2019

Master 2 - Mention Informatique Parcours Bioinformatique et Modélisation

Supervisor :

Yitzhak PILPEL

Weizmann Institute of Science, Israel

Influence of mutation rate variability on adaptive process on different fitness landscapes

SORBONNE UNIVERSITÉ, FRANCE

GABRIELA LOBINSKA (3770406)

Acknowledgements

I am grateful to Prof. Yitzhak Pilpel for welcoming me in his research group, supervising my project and his patience and guidance throughout my stay.

I would also like to thank the whole research team for creating a friendly and welcoming atmosphere.

Table of Contents

1 Introduction 3
1.1 Darwin's theory 3
1.2 Fitness Landscapes 3
1.2.1 Definition 3
1.2.2 Biological fitness landscapes 3
1.2.3 Artificial fitness landscapes 3
1.3 Evidence for mutation rate heterogeneity 3
1.3.1 Mutation rate inheritance 4
1.3.2 Mutation rate heterogeneity within the population 4
1.3.3 Mutation rate variability in nature 4
1.4 Current work on mutation rate variation patterns 4
1.4.1 Evolutionary Biology studies 4
1.4.2 Parallel problems in computer science 4
1.5 Description of the laboratory 5
1.6 Work Performed 5
2 Methods 6
2.1 Mutation rate variation patterns 6
2.2 Fitness Landscapes 6
2.2.1 Step Fitness Landscapes 7
2.2.2 Transcription Factor binding sites fitness landscapes 8
2.2.3 Protein Fitness Landscapes 9
2.3 Analytical Approach 11
2.3.1 Fixed mutation rate 11
2.3.2 Randomly variable mutation rate 12
2.3.3 Inheritable mutation rate 13
2.3.4 Inheritable mutation rate with noise 13
2.3.5 Computational Implementation 13
2.4 Simulation Approach 14

2.5 Parameters 15
2.5.1 Mutation rate level 15
2.5.2 Distribution of mutation rate level 15
2.5.3 Initial genotype 16
2.5.4 Number of mutations per cell division 16
2.5.5 Other parameters 16
3 Results 17
3.1 Step Landscapes 17
3.1.1 Simulation Approach 17
3.1.2 Analytical Approach 19
3.2 Transcription Factor Binding Sites Landscapes 21
3.2.1 Simulation approach 21
3.2.2 Analytical Approach 22
3.3 Artificial Protein Landscape 23
3.3.1 Simulation Approach 23
3.3.2 Analytical Approach 25
4 Discussion 27
4.1 Enhanced performance of the noisily inheritable mutation rate and link with the multi-armed bandit problem 27
4.1.1 Correlation between mean population mutation rate and mean population fitness 28
4.1.2 Population heterogeneity along evolutionary time 29
4.2 Differences in behaviour of the four mutation rate strategies depending on the landscapes 30
4.3 Future directions 30
5 Bibliography 31
6 Supplementary Materials 33

1 INTRODUCTION

1.1 DARWIN'S THEORY

According to the Darwinian theory, evolution is driven by randomly occurring, rare changes in the genetic material. A small minority of those mutations are beneficial and confer a selective advantage to their bearers. The genotype frequencies in the population are thus shifted in favour of adaptive genotypes over evolutionary time. The capacity of a population to efficiently explore the available phenotypes and evolve towards the optimal one is known as evolvability [1].

The phenotype is the result of the interaction between the environment and the organism's genome.

1.2 FITNESS LANDSCAPES

1.2.1 Definition

A fitness landscape is a function which maps a genotype to a phenotype. Fitness landscapes can be plotted on a genotype map, where mutational neighbours are placed next to each other. This leads to various geometric patterns which are often described using the metaphor of a landscape, describing local fitness maxima as 'peaks' and local fitness minima as 'valleys' [2].

1.2.2 Biological fitness landscapes

Biological fitness landscapes are landscapes which usually map a genetic or amino acid sequence to an expected fitness. Mapping biological landscapes is often not achievable due to their extremely high combinatorial space. Indeed, the alphabet size for biological sequences is either 4 (for genetic sequences) or 20 (for amino acid sequences), which leads to either 4^N or 20^N possible sequences. These quantities of genotypes quickly become impossible to survey as N increases, even if it remains small.

To bypass the combinatorial problem, studies have either focused on small functional sequences, such as transcription factor binding site sequences [3], or on immediate neighbours of wild-type sequences [4], [5]. The recent introduction of high throughput technologies, such as protein binding microarrays, have made it possible to assign each sequence to a phenotype, for example binding affinity, which can then serve as an approximate for fitness [6]. Nevertheless, the number of published fitness landscapes remains low.

1.2.3 Artificial fitness landscapes

Since empirical studies are sparse, it is often difficult to assess which of the models for generating fitness landscapes are most representative of the biological reality. The most used one in biological simulations is the NK-model [7], [8]. In the NK-model, the K value determines the extent of epistasis between the positions of the artificial genome. This in turns regulates the ruggedness of the landscape. Epistasis, and by extension ruggedness, has been found to be a defining characteristic of the fitness landscape [9]. This is not surprising since nucleotides or amino acids rarely perform a function on their own, but rather through interactions with other nucleotides or amino acids in the sequence.

It is also possible to generate landscapes based on partial landscapes present in the literature. One approach is to create computationally a complete landscape with a smaller number of genotypes which imitate some of the properties known to be specific to fitness landscapes, such as variable ruggedness. Other studies have used machine learning methods to predict the fitness of genotypes missing from empirical screens [10].

1.3 EVIDENCE FOR MUTATION RATE HETEROGENEITY

Populations navigate the fitness landscapes by mutating their genotypes. The optimal mutation rate reflects a balance between a high extremum and a low extremum and its exact level has been the subject of debate within the scientific community [11]–[20]. A disproportionately high mutation rate leads to the accumulation of deleterious mutations at a

faster pace than natural selection can eliminate them. On the other hand, a very low mutation rate prevents the generation of even the beneficial mutations and stalls the adaptive process. However, those studies have considered the mutation rate as a fixed parameter: no mutation rate inheritance between parent and offspring, which means no mutation rate variation in time; and no mutation rate heterogeneity within the population.

1.3.1 Mutation rate inheritance

Mutation rate depends on the sequence of the DNA polymerase, on the sequence of DNA repair enzymes, the regulatory genetic elements that determine their transcription levels, and many other factors which are directly inheritable through inheritance of the genetic sequence. Therefore, mutation rate differences and evolution between parent and offspring are expected.

A recent paper by [21] has shown that mutation rate could even be inherited epigenetically. Ada is a DNA repair enzyme which is present at very low concentrations in *E. coli*. It is also autoregulated. Therefore, it needs to be present of the cell to trigger its own production. The offspring's mutation rate will therefore be dependent on the number of Ada copies transmitted by their parent cell.

1.3.2 Mutation rate heterogeneity within the population

Since mutation rate is inheritable, it follows that it will be heterogenous within the population since not all offspring will inherit the same mutation rate from their parents.
Stochastic fluctuations of DNA repair enzymes, as well as any other protein involved in mutation repair, can lead to mutation heterogeneity within the population. Protein binding to transcription factor binding sites or to each other to form complexes is also a stochastic process and also result in mutation rate variance within the population.

1.3.3 Mutation rate variability in nature

Variation in mutation rate has been observed in nature [22]. In addition to the Ada protein example, mentioned above, examples of stress induced mutagenesis and hyper-mutators have already been extensively described [23]–[26].
Therefore, the mutation rate should not be considered a fixed parameter and would be more accurately described by a value sampled from a probability distribution.

1.4 CURRENT WORK ON MUTATION RATE VARIATION PATTERNS

1.4.1 Evolutionary Biology studies

Natural variability of the mutation rate has already been suggested to lead to different evolutionary dynamics, in particular increased population fitness when compared to studies of fixed mutation rates. One reason for this is that variability within the population generates subpopulations of organisms with higher than average and lower than average mutation rates. Depending on whether it is more advantageous to change or maintain the currently prevailing genotype, these subpopulations would display a selective advantage which could increase the mean population fitness. A small proportion of the population with higher mutation rates can also lead to faster multi-locus adaptation and generally more thorough exploration of the sequence space. Inheritable mutation rate favours the accumulation of mutations when multi-locus adaptation is required through maintaining a subpopulation with slightly higher mutation rate, and conversely favours the preservation of the genotype when the fitness optimum is already achieved. [22]

1.4.2 Parallel problems in computer science

The problem studied in this project could be seen as a variation of the multi-armed bandit problem in computer science. In this problem, the agent is faced with a set of choices, each one of them resulting in a gain sampled from a probability distribution specific to that choice. The agent needs to maximise its gain by making the optimal choice at each time point. At each time point, the agent can make a decision based on the memories of the choices it has already made in the past ('exploitation') or make a decision it has not made before, on the

gamble that it could bring him more gain than its previous decisions ('exploration'). The optimal balance between those strategies has been the subject of many studies and is generally formalised as the probability of exploring a new strategy ε [27]. In biology, the agents are cells from the population, that need to decide whether to mutate to increase their fitness ('exploration') or keep their genotype ('exploitation'). The mutation rate corresponds to the ε-factor from the multi-armed bandit, and the fitness corresponds to the gain.

However, biological evolution differs from the multi-armed bandit problem in that the mutation rate (or ε-factor) evolves through repeated cycles of random adjustment and then selection of the best performing agents, as opposed to a decision taken by the agent. Nevertheless, insights from this problem might prove valuable for understanding adaptation.

1.5 DESCRIPTION OF THE LABORATORY

The Pilpel laboratory is part of the Weizmann Institute of Science, in Rehovot, Israel. It is an evolutionary biology laboratory which is organised around two main axes: the study of translation and the study of evolvability, which is the axis most relevant to the present thesis. An important project of the laboratory has been the organisation of the Evolthon competition, in which participants needed to adapt a provided yeast strain to an environmental challenge (low temperature). Each of the participants adopted a slightly different strategy which allowed to study the best methods to promote evolvability [28].

The laboratory also studied non-genetic mechanisms which aid evolvability. In particular, it was the first research group to demonstrate RNA editing of protein coding sequences in bacteria [29]. It also investigated the mechanistic process mistakes in translation in E.coli and in yeast [30].

Ongoing projects related to the study of evolvability, and more closely related to the present study, include:

- The analysis of tRNA landscapes in order to understand why the preponderant tRNA genotype does not correspond to the fitness landscape optimum. In reality, it corresponds to the 'flattest' sequence, which is the sequence whose fitness is the most robust to mutations.
- Simulation of cell populations in order to explore whether the numbers of mutational neighbours of a sequence in the fitness landscape is evolutionarily advantageous
- Search for 'evolvability' genes – genes which, when present in yeast, result in higher rates of evolvability
-

1.6 WORK PERFORMED

In this project, we focused on four mutation rate variation patterns:

- **Fixed mutation rate**, which represents the most commonly used model for mutation rate in the current literature
- **Randomly variable mutation rate**, which models stochastic fluctuations in DNA repair proteins concentrations and fluctuations in DNA polymerase efficiency
- **'Perfectly' inheritable mutation rate,** which models the inheritance of genetic sequences coding for DNA polymerases and DNA repair proteins
- **'Noisily' inheritable mutation rate,** which models the epigenetic inheritance of factors impacting the mutation rate, such as DNA repair protein concentrations

We studied the evolution of the adaptive process on three different landscapes: the 'steps' landscapes, which is a set of simplified landscapes which allow us to study all possible trajectories from the initial genotype to a target genotype; transcription factor binding site fitness landscapes, taken from [3]; and an artificial protein fitness landscape, which we modelled from information provided by [4] and [31]. This allowed us to contrast the influence of the four mutation rate variation patterns on the adaptive process in different landscape geometries.

We used two approaches: an analytical, deterministic model, and a computational simulation. Those two approaches each come with their advantages and disadvantages and gave complementary insights.

We found that the noisily inheritable mutation rate performs better than the other three strategies in terms of achieving the target genotype and mean population fitness. This is consistent with existing literature on the optimal exploration/exploitation setting in the multi-armed bandit problem.

The analytical model has shown that the mutation variation patterns involving inheritance have the potential of being stuck in local optima. However, they achieve high mean population fitness by leading to a reduction of the mean population mutation rate once an optimum is reached.

2 METHODS

The analytical approach considers the population as infinite. It therefore avoids stochastic noise inherently present when studying a finite population. However, taking into account genetic drift models more closely biological reality. The analytical approach is less computationally expensive, but the exploration of all possible genotypes can be prohibitive for larger fitness landscapes.

2.1 MUTATION RATE VARIATION PATTERNS

We investigated four mutation rate variation patterns:

- The **fixed mutation rate**, which represents the default mutation rate variation pattern used in theoretical studies [11]–[20]. Under this pattern, the mutation rate is set as a constant parameter and does not vary between individuals in the population or between generations.
- The **randomly variable mutation rate**, which models random fluctuations in DNA repair enzymes concentrations as well as other stochastic events which can affect the mutation rate, without memory between the parent and the offspring: each cell samples its mutation rate from a probabilistic distribution independently from its parent cell.
- The **perfectly inheritable mutation rate**, which models the inheritance of biological sequences that code for any protein that can affect the mutation rate. There is variation between individuals in the population at the initial time point, but each cell then inherits its parent's mutation rate.
- The **noisily inheritable mutation rate**, which models epigenetic inheritance, that is inheritance through any other means than the transmittance of genetic sequences. The offspring's mutation rate is sampled from a probability distribution whose parameters are dependent on the parent's mutation rate. This allows for evolution of the mutation rate along generations.

We show a schematic representation of each mutation rate pattern as implemented in the simulation approach in 2.4.

2.2 FITNESS LANDSCAPES

The choice of a fitness landscape is crucial for studying the progression of the evolutionary process. The geometry of the fitness landscape around the initial genotype can determine the probability of events such as clonal interference or getting stuck in a local fitness optimum.

In this study, we first worked with a simplified 'step landscape' which allowed us to study the influence of the neighbourhood of the initial genotype on the evolutionary process for each of the four mutation rate variation patterns.

Then, we moved to more biologically realistic mutational landscapes: first; we used landscapes of transcription factor binding sites established by [3]. Then, we attempted to construct a protein fitness landscape based on studies by [4] and [31].

Studying both a DNA binding site and a protein fitness landscape was important since the relationship between fitness and genotype is very different in those two cases. For DNA binding sites, the function depends on the physical shape of the binding site sequence. This explains why similar sequences have similar binding affinities, which leads to "small world" networks described by [3].

Protein sequences are different since their function mainly depends on their correct folding during translation and possible allosteric changes for enzymatic activities. Therefore, even very similar sequences can have vastly different fitness if the mutated residues are necessary for acquiring and maintaining the correct structure. On the other hand, very different structures can have similar fitness if the residues through which they differ are structural and with similar physicochemical properties to that of the wild-type sequence. This leads to a 'threshold model' proposed by [4] where the protein fitness tolerates some mutations before abruptly decreasing if the changes are such that the protein overall stability is compromised.

2.2.1 Step Fitness Landscapes

We designed simplified landscapes in order to study the effect of the geometry of the landscape on the performance of each strategy. In biological conditions, the ensemble of possible fitness landscapes is infinite. We applied several simplifications to this ensemble in order to obtain a set of landscapes that would be both possible to explore computationally and representative of all possible landscape geometries.

The first simplification was to consider the landscape as a linear vector, with each genotype having only two neighbours instead of the 3^n neighbour genotypes (with n the length of the genome) possible neighbours it would have under real biological conditions. In order to avoid "sink" genotypes, the landscape "wraps around" itself, with the first genotype of the sequence being the neighbour of the last genotype of the sequence.

The second simplification was to discretise the space of possible fitness values. In biological conditions, the fitness value of each genotype can take any value. Here, we fix a parameter k, the set of possible fitness values being all strictly positive integers lower than or equal to k+1. Moreover, the target genotype has the strictly maximal fitness value, which is equal to k+1.

The fitness landscape is then constructed as a collage of three parts: an initial genotype; a target genotype; and a sequence of genotypes, a 'gap' which separate them. In this case, we are only interested in studying one path leading from the initial genotype to the target genotype at a time. This requires the path leading from the initial genotype to the target genotype to be identical regardless of which of its two neighbours is explored first. Thus, the 'gap' sequence needs to be symmetrical around the initial genotype.

All these considerations led us to establish the following method to generate the simplified landscapes for a given parameter k:

1) Generate all k^k possible combinations of integers between 1 and k included. Those will be the 'gap' sequences.
2) For each possible fitness value of the initial genotype (integers between 1 and k included), set the initial genotype as the central value; then for each 'gap' sequence, add the sequence from both sides of the initial genotype so that it is a symmetry axis
3) Complete the landscape by adding the target genotype with a fitness value of k+1

For this simulation, we used k = 3 which gave us 81 landscapes with fitness values between 1 and 4. A schematic representation of the building of a landscape is shown in Figure 1. 81 is the largest number of landscapes which we can study while still being able to easily visualise the results. The generated landscapes are shown in Figure 2.

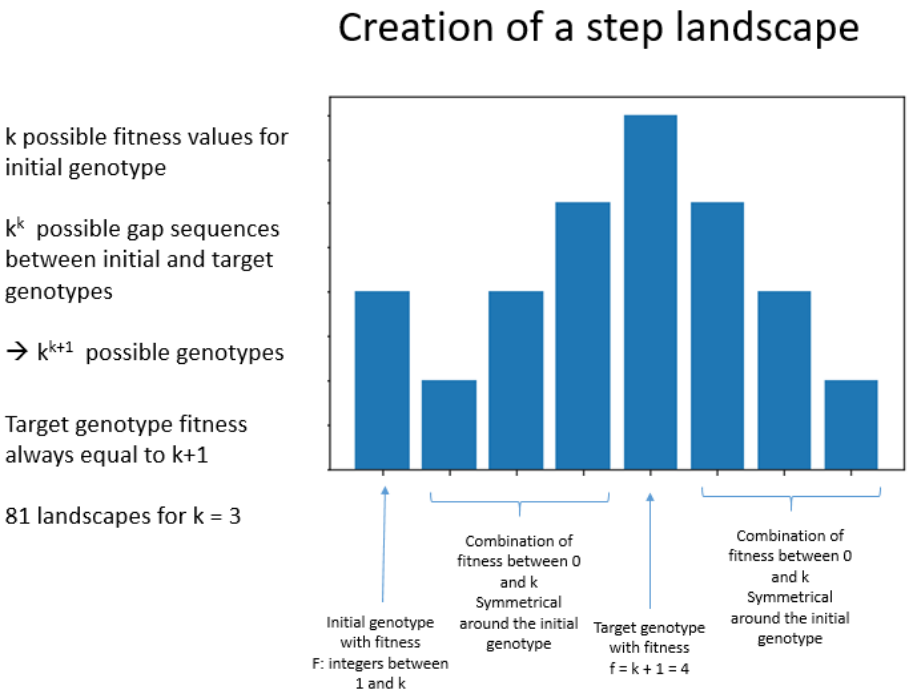


Figure 1: The composition of a step landscape, for k = 3. This method of generating simplified landscapes allows for the exhaustive study of all possible trajectories between the initial and the target genotype.

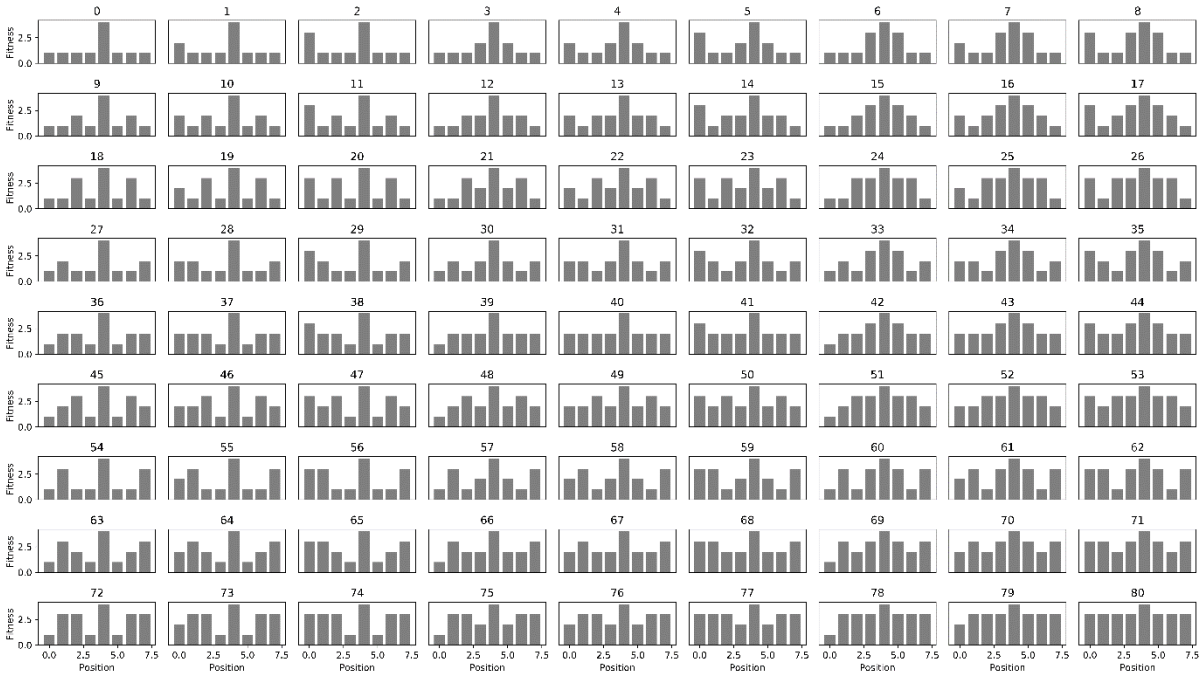


Figure 2: All step landscapes generated for k = 3. A k-parameter at 3 resulted in 81 landscapes, which is a number sufficiently small to be easily visualisable.

2.2.2 Transcription Factor binding sites fitness landscapes

Transcription binding sites are short in sequence (6-8 nucleotides) and therefore it is possible to survey all possible binding sites experimentally. In particular, [6] have analyzed 1180 transcription factor binding sites from species spanning all three kingdoms of life.

The raw data used by [6] consisted of binding affinities of a transcription factor to each of the 32,896 possible binding sequence motifs of length 8, as measured by protein binding microarrays. 1180 transcription factors were surveyed, of which we eliminated 13 because

they bound less than 20 sequences, which we considered too low to construct a landscape. For each transcription factor, they first asked which of the sequences bound specifically to the transcription factor by applying a cutoff to the measured binding affinities. This eliminated most sequences for all transcription factors: the maximal number of sequences retained for a transcription factor was 2361. In general, there was a strong preference towards a small (< 500) number of retained sequences. The distribution of the numbers of sequences with binding phenotype for all transcription factors are shown in Figure 3.

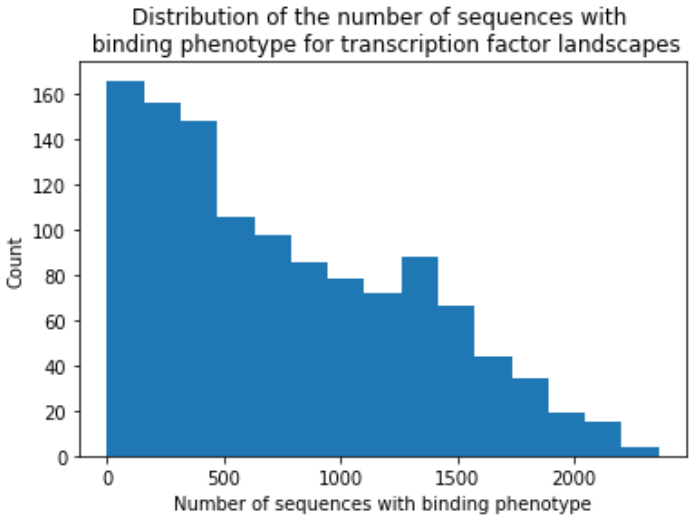


Figure 3: The distribution of the number of sequences with a binding phenotype for all landscapes analyzed by [3]. Most of the fitness landscapes contained a very small proportion of all possible sequences.

Next, [6] constructed networks from the remaining sequences. Nodes were sequences that were linked by an edge if they differed by one base pair.
In this study, we made the strong assumption that the fitness of each sequence could be approximated by its binding affinity to the transcription factor. The sequences that did not have a binding phenotype were assigned a fitness equal to $10^{-50.}$ For each landscape, we set the initial sequence to be the sequence with the minimal fitness that still exhibited a binding phenotype.

2.2.3 Protein Fitness Landscapes

Protein landscapes are much trickier to obtain due to their larger sequence size as well as the larger alphabet – 20 possibilities per site instead of 4 per site for DNA sequences. Therefore, the fitness landscape studies have focused on small subsequences of the protein which are inherently biased since they are all neighbors of the wild-type sequence.
Nevertheless, some attempts have been made to characterize protein fitness landscapes. Computational machine learning methods have been suggested in order to characterize the fitness landscape by inferring the fitness from sequences for a number of sequences, neighbors of the wild-type sequence or chimeric constructions from known proteins [32], [33]. Statistical methods for predicting the fitness of artificial sequences have given promising results. Experimental methods have tried to characterize the fitness landscape by focusing on positions in the active site [34] or by generating a large amount of random mutations around the wild-type sequence [4], [5].

We used two studies to create a landscape as similar as possible to the real GFP fitness landscape.

2.2.3.1 Number of genotypes

When constructing our fitness landscape, we assumed that each genotype has only 5 possible neighbors instead of the 20**N neighbors of a real amino acid sequence of length N.

This was necessary since it is not possible to construct a landscape with this number of genotypes.

First, we set an initial genotype as a first, central genotype. Then, we progressively added neighbor genotypes: 5 1-neighbours, each of whom had 5 neighbors, which where thus 2-neighbors of the initial genotype, etc. until reaching 4 neighbors of the initial genotype.
This resulted in 781 genotypes in that step, which corresponds to the sum:

$$\#\ genotypes = \sum_{n=0}^{4} 5^n$$

A decision which we had to take was whether k-neighbors of the same order were neighbors of each other, and how many genotypes from the k+1-neighbour set were neighbors of a given genotype from the k-neighbor set.

In biological conditions, this depends on the length of the amino acid sequence: for short sequences, k-neighbors are different variations of amino acids at a small number of sites. Therefore, two sequences will able to mutate into each other with a small number of mutations. For long sequences, it is very improbable that two k-neighbor sequences will differ in the same k sites and will need a higher number of mutations to mutate into each other.

Here, we chose to assume in a first approach that the sequence is long enough that all k+1-neighbors have only one k-neighbor, and that none of the k-neighbors are neighbors of each other.

2.2.3.2 Setting the fitness of each genotype

The study by [4] revealed the large role of negative epistasis in the GFP fitness landscape. The wild-type sequence was decreasingly robust to mutations as their number increased. This led [4] to suggest a threshold model where each mutation leads to a decrease in protein stability, until it takes it under a threshold which then results in the complete inactivation of the protein. The authors provided fluorescence distributions for each of the surveyed k-neighbor for k ranging from 1 to 11 as well as the proportion of non-fluorescent proteins.

For each k-neighbor set, we set approximatively the same proportion of genotypes to a null fitness as was observed by [4]. It was not possible to recreate exactly the same proportion, since when removing some nodes at random some trajectories in the landscape are eliminated, thus removing more nodes. The number of genotypes with a fitness of 0 for each k-neighbor set is detailed in Table A.

k	Number of neighbours	Theoretical Number with phenotype	Real number with phenotype	Theoretical proportion with phenotype	Real proportion with phenotype
1	5	5	5	0.93	1.
2	25	23	19	0.91	0.76
3	125	99	91	0.80	0.72
4	625	399	324	0.64	0.51

Table A: Number of effective neighbors compared to theoretical neighbors with phenotype in the GFP landscape.

2.2.3.3 Setting the adaptive genotype

We assumed that the landscape for red fluorescent protein was similar to that of the green fluorescent protein.

We duplicated the landscape, conserving the same geometry. We then linked the two landscapes by making 5 random 4-neighbours of each landscape neighbors of each other. [31] found that a small number of mutations was necessary to perform the switch from the green fluorescent to the red fluorescent protein. They observed that only a few from the eleven

mutations were necessary to recapitulate the evolution from green to red fluorescence and the rest of the mutations are 'fine-tuning' mutations which enhance the newly evolved phenotype. This was coherent with our strategy: indeed, our landscape models the need of a small number of mutations, and then additional 'fine-tuning' mutations to go from the 4-neighbours of the red fluorescent genotype to the central node of the red fluorescent network.

Lastly, to model the fact that the red fluorescent protein is adaptive, we multiplied the fitness in one of the halves of the landscapes to be 110% the fitness of the corresponding genotype in the other half.

A schematic representation of the landscape is shown in Figure 4.

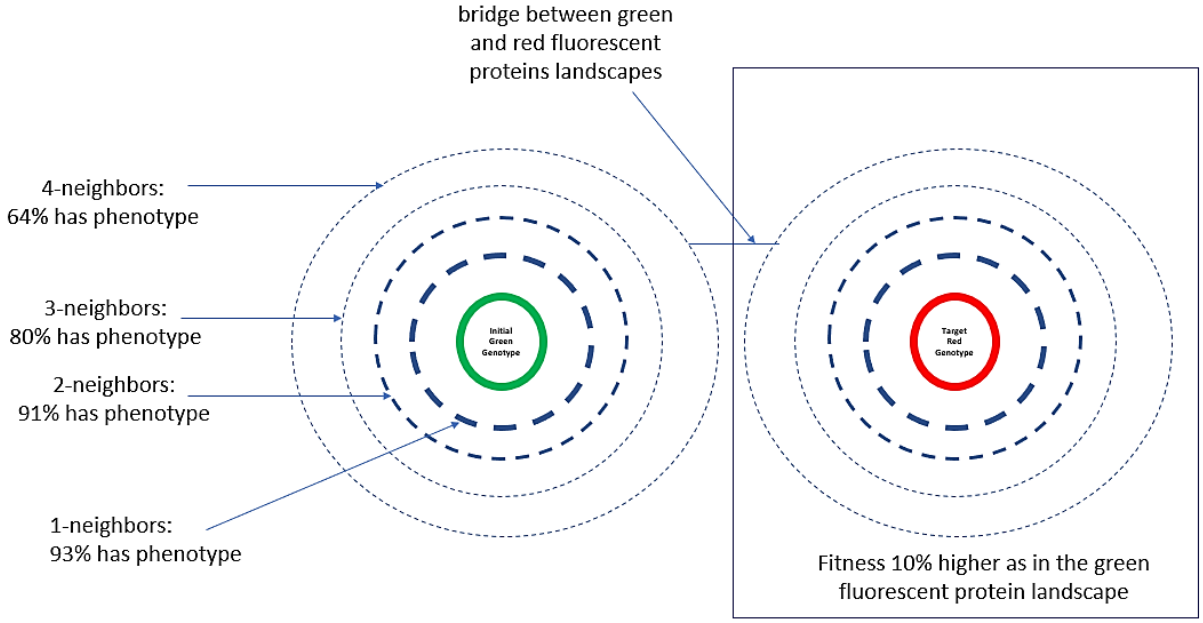


Figure 4: Schematic representation of the artificial GFP-like landscape. The two local peaks corresponding to the green fluorescent phenotype, and the red fluorescent phenotype, surrounded by genotypes of similar fitness and some of null fitness, as predicted by the threshold model. Having a phenotype means having a non-zero fitness. As established by [4], the loss of phenotype in the GFP case is sudden and increasingly likely with as the number of mutations from the wild-type sequence increases.

2.3 Analytical Approach

2.3.1 Fixed mutation rate

In the analytical approach, we model an infinite population of well-mixed, asexually reproducing cells on a given fitness landscape with k genotypes, and for a given mutation rate μ, with a system of equations. Each one of these equations corresponds to the frequency of one genotype. We perform two steps: a 'mutation' step, then a 'selection' step.

At each time point, the frequency of a genotype is composed of an incoming flux, corresponding to the cells that mutated to that genotype, and a negative, outgoing flux, corresponding to the cells that mutated from that genotype to a different genotype.

For a genotype g, the incoming flux is the sum of the frequencies of all neighbors of g multiplied by the probability that they will mutate to g. This probability is equal to the mutation rate μ to the power of their distance from g, from 1 to the maximal distance I. We also need to divide it by the number of other neighbors of the genotype which would mutate to g. Mathematically, we write:

$$incoming_g^{t+1} = \sum_i^I \frac{\mu^i}{\#N(\eta_g^i)} \sum_{\eta_g^i} X_{\eta_g^i}^t$$

In this equation, μ is the mutation rate, a parameter; η_g^i are the neighbors of g at distance i; and $X_{\eta_g^i}^t$ is the frequency of η_g^i at time t. $\#N(\eta_g^i)$ is the number of neighbors of a neighbor η_g^i.
For a genotype g, the outgoing flux is the negative sum of the frequency of genotype g multiplied by the probability that it will mutate to a neighbor genotype, from distance 1 to maximal distance I.

$$outgoing_g^{t+1} = \sum_i^I \mu^i * X_g^t$$

In this equation, μ is the mutation rate, and X_g^t is the frequency of genotype g at time t.

The final equation which gives the frequency of a genotype g at time t+1, given the frequencies of all genotypes at time t is therefore:

$$X_g^{t+1} = X_g^t + incoming_g^{t+1} - outgoing_g^{t+1}$$

Once we have calculated the above equation for each of the possible genotypes given the landscape, we have completed the 'mutation' step of the analytical approach. We now perform the 'selection' step, in which we multiply each frequency by its fitness, and normalize the population size to 1. We calculate this for each genotype, until convergence, that is no difference between all the values at time t to time t+1.

2.3.2 Randomly variable mutation rate

For variable mutation rate, we subdivide the population into four categories, corresponding to mutation rates μ_1, μ_2, μ_3, μ_4. We chose to consider four categories because it is the minimal number of categories to model a binomial distribution of mutation rates. Indeed, if we set $\mu_2 = \mu_3$, we can set the values of μ_i in such a way that at the initial state, half of the population will have a given mutation rate, a quarter a mutation rate inferior by a standard deviation, and a quarter a mutation rate superior by a standard deviation.

The basic reasoning behind the equations presented above remains the same. However, we now have four times as many equations since each genotype is now subdivided into cells with each of the possible mutation rates.

The equation for the incoming factor, for a genotype g and mutation rate m, is the sum of the contributions of the neighbors of g, but now it also takes into account variable contributions from neighbors of different mutation rates.

$$incoming_{g,m}^{t+1} = \sum_m \sum_i^I \frac{\mu_m^i}{\#N(\eta_g^i)} \sum_{\eta_{g,m}^i} X_{\eta_{g,m}^i}^t$$

In this equation, μ is the mutation rate, a parameter; η_g^i are the neighbors of g at distance i; and $X_{\eta_g^i}^t$ is the frequency of η_g^i at time t. $\#N(\eta_g^i)$ is the number of neighbors of a neighbor η_g^i.

The equation for the outgoing factor is almost identical except for an added dependence on the mutation rate category m:

$$outgoing_{g,m}^{t+1} = \sum_i^I \mu_m^i * X_{g,m}^t$$

The final equation is still:

$$X_{g,m}^{t+1} = X_{g,m}^t + incoming_{g,m}^{t+1} - outgoing_{g,m}^{t+1}$$

Since for this mutation rate variation pattern there is no memory of the mutation rate category for the new generations, for each time t, before proceeding with the 'selection' step, we sum X_g^{t+1} for all four m, which we then redistribute equally into each $X_{g,m}^{t+1}$. Mathematically:

$$X_{g,m}^{t+1} = \frac{1}{4} * \sum_{M} X_{g,M}^{t+1}$$

For the 'selection' step, we proceed as with the fixed mutation rate variation pattern.

2.3.3 Inheritable mutation rate

For the inheritable mutation rate, we still have heterogeneity in the population, therefore we still have four times as many equations as in the fixed mutation rate model. The framework is very similar to the randomly mutation rate variation pattern, described above. We simply do not perform the redistribution step done at the end of the 'mutation' step of the variable mutation rate model.

2.3.4 Inheritable mutation rate with noise

For the 'noisily' inheritable mutation rate, the framework is similar to the inheritable mutation rate model. However, we introduce a parameter, called 'noise', which is the proportion of cells which do not mutate to the same m category as they originate from. For the incoming factor, the category for genotype g and mutation rate category m receives contributions mainly from the same mutate rate category (multiply the frequency by 1-noise) but also from other mutation rate categories (whose frequency contributions will be multiplied by noise/3, since 3 is the number of mutation rate categories different than their own).
The incoming factor's equation is now:

$$incoming_{g,m}^{t+1} = (1 - noise) * \sum_{i}^{I} \frac{\mu^i}{\#N(\eta_{g,m}^i)} \sum_{\eta_{g,m}^i} X_{\eta_{g,m}^i}^t + noise * \sum_{\mathcal{M}!=m} \sum_{i}^{I} \frac{\mu^i}{\#N(\eta_{g,\mathcal{M}}^i)} \sum_{\eta_{g,\mathcal{M}}^i} X_{\eta_{g,\mathcal{M}}^i}^t$$

The second term of this equation represents the contributions of the terms that are not from mutation category m. Here the 'noise' factor is in front of the operator, hence it does not need to be divided by three.

The outgoing term is composed of the cells that mutate from genotype g to reach a neighbor genotype, but with the same mutation category m (those will be multiplied by a 1-noise factor) and cells that mutate from genotype g to reach a neighbor genotype, but with a different m (each one of those will be multiplied by a factor of noise/3). The outgoing term's equation is now:

$$outgoing_{g,m}^{t+1} = (1 - noise) * \sum_{i}^{I} \mu_m^i * X_{g,m}^t + noise * X_{g,m}^t$$

The final equation is thus:

$$X_{g,m}^{t+1} = (1 - noise) * X_{g,m}^t + noise * \sum_{\mathcal{M}!=m} X_{g,M}^{t+1} + incoming_{g,m}^{t+1} - outgoing_{g,m}^{t+1}$$

2.3.5 Computational Implementation

For the computational implementation, it is possible to implement only the noisily inheritable mutation rate and obtain the other mutation rate variation patterns by adjusting parameters as follows:

- For **fixed mutation rate**, we set all the mutation rates in all categories to the same level
- For **randomly variable mutation rate**, we set the noise level to 0.75: this way, each generation, the probability of remaining with the parent mutation rate is equal to the probability of migrating to another mutation rate level category
- For **inheritable mutation rate**, we set the noise level to 0.

This avoids for artifacts in results from different structure of the simulation. For examples, rounding in very small values could give rise to different results for fixed mutation rate (where

there only one frequency value for each genotype) and the variation patterns where there is heterogeneity in the population (where we have four frequencies for each genotype).

2.4 SIMULATION APPROACH

The population consists of N cells, each of them represented by its genotype. Each generation, we calculate the fitness of each cell according to its genotype and given the prevailing fitness landscape. We then perform a sampling with replacement of cells according to their fitness such the higher fitness cells are more likely to reproduce.

The formula giving the expected number of offspring X of cell of fitness f given a generation G of N cells is:

$$E(X) = \frac{f}{\sum_{i \in G} f_i} * N$$

For each cell in the draw, we adjust the mutation rate according to the strategy of the simulation run. In this study, we assume that the mutation rate is inversely, linearly determined by the concentration in the cell, m of a certain DNA repair protein. This protein is assumed to be distributed according to a Poisson distribution, similarly to Ada [21]. The mutation rate μ is calculated according to:

$$\mu = 1 - \frac{m}{100}$$

The manner of adjusting the mutation rate according to the mutation rate variation pattern is as follows:

- **Fixed mutation rate**: m is equal to m_{input} for each cell and throughout the duration of the simulation
- **Randomly variable mutation rate**: for each cell division, m is sampled from a Poisson distribution of parameter m_{input}
- **Perfectly inheritable mutation rate**: for the initial population, m for each cell is sampled from a Poisson distribution of parameter m_{input}. For each subsequent cell division, m is inherited from the parent cell.
- **Imperfectly inheritable mutation rate**: for the initial population, m for each cell is sampled from a Poisson distribution of parameter m_{input}. For each subsequent cell division, m is sampled from a Poisson distribution of parameter calculated according to:

$$m = 100 * (1 - \mu_{parent})$$

For each position in the genome, we compare m with a decision variable sampled from a uniform distribution; if this variable is smaller than the mutation rate, this position mutates. A schematic representation of the four mutation rate variation patterns in shown in Figure 5.
The simulation is then iterated for an *ngen* number of generations.

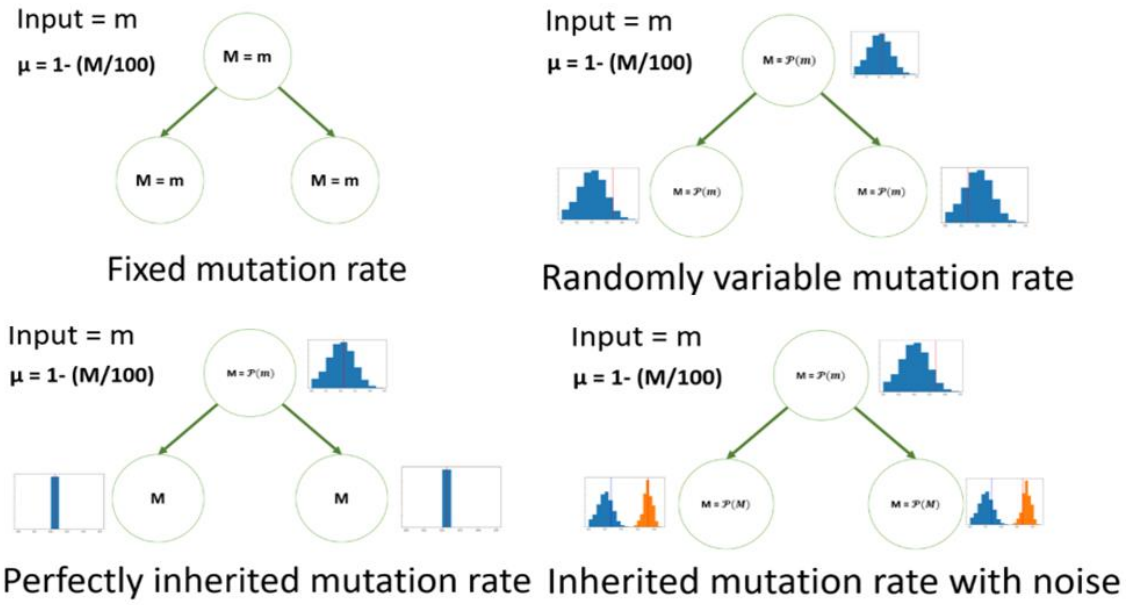


Figure 5: Schematic representation of the four studied mutation rate variation patterns as implemented in the simulation approach.

2.5 PARAMETERS

2.5.1 Mutation rate level

According to the quasi-species theory, an evolving system will aim to balance convergence to a robust genotype, that is a genotype whose neighbors have similar fitness to itself, and a fit genotype, which is the genotype with optimal fitness, even if the genotype is surrounded by less fit genotypes.

If we consider a landscape containing two peaks, a 'flat' peak – corresponding to the robust genotype, and a 'fit' peak – corresponding to the fittest genotype, theoretical predictions postulate that the population will converge to the flat peak if the mutation rate is higher than the 'quasi-species threshold' which can be approximated by:

$$\mu_{threshold} = \frac{1}{L}$$

where L is the length of the genome.

Out of the four studied landscapes, only two had 'genomes' with a defined "length": the NK landscape and the transcription factor binding sites landscapes. For those landscapes, a low mutation rate, below the quasi-species threshold, at 0.01 and a high mutation rate, above the quasi-species threshold, at 0.2.

We found that for the step landscapes, this level of mutation was also reasonable: the step landscapes required the same amount of mutations to reach the target genotypes from the initial genotypes as for the transcription factor binding sites landscapes (approximately 3.7 [3]).

Lastly, for the protein fitness landscape, twice as many mutations are needed in order to reach the target genotype from the initial genotype. However, since the number of trajectories is more limited, we chose to use the same mutation rate levels.

2.5.2 Distribution of mutation rate level

As discussed before, we chose to model the distribution of the mutation rate with a Poisson distribution for the simulation approach because it represented more closely a biological example of a DNA repair protein, Ada. [21]

For the analytical approach, we set the four mutation rates in such a manner that the distribution of the frequencies for each mutation rate was 0.25 – 0.5 – 0.25, which corresponds to a binomial distribution. We also set it that the variation between the lower and the

intermediate mutation rate was equal to the variation between the intermediate and the higher mutation rate, and that the mean of all mutation rates was equal to the fixed mutation rate. Therefore, the mutation rates for low levels were calculated according to:

$$\begin{cases} m_1 = 0 \\ m_2 = 0.01 \\ m_3 = 0.01 \\ m_4 = 0.02 \end{cases}$$

And for high levels:

$$\begin{cases} m_1 = 0 \\ m_2 = 0.2 \\ m_3 = 0.2 \\ m_4 = 0.4 \end{cases}$$

2.5.3 Initial genotype

For the step landscapes and the protein fitness landscape, the initial genotypes were obvious as per the construction of the landscape.

For the transcription factor binding sites landscapes, we set the initial genotype as the genotype with minimal fitness that is higher than 0. The number of sequences with null fitness was very large; we decided to start with a sequence that already has a binding phenotype to avoid many generations of populations evolving neutrally over genotypes with identical, null fitness.

2.5.4 Number of mutations per cell division

We were confronted with the choice on whether to allow multi-locus mutations in our model and simulations. On the one hand, theoretical studies have shown that some evolutionary dynamics might only be observable when multi-locus mutations were allowed. Indeed, those were necessary in order to fully reveal the effect of population mutation rate heterogeneity. On the other hand, biological sequences are sufficiently short on a genome scale, and mutation rates so small that multi-locus mutations remain improbable.

We also faced a problem of computational complexity: for the analytical model, considering k-neighbors of a sequence of length 8 with an alphabet of 4 required simplifications of the model. In the end, we settled multi-locus mutations.

The number of allowed mutations per approach and per landscape is detailed in Table B. For each case, we sought to allow the maximal number of mutations per cell division.

Landscape	Maximal number of mutations per cycle in simulation	Maximal number of mutations per cycle in analytical model
Steps	4	3
TF binding site	8	3
GFP-like	4	3

Table B: Maximal number of mutations for each approach and each landscape.

2.5.5 Other parameters

The number of cells in the population was 200 for the simulation, and the number of generations for which the simulation was iterated was 200. For each simulation setup, we ran 50 runs in order to limit the effect of stochasticity on the results.

3 RESULTS

3.1 STEP LANDSCAPES

3.1.1 Simulation Approach

For each of the step landscapes, we plotted the distribution of frequencies of the target genotype after 200 generations for 50 runs. We binned the results into categories of range 20. These distributions were then used as an input for hierarchical clustering, which allowed us to determine the number of groups of landscapes with similar behaviours. The dendrograms obtained from the hierarchical clustering of results for high and low mutation rate are shown in Supplementary Materials (S1 and S2). The target genotype histograms along with the landscapes coloured by their cluster are shown in Figure 6 for high mutation rate.

The main result we obtain is the superior performance of the noisily inherited mutation rate for all landscapes. More than 90% of cells in all simulation runs have reached the target genotype for all landscapes for this mode of mutation. For inheritable mutation rate variation pattern, on most landscapes, we observe a normal distribution of the target genotype around the category where 50-60% of cells have reached the target genotype.

Inheritable mutation rate simulation runs have all performed similarly for landscapes in the ‘orange’ cluster. They reached intermediate target genotype frequencies, with a large difference between the runs where it performed the best and the runs where it performed the worst. It is possible that this result is due to drift, and a different mutation rate distribution was selected at each different simulation run. In the landscapes from the ‘green’ clusters and the ‘blue’ clusters, inheritable mutation rate simulations have performed worse than the other mutation rate variation patterns, with the lowest performances in landscapes from the ‘green’ cluster. These landscapes were characterised by a high initial fitness followed by a sudden drop and a fitness valley before the target genotype. Arguably, these were due to the selection of a smaller mutation rate at the beginning of the simulation, which made the crossing of the valley more difficult. This hypothesis is strengthened by the fact that the lowest performances of the inheritable mutation rate were seen in landscapes 2, 29, 11 and 5 which had the deepest valleys.

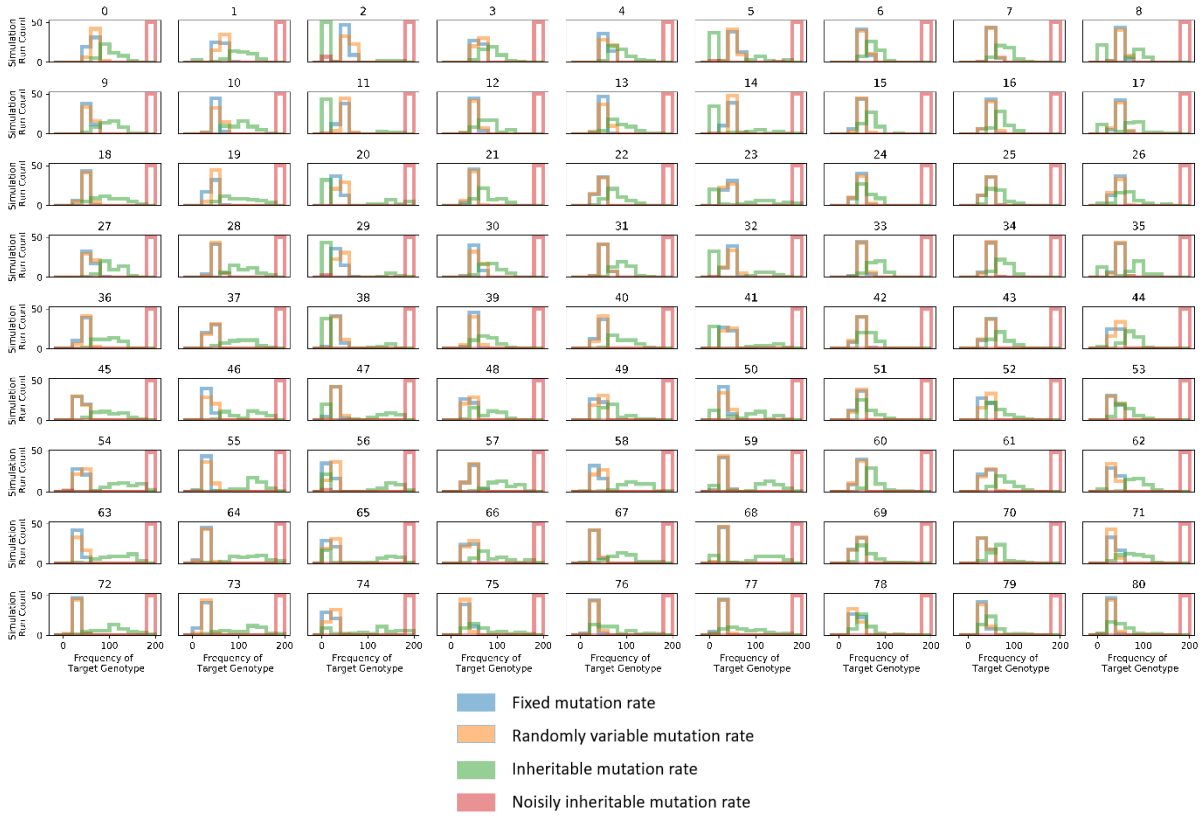
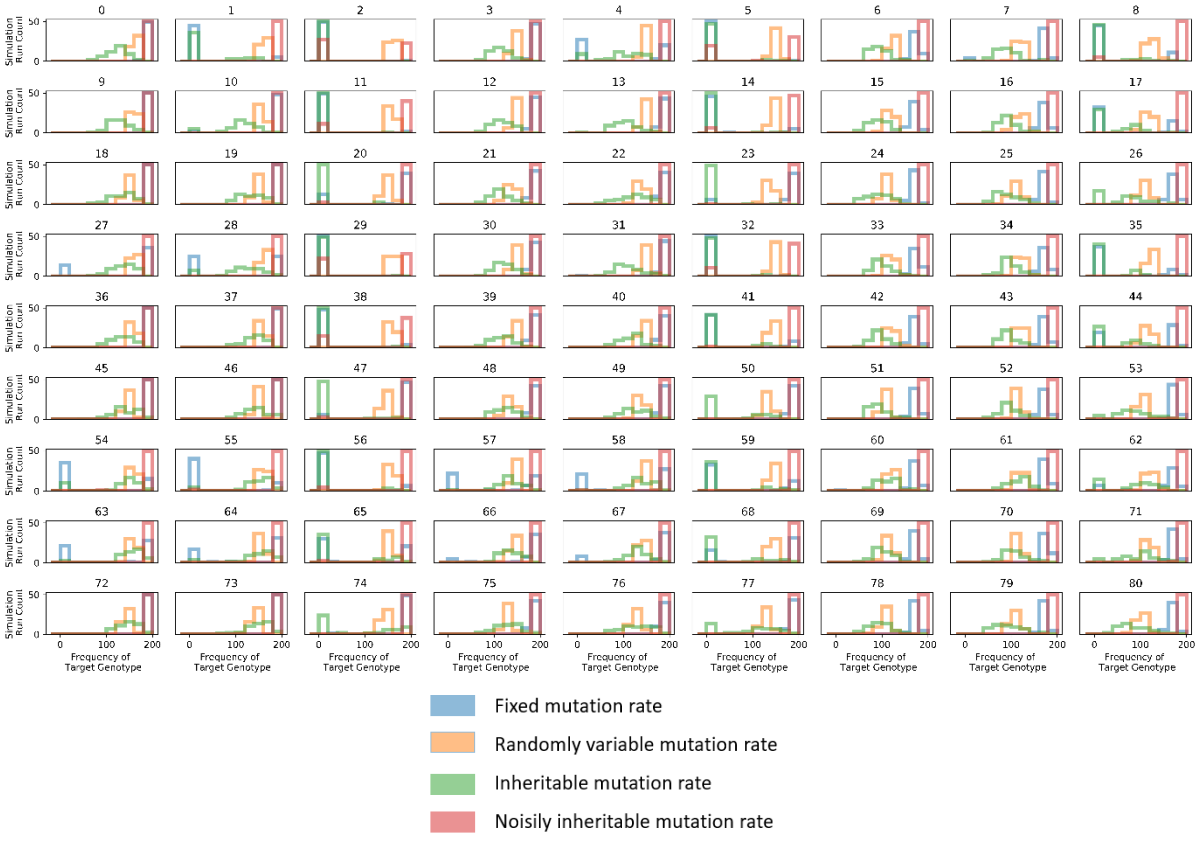


Figure 6: Target genotype frequency distributions for high mutation rate (top) and corresponding landscapes coloured by clusters obtained through hierarchical clustering (bottom).

Fixed and randomly variable mutation rate variation patterns performed very similarly for all landscapes, achieving low to intermediate frequencies of the target genotype at the end of the simulation.

For low mutation rate, we proceeded in the same manner as for the high mutation rate. The target genotype histograms along with the landscapes coloured by their cluster are shown in Figure 7.



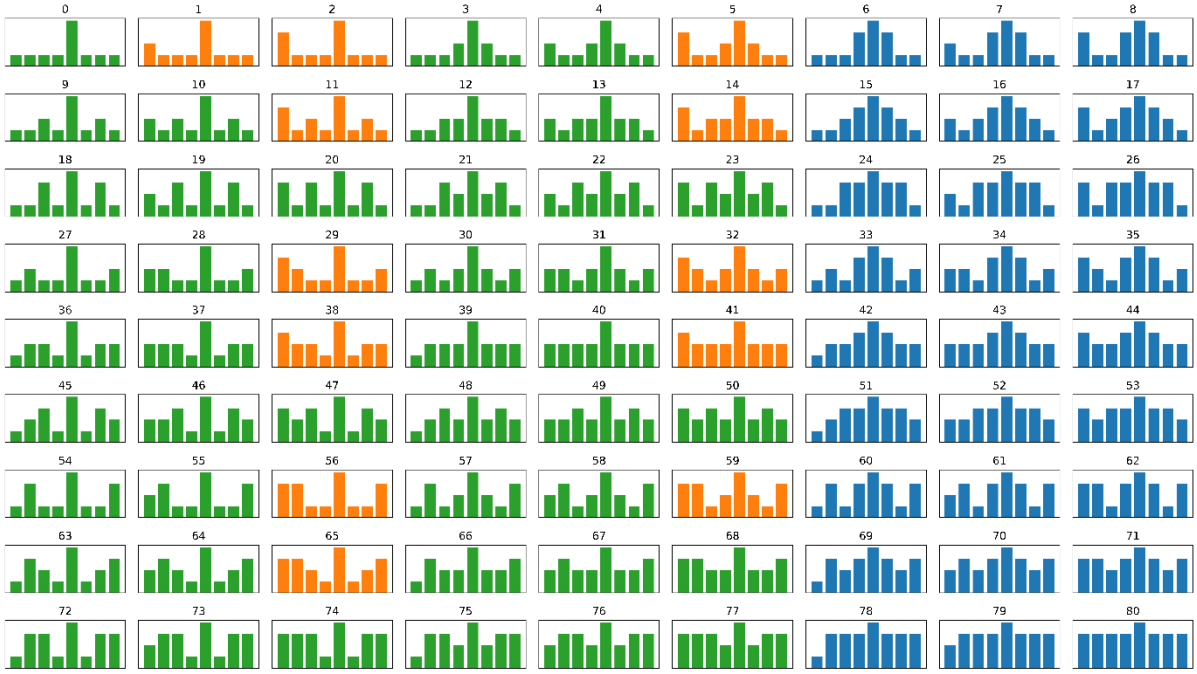


Figure 7: Target genotype frequency distributions for low mutation rate (top) and corresponding landscapes coloured by clusters obtained through hierarchical clustering (bottom).

Once again, simulation with noisily inherited mutation rate fared the best of the four mutation rate variation patterns. Only on a handful of landscapes (such as 2,29,5, see Figure 8) with very deep valleys the proportion of runs achieving more than 0.9 target genotype frequency was at 0.5-0.6 instead of 1.

For low mutation rate, the neighbourhood of the target genotype seemed to be more important than the neighbourhood of the initial genotype. Landscapes from the 'blue' cluster, on which fixed mutation rate performed well, achieving 0.7-0.8 target genotype frequency and medium performance from inheritance and randomly variable rates, were characterised by high fitness around the target genotype.

A difference with high mutation rate was that fixed mutation rate and randomly variable mutation rate variation patterns no longer performed at the same level. This could be due to multi-locus adaptation that can be achieved by randomly variable mutation rate [22] for landscapes 53, 54, 63, 64 which allows those populations to 'jump' over the valley.

As for the high mutation rate, we see a poor performance of the inheritable mutation rate for landscapes with high initial fitness followed by a deep fitness valley before the target genotype.

3.1.2 Analytical Approach

We ran our analytical model for high and low mutation rate levels. For 50 generations, we plotted the frequency of the target genotype (Figure 8). For the other 4 genotypes from the landscape, the corresponding plots are shown in Supplementary Materials (S3 and S4).
Results from the analytical model corresponded roughly to the results from the simulation.

For high mutation rate, we can see the poor performance of the inheritable mutation rate in landscapes from the third column and landscapes from the sixth column of the landscape grid. We can also see the near perfect performance of noisily inherited mutation rate for all landscapes, as well as the similar, intermediate performance of fixed and randomly variable mutation rate variation strategies.

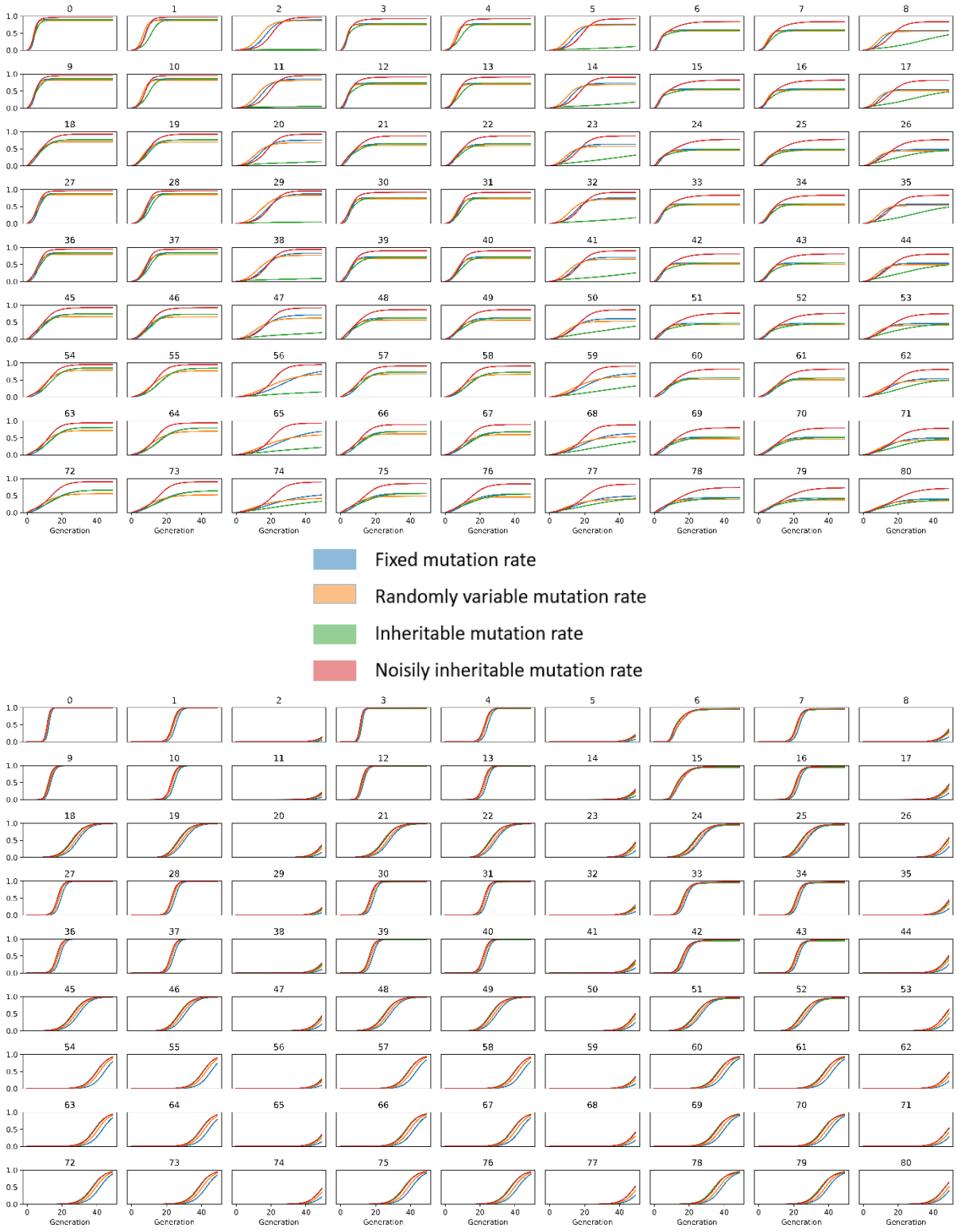


Figure 8: Evolution of the frequency of the target genotype along 50 generations, as calculated by the analytical model. Results for high mutation rate are shown at the top, results for low mutation rate are shown at the bottom.

We also observe that in some landscapes, especially from the third and ninth column of the landscape grid, the randomly variable and fixed mutation rate variation patterns result in faster convergence to the target genotype, even though they plateau at lower frequencies than the slower noisy inheritance mutation rate variation pattern.

For low mutation rate, we observe similar performance of the inheritable, noisily inheritable and randomly variable mutation rates.

3.2 TRANSCRIPTION FACTOR BINDING SITES LANDSCAPES

3.2.1 Simulation approach

The transcription factor binding site landscape set consisted of 1,167 landscapes from all kingdoms of life [3]. Each of these landscapes consisted of 65,536 genotypes. Due to this large number, we are unable to present the results in the same form as we did for step landscapes, with detailed accounts of the frequencies for each genotype and their evolution along generations. Moreover, inspection of individual simulation runs has revealed that the simulations rarely converge to the target genotype, but rather reach an equilibrium at the fitter genotypes of the population.

Therefore, we decided to consider the mean population fitness at the final generation for each fitness landscape. For each pair of strategies, we used a two-sided two-sample t-test to compare the mean fitness of each strategy at the final generation (200). This was possible because we assumed that the fitness distributions for each strategy would be normally distributed. The variances of the fitness distributions could not be assumed to be equal; however, the t-test is robust to unequal variances between two samples if the sample size is equal [35], which is the case here. We used a stringent significance threshold of 0.01, to which we applied a Bonferroni correction for multiple testing, which gave us a significance threshold of 1.4*10e-6.

The number of landscapes, along with the significance of the comparison between each pair of strategies, is shown in Table C.

Number of genotypes	Fixed vs. Random	Fixed vs. Inheritable	Fixed vs. Noisily Inheritable	Random vs. Inheritable	Random vs. Noisily Inheritable	Inheritable vs. Noisily Inheritable
1033	No (0.49)	No (0.49)	Yes (6.2e-9)	No (0.49)	Yes (5.3e-9)	Yes (7.9e-9)
111	No (0.49)	No (0.51)	No (0.16)	No (0.51)	No (0.15)	No (0.15)
6	No (0.54)	No (0.20)	No (6.8e-5)	No (0.25)	No (4.6e-3)	Yes (5.7e-7)
4	No (0.27)	No (0.21)	Yes (5.2e-7)	No (0.65)	No (9.7e-5)	No (3.1e-5)
4	No (0.67)	No (0.46)	No (1.1e-5)	No (0.47)	Yes (7.0e-7)	No (4.8e-6)
4	No (0.66)	No (0.75)	Yes (8.7e-7)	No (0.61)	No (2.2e-6)	Yes (5.5e-7)
3	No (0.77)	No (0.35)	Yes (5.8e-7)	No (0.43)	Yes (7.5e-7)	No (9.4e-6)
2	No (0.36)	No (0.28)	No (1.6e-6)	No (0.80)	Yes (3.6e-7)	Yes (4.2e-7)

Number of genotypes	Fixed vs. Random	Fixed vs. Inheritable	Fixed vs. Noisily Inheritable	Random vs. Inheritable	Random vs. Noisily Inheritable	Inheritable vs. Noisily Inheritable
751	No (0.21)	No (0.49)	No (0.41)	No (0.23)	No (0.25)	No (0.43)
175	Yes (4.3e-8)	No (0.51)	No (0.17)	Yes (3.8e-8)	Yes (2.8e-8)	No (0.16)
46	Yes (1.1e-7)	No (0.50)	No (0.40)	Yes (1.2e-7)	No (8.7e-2)	No (0.41)
45	No (7.7e-3)	No (0.58)	No (0.14)	No (8.5e-3)	Yes (1.3e-7)	No (0.19)
45	No (0.04)	No (0.27)	No (0.46)	Yes (2.1e-7)	No (0.13)	No (0.41)
43	Yes (2.6e-7)	No (0.27)	No (0.35)	No (2.7e-2)	No (0.13)	No (0.50)
29	No (1.0e-2)	No (0.29)	No (0.15)	Yes (9.8e-8)	Yes (1.4e-7)	No (0.33)
29	Yes (3.1e-7)	No (0.36)	No (0.30)	No (4.7e-3)	Yes (1.1e-7)	No (0.12)
2	Yes (2.7e-7)	No (0.27)	No (0.35)	No (2.6e-2)	No (0.13)	No (0.50)
2	Yes (5.6e-10)	No (5.8e-2)	Yes (1.5e-7)	Yes (8.4e-11)	Yes (2.5e-14)	No (1.6e-3)

Table C: Number of landscapes for which each pair of the studied mutation rate variation patterns exhibited significant differences, with $\alpha = 1.4*10e-6$, for high mutation rate (top) and low mutation rate (bottom). The mean p-value is indicated between parentheses in each case.

A representative histogram from the two major categories for each mutation rate level is shown in Supplementary Materials (S5 and S6), along with the time evolution of mean population fitness.

For high mutation rate, we observe that the noisily inheritable mutation rate is significantly different than the three other strategies for almost all the surveyed landscapes (1033 out of 1,167, about 90%), while the other pairs of strategies were comparable. We further investigated to see whether the effect was a higher population fitness or a lower population fitness: it was higher for all 1033 landscapes. This was consistent with results from the step landscapes. For 111 landscapes, all four of the strategies performed equivalently.

For low mutation rate, we observe that for more than half (751 out of 1,167 landscapes, that is 64%) landscapes all the mutation rate variation patterns performed equivalently. The random mutation rate variation pattern seemed to differentiate itself from the other patterns the most, since all the other categories were different combinations of at least one significantly different pair involving the randomly variable mutation rate variation pattern. This could be due to the heterogeneity of the mutation rate, leading to a heterogeneity of genotypes in the population around the fitness peak. The other patterns would have fixed mutation rates by that point in the simulation.

3.2.2 Analytical Approach

We ran the analytical model on one representative landscape from each of the two major mutation rate categories for each mutation rate level.

Evolution of mean population fitness for high mutation rate is shown in Figure 9. The results were consistent with results obtained through the simulation: noisily inheritable mutation rate resulted in higher mean population fitness. Fixed and inheritable mutation rate resulted in the same mean population fitness: in the simulation approach also, we never observed any significant difference between fixed and inheritable mutation rate variation patterns. The only difference between the simulation approach and the analytical approach was the lower performance of the randomly variable mutation rate variation pattern in the

analytical approach. In the simulation approach, it performed slightly better than inheritable/fixed mutation rates.

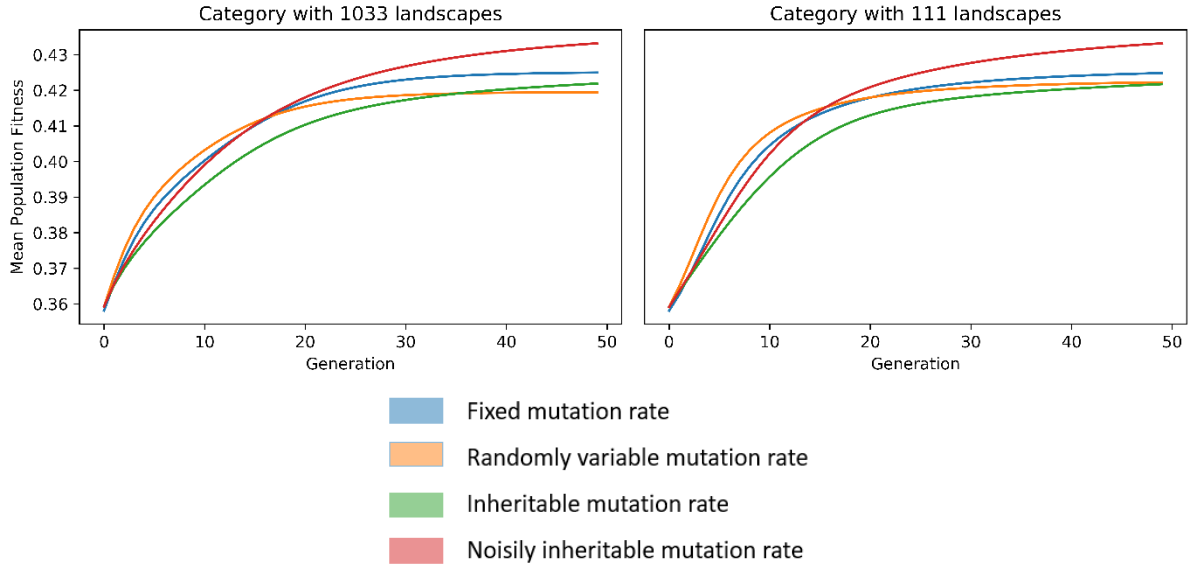


Figure 9: Evolution of mean population fitness over 50 generations for high mutation rate in analytical approach. The two plots each correspond to a different landscape, representative of the two major categories of behaviours determined by the simulation approach. Left corresponds to category with 1033 landscapes (see Table C), right to the category with 111 landscapes.

For low mutation rate, the behaviours were more similar between mutation rate variation patterns (see Figure 10). Except for a slight underperformance of the inheritable mutation rate, the four mutation rate convergence patterns were almost confounded.

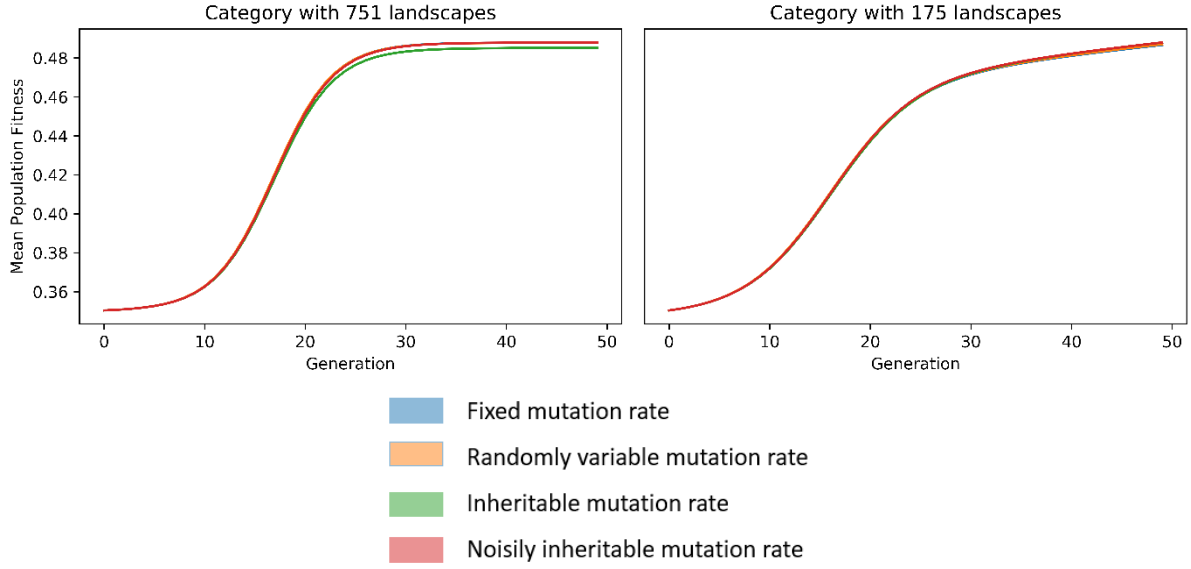


Figure 10: Evolution of mean population fitness over 50 generations for low mutation rate in analytical approach. The two plots each correspond to a different landscape, representative of the two main categories of behaviours determined by the simulation approach. Left corresponds to category with 751 landscapes (see Table C), right to the category with 175 landscapes.

3.3 Artificial Protein Landscape

3.3.1 Simulation Approach

We ran 50 runs of the simulation, with 200 cells and for 200 generations, for high and low mutation rate levels. We plotted the evolution of the mean population fitness along

generations as well as the distribution of final mean population fitness. Results are shown in Figure 11.

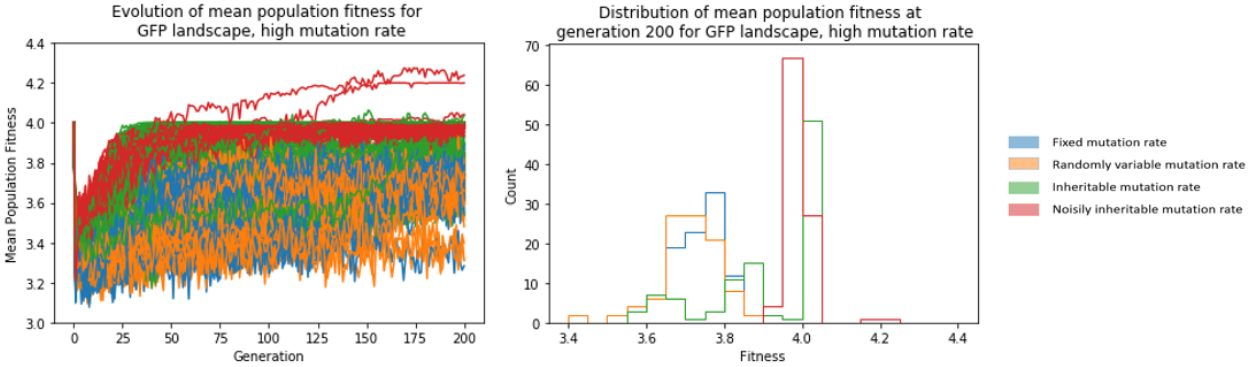


Figure 11: Evolution of mean population fitness (left) and distribution of final mean population fitness (right) for GFP landscape, at high mutation rate.

Pair of mutation rate variation pattern	T-test statistic	P-value	Significance
Fixed vs. Random	2.23	0.02	Not significant
Fixed vs. Inheritable	-10.09	1.44e-19	Very Significant
Fixed vs. Noisily Inheritable	-29.83	3.54e-75	Very Significant
Random vs. Inheritable	-11.58	5.25e-24	Very Significant
Random vs. Noisily Inheritable	-31.62	2.51e-79	Very Significant
Inheritable vs. Noisily Inheritable	-7.14	1.74e-11	Very Significant

Table D: Comparison of mean population fitness distributions at final simulation time point, for high mutation rate and GFP landscape. The α-threshold for significance was set at 8.3e-3 (0.05 Bonferroni corrected for multiple testing). When the t-test statistic is negative and the corresponding p-value is significant, this means that the second mutation rate variation pattern performed better.

For high mutation rate, we observed no significant difference between the final mean population fitness distributions for fixed and randomly mutation rate variation patterns (two-tailed t-test, statistic = 2.23, p = 0.02). However, there was a significant difference between the noisily inheritable mutation rate final mean population fitness distributions and all the other mutation rate variation patterns. P-values for t-tests, for all pair of mutation rate variation patterns are shown in Table D.

The noisily mutation rate performed the best out of the four mutation rate variation patterns, followed by the perfectly inheritable mutation rate variation patterns. However, the latter had very variable performance. These results are consistent with the results obtained for steps landscapes, especially landscapes such as 63 or 64 (see Figure 6) which represent a

progressive fitness decrease to a deep valley, before a higher target genotype fitness, which corresponds to our landscape.

For low mutation rate, the evolution of mean population fitness is shown in Figure 12, along with the distributions of mean population fitness and their comparison in Table E. All the strategy pairs were significantly different. The randomly variable mutation rate performed worst, followed by the fixed mutation rate. It seems like the perfectly inheritable mutation rate performed the best.

We also observe a decrease in fitness at the initial generations of the noisily inheritable mutation rate, possibly due to getting stuck in local optima before genotypes with null fitness. This is consistent with step landscapes, where randomly variable/fixed mutation rate variation patterns performed better in the initial generations than the other mutation rate variation patterns.

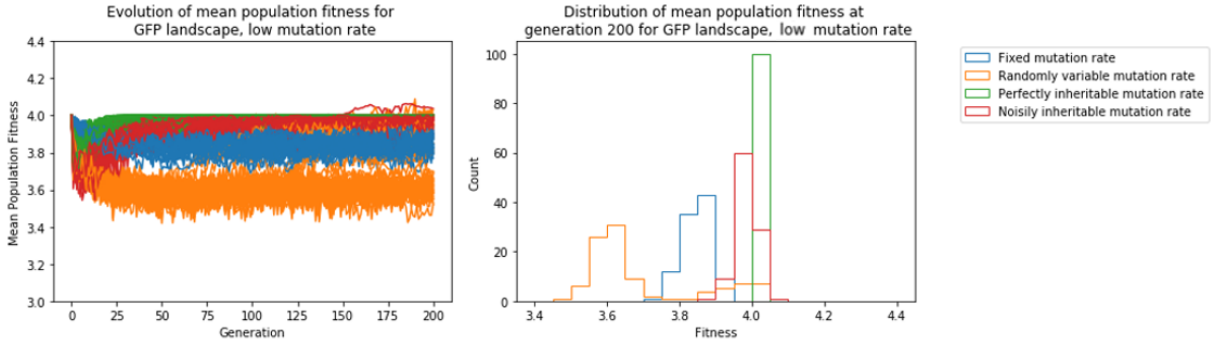


Figure 12: Evolution of mean population fitness (left) and distribution of final mean population fitness (right) for GFP landscape, at high mutation rate.

Pair of mutation rate variation pattern	T-test statistic	P-value	Significance
Fixed vs. Random	9.79	1.03e-18	Very Significant
Fixed vs. Inheritable	-38.15	3.41e-93	Very Significant
Fixed vs. Noisily Inheritable	-28.12	4.56e-71	Very Significant
Random vs. Inheritable	-19.68	1.73e-48	Very Significant
Random vs. Noisily Inheritable	-18.37	1.12e-44	Very Significant
Inheritable vs. Noisily Inheritable	6.14	4.32e-09	Very Significant

Table E: Comparison of mean population fitness distributions at final simulation time point, for low mutation rate and GFP landscape. The α-threshold for significance was set at 8.3e-03 (0.05 Bonferroni corrected for multiple testing).

3.3.2 Analytical Approach

For the analytical approach, we plotted the mean population fitness for high and low mutation levels, shown in Figure 13, and distributions of genotypes in the half of the fitness

landscape corresponding to the ‘wild-type’ genotype and the half corresponding to the ‘adaptive’ genotype, in Figure 14, along with their frequency evolution along generations.

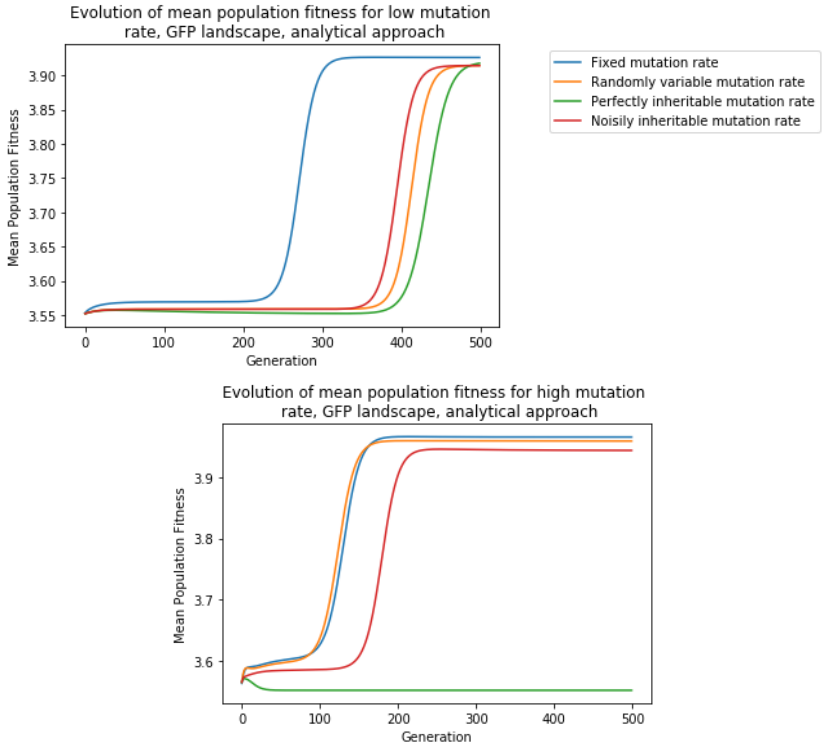


Figure 13: Evolution of the mean population fitness for GFP landscape, low mutation rate (left) and high mutation rate (right).

The fixed mutation rate was faster and reached higher mean population fitness than the three other mutation rate variation patterns, which performed similarly.
For high mutation rate, the inheritable mutation rate variation rate pattern failed to adapt and converged to the wild-type genotype. Similarly, the noisily inheritable mutation rate variation resulted in longer persistence of the wild-type genotype. However, it also resulted in highest frequencies of the target genotype while the other mutation rate variation patterns were split into a multitude of small peaks over the adaptive part of the landscape.

For fixed mutation rate, the frequency of wild-type genotype decreased immediately and exponentially. For the other mutation rate variation patterns, the wild-type frequency plateaued for 200 generations (high mutation rate) or 400 generations (low mutation rate) before decreasing.

It thus seems like the heterogeneity of mutation rate can lead to persistence of the wild-type genotype for longer, and leads to binary-like population dynamics, where the population switches from one genotype to the other at high frequency.
For low mutation rate levels, we again see a dynamic where a switch seems to take place between the ‘wild-type’ half of the landscape and the ‘adaptive’ part of the landscape for the inheritable/inheritable with noise mutation rate variation patterns. For the fixed/randomly variable mutation rate variation patterns, we observe the same dynamic, but again, less marked.

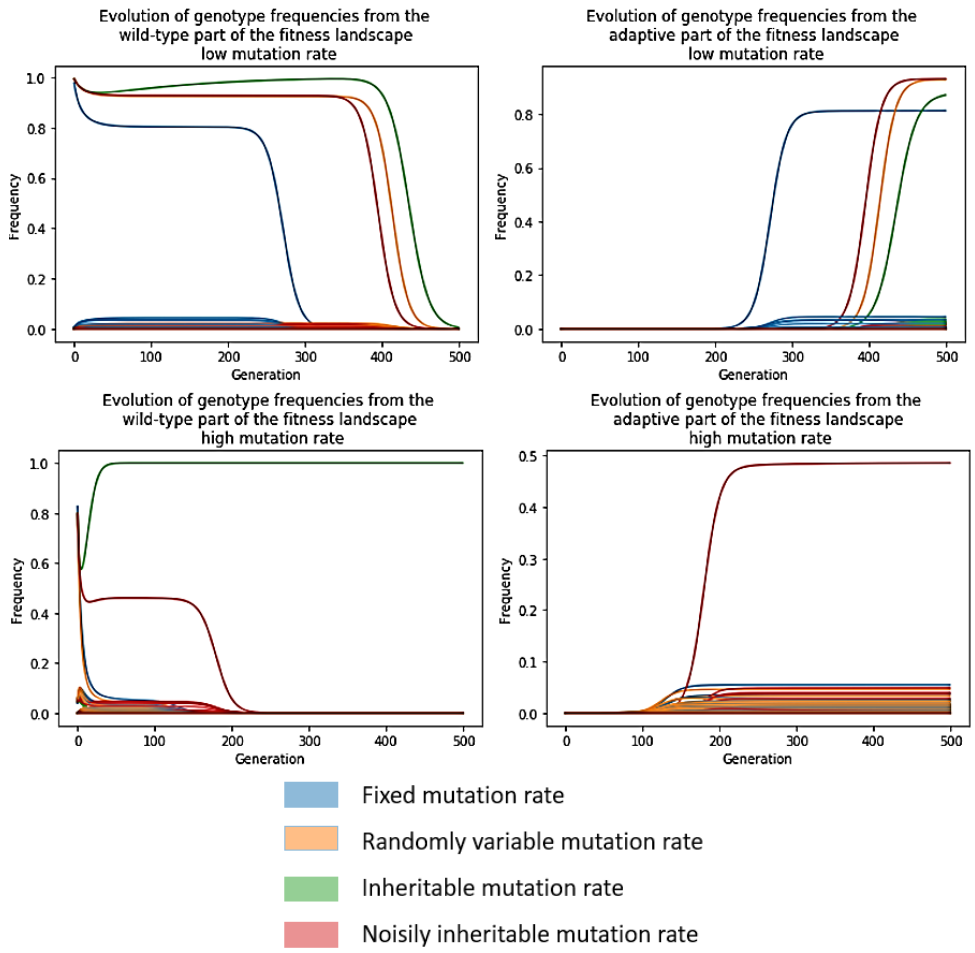


Figure 14: Evolution of genotype frequencies from the wild-type part of the fitness landscape (left) to adaptive part of the fitness landscape (right) for low mutation rate (top) and high mutation (bottom).

The apparent discrepancy between the simulation approach and the analytical approach can be explained by the stochasticity of the simulation. In the GFP landscape, there are a lot of null fitness genotypes which are all taken into account in the analytical approach. In the simulation, however, only some of them will actually occur, which leads to higher mean fitness of one neighbours of a genotype, which can rescue the adaptation in some cases.

4 DISCUSSION

We found that the noisily inheritable mutation rate results in enhanced adaptation process on several landscapes. The behaviour of all four of the mutation rate variation patterns has demonstrated the high dependency of the landscape on the adaptive process.

4.1 ENHANCED PERFORMANCE OF THE NOISILY INHERITABLE MUTATION RATE AND LINK WITH THE MULTI-ARMED BANDIT PROBLEM

We have found that the noisily inheritable mutation rate variation pattern leads to higher mean population fitness for transcription factor binding sites fitness landscapes and higher frequencies of target genotype for step landscapes, as well as an artificial protein landscape in the simulation approach.

We also plotted the mean population mutation rate for transcription factor binding sites and the protein landscapes. We found that the mean mutation rate increased, then decreased (see Figure 15) consistently with optimal methods of adjusting ε in order to optimize the exploration/exploitation balance. Indeed, the excellent performance of the noisily inherited mutation rate can be understood intuitively: when fitness is low, cells with high mutation rate

are selected for since they are more likely to have acquired fitter genotypes. This enhances 'exploration' and makes it more likely that some cells will find the target genotype. On the other hand, once the target genotype is reached, cells with low mutation rate will be selected for since they will be more likely to have kept the fitter genotype. This in turn enhances the 'exploitation' mechanism.

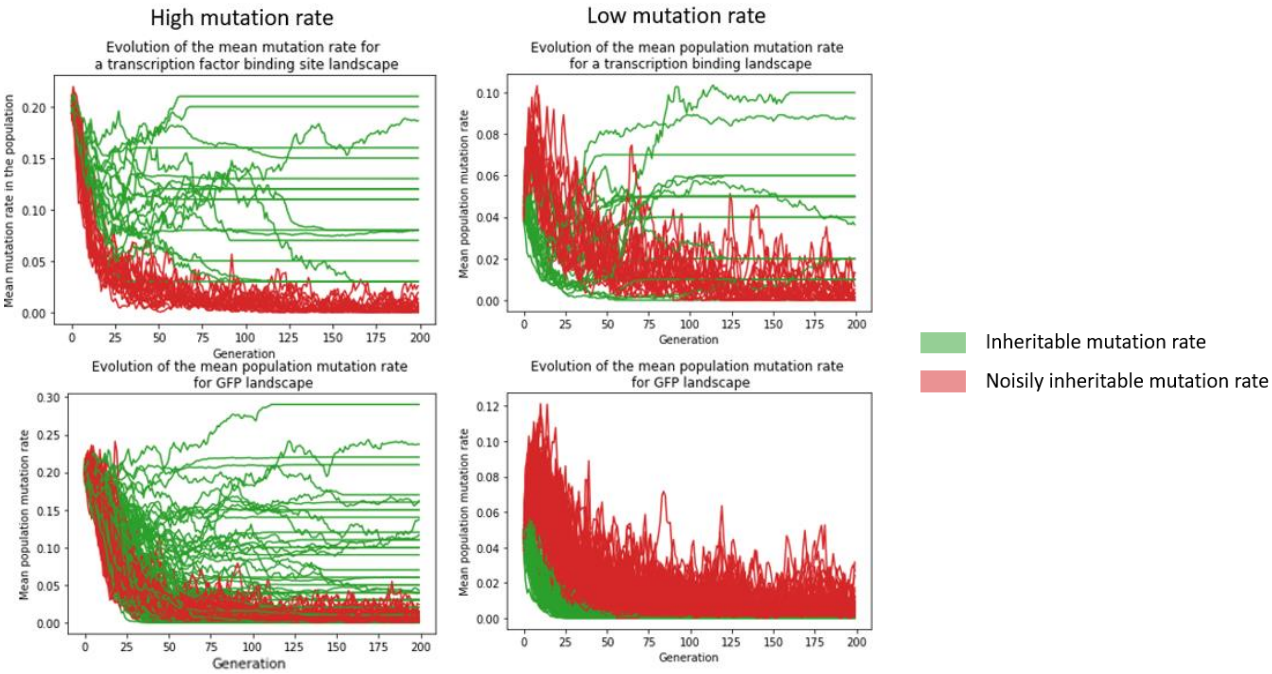


Figure 15: Evolution of the mean population mutation rate as a function of time. Consistently with the optimal methods of adjusting ε, the mutation rate increases during the adaptive process then decreases once a target genotype has been reached.

We will now discuss additional evidence for this proposed mechanism of better performance of the noisily inheritable mutation rate. If the noisily inheritance of mutation rate enhances the exploration/exploitation transition, we should see an increase of mean population mutation rate corresponding to the 'exploration' phase, followed by a decrease corresponding to the 'exploitation' phase. This was discussed above. We would also need to a higher number of genotypes in the population vs. the other mutation rate variation patterns, followed by a smaller number of genotypes; and a correlation between mean population mutation rate and mean population fitness.

4.1.1 Correlation between mean population mutation rate and mean population fitness

We plotted the mean population mutation rate against the mean population fitness in Figure 16, for one transcription factor binding site landscape and the GFP landscape.

We observed very strong negative correlation between the mean population fitness and the mean population mutation rate, for the noisily inherited mutation rate variation pattern. This is also consistent with the methods to adjust ε in the multi-armed bandit problem. When the mean population fitness against the mean population mutation rate, we found a very strong negative correlation for both inheritable and noisily inheritable mutation rate variation patterns.

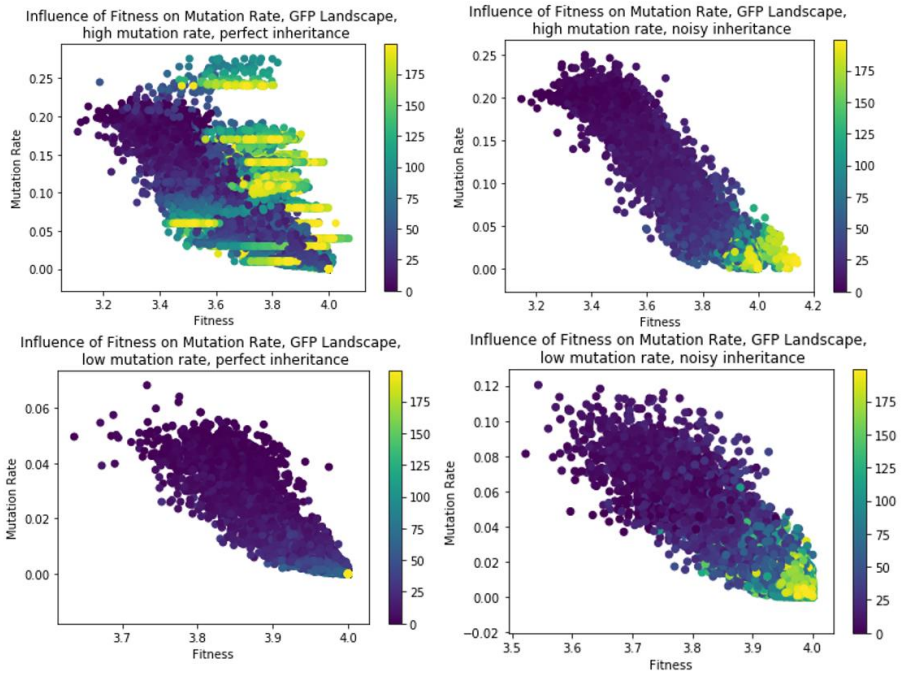


Figure 16: Correlation between mutation rate and mean population fitness for high mutation rate level (top) and low mutation rate level (bottom), for perfect inheritance mutation rate variation pattern (left) and noisy inheritance mutation rate variation pattern (right). The scatter dots are coloured according to generation time.

4.1.2 Population heterogeneity along evolutionary time

If the noisily inherited mutation rate leads to increased exploration, then we should see more genotypes in the population for the noisily inherited mutation variation pattern than for the other mutation rate variation patterns. This we only observed for low mutation rates (see Figure 17). For high mutation rate, it seems like the exploration does not need to be enhanced: we see no difference between the number of genotypes for noisily inherited mutation rate and the other variation patterns, and the mutation rate starts decreasing immediately (see Figure 17).

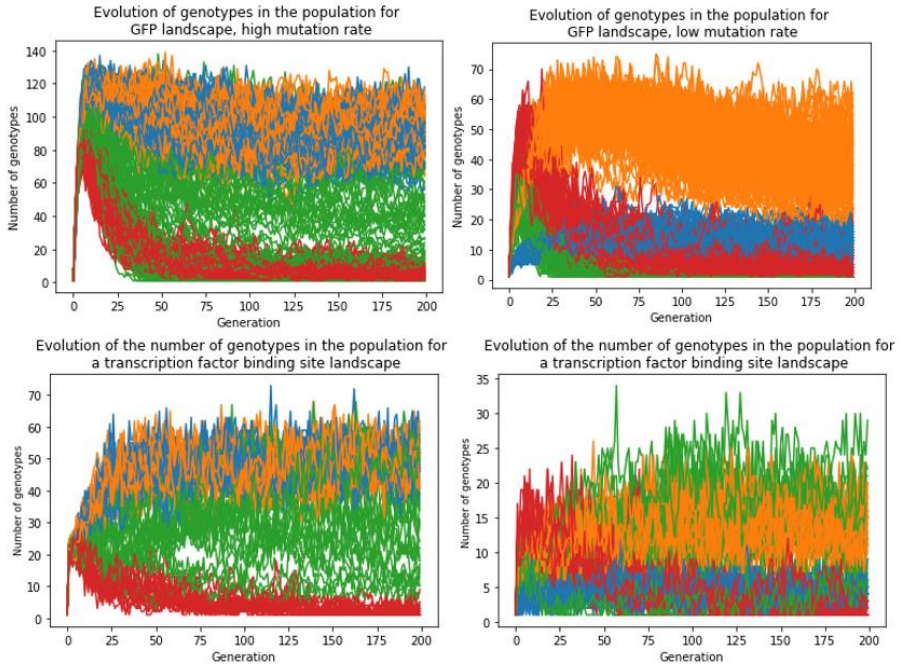


Figure 17: Number of genotypes as a function of time, for high mutation rate (left) and for low mutation rate (right), for GFP landscape (top) and for a transcription factor binding site (bottom).

The noisily inherited mutation rate variation pattern did not result in even higher numbers of genotypes for high mutation rate ('exploration' phase) but resulted in lower number of genotypes for both levels of mutation rate ('exploitation' phase).

On the other hand, the 'exploitation' part seems to be enhanced for both high and low mutation rates. The number of genotypes decreases sharply once the population has found some fitter peaks and is lower for both levels of mutation and for both types of landscapes than any other mutation rate variation pattern.

4.2 DIFFERENCES IN BEHAVIOUR OF THE FOUR MUTATION RATE STRATEGIES DEPENDING ON THE LANDSCAPES

In this project, we showed that the optimal mutation rate variation pattern is dependent on the fitness landscape. For the GFP landscape, we have found that the optimal mutation rate variation pattern did not correspond to the optimal mutation rate variation pattern of the transcription factor binding sites landscapes. If we suppose that biological organisms can evolve their mutation rate variation pattern, we must first think about which fitness landscape is the most representative of the biological reality.

Mutations can be neutral, deleterious or advantageous. Kimura postulated in 1968 [36] that most of the mutations are neutral or quasi-neutral. However, it is also the fact that biological landscapes are very sparse: out of the near infinity of possible sequences, only a few of can give rise to functional sequences.

It thus seems that our GFP landscape was the most representative of biological reality, and thus the fixed mutation rate could be the optimal mutation rate in biological systems. The noisily inheritance rate, however, can still be beneficial in this setup since it leads to very fast switching between the adaptive and the wild-type genotypes, while the other mutation rate variation patterns seem to lead to a multitude of local peaks. This could be important for the evolution of interactions with other components of the genome.

4.3 FUTURE DIRECTIONS

We will continue the analytical study of the impact of the landscape geometry on the optimal mutation rate variation pattern and try to translate it to biological reality. In this project, we considered the three non-fixed mutation rate variation patterns as distinct strategies. In reality, they can be all summarised as a noisily inheritable mutation rate pattern, with noise set to 0 to model perfectly inheritable mutation rate and noise set to infinity to model randomly variable mutation rate. Future work will consist to find general features of fitness landscapes in order to determine which mutation rate variation pattern is optimal for each.

We will also try and uncover evidence for noisily variation in the mutation rate. In biological systems, we assume that mutations are random. Therefore, a potential noise would have to cater to all possible fitness landscapes that are being explored by the organism. It is hence possible that the mutation rate noise that has evolved in nature is suboptimal, but more versatile.

Here, due to space constraints, we only considered noise in genetic mutation rate, that is the rate of modification of DNA sequences. However, phenotypic mutations, such as errors in transcription and translation, have been shown to have an important role in evolution [2]. Again, the optimal level of phenotypic mutation rate and its evolution can be determined through an approach very similar to what we presented in this work and represent a major future direction in this study.

5 BIBLIOGRAPHY

[1] M. Kirschner and J. Gerhart, “Perspective Evolvability,” *Signals*, 1998.

[2] J. L. Payne and A. Wagner, “The causes of evolvability and their evolution,” *Nature Reviews Genetics*. 2019.

[3] J. Aguilar-Rodríguez, J. L. Payne, and A. Wagner, “A thousand empirical adaptive landscapes and their navigability,” *Nat. Ecol. &Amp; Evol.*, vol. 1, p. 45, Jan. 2017.

[4] K. S. Sarkisyan *et al.*, “Local fitness landscape of the green fluorescent protein,” *Nature*, 2016.

[5] C. Li, W. Qian, C. J. Maclean, and J. Zhang, “The fitness landscape of a tRNA gene,” *Science*, vol. 352, no. 6287, pp. 837–840, May 2016.

[6] J. Aguilar-Rodríguez, L. Peel, M. Stella, A. Wagner, and J. L. Payne, “The architecture of an empirical genotype-phenotype map,” *Evolution (N. Y).*, 2018.

[7] S. Kauffman and S. Levin, “Towards a general theory of adaptive walks on rugged landscapes,” *J. Theor. Biol.*, vol. 128, no. 1, pp. 11–45, 1987.

[8] S. Kauffman, “Homeostasis and differentiation in random genetic control networks,” *Nature*. 1969.

[9] J. A. G. M. De Visser and J. Krug, “Empirical fitness landscapes and the predictability of evolution,” *Nature Reviews Genetics*. 2014.

[10] P. A. Romero, A. Krause, and F. H. Arnold, “Navigating the protein fitness landscape with Gaussian processes,” *Proc. Natl. Acad. Sci. U. S. A.*, 2013.

[11] M. Nilsson and N. Snoad, “Optimal mutation rates in dynamic environments,” *Bull. Math. Biol.*, 2002.

[12] J. Clune, D. Misevic, C. Ofria, R. E. Lenski, S. F. Elena, and R. Sanjuán, “Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes,” *PLoS Comput. Biol.*, 2008.

[13] F. Y. M. Wan, A. V. Sadovsky, and N. L. Komarova, “Genetic Instability in Cancer: An Optimal Control Problem,” *Stud. Appl. Math.*, 2010.

[14] I. Eshel, “Clone-selection and optimal rates of mutation,” *J. Appl. Probab.*, 1973.

[15] T. Miura and P. Sonigo, “A mathematical model for experimental gene evolution,” *J. Theor. Biol.*, 2001.

[16] B. O’Fallon, “Two optimal mutation rates in obligate pathogens subject to deleterious mutation,” *J. Theor. Biol.*, 2011.

[17] R. V. Belavkin, A. Channon, E. Aston, J. Aston, R. Krašovec, and C. G. Knight, “Monotonicity of fitness landscapes and mutation rate control,” *J. Math. Biol.*, 2016.

[18] J. F. Y. Brookfield, “Mutation Rates: Simpler Than We Thought?,” *Current Biology*. 2018.

[19] M. Stich, S. C. Manrubia, and E. Lázaro, “Variable mutation rates as an adaptive strategy in replicator populations,” *PLoS One*, 2010.

[20] H. Mühlenbein, “How genetic algorithms really work. Mutation and Hillclimbing,” *Proc. Parallel Probl. Solving from Nat. 2,* 1992.

[21] S. Uphoff, N. D. Lord, B. Okumus, L. Potvin-Trottier, D. J. Sherratt, and J. Paulsson, “Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation,” *Science (80-.).*, 2016.

[22] S. Bonhoeffer, S. I. Mayer, and H. K. Alexander, “Population Heterogeneity in Mutation Rate Increases the Frequency of Higher-Order Mutants and Reduces Long-Term Mutational Load,” *Mol. Biol. Evol.*, vol. 34, no. 2, pp. 419–436, Nov. 2016.

[23] T. G. Hammerstrom, K. Beabout, T. P. Clements, G. Saxer, and Y. Shamoo, “Acinetobacter baumannii repeatedly evolves a hypermutator phenotype in response to tigecycline that effectively surveys evolutionary trajectories to resistance,” *PLoS One*, 2015.

[24] L. R. Mulcahy, J. L. Burns, S. Lory, and K. Lewis, “Emergence of Pseudomonas aeruginosa strains producing high levels of persister cells in patients with cystic fibrosis,” *J. Bacteriol.*, 2010.

[25] J. M. Kang, N. M. Iovine, and M. J. Blaser, “A paradigm for direct stress-induced mutation in prokaryotes,” *FASEB J.*, 2006.

[26] M. Karami-Zarandi, M. Douraghi, B. Vaziri, H. Adibhesami, M. Rahbar, and M. Yaseri, “Variable spontaneous mutation rate in clinical strains of multidrug-resistant Acinetobacter baumannii and differentially expressed proteins in a hypermutator strain,” *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, 2017.

[27] N. Cesa-Bianchi, “Multi-armed Bandit Problem,” in *Encyclopedia of Algorithms*, 2014.

[28] S. K. Strauss *et al.*, “Evolthon: A community endeavor to evolve lab evolution,” *PLoS Biol.*, 2019.

[29] D. Bar-Yaacov *et al.*, “RNA editing in bacteria recodes multiple proteins and regulates an evolutionarily conserved toxin-antitoxin system,” *Genome Res.*, 2017.

[30] E. Mordret *et al.*, “Systematic detection of amino acid substitutions in proteome reveals a mechanistic basis of ribosome errors,” *bioRxiv*, 2018.

[31] S. F. Field and M. V. Matz, “Retracing evolution of red fluorescence in GFP-like proteins from faviina corals,” *Mol. Biol. Evol.*, 2010.

[32] P. A. Romero and F. H. Arnold, “Exploring protein fitness landscapes by directed evolution,” *Nature Reviews Molecular Cell Biology*. 2009.

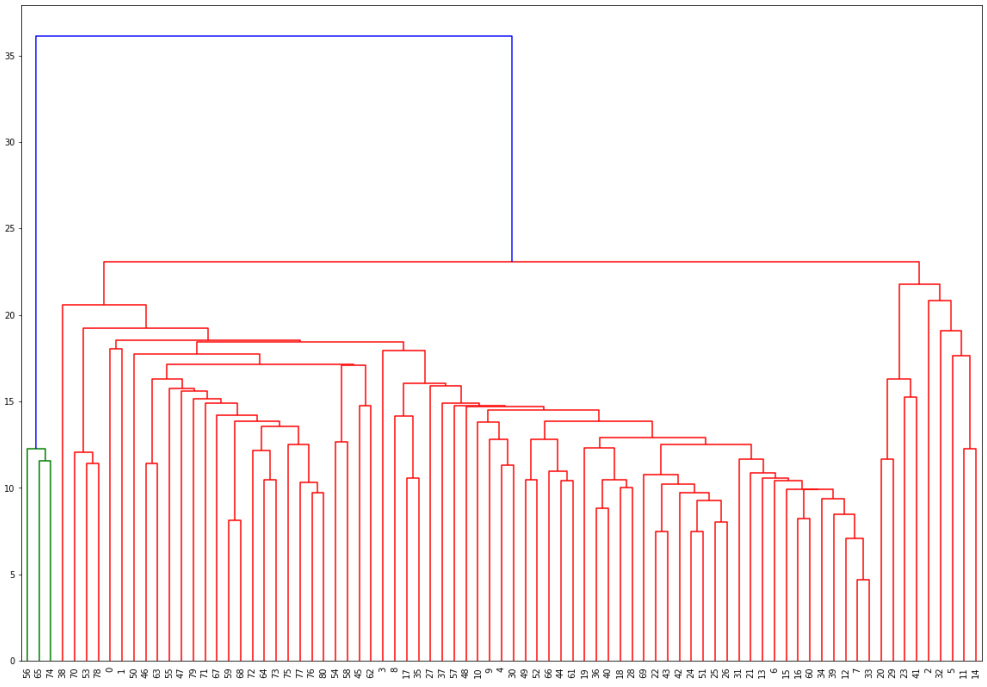
[33] C. R. Otey, M. Landwehr, J. B. Endelman, K. Hiraga, J. D. Bloom, and F. H. Arnold, “Structure-guided recombination creates an artificial family of cytochromes P450,” *PLoS Biol.*, 2006.

[34] S. Banerjee *et al.*, “Mispacking and the Fitness Landscape of the Green Fluorescent Protein Chromophore Milieu,” *Biochemistry*, 2017.

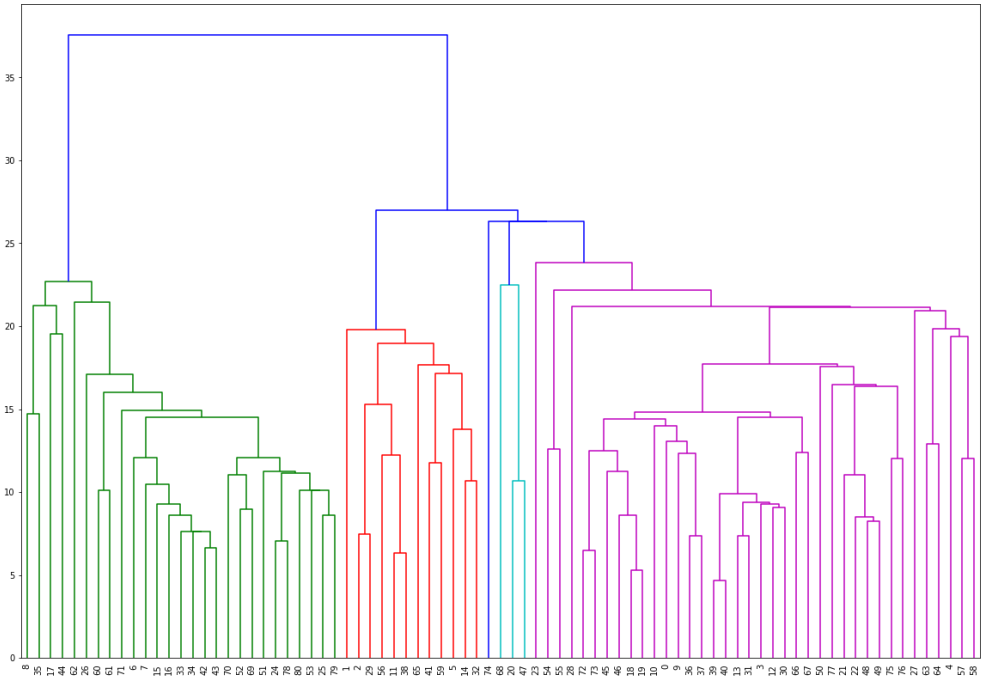
[35] C. A. Markowski and E. P. Markowski, “Conditions for the effectiveness of a preliminary test of variance,” *Am. Stat.*, 1990.

[36] M. Kimura, “Evolutionary rate at the molecular level,” *Nature*, 1968.

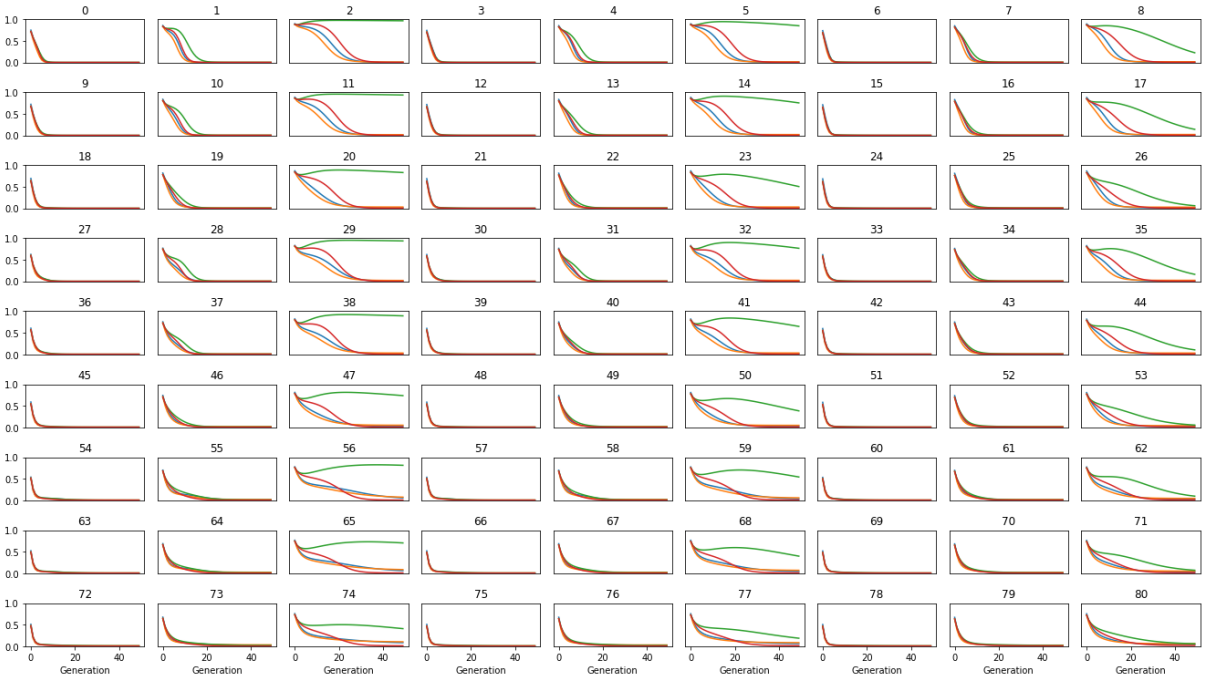
6 SUPPLEMENTARY MATERIALS

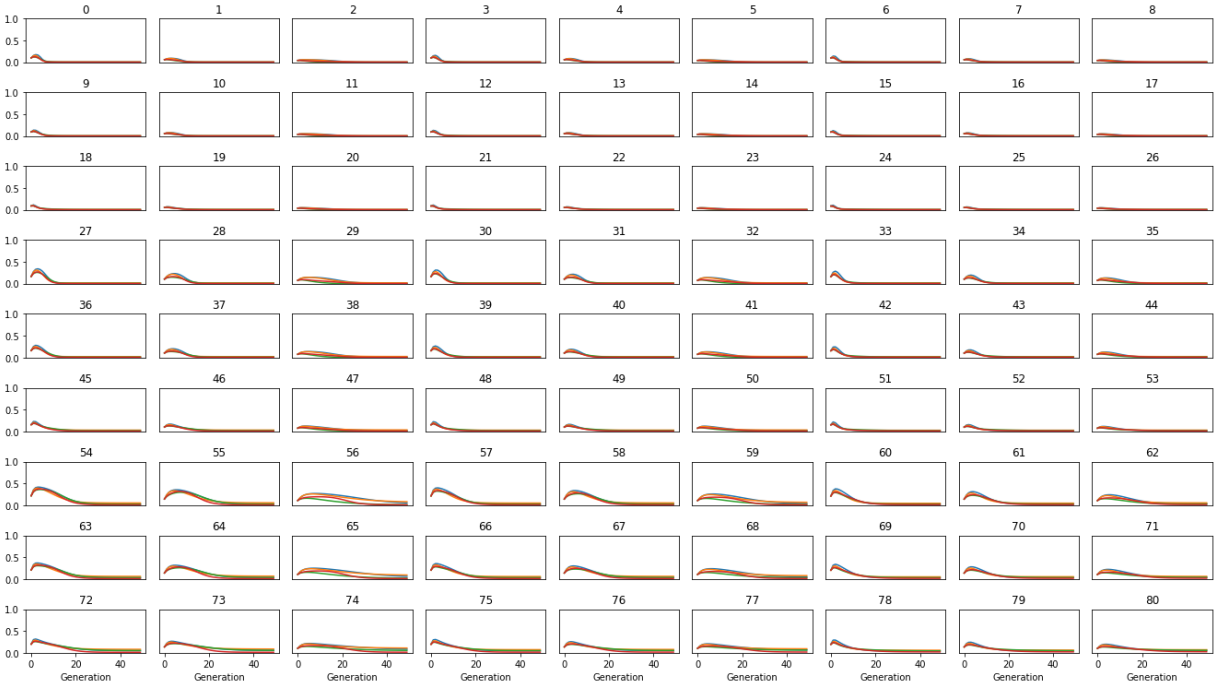


S1: Dendrogram resulting from hierarchical clustering of simulation results for step landscapes, high mutation rate. Leaf labels are landscape identification numbers from Figure 2 in the main text.

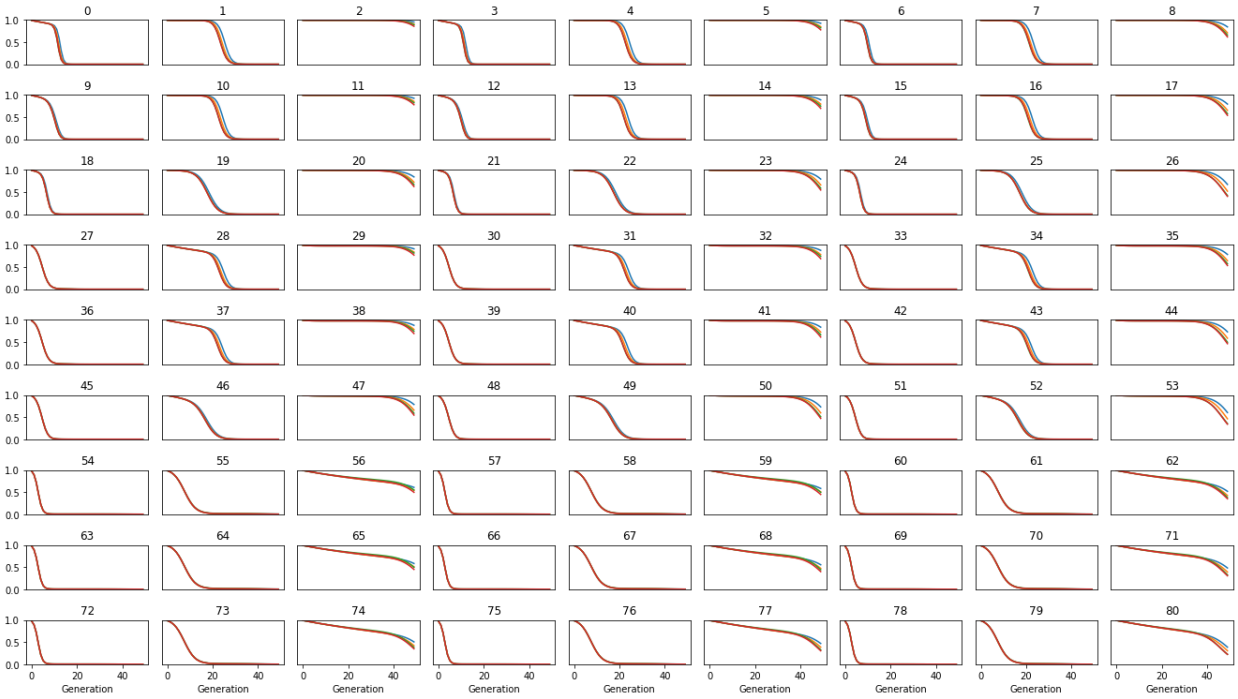


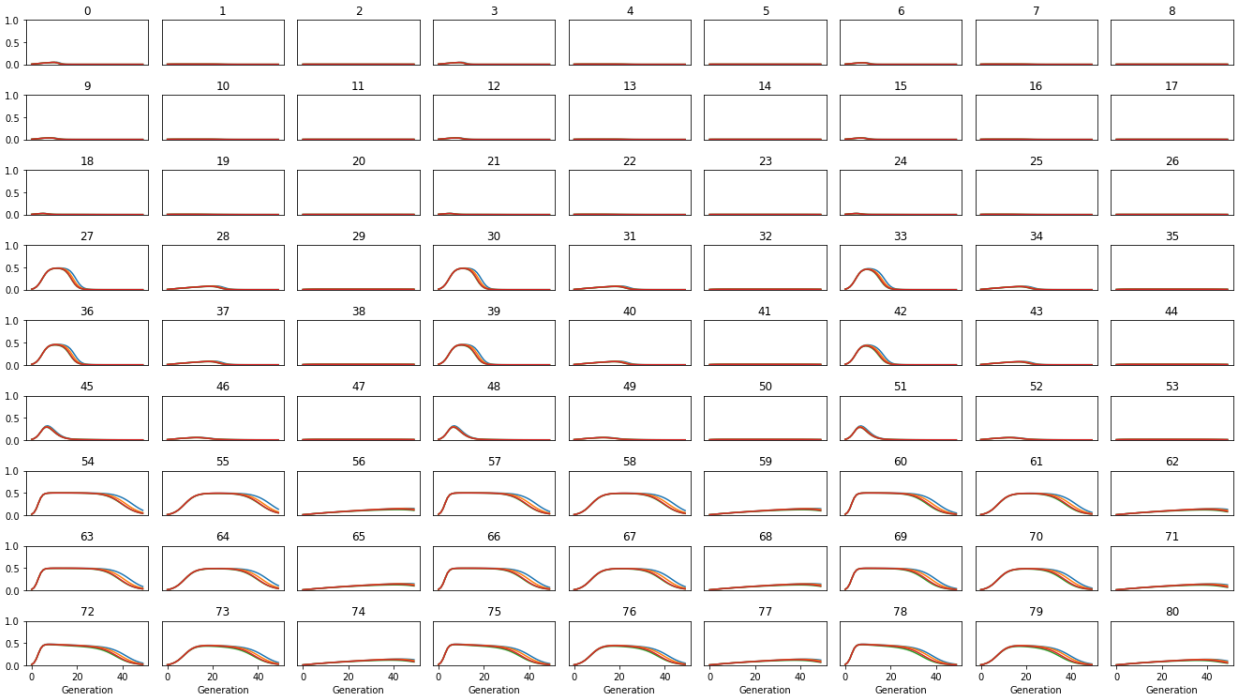
S2: Dendrogram resulting from hierarchical clustering of simulation results for step landscapes, low mutation rate. Leaf labels are landscape identification numbers from Figure 2 in the main text.

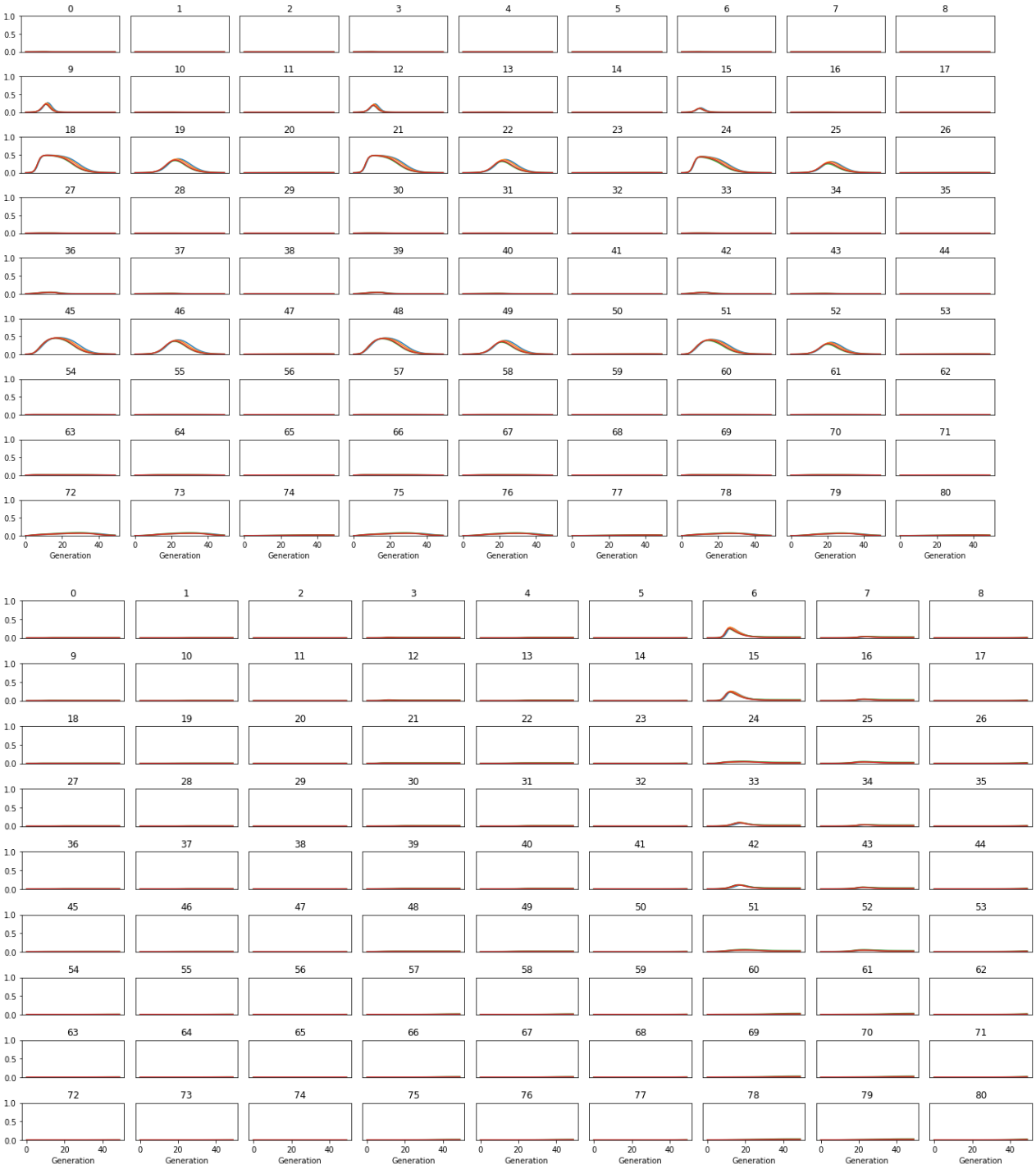

0 1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35
36 37 38 39 40 41 42 43 44
45 46 47 48 49 50 51 52 53
54 55 56 57 58 59 60 61 62
63 64 65 66 67 68 69 70 71
72 73 74 75 76 77 78 79 80
1.0
0.5
0.0
0 20 40
Generation


0 1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35
36 37 38 39 40 41 42 43 44
45 46 47 48 49 50 51 52 53
54 55 56 57 58 59 60 61 62
63 64 65 66 67 68 69 70 71
72 73 74 75 76 77 78 79 80
1.0
0.5
0.0
0 20 40
Generation

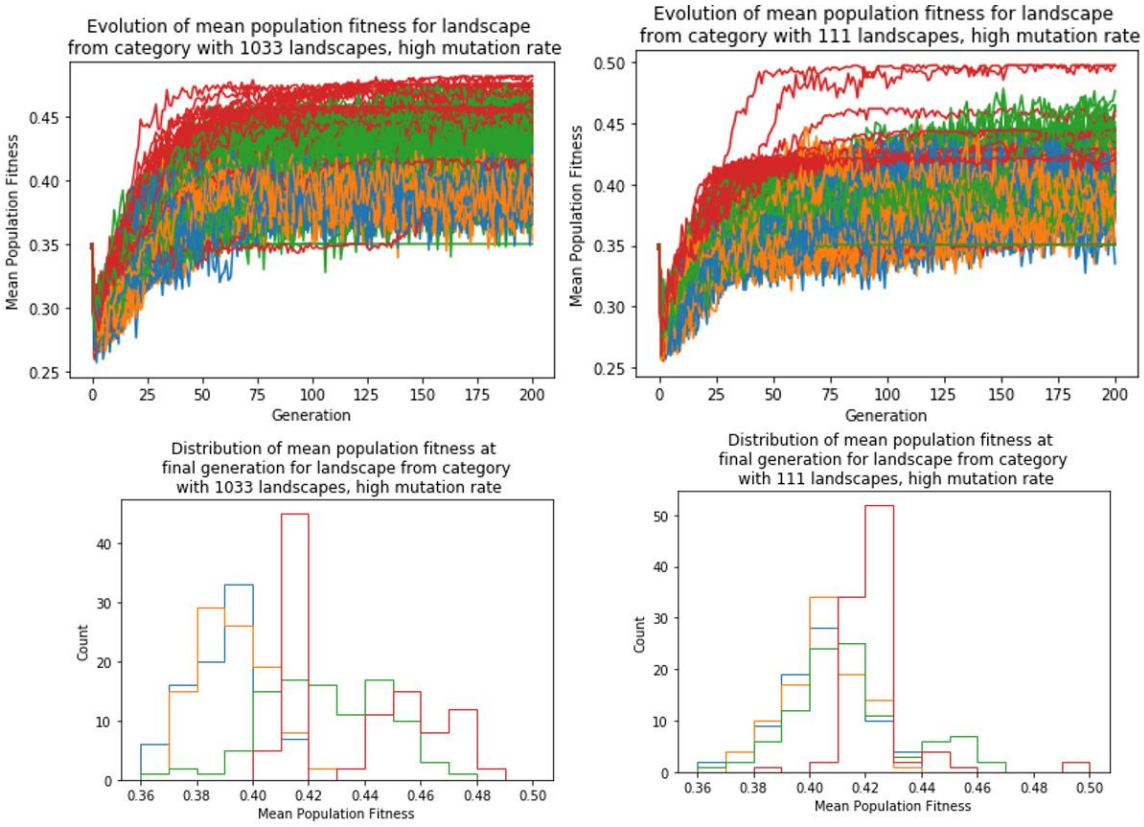
S3: From top to bottom, evolution of genotype frequencies: the initial genotype; k1 genotype; k2 genotype and k3 genotype, for the four studied mutation rate variation patterns and high mutation rate.


0 1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35
36 37 38 39 40 41 42 43 44
45 46 47 48 49 50 51 52 53
54 55 56 57 58 59 60 61 62
63 64 65 66 67 68 69 70 71
72 73 74 75 76 77 78 79 80
1.0
0.5
0.0
0 20 40
Generation

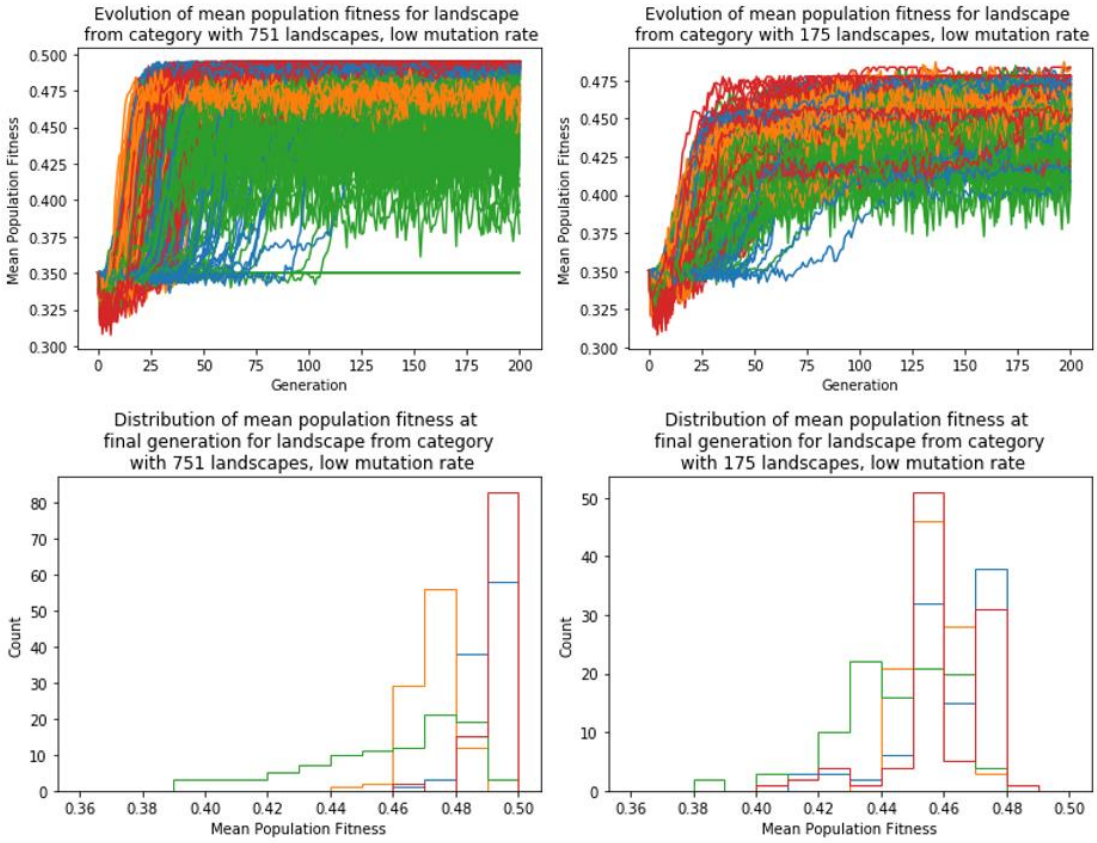

0 1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35
36 37 38 39 40 41 42 43 44
45 46 47 48 49 50 51 52 53
54 55 56 57 58 59 60 61 62
63 64 65 66 67 68 69 70 71
72 73 74 75 76 77 78 79 80
1.0
0.5
0.0
0 20 40
Generation



S4: From top to bottom, evolution of genotype frequencies: the initial genotype; k1 genotype; k2 genotype and k3 genotype, for the four studied mutation rate variation patterns and low mutation rate.



S5: Mean population fitness evolution along evolutionary time, and mean population fitness distribution at the final generation for the four studied mutation rate variation patterns, high mutation rate. On the left, for a representative landscape from the category with 1033 landscapes; on the right, for a representative landscape from the category with 111 landscapes. (see main text, section 3.2.1)



S6: Mean population fitness evolution along evolutionary time, and mean population fitness distribution at the final generation for the four studied mutation rate variation patterns, low mutation rate. On the left, for a representative landscape from the category with 751 landscapes; on the right, for a representative landscape from the category with 175 landscapes. (see main text, section 3.2.1)