Master Thesis

Experiments in In Silico Evolution with Aevol

Brian Davis

Supervisor: Berenice Batut

April 2020

Albert-Ludwigs Universität Freiburg

Bioinformatics Group

Professor Dr. Rolf Backofen

Abstract

Not all aspects of evolution are fully understood, and one area of active interest is reductive evolution, in which the genome of an organism evolves to become significantly smaller over time. Among the more well-known examples of this phenomena is *Prochlorococcus*, a marine cyanobacteria whose genome is up to 50% smaller than its closest living relative, *Synechococcus*. To study the mechanisms behind reductive evolution in the lab would be too costly, expensive, and slow, so we turn instead to *in silico* evolution. This method seeks to simulate organisms and their evolution in software, allowing for greater control of the environment, mutation rates, and other variables, as well as providing a full record of all organisms in a lineage. In this thesis we use the in silico tool *Aevol* to study reductive evolution, particularly by looking at how varying parameters (e.g. mutation and selection rates, population size) impacts the structure of the genome as well as measures such as robustness, evolvability, and fitness. Through these methods we hope to shed some light on the underlying mechanisms which lead to reductive evolution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Reductive evolution is the process of the genome of an organism shrinking over time, with respect to both the number of base pairs and genes. Some species of bacteria have experienced reductive evolution over the course of millions of years, and this reduction in their genome has lead to a loss of genes, certain regulatory abilities, etc. For example, some strains of the marine cyanobacteria *Prochlorococcus* have experienced a reduction of nearly 40% of their base pairs when compared to larger strains of their closest living relative, *Synechococcus*[23]. Despite being extensively studied, the mechanisms and full impact of reductive evolution are not fully understood and are an area of ongoing research.

Although it would provide more conclusive evidence, performing *in vivo* experiments is often impractical because of the difficulty or impossibility of reproducing natural environmental conditions in a lab. Such experiments are often too costly in terms of both time and resources. As an alternative, *in silico experimental evolution* is one option that can be used to study the conditions under which an organism's genome may become reduced. In this method, organisms and their evolution over thousands or millions of generations are simulated in software. In this manner, one can control and evaluate every aspect of their evolution over time and a full record of their lineage may be maintained and studied, allowing one to go back and closely examine every step of the evolutionary period for a greater understanding of the factors that lead to specific effects on the genome. The in silico tool *Aevol* is one such platform which realistically models bacterial genomes and

evolution, allowing one to draw conclusions about their real-world counterparts. In the following thesis, we present the results of our experiments in artificial evolution which aim to identify and evaluate several factors which potentially lead to changes in genome structure and a reduced genome in simulated bacteria using the Aevol platform.

1.1 Problem Statement

Among the difficulties of studying reductive evolution with in vivo evolutionary experiments, one of the most difficult obstacles to overcome is the lack of a full ancestral record. This lack of a full phylogeny can make it difficult or impossible to tell exactly when and how a specific event occurred, or a trait evolved or was lost, as illustrated in Figure 1.1 below. In this exam-

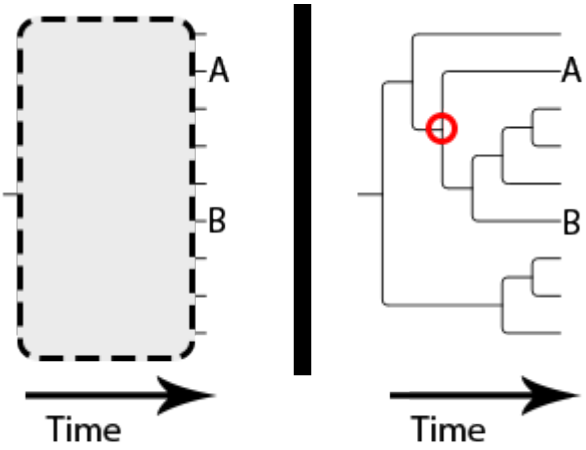

Figure 1.1: An illustration of unknown phylogeny. Since the phylogenetic information under the shaded box is typically not known, the point of divergence (red circle) can't be determined.

ple, we are comparing two related organisms A and B and we are trying to determine when and how a specific trait was gained or lost by one of the organisms. This may be useful, for example, if we are attempting to estimate the relative importance (due to conservation over many generations) of some trait. Without the phylogenetic information (under the shaded box) we may not be able to identify the point in their evolutionary history at which the two organisms diverged, making time estimates difficult or impossible.

Another major downside to *in vivo* evolutionary experiments is that they are slow. For example, the well-known E. coli Long-Term Evolution Experiment (LTEE) by Profesor Lenski at Michigan State University has been ongoing since February of 1988 and only passed generation 65,000 in 2016, 28 years later.

As an alternative to in vivo experiments, *in silico* evolutionary experiments are well-suited to the task of studying reductive evolution. Generations of organisms may be evolved within a very short time period, and a full "fossil record" of each lineage may be kept on disk for further analysis.

The in silico tool Aevol has a realistic artificial chemistry model which was developed specifically to study genome structure. It contains tools to analyze the robustness, fitness, and evolvability of digital organisms over time.

1.2 Report outline

This chapter serves as the introduction to the thesis and the research problem we are facing. In Chapter 2, we provide some necessary background information on reductive evolution, in silico evolution in general, and Aevol in particular. Chapter 3 describes our experimental setup. Chapter 4 provides the results and analysis of our experiments, and Chapter 5 presents our conclusions.

Chapter 2

Background

In this chapter we will examine some of the theoretical background information required for this thesis. We begin with an examination of experimental evolution and move on to a discussion of Aevol, the specific tool that was used in this thesis. We close the chapter by examining how Aevol can be used to study reductive evolution and discuss the current state of the literature surrounding reductive evolution.

2.1 Experimental Evolution

As discussed in Section 1.1, in vivo experiments, although sometimes more realistic, have their own set of difficulties. Some examples of these difficulties include recreating challenging environmental conditions (e.g. simulating the open ocean in a lab) and identifying and/or simulating the multiple selection pressures acting on genomes in the real world [3]. These challenges add enormous difficulty and complexity to conducting proper experiments and isolating the specific factors which lead to particular outcomes.

In silico evolution simulates organisms in software, thus allowing for far greater control and analysis of the environment and other experimental conditions. In contrast to in vivo experiments, a greater amount of control is also available with regard to the way organisms may interact, reproduce, and evolve. For example, a genome may be created completely from scratch or an existing genome may be fed into the simulation. Reproduction rates can depend on overall fitness, on relative fitness, or some other criterion.

Factors such as the mutation rates or selection pressure are then parameters for the model and may be kept constant or allowed to vary over time. Given that these are parameters of the system, they may be tightly controlled, leading to a clearer picture of the factors influencing different outcomes. An underlying deterministic model can also allow for a reconstruction of the system from any given point, allowing one to easily create a record of events, including phylogenetic trees.

Why use Aevol

Many silico evolution tools have been used to test various aspects of evolution: Tierra and Avida, in which the genetic units are computer programs fighting for CPU time [20] and [19] were some of the first; DOSE, an ecology-conscious method of checking for heterozygosity (variation within a population) [7] is a more recent example; and many more (see [18] for a more full review).

The in silico tool *Aevol* was developed to "study the evolution of the size and organization of bacterial genomes in various scenarios"[3]. The program has been expanded upon and tested in a variety of scenarios over the years. Examples of such experiments include: testing the predictability of evolution with high mutation rates as in viruses [4], determining whether selection is able to overcome evolution's drive towards more complex organisms [15], examining the role of mutators in reorganizing the genome in order to overcome mutational load [21], examining the effects of population shape on levels of cooperation [17], modeling regulatory networks [22] and more.

As an in silico experimental evolution tool, Aevol embodies several of the advantages of in silico evolution in general. There is a "fossil record" of each generation, experimental conditions are tightly controlled, and experiments are easily repeatable. The encoding/decoding strategy of Aevol follows a biologically realistic model, in the sense that there are many degrees of freedom between an organism's genome and its proteome. Many genes may encode for very simple proteins (e.g. if the genes contain the same or similar sequences), and by contrast, overlapping genes may code for complex proteomes.

In the following sections, the in silico experimental evolution tool Aevol

will be examined in greater detail.

2.2 Aevol

Aevol follows a "sequence-of-nucleotides" model [3] in which organisms are simulated with a binary genome which can either be generated at random or input as a previously-generated sequence. Aevol essentially consists of three steps: 1) decoding the genome of these organisms to produce artificial proteins, 2) selecting the most fit individuals and 3) reproduction of these fittest individuals with possible variations (mutations, rearrangements, etc.). In the sections below we will examine each of these steps in greater detail.

2.2.1 Aevol's Architecture

Aevol's three steps—decoding the genome, selection, and reproduction—are illustrated in Figure 2.1.

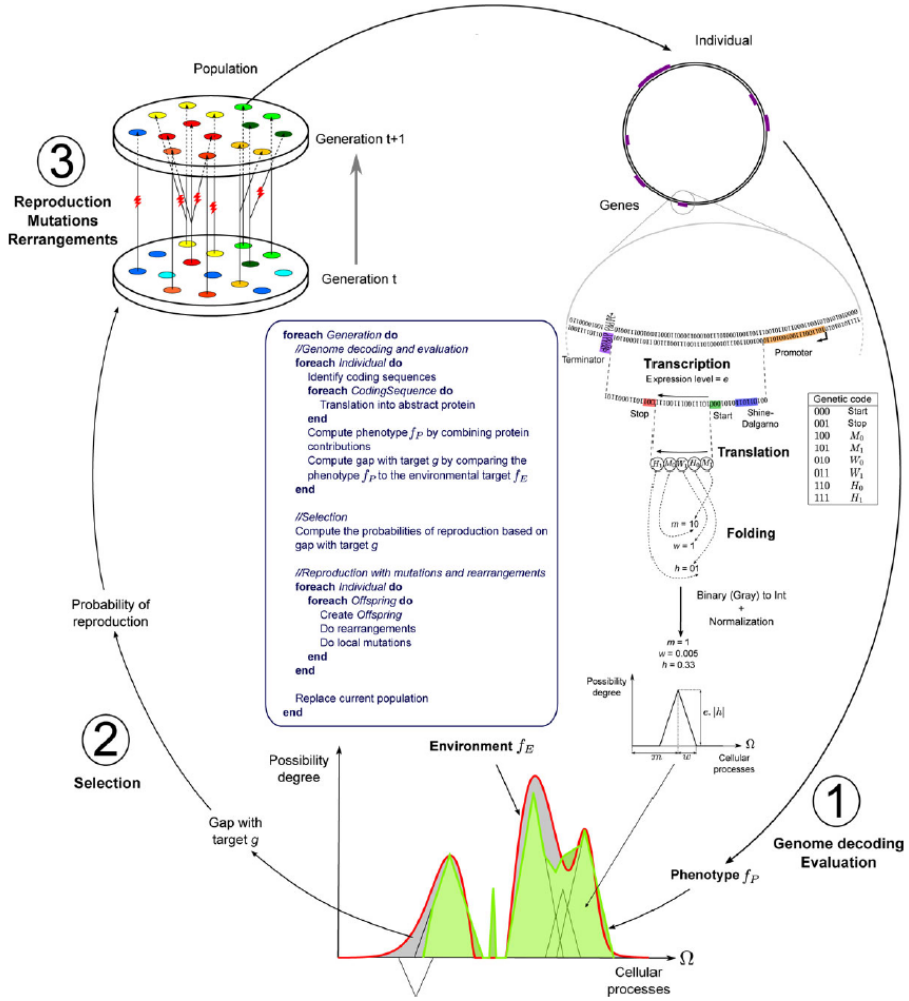

Figure 2.1: Overview of Aevol's architecture, from [3]

Decoding the Genome

In Aevol, a genome consists of a string of binary characters where 0 is complementary to 1. Each organism in the (initially clonal) population has a double-stranded circular genome which is either generated randomly or which was provided as input. To decode the genome and produce the phenotype, the sequence is searched for transcribed regions. Transcribed regions are denoted by promoter and terminator sequences. The promoter sequence is a sequence whose Hamming distance d is within $d_{max} = 4$ mismatches of the predefined consensus sequence 01010110011100100101 10. Termina-

tors are sequences which can form a stem-loop structure with a stem size of 4 bases and a loop length of 3 bases (i.e. $abcd{***}\overline{dcba}$ where a is complementary to $\overline{a}$, b is complementary to $\overline{b}$, etc.). Lastly, the initiation and termination signals are sought, which are simply Shine-Dalgarno-like signals (i.e. $011011 * * * *000$ to start and $011011 * * * *001$ to stop). Lastly, an expression level e is assigned to each coding region, following the formula $e = 1 - \frac{d}{d_{max}+1}$ where d is again the Hamming distance between the coding region and the consensus sequence given above and d_{max} is the maximum allowable distance (i.e. 4).

Once an initiation sequence is found, the following bases are read three at a time (codon by codon) until a stop codon (by default 001) is found. If a stop codon is not found, then no protein is produced. Since a transcribed region may have multiple initiation signals, operons are therefore allowed. The codons following an initiation signal encode for three parameters according to the genetic code given in Figure 2.1: m (mean), w (half-width), and h (height), which together define a triangle representing a "cellular process".

A cellular process is simply an abstract representation of some phenotypic function and is represented by the ordered set $\Omega = [a, b] \subset \mathbb{R}$. Together, these cellular processes make up the organism's proteome. To keep things simple, Ω is a one-dimensional space in the interval $[0, 1]$, i.e. a "cellular process" is simply a real number, and the genomic encoding of each cellular process determines the function $f(x) : \Omega \to [0, 1]$. The mean m gives us the specific cellular process in the range $[0, 1]$. The width w describes the "scope" of the process, i.e. the *pleiotropy* of the process, meaning the subset of the protein that is in the interval $(m - w, m + w) \subset \Omega$. The height determines the degree of possibility of the process, i.e. its relative strength.

The codons are read one after the other and their Gray codes[1] are used to compute the real numbers m, w, and h as follows. Each parameter (m, w, h) is assigned two codons in the genetic code (see Figure 2.1), for example $w_0 = 010$ and $w_1 = 011$. Any w_0 codons become a 0 in the Gray code, and vice versa with 1s. So if, for example, when reading the coding sequence, the codons w_1, w_0, w_1, w_0 are read, the Gray code becomes 1010, which is 12 in decimal. This is done for m, w, and h, and the resulting values

[1] A binary encoding such that two successive values (e.g. 2, 3) only differ by at most one bit (e.g. 0011, 0010). See `https://en.wikipedia.org/wiki/Gray_code`

are then normalized to be in the proper range. w's range is specified in the parameter file (as `MAX_TRIANGLE_WIDTH`), h must be in the range $[-1, +1]$ (indicating that both activating and inhibiting processes are allowed) and m must be in the range $[0, 1]$ (the range of Ω).

Given the fact that there are likely multiple coding sequences in a genome, several triangles (cellular processes) are translated from the genome, each parameterized by its own m, w, and h. These triangles form the phenotypic function f_P. *Fuzzy logic* is used to find the overall contribution of each cellular process, using the Lukasiewicz fuzzy operators[2]. Roughly speaking, the activating proteins are added up, as are the inhibiting proteins, and the difference between these two totals represents the final function f_P. More formally, if f_i is the possibility distribution of the i-th activator protein and f_j is the possibility distribution of the j-th inhibitor protein, then the phenotype of the individual is defined as:

$$f_P = max\left(min\left(\sum_i f_i(x), 1\right) - min\left(\sum_j f_j(x), 1\right), 0\right)$$

Selection

After the genome is decoded, the organisms are tested for their fitness. Fitness in Aevol is related to the gap between the phenotype of a sequence f_P and the environmental target function f_E, as illustrated in Figure 2.1. This environmental target function f_E is a user-defined set of Gaussians which are specified in a parameter file, with each Gaussian being identified by three parameters: its height, its location along the axis, and its width. The difference between the phenotype (as calculated above) and the environmental function is the "metabolic error", labeled g in the figure and is more formally defined as: $g = \int_a^b f_E(x) - f_P(x)dx$. The idea is illustrated in Figure 2.2

Aevol contains several selection schemes but here we will only consider the `fitness_proportionate` scheme, since this was the only selection scheme employed in our experiments. In this scheme, the probability of

[2]See `https://en.wikipedia.org/wiki/Lukasiewicz_logic` for an introduction.

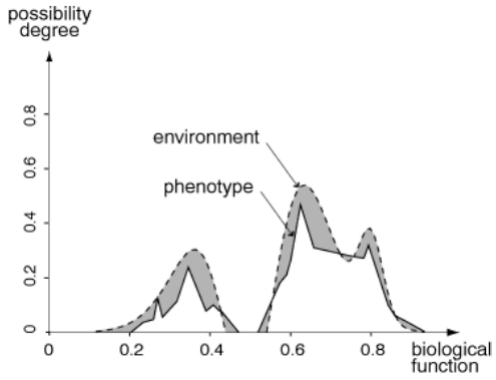

Figure 2.2: Overview of Aevol's conception of fitness, from [12]

reproduction for each organism i is proportionate to its fitness, namely:

$$P(\text{reproduction}) = \frac{e^{-k*g}}{\sum_{i=1}^{N} e^{-k*g_i}}$$

where k is a user-definable parameter which determines the selection intensity and g is the metabolic error.

Reproduction

Once the fittest organisms in the population are found and their probabilities of reproducing are calculated as described in the previous Section, new organisms are produced. This is done for each potential parent organism by drawing from a multinomial distribution with the probability of reproduction given above. The population size N is kept constant and a record of each generation is kept so that the phylogenetic lineages can be recreated. Since the population size is held constant, this implies that a single organism with a high probability of reproduction may produce multiple offspring and an organism with low probability of reproduction may produce none.

When new organisms are created and their genomes are copied from their parent organisms, it is at this stage that some of the driving forces in evolution occur, namely the possibility for variation through mutation, indels, and frameshifts. Offspring will receive their parents' genome but their genome may be subject to perturbations due to stochastic effects. Mutation rates are set in the parameter file and include point mutations, insertions

and deletions (indels), and rearrangements (duplication, deletions, translocations, and inversions).

The mutation, indels, rearrangement, etc. events are carried out by first determining the number μ of such events which will occur, based on the mutation rate specified in the parameter file and drawing from a binomial distribution (e.g. $B(L, \mu_{\text{point}})$ for point mutations, $B(L, \mu_{\text{large deletions}})$ for large deletions, etc. where L is the size of the genome). Then a random point (or points, in the case of e.g. rearrangement) is chosen and the event is carried out, with the order of these events shuffled randomly.

2.3 Analyzing with Aevol

Once the experiments have completed, Aevol by default produces several statistics files which include information about genome size, the percentage of coding DNA, number of genes, average metabolic error, and many other statistics. It further includes a number of post-treatments that allow one to analyze specific individuals or the population at large, including tools for determining robustness, evolvability, coalescence, and the lineages.

One of the key features of Aevol is the ability to look back in time at the "archaeological record" of previous organisms, which is stored on disk, in order to perform various analyses. This is primarily done with a myriad of post-treatments", i.e. supplementary programs. These post-treatments generally require a `lineage` file, which shows a record back in time of the line of descent for an individual. One may specify either the best-ranked individual (i.e. the fittest) or a specific individual by their unique identification number. The basic idea is illustrated below in Figure 2.3.

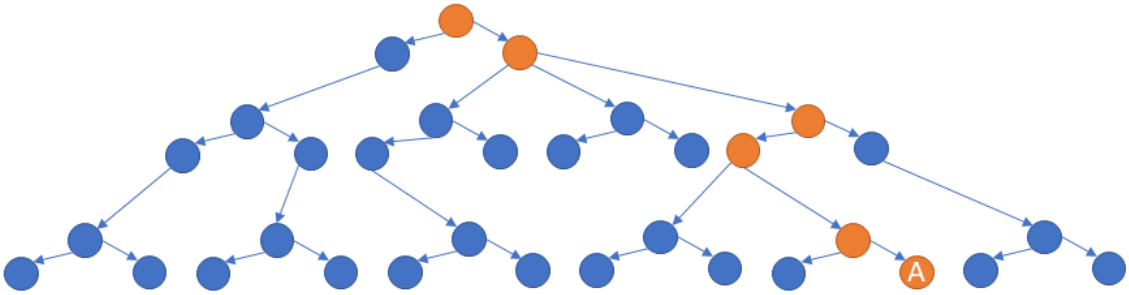

Figure 2.3: A basic illustration of a lineage. The ancestors of the individual (labeled 'A') can be traced back through the previous generations for all of its ancestors (all in orange).

In the following subsections we will examine other statistics that Aevol calculates, as these factors play a major role in reductive evolution.

2.3.1 Evolvability

Evolvability is usually defined as the ability of a system (in this case an organism) to evolve. In other words, a system has evolvability "if mutations in it can produce heritable phenotypic variation"[25]. However, of critical importance is that this is not simply having a large amount of genetic diversity, but rather *adaptive* diversity which provides some benefit.

2.3.2 Robustness and Antirobustness

Robustness is broadly defined as the ability of an organism to withstand disruptions or perturbations without affecting its phenotype. There are differing kinds of robustness, as well: one may describe robustness in terms of mutational robustness–which describes the extent to which an organism's phenotype is not affected by stochastic mutational events–or one may speak of environmental robustness, which describes the ability of an organism to maintain its phenotype in diverse environments with little or no loss in fitness.

Also important is the idea of antirobustness, wherein an organism may actually *thrive* on such perturbations. This has been suggested as a possible method of minimizing the effects of *Muller's ratchet*, wherein deleterious

mutations accumulate in a population as a result of genetic drift[11]. A large number of deleterious mutations may provide fodder for selection to fix more beneficial mutations in the population[21].

Robustness vs. Evolvability

There is, then, a seeming trade-off between robustness and evolvability. The more robust a system is, the less phenotypic variation is generated by random mutation events, and thus less evolvability. However, a key factor is to distinguish between genomic and phenotypic robustness; a strong phenotypic robustness promotes structural evolvability, as the likelihood that a mutation is deleterious is smaller in populations with more robust phenotypes. For a fuller discussion, see [25].

2.3.3 Fitness

2.4 Related Work

Much work has already been done in the field of reductive evolution, and in this section we will look at the current state of the literature.

Liard et al. showed in [15] that selection for fitness is not necessarily enough to overcome the tendency of organisms to become more and more complex. They describe the problem of a "complexity ratchet," that is, that the tendency of organisms becoming more and more complex as irreversible once the organisms had reached a certain complexity level. In their words:

> "Since gene deletion is obviously deleterious [in this scenario], the only available evolutionary path for already complex organisms is a headlong rush toward increasing complexity by acquiring new genes. Hence the ratchet clicks, further widening the fitness valley that separates the current genome from a simple one, soon making it so wide it is very unlikely to be crossed."

Echoing the findings of Knibbe et al. [13] they found, however, that limiting *robustness* can overcome the tendency of organisms to become more and more complex, because this places an upper bound on the amount of information that an organism can transmit in its genome. Increasing the mutation rate forced gene elimination despite the fitness loss because it lowered

the information content of the genome. A mutation rate of even $\mu = 10^{-4}$ resulted in nearly 40% of their organisms developing a simpler genome.

Knibbe et al. also found, via in silico experimentation, that the accumulation of non-coding DNA strongly depends on the mutation rate. This in turn affects the selection tradeoff between reliably passing on the existing genome and having the mutational variability to adapt to new challenges. Under higher mutation rates, their organisms closely resembled viral genomes in that they had almost no non-coding sequences. When the selection strength was larger, genomes were larger.

Koskiniemi et al. [14] performed in vivo experiments on the bacterium *Salmonella enterica* in which the effects on fitness of random deletions was measured. Some 25% of the deletions actually caused an increase in fitness under some conditions, suggesting that there is a certain cost associated with having superfluous genes and thus gene loss may be selected for under certain conditions.

Chapter 3

Methods

With an understanding of the basics behind us, in this chapter we provide an overview of our contributions and proposed solution.

3.1 Contributions

3.2 Experimental Designs

To assist in determining which conditions might lead to reductive evolution, we need to isolate individual variables and change just one thing at a time in order to see what effect, if any, it has on the final genome. To that end, we designed and conducted a series of experiments in which a "wild type" genome was allowed to evolve in differing conditions for 500,000 generations before analyzing the effects. To first create the wild type, a genome was generated randomly in Aevol which had at least one coding gene and which was allowed to evolve for 10 million generations in a non-varying environment. By the beginning of its use in our experiments, the wild type comprised 13,237 base pairs with 132 functional genes (i.e. genes which produce a gene product).

In our experiments, we allowed the wild types to continue to evolve in the same environment in a total of 6 different conditions: with an increased/decreased selection strength, an increased/decreased mutation rate, and an increased/decreased population size. For all experiments the environmental target function did not vary with respect to the time of the

wild type generation nor with respect to any of the performed experiments. The three Gaussian functions are characterized by a height h, mean m, and width w:

$$y(x) = h * e^{\left(\frac{-(x-m)^2}{2*w^2}\right)}$$

An approximation of the target function used in our experiments, f_E, can be seen below in Figure 3.1.

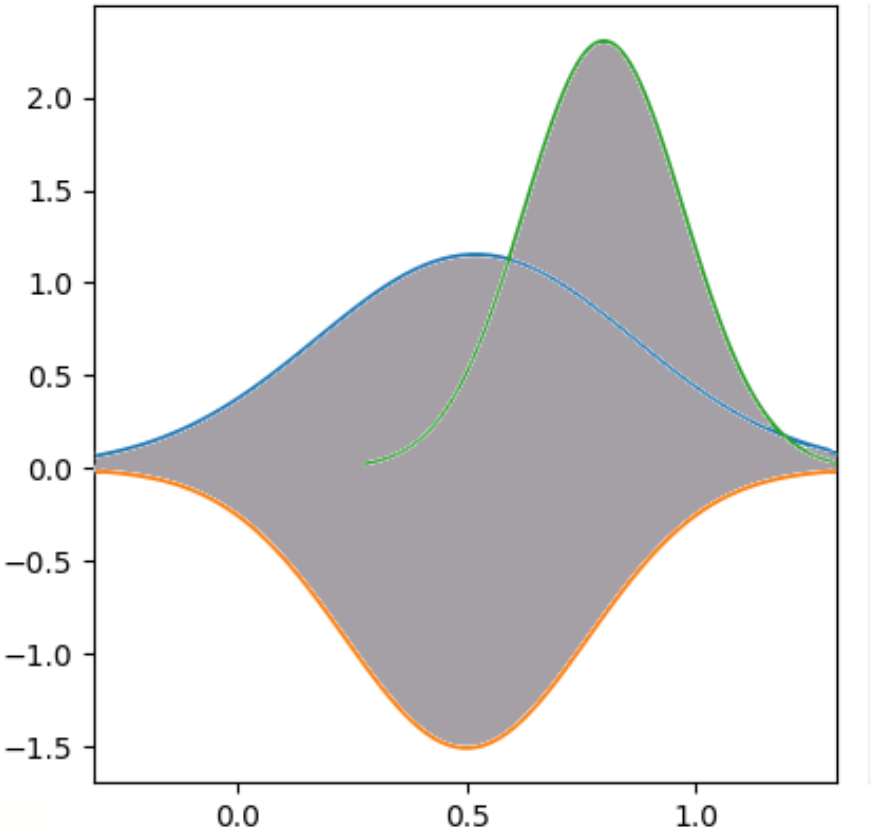

Figure 3.1: A visual approximation of the target function f_E for our experiments.

The parameters for the three Gaussians are given in the table below.

	h	m	w
Gaussian 1	1.2	0.52	0.12
Gaussian 2	-1.4	0.5	0.07
Gaussian 3	0.3	0.8	0.03

Table 3.1: The Gaussian environmental target function parameters.

In addition to the six tested changed conditions, a control condition was performed in which the wild type was simply allowed to continue to evolve for the 500,000 generations with no change in any of the parameters from the generation of the wild type. To minimize bias, for each condition we performed five runs each (i.e. 5 rounds of mutation up, 5 rounds of mutation down, etc.), where each run had a differing random seed to control for the deterministic effects of the pseudorandom nature of Aevol's stochastic processes. This lead to a total of 35 experiments, all of which were carried out on a cluster from bwCloud[1].

The resulting data was processed using a combination of Python[2], Pandas[3], and Jupyter Notebooks[4].

3.2.1 Inputs

In Table 3.2, our parameter values for all input parameters may be found. Please note that for μ, this represents the mutation rates for point mutations, small insertions, and small deletions. The rearrangement rates were not changed under any condition and were always $1e-6$ for duplications, deletions, translocations, and inversions.

As can be seen in Table 3.2, only one parameter varied per condition in order to isolate potential influences on the outcome. A multiplier of 4 was chosen for the differing conditions relative to the control condition (e.g. $N_{\text{population up}} = 4096 = 4 * N_{\text{control}} = 4 * 1024$). This choice of a 4x multiplier was chosen following the model of Carde et al.[6]. In all conditions, the environment did not vary, and once the experiment was begun, the above parameters were held steady as well.

[1] `https://www.bw-cloud.org/`
[2] `https://www.python.org/`
[3] `https://pandas.pydata.org/`
[4] `https://jupyter.org/`

Condition	Parameter		
	μ	k	N
control	$1.00E{-}7$	1000	1024
mutation up	$4.00E{-}7$	1000	1024
mutation down	$2.50E{-}8$	1000	1024
selection up	$1.00E{-}7$	4000	1024
selection down	$1.00E{-}7$	250	1024
population up	$1.00E{-}7$	1000	4096
population down	$1.00E{-}7$	1000	256

Table 3.2: Table of input parameters. μ is the mutation rate, k is the selection strength, and N is the population size.

3.3 Evaluation Strategy

In this section, we will examine the criteria we will be using to evaluate the results. The primary criteria will be examining the evolved genome's evolvability, robustness, and structure, with a statistical analysis of the changed conditions vs. the control condition.

3.3.1 Statistical Analysis of the Conditions

We must first determine how to tell whether the results of the condition (e.g. mutation up, selection down, etc.) were significantly different from the control condition, statistically speaking. To do this, for some test variable (e.g. robustness, evolvability, etc.) we will first calculate the mean of all seeds for that condition and compare it to the mean across all seeds of the control condition. For the comparison, we will use the "Mann-Whitney U" test. The Mann-Whitney U test is similar to the Wilcoxon signed-rank test, but it is used when the distribution of the two samples cannot be assumed to be normally distributed. The Mann-Whitney U test can also be used when the sample sizes are different, for example when the number of base pairs is different between two organisms.

The purpose of the Mann-Whitney U test is to check whether two independent samples were selected from populations having the same distribution. Similar to other rank-sum tests, the test consists of the following

steps:

1. Assign numeric ranks to all observations;
2. Add up the ranks for the observations from the first sample, giving us R_1
3. The statistic U_1 for the first sample is given by:

$$U_1 = R_1 - \frac{n_1 * (n_1 + 1)}{2}$$

where n_1 is the sample size for the first sample and R_1 is the sum of the ranks from the first sample.

The U statistic for the second sample (i.e. U_2) is computed analogously. We then use $U = U_1 + U_2$ which, for sample sizes greater than 20, is approximately normally distributed. Then the standard score is given by:

$$z = \frac{U - m_U}{\sigma_U}$$

where m_U and σ_U are the mean and standard deviation of U.

$$m_U = \frac{n_1 * n_2}{2} \text{and } \sigma_U = \sqrt{\frac{n_1 * n_2 \left(n_1 + n_2 + 1\right)}{12}}$$

We used the Python library function `scipy.stats.mannwhitneyu` to calculate this statistic, which also calculates the p-value. If the p-value is below 0.05, we may reject the null hypothesis H_o that the two samples (i.e. the control and the condition) are from the same distribution.

3.3.2 Fitness

Fitness in Aevol is closely tied in to the "metabolic error". This error, g, is calculated as the gap between the environmental function f_E and the phenotype of the organism, f_P. Once g is determined as described above in Section 2.2.1, it may be used to calculate the actual fitness of the organism according to the equation:

$$\text{fitness} = exp(-k * g)$$

where k is the *selection coefficient* variable set by the user in a parameter file. Aevol provides fitness statistics both for the fittest individual and for the population at large.

3.3.3 Robustness

Aevol calculates statistics for both mutational robustness as well as antirobustness. Robustness in Aevol is calculated similarly to evolvability: a `lineage` file for an individual is fed to the post-treatment `misc_ancestor_robustness`, which generates a large number (specifiable by the user) of offspring, whose fitness is then measured. The percentage of these offspring which are neutral (i.e. whose phenotype is not affected by the mutations) is the robustness, and the percentage of positive offspring determines the antirobustness.

In our experiments, at the end of the run of 500,000 generations we generated a lineage file for the best individual at generation 500,000, i.e. the individual whose metabolic error was smallest. This lineage file is then fed in to the post-treatment `aevol_misc_ancestor_robustness` and robustness statistics are generated for each generation in the lineage file. The statistics produced by this post-treatment are summarized in Table 3.3 below.

Statistic
Fraction of positive offspring
Fraction of neutral offspring (aka reproductive robustness)
Fraction of neutral mutants (aka mutational robustness)
Fraction of negative offspring
Cumulative delta-gaps of positive offspring
Cumulative delta-gaps of negative offspring
Delta-gap for the best offspring
Delta-gap for the worst offspring
Cumulative delta-fitness of positive offspring
Cumulative delta-fitness of negative offspring
Delta-fitness for the best offspring
Delta-fitness for the worst offspring

Table 3.3: Table of robustness statistics calculated by Aevol for the best individual with the provided lineage.

We may then compare these statistics for both the control and specific condition we wish to compare (e.g. mutation up). Because this data is somewhat noisy (owing to the fact that the fitness may change rapidly)

we will use box and whisker plots to show the overall spread rather than plotting the robustness generation by generation.

Evolvability

In Aevol, evolvability is calculated by generating a large set of offspring for a specific individual (one whose lineage was generated using the post-treatment `aevol_misc_lineage`) at regular periods along their lineage and then determining the number of "positive offspring". Positive offspring are defined as those whose fitness is greater than their parent's. The evolvability of an individual is then the sum total of all improvement of all of the beneficial offspring, i.e.:

$$\text{evolvability} = \frac{|\text{positive offspring of } i|}{|\text{total offspring of } i|} * \sum \Delta_{fitness}^{\text{positive offspring}}$$

where $\Delta_{fitness}^{\text{positive offspring}}$ is the cumulative sum of the fitness increase for the positive offspring. Thus, evolvability in Aevol accounts for both the likelihood of a positive mutation and the average improvement provided by said mutation. Practically speaking, to find an organisms evolvability in Aevol one must give the post-treatment `misc_ancestor_robustness` a lineage file and then multiply the fraction of the number of positive offspring (column 2) by the cumulative total of the fitness gap g of the positive offspring (column 10).

3.3.4 Structure

Another strength of Aevol is its ability to analyze changes in the structure of DNA and RNA. As with fitness, Aevol produces statistics about individuals and the population at large for many aspects of genome structure, including: the number of coding vs. non-coding bases (i.e. they respectively do or do not code for at least one protein), the average size of coding and non-coding DNAs, the number of genes, the number of "essential" base pairs (i.e. those that are part of a functional coding sequence), etc. Tables 3.4 and Table 3.5 summarize the different statistics and where they are to be found Aevol's output files.

Stat File	
`stat_genes_`⟨`best/glob`⟩	`stat_bp_`⟨`best/glob`⟩
number of coding RNAs	number of bp not in any CDS
number of non-coding RNAs	number of bp not included in any functional CDS
average size of coding RNAs	number of bp not included in any non-functional CDS
average size of non-coding RNAs	number of bp not included in any RNA
number of functional genes	number of bp not included in any coding RNA
number of non-functional genes	number of bp not included in any non-coding RNA
average size of functional genes	number of non-essential bp
average size of non-functional genes	number of non-essential bp including non-functional genes

Table 3.4: Statistics found in `stat_genes` and `stat_bp` output files from Aevol. ⟨best/glob⟩ indicates that statistics are available for both the best individual and the average across the whole population.

"Essential" base pairs are those whose mutation would change the phenotype of the organism.

Stat File	
stat_fitness_⟨best/glob⟩	**stat_mutation_⟨best/glob⟩**
population size	number of local mutations
fitness	number of chromosomic rearrangements
genome size	number of switches
metabolic error	number of indels
parent's metabolic error	number of duplications
metabolic fitness	number of deletions
secretion error	number of translocations
parent's secretion error	number of inversions
secretion fitness	
amount of compound present in grid-cell	

Table 3.5: Statistics found in `stat_fitness` and `stat_mutation` files from Aevol. ⟨best/glob⟩ indicates that statistics are available for both the best individual and the average across the whole population.

3.4 Expected Results

Given the state of the literature and other experiments, as partially described in Section 2.4, the table below summarizes our expected results.

Experiment Predictions						
Result	**Condition**					
	μ_+	μ_-	k_+	k_-	N_+	N_-
Genome Size	$-^{[5,\ 16]}$	$+^{[5,\ 9]}$	$+^{[3]}$	$-^{[3]}$	$-^{[2]}$	$+^{[2]}$
Fitness	$+^{[1,\ 24]}$	$+^{[24]}$	$+^{[2]}$	$-^{[2]}$	$+^{[8,\ 24]}$	$-^{[8,\ 24]}$
Amount of non-coding DNA	$-^{[13]}$	$+^{[13]}$	$+^{[3]}$	$-^{[3]}$	$-^{[3]}$	$+^{[3]}$
Number of genes	$-^{[13]}$	$+^{[13]}$	$-^{[13]}$	$+^{[13]}$	$-^{[2]}$	$+^{[2]}$
Average size of genes	$-^{[15]}$	$+^{[15]}$	$-^{[3]}$	$+^{[3]}$	$-^{[2]}$	$+^{[2]}$
Robustness	$-^{[13]}$	$+^{[13]}$	$-^{[3]}$	$+^{[3]}$	$-^{[10]}$	$+^{[10]}$
Evolvability	$+^{[13]}$	$-^{[13]}$	$+^{[3]}$	$-^{[3]}$	$-^{[26]}$	$+^{[26]}$

Table 3.6: Our predictions for the experiments. μ is the mutation rate, k is the selection rate, and N is the population size. μ_+ indicates an increased mutation rate, μ_- a decreased rate, etc. A $+$ in the main grid space indicates we expect an increase (over the control condition) for that condition, and a $-$ indicates an expected decrease for that condition.

Chapter 4

Results and Discussion

This chapter presents the results of our experiments as performed according to the description in Chapter 3. We then discuss our results and conclude the chapter by relating our results to real-world organisms, as well a short discussion of potential limitations of the experiments.

4.1 Results

In this section we present the results of our experiments. We begin by presenting the results of our statistical analysis before moving on to show the results in various plots. It is of minor note that in many of the figures, a rolling window was used to smooth the values, resulting in many of the plots only showing data starting after a few thousand generations (often 10,000). It should be noted, however, that organisms were continuously evolved for 500,000 generations.

4.1.1 Genome Size

In this section we will examine the results of the different conditions, focusing on which, if any, lead to a reduced genome. Figure 4.1 presents the main findings regarding genome size. In the figure, the blue line represents the control condition and the other colors show the changed conditions: mutation up/down, selection up/down, and population up/down. As can be seen from the figure, our expectations in Section 3.4 did not hold up, as all conditions actually *increased* over their original size.

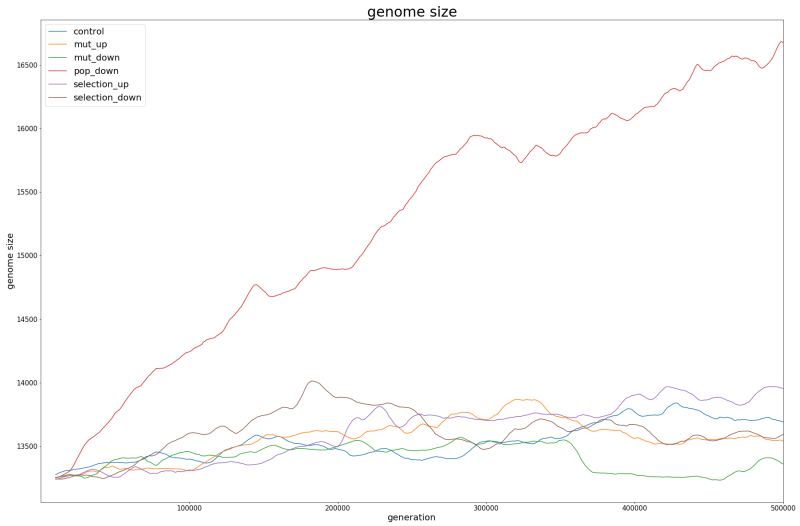

Figure 4.1: Genome size in number of bases of all conditions. Average taken across all five seeds for each condition.

In fact the *population down* condition had a runaway increase in the number of bases, reaching over 16,500 bases, at least a 25% increase over the original wild type's roughly 13,200. Surprisingly, even after 500,000 generations it seems that the upper limit may still not have been reached. The statistical results are given below in Table 4.1.

Genome Size		
	rank sum U	**p-value**
Mutation Up	110508011469.5	< 0.01
Mutation Down	71008638349.5	< 0.01
Population Down	16477791900.5	< 0.01
Selection Up	100151680984.5	< 0.01
Selection Down	88533681875.0	< 0.01

Table 4.1: Genome size statistics. Each condition is compared with the *control* condition.

The next most obvious observation is that of the remaining conditions,

all but the *selection up* condition (which had an an increase of 11% over the control condition) ended up with fewer base pairs than the control condition, though all increased slightly over their starting point. Also noteworthy is that the mutation down condition appears to have had a steady increase in the number of base pairs until a maximum of just over 14,000 around generation 350,000 before having a fairly sharp decline back to nearly the original size.

Examining the percent changed from the *control* condition shown in Figure 4.2 we see that at points, the *mutation down* condition was nearly 5% smaller than the *control* condition.

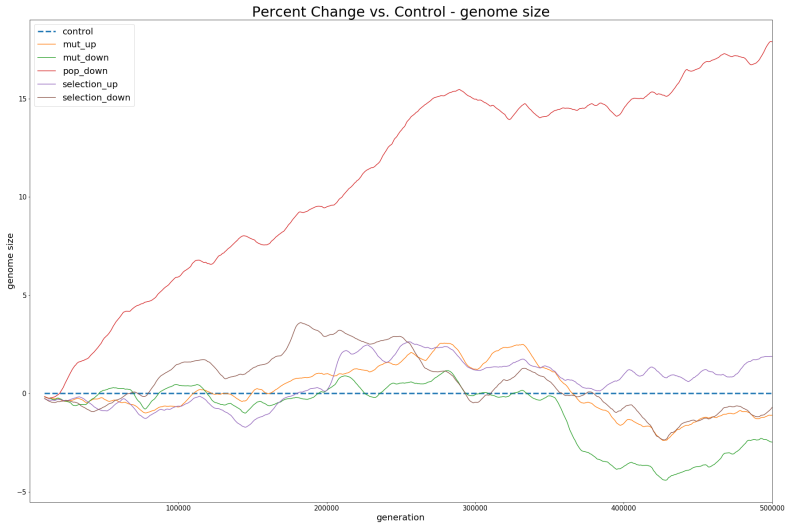

Figure 4.2: Genome size's percent change from the *control* condition.

4.1.2 Metabolic Error and Fitness

Figure 4.3 shows the mean fitness of the population for the control, mutation up/down, and population down conditions. Selection up/down were excluded from this graph because they were significantly outside of the range of the other conditions and made the results impossible to graph, but they are included in Figure 4.4, a histogram of the results.

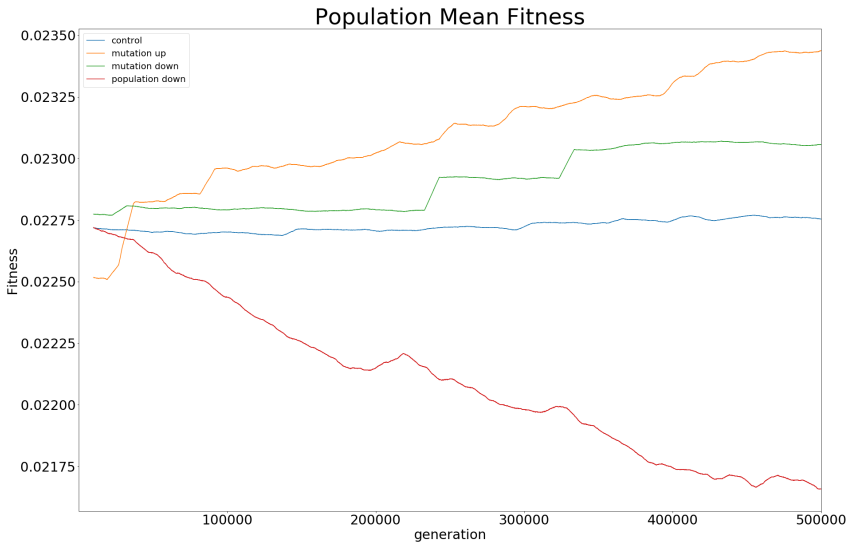

Figure 4.3: Plot of the mean fitness over time for the control, mutation up/down, and population down conditions.

We can see in the figure that the fitness of the control condition (in blue) pretty consistently stayed at the same level, likely owing to the fact that the wild genotype had already been allowed to evolve for 10 million generations in this environment, so the phenotype closely lines up with the target before the simulations even began. Small fluctuations occurred due to mutations, insertions, etc. but since we were beginning with a clonal population of the best organism after 10 million generations in this environment, the average fitness remained steady.

More interesting is the population down condition, where the average fitness in the whole population sharply declined, sinking to 4.5% below the control condition at generation 500,000 . It seems that with the smaller population size, genetic drift may be more strongly at work in continually increasing the gap between the phenotype and environmental function, as the lack of variety inherent in a smaller population causes a cascade of increasingly deleterious consequences.

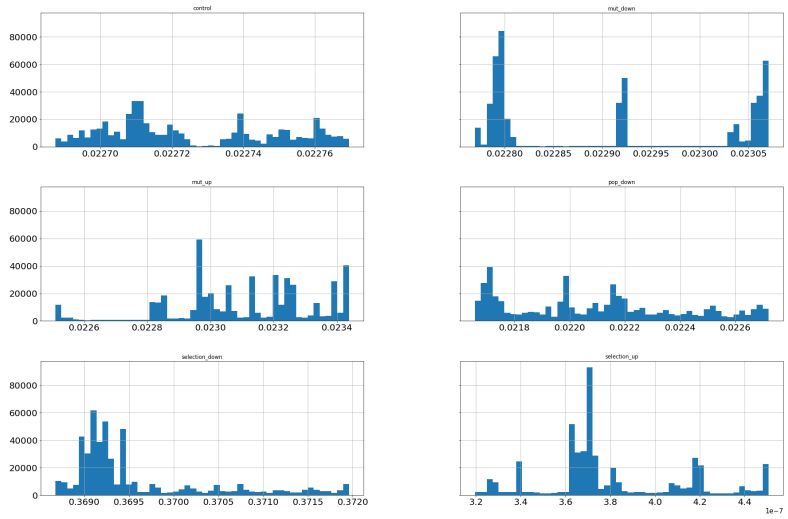

Figure 4.4: Histogram of all conditions' fitness. Note that for the selection up condition, the scale is 1e-7.

Figure 4.4 shows a histogram of all fitness values for all conditions. Although the *mutation down* condition ended the simulations in the most similar position to the *control* condition, Figure 4.4 makes it clear that the *control* condition remained much steadier throughout the simulation, regularly jumping from value to value, whereas the *mutation down* condition, as expected, tended to hover at certain locations in between rarer mutation events.

Figure 4.5 shows the mean metabolic error across all seeds for the whole population over time with respect to each condition.

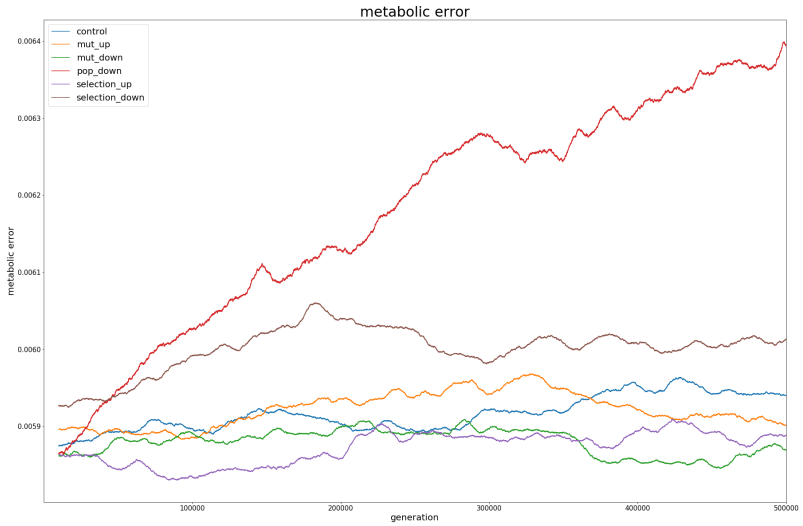

Figure 4.5: Plot of the metabolic error over time for all conditions, average of all seeds

Table 4.2 below gives the mean fitness scores for differing conditions. The selection up/down conditions seem to be somewhat anomalous.

	mean	**standard deviation**
control	0.022725416617759595	2.3443283583311204e-05
mutation up	0.02310700617283088	0.00021918255246890017
mutation down	0.022908786914463554	0.0001180940074554785
selection up	3.803823317637301e-07	3.130503751740127e-08
selection down	0.3695935240107418	0.0008101731945386823
population down	0.02209722426327717	0.00031042637566743275

Table 4.2: Fitness means and standard deviations across all seeds.

4.1.3 Genome Structure

In this section we examine the effects of the differing conditions on the structure of the genome as measured by the criteria in Tables 3.4 and 3.5.

Non-coding DNA

One important factor in genome structure is the amount of non-coding DNA, i.e. the number of bases which are part of a genome but which do not encode protein sequences. Aevol gives the number of bases which are not in any coding sequence, as well as the total genome size, so one may easily calculate this percentage. The results from our experiments are shown in Figure 4.6. Confirming previous suspicions, it seems that as the genome size of the population down condition rapidly expanded, most of those expansions were non-coding.

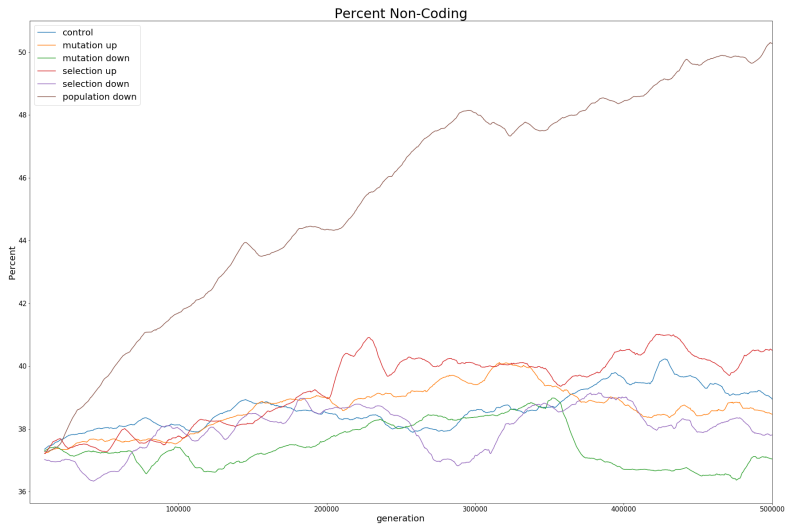

Figure 4.6: Plot of non-coding DNA over time for all conditions, average of all seeds.

Number of Genes

Figure 4.7 below illustrates the mean number of genes across the population for each condition.

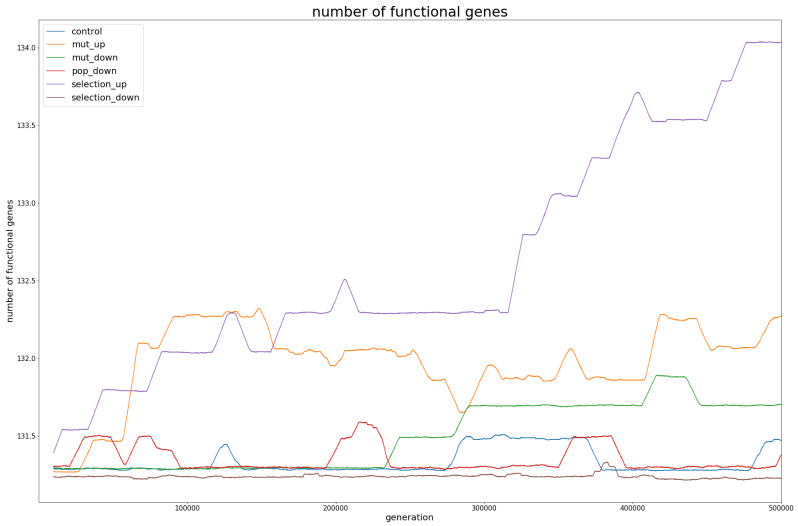

Figure 4.7: Plot showing the mean number of functional genes over time across all seeds.

As seen in the figure, the increased selection condition showed the greatest increase in the number of functional genes in the whole population, about 3% (130 to 134). Interestingly, the selection down condition did not change the number of genes at all, which is to be expected, since with a lower selection pressure, any newer genes may not be selected for reproduction. Whereas most of the other curves are fairly flat, the mutation up condition fluctuates relatively rapidly, owing to the quick increase and decrease in the number of base pairs.

Average Size of Functional Genes

Figure 4.8 shows the average number of base pairs for each functional gene for the best individual, mapped out over the 500,000 generations.

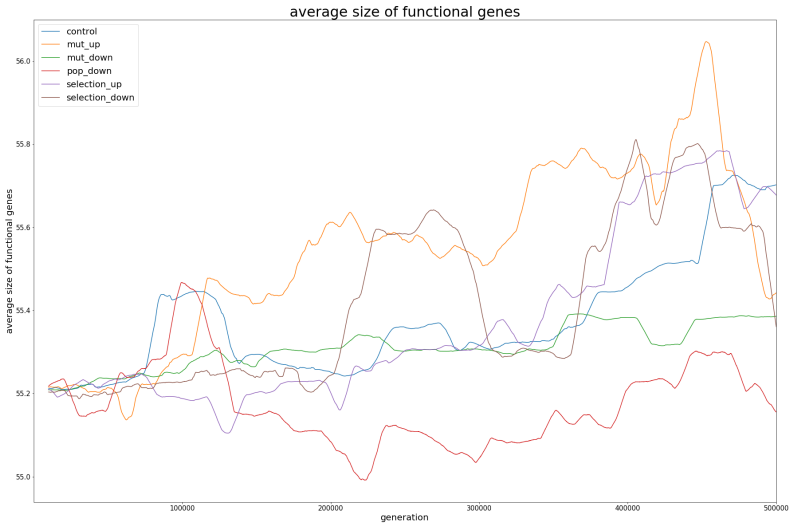

Figure 4.8: Plot showing the average size of functional genes over time for all seeds.

The population down condition continues to be the outlier in terms of genome structure, with the average size of the functional genes remaining noticeably lower than for any other condition. It is worth pointing out, however, that the difference between the smallest average and the largest average is only 1 base pair.

4.1.4 Evolvability

In Figure 4.9 below, we see the results of the experiments on evolvability for the best individual's (at generation 500,000) lineage.

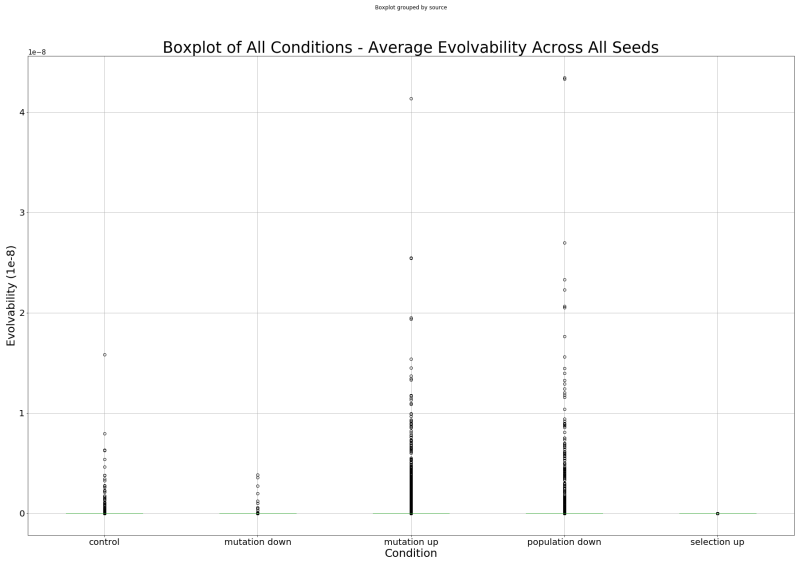

Figure 4.9: A box plot showing the mean evolvability spread of all seeds and all conditions. Higher numbers are more evolvable.

The figure illustrates that, for each condition, overall the best individual still had quite low evolvability. For the selection up condition, however, the deviation from zero was even smaller, as illustrated in Table 4.3 below, which provides the mean and standard deviation of the evolvability of the best individual in each condition.

	mean	**standard deviation**
control	2.035523813975702e-11	3.2787490312081233e-10
μ_{up}	9.97655150597127e-11	8.41749150634588e-10
μ_{down}	6.064697990806935e-12	1.2471674281540128e-10
k_{up}	1.9498939718794653e-15	6.394659146405824e-14
k_{down}		
N_{up}	8.837632116681012e-11	1.0799303682398726e-09
N_{down}		

Table 4.3: Table illustrating the mean and standard deviation of the evolvability for each condition. μ is the mutation rate, k is the selection rate, and N is the population size.

4.1.5 Robustness

Recall from Section 2.3.2 that robustness is measured by the fraction of neutral mutants of an individual. In the following figure, we see a bar plot showing the spread of neutral mutants for the best individual for the control condition as well as the six variations.

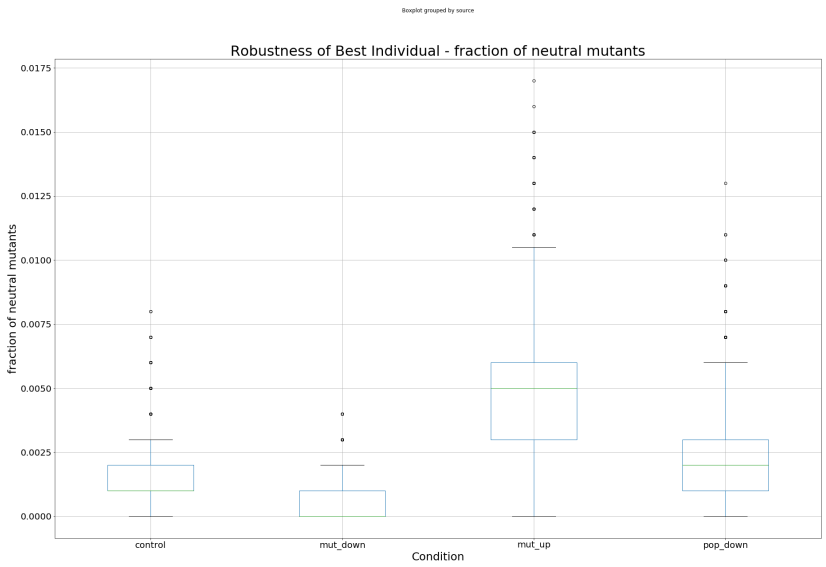

Figure 4.10: Bar graph showing the spread of neutral mutants for the best individual at generation 500,000, all conditions.

The mutation up condition clearly had the largest percentage of neutral mutants, at 0.5%.

4.2 Discussion

Overall, the mean fitness of our digital organisms was nearly unchanged from generation 1 to 500,000. One explanation for this might be that the organisms were allowed to continue to evolve in the same environment in which their wild types were generated.

4.2.1 Relation to Real-World Biological Entities

4.2.2 Limitations of Results

One limitation to consider is that only one parameter varied per condition. It may be possible that it is only under a combination of conditions (e.g. low selection *and* high mutation rates) does reductive evolution occur.

Another limitation is that the environments did not vary in our experiments. This could potentially have a large effect on robustness and evolvability, which are strong influencers of reductive evolution.

Aevol as a modeling software is limited in that it relies, like all models, on several simplifications. The population sizes tested here, even in the population up condition, are still much smaller than would be found in real world populations.

Chapter 5

Conclusion and Future Work

In this section we will summarize our overall conclusions drawn from the work and highlight a few possibilities for further research.

5.1 Conclusion

5.2 Future work

Bibliography

[1] Thomas Bataillon. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? *Heredity*, 84(5):497–501, 2000.

[2] Bérénice Batut, Carole Knibbe, Gabriel Marais, and Vincent Daubin. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature reviews. Microbiology*, 12(12):841–850, 2014.

[3] Bérénice Batut, David P. Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC bioinformatics*, 14 Suppl 15:S11, 2013.

[4] Guillaume Beslon, Vincent F Liard, and Santiago F Elena. Testing evolution predictability using the aevol software. 16th international meeting of the European Society of Evolutionary Biology (ESEB 2017) , August 2017. Poster.

[5] Katie Bradwell, Marine Combe, Pilar Domingo-Calap, and Rafael Sanjuán. Correlation between mutation rate and genome size in riboviruses: mutation rate of bacteriophage qβ. *Genetics*, 195(1):243–251, 2013.

[6] Quentin Carde, Marco Foley, Carole Knibbe, David P. Parsons, Jonathan Rouzaud-Cornabas, and Guillaume Beslon. How to reduce a genome? alife as a tool to teach the scientific method to school pupils. *Artificial Life Conference Proceedings*, (31):497–504, 2019.

[7] Clarence FG Castillo and Maurice HT Ling. Digital organism simulation environment (dose): A library for ecologically-based in silico experimental evolution. *Advances in Computer Science : an International Journal*, 3(1):44–50, 2014.

[8] A.D. Cutter. *A Primer of Molecular Population Genetics*. Oxford University Press, 2019.

[9] John W Drake. A constant rate of spontaneous mutation in dna-based microbes. *Proceedings of the National Academy of Sciences*, 88(16):7160–7164, 1991.

[10] Santiago F Elena, Claus O Wilke, Charles Ofria, and Richard E Lenski. Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution*, 61(3):666–674, 2007.

[11] Isabel Gordo and Brian Charlesworth. On the speed of muller's ratchet. *Genetics*, 156(4):2137–2140, 2000.

[12] Carole Knibbe. *Evolution of genome structure by indirect selection of the mutational variability: a computational approach*. Theses, INSA de Lyon, December 2006.

[13] Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, and Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding dna. *Molecular Biology and Evolution*, 24(10):2344–2353, 08 2007.

[14] Sanna Koskiniemi, Song Sun, Otto G Berg, and Dan I Andersson. Selection-driven gene loss in bacteria. *PLoS genetics*, 8(6), 2012.

[15] Vincent Liard, David Parsons, Jonathan Rouzaud-Cornabas, and Guillaume Beslon. The complexity ratchet: Stronger than selection, weaker than robustness. *Artificial Life Conference Proceedings*, 1(30):250–257, 2018.

[16] Gabriel AB Marais, Alexandra Calteau, and Olivier Tenaillon. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, 134(2):205–210, 2008.

[17] Octavio Miramontes, Franois Taddei, Ariel B. Lindner, Antoine Frenoy, and Dusan Misevic. Shape matters in cooperation. *Artificial Life Conference Proceedings*, 1(28):340–341, 2016.

[18] Vadim Mozhayskiy and Ilias Tagkopoulos. Microbial evolution in vivo and in silico: methods and applications. *Integrative Biology*, 5(2):262–277, 10 2012.

[19] C. Ofria and C. O. Wilke. Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, 10(2):191–229, 2004.

[20] T. S. Ray and J. Hart. Evolution of differentiated multi-threaded digital organisms. In *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, volume 1, pages 1–10 vol.1, 1999.

[21] Jacob Rutten, Paulien Hogeweg, and Guillaume Beslon. Adapting the engine to the fuel: mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC Evolutionary Biology*, 19, 12 2019.

[22] Yolanda Sanchez-Dehesa, Loïc Cerf, Jose Maria Pena, Jean-François Boulicaut, and Guillaume Beslon. Artificial Regulatory Networks Evolution. In *Proc 1st Int Workshop on Machine Learning for Systems Biology MLSB 07*, pages 47–52, Evry, France, September 2007.

[23] Zhiyi Sun and Jeffrey Blanchard. Strong genome-wide selection early in the evolution of prochlorococcus resulted in a reduced genome through the loss of a large number of small effect genes. *PloS one*, 9:e88837, 03 2014.

[24] Ali R Vahdati, Kathleen Sprouffske, and Andreas Wagner. Effect of population size and mutation rate on the evolution of rna sequences on an adaptive landscape determined by rna folding. *International journal of biological sciences*, 13(9):1138, 2017.

[25] Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, 2008.

[26] Tanita Wein and Tal Dagan. The effect of population bottleneck size and selective regime on genetic diversity and evolvability in bacteria. *Genome biology and evolution*, 11(11):3283–3290, 2019.