a population of self-replicating entities evolving by selection, mutation, and genetic drift according to a serial-transfer protocol of experimental evolution. We ought to investigate whether the bottleneck regime can affect the population's rate of adaptation and the predictability of its genetic composition .

We resort to the study of evolution over a fitness landscape (7). We build a genotype-to-fitness map where the genetic composition of an evolving population can be registered. Under this path-dependent perspective, the dynamic of the genetic variation describes an adaptive trajectory over this genotypic space (8). By generating an ensemble of trajectories over several distinct fitness landscapes we can assess different measurements of the predictability of the system (9).

In summary, two key factors characterize our approach:

1) A genotypic space is used to map the population dynamics. Such formulation allows us to:

- record the genetic states visited by the population – an adaptive trajectory in this map;
- utilize statistical measures of predictability in the ensemble of trajectories;
- investigate how the correlation between the elements of the genome (epistasis) can interfere in the accessibility of paths.

2) A bottleneck regime is implemented for the evolving population. We investigate its implications for the adaptive rate and the predictability of the trajectories.

By modeling any system, we should ask ourselves what are the assumptions adopted a priori and the range of its applicability. Efforts were made for those points being highlighted throughout this work. In Chapter 1 it is presented a brief historical context of the modern evolutionary theory, converging to the particular topics of our research. Here is discussed the conceptual framework of a fitness landscape, the protocol utilized in experimental evolution, and the notion of predictability. In Chapter 2, we work out the basic concepts of population genetic theory to better formalize the evolutionary processes involved in the dynamics. It is also discussed the properties of the genotype space and how they affect the structure of the fitness landscape and, hence, the accessibility of paths. In Chapter 3, we describe the methods and models chosen to simulate our dynamics, the adopted protocols to compare adaptation rates and generate the ensembles of trajectories, and the statistical and computational measurements of the predictability of the system. In Chapter 4, the analytical and simulation results of our research are presented and discussed alongside previously published works in the experimental and theoretical fields. We highlight the implications of a bottleneck regime at the macroscopic and microscopic level of the evolving population: from the fitness trajectories achieved by adaptation, to the change in genetic diversity, to the statistical properties of the evolutionary pathways. Finally, in chapter 5, we summarize the main results obtained along this dissertation, and discuss the perspectives for future investigations.

1.1 HISTORICAL VIEW

1.1.1 Darwinism and Mendelian Genetics

In the documented history of human-kind, efforts in the understanding of life's origin and diversity are present in many ancient civilizations. However, the history of a modern evolutionary theory as we know it today began recently.

In the early XIX century, Naturalists tackled the discussion of life variability using an exclusive observational approach. Through the analyses of fossil records and a meticulous morphological characterization of organisms, the idea of species sharing common origins and the fact that different variations were present in the same species was already granted. To explain why some traits prevailed while others disappeared, the Lamarckian theory of differentiation claimed that the environment gives rise to change in animals by imposing a higher utilization of some organs, granting them a greater development (10). To Lamarck, these changes would

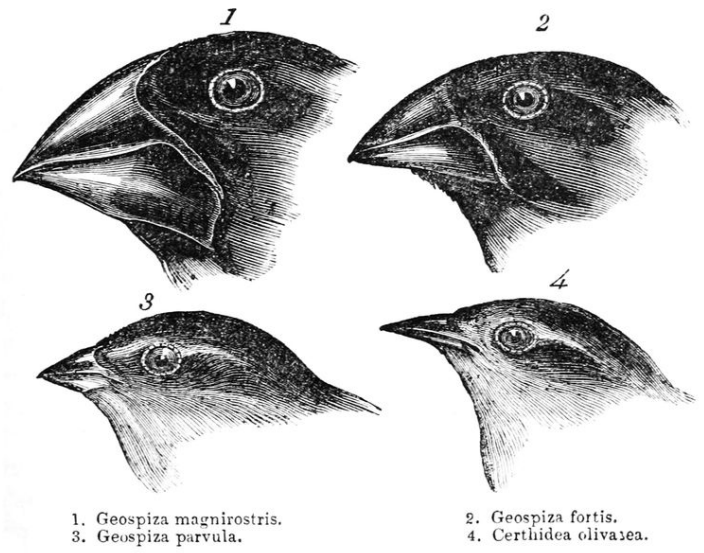
then be continuously acquired over the organisms' life and passed to their offsprings, giving origin to new species. From this perspective, the environment was understood as a source of variation.

This predominant view was challenged in 1858, when Charles Darwin and Alfred Russel Wallace announced an alternative process they have attained, independently (10), to explain the fate of the observed variation of traits. In their view, the traits - or phenotypes - are transferred exclusively by inheritance from the ancestral parents, with no change along the organisms' life, and has specific effects not only on the physical characteristics but also on the behavior of the offspring. In a given environment, some traits would bear a relative advantage among the others in terms of reproduction success and surviving stages of life development - such general conditions are often referred to as fitness. Therefore, the ones with 'bad' traits within a given species would leave fewer descendants, on average, than the 'fitter' ones.

In principle, many variations could be established simultaneously, but resource and spatial limitation impose a limit on the number of prevailing traits. Given enough time, subsequent generations would present a diversity of these variants, with frequency proportional to their respective fitness.

This new perspective brought about a paradigmatic change to the understanding of life (11). The environmental conditions *together* with the variation of traits would determine the perpetuation of the organisms. Coined Natural Selection, this process would shape the diversity of a given population over time.

Figure 1 – Simple examples of variation of traits among the same species. The beak of flinches changed gradually as their subsequent offspring adapted to different diets such as seeds, insects, and fruits.



Source: Public domain image from (12).

In the words of Pugliucci (11), natural selection and the inheritance of traits were "the first two conceptual pillars of modern evolutionary theory". Their presentation and further applications were better formalized in Darwin's book *On the Origin of Species (1859)* (10). Since a change in the environment can affect the distribution of fitness among traits within a population, such dependence could be straightforwardly related to open problems involving the discontinuity of fossil records, geographically isolated animals (Figure 1), and the commonly used process of animal and plant breeding - aka artificial selection. But a convincing hypothesis about the fundamental origin of the variations was lacking so far. "For a population to evolve by natural selection, the members of the population must vary - if all organisms are identical, no selection can occur" (13).

Contemporaneous to those ideas, Gregor Mendel (13) was conducting a controlled experimental work on garden pea plants, interbreeding harvests with different traits. The pure plant lineages differed, for example, by seed shape, flower color, and height. As new generations sprouted, Mendel observed a proportionality pattern in the distribution of the different traits. By carefully analysing the data, he hypothesized that each plant contained a pair of units that determined its trait — for example, flower color.

Mendel's theory was the first to conceive the inheritance of traits as discrete hereditary components. Mendel's units and the variations it can take (e.g. pink or white colors in the pea plant petals) are known, respectively, as the genes and alleles of population genetics (13). Thus recombination of the parent's genes determines which alleles will be present in the offspring.

Although Mendel's discoveries mark a turning point in our understanding of inheritance, its importance took almost forty years to be noticed by the scientific community and one more decade to be related to Darwinism's initial variation problem.

1.1.2 Modern Synthesis and the Genome

At the turn of the 20th century, when Mendel's ideas were rediscovered (11), Darwinians claimed that the heredity of discrete traits - such as the flower color - studied by Mendel, was incompatible with the gradualism view of natural selection as it could not explain the variation in continuous traits - such as body size. Thus, it could not be responsible for the long term change in species (7). In contrast, Mendelians emphasized that a discontinuous variation of traits was not only universal for evolution but that major adaptive change could be produced by single hereditary steps (13). Over the years, the accumulation of empirical evidence favoring

Mendelianism brought an urgent need to unify both theories.

Achievements in this direction were made in the 1920s, through the efforts of R.A. Fisher, J.B.S. Haldane and Sewall Wright (8). They developed formal mathematical models to explore how the mechanisms of selection, drift, and mutation, would modify the population's genetic composition, obeying the Mendelian rules of inheritance.

In particular, Fisher's (14) *The correlation between relatives on the supposition of Mendelian inheritance (1918)* is "one of the most important papers ever written in evolutionary biology" (11). Fisher demonstrated that if a large number of independent genes give small contributions to the change of a continuous trait, the sum of trait effects would approximate a normal distribution in a population - as happens through the Central Limit Theorem with the sum of many independent random contributions (15). "Since the Darwinian process was widely believed to work on continuously varying traits, the demonstration that the distribution of such traits was compatible with Mendelism was an important step towards reconciling both theories" (13).

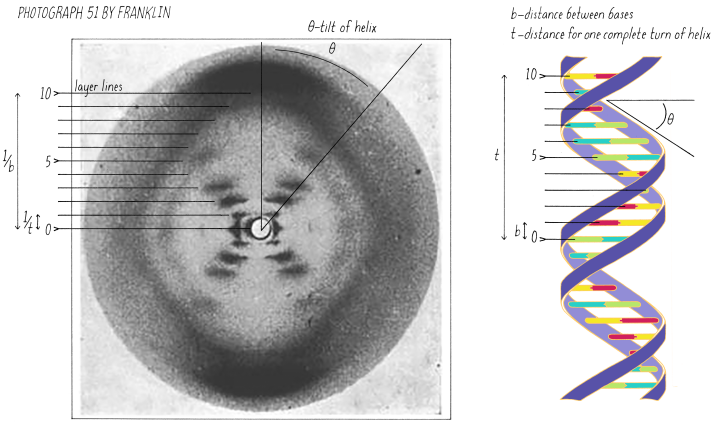
The conceiving of formal models to investigate the process of evolution played a key part in the formation of the Modern Synthesis. Remarkably, they were capable to integrate the Mendelian nature of mutation without mention to the real nature of the genomic and molecular basis of inheritance - which was unknown at the time (11).

In the 40s, it was already known that the DNA carries the hereditary genetic material, that it was located on chromosomes, and that each cell comprising any organism contains a chromosome set (16); but at the time there was no X-ray crystallography or electron microscopy to further increase the resolution and reveal its structure. Within a theoretical perspective, in the series of lectures entitled *What is Life? (1943)* (17), Erwin Schrödinger highlighted that if the genetic inheritance was stored in objects with the size of single molecules, it should have an aperiodic crystal structure in order to store the abundant information about the individual development and retain its stability.

This informational approach to the problem inspired many molecular biologists in the field (17), including J. Watson and F. Crick. In the early 50s, they were working on a model of code-transcription to fulfill the genetic heredity theory, when Rosalind Franklin et al. took the famous Photo 51 from a crystalline gel composed of DNA fiber (Figure 2).

With the refinement of crystallography resolution techniques, it revealed a double-helix string connected through bridges. Watson and Crick integrated this novel structure into their models of code-script information and two years later (16) announced the discovery of the genome structure present in every living being: a codified string, whose length and sequence

Figure 2 – Photo 51 and double-helix DNA structure.



Source: Image from (16).

combination of its elements could provide the expressed traits (8). The specific sequence of nucleotides – or genotype – is the inherent hereditary information passed through generations, and the change in its code can cause the variation needed for natural selection to act upon.

Almost one century separates the conception of Darwinism from the discovery of the genome. In the last third of the 20th century, Modern Synthesis was constantly modified owing to the abundance of data provided by the genomic revolution (7), thus turning feasible fast whole-genome sequencing. Today we know that many of its common assumptions can not be arbitrarily generalized. For example, it has been found that the genome is not always a well-organized set of genes and that not all genes necessarily have a single function[1]. Pigliucci(11) argues that a new perspective of evolution itself may integrate genomics, complex theory, and evolutionary genetics into an "Extended Evolutionary Synthesis". Despite that, the succession of events here exposed consolidates the efforts of many scientists from many distinct fields and generations to build a reasonable and accurate understanding of life itself. The Modern Synthesis' paradigm sheds light on the machinery shared by all living organisms - even those separated by hundreds of millions of years of evolutionary history.

1.2 GENOTYPE TO FITNESS MAPPING

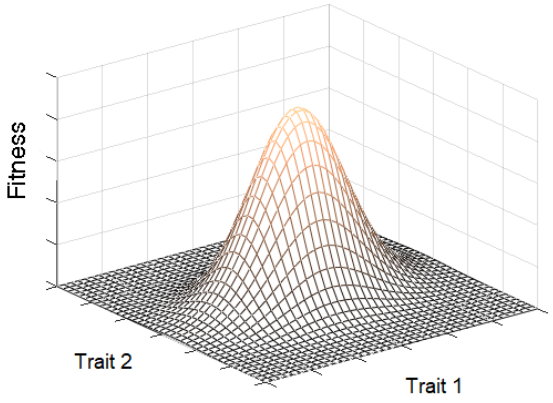
In the early development of the Modern Synthesis, most of the models were limited to the analyses of few genes (or loci), assuming that their effects on the traits were independent of

[1] See reference (11) for many more.

the rest of the genome (8). The first attempts to formalize the effects of genes at multiple loci were presented in the early 30s.

Fisher and Wright were protagonists of a long-standing debate regarding the effect of mutations on the variance of traits and, ultimately, on fitness. Both of them shared the view that the fitness of a population depends on the combined states of their traits. Therefore, there should exist an optimum combination of those traits, such that its fitness is the maximum possible (7). Their ideas were better visualized as a landscape-like figure (Figure 3), where each "point" in the base grid represents a specific combination between the axis traits, and the height corresponds to their respective fitness. Populations experiencing mutations and genetic drift would then move at random directions and lengths over this surface, but, since selection will act on traits to gradually change them to fitter ones, adaptation was seen, they assumed, as an optimization process over this landscape in the search for fitness peaks.

Figure 3 – Simple representation of a phenotypic space. The height is proportional to the fitness of combined traits. As selection increases the frequency of fitter individuals, adaptation is an optimization process towards the peak.

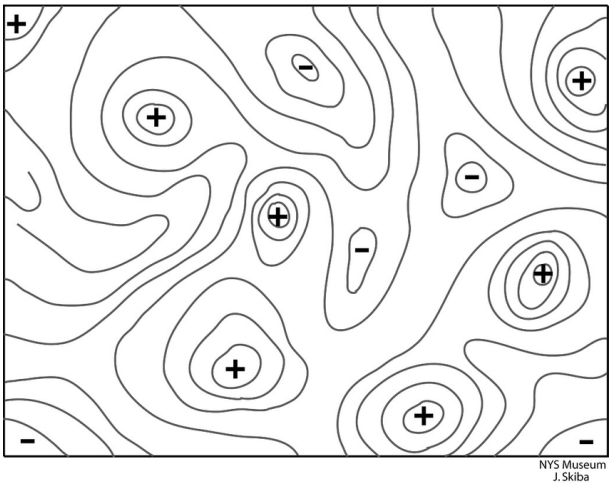


Source: The author (2020).

In Fisher's additive view of genetics, he considered that the independent dimensions should be seen as the phenotypic variation of traits and, as such, they could take continuous values - as it is continuous the length of a beak, the body size or the consumption rate of 'food'. "If there are many different ways to change a phenotype, it becomes very unlikely that a random change acquires the right combination of traits in the right way to improve fitness" (18). Moreover, the model should be limited to a few independent traits, since the existence of maxima in that space decreases as the dimensions increases - as it happens with a saddle point in 3 dimensions. In Fisher's Geometric Model a continuous surface produces a smooth single-peaked fitness landscape.

On the other hand, Wright's (19) adaptive landscape diagram (Figure 4) was the first concept of a discrete genotypic space and a tentative to reconcile a higher set of genes in the theory (7). "Within a particular environment, each genotype can be assumed to have a particular fitness;" furthermore, "genotypes that are relatively close will tend to have more similar fitnesses than genotypes chosen at random" (20). This hypothesis represents Wright's emphasis on the assumption of non-additive interactions between the genes within a single genome - in other words, the effect of a single gene could be dependent on the genetic background it appears - what today is known as epistasis. Through an intuitive portray of this multidimensional map in fewer dimensions, Wright visualized a picture of a rugged landscape that displays multiple peaks and valleys. This new topology implied some meaningful aspects to evolutionary theory that was later confirmed: in a given environment, there could exist multiple optimal combinations of traits for an evolving population.

Figure 4 – Wright's landscape map representation with peaks (+) and valleys (-). Evolving populations would travel through this discrete space in search for optimum combinations of traits.



Source: Image from (19).

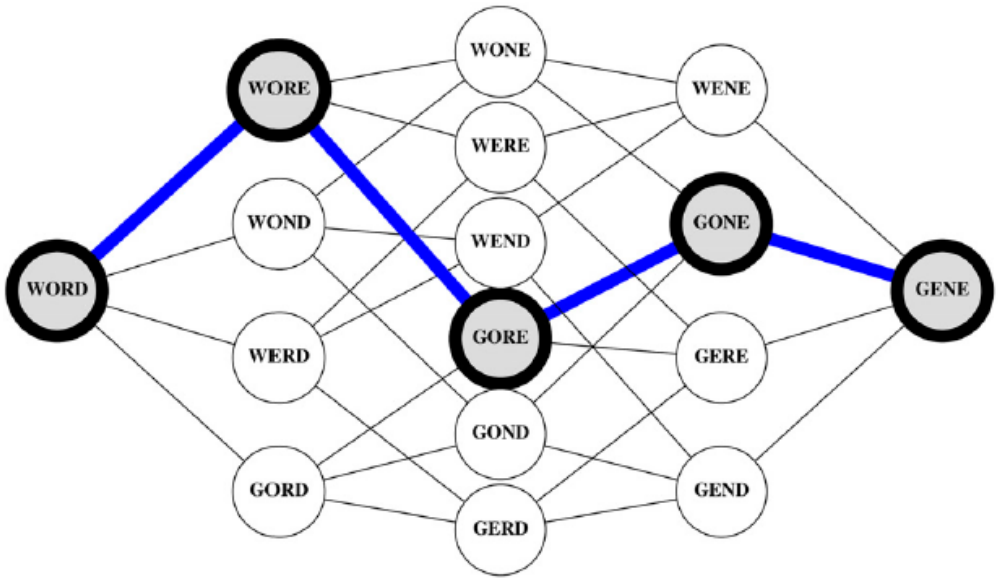
Despite its recognizable importance to evolutionary theory and its great heuristic value, the adaptive landscape as proposed by Wright was the subject of many criticisms. (7, 11, 20, 21). Besides his 'crude' low-dimensional projection, there was no understanding of adaptation at the molecular level for multilocus systems at the time and, consequently, much of Wright's conceptions had a pure conjecture status. In particular, the fundamental elementary units of the model's genotype axis were still unknown and this lack of definition rendered the notion of distance and nearness - based on the assumption of walks through discrete steps - as empty.

In 1970, Maynard Smith addressed this problem by introducing the notion of a mutational pathway in a protein space (21, 22). Proteins differ from one another primarily in their sequence

of amino acids, which in turn comprises a sequence of nucleotides. From all the possible sequences constructed with the 20 existing amino acids, only a subset of them are regarded as functional - or beneficial. Given a protein of fixed length, Maynard asked about the frequency and distribution of amino acid sequences which are indeed functional.

In the original paper, he presented his hypotheses in a ludic way, in the form of a word game: given a fixed number of letters and a specific language, only a subset of combinations encode for existing words.

Figure 5 – Maynard's word game. Single letter transitions from WORD to GENE, over states of existing english words.

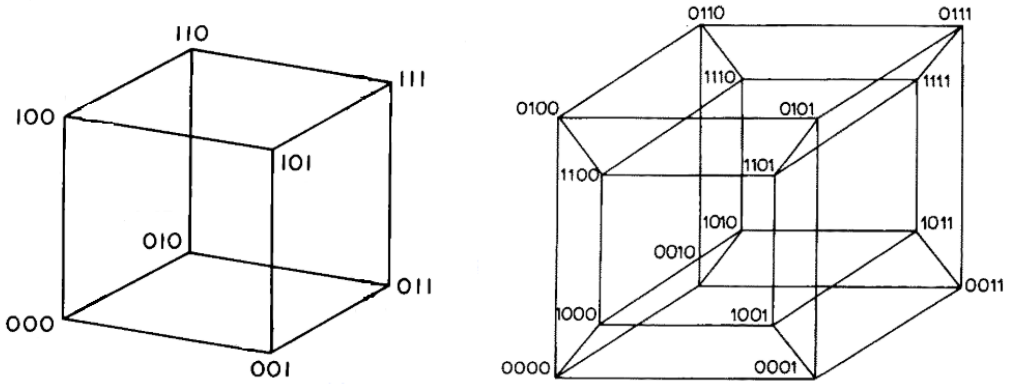


Source: Image from (21).

In the analogy above, letters are seen as amino acids, English words correspond to functional proteins and transitions happens towards (previously defined) fitter words (as in Figure 5). He argued, with outstanding simplicity, that protein evolution can only happen along paths connecting functional proteins since natural selection would tend to purge the pathways connected through non-functional ones. Likewise, he showed that the probability of obtaining the next functional protein decreases as we increase the number of simultaneous amino acid changes. Maynard's arguments and metaphor, although echoing with Fisher and Wright's ideas, were neglected by the scientific community until 1990 (7, 21). When microbial populations were followed in the lab for many generations, adaptation regarded as discrete changes in sequence space was observed. This has renewed the interest in modeling adaptation as an adaptive walk either in phenotypic or genotypic space.

Under his assumptions, the high-dimensional space now gets a more mathematical description. Given a sequence of fixed length L and an alphabet with A letters, a sequence space is a generalization of a Hypercube H_L^A. The Figure 6 shows the simplest case of a binary alphabet $A = \{0, 1\}$: each combination represents a unique state connected to its neighbors through single-step transitions. By assigning a fitness value to each state, evolution is seen as a "walk" in this network and adaptation as a directional "climb" to fitter states. Even with this mathematical formalization of the space, expressions such as "peaks", "smoothness", and "crossing of valleys", although inadequate, are still used by the community in a heuristic manner for their highly intuitive values. Our research is no exception.

Figure 6 – Hypercube representation for L=3 and L=4.



Source: Modified from (23).

Inspired by the protein space conception, Stuart Kauffman et al. (23, 24) developed a statistical computational model for constructing a sequence space of random epistatic interactions - the NK model - where the ruggedness of the landscape (i.e. their average number of fitness peaks) could be "tuned" by the free parameter K, allowing the analysis of adaptation for many families of landscapes. In our work, we use the NK model to assign fitness values to the set of all genotype sequences of a genome of length L^2.

It is important to highlight that, in reality, the relationship between genotype and fitness can be exceedingly complex, especially in the case the environment plays a role and thereby influences the reproductive success (25). Such complexity increases in multicellular organisms, where the early stages of embryo development are determinant to establish the traits (2, 11). Besides that, within a continually changing environment, a fitness attributed to a genotype

[2] Since N is universally related to the population size, we adopt L for the genome length instead of the original choice of the NK model.

could change drastically, and hence, the "heights" of the landscape would vary accordingly - which is sometimes referred through the analogy of a "Sea-landscape". Thus, if the time scale needed for a population to evolve is longer than the change of the landscape itself, it is unfeasible to ascribe absolute or relative values of fitness to genes.

"Nevertheless, a static fitness landscape that depends only on the genotype can be a good approximation on short time scales, if all the mentioned factors remain essentially constant, or under laboratory conditions that can be held constant over many generations" (8). Furthermore, our approach places substantial emphasis on experiments with simple unicellular microbes, for which the variety of expressed traits are less abundant and genomic variation can be carefully registered (5).

Today, the techniques of sampling of genetic material have, so far, been improved in time and range, making the construction of empirical fitness landscapes feasible. Properties of the landscape - such as peak fraction, deviation from additivity, and accessibility of paths (26) - are investigated to better understand the restrictions that the sequence space itself imposes to evolutionary dynamics, integrating a path-dependent analysis to evolutionary theory.

1.3 EXPERIMENTAL EVOLUTION AND BOTTLENECKS

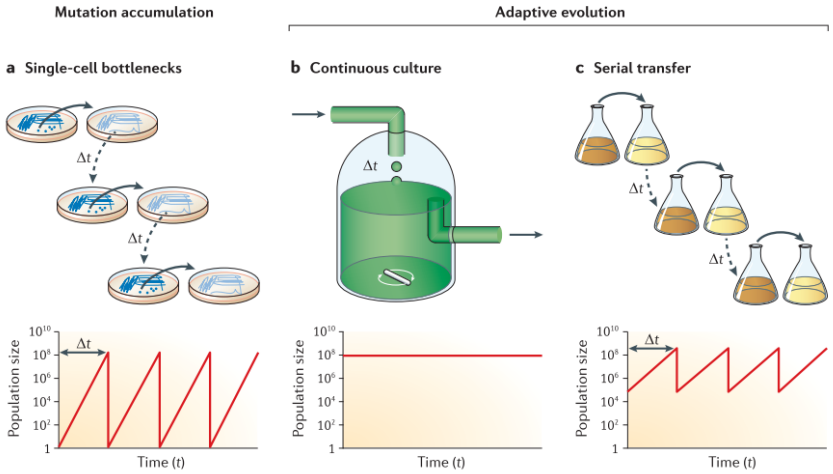
For almost 30 years, researchers have developed experimental protocols to cultivate microorganisms in laboratory conditions to study and monitor the evolutionary process. Controlled and replicated experiments are used to test hypotheses and investigate how their genotypic and phenotypic properties evolve over many generations.

The most suited candidates for such experiments have been viruses, bacteria, fungi, and unicellular algae. As exhaustively addressed by Elena and Lenski (27), among the reasons to use those organisms in lab conditions we highlight: their easiness to propagate and enumerate; their fast reproduction rate, rapidly achieving many generations; the possibility of being stored in a frozen state and later revived, permitting a direct comparison of ancestral and evolved types; and the possibility of replication experiments. The environmental conditions from a laboratory are easier to manipulate, such as resources and temperature, as well as the genetic composition of founding populations. Additionally, the techniques presented today allows a fast and precise genetic analysis and manipulation.

The first and longest ongoing evolution experiment started in 1988, with Richard Lenski (2, 28, 29) cultivation of 12 identical populations of *E.coli* subjected to constant lab conditions.

This Long Term Evolution Experiment (Long-Term Evolution Experiment (LTEE)) has been carried out for more than 30 years, and became the initial standard protocol to build such systems. At the beginning of the LTEE, a sample initially containing an identical genotype (the common ancestor) is grown in a food-rich environment up to reaching the carrying capacity. After this growth phase, the population is subjected to a bottleneck: a small fraction of the population is randomly sampled and then introduced into an identical environment to form the next founder population. This procedure is repeated over many generations. A sample of the ancestor as well as any other individual at any generation can be frozen and stored. Adaptation can be quantified by measuring changes in fitness. With microorganisms, fitness can be measured using "head-to-head" competition between frozen fossils. As they compete for a pool of resources, their population growth rates are measured (27).

Figure 7 – Distinct protocols of experimental evolution to achieve: a) accumulation of mutations through single-individuals bottlenecks; b) adaptation of a constant population size; and c) populations adapting under regular, periodic bottlenecks.



Source: Image from (5).

In the years that followed, a proper formalization of these protocols was built to infer different parameters of the evolving population, such as the rate at which new genetic mutations spontaneously occur from parent to offspring (Figure 7.a). In many of them, the serial passaging with periodic bottlenecks regime is an inherent aspect.

The effects of a bottleneck in an evolving population may vary. Given its random sampling nature, bottlenecks are directly related to the decrease in genotype diversity, but while this imposes a risk to the perpetuation of the surviving population, it can also increase the potential to explore new domains of the fitness landscape by increasing the strength of genetic drift and thereby avoiding the population to be stuck in local fitness maxima of the landscape (3).

In experimental evolution, these regular, periodic bottlenecks occur with periods of sustained growth phase which enhances the probability of new beneficial mutations rise and increase in frequency. Straightforwardly, the serial-transfer protocol is a method to induce and register adaptation at shorter time scales.

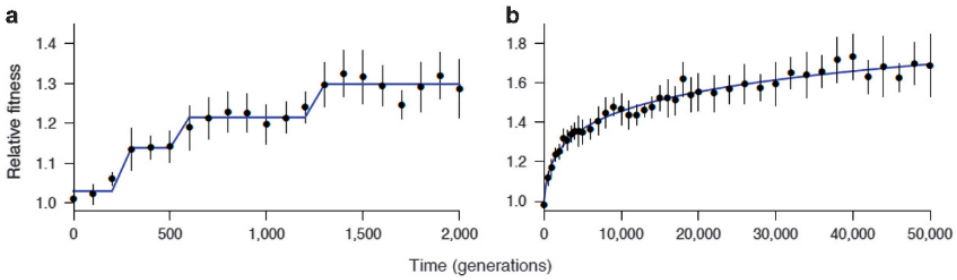
However, the description above leaves aside many non-trivial aspects of the adaptive dynamics; there are many chance events whose consequences depend on the time and the genetic background they occur. For instance, within the real lab resource and space limitations, there is no guarantee that beneficial mutations will be dominating the final population size at the end of the growth phases, which raises the probability that bottlenecks could in fact eliminate those beneficial mutations during the random sampling procedure, thus decreasing the adaptation rate. Taking these facts into account, Lindi Wahl et al. (30) investigated the likelihood that single beneficial mutations are lost due to bottleneck procedures. They inferred that, although random sampling reduces the fixation probability[3], the sustained periods of exponential growth rate can compensate the fixation rates. Following this analyses, they searched for a dilution ratio - the ratio between final and initial population size - that could minimize the extinction chance, thus optimizing the adaption rate in serial passaging; and found it to be e^{-2}.

Another open question explored by evolution experiments is whether an evolving populations would continuously adapt to the invariant lab conditions, with an ever increase of mean fitness. In fact, it was one of the formal questions that motivated Lenski to build the LTEE (29). Wiser et al. (31) analyzed the mean fitness trajectories of many replicates of *E.coli* populations for more than 50,000 generations as showed in Figure 8. The rapid growth of fitness in the earlier stages of adaptation followed by an ever-decreasing rate of fitness increase is present in all replicated populations and agrees with the theoretical predictions on the magnitude of the contribution of individual mutations to fitness.

As stated by Lenski (27), "such dynamics indicate that populations, after being placed in a new environment, evolves from a region of low fitness towards an adaptive peak". The fact that fitness keeps increasing implies that further adaptation is being achieved by the populations, as they reach higher fitter states.

[3] The probability that a mutation spread and dominate a population

Figure 8 – Fitness trajectories of evolving *E. coli* in the LTEE. a) One of the 12 replicate populations, b) mean fitness of all 12 populations.



Source: Image from (28).

1.4 CONTINGENCY AND PREDICTABILITY

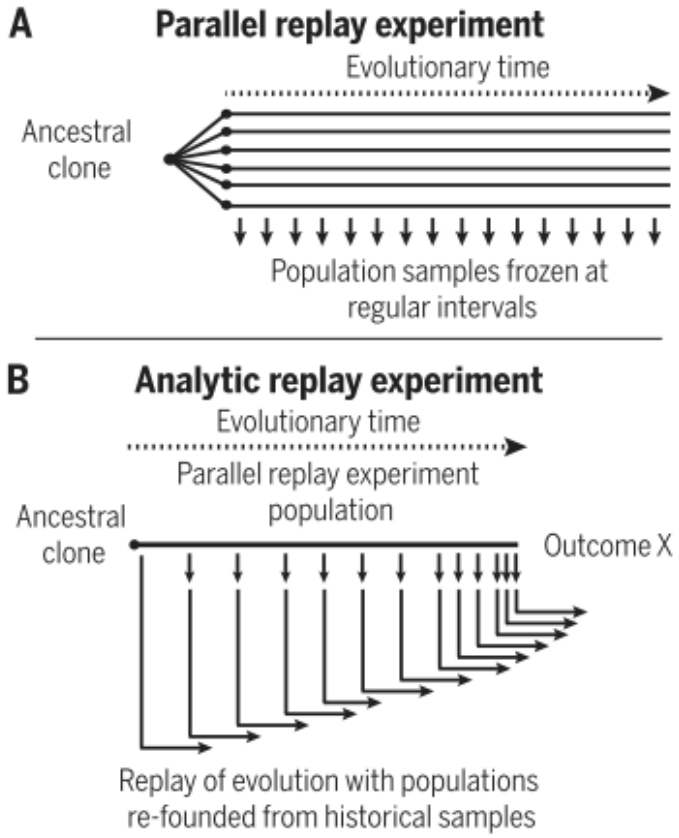
In the classic view of evolution, selection favors the perpetuation of the individuals most adapted to a particular environment, while the stochastic processes of mutation and genetic drift may alter the population's composition in a random manner (3, 32). Thus, populations that start from the same state and evolve under identical conditions might follow different evolutionary trajectories. The idea that random events could have lead populations to develop strikingly different characteristics from what they present today, is a theme of wonder.

Analogously, the same final outcome can be achieved through different paths. The occurrence of a given evolutionary path is greatly influenced by the ordering that mutations occur and their effects - which in its turn can be dependent on the genetic background at the moment they first arise (4, 9).

This susceptibility to the details of historical events is addressed by biologists through the concept of contingency, coined by Stephen Jay Gould in his analogy of "replaying life's tape" (2). This property has been identified as "intrinsic to path-dependent systems in which there are multiple possible paths from an initial state, multiple possible outcomes, and probabilistic causal dependence that links the two" (32).

Straightforwardly, the only feasible way to investigate this aspect of a system is through the replication of an experiment with initially identical populations and also evolving under identical conditions. While this does not represent a problem from a theoretical perspective, only recently this methodology has been implemented in microbial evolution experiments, where the well-controlled lab conditions and the possibility of frozen fossil records make feasible the production of replicates (see Figure 9).

Figure 9 – Designs of microbial evolution experiments to explore historical contingency in parallel replay experiments. a) Initially identical replicate populations are evolved under the same conditions to see whether evolution is parallel or divergent. b) Analytic replay experiments are used to assess the contingency of a given outcome observed in a parallel replay experiment by replaying the population's evolution from various points in its history to see whether the likelihood of that outcome changes over time.



Source: Image from (2).

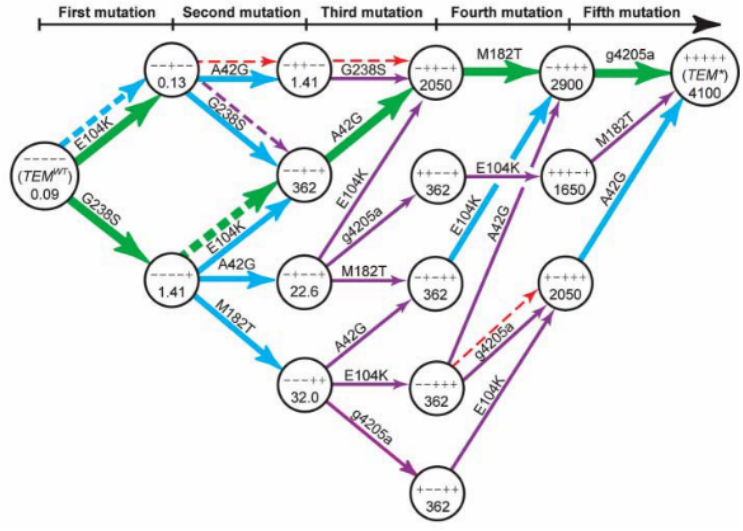
In a broader sense, contingency measurements need to be defined by specifying their particular observables. For instance, in cases where distinct genotypes can culminate in the same traits, the convergence to these traits will imply less contingency, but not at the genotypic level. For example, in the LTEE replicate experiment, all its 12 identical initial populations "have evolved much higher fitness, faster growth rate, and larger cells size than the ancestor" (2). On the other hand, each of them "has accumulated a unique set of mutations".

In the present work, we are solely concerned with surveying the contingency within the trajectories over a fitness landscape - in other words, we focus on the microscopic genotypic change of the system. For that purpose, we utilize statistical and computational measures to characterize aspects of contingency related to the repeatability and similarity of such paths (9).

A recent study on bacterial resistance to antibiotics has brought renewed interest in this path-dependent perspective. Weinreich et al. (4) examined the combination of 5 mutations in the β-lactamase enzyme of bacteria, which jointly increase the antibiotic resistance response. The accumulation of such mutations could only happen one at a time but in any particular order. Thus, from the ancestor mutant-free to the highest-resistance mutator there were $5! = 120$ distinct trajectories up to reaching the highest fitness peak. Here, fitness is defined as a proxy for antibiotic resistance.

By exhaustive analyses of all mutant combinations, they constructed a five-locus fitness landscape. It became clear that the fitness of a specific mutation was dependent on the presence (or absence) of the other four mutations: their contributions to antibiotic resistance could increase, decrease and even turn over negatively to that end - that is, becoming a deleterious mutation instead. Thus, the most notorious aspect of their research was the large number of distinct deleterious combinations from these 5 point mutations. From the 120 possible paths, only 18 were attainable through replicate experiments (Figure 10) while the remaining trajectories had negligible probabilities of realization.

Figure 10 – A subspace of the full fitness landscape comprising the accessible pathways from β-lactamase towards higher fitness states. The symbols (+) and (-) indicates the presence or absence of mutations at a specific locus. The number inside each circle is the fitness of their specific combination.



Source: Image from (4).

This mutational background dependence is referred to in the literature as epistasis (4, 33) and, more specifically, this reverse change between beneficial and deleterious as sign epistasis. As discussed in Section 2.5, epistasis has a direct relationship with the topography of the landscape and, consequently, with the number of accessible mutational pathways. As being

reported by many empirical studies, at the small genomic scales, sign epistasis is commonly observed (3).

In sum, the complex correlation between mutations - captured by its fitness landscape-mapping - imposes restraints on the contingency of genetic changes. "Evolution is more likely to be historically insensitive and repeatable if the adaptive landscape offers few alternative paths or many that lead to similar outcomes" (2).

2 CONCEPTS

This chapter focuses on the basic concepts of population genetic theory to understand qualitative and semi-quantitative aspects of an evolving population. Based on the mathematical framework developed in its early years, many unrealistic assumptions are adopted initially, and, as we progress, we move away from those idealized cases to more complex regimes where numerical treatment is adequate. The reader focused on the implications of this research, may go straight to the discussions of the last two subchapters.

We start by discussing simple stochastic processes utilized to model population growth and integrate the implementation of selection and mutation. We highlight how the combination of these mechanisms alters the regime by which the population evolves. In the end, we argue the implementation of gene architecture in the models and its importance to the development of a path-dependent perspective to study the evolutionary process.

Much of the analytical work exposed here follows the steps of J. Haldane, R. Fisher, and especially of John Gillespie, Motoo Kimura and James Crow, from whose textbooks (34, 35) contains many of the presented derivations. In comparison, the adoption of a fitness landscape and its accessibility implications goes back to S. Wright and Maynard Smith's conceptions.

2.1 POPULATION GROWTH AND DRIFT

In biology, a population is the summation of living organisms of the same group or species. By multiplying, individuals of the same species pass on their hereditary traits to their descendants. The first and simplest life forms to arise on Earth replicate themselves, and at some point in life's history, newly developed forms needed the combination of two parents to generate offspring. Even with a low life expectancy, a set of individuals will grow in number as long as food, space, and other resources are available.

Throughout this work, we focus on modeling a population of asexual haploids organisms, i.e., life forms that reproduce exclusively by replication. Although, in many models, recombination can be easily integrated. Furthermore, we are only interested in birth and death events, with no mention of life development between them. Such a choice is an excellent approximation to microbial organisms that reproduce much faster than individuals' life cycles. Thus, on

average, a population with a number of individuals that fluctuates around N, is said to have a size of N.

Take *E. coli*, for example, the most widely studied bacteria in the world. Living in the guts of every human being, it replicates to form two to six copies of itself over a specific time scale in ideal lab conditions (1).

Being N_t the number of individuals at a discrete-time t, and c the number of copies that it divides at each generation, we have by recursive relations

$$\begin{aligned} N_t &= cN_{t-1} \\ &= c^2 N_{t-2} = \dots \\ &= c^t N_0, \end{aligned} \tag{2.1}$$

where N_0 is the initial population's size at generation $t = 0$.

Given the rapid increase in numbers over very low time scales, Equation 2.1 can be approximated by a continuous growth version, where individuals grow with a rate r

$$\begin{aligned} \frac{dN(t)}{dt} &= rN(t) \\ \Rightarrow N(t) &= N(0)e^{rt}, \end{aligned} \tag{2.2}$$

thus, over this approximation, population's size $N(t)$ has a exponential growth.

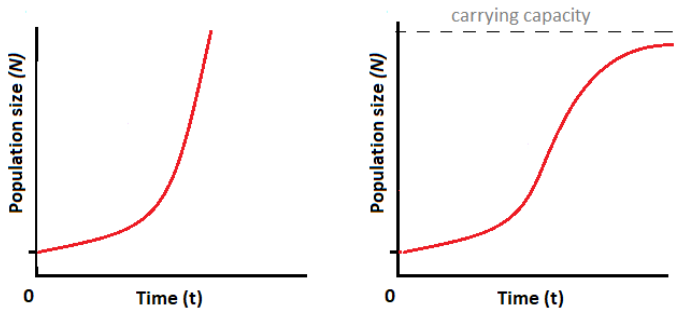
But of course, a population can not grow indefinitely. The limitations of space and resources impose an upper limit on the final population size, usually dubbed as the system's carrying capacity $\bar{K}$. Straightforwardly, one can implement such a limit simply by truncating an exponential growth as soon as it is reached. In a more realistic assumption, population growth slows down as it approaches $\bar{K}$. The simplest implementation of the latter case assumes a linear dependence on $\bar{K}$ and is known as the logistic equation,

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{\bar{K}}\right). \tag{2.3}$$

Note that while $N << \bar{K}$, it keeps an exponential growth rate.

As long as environmental conditions changes do not interfere with the carrying capacity, an established population may maintain their numbers invariant over long time scales.

Figure 11 – Simple distinction between exponential and logistic growth.



Source: The author (2020).

So far, these models express a growing population in pure absolute numbers. However, individuals among the same species can bear distinct traits that can ultimately affect their reproduction rate. To further model such evolutionary processes, references must be made to the heritage passed on to the next generation, i.e., a model that can trace what descendants came from what parents. Given our lack of knowledge about the real hereditary tree, we employ stochastic models to approximate this fundamental aspect of evolution by considering the next generations as random samples from the previous ones.

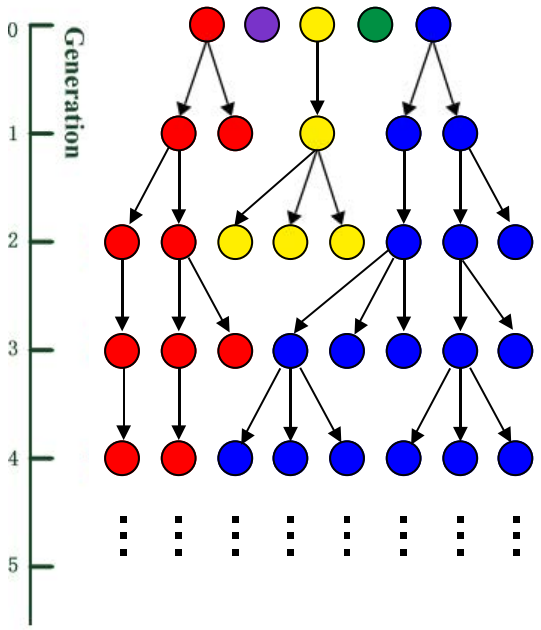
To take this next step, let us begin with the simplest and intuitive model, a stochastic branching process. It consists of a set of random variables $k = 0, 1, 2, \ldots$, where each of the parents in generation n produces a random number k of individuals in generation $n + 1$. Accordingly, in the simplest case, the random number k is obtained from a fixed probability distribution p_k.

First, we distinguish between individual traits by what we shall call for now as their lineage - schematically represented in Figure 12 as the colors. Second, we assume the simplest case where p_k does not vary from individual to individual. Thus, we are only interested in the reproduction's dynamics itself and how it affects the lineages' fate.

We may dive into these aspects by following the fate of a *single* lineage (color) while it evolves in a branching process much like the above. We follow the line of Haldane, and Fisher (35) by assuming that the population is already at its carrying capacity and that population size N is very large but countable.

The probability of extinction of a lineage can be derived as soon it appears in the population. Let $p_0, p_1, p_2, ..$ be the probabilities that the lineage will leave $0, 1, 2, ...$ descendants in the next generation, i.e. $\sum_k p_k = 1$. In particular, p_0 is the probability that given lineage will go extinct.

Figure 12 – Merely illustrative figure showing the perpetuation of a lineage.



Source: The author (2020)

Such assumption have a probability generating function (Probability Generating Function (PGF)):

$$f(x) = p_0 + p_1 x + p_2 x^2 + ... = \sum_{k=0}^{\infty} p_k x^k, \tag{2.4}$$

where the probability of leaving k mutant genes for the next generation is given as the coefficient of x^k.

The mean number of descendants, i.e. the mean of this distribution is given by $\frac{df}{dx}\Big|_{x=1} = f'(1)$, since

$$f'(x) = p_1 + 2p_2 x + 3p_3 x^2 + ... = \sum_{k=0}^{\infty} k p_k x^k$$

$$f'(1) = \sum_{k=0}^{\infty} k p_k =< k >= \mu.$$

Assuming that the subsequent offspring distribution is independent of the previous generations, in the next generation

$$f\Big(f(x)\Big) = \sum_{k=0}^{\infty} p_k (f(x))^k \tag{2.5}$$

and analogously, after n generations $(n = 0, 1, 2, ...)$

$$\underbrace{f(f(f(\dots f}_{n \ \ times}(x)\dots)) = f_{n-1}(f(x)) = f(f_{n-1}(x)) = f_n(x), \tag{2.6}$$

which one denotes by $f_n(x)$, with $f_1(x) = f(x)$ (35). Therefore, the statistics of the *n*th generation is the compound of its generating functions. Hence, by the chain rule, the average number of descendants at the nth generation is (15)

$$\begin{aligned}\mu_n &= f_n'(1) \\ &= f_{n-1}'\Big(f(1)\Big)\, f'(1) \\ &= f_{n-1}'(1)\, f'(1) \\ &= f_{n-1}'(1)\, \mu \\ &= f_{n-2}'(1)\, \mu^2 = \dots \\ &= \mu^n.\end{aligned}$$

So if each individual is expected to have more than one offspring, then the population will increase. If each individual is expected to have either one or no offspring, then the population will remain constant or decrease until eventually die out. On average, if the lineage leaves $\mu < 1$ descendants per generation it is bound to go extinct as n increases, in contrast to the case $\mu > 1$.

$$\begin{cases}\mu < 1 & \mu_n \to 0, \\ \mu = 1 & \mu_n \to 1, \\ \mu > 1 & \mu_n \to N.\end{cases}$$

Of course, in this large population limit, conclusions about $\mu > 1$ may be misleading. The lineage might have an appreciable probability of extinction, despite its replication rate. We refine these qualitative results by investigating the probability of the lineage being extinct *by* generation n (stress on the italic). Let

$$\begin{aligned}\theta_n &= \text{Prob (nth generation has no individuals)} \\ &= \text{Prob (extinction occurs } \textit{by} \text{ nth generation)} \\ &= f_n(0) \\ &= f_{n-1}(f(0)) \\ &= f(f...(0)...)) \\ &= f(f_{n-1}(0)) \\ \theta_n &= f(\theta_{n-1}).\end{aligned} \tag{2.7}$$

Since such an event could happen in any previous generation, we add the probabilities:

$$\text{Prob (extinct } \textit{by} \text{ nth generation)} = \text{Prob (extinct } \textit{by} \text{ (n - 1)th)} + \text{Prob (extinct } \textit{at} \text{ nth).}$$

$$\text{or } \theta_n = \theta_{n-1} + \text{Prob (extinct } \textit{at} \text{ nth)}$$

$$\Rightarrow \theta_n \geq \theta_{n-1}, \ \forall \text{ n.}$$

Assuming the non-trivial case where $\theta_0 = 0$, we have that

$$0 = \theta_0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq ... \leq 1, \tag{2.8}$$

thus θ_n is a non-decreasing sequence bounded by 1 (its a probability), hence, there exists a value θ such that θ_n converges to θ, as $n \to \infty$. We call θ the probability of ultimate extinction, and it is the solution of Equation 2.7 when taking this limit:

$$\begin{aligned} \lim_{n\to\infty} \theta_n &= \lim_{n\to\infty} f(\theta_{n-1}) \\ \theta &= f(\theta), \end{aligned} \tag{2.9}$$

with $\theta \in [0, 1]$. Remember that $f(\theta)$ is just the probability generating function in Equation 2.4,

$$f(\theta) = \sum_{k=0}^{\infty} p_k \theta^k. \tag{2.10}$$

It is easy to see that $f(1) = 1$ is a trivial solution. To investigate the existence of other roots, i.e., if there exists other points in which the $f(\theta)$ curve intersects Equation 2.9, we note that its derivative $f'(\theta)$ is strictly increasing for $0 < \theta < 1$ and $f''(\theta)$ is convex on that same interval; as shown in Figure 13. Thus, for $f(0) = p_0 > 0$, the existence of another solution depends on the slope of the curve at $\theta = 1$, which happens to be $f'(1) = \mu$, the average number of descendants.

Figure 13 – Depending on the slope at $\theta = 1$, the PGF $f(\theta)$ may have one additional solution at $\theta*$.

Source: The author (2020).

In the line of our previous discussion, we see that ultimate extinction is inevitable when the mean number of offspring is $\mu \leq 1$. On the other hand, despite the average increase in individuals' frequency with $\mu > 1$, there is still a probability $0 < \theta^* < 1$ of such lineage disappears over time. This result evinces the role of random genetic drift[1], a stochastic effect

[1] The "genetic" part is referent to our particular problem.

inherent to the sampling process. It is known that when there are few copies of a lineage, the effect of genetic drift is larger (34). We can calculate the variance of the process to investigate this observation mathematically. Still, sadly, the branching process variance is rather cumbersome and makes no explicit reference to the size N of the population.

As such, we use this opportunity to address another model of sampling. In the Wright-Fisher model (34), the descendants of generation $n+1$ are randomly sampled with replacement from the parents' generation n with some probability $p \in [0,1]$. In its non-overlapping generation version, all individuals reproduce and die simultaneously, i.e., once generation $n+1$ is obtained, it replaces the previous one. With a constant population size N, let $X = 0, 1, 2, ..., N$ be a random variable assigning the number of copies of one of the lineages present. The probability to sample a number $X = j$ in N independent trials follows a binomial distribution $X \sim Bin(N, p)$, where

$$Prob\left(X_{t+1} = j | X_t = i\right) = \frac{N!}{j!(N-j)!} p^j (1-p)^{N-j}, \tag{2.11}$$

which is known to have mean $E[X] = Np$ and variance $Var[X] = Np(1-p)$. Here, p is simply the fraction of the individuals present of a given lineage,

$$p = \frac{p_k}{\sum_k p_k} = \frac{i/N}{\frac{i}{N} + \frac{N-i}{N}} = \frac{i}{N}, \tag{2.12}$$

which makes the sample process equivalent to randomly picking balls from a box with replacement[2].

To investigate the stochasticity of the process it is essential to characterize the variability of X over time. To obtain a rough measure of the magnitude of the effects of random genetic drift, we describe the short-term fluctuations by quantifying the expected mean and variance of its frequency(34). Defining a generation transition as $\Delta t = \frac{1}{N}$, let $p' = 0, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N}, 1$, be the allowed frequencies of X. The mean and variance of a sample proportion of p' is given by

$$\begin{aligned} E[p'] &= E\left[\frac{X}{N}\right] = \frac{1}{N} E[X] = p \\ &\quad and \\ Var[p'] &= E\left[\left(\frac{X}{N}\right)^2\right] - \left(E\left[\frac{X}{N}\right]\right)^2 = \frac{1}{N^2} Var[X] = \frac{p(1-p)}{N}. \end{aligned} \tag{2.13}$$

[2] This case represents the absence of selection.

Here we obtain another example of random genetic drift. While the average lineage frequencies are invariant over generation times, the actual lineage frequencies change at a rate inversely proportional to population size. Thus, the Wright-Fisher model indicates that random drift is stronger in small populations. Furthermore, its dependency on the current frequency implies that rare lineages (very low frequency) have a higher chance of being extinct in subsequent steps.

The phenomenon of drift is not exclusive to these mathematical models. In real populations, random drift is inherently bound to the reproduction process, since chance has a role in determining whether a given individual survives and reproduces - which the stochastic nature of sampling and the distribution of offsprings tries to approximate. Random drift does not have a preferential direction, it can increase or decrease a frequency, leading a lineage to reach fixation[3] or being lost. In this sense, it is a mechanism that lowers the diversity present in a population.

Of course, random drift is only one of the many evolutionary mechanisms which can change a population's composition. As we shall see, the selective advantage between traits modify the probabilities of fixation and extinction of lineages, and mutation rates supply the population diversity over generations.

2.2 SELECTION AND SURVIVAL

In a given environment, lineages may carry traits that confer an advantage over the others. Let it be a higher growth rate, a resistance to some pathogen, or even the efficiency to assimilate resources. In a broader sense, despite the particularity of the causes, the ultimate consequence of a species favored by Natural Selection is its descendants' perpetuation. In other words, selection must increase the reproduction process's success and later survival of individual development. As such, one can relate the relative advantage among traits - or selective advantage - to the simple measure of reproductive success or fitness.

In the previous section, we called the carries of these distinct traits as lineages. As addressed in the discussion of a genotype-to-fitness mapping (Section 1.2), we assume that the genetic code is the sole responsible for expressing such traits. Therefore, from now on, we shall address the distinct lineages as distinct configurations of the genetic code - or genotypes.

Selection is implemented in any model through the adoption of a fitness distribution $F(s)$.

[3] Fixation is attainable when all or the majority of the population is comprised of a given lineage.

Assuming that the advantage is purely additive and does not change regarding individuals' frequency, in an isogenic population with fitness $F = 1$, a distinct genotype is said to have fitness $F_\sigma = 1 + s$. If $s > 1$, this genotype is favored by selection to increase its frequency.

However, since reproduction is a stochastic sample process, all genotypes are susceptible to random genetic drift. Even the fittest among them can be purged in subsequent generations - especially in small populations. So it is natural to ask the probability of a genotype with relative fitness $F_\sigma = 1 + s$ to escape random drift.

To understand how this advantageous effect acts on the probability of genotype survival let us return to the branching process approach of following the fate of a single copy (35). We now assume a particular form of distribution p_k for the number of descendants of this genotype. For large N, a simple but realistic assumption is that the values $p_0, p_1, p_2, \ldots$ follows a Poisson distribution such that the probability of leave k descendants, has mean $\lambda = 1 + s$, i.e.,

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{2.14}$$

The probability generating function, Equation 2.4, for this distribution is

$$\begin{aligned} f(x) &= \sum_{k=0}^{\infty} p_k x^k \\ &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} x^k \\ &= \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda} \\ &= e^{\lambda x} e^{-\lambda} \\ &= e^{\lambda(x-1)} \\ &= e^{(1+s)(x-1)}. \end{aligned} \tag{2.15}$$

From Equation 2.9, extinction occurs with probability θ^*, which are solutions of

$$f(\theta) = e^{(1+s)(\theta-1)} = \theta. \tag{2.16}$$

Defining its complementary by $\pi = 1-\theta$, we may obtain the probability of ultimate survival by solving the equation,

$$f(1-\pi) = e^{-(1+s)\pi} = 1 - \pi, \tag{2.17}$$

which is a transcendental equation for π. We then solve for the selective advantage s:

$$-(1+s)\pi = ln(1-\pi)$$
$$s = -1 - \frac{ln(1-\pi)}{\pi}.$$

Assuming that π is small, and hence s, we expand the natural log and neglect the terms of order $O(\pi^3)$ and higher, such that

$$\begin{aligned} s &= -1 - \frac{1}{\pi}\left[-\pi - \frac{\pi^2}{2} + O(\pi^3)\right] \\ \Rightarrow \pi &\approx 2s. \end{aligned} \tag{2.18}$$

Therefore, the probability of ultimate survival of an individual genotype is approximately equal to twice its selective advantage when the selective advantage s is small.

As we will see in the next section, with the assumptions made so far, when a new arising genotype survives drift, it rapidly grows until it reaches fixation. Therefore, to new arising genotypes, the probability of survival and the probability of fixation are interchangeable π in this simple scenario (36).

Based on the branching-process method, the above treatment, though simple and intuitively appealing, has assumptions that may not render the same approximation on other models. To investigate the generality of its results, work in deriving π for other models had been extensively done (35, 36). Motoo Kimura was the first to investigate such aspects of the Wright-Fisher model through a diffusion approximation. Its famous demonstration considers more generally the probability, $\pi(p,t)$, that a genotype becomes fixed in the population by the generation t, given that its frequency is p at $t=0$. When the population size is large enough, gene *frequency* change is treated as a continuous stochastic process in a good approximation.

As such, the Kolmogorov backward equation can be applied to the problem to describe how the probability $\pi(p,t)$ changes over time:

$$\frac{\partial \pi(p,t)}{\partial t} = \frac{V_{\delta p}}{2}\frac{\partial^2 \pi(p,t)}{\partial p^2} + M_{\delta p}\frac{\partial \pi(p,t)}{\partial p}, \tag{2.19}$$

where $M_{\delta p} = E[\delta p]$ and $V_{\delta p} = E[\delta p^2]$ are the diffusion coefficients of the rate of change in gene frequency per generation. Considering the limit to obtain the probability of ultimate survival, we have that

$$\lim_{t\to\infty} \pi(p,t) = \pi(p), \tag{2.20}$$

for which $\frac{\partial \pi}{\partial t} = 0$. And, hence, we need to find a solution $\pi(p)$ that satisfies the equation

$$\frac{V_{\delta p}}{2}\frac{d^2\pi(p)}{dp^2} + M_{\delta p}\frac{d\pi(p)}{dp} = 0, \tag{2.21}$$

with trivial boundary conditions

$$\pi(0) = 0 \quad \text{and} \quad \pi(1) = 1, \tag{2.22}$$

since p is a frequency. We can rewrite the above differential equation as

$$\frac{d}{dp}(log\frac{d\pi}{dp}) = -2\ \frac{M_{\delta p}}{V_{\delta p}}. \tag{2.23}$$

Therefore, solutions are bounded to the definitions of $M_{\delta p}$ and $V_{\delta p}$. To identify these appropriate diffusion variables, we need to establish δp, the change in the proportion over a time interval of length δt, as $\delta t \to 0$. To the Wright-Fisher population growth model, it is intuitive to take $\delta t = 1/N$, i.e., one generation.

Going back to the binomial process (Equation 2.11), let us now consider the weighted probability of success as

$$p_s = \frac{w_k p_k}{\sum_k w_k p_k} = \frac{(1+s)i}{(1+s)i + (N-i)}. \tag{2.24}$$

Thus, the number of offspring of the fittest lineage in the next generation is provided by $X \sim Bin(N, p_s)$. Note that when $s = 0$ we restore the neutral case in which the success probability is simply $p = \frac{i}{N}$ for $X = i$.

So, the $M_{\delta p}$ can be defined as the expected value of the change of gene proportion from $p = X/N$ to the next generation, in other words,

$$M_{\delta p} = E[\delta p] = \frac{1}{N}E[X_{t+1} - X_t] = \frac{1}{N}\Big(Np_{t+1} - Np_t\Big) \tag{2.25}$$

but, while $p_t = i/N$, realize that p_{t+1} is chosen from p_s, hence

$$\begin{aligned} Np_s - i &= \frac{N(1+s)i}{(1+s)i + (N-i)} - i \\ &= \frac{Ni + Nsi - Ni - si^2}{N + si} \\ &= \frac{Nsi - si^2}{N + si} \\ &= \frac{N^2 s(\frac{i}{N} - \frac{i^2}{N^2})}{N(1 + s\frac{i}{N})} \end{aligned} \tag{2.26}$$

substituting $p = \frac{i}{N}$, where $p \leq 1$, and considering the limit of s small, such that $sp \ll 1$, we obtain

$$M_{\delta p} = sp(1-p). \tag{2.27}$$

And, as showed by Kimura (35), since $E^2[\delta p]$ is negligible small, we have

$$E[\delta p^2] = Var[\delta p] + E^2[\delta p] \sim Var[\delta p], \tag{2.28}$$

where $Var[\delta p]$ is the same as in the no-selection case of Equation 2.13, thus

$$V_{\delta p} = \frac{p(1-p)}{N}. \tag{2.29}$$

Substituting these results in 2.23, and integrating from 0 to p, one has

$$\begin{aligned} \frac{d}{dp}(log\frac{d\pi}{dp}) &= -2Ns \\ log\frac{d\pi}{dp} &= -2Nsp + C_1 \\ \frac{d\pi}{dp} &= C_1\, e^{-2Nsp} \\ \pi(p) &= \frac{C_1}{2Ns}\left(1 - e^{-2Nsp}\right) + C_2. \end{aligned} \tag{2.30}$$

Utilizing the boundary conditions 2.22, we can solve the constants:

$$\begin{aligned} \pi(0) = 0 &\implies C_2 = 0 \\ \pi(1) = 1 &\implies C_1 = \frac{2Ns}{\left(1 - e^{-2Ns}\right)}. \end{aligned} \tag{2.31}$$

And finally, substituting the above result, the probability of surviving drift in terms of the population size and selection coefficient is

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}. \tag{2.32}$$

If a single genotype are present in the population, i.e., $p = \frac{1}{N}$, the probability of fixation is given by

$$\pi = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}. \tag{2.33}$$

As standard, by considering s small and using a Taylor expansion,

$$\pi = \frac{2s}{1 - e^{-2Ns}}. \tag{2.34}$$

And for large N, we recover the result $\pi \approx 2s$.

As emphasized in the text, each model relies on its assumptions to better represent real-life observations. As stressed by Kimura and Crow (35), if we assume a population which size fluctuates or changes periodically, the survival probability becomes $\pi \approx 2s\frac{N_e}{N}$, where N_e corresponds to some definition of effective population size. An extensive revision made by Wahl and Patwa (36), summarizes many different approaches to determine the probability of survival of a genotype (or to escape drift) as a function of its selective advantage. Strikingly, all the models' predictions point to a linear dependence on $\pi \approx cs$, differing only by a scalar magnitude $c \in R$. This semi-quantitative result implies that, despite the distinction in the approaches, the analyses are valid for a large class of exchangeable models, as they converge to expected general results. Thus, even if we can not make direct inferences about the real microscopic behavior, the probabilistic framework underlying the assumptions made so far obtains good approximations.

Lastly, we would like to address the interplay between selection and drift. If there is no selective advantage $s = 0$, the probability that a single genotype reaches fixation is equal to $\pi = \frac{1}{N}$. So, there should be a threshold for fitness values, small enough such that

$$\pi = 2s \approx \frac{1}{N}, \tag{2.35}$$

Therefore, for a fixed s, drift dominates in very small populations, where $Ns \ll 1$, despite selection. On the other hand, in the $Ns \gg 1$ limit, corresponding to large populations sizes, selection plays a role by increasing the frequency of the fitter genotypes (37). Given enough time, the fitter individual will rapidly sweep through the population until fixation. These selective sweeps are discussed in the next section, but we highlight its contribution to the population's mean fitness for now.

Adaptation by natural selection occurs through the spread and substitution of new genotypes that improve the performance of an organism and its reproductive success in its environment.

So, as long as the population is supplied with new genotypes, selection can act on this diversity to increase the frequency of the fitter ones and, ultimately, drive the whole population mean fitness to higher values. This mean fitness increase in populations is synonymous of its adaptation to the pressures of environmental conditions (3, 27, 28, 5).

2.3 MUTATION AND ADAPTATION REGIME

A mutation refers to the change of the discrete hereditary information in the macromolecules of DNA or RNA. Although it can happen along with the organism's development or maturation through external factors such as wavelength radiation incidence and cancer, it is also an inherent aspect of the gene transfer process in life reproduction from parents to their offspring.

During reproduction, there exists a chance U for the descendant being born with a different genetic code [4] from their parents. Hence, considering a population of size N, NU is the influx of new mutants in the evolving population at each generation, providing a genotypic diversity upon which selection can act. Newly arising mutations (*de novo* mutations) are regarded as beneficial or deleterious depending on their contribution to the population's overall fitness (37). As we already have seen, selection can increase or decrease these genotypes' frequency depending on their fitness. Given sufficient time, beneficial mutations can spread through the population until it eventually fixes. A mutation is said to fix when the majority of the organisms inherit such hereditary genes.

While drift and selection decrease genetic diversity, mutation is a fundamental source of variation. To understand the effect of these processes on the composition of the population, let us study rare mutations' fate. By rare, we mean that a new single genotype arises, through the mutation process, in an isogenic population. This case recovers our previous sections' scenario where only coexist two distinct genotypes, one of which is a single copy.

Assuming that mutation rates are constant in time, we ask the waiting time for a *de novo* beneficial mutation with relative fitness $1 + s$ to rise and fix.

Over the generations, mutations arise at rate NU but can be lost through genetic drift (see Figure 14), the probability they survive drift is $\sim s$ (as worked in Section 2.2). Thus, on average, a mutation takes a time $t = \frac{1}{NU}$ to arise, and a time

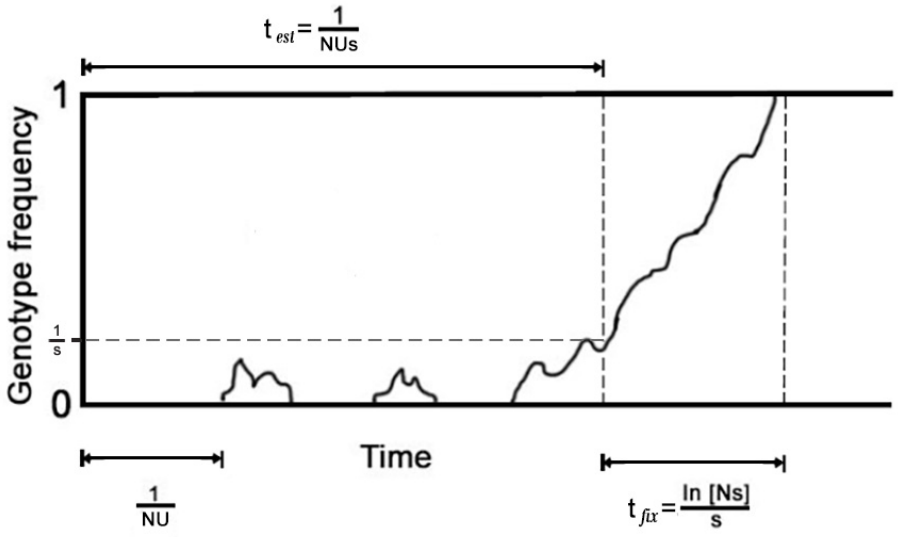
$$t_{est} \sim \frac{1}{NUs}, \tag{2.36}$$

to get established within the population. The establishment refers to a threshold frequency needed for a genotype to no longer be extinct by drift. It does not have a priori definition, is an *ad hoc* threshold imposed by an arbitrary confidence interval. Upon our assumptions from

[4] In principle, not so different, since simultaneous mutations have lower probabilities of occurrence.

previous sections, the mutant lineage must reach a size $n \sim 1/s$ before it becomes "safe" from extinction and begins to grow mostly deterministically (37).

Figure 14 – Mutations arise at short time scales but are rapidly purged by genetic drift. With a probability proportional to their fitness s, mutations can reach a threshold frequency to survive drift and, once done, they rapidly reach fixation at time t_{fix}.



Source: The author (2020).

Once established, i.e. once its reachs a size $1/s$,it is known to rapidly grows in frequency until it fixes. An exponential growth usually approximates such an event,

$$n(t) = \frac{1}{s}e^{st}, \quad \text{where } n(t=0) = 1/s. \tag{2.37}$$

Therefore, from the establishment, the time it takes to reach fixation, i.e. $n(t) \approx N$, from this point onward is given by

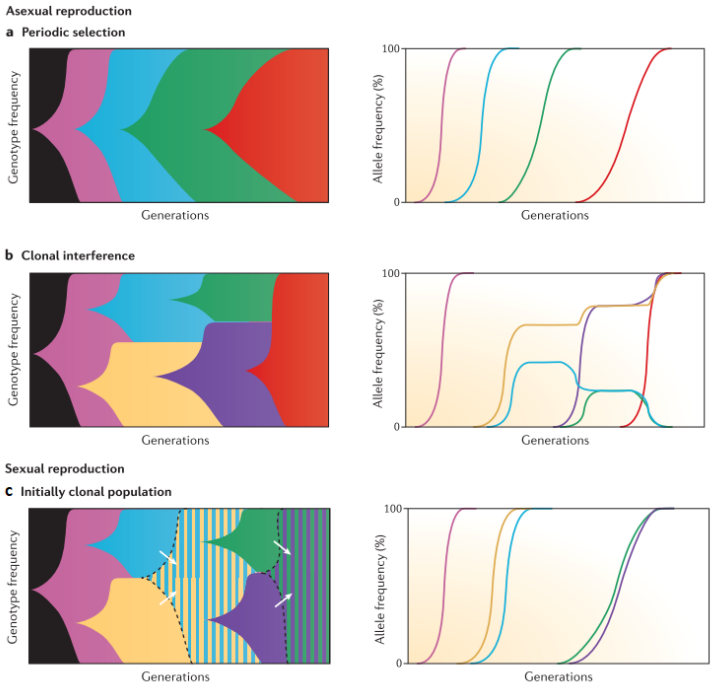
$$t_{fix} \sim \frac{\ln[Ns]}{s}. \tag{2.38}$$

These results help us understanding the rate at which the population evolves. It is a semi-quantitative approximation that does not depend on the microscopic details of the processes - which we can not access. The relation between these waiting times is used to delineate adaptive regimes (3, 33, 37).

In the simplest case, when mutations are rare $U \ll 1$, it follows that

$$\begin{aligned} t_{est} &\gg t_{fix} \\ \frac{1}{NUs} &\gg \frac{\ln[Ns]}{s} \\ NU &\ll \frac{1}{\ln[Ns]}. \end{aligned} \tag{2.39}$$

Figure 15 – Representation of the discussed regimes in asexual and sexual populations. **a)** In SSWM, the population changes its whole genotype configuration, one at a time. **b)** If *de novo* mutations establishes before the previous fix, they compete for fixation. **c)** From clonal interference, we see how recombination decreases the competing aspect by assimilating both genotypes in the same descendant.



Source: Modified from (5).

As long as the last inequality holds, *de novo* beneficial mutations will fix in the population before a new one arises and establishes. In this strong-selection-weak-mutation (Strong-Selection Weak-Mutation (SSWM)) regime, the population can be seen as an isogenic entity whose dynamics can be visualized as transitions from one genetic configuration to another; analogously to a random walker restricted to move over states of fitness increase. This rapid and sole substitution is called a selective sweep, as represented in Figure 15a. Since deleterious mutations are purged, and there is effectively only one beneficial mutation present at a time, in the SSWM regime the population mean fitness always increases and fast.

However, "genetic dynamics in evolution experiments rarely seem to be in this simple regime" (5). Typically, before a mutation can sweep to fixation, *de novo* beneficial mutations arise in a different lineage and become established, leading to a competition between each other (Figure 15b). As equation 2.39 suggests, this effect is pronounced in large populations or at higher mutation rates. In asexual populations, this competition between these multiple mutations slows down their fixation rates, since mutations must now "displace fitter competitors rather than only its less fit ancestor" (5). This phenomenon is called clonal interference

(Clonal Interference (CI)), common in asexual populations where there is no recombination. Recombination is the process of combining the genes from the parents to the emerging offspring. By doing so, previously competing mutations can coexist in the offspring, leading to an increase in fitness (Figure 15c). It is one of the factors that make sex an advantageous process (38).

One attempt to overcome these complex but more common scenarios is to integrate genomic architecture aspects into the models.

The above solution's principal motivation is that given the genetic composition of a population at time t, we can make an explicit reference to the arising genotypes accessible through mutational events at a time $t + 1$. This multilocus approach allows us to look at evolutionary dynamics as changes in the population's genetic configurations.

2.4 GENE AND SEQUENCE SPACE

In this section, we pass to integrate the genomic architecture into the previous models. In sum, it is possible to map a genetic code in a sequence space upon which a population, with a given fitness distribution $F(s)$, evolves. The genotypic map concept was born to build a mathematical framework that underlies the evolutionary mechanisms and the fundamental structure encoding the traits - the genome (7).

The genome itself is a collection of "sites"(loci) which codes the information to be read by sequence reading proteins that, ultimately, express itself through the phenotypic traits (8). As discussed in Section 1.2, this relation is non-trivial to more complex organisms in nature but can be a good approximation for simple lifeforms in a laboratory environment.

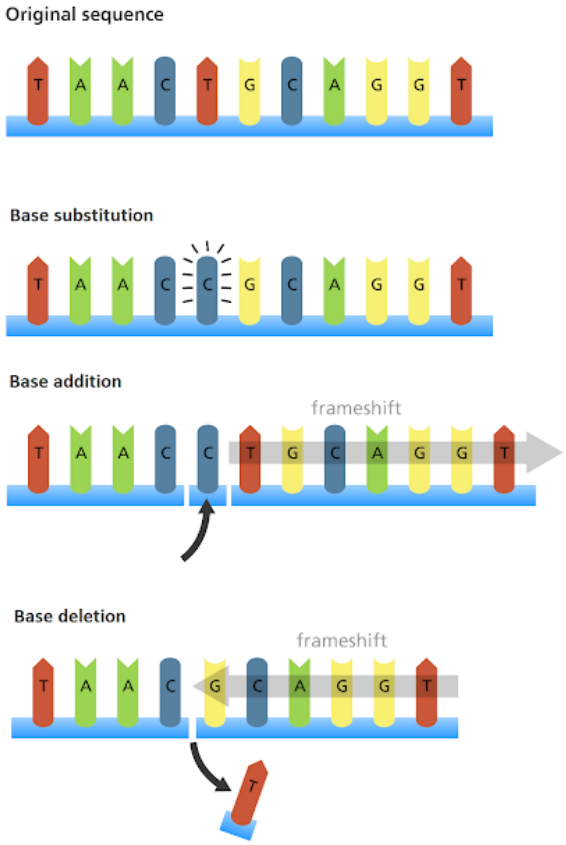
In the molecular context, by modeling RNA Polymerase, each locus can be filled with one of the four possible nucleotides $A = \{A, T, G, U\}$. In a given genome with a length of L loci, there are 4^L unique combinations of this four-letter alphabet. In the case of a protein, the L elements correspond to amino acids in the primary sequence, and so there are 20 possible amino acid states to be considered. A binary alphabet $A = \{0, 1\}$ can be used as well, merely to indicate if a mutation is present or absent in the original sequence (24).

Mutations are, therefore, a change in such a code.

In most models, mutations are copying errors in the form of point mutations (base sub-

[5] Available in: <http://biology4alevel.blogspot.com/2016/06/133-genetic-mutations.html>.

Figure 16 – Visualization of a genome structure as a chain of nucleotides, and common types of mutational process.



Source: Public image accessed in 2020. [5]

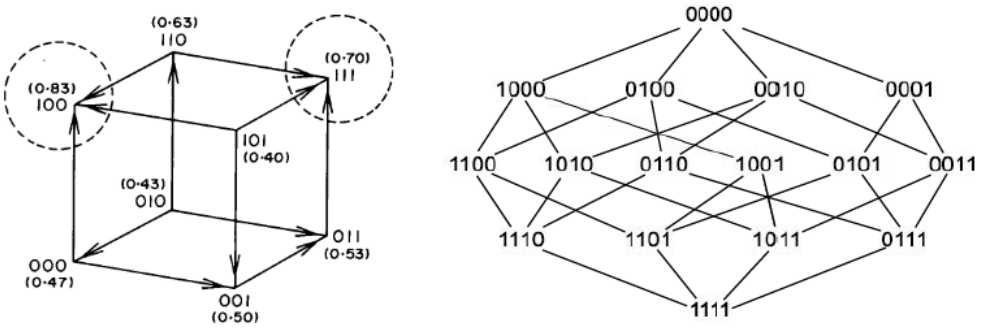
stitution in Figure 16). Although other forms of modification of genetic information are well known (such as insertion and deletion), the heuristic arguments for their negligible occurrence relies on the mechanism of self-correction present in the code transcription process (8).

Thus, mutations are modeled as 1-single letter transitions between sequence states. This assumption is also based on Wright, Fisher, and Maynard's arguments about the notion of nearness between elements of a genotypic space (Section 1.2), and it starts to gives us a notion of distance and neighborhood.

Many studies, including the present one, restrict their scope to a binary alphabet because it is more easily simulated and sometimes allows analytical treatment. In many instances, they are also mapped by many combinatorial problems (25). This gives origin to a binary space connected through single digits flip - or a L-dimensional hypercube H_L^2.

This space is equipped with a Hamming metric d, and because of that, it is usually called a Hamming graph. One measures distance in units of single flips to go from one sequence to another. Therefore, a given sequence has L immediate neighbors, i.e., states that can be

Figure 17 – Hypercube representation for L=3 and L=4. Only in the first case, fitness is assigned to each vertex in parenthesis. Arrows heads represent transitions to fitter states, while the circles outline a maximum state.



Source: Modified from (23, 39)

reached through a single flip, $d = 1$. Likewise, the longest possible distance is a chain of single transitions until all the digits of a sequence have been flipped, thus $d = L$. This space is characterized by short distances and high dimensionality (1).

In order to quantify the reproductive value of a certain genotype σ, a fitness value $F(\sigma) \in R$ is assigned to each sequence following a given fitness distribution $F(s)$. This final mapping is called a fitness landscape (23).

For consistency, let us visit again the canonical case of a *de novo* beneficial mutation σ_g arising in an isogenic population. Let the genotype of the population be, for example, $\sigma = \{000\}$. With the restriction of single-step transitions new mutations can only emerge at the next generation with one of the following configurations $\sigma_g = \{001\}, \{010\}$ or $\{100\}$. In either choice, change is occurring at a single locus.

In the SSWM regime, natural selection will increase fitter individuals' frequency by increasing the average number of offspring per generation. Given enough time, the whole population will jump to the genetic state σ_g. Therefore a population undergoing adaptation propagates through the space of genotypes along a monotonic path of increasing fitness. The process stops when a fitness peak is reached. As conceived by S. Wright (7, 19), one visualizes evolution as trajectories over the fitness landscape and adaptation as a search for fitness peaks.

A definition of a trajectory on the CI regime is more troublesome: the population is now better described as a cloud around a given sequence moving on the landscape, but in some cases can divide itself into subpopulations, each of them describing their particular trajectories, which can culminate in the same or even different fitness peaks.

The existence of multiple peaks opens a new discussion about the landscape topology and

is intrinsically related to multilocus systems analyses. Many experimental lines (4, 3, 5, 40) have observed that the selective effect contributions to the states of a specific locus can change regarding the states of other loci. This degree of correlation between a mutation and the genetic background at which they appear is known as epistasis. As we shall discuss in the next section, epistasis is responsible for determining landscape topography, which, ultimately, affects the evolutionary trajectories.

Lastly, let us discuss some implications of using this model to approximate empirical observations. First of all, not all the loci on a genome results in an expression: some are responsible for the regulatory activity of reading-and-coding, catalytic tasks in the cells, and some others are regarded as inactive - many of those being remnants of some ancestor (8). A complete gene has several loci, typically of the order of $\sim 10^4$ in simple prokaryotes, reaching about $\sim 10^9$ in humans - where, in the latter, only ~ 3 percent code for proteins (1). Despite that, even if we could neglect the 'irrelevant' ones, making a full analysis of all loci combinations is unfeasible.

As stated by Drossel (8), "because the structure of the full fitness landscape is unknown and complex beyond any modeling capability, toy landscapes are introduced that may hopefully reflect some features of the real landscape."

We utilize one of these models, the Kauffman's NK model, which conveniently integrates epistasis effects. That being said, what is usually considered when short sequences (e.g. $8 \leq L \leq 16$) are studied, is that we are making a local but complete analysis of the real sequence space (3). In this focal set of loci, variation is attainable in a typical time scale and mutations share some effect on the population's fitness.

2.5 EPISTASIS AND ACCESSIBILITY OF PATHS

With the presentation of the previous contents, the dynamic of genetic variation can be seen as directional state transitions over a configuration space; or adaptive trajectories over a fitness landscape.

In general, the fitness function $F(\sigma)$, attributed to a genotype σ, can not be decomposed into a sum of *independent contributions* from each locus L (3, 24). Therefore, the fitness effect of a mutation at a given locus may depend on the mutations at other loci - in other words, it may depend on the genetic background. At the genotypic level, epistasis is defined

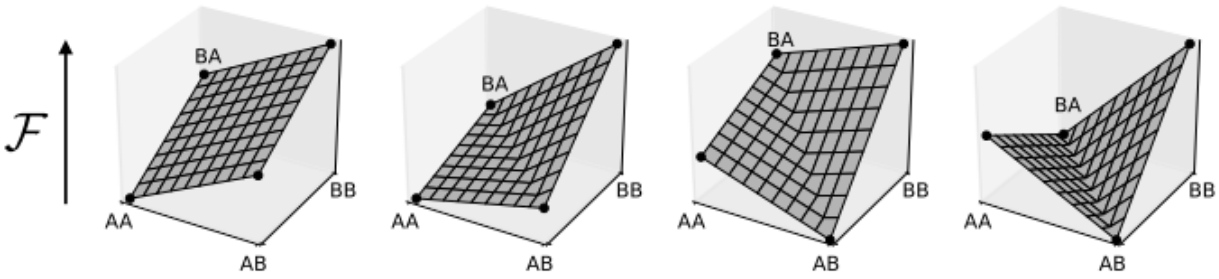
as this interactive coupling of sequences (33).

In many analytical models, $F(s)$ assumes the form of a typical probability distribution of available mutations (37)[6]. In these models, no mention is made about the mutational proximity between mutations as well as their fitness effects correlation. Therefore, many of the analytical models do not integrate epistasis effects.

At the genotypic level, epistasis is characterized according to the differences in the combined fitness effects of mutations. We follow the intuitive and pedagogic approach of Manhart et al. (33), by analyzing the four possible types using a two-letter, two-locus model.

In Figure 18, the sequence AA evolves towards the fittest sequence BB along single-letter transitions. In the first case 18a, the fitness effect of a substitution at locus x_2 is the same regardless of the state of locus x_1, and vice versa. Thus the fitness of each sequence can be decomposed into a sum of additive contributions from each locus: $F(\sigma) = F(x_1) + F(x_2)$, i.e., there is no epistasis.

Figure 18 – The simplest representation of genetic epistasis. From left to right: a) no epistasis, b) magnitude epistasis, c) sign epistasis and d) reciprocal sign epistasis. The fitness of each loci combination is given by its height.



Source: Image from (33).

In the case of 18b, the fitness effect at locus x_2 differs in magnitude but not in sign depending on whether locus x_1 has A or B, thus both the combinations remain beneficial. This case is known as magnitude epistasis (3, 33). With magnitude or no epistasis, a landscape has a single sequence state comprised of the optimum combination between its loci's states. Thus, generating a single-peaked fitness landscape.

The next case of Figure 18c shows how the substitution at locus x_2 can have opposite effects on fitness depending on the state of locus x_1: it is deleterious if A, but beneficial if B. Since this interaction causes a reverse change on the effects of mutations on fitness, it

[6] Usually a different distribution is used for beneficial and deleterious mutations.

is coined as sign epistasis. If sign epistasis is predominant at multiple loci, it is known as reciprocal sign epistasis 18d. As many studies have addressed, reciprocal sign epistasis is a necessary condition for the existence of multiple fitness maxima (4, 33, 40).

To investigate the effects of contingency on the overall dynamics of a population evolving in a fitness landscape, many theoretical works utilize the adaptive walker of an SSWM regime to characterize the accessibility of paths as a function of the degree of epistasis (3, 9, 41). As discussed in the previous section, under the walker's move-rule, transitions happen at random but only between states of ever-increasing fitness $F(\sigma_i) < F(\sigma_{i+1})$, such paths have therefore been termed selectively accessible.

This restriction is unaltered for transitions occurring between sequence states correlated through magnitude or no epistasis. On the other hand, sequence states correlated through sign or reciprocal-sign epistasis imposes further restraints on its movements by reducing the number of accessible paths - which in turn increases its determinism.

However, when analyzing landscapes with more loci, other aspects regarding the contingency of genotypic change emerge, since the number of peaks in a rugged landscape increases with the number of loci (3, 23, 33).

While in the first case a smooth single-peaked landscape has only one maximum to be reached, multiple peaks emerge from the latter case. Given its inability to move 'downhill', adaptive walkers suffer from a suboptimal search of higher peaks, getting trapped in the local maxima solutions. "The higher the number of local optima in the fitness landscape the smaller the probability that populations with the same starting genotypes will reach the same optimum" (40). Hence, in the SSWM regime, contingency is an interplay between the restriction to the accessible paths and the abundance of final outcomes.

Those move-rule constraints are relaxed for populations under the CI regime. Following the statements above, this might imply that clonal interference necessarily reduces the determinism, given the dispersion of its individuals among the available states. Despite that, determinism can be enhanced through the effects of two factors (3): that individuals reaching lower fitness states have a higher probability of leaving fewer to none descendants, making the probability of transition between these states more negligible; and that among the available fitter states, competition between multiple mutations might promote the fittest ones if given enough time.

In this sense, we can see the SSWM regime as a limit case when these probabilities are

further enhanced. Allied with its walker-like statistics, this regime is usually adopted to investigate the structures of the landscape itself (23, 25), which is independent of the particularities of a population's dynamics.

On the other hand, the non-trivial behavior of multiple mutations can only be investigated through numerical means. In this work, we characterize a pathway followed by the population on this regime and, through an ensemble of many independent runs, we infer statistical measures regarding the repeatability of the taken pathways and their degree of similarity (3, 9, 33).

We further note that no mention is made to the natural causes of epistasis. Following many lines of works, the focus is given only to its consequences.

3 METHODS

This chapter presents the adopted models and protocols that guide our work. We start with the NK model that generates the fitness landscape in which our dynamics occur, and the Wright-Fisher model that simulates a population evolving under a periodic bottleneck regime. Following, we expose the criteria to compare bottleneck ratios and adaptation rates; as well as measurements of Hill numbers to infer the impact of bottlenecks on the population genetic diversity. Lastly, concerning the path-dependent approach, it follows our definition of a trajectory over the fitness landscape, the protocol to generate an ensemble of them, and the statistical and computational methods used to analyze the contingency properties of such ensembles.

3.1 NK LANDSCAPE MODEL

In this section, we build our fitness landscape, a mapping from a set of genotypes into real fitness values; equipped with some notion of dimensionality, distance, and neighborhood.

Our analyses have concentrated on the NK model of random epistatic interactions, introduced by Kauffman et al. (23, 24), to study the influence of the landscape ruggedness on adaptive evolution. It belongs to a class of mathematical models applied to systems composed of many unitary parts. The state of the system is a function of each part's states - and some degree of correlation between them.

It bears close resemblance with spin-like models of disordered magnetic materials, common in solid-state physics, where each of the atom spins can assume one of two possible states $\{\pm 1\}$ (42). Likewise, both systems share the existence of frustration.

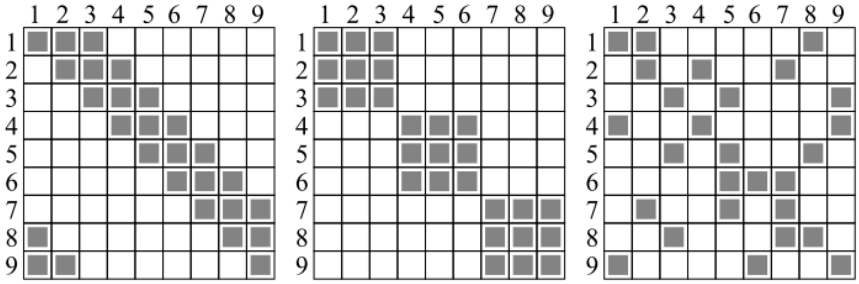
In our case, we consider a population of asexually reproducing haploid organisms. Each organism in the population is represented by L loci, or genes, where each of the loci can be in either of the two possible states, designated by 0 or 1. By definition, the fitness of an organism is the fitness of its genotype sequence σ, and calculated as (23)

$$F(\sigma) = \frac{1}{L}\sum_{i=1}^{L} f(x_i), \quad \text{where } \sigma = \{x_1, x_2, .., x_L\} \text{ and } x_i = \{0, 1\}. \tag{3.1}$$

Each locus' fitness contribution, $f(x_i)$, is drawn from a uniform probability distribution $U(0, 1]$, and the fitness of a genotype sequence is simply the average among the L loci, as shown above. Once one ascribes a fitness value to all 2^L possible genotypes, we obtain one realization of our discrete sequence space or fitness landscape.

However, one meaningful aspect of this model is that the correlation between the loci - and therefore, between fitness values - can be tuned through the parameter K. Going back to the disordered spin analogy, this would be equivalent to change the degree of frustration of the system. This is one of the strong motivations for its adoption in researches involving spin-glass like structures (42, 43). Hence, the fitness contribution of a specific locus $f(x_i)$ depends not only on its own state but also on the states of K other genes. There are different ways to implement this aspect of the model, called in the literature as its structure (25). Figure 19 summarizes some common choice.

Figure 19 – Illustration of genetic structure implementations to L=9 and K=2. From left to right: adjacent, block and random structures. By following a specific row, we represent the dependence between the row locus and the other loci as a grayed square. For example, in the first panel, locus number five is correlated with loci 5, 6 and 7. As expected, the diagonal of all structure choices is always grayed.



Source: Image from (25).

Without loss of generality, we choose the random neighbor structure. In the simplest case $K = 0$, the fitness contribution attributed to the locus x_i depends exclusively on its own state being 0 or 1 (resulting in a diagonal-only grayed visualization), thus we draw 2 real values $f(x_i)$, each of them for each state of locus i. In the example of Figure 19 (third panel), with $K = 2$, the fitness attributed to gene 4 depends not only on its own state but also on the states of genes 1 and 9. Thus, there are 2^3 distinct state combinations, and hence, 2^3 real values are drawn for each state of locus i. Analogously, in the $K = L - 1$ limit case, with a full grayed figure, a locus x_i has a range of 2^L possible fitness real values, since it also depends on the possible states of the $L - 1$ remaining genes.

In sum, for a given K, each realization of a landscape randomly generates a random

neighborhood structure (such as the example above) and, together with that, a real valued matrix F consisting of L numbering loci rows and 2^{K+1} columns, where each of its entry receives a fixed real value drawn from $U(0,1]$. These two 'ingredients' define the fitness contribution of each locus to a all genotypes σ.

To understand how the parameter K of the model is related to the epistasis, and ultimately to the ruggedness of the landscape, we can analyze the autocorrelation function of *fitness' effects* between states. Define

$$\rho(d) = autocorrelation\Big[\sigma_i(x), \sigma_{i+d}(x)\Big],$$

as the probability that the fitness contribution from a specific locus x remains unchanged after d neighbors. This can only happen if neither x nor its K coupled neighbors are chosen to mutate. Since each locus has an equal probability of being chosen, in the immediate neighborhood ($d = 1$), we have: $\rho\,(1) = 1 - \frac{K+1}{L}$, where $\frac{K+1}{L}$ is the probability that a mutation *does happen* in x or in any of its K neighbors. Thus,

$$\rho\,(1) = \begin{cases} 1 - \frac{1}{L} & K = 0, \\ 0 & K = L - 1, \\ \frac{L-K-1}{L} & \text{for arbitrary K.} \end{cases} \tag{3.2}$$

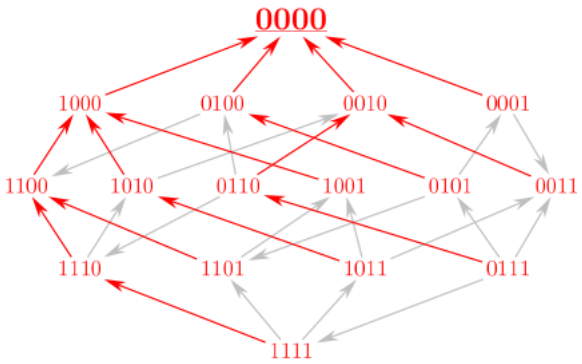
We can generalize for an arbitrary distance d, by assuming that the next mutation necessarily occurs at a different locus from the previous one. The chance that two sequential mutations do not affect the contribution of the same locus x is given by $(L-K-1)(L-K-2)/[L(L-1)]$ and, as demonstrated by Campos et al. (44), the autocorrelation function $\rho(d)$ can be obtained by induction, and written as:

$$\rho(d) = \frac{(L-K-1)!(L-d)!}{L!(L-K-d-1)!} \tag{3.3}$$

In other words, $\rho(d)$ measures how similar the fitness values of "d-mutant" variants are (23). In the highly correlated case ($K = 0$), each locus is independent of all other loci, and their contributions to fitness are simply additive. Hence, neighbors have similar fitness values, and fitness values smoothly vary as we move on to the landscape. Furthermore, there exists a single sequence 'comprised' of the fitter state of each locus. For example, if $f(0) > f(1)$ for every locus, there is a single optimum at the sequence $\sigma = all\,(0)$. Following the discussion

on Section 2.5, a sequence space map free of epistasis results in a smooth, additive landscape with a single peak, as in Figure 20.

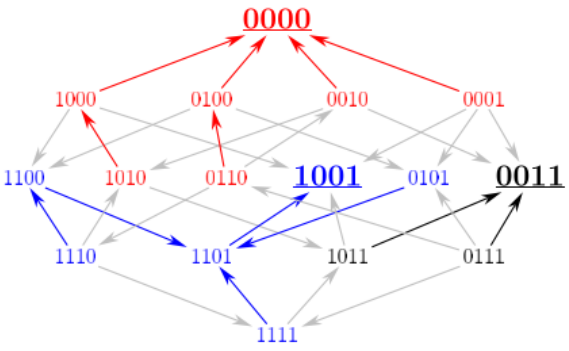
Figure 20 – Landscape's smoothness representation in the Hypercube. Maxima are enlarged and underlined. Arrows heads represent transitions to fitter states, while colors outline the basin of attraction of a peak.



Source: Modified from (39).

In contrast, for $K = L - 1$, each locus' fitness contribution depends on the remaining loci comprising sequence, therefore altering any locus' state alters each locus' fitness contribution to a new random value. Hence, fitness contributions are the most randomly possible, with each sequence having an independent, non-correlated value. This situation results in a full uncorrelated landscape, with $2^L/(L+1)$ local maxima, on average (Figure 21).

Figure 21 – Landscape's ruggedness representation in the Hypercube. Maxima are enlarged and underlined. Arrows heads represent transitions to fitter states, while colors outline the basin of attraction of a peak.



Source: Modified from (39).

For large values of L, this fully random case resumes to Derrida's random energy model of spin glasses, where the uniform distribution converges to a gaussian (42). Only in these limiting cases, $K = 0$ and $K = L - 1$, general aspects of the landscapes can be analytically derived, such as the number of peaks, or maximum fitness.

As argued before, along with many experimental lines (2, 4, 3), there is strong evidence of partially-correlated landscapes. Therefore, the range of $0 < K < L - 1$ is of substantial interest to this and further studies[1]. Since we aim to analyze general aspects of landscapes, this controlled correlation brings a strong motivation for adopting the NK-Model.

Furthermore, as demonstrated in (44), the fitness correlation function is independent of the chosen interaction scheme. And since it is a function of K and L only, it serves as a base to compare results from distinct combinations of the genome's length and epistasis.

3.2 WRIGHT-FISHER MODEL

To implement an evolving population, a replication process and an appropriate definition of generation must be chosen. From the vast pool of models serving those purposes - and to account for the evolutionary processes of random genetic drift, mutation, and selection - we use the Wright-Fisher model with discrete non-overlapping generations.

We consider a finite population of haploid individuals that grow exponentially for a period of length τ. During the growth phase, the population size $N(t)$ changes according to $N(t) = N(0)\, e^{rt}$ with $r = ln\, 2$ and $t = 0, 1, ..., \tau$. To account for selection, individuals at time $t < \tau$ contribute to form the next generation $t + 1$ with probability p_k proportional to their fitness (38). With n_k being the number of individuals of type (or equivalently, class) k, this probability equals

$$p_k = \frac{n_k f_k}{\sum_k n_k f_k} \tag{3.4}$$

where $\sum_k p_k = 1$, $\sum_k n_k = N(t)$, and f_k being the fitness assigned to genotype k. Hence, the composition of the population at time $t + 1$ is built from the population at time t, in a process of random sampling with replacement.

To implement this sample, for each trial, draw a number X from the uniform distribution $U(0, 1]$ and add the weighted probabilities p_k in any particular order until obtains the individual j which satisfies the condition

$$\left(\sum_{k=1}^{j} p_k \right) - X \geq 0 \quad with\, j = 1, .., k. \tag{3.5}$$

To account for mutation, the chosen individual to reproduce from each trial has a fixed probability of $1 - U$ to make an exact copy of itself or has a probability U to leave an

[1] In this sense, the asymptotic limits of $K = 0$ and $K = L - 1$ serves as a null model.

offspring with a different genetic state. In the latter case, since we are working with single-step mutations, the resulted state must be at a Hamming distance $d = 1$ from the former - therefore, it must be one of its L neighbors with an equal probability of $1/L$, i.e.,

$$Prob = \begin{cases} 1 - U & d = 0, \\ U & d = 1, \\ 0 & d > 1. \end{cases}$$

These processes are repeated a number of $N(t+1)$ times and, once done, generation $N(t)$ is replaced. Finally, at the end of the growth phase, the population of $N(\tau)$ individuals is subjected to a bottleneck protocol, which is simply a random sample of $N(0)$ individuals without replacement.

However, the aforementioned implementation has a high computational cost for large population sizes. Similarly, the process of random sampling with selection can be implemented more efficiently by using a multinomial distribution, such as,

$$\boldsymbol{N}(t+1) \sim Multinomial\left[N(t+1); p_1, ..., p_k\right]. \tag{3.6}$$

That way, the probability to sample a specific configuration $\boldsymbol{N} = \{n_1, n_2, \ldots, n_k\}$ is given by

$$P(n_1, n_2, .., n_k) = \frac{N(t+1)!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}, \tag{3.7}$$

here, $\sum_k n_k = N(t+1)$. Equivalently, the number of mutants within a given class k is taken from a binomial distribution

$$N_{U,k} = Binomial\left[n_k; U\right]. \tag{3.8}$$

3.3 SIMULATION PROTOCOLS

We simulate the evolution of haploid organisms as they undergo repeated cycles of growth and dilution. At the start of each simulation, the population consists of N_0 identical clones. The population goes through rounds of growth according to the adopted protocol, and each individual produces offspring depending on its fitness, $F(\sigma)$. After τ generation steps, a bottleneck, modeled as a random sampling, reduces the population size back to N_0, hence initiating

another round of exponential growth. This procedure is iterated until the desired number of generations is reached.

To capture the statistical properties of an evolving population visiting the states of the genotype space, our simulation was designed to ensure that populations of different parameter sets experienced the same fitness landscape. To determine the generality of our results, in all the proceedings below we sampled a number of 50 distinct randomly-drawn fitness landscapes, and *for each of them* we study ~ 1000 evolutionary trajectories for populations of different bottleneck sizes. This way, we neglect the implications of the particularities of a single landscape and drive our attention to the sensitivity of the model to its free parameters only: the sequence length L, the mutation transition rate U, the epistasis' parameter K, and the bottleneck dilution ratio $D = N_0/N_f$.

For the simulations regarding the adaptation rate, we mainly use a mutation rate of $U = 10^{-4}$, genome length $L = 8$ and epistasis $K = 2$. The last two parameters are increased only to investigate the sensitivity of the observed patterns to the chosen values.

For the simulations regarding the predictability of the trajectories, most results were obtained with a mutation rate of $U = 5 \times 10^{-2}$, genome's length $L = 8$ and epistasis $K = \{1, 2, 3, 4\}$. To further investigate and compare landscapes with increased ruggedness and length, we also utilized a varying combination of K and L such that the landscape's fitness correlation $\rho = 1 - (K + 1)/L$ remained constant - in particular we take $\rho = 0.5$ and 0.75. Additionally, regarding the predictability of the endpoints, lower mutation rates were considered, ranging from $U = 5 \times 10^{-3}$ to 5×10^{-5}.

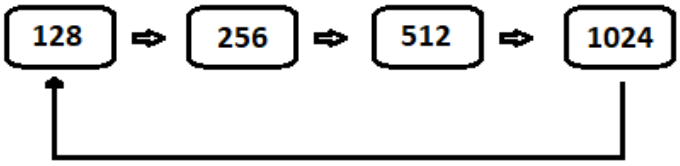
Following, we present the adopted bottleneck parameters; our definition of adaptation and its respective time units; and our criteria to build an adaptive trajectory.

3.3.1 Bottleneck parameters

In the Wright-Fisher model, we adopt a constant growth rate of $r = ln(2)$. Starting with a number of $N_0 \equiv N(0) = 2^n$ initial individuals with $n \in N$, the population doubles in size at each generation from $2^n \rightarrow 2^{n+1} \rightarrow 2^{n+2}$.. until it reaches a previously defined $N_f \equiv N(\tau)$.

By fixing N_f and varying N_0 we can investigate the effects of different serial-transfer regimes; ranging from a severe bottleneck with a long growth phase $N_0 << N_f$, to the opposite situation when $N_0 \approx N_f$. Similarly, we can maintain N_0 fixed and vary N_f to emphasize the effects of the growth phase's length.

Figure 22 – Serial-transfer example with $N(\tau) = 2^{10}$ and $N(0) = 2^7$



Source: The author (2020).

Over the simulations, the population's size can range from the order of some tens to 10^4. Looking at the interplay between drift and selection of Equation 2.35, we see that the populations subjected to more severe bottlenecks are more affected by random genetic drift than the contrary, at least at the early generations of their growth phase.

3.3.2 Measurements of Adaptation rate

When investigating the effects of bottleneck protocols on the adaptation rate, we must explicitly define the time unit of this rate. Here, we consider the rate at which the mean population fitness increases, that is, the rate at which *de novo* beneficial mutations occur, survive, and increase mean fitness. The mean fitness is measured immediately after each bottleneck event, establishing a standard bottleneck time scale $t_{bottleneck}$ regarding the numbers of bottlenecks. One may ask, despite that, if the observable rates differ if we consider the adaptation rate per replication event or per generation (doubling), in which these time units relate through

$$t_{generation} = \tau t_{bottleneck} \tag{3.9}$$

where τ is the number of discrete generations along the growth phase, and

$$t_{birth} = 2N_0(2^\tau - 1)t_{bottleneck} \tag{3.10}$$

accounts for the total number of new descendants at the end of the growth phase. Each of these rates may have practical relevance: if the population is resource-limited, a limited total number of new births may be possible, and thus the adaptation rate *per birth* might be the critical factor in an evolutionary rescue scenario; if we are concerned with environmental change that occurs at a pace set by calendar time, the adaptation rate *per generation* time may be relevant; if we consider an experimental population for which the bottleneck process itself is labor-intensive, *per bottleneck* will be the appropriate rate to compare across cases.

3.3.3 Ensemble of Trajectories

At the genotypic level, evolution is modeled by trajectories over the fitness landscape, defined as paths of connected walks through a succession of neighboring sequences. The Hamming distance $d(\sigma_i, \sigma_j)$ is the minimal number of mutations required to change the genotype from i to j. By starting with a population of identical genetic individuals at σ_0, the succession

$$\sigma_0 \rightarrow \sigma_1 \rightarrow \ \ldots \ \rightarrow \sigma_n$$

is called a path ϕ, if $d(\sigma_i, \sigma_{i+1}) = 1$ for all i.

As discussed in Section 2.3, under the SSWM[2] regime, the population is nearly analogous to a single walker transitioning over the sequence states - thus describing a monotonic path where $F(\sigma_{i+1}) > F(\sigma_i)$.

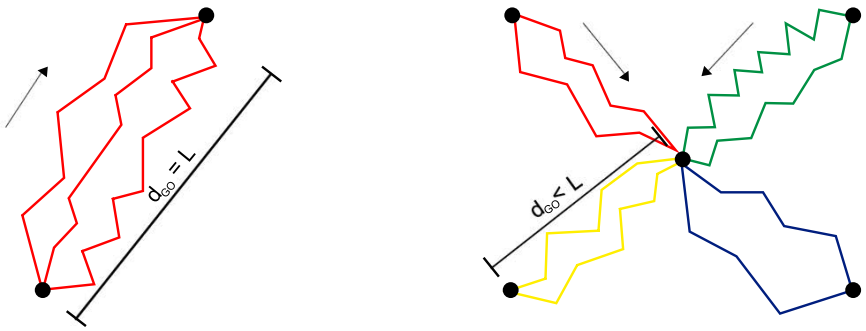
In contrast, for the parameters used in our simulations, our population is under the CI[3] regime, where multiple genotypes are present and competing for fixation. In this case, the population is better described as a cloud of mutants around a given sequence, and to characterize a pathway under this regime, we are concerned with the fittest genotype along the evolutionary trajectory. Therefore, when a fitter genotype appears, it is added to the path. Furthermore, before analysis, the trajectories are purged of loops, i.e., if a sequence appears more than once in an evolutionary pathway, which characterizes a loop structure, the pathway is redefined with the loop removed. This is to avoid situations in which a fitter genotype is generated and then lost (not increase in frequency), not genuinely characterizing a displacement of the population to a new fitter sequence.

The simulation starts with an identical population from an initial genotype σ_0, and it stops once the fitter genotype reach the global optimum (Global Optimum (GO)), thus $\sigma_n = \sigma_{GO}$. All the statistical measures utilized need fixed initial and ending points. For establishes the initial condition, we adopt two procedures based on its Hamming distance from the GO: $d(\sigma_0, \sigma_{GO}) = d_{GO} \leq L$. On the first one, an initial condition is naturally and uniquely attained for $d_{GO} = L$, since the further and only possible state from the global maximum is the antipode of σ_{GO}. After many independent trials, an ensemble of the trajectories is then considered to produce the statistical analysis.

However, the number of steps in the evolutionary pathway is variable and dictated by the dynamics itself, and for high values of K or L, it becomes computationally costly to keep this

[2] Strong-Selection Weak-Mutation
[3] Clonal Interference

Figure 23 – Representation of simulations starting from sequences at a length d_{GO}. a) For $d_{GO} = L$; and b) for $d_{GO} < L$. In the latter case, measures are obtained for the ensemble of each starting point, and then averaged.



Source: The author (2020).

approach. It also enhances the probability that populations get trapped in local maxima, thus never reaching the GO.

To handle this obstacle, we follow the protocol utilized by Szendro et al. (41) by setting a fixed Hamming distance $d_{GO} < L$ and choosing different initial conditions among the possible ones. We carry out independent evolutionary runs for each of the starting points and determine the statistical measures separately. Next, these results are averaged over the different starting points.

Both approaches are highlighted in the presentation of the results in Section 4.2.3.

3.4 DIVERSITY MEASURES

With the possibility of multiple mutations and the effects of bottlenecks on the population composition, it is fundamental to understand how bottleneck sizes affect the lineage genetic diversity. Several measures of genetic variation can be used to assess the amount of genetic variation among individuals within as well as between populations. The simplest and most commonly used are the genotypic richness (number of different genotypes in the population) n, heterozygosity H, and various related measures of entropy, such as the Shannon entropy (45). These measures can be unified through the use of Hill numbers, which capture essential properties of genetic diversity in a population (46).

The Hill diversity number of order a is defined as

$$D_a = \left(\sum_i p_i^a \right)^{\frac{1}{1-a}} \tag{3.11}$$

where p_i is the frequency of genotype i in the population. These diversity numbers can be understood as the weighted sum of each p_i to the power $a-1$, where the weights are themselves the p. We then take the $(a-1)$-th root of that sum, thus D_a is simply a *weighted* $(a-1)$-*Norm of the vector of genotype frequencies*. The Hill diversity numbers can be easily related to the three commonly used measures of diversity, and most importantly, they provide a unified understanding (and consistent units) across which the other measures of diversity can be compared:

- $D_0 =$ number of different genotypes $= n$;
- $D_1 = \exp(S)$, where S is the Shannon entropy[4];
- $D_2 = \frac{1}{1-H}$, where $H = 1 - \sum_i p_i^2$ is the heterozygosity.

Note that the D_a of any order measures diversity in units of genotypes, and can be interpreted as an effective number of genotypes or number of lineages in the population. While D_0 simply counts the number of distinct genotypes in the population, as the order a increases, the contribution of rare types to the corresponding diversity D_a is reduced.

3.5 STATISTICAL MEASURES

As mentioned, evolution studied under a path-dependent analysis allows statistical tools to explore aspects of contingency (2, 3, 9). Among them are the measures of Predictability P_2 and the Mean Path Divergence $\bar{D}$.

P_2 is a natural measure of the repeatability of a system of many possible outcomes. With the ensemble of trajectories at hand, let $p(\phi)$ be the weight or the frequency of path ϕ observed in independent trials. Thus, the probability of observing this path twice in two replicate runs is $p^2(\phi)$ and, hence, the sum over all the paths

$$P_2 = \sum_\phi p^2(\phi) \tag{3.12}$$

[4] For its less intuitivity, a demonstration is shown in Appendix B.

is the probability of observing any pathway twice in two replicate runs (3). This quantity is a simple metric for pathway repeatability that varies between

$$P_2 = \begin{cases} 1 & \text{for the occurrence of a single path,} \\ 1/n & \text{for n equally likely paths.} \end{cases}$$

Equivalent results are obtained for similar observables such as the entropy measure $S = -\sum_{\phi} p(\phi)\, ln\, p(\phi)$ (41) . Common to all of them, determinism is quantified about how often each outcome occurs.

With P_2 or its equivalent measures, paths that diverge by at least a single sequence are treated as distinct. To better refine the contingency's notion, it is possible to quantify the degree of similarity between the paths - or, analogously, of dissimilarity. The alternative idea proposed by Lobkovsky et al. (9), is that "many similar clustered paths represent a high degree of repeatability, whereas a small number of different paths signifies a lower one". We utilize a measure of divergence among accessible paths to quantify this diversity more precisely. One define the mean pathway divergence $\bar{D}$ as

$$\bar{D} = \sum_{\phi_1 \neq \phi_2} d(\phi_1, \phi_2)\, p(\phi_1) p(\phi_2), \tag{3.13}$$

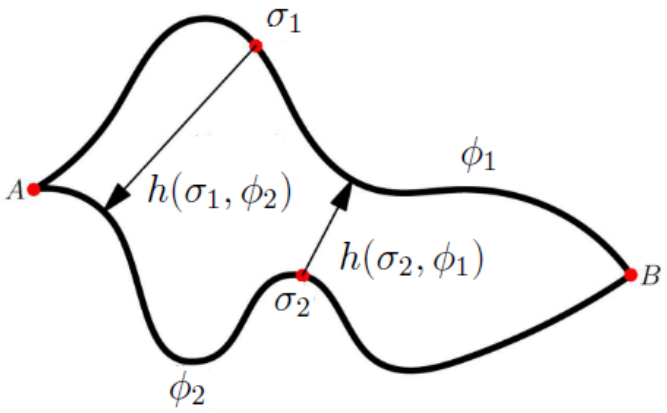
where the sum is over all pairs of distinct paths in an ensemble, $p(\phi)$ is the previously defined probability of path ϕ, and $d(\phi_1, \phi_2)$ is the distance, or divergence, between two paths ϕ_1 and ϕ_2. A natural definition of this divergence should account for the inner-path distance between trajectories, thus should be a function of the Hamming distance between all genotypes of pair of paths (33):

$$d(\phi_1, \phi_2) = \frac{1}{L(\phi_1) + L(\phi_2)} \left(\sum_{\sigma_1 \in \phi_1} h(\sigma_1, \phi_2) + \sum_{\sigma_2 \in \phi_2} h(\sigma_2, \phi_1) \right). \tag{3.14}$$

In definition above, $L(\phi)$ is the length (number of steps) of path ϕ, and $h(\sigma_1, \phi_2)$ is the shortest Hamming distance between a genotype σ_1 and all genotypes $\sigma_2 \in \phi_2$. To be clear, *for each* genotype σ_1 comprising pathway ϕ_1, one estimate its Hamming distance to *every* genotype $\sigma_2 \in \phi_2$. The lowest distance is then stored, and the process is repeated until all genotypes in ϕ_1 are rated. The process is then repeated in reverse, from ϕ_2 to ϕ_1 (see Figure 24). The divergence $d(\phi_1, \phi_2)$ is taken as the mean value of those shortest Hamming distances.

The mean path divergence $\bar{D}$ therefore captures not only how many paths are available, but weighs them by their spatial proximity.

Figure 24 – Representation of two distinct paths with same initial and final points. On the right, the Hamming distance is measured for each point $\sigma_1 \in \phi_1$ to every point $\sigma_2 \in \phi_2$, and vice-versa. The shortest measure is stored.



Source: Modified from (9).

The measures P_2 and S can be used to quantify the repeatability of "any replicate experiment in which the outcome belongs to a discrete set" (3), for example, the genotypic trajectory of evolution or its endpoints. While $\bar{D}$ is a refinement to compare trajectories in terms of similarity.

3.6 MULTIDIMENSIONAL SCALING

The Multidimensional Scaling (Multidimensional Scaling (MDS)) technique allows a lower-dimensional visualization of a higher dimensional data set. Our last method is completely computational and can be applied to many different kinds of data. Its central motivation is to map the information about the pairwise distances among a set of N objects, into a configuration of N points mapped into an abstract Cartesian space. This analysis can make explicit some patterns from the data set otherwise inaccessible in the high dimensional case.

The technique presumes a measure of dissimilarity (47). In Euclidean space, it could be the real length connecting two points. In statistical theory, it is often the covariance between two random variables. In our case, the divergence $d(\phi_x, \phi_y)$, defined as a function of the inner-path distance from Equation 3.14, has an appropriate definition of dissimilarity between trajectories.

These distances are the entry of the dissimilarity matrix:

$$\boldsymbol{\Delta} = \begin{bmatrix} d(\phi_1,\phi_1) & d(\phi_1,\phi_2) & ... & d(\phi_1,\phi_n) \\ d(\phi_2,\phi_1) & d(\phi_2,\phi_2) & ... & d(\phi_2,\phi_n) \\ \vdots & & \ddots & \vdots \\ d(\phi_n,\phi_1) & d(\phi_n,\phi_2) & ... & d(\phi_n,\phi_n) \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & ... & d_{1n} \\ d_{21} & d_{22} & ... & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & ... & d_{nn} \end{bmatrix} \tag{3.15}$$

where $d_{ij} = d_{ji}$ and $d_{ii} = 0$.

The goal of MDS is, given $\boldsymbol{\Delta}$, to find M vectors $x_1, x_2, .., x_M \in R$ such that

$$||x_i - x_j|| \approx d_{ij} \quad \text{for all i,j = 1, 2, .., M} \tag{3.16}$$

The algorithm for recovering coordinates from dissimilarities between pairs of points is as follows[5]:

1) Form the squared matrix of dissimilarities $\boldsymbol{\Delta}^2 = [d_{ij}]^2$;

2) Compute the matrix $\boldsymbol{B_\Delta} = \boldsymbol{J\Delta^2 J}$, where $\boldsymbol{J}$ is the centering matrix $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n^T}$, where $\mathbf{1_n}$ is a vector of ones.

3) Find the spectral decomposition of $\boldsymbol{B_\Delta}$, $\boldsymbol{B_\Delta} = \boldsymbol{Q\Lambda Q'}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix formed from the eigenvalues of $\boldsymbol{B_\Delta}$, and $\boldsymbol{Q}$ is the column of corresponding eigenvectors;

4) Find $\boldsymbol{X} = \boldsymbol{Q\Lambda^{\frac{1}{2}}}$; the coordinates of the points are given by the rows of $\boldsymbol{X}$.

The higher the value of a eigenvalue λ of $\boldsymbol{\Lambda}$, the better its eigenvector approximates Equation 3.16. By choosing a number $m < M$ of the highest ones, as long as the sum of the eigenvalues in $\boldsymbol{\Lambda_m}$ approaches the sum of all eigenvalues in $\boldsymbol{\Lambda_M}$, in other words, as long as

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{M} |\lambda_i|} \approx 1 \tag{3.17}$$

is satisfied, the chosen dimension brings a good approximation for the input data. In sum, this technique allows a representation, usually in two or three dimensions, that better preserves the divergence 3.14.

Consequently, the analysis of the result is not based on the point coordinates and instead relies on the visualization of the patterns and clusters (48).

[5] The derivations behind such steps are worked out in Appendix A.

4 RESULTS AND DISCUSSION

4.1 ANALYTICAL RESULTS

For consistency, we began our analysis by inferring some general properties of the used models: the expected fitness values of an NK landscape, and the relation between a Wright-Fisher model with constant population size and our bottleneck protocol.

4.1.1 Expected fitness values

The expected fitness value of the global optimum (GO) of the NK fitness landscape can be predicted analytically for any sequence length L, for the two asymptotic cases $K = 0$ (a smooth landscape) and $K = L - 1$ (a maximally rugged landscape). These expectations are not only of theoretical interest, but, importantly, allow us to independently validate the computational implementation of the fitness landscape.

We use an additive fitness function in which the contribution of locus x_i, $f(x_i)$, is drawn from a uniform distribution, $U(0, 1]$. When $K = 0$, the contribution of locus x_i to the GO fitness, f_{max}, is the maximum of the two possible fitness values at locus i. It is straightforward to demonstrate that the expected value of the maximum of two draws from $U(0, 1]$ is 2/3. Thus the GO fitness f_{max} is given by $1/L$ times the sum of L independent, identically-distributed random variables, each of which has expected value 2/3. The expected value of the GO, $E[f_{max}]$, is thus 2/3.

When $K = L - 1$, each of 2^L possible sequences is independently assigned a fitness value. *Each* of these fitness values is computed as $1/L$ times the sum of L draws from $U(0, 1]$ (Equation 3.1). The cumulative density function (Cumulative Density Function (CDF)) for the sum of L draws from $U(0, 1]$ is given by the Irwin-Hall distribution (49, 50):

$$H_L(x) = \frac{1}{L!} \sum_{i=0}^{\lfloor x \rfloor} (-1)^i \binom{L}{i} (x - i)^L \ . \tag{4.1}$$

To determine the expected value of the GO, we first compute the CDF of the GO, $M(x)$. If we let F denote the random variable for the fitness of a sequence, $M(x)$ gives the probability

that the maximum of 2^L independently drawn values of F is less than or equal to x.

$$\begin{aligned} M(x) &= \text{Prob(maximum of } 2^L \text{ values of } F \leq x) \\ &= \text{Prob(maximum of } 2^L \text{ draws from } H_L \leq Lx) \\ &= \text{Prob(each of } 2^L \text{ draws from } H_L \leq Lx) \\ &= \prod_{k=1}^{2^L} H_L(Lx) \\ &= \prod_{k=1}^{2^L} \frac{1}{L!} \sum_{i=0}^{\lfloor Lx \rfloor} (-1)^i \binom{L}{i} (Lx - i)^L \ . \end{aligned} \tag{4.2}$$

The expected value of the GO is then given by integrating the product of x with the probability density function associated with $M(x)$:

$$E[f_{max}] = \int_0^\infty x \frac{d}{dx} M(x) dx \ . \tag{4.3}$$

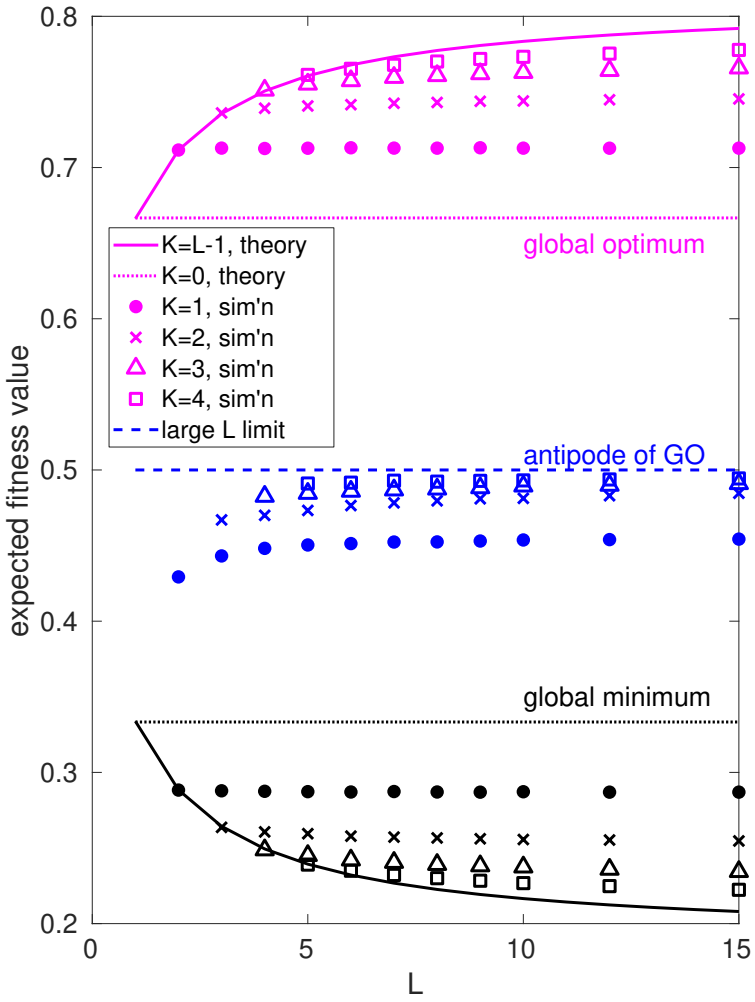
In Figure 25, we illustrate these analytical predictions of $E[f_{max}]$, for $K = 0$ and $K = L-1$, along with results for the observed value of f_{max}, averaged over 100,000 simulated landscapes. As expected, the analytical predictions agree with simulation results when $K = 0$ or $K = L-1$, and in all other cases give upper and lower bounds on $E[f_{max}]$. We further observe that for the relatively short sequences we investigate, the GO fitness depends strongly on the ruggedness of the landscape (K) and weakly on sequence length (L).

For comparison, we also plot the global minimum fitness (black) and the fitness of the antipode of the GO (blue) in Figure 25. By analogous arguments, is straightforward to demonstrate that the expected value of the global minimum is simply $1 - E[f_{max}]$. When $K = 0$, the global minimum corresponds to the antipode of the GO, and thus the expected value of the global minimum and the antipode of the GO is 1/3 (black dotted line). When $K = L - 1$, the fitness of the GO antipode is given by a randomly chosen fitness from the landscape, conditioned by the fact that the chosen fitness is not the GO. Thus the GO antipode fitness approaches the mean landscape fitness, 1/2, as L increases. For smaller values of L, the expected fitness of the GO antipode is less than 0.5, because this conditioning has a more pronounced effect when L is small.

4.1.2 Wright fisher comparison

Consider a discrete time Wright-Fisher model with constant population size N, in which the ith individual in the population has absolute fitness W_i. This parent individual i has

Figure 25 – Global optimum, antipode of global optimum and global minimum fitness values, versus sequence length, L. The analytical prediction (Equation 4.3, evaluated numerically) for the global optimum fitness, $E[f_{max}]$, is shown for $K = L-1$ (magenta solid line), along with the analytical prediction of $E[f_{max}] = 2/3$ for $K = 0$ (magenta dotted line). Analogous analytical predictions for the global minimum fitness are shown for comparison (black lines). Simulation results are shown for comparison for $K = 1$, 2, 3 and 4 (magenta and black symbols as indicated). Simulation results for the expected fitness of the antipode of the global optimum are shown in blue, along with the asympotic expectation (0.5 for large L and $K = L-1$, dashed line). Simulation results show the mean across 100,000 randomly generated fitness landscapes in each case. Error bars for simulation results are similar to symbol heights and omitted for clarity.



Source: The author (2020).

a Poisson-distributed number of offspring with expected value W_i, and these offspring are sampled to form the next generation. It is standard to assume that the number of offspring is large, such that offspring can be sampled with replacement, that is, each offspring is selected independently with a fixed probability. To maintain a constant population size, the sampling probability must be $1/\bar{W}$, where $\bar{W}$ is the mean population fitness. The probability generating function (PGF) for the descendants of individual i in the next generation, $F_i(x)$, is then given by the composition of the Poisson offspring PGF (Equation 2.15), $\exp(W_i(x-1))$ and the

binomial *sampling* PGF, $(1-p)+px$ (where $p = 1/\bar{W}$ is the sampling probability):

$$\begin{aligned} F_i(x) &= \exp\left[W_i(x'-1)\right] \\ &= \exp\left[W_i\left(\left(1-\frac{1}{\bar{W}}\right)+\frac{1}{\bar{W}}x-1\right)\right] \\ &= \exp\left[\frac{W_i}{\bar{W}}(x-1)\right]. \end{aligned} \tag{4.4}$$

Thus, the net effect of this process – a large, Poisson-distributed number of offspring, followed by independent sampling with a constant probability – is the same as a Poisson distribution of descendants with mean $\frac{W_i}{\bar{W}}$.

In the simulations to follow, the population size doubles for τ generations, with each individual contributing offspring to the next generation in proportion to their relative fitness. The population is then sampled with sampling probability $2^{-\tau}$. When $\tau = 1$, the contribution of the ith member of the initial population (of size N_0) to the next population of size N_0 (after one cycle of growth and one bottleneck) is therefore given by:

$$f_i(x) = \exp\left[\frac{2W_i}{\bar{W}}\left(\left(1-\frac{1}{2}\right)+\frac{1}{2}x-1\right)\right] \tag{4.5}$$

$$= \exp\left[\frac{W_i}{\bar{W}}(x-1)\right] = F_i(x). \tag{4.6}$$

Thus, in the results to follow, cases illustrated for $\tau = 1$ (often the extreme or asymptotic cases) are equivalent to a standard discrete time Wright-Fisher model at fixed population size N_0. In other words, the case $\tau = 1$ reveals the behaviour of a Wright-Fisher population in the absence of population bottlenecks.

4.2 SIMULATION RESULTS

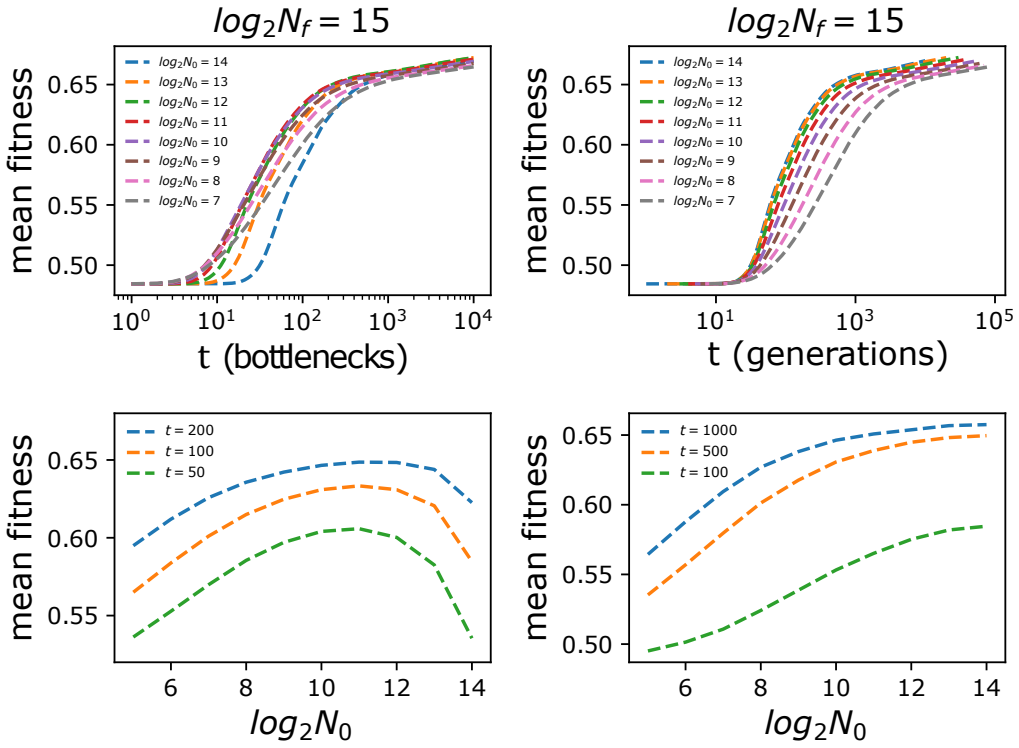
We will first investigate the role of population bottlenecks in adaptation. In particular, we are interested in the conditions that optimize the rate of adaptation. In a second section, we turn our attention to the characterization of evolutionary pathways, and how their statistical properties are affected by the bottlenecks.

4.2.1 Fitness trajectories

Figure 26 shows average fitness trajectories for different bottleneck ratios. The data correspond to an average of 1000 replicates each for over 50 distinct randomly-drawn fitness

landscapes. Here, the population size after the growth phase is fixed, $N_f = 2^{15} = 32768$. The fitness trajectories are shown in time units of bottlenecks (left panels) and in time units of generations (right panels).

Figure 26 – Fitness trajectories, that is, mean population fitness versus time for different bottleneck ratios (upper panels) and fitness versus bottleneck size at different times (lower panels). Time is measured in units of bottlenecks (left panels) and generations (right panels). The parameter values are mutation rate $U = 10^{-4}$, sequence size $L = 8$, epistasis parameter $K = 2$, and N_f is set at $N_f = 2^{15} = 32768$. The bottleneck sizes are indicated in the legends. In the bottom panels the curves correspond to fixed numbers of bottlenecks (or generations) as indicated in the legends.



Source: The author (2020).

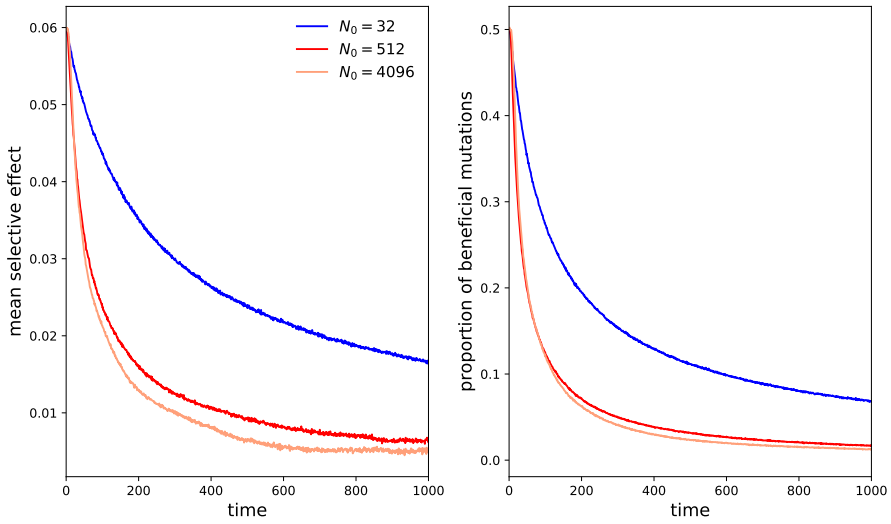
We first observe that fitness trajectories are concave, thus evincing a pattern of diminishing-returns in which fitness quickly grows at the early stages of adaptation and then slows at later times (note the semi-log scale of the plot) (51, 52).

This feature is in contrast with results obtained in others fitness landscape models (53), in which the availability of beneficial mutations and their selection coefficients remain the same as adaptation proceeds, allowing population fitness to grow exponentially over time. Our results are, however, compatible with empirical observations of *E. coli* populations over long-term adaptation (28, 31, 5, 54).

As previously seen, adaptation occurs through the accumulation of beneficial mutations of ever-increasing fitness. To investigate its correspondence with the observed pattern, we

measure the supply of beneficial mutation available to be acquired by the population and their respective mean selective advantage (Figure 27) as a function of the time between bottlenecks. The rapid drop of both quantities is a clear signature of the pattern of diminishing returns (51), and explains why the rate of increase of fitness slows down with time.

Figure 27 – Mean selective effect and proportion of beneficial mutations as a function of time in units of bottleneck events. Since fitness is a relative measure, the selective effects of the beneficial mutations correspond to the fitness advantage they confer at the genetic background they arise. Its clear from the plot, that those arising at a later time have a smaller effect. Also, note that both measures decrease under severe bottleneck regimes, which suggests a decrease in adaptation rate. The different curves correspond to distinct values of N_0, as indicated in the legends. The other parameter values are $N_f = 32768$, mutation rate $U = 10^{-4}$, sequence size $L = 8$ and epistasis parameter $K = 2$.



Source: The author (2020).

We do not expect the bottleneck ratio to produce any considerable discrepancy in the long-term fitness attained by the populations; however bottlenecks do play a clear role at earlier stages of adaptation. These findings are better summarized in the lower panels of Figure 26. Here, curves correspond to distinct times at which fitness was reported. As explained in Section 3.3.2, different time units may render different results: when measured in bottleneck events (left panels), the mean fitness has a maximum at intermediate bottleneck protocols; however, if time is measure in generations (right panels), mean fitness grows monotonically with bottleneck size N_0. In the latter case, larger effective population sizes lead to a higher rate of adaptation, although this increase begins to saturate for very large population sizes. Thus, the adaptation rate *per generation* is simply maximized by experimental protocols that maximize the supply of beneficial mutations NU, that is, by the largest effective population size.

To understand the role of bottlenecks in adaptation, three main factors must be considered. First, the severity of the bottleneck itself clearly poses an obstacle for the survival of a new lineage. Second, unless the population goes extinct or can grow infinitely large, bottlenecks must occur hand-in-hand with periods of sustained growth; these periods of growth favor the survival of beneficial lineages. Finally, bottlenecks, and the inherent changes in population size they confer, change the rate of supply of beneficial mutations NU, at each generation step.

In sum, our results indicate that there is a trade-off between sampling the population too frequently and imposing infrequent, but more severe, bottlenecks. Moreover, these results predict that sampling about $10-20\,\%$ of the population will maximize the speed of adaptation *per bottleneck*; this agrees with previous studies addressing the impact of bottleneck ratios on the fixation probability of beneficial mutations, predicting that the optimum ratio occurs around a dilution $D \sim 1/e^2$ (6, 30). [1]

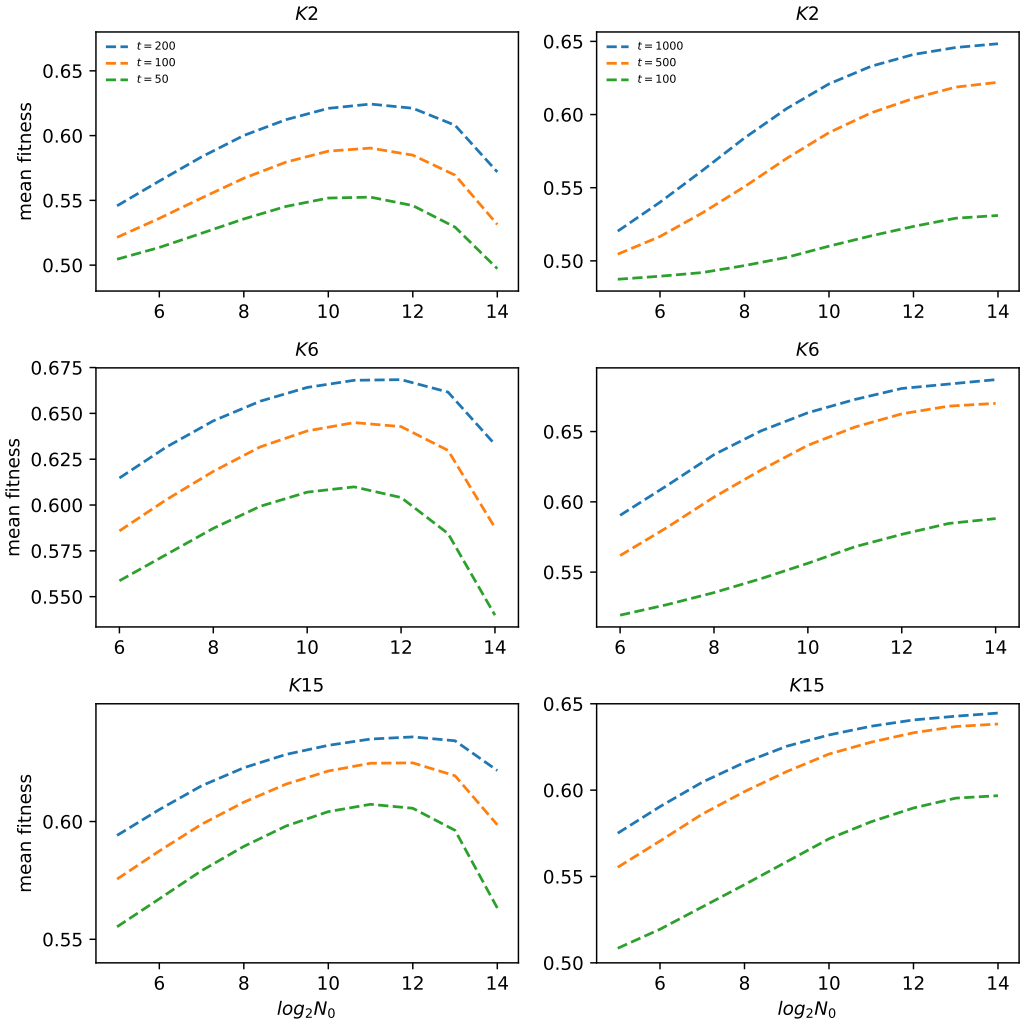
The existence of an optimal bottleneck ratio seems to be inherent to the dynamics when time is given in units of bottlenecks, being observed at different times and across landscape structures; Figure 28 shows analogous results for varying degrees of epistasis in the landscape. The lowest panels corresponds to the extreme condition, where $K = L - 1$, at which any single mutation changes the fitness value of a given genotype in a random manner, and so the fitness landscape is said to be completely uncorrelated. We notice that under this extreme case, the optimal adaptation seems to be slightly shifted towards larger bottleneck sizes when compared to low and intermediate correlated fitness landscapes.

Finally, when time is measured in units of birth events (see Figure 35 in the Appendix C), one recovers the scenario shown on the left in Figures 26 and 28, and once again the highest adaptation rates are found at intermediate bottleneck sizes. In the results above we changed the bottleneck ratio by varying N_0, while holding N_f fixed. Analogous results, with N_0 constant while N_f varies, are shown in Figure 37. Whether time is measured in units of bottlenecks and generations, the mean population fitness displays a monotonic increase with N_f in this case. Overall, we conclude that when time is measured in generations, the adaptation rate increases monotonically in larger populations, irrespective of population bottlenecks.

We now pass to understanding the optimal adaptation rate at intermediate bottleneck sizes by investigating its effects on the population's genetic diversity and fitness variance.

[1] Although the cited study does not account for clonal competition and epistasis.

Figure 28 – Dependence of mean fitness on bottleneck size, measured at different times, for varying degrees of epistasis. Time is expressed in units of bottlenecks (generations) on the left (right) panels. The parameter values are $N_f = 2^{15}$, mutation rate $U = 10^{-4}$, sequence size $L = 16$ and epistasis parameters as indicated in the titles.



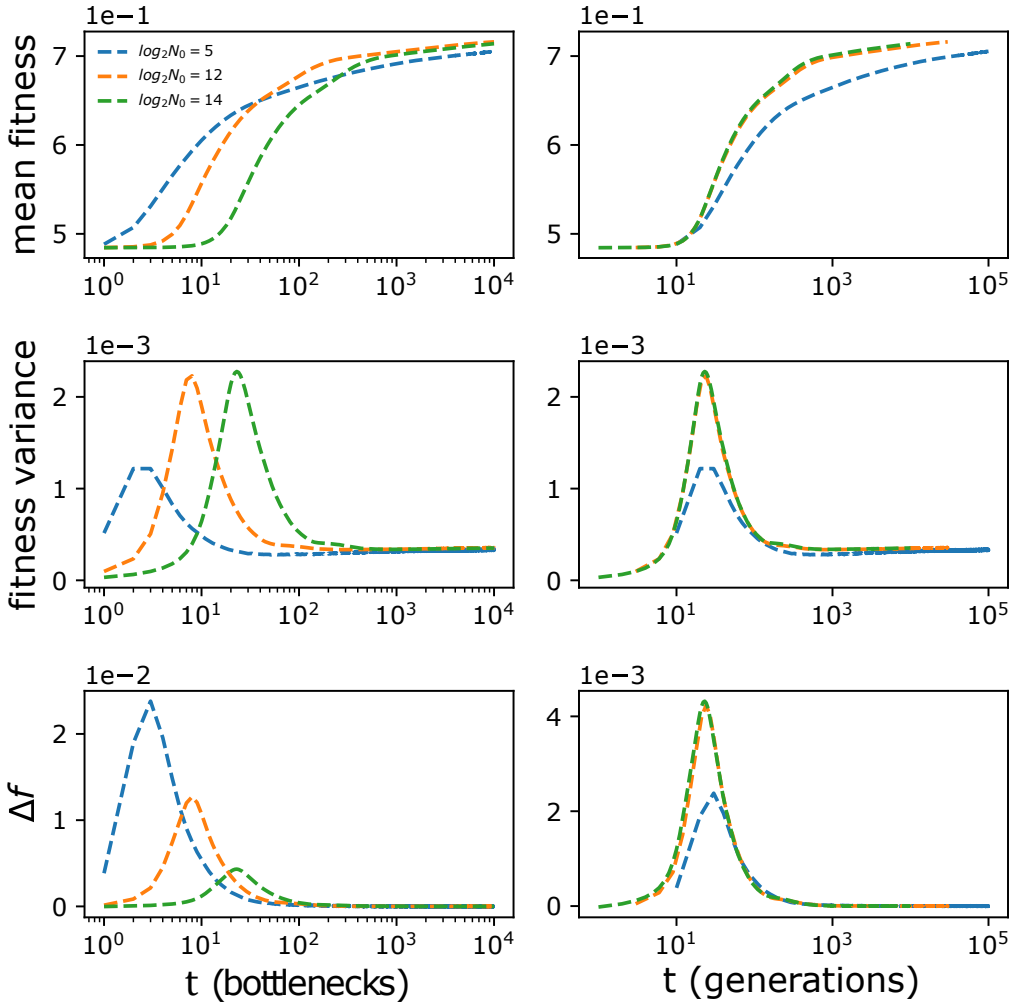
Source: The author (2020).

4.2.2 Effects of bottlenecks on genetic diversity

Figure 29 shows the time evolution of the mean population fitness, the fitness variance as well the fitness increase at each time step. Three different bottleneck sizes are compared: $N_0 = 2^5 = 32$, $N_0 = 2^{12} = 4096$ and $N_0 = 2^{14} = 16384$. In all cases the population size at the end of the growth phase is $N_f = 2^{15}$, and mean fitness and fitness variance are compared (in this figure) just after each population bottleneck, for a population of size N_0.

First, in the earlier stages of adaptation, while adaptation occurs at a faster pace, the fitness variance is considerably enhanced. As expected, in each population the adaptation rate (slope in top panel) is greatest when the fitness variance is maximized; we can confirm this

Figure 29 – Mean population fitness, fitness variance and change in fitness Δf as a function of time. Time is expressed in units of bottlenecks (generations) on the left (right) panels. The parameter values are $N_f = 2^{15} = 32768$, mutation rate $U = 10^{-4}$, sequence size $L = 8$ and epistasis and $K = 2$. The bottleneck sizes are $N_0 = 32$ (blue dashed-lines), $N_0 = 4096$ (orange dashed-lines) and $N_0 = 16384$ (green dashed-lines). Δf is simply the mean population fitness at time $t + 1$ minus the mean population fitness at time t, for t in the units indicated.



Source: The author (2020).

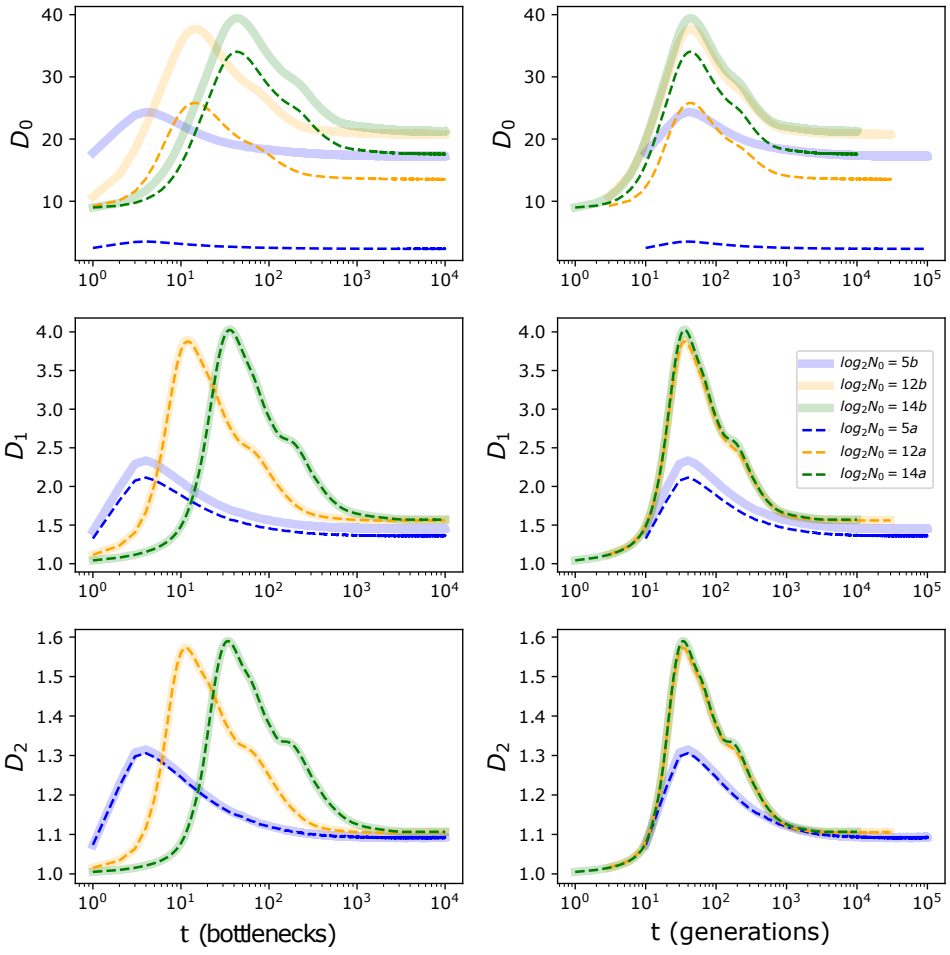
by comparing the fitness variance with the change in fitness per time step (bottom panel). We also note in the centre panels that the peak fitness variance is reduced for the smallest population $N_0 = 32$, but is not sensitive to the bottleneck for less severe bottleneck ratios. As the fitness variance is measured just after a bottleneck, this loss in fitness variance is attributed to the loss of lineages through the bottleneck; severe bottlenecks retard the adaptation rate through the loss of fitness variance.

As can be corroborated from Figure 29, measuring time in generations reconciles the time courses of mean fitness and fitness variance, and isolates the effects of the bottlenecks. We see clearly here that only the most severe bottleneck ratio in this example (2^{-10}) reduces fitness

variance and thus retards the adaptation rate. Importantly, we note that when $N_0 = 2^{14}$ and $N_f = 2^{15}$ (green lines), the simulated population dynamics are mathematically equivalent to a discrete time Wright-Fisher process, in a population of fixed size 2^{14} (see Analytical Results, above). Thus, the green lines plot the time course of adaptation *that would be obtained in the absence of population bottlenecks*. We see that when time is measured in units of generations, the rate of adaptation is almost insensitive to bottlenecks, once the bottleneck size exceeds several thousand individuals.

Since the fitness variance is related to the genetic diversity within the population, we investigate the change in genetic composition by measuring its diversity (in Hill numbers) immediately after, and before a bottleneck event.

Figure 30 – Hill diversity numbers D_0, D_1 and D_2 versus time. Time is expressed in units of bottlenecks (generations) on the left (right) panels. The measures of diversity are presented at the end of the growth phase (solid lines) and just after the bottleneck protocol (dashed lines). The parameter values are $N_f = 2^{15} = 32768$, mutation rate $U = 10^{-4}$, sequence size $L = 8$ and epistasis parameter $K = 2$. The bottleneck sizes are $N_0 = 32$ (blue lines), $N_0 = 4096$ (orange lines) and $N_0 = 16384$ (green lines) as indicated in the legends. The symbol a in the legend means just after bottlenecks, whereas b means just before bottlenecks.



Source: The author (2020).

Remember that the Hill numbers D_a, gives smaller weight to rare entities as we increase its parameter a. Therefore, D_0 accounts for the absolute genetic diversity within a population, corresponding to the number of distinct lineages; while D_1 and D_2 ignores the lowest frequency genes, putting less emphasis on rare genotypes.

In Figure 30, the Hill numbers D_0, D_1 and D_2 are plotted over time for the same populations ($N_0 = 2^5$, $N_0 = 2^{12}$ and $N_0 = 2^{14}$, with $N_f = 2^{15}$). Values computed just before the population bottleneck (solid lines), and immediately after the bottleneck protocol (dashed lines) are shown. Taking time in units of generations allows us to isolate the effect of the bottlenecks on genetic diversity.

We see that the first and second order Hill numbers, D_1 and D_2, change only negligibly over the course of a single bottleneck (solid versus dashed lines). In contrast, the zero-th order Hill number D_0 is greatly reduced, especially when the dilution ratio N_0/N_f is small (blue lines correspond to a ratio of $2^{-10} \approx 1 \times 10^{-3}$). This loss of rare lineages, revealed by D_0, has long term consequences for the population; although a single bottleneck has little effect on D_1 or D_2, we see that both of these diversity measures are greatly reduced in the $N_0 = 32$ population. Thus the loss of rare lineages through the bottleneck feeds forward, resulting in overall reduced diversity at later times, even as measured by higher order metrics that are less sensitive to rare lineages, and despite the supply of mutations happening along with the sustained growth phase.

4.2.3 Effects of bottlenecks on genetic contigency

At this point, we aim to carry out a more detailed analysis of the evolutionary pathways at the genotypic level.

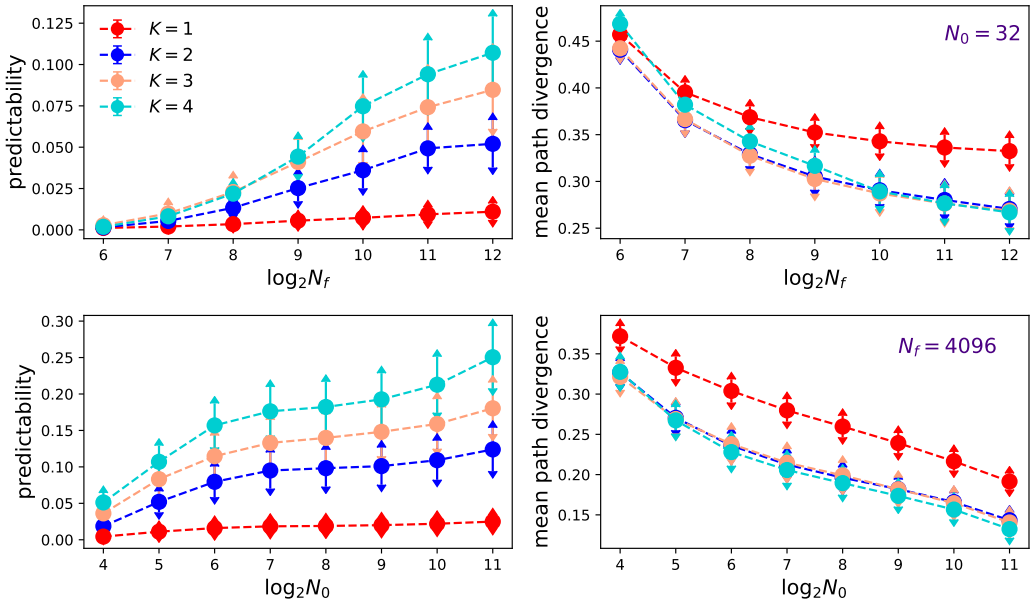
As fitness landscapes are complex and may have multiple peaks, some mutational paths may lead to 'dead ends' with no, or at least fewer, opportunities to further improve. In other cases, certain mutations may open up new opportunities for evolution that could not be accessed if other routes were taken. As mentioned previously, the two quantities, predictability and mean path divergence, may elucidate these dynamics.

Computationally, simulations that quantify predictability and mean path divergence are costly, as they require that the starting and ending points of the trajectories are the same. We first investigate adaptive trajectories assuming that the population starts from the antipode of σ_{max}, the global optimum of the fitness landscape. In this case, the Hamming distance

from the starting point to the global optimum takes its maximum value, being equal to the sequence size L, $d_{GO} = L$. While the route is smooth for an additive landscape ($K = 0$), it becomes increasingly tortuous as the landscape becomes more rugged. For low mutation rates, the population is easily trapped by local optima of the landscape, and so the time needed to reach the global optimum rises substantially.

As an overall result, we observe from Figure 31 a monotonic increase of predictability with N_0, along with the decline in mean path divergence. This corroborates the observed negative correlation between predictability and mean path divergence on simulations over two empirical fitness landscapes (55), although such a claim may not be generalizable, as such correlations may strongly depend on the topological properties of the underlying fitness landscape. We note further that predictability rises steeply with N_0 when N_0 is small, but saturates at larger N_0 values. Therefore, we find that irrespective of the timing of population bottlenecks, predictability increases and saturates with an increasing mutational supply.

Figure 31 – Predictability and mean path divergence. In the upper panels both quantities are shown as a function of the population size at the end of the growth phase N_f. In these panels the population size after the bottleneck is set at $N_0 = 32$. In the lower panels both quantities are shown as a function of the population size after the bottleneck N_0. In these panels the population size N_f is set at $N_f = 4096$, whereas the mutation rate is set at $U = 5 \times 10^{-2}$ and sequence size at $L = 8$. The epistasis parameter K is displayed in the legends.



Source: The author (2020).

It is generally accepted that large populations will tend to evolve more rapidly than smaller ones. This is caused by two related factors. First, large populations have an increased supply of beneficial mutations each generation NU, which decreases the waiting time for new ad-

vantageous mutations to arise. Second, large populations have increased access to mutations that confer large benefits (as seen in Figure 27). These factors imply that larger populations gain an advantage by taking larger adaptive steps during population evolution.

Because large populations tend to fix the most advantageous mutations first, they thereby follow a very limited set of adaptive trajectories. In contrast, smaller populations become fixed for a wider range of possible beneficial mutations which leads to increased variation in adaptive trajectories across populations.

The simulation results also demonstrate that the predictability seems to be much more sensitive to the ruggedness of the fitness landscape than the mean path divergence. While the predictability rises monotonically with K, the mean path divergence seems to be bounded already at intermediate ruggedness, and the curves for $K = 2, 3$ and 4 nearly collapse. Overall, when considering adaptive trajectories that are constrained to start at the antipode and end at the GO, any increase in epistasis (landscape ruggedness) very quickly reduces the divergence of pathways, and more gradually increases predictability.

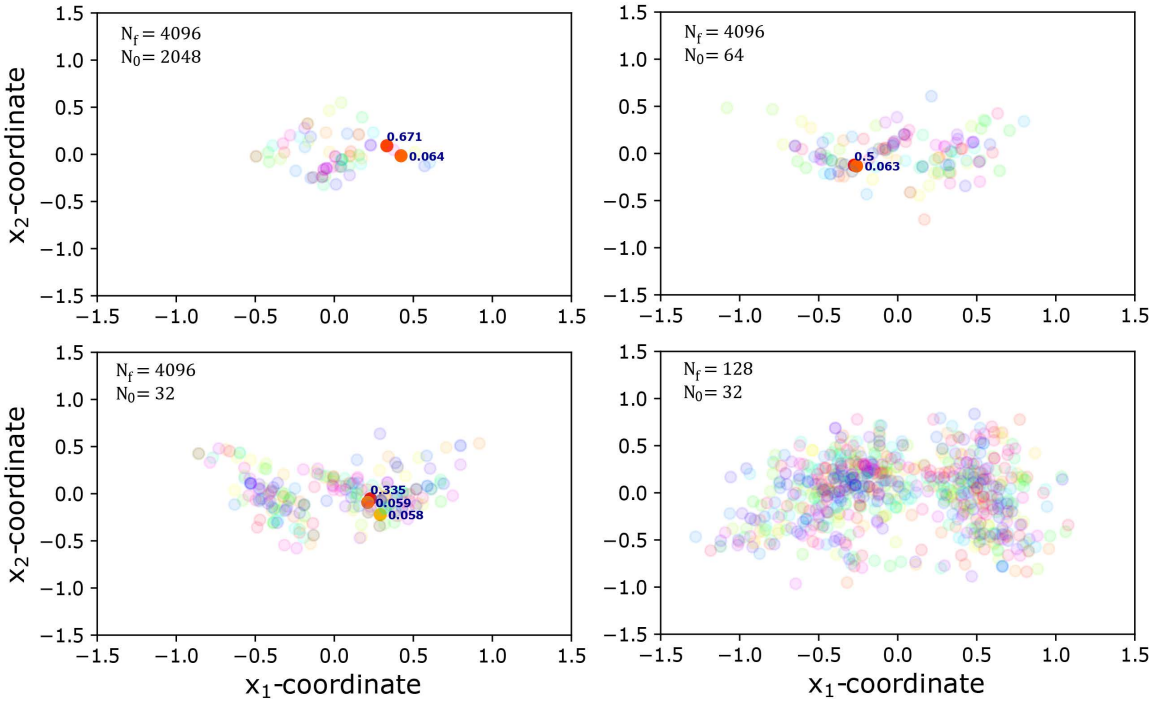
Bearing the difficulty of directly access the microscopic change, we recur to the computational analysis of multidimensional scaling in the hope that its patterns can contrast or corroborate our hypotheses. Figure 32 provides two-dimensional representations of the ensemble of distinct evolutionary pathways collected in one thousand replicates each starting with an isogenic population at the antipode of σ_{max}, and ending at the global optimum σ_{max} of the fitness landscape.

For direct comparison, parameters of Figure 31 were adopted, i.e.: $U = 5 \times 10^{-2}$, $L = 8$ and $K = 2$. In Figure 32, N_f is set at $N_f = 4096$, whereas we vary $N_0 = \{2048, 64, 32\}$ (first three panels). An extreme case, of $N_0 = 32$ to $N_f = 128$, is ploted on the last panel (bottom right). Paths that are used with a frequency of 5% or higher are highlighted.

For the largest population size $N_f = 4096$, the distribution of trajectories becomes substantially more compact as we increase N_0. Remember that the axis distances are approximations of the inner-path distance (Equation 3.16), thus this closeness reflects the decrease in the mean path divergence reported in Figure 31. Furthermore, we also note its effects in the paths that achieve high frequencies; the most frequent path is taken in nearly 33% of the independent runs, to 50% and 67% at higher N_0. This change is likewise captured by the measure of predictability in the right panel of Figure 31.

Thus, when the bottleneck size, N_0, is increased, the bundle of evolutionary paths becomes less widespread over genotype space, and some paths are more frequently used. These two

Figure 32 – Multidimensional scaling plot of the evolutionary pathways. In the upper panels N_f is set at $N_f = 4096$, whereas $N_0 = 2048$ (left upper panel) and $N_0 = 64$ (right upper panel). In the lower panels, N_0 is set at $N_0 = 32$, whereas $N_f = 4096$ (left lower panel) and $N_f = 128$ (right lower panel). The data correspond to a fixed fitness landscape with epistasis parameter $K = 2$. Those evolutionary pathways that achieve a frequency higher than 0.05 are highlighted in the plot (dark circles, with numbers indicating path frequency).



Source: The author (2020).

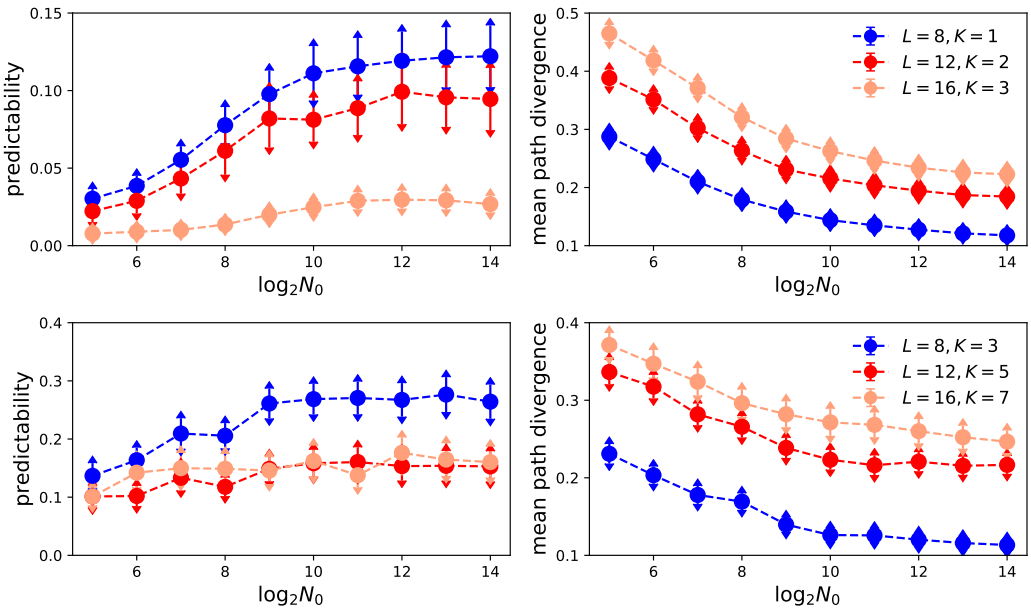
facts - that the paths are less scattered over genotype space, and that the frequency of the most-used paths increases - explain both the increase in predictability and decrease in the mean path divergence seen in the panels of Figure 31.

The lower panels of Figure 32 show analogous results if the bottleneck size $N_0 = 32$ is fixed while N_f is increased. Thus, the constraint of pathways can also be ascribed to the expansion of the growth phase, boosting the determinism of the evolutionary process. On the right, for $N_f = 128$, the effective population size is correspondingly small. As expected, the distribution of evolutionary paths is quite diffuse in this extreme case; besides the higher dispersion of points, none of the paths reach a frequency of 5%.

The main limitation of the previous protocol is that the study cannot be generalized to larger sequence sizes or to less restrictive ranges of correlation in the fitness landscape, because the population tends to become trapped at local maxima, making the global optimum essentially inaccessible. To address this issue, as discussed in Section 3.3.4, instead of initiating

the dynamics from the antipode of the GO, the evolutionary process is now initiated from genotypes that are placed at a given Hamming distance, $d_{GO} < L$, from the global optimum.

Figure 33 – Predictability and mean path divergence against the bottleneck size N_0. In all plots, the population size at the end of the growth phase is $N_f = 32768$, and the mutation rate is $U = 5 \times 10^{-3}$. The sequence size L and epistasis parameter K are both varied such that the correlation $\rho = 1 - (K+1)/L$ is kept constant. In the upper panels $\rho = 0.75$, whereas in the bottom panels $\rho = 0.5$. The Hamming distance from starting points to the global optimum is five, $d_{GO} = 5$. The simulation data plot an average over 10 distinct fitness landscapes, and 10 random starting points for each landscape.



Source: The author (2020).

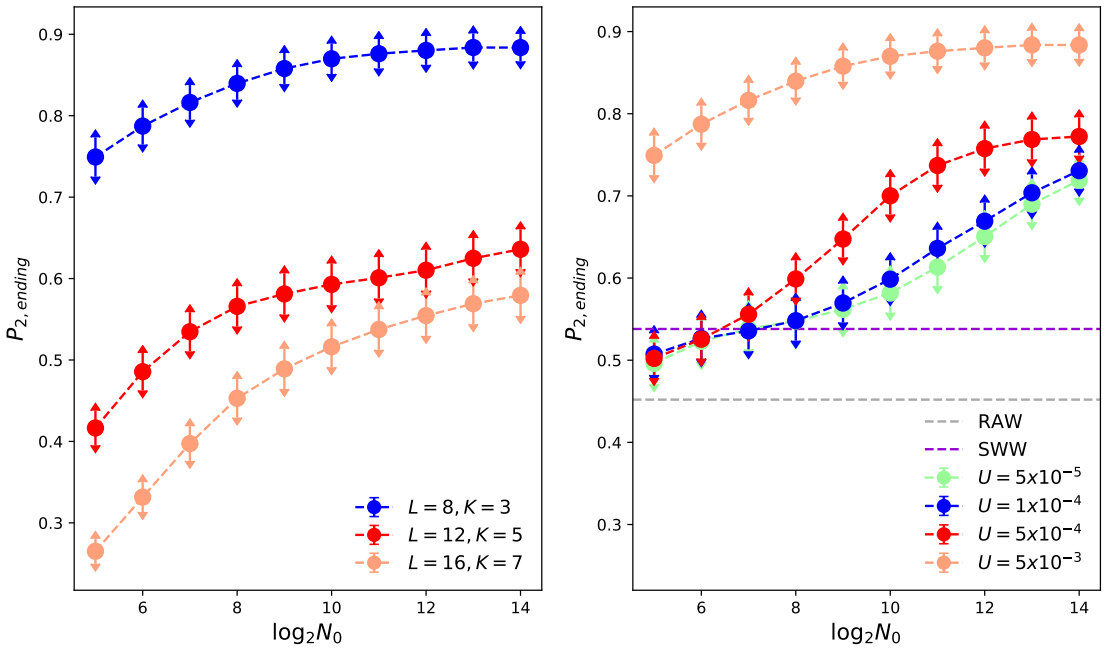
Results are shown in Figure 33 for $d_{GO} = 5$, where the sequence size, as well as the epistasis parameter K, are varied such that the degree of correlation among fitness effects of neighbors mutations, $\rho = 1 - (K+1)/L$, remains unchanged. In the plot, this correlation is set at $\rho = 0.75$ (upper panels) and $\rho = 0.5$ (lower panels).

In spite of the Hamming distance from the starting points to the global optimum being the same, $d_{GO} = 5$, the population have a number $L!$ of distinct combination of paths to reach a same final state. Thus, we observe that predictability decreases while the mean path divergence increases with the sequence size L. Note that here both quantities are measured with respect to the paths, i.e., the starting and ending points for each ensemble of trajectories are the same. The data refer to an average of over 10 distinct starting points, all chosen at the same distance d_{GO} from the global optimum.

In Figure 34, the predictability with respect to the ending points, denoted by $P_{2,ending}$, is shown as a function of N_0. The quantity $P_{2,ending}$ gives the probability that two randomly

chosen paths terminate at the same genotype. Here, the evolutionary trajectories begin at a fixed distance from the global optimum, and are simulated up to a fixed time, $t_{bottleneck} = 2,000$. In the left panel, the dependence on the sequence size L is shown as the correlation ρ is fixed. Note that for $L = 8$ and $K = 3$ the predictability is very close to one, meaning that the global optimum is easily accessible. As L increases, the accessibility of the global optimum decreases and so does the predictability.

Figure 34 – Predictability with respect to the ending points. In all plots, the population size at the end of the growth phase is $N_f = 32768$. In the left panel, the mutation rate is $U = 5 \times 10^{-3}$ whereas the sequence size L and epistasis parameter K are both varied such that the correlation is $\rho = 0.5$. In the right panel, the sequence size is $L = 8$ and the epistasis parameter is set at $K = 3$. The Hamming distance from starting points to the global optimum is five, $d_{GO} = 5$. The simulation data refers to an everage over 10 distinct fitness landscapes, and 10 random starting points for each landscape. The dashed-lines correspond to the predictability with respect to ending points for random adaptation walks (RAW) and for S-weighted walks (SWW).



Source: The author (2020).

In the right panel we investigate the role of the mutation rate. Higher mutation rates lead to higher predictability with respect to the ending points as they also allow the population to more easily escape from local maxima of the landscape and ultimately reach the global optimum. When the mutation rate U and bottleneck size are sufficiently small, predictability becomes independent of the mutation rate. In order to understand this effect, we additionally simulated two versions of adaptive walks, named "random adaptation walks" (Random Adaptation Walks (RAW)) and the "S-weighted walks" (S-Weighted Walks (SWW)) (56, 55). In the former, the walker randomly chooses one of its fitter neighbors, whereas in the SWW version the next step

is chosen with probability proportional to the fitness advantage of its fitter neighbors. From these results, we find that in the limit $N_0/N_f \ll 1$ the predictability $P_{2,ending}$ lies between the two adaptive walk variants. We hypothesize that when the population size and mutation rate are sufficiently small, the simulated populations exist in the strong-selection weak-mutation regime (57), and thus the dynamics are well-approximated by adaptive walk dynamics. In particular, the predictability with respect to the ending points, $P_{2,ending}$ at low mutation rates is sensitive to the earliest stages of adaptation, in which the selective effects of mutations are expected to be larger. This explains why the dynamics in those limits are well captured by adaptive walk dynamics.

In all of the scenarios illustrated here, predictability exhibits a monotonic dependence on the bottleneck size N_0, meaning that the underlying dynamics become increasingly deterministic for larger populations, not only with respect to the paths but also with respect to the ending points.

5 CONCLUSIONS

In this work, we used an epistatic fitness landscape to explore two hypotheses: whether the adaptation rate is affected by population bottlenecks, and whether population bottlenecks reduce the predictability of adaptation.

In the first results, we obtain that adaptive trajectories in populations experiencing regular bottlenecks can be reconciled when time is scaled in units of generations (in our case, population doublings). We demonstrate that the time course of fitness increase, fitness variance and genetic diversity are all insensitive to population bottlenecks when time is expressed in population generations, provided the bottleneck size exceeds a few thousand individuals (Figures 29 and 30). Thus, in contrast with previous results that adaptation *per bottleneck* is fastest at intermediate bottleneck ratios (6), we demonstrate that the adaptation rate *per generation* is simply maximized by experimental protocols that maximize the supply of beneficial mutations, that is, by the largest effective population size (Figure 26, right panels).

We demonstrate that small bottleneck sizes can retard adaptation through the elimination of rare lineages, but this effect disappears when N_0 is of order of one thousand individuals, rather than tens of individuals. Thus, for most microbial populations, the adaptation rate *per generation* will be largely insensitive to the bottleneck ratio. Overall these results imply that the "natural" time unit for adaptation is generations, irrespective of the number of generations that elapse between population bottlenecks, as long as the bottleneck size is not too small.

When adaptation rate is measured in bottleneck events, an optimum at intermediate bottleneck size is obtained. The optimum ratio at $10-20\%$ approximates the analytically derived results by Wahl et al. (6, 30) for the probability of survival and fixation of a lineage. This may have practical applications to studies where the maintenance of specific beneficial mutations, or their tracking, is of importance. In natural populations, where bottlenecks are related to extinction events or host-to-host contagious, this sample ratio might determine the fate of the surviving individuals or the efficiency of a host invasion.

We also investigated the role of population bottlenecks at a microscopic level by tracking the evolutionary process at the genetic level. We observe a higher level of determinism of the adaptive process when either the bottleneck size N_0 or the population size at the end of the growth phase N_f is increased. In other words, predictability is maximized in larger populations irrespective of population bottlenecks. This dependence was also reported when

repeatability was measured *at the fitness* level in experiments with unicellular algae (58), for which large populations presented a higher degree of repeatability, and small populations were more susceptible to chance events.

More generally, the increase in predictability that we observe in large populations can be ascribed to an increased determinism in the underlying process due to the combined effects of three factors: more beneficial mutations are generated, these mutations have longer times to increase in frequency, and there is greater interference among mutant lineages, thus promoting those that confer larger selective advantage.

In all the scenarios investigated here, we observed a monotonic dependence of predictability on both N_0 and N_f. In contrast, Szendro et al. observed a non-monotonic dependence of predictability measures on population size (41) for the Rough Mount Fuji landscape, which is not expected to display a pattern of diminishing returns. As well as this key difference between the topographies of the fitness landscapes, here we are mostly not concerned with the SSWM regime, which in the range of mutation rates here considered is only attainable when $N_0/N_f \ll 1$. Another point to highlight is that the decrease of predictability with population size observed in (41) was due to the appearance of second-step mutations, increasing when the value of NU^2 was in the range $10^{-6} - 10^{-7}$. For computational efficiency in the very large fitness landscapes simulated here, we have studied comparatively large mutation rates, in which the value of NU^2 is typically several orders of magnitude larger than this threshold, and thus the appearance of second-step mutations is practically assured. We hypothesize that in principle, as the mutation rate increases, alternating regimes in which predictability increases or decreases with population size may be possible, as higher-order mutational neighbourhoods become newly accessible. This would be a clear avenue for future works.

The main results presented in this dissertation have been published in the article:
- Ref. (59): Robustness and predictability of evolution in bottlenecked populations,
Osmar Freitas, Lindi M. Wahl, and Paulo R. A. Campos, Phys. Rev. E (2021)
DOI: 10.1103/PhysRevE.103.042415

REFERENCES

1 NOWAK, M. A. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006. ISBN 9780674023383. Disponível em: <http://www.jstor.org/stable/j.ctvjghw98>.

2 BLOUNT, Z. D.; LENSKI, R. E.; LOSOS, J. B. Contingency and determinism in evolution: Replaying life's tape. *Science*, v. 362, n. 6415, 2018. ISSN 10959203.

3 De Visser, J. A. G.; KRUG, J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, Nature Publishing Group, v. 15, n. 7, p. 480–490, 2014. ISSN 14710064. Disponível em: <http://dx.doi.org/10.1038/nrg3744>.

4 WEINREICH, D. M.; DELANEY, N. F.; DEPRISTO, M. A.; HARTL, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, v. 312, n. 5770, p. 111–114, 2006. ISSN 00368075.

5 BARRICK, J. E.; LENSKI, R. E. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, Nature Publishing Group, v. 14, n. 12, p. 827–839, 2013. ISSN 14710056. Disponível em: <http://dx.doi.org/10.1038/nrg3564>.

6 WAHL, L. M.; GERRISH, P. J.; SAIKA-VOIVOD, I. Evaluating the impact of population bottlenecks in experimental evolution. *Genetics*, v. 162, n. 2, p. 961–971, 2002. ISSN 00166731.

7 ORR, H. A. The genetic theory of adaptation: A brief history. *Nature Reviews Genetics*, v. 6, n. 2, p. 119–127, 2005. ISSN 14710056.

8 DROSSEL, B. Biological evolution and statistical physics. *Advances in Physics*, v. 50, n. 2, p. 209–295, 2001. ISSN 14606976.

9 LOBKOVSKY, A. E.; KOONIN, E. V. Replaying the tape of life: Quantification of the predictability of evolution. *Frontiers in Genetics*, v. 3, n. NOV, 2012. ISSN 16648021.

10 DARWIN, C. *On the Origin of Species by Means of Natural Selection*. London: Murray, 1859. Or the Preservation of Favored Races in the Struggle for Life.

11 PIGLIUCCI, M. Do we need an extended evolutionary synthesis? *Evolution*, v. 61, n. 12, p. 2743–2749, 2007. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2007.00246.x>.

12 GOULD, J. *From "Voyage from the Beagle"*. <http://darwin-online.org.uk/converted/published/1845_Beagle_F14/1845_Beagle_F14_fig07.jpg>.

13 OKASHA, S. Population Genetics. In: ZALTA, E. N. (Ed.). *The Stanford Encyclopedia of Philosophy*. Winter 2016. [S.l.]: Metaphysics Research Lab, Stanford University, 2016.

14 FISHER, R. A. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, Royal Society of Edinburgh Scotland Foundation, v. 52, n. 2, p. 399–433, 1918.

15 FELLER, W. *An introduction to probability theory and its applications. Vol. II.* New York: John Wiley & Sons Inc., 1971. xxiv+669 p. (Second edition).

16 VALE, R. *The Explorer's Guide to Biology: The structure of DNA*. <https://explorebiology.org/collections/genetics/the-structure-of-dna>.

17 MOBERG, C. Schrödinger's what is life?—the 75th anniversary of a book that inspired biology. *Angewandte Chemie International Edition*, v. 59, n. 7, p. 2550–2553, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201911112>.

18 WAGNER G., Z. J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet*, n. 12, p. 204–213, 2011.

19 WRIGHT, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth international congress of Genetics*, v. 1, p. 356–366, 1932.

20 KAPLAN, J. The end of the adaptive landscape metaphor?. *Biol Philos*, n. 23, p. 625–638, 2008.

21 OGBUNUGAFOR, C. B. A Reflection on 50 Years of John Maynard Smith's "Protein Space". *Genetics*, v. 214, n. April, p. 749–754, 2020.

22 SMITH, J. M. Natural selection and the concept of a protein space. *Nature*, v. 225, p. 563–564, 1970.

23 KAUFFMAN, S. A.; WEINBERGER, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, v. 141, n. 2, p. 211–245, 1989. ISSN 10958541.

24 KAUFFMAN, S.; LEVIN, S. Towards a General Theory of Adaptive Walks on Rugged Landscapes Section of Ecology and Systematics , and Ecosystems Research Center ,. *New York*, v. 128, n. 1, p. 11–45, 1987. ISSN 0022-5193.

25 HWANG, S.; SCHMIEGELT, B.; FERRETTI, L.; KRUG, J. Universality Classes of Interaction Structures for NK Fitness Landscapes. *Journal of Statistical Physics*, Springer US, v. 172, n. 1, p. 226–278, 2018. ISSN 00224715. Disponível em: <https://doi.org/10.1007/s10955-018-1979-z>.

26 LOBKOVSKY, A. E.; WOLF, Y. I.; KOONIN, E. V. Predictability of evolutionary trajectories in fitness landscapes. *PLOS Computational Biology*, Public Library of Science, v. 7, n. 12, p. 1–11, 12 2011. Disponível em: <https://doi.org/10.1371/journal.pcbi.1002302>.

27 ELENA, S. F.; LENSKI, R. E. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, v. 4, n. 6, p. 457–469, 2003. ISSN 14710056.

28 LENSKI, R. E. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME Journal*, Nature Publishing Group, v. 11, n. 10, p. 2181–2194, 2017. ISSN 17517370. Disponível em: <http://dx.doi.org/10.1038/ismej.2017.69>.

29 FOX, J.; LENSKI, R. From here to eternity—the theory and practice of a really long experiment. *PLoS biology*, v. 13, p. e1002185, 06 2015.

30 WAHL, L. M.; GERRISH, P. J. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution*, v. 55, n. 12, p. 2606–2610, 2001. ISSN 00143820.

31 WISER, M.; RIBECK, N.; LENSKI, R. Long-term dynamics of adaptation in asexual populations. *Science (New York, N.Y.)*, v. 342, 11 2013.

32 DESJARDINS, E. Historicity and experimental evolution. *Biol Philos*, n. 26, p. 339–364, 2011.

33 MANHART, M.; MOROZOV, A. V. Statistical physics of evolutionary trajectories on fitness landscapes. *First-Passage Phenomena and Their Applications*, p. 416–446, 2014.

34 GILLESPIE, J. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 1998. (A Johns Hopkins paperback : Science). ISBN 9780801857553. Disponível em: <https://books.google.com.br/books?id=zb1qAAAAMAAJ>.

35 CROW, J.; KIMURA, M. *An Introduction to Population Genetics Theory*. Burgess Publishing Company, 1970. ISBN 9780808729013. Disponível em: <https://books.google.com.br/books?id=MLETAQAAIAAJ>.

36 PATWA, Z.; WAHL, L. M. The fixation probability of beneficial mutations. *Journal of the Royal Society Interface*, v. 5, n. 28, p. 1279–1289, 2008. ISSN 17425662.

37 DESAI, M. M.; FISHER, D. S. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics*, v. 176, n. 3, p. 1759–1798, 2007. ISSN 00166731.

38 CAMPOS, P. R.; WAHL, L. M. The effects of population bottlenecks on clonal interference, and the adaptation effective population size. *Evolution*, v. 63, n. 4, p. 950–958, 2009. ISSN 00143820.

39 FRANKE, J.; KLöZER, A.; VISSER, J. A. G. M. de; KRUG, J. Evolutionary accessibility of mutational pathways. *PLoS Computational Biology*, Public Library of Science (PLoS), v. 7, n. 8, p. e1002134, Aug 2011. ISSN 1553-7358. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1002134>.

40 POELWIJK, F. J.; TăNASE-NICOLA, S.; KIVIET, D. J.; TANS, S. J. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, v. 272, n. 1, p. 141–144, 2011. ISSN 00225193.

41 SZENDRO, I. G.; FRANKE, J.; De Visser, J. A. G.; KRUG, J. Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences of the United States of America*, v. 110, n. 2, p. 571–576, 2013. ISSN 00278424.

42 DERRIDA, B. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, American Physical Society, v. 24, p. 2613–2626, Sep 1981. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevB.24.2613>.

43 De Oliveira, V. M.; FONTANARI, J. F. Landscape statistics of the p-spin Ising model. *Journal of Physics A: Mathematical and General*, v. 30, n. 24, p. 8445–8457, 1997. ISSN 03054470.

44 CAMPOS, P. R.; ADAMI, C.; WILKE, C. O. Optimal adaptive performance and delocalization in nk fitness landscapes. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 304, n. 3-4, p. 495–506, 2002.

45 CHAO, A.; JOST, L.; HSIEH, T.; MA, K.; SHERWIN, W. B.; ROLLINS, L. A. Expected shannon entropy and shannon differentiation between subpopulations for neutral genes under the finite island model. *PloS one*, Public Library of Science, v. 10, n. 6, p. e0125471, 2015.

46 HILL, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology*, Wiley Online Library, v. 54, n. 2, p. 427–432, 1973.

47 O'CONNELL, A. A.; BORG, I.; GROENEN, P. *Modern Multidimensional Scaling: Theory and Applications*. [S.l.: s.n.], 1999. v. 94. 338 p. ISSN 01621459. ISBN 9780387251509.

48 HOUT, M. C.; PAPESH, M. H.; GOLDINGER, S. D. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, v. 4, n. 1, p. 93–103, 2013. ISSN 19395078.

49 IRWIN, J. On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's type II. *Biometrika*, v. 19, n. 3/4, p. 225–239, 1927.

50 HALL, P. The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, v. 19, n. 3/4, p. 240–245, 1927.

51 GORDO, I.; CAMPOS, P. R. Evolution of clonal populations approaching a fitness peak. *Biology letters*, The Royal Society, v. 9, n. 1, p. 20120239, 2013.

52 CHOU, H.-H.; CHIU, H.-C.; DELANEY, N. F.; SEGRÈ, D.; MARX, C. J. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, American Association for the Advancement of Science, v. 332, n. 6034, p. 1190–1192, 2011.

53 KRYAZHIMSKIY, S.; TKAČIK, G.; PLOTKIN, J. B. The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 106, n. 44, p. 18638–18643, 2009.

54 GOOD, B. H.; MCDONALD, M. J.; BARRICK, J. E.; LENSKI, R. E.; DESAI, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature*, Nature Publishing Group, v. 551, n. 7678, p. 45–50, 2017.

55 REIA, S. M.; CAMPOS, P. R. Analysis of statistical correlations between properties of adaptive walks in fitness landscapes. *Royal Society Open Science*, The Royal Society, v. 7, n. 1, p. 192118, 2020.

56 ORR, H. A. A minimum on the mean number of steps taken in adaptive walks. *Journal of Theoretical Biology*, Elsevier, v. 220, n. 2, p. 241–247, 2003.

57 GILLESPIE, J. H. A simple stochastic gene substitution model. *Theoretical population biology*, Elsevier, v. 23, n. 2, p. 202–215, 1983.

58 J REID J, C. N. L. Repeatability of adaptation in experimental populations of different sizes. *Proc Biol Sci*, 2015.

59 FREITAS, O.; WAHL, L. M.; CAMPOS, P. R. A. Robustness and predictability of evolution in bottlenecked populations. *Phys. Rev. E*, American Physical Society, v. 103, p. 042415, Apr 2021. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.103.042415>.

APPENDIX A – MATRIX ALGEBRA OF MDS

We formalize the algorithm of Section 3.6, used to obtain the coordinate axis of the multidimensional scaling technique. Here, we explicit the contents in the references (48, 47).

Let $\boldsymbol{X}_{n\times m}$ be the matrix of coordinates of points. Each row i of $\boldsymbol{X}$ gives the coordinates of point i on m dimensions, that is, $x_{i1}, x_{i2}, ..., x_{im}$. In MDS we are concerned with the distances among all n points. We can use the matrix algebra to obtain a compact expression for computing the squared Euclidean distances between all points. The squared Euclidean distance is defined by

$$d_{ij}^2(\boldsymbol{X}) = d_{ij}^2 = \sum_{a=1}^{m}(x_{ia} - x_{ja})^2 = \sum_{a=1}^{m}(x_{ia}^2 + x_{ja}^2 - 2x_{ia}x_{ja}) \tag{A.1}$$

Suppose that $\boldsymbol{X}$ contains the coordinates of three points in two dimensions. Now the matrix of squared distances, denoted by $\boldsymbol{D^2(X)}$, is

$$\begin{aligned}\boldsymbol{D}^2(\boldsymbol{X}) &= \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 \\ d_{21}^2 & 0 & d_{23}^2 \\ d_{31}^2 & d_{32}^2 & 0 \end{bmatrix} = \sum_{a=1}^{m} \begin{bmatrix} x_{1a}^2 & x_{1a}^2 & x_{1a}^2 \\ x_{2a}^2 & x_{2a}^2 & x_{2a}^2 \\ x_{3a}^2 & x_{3a}^2 & x_{3a}^2 \end{bmatrix} \\ &+ \sum_{a=1}^{m} \begin{bmatrix} x_{1a}^2 & x_{2a}^2 & x_{3a}^2 \\ x_{1a}^2 & x_{2a}^2 & x_{3a}^2 \\ x_{1a}^2 & x_{2a}^2 & x_{3a}^2 \end{bmatrix} - 2\sum_{a=1}^{m} \begin{bmatrix} x_{1a}x_{1a} & x_{1a}x_{2a} & x_{1a}x_{3a} \\ x_{2a}x_{1a} & x_{2a}x_{2a} & x_{2a}x_{3a} \\ x_{3a}x_{1a} & x_{3a}x_{2a} & x_{3a}x_{3a} \end{bmatrix} \\ &= \boldsymbol{c1' + 1c' - 2\sum_{a=1}^{m} x_a x_a'} = \boldsymbol{c1' + 1c' - 2XX'}\end{aligned} \tag{A.2}$$

where $\boldsymbol{x}_a$ is column a of matrix $\boldsymbol{X}$, $\boldsymbol{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{c}$ is a vector that has elements $\sum_{a=1}^{m} x_{ia}^2$, the diagonal elements of $\boldsymbol{XX'}$. Calling $\boldsymbol{B = XX'}$, we can find the MDS coordinates through its eigendecomposition

$$\boldsymbol{B = Q\Lambda Q' = (Q\Lambda^{\frac{1}{2}})(Q\Lambda^{\frac{1}{2}})' = XX'} \tag{A.3}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{Q}$ are the eigenvalues and eigenvectors of $\boldsymbol{B}$, respectively. Then, the coordinate matrix of classical scaling is given by $\boldsymbol{X = Q\Lambda^{\frac{1}{2}}}$.

Now, if we start with a given matrix of distances $\boldsymbol{D}$, we can find $\boldsymbol{B}$ by we multiplying the left and the right sides of Equation A.2 by the centering matrix $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$ and by the factor $-\frac{1}{2}$:

$$\begin{aligned} -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}^{\mathbf{2}}\boldsymbol{J} &= -\frac{1}{2}\boldsymbol{J}(\boldsymbol{c}\mathbf{1}' + \mathbf{1}\boldsymbol{c}' - \mathbf{2}\boldsymbol{X}\boldsymbol{X}')\boldsymbol{J} \\ &= -\tfrac{1}{2}\boldsymbol{J}\boldsymbol{c}\mathbf{1}'\boldsymbol{J} - \tfrac{1}{2}\boldsymbol{J}\mathbf{1}\boldsymbol{c}'\boldsymbol{J} + \boldsymbol{J}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{J} \\ &= -\tfrac{1}{2}\boldsymbol{J}\boldsymbol{c}\mathbf{0}' - \tfrac{1}{2}\,\mathbf{0}\boldsymbol{c}'\boldsymbol{J} + \boldsymbol{J}\boldsymbol{B}\boldsymbol{J} \\ &= \boldsymbol{B} \end{aligned} \tag{A.4}$$

The first two terms are zero, because centering a vector of ones yields a vector of zeros. Since relative distances do not change under translations, we assume that **X** has column means equal to 0. Thus, centering around **B** can be removed because **X** is column centered, and hence so is **B** (48). The method of multidimensional scaling only differs from this procedure in that the matrix of squared distances $\boldsymbol{D}^{\mathbf{2}}$ is replaced by the squared dissimilarities $\boldsymbol{\Delta}^{\mathbf{2}}$ (47).

APPENDIX B – HILL NUMBER DERIVATION

For consistency, we expose in this appendix the derivation of the Hill number D_1, of Section 3.4. What follows is nothing more than a reproduction of what can be found in the original Hill's paper (46), and it is presented in this dissertation by seeking its completeness.

Given an event with probabilities $p_0, p_1, \ldots, p_i$, where $\sum_i p_i = 1$, we have

$$S = -\sum_i p_i \, ln \, p_i, \tag{B.1}$$

as its the respective Shannon entropy. And, as defined by Hill:

$$D_a = \Big(\sum p_i^a \Big)^{\frac{1}{1-a}} \quad \text{and} \quad D_1 = \exp S. \tag{B.2}$$

So, its left to show that, in the limit $D_1 = \lim_{a \to 1} D_a$, we should obtain that

$$\lim_{a \to 1} \Big(\sum p_i^a \Big)^{\frac{1}{1-a}} = \exp \Bigg[-\sum_i p_i \, ln \, p_i \Bigg]. \tag{B.3}$$

By taking the logarithm of both sides and making the substition $a = b + 1$, we have that

$$\begin{aligned} -\sum_i p_i \, ln \, p_i &= \lim_{a \to 1} \frac{1}{1-a} \, ln \Big(\sum_i p_i^a \Big) \\ &= -\lim_{b \to 0} \frac{1}{b} \, ln \Big(\sum_i p_i^{b+1} \Big) \\ &= -\lim_{b \to 0} \frac{1}{b} \, ln \Big(\sum_i p_i^b \, p_i \Big) \\ &= -\lim_{b \to 0} \frac{1}{b} \, ln \Big(\sum_i p_i^b \, \exp(b \, ln \, p_i) \Big). \end{aligned} \tag{B.4}$$

Since b is small, we can expand the exponential to obtain

$$\sum_i p_i \, ln \, p_i = \lim_{b \to 0} \frac{1}{b} \, ln \Big(\sum_i p_i \Big(1 + b \, ln \, p_i \Big) \Big). \tag{B.5}$$

Finally, remembering that $\sum_i p_i = 1$, we can expand the resulted logarithm, thus

$$\begin{aligned} \lim_{b \to 0} \frac{1}{b} \, ln \Big(\sum_i p_i + b \sum_i p_i \, ln \, p_i \Big) &= \lim_{b \to 0} \frac{1}{b} \, ln \, \Big(1 + b \sum_i p_i \, ln \, p_i \Big) \\ &= \sum_i p_i \, ln \, p_i, \qquad QED. \end{aligned} \tag{B.6}$$

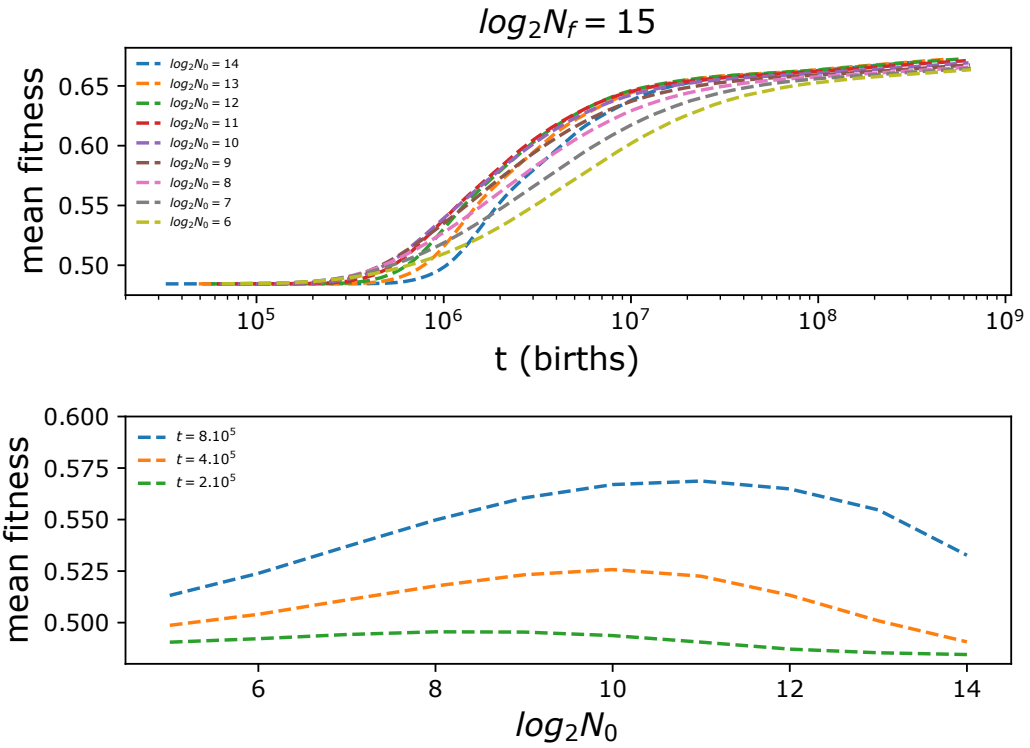
APPENDIX C – COMPLEMENTARY RESULTS

For their complementary nature, we separate some results from the main body of Section 4.2.1. Here we expose the results reported for the adaptation rate in units of births, for intermediate genome's length $L = 12$, and when N_0 is fixed with varying N_f.

TIME MEASURED IN UNITS OF BIRTHS

As stressed along the work, the adaptation rate depends on the definition of its time units. When time is measured in units of births, we obtain the similar scenario shown on the left in Figures 26 and 28, and once again the highest adaptation rates are found at intermediate bottleneck sizes.

Figure 35 – Fitness trajectories for different bottleneck ratios (upper panel) and fitness at fixed times for various bottleneck sizes (lower panel). Time is measured in units of births. The parameter values are $N_f = 32768$, mutation rate $\mu = 10^{-4}$, sequence size $L = 8$ and epistasis and $K = 2$. The bottleneck sizes and times at which fitness are recorded are indicated in the legends.

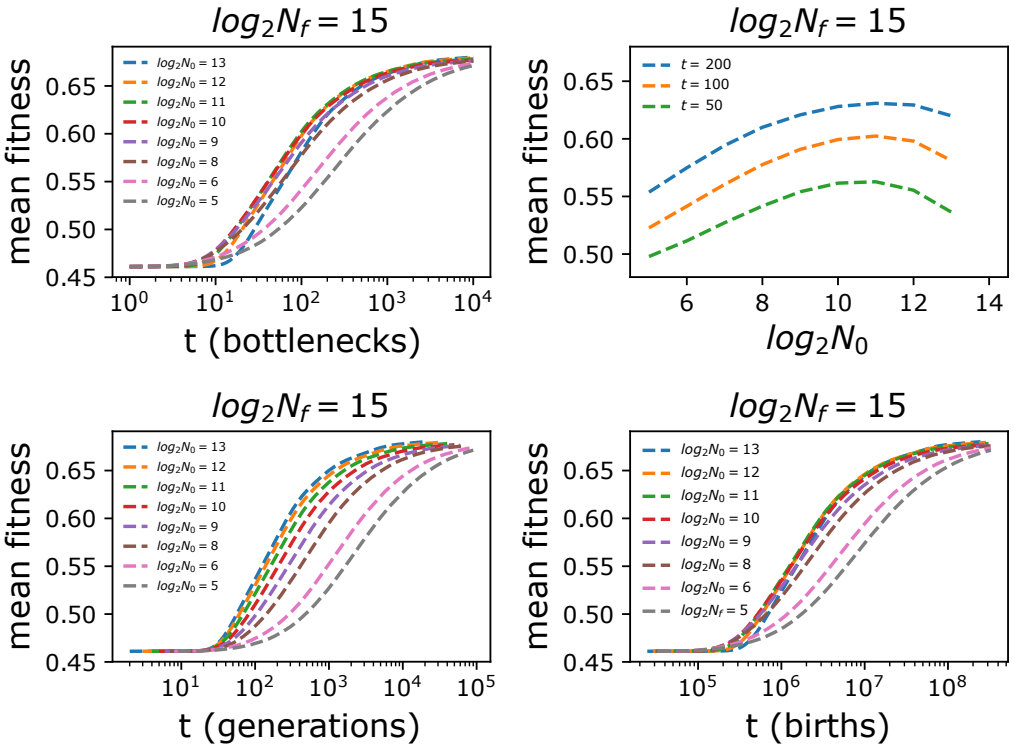


Source: The author (2020).

RESULTS FOR $L = 12$

At the first steps of our work, we investigated if the pattern of adaptation rates was invariant over the adopted parameters. In Figure 36, the genome length is set at $L = 12$, an intermediate parameter compared with the results in 26 and 28. As we can see, the pattern is maintained in along all time units.

Figure 36 – Fitness trajectories for different bottleneck ratios. The data correspond to an average over 50 fitness landscapes. The parameter values are $N_f = 32768$, mutation rate $\mu = 10^{-4}$, sequence size $L = 12$ and epistasis parameter $K = 2$. The bottleneck sizes are indicated in the legends. Time is measured in units of bottlenecks (upper panel), generations and births (lower panels). In the right upper panel the curves correspond to fitness reported at different times.



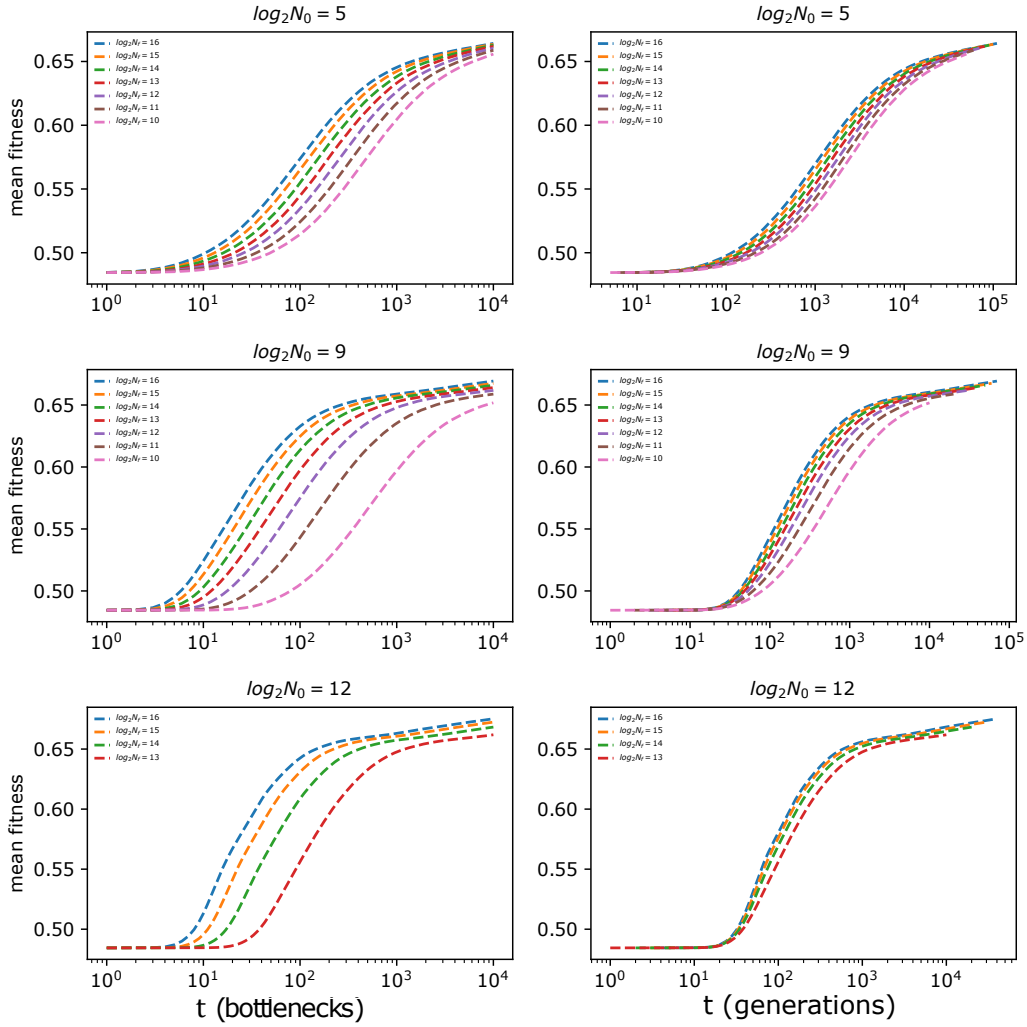
Source: The author (2020).

FIXED N_0 WHILE CHANGING N_f

In the results of Section 4, we changed the bottleneck ratio by varying N_0, while holding N_f fixed. Analogous results, with N_0 constant while N_f varies, are shown in Figure 37. Whether time is measured in units of bottlenecks and generations, the mean population fitness displays a monotonic increase with N_f in this case. Overall, this further corroborates our conclusions that when time is measured in generations, the adaptation rate increases monotonically in

larger populations, irrespective of population bottlenecks.

Figure 37 – Fitness trajectories for fixed bottleneck sizes N_0. The different curves correspond to distinct values of N_f, as indicated in the legends. Time is measured in units of bottlenecks (left panels) and doublings (right panels). The other parameter values are mutation rate $U = 1 \times 10^{-4}$, sequence size $L = 8$ and epistasis parameter $K = 2$.



Source: The author (2020).