Evolutionary Stabilization of Microbial Genomes and Mutation Rates in Variable Environments

Chibuikem Nwizu
Department of Applied Mathematics, Brown University

Submission for Applied Mathematics-Biology Science Concentration
Adviser: Dr. Sohini Ramachandran, Ph.D.
First Reader: Dr. Daniel Weinreich, Ph.D.
Second Reader: Dr. Anastasios Matzavinos, Ph.D.
Spring 2017

Abstract

The goal of this project is to understand how fluctuating environments maintain genetic diversity and stabilize mutation rates in bacteria. Can it explain the diversity of genomes and the stability of mutation rates in natural prokaryotic populations? This thesis presents a theoretical model of the evolutionary dynamics of genomic complexity/diversity and mutation rate in variable environments. In the thesis, we explore the dynamics of this model and determine the environmental conditions in the model under which mutation rates are stabilized. The results of these simulations, not only support experimental observations, but also use the ecological framework of specialists/generalists to understand the results. We can use this new model to inform existing theory and empirical observations on microbial genome diversity, microbial genome complexity, and mutation rate evolution.
Contents

1 Introduction 3
   1.1 Microbial Diversity and the Prokaryotic pan-genome 3
   1.2 Mutation rate stabilization 4

2 Model 5
   2.1 Model Formulation 5
   2.2 The Simple Model 8
       2.2.1 Equilibrium dynamics of 1 Loci 8

3 Results 9
   3.1 Non-Equilibrium Behavior 9
   3.2 Equilibrium Behavior 11
       3.2.1 Equilibrium Behavior of Genome Complexity 11
       3.2.2 Equilibrium Behavior of Genome Diversity 13
   3.3 The Fate of a Mutator in variable Environments 15

4 Discussion & Conclusion 18

5 Methods 20

6 Acknowledgements 20

References 20

7 Supplementary Materials 24
1 Introduction

1.1 Microbial Diversity and the Prokaryotic pan-genome

We are motivated by the question: why do populations have such high intra-species genomic diversity? In nature, if one sequenced the DNA of two naturally occurring E. coli bacteria found in the same soil but separated by only a centimeter, despite them being from the same species, one would find that their genomes look very different. Trosvik et. al. 2002, report some prokaryotic species have as little as 70% of shared DNA. Furthermore, in soil in 30-100 cm$^3$, up to 3,000-11,000 genomes may exist [38]. Recent advances in sequencing technology have allowed us to look, with greater resolution into the genomes of prokaryotes. When looking at the total number of genes observed in a prokaryotic species as a function of the number of genome sequenced, one observes the number of genes increases rapidly, initially; however as the number of genome sequenced continues to increase, the number of new genes converges. In ecology, such curves are called a rarefaction curve (Figure 1).

![Figure 1](image-url)

Figure 1: (A) The number of genes in the core genome. (B) The number of genes in the accessory genome. (C) The number of genes in the pan-genome

From these curves we observe, a baseline level of genes found in all members of the species (Figure 1 A). The literature has called this the core genome. We also see in rarefaction curves the number of genomes not present in all members of a species (Figure 1 B). The literature calls this the accessory genome [21]. Together, both the core and accessory genome make up what is known as the pan-genome (Figure 1 C).

The idea of a pan-genome is unique to prokaryotes but poorly understood [27, 6]. Furthermore, to understand the factors that impact the pan-genome’s size and complexity, one must understand the magnitude, dynamics and controlling factors of the accessory genome. The pan-genome expands the definition of what it means to be a member of a particular prokaryotic species: a prokaryote
is a member of a species because it has any combination of genes that are found in the pan-genomic pool of genes of a particular species [19].

In bacteria, a number of reasons explain why this happens and they are all related to this idea of horizontal gene transfer. Normally genes are transferred vertically—from mother to daughter. In bacteria, however, DNA can be swapped via conjugation, transduction (the incorporation of DNA via viruses), chromosomal rearrangement, and even collection of DNA from the environment [8, 35, 7, 39, 3]. All these mechanisms allow for genes to be constantly passed around, creating that diversity pool in the pan-genome. This is a popular view of what maintains this diversity [28].

This presents a problem when considering this variation in light of our understanding of evolution. Currently, we understand evolution works on variation within a population. Selection of beneficial and variation is effectively reduced due to the selection of beneficial variants. In this current understanding of how evolution works, variation is hard to maintain.

In this paper we examine the stabilizing effect of varying environments on pan-genomic diversity and genome complexity. Here we define complexity as the number of functional genes in the genome and diversity as the number of unique genomes in the population. We predict that patterns of genome diversity observed in nature can be reproduced modelling the evolution of the accessory genome in variable environments. We know that the tendency is to loose genes that are not required at a particular time either through selection or mutation. We investigate whether varying what is required at a particular time is enough to maintain genome complexity and diversity.

### 1.2 Mutation rate stabilization

In population genetics, classical theory on mutation rate evolution is already established. Due to the highly deleterious nature of mutations, mutation rates are expected to be low in most populations. Therefore, one should not expect to see a high mutation rate in a population in constant environments. Though having a high mutation rate would confer to an increase in deleterious mutations, it also confers to an increase of beneficial mutations. These beneficial mutations increase the fitness of a population, lead to beneficial adaptations, and drive evolution [31].

Giraud et al. 2001, however, observed the opposite phenomenon. When an E. coli mutator strain was allowed to adapt in an in vivo constant environment (in a mouse gut), the high mutation rate was initially beneficial, for it allowed it to adapt faster in that environment compared to the wild type strain. Interestingly, when that adapted mutator strain was moved to a second environment, the neutral/beneficial mutations it had accumulated in one environment had become deleterious in the other environment. Subsequent, studies corroborated these results, showing mutators perform better in lab and clinical environments (where the environment is constant) than in fluctuating environments [33, 40]. Thus, there is a profound gap between the theory of mutation rate evolution and what is observed. This suggests that the current theory is in need of revision.
Recent studies have given possible molecular reasons for the favoring of mutators in constant environments. These studies have shown that, as expected, high mutation rates confer many loss of functions mutations that break genes. However, assuming that the gene is not essential to survival, this loss of function may lead to an increase in fitness, for the organism is no longer wasting energy using/maintaining an unnecessary function [21, 17]. Loss of function mutations may (albeit very rarely) also lead to the creation of a de novo function that confers a fitness advantage in a particular environment [13]. If the environment changes, however, and the function lost is necessary, any benefit for losing that gene is lost, and the mutation becomes deleterious. Mutators, therefore, overspecialize in constant environments and are selected for, but when the environment changes, they are selected against [13, 17].

2 Model

2.1 Model Formulation

In our model, we represent the a microbial accessory genome as a $1 \times n$ vector:

$$g = [x_1, x_2, \ldots, x_n], x_j \in \{1, 0\}$$

where $n$ is a fixed number and represents the number of loci in the accessory genome. In our model, each gene occupies one loci position. The value of $x_j$ indicates the functionality of the gene at location $j$. A 1 indicates that the gene at position $j$ is functional, while 0 indicates that the gene at position $j$ is not functional.

Next, in our model, we represent an entire population as an $N \times n$ matrix

$$G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{bmatrix} = \begin{bmatrix} x_{11}, x_{12}, x_{13}, \ldots, x_{1n} \\ x_{21}, x_{22}, x_{23}, \ldots, x_{2n} \\ \vdots \\ x_{N1}, x_{N2}, x_{N3}, \ldots, x_{Nn} \end{bmatrix}$$

where element $x_{ij}$ of $G$ is the $j$th gene of individual $i$ in the population.

In our model we represent the environment as a $1 \times n$ vector similar to the microbial genome:

$$E = [e_1, e_2, \ldots, e_n], e_j \in \{1, 0\}$$

Where $n$ is a fixed number and represents the total number of possible genes in the accessory genome that the environment can act upon. The value of $e_j$ indicates the whether or not the environment requires the gene at location $j$. A 1 indicates that the environment requires the gene at position $j$, while a 0 indicates that the environment does not require the gene at position $j$. Unlike the genome in our model, $E$ is strictly a row vector, meaning that all individuals in the populations are subjected to the same environmental pressures.
To assess the how well adapted each individual is to an environment, we must assign each individual a fitness score. We assume every functional gene has some energy expenditure associated with its maintenance. If the environment requires a function and an individual has this function in the genome, the individual enjoys the benefit of utilizing the function. Likewise, if the environment does not require a function and an individual lacks the function, the individual still the benefits, for they do not have to expend energy to maintain the function. Both of these cases illustrate how an organism’s fitness changes when the both the genes required by the environment and the individual’s functional genome are exactly the same. This is not what is observed in nature and in experiments, for we know there is variability with populations. Moreover, it is this intra-population variability drives selection.

In our model, differences between the environment and the individual drives selection (Figure 2 A). If the environment requires a function and an individual does not have the function in the genome, relative fitness decreases by a substantial amount. If the environment does not require a function and an individual does have the function in the genome, relative fitness decreases by a smaller amount (Figure 2 B). We define \( s_{\text{gains}} \) as the relative benefit of maintaining a functional gene when the environment does require it. Likewise, we define \( s_{\text{baggage}} \) as the relative benefit of not maintaining a functional gene when the environment does not require it. Consequently, we define the fitness score of individual \( i \) as

\[
F_i = (1 + s_{\text{baggage}})^{b_i} + (1 + s_{\text{gains}})^{g_i}
\]  

(1)
where

\[ b_i = \sum_j 1_b(x_{ij}, e_j) \]
\[ g_i = \sum_j 1_g(x_{ij}, e_j) \]

and \(1_b\) and \(1_g\) are indicator functions that are 1 when \(x_{ij} = e_j = 0\) (in the case of \(1_b\)) or \(x_{ij} = e_j = 1\) (in the case of \(1_g\)), but are zero otherwise.

We employ a Wright-Fischer Model for selection to model selection in our population. In Wright-Fischer, individuals are stochastically chosen to reproduce in proportion to their fitness in the current environment—the value represented in equation 1. Using Wright-Fisher, the allelic composition of the population changes, but the total population size remains fixed at N. We will not discuss how Wright-Fisher selection works in this paper, only mention its use. Again, by employing Wright-Fisher, we incorporate the observed stochasticity of selection into our model: Though an organism may have a high fitness score, it does not guarantee their survival in an environment, only increases their probability of surviving.

We also include mutations in our model. Mutations in nature occur as an error in the copying of genetic information from parent to offspring—manifesting itself from small events like single nucleotide substitution, to large large events like a chromosome insertion/deletion. As previously noted, mutations are a major driver for genetic variability within a population. Likewise, within our model, mutations drive intra-population variability.

We define \(\mu\) as the per loci per individual knockout mutation rate (that is \(1 \rightarrow 0\)). In the accessory genome, we model the number of knockout mutations as a Binomial random variable

\[ X_{KO} \sim \text{Bino}(n_{func}, \mu) \]

where \(n_{func}\) is the total number of functional sites available in accessory genomes across the population.

Likewise, we define \(\mu_r\) as the per loci per individual reverse mutation rate (that is \(0 \rightarrow 1\)) occur per loci per individual. We assume that the reverse mutations rate are proportional to knockout mutation rate

\[ \mu_r = r \cdot \mu, \quad r \in [0, 1) \]

We model the number of reverse mutations also as a Binomial random variable

\[ X_{Rev} \sim \text{Bino}(n_{non-func}, \mu_r) \]

where \(n_{non-func}\) is the total number of non-functional sites available in accessory genomes across the population.

Mutations can occur in our core genome (though it is not explicitly modeled). We also assume that a mutation to the core genome is automatically lethal. We model the number of lethal mutations as a Poisson random variable

\[ X_{Leth} \sim \text{Poi}(\lambda) \]
where $\lambda = n_{\text{essential}} \ast \mu$, and represents the average number of times mutation occurs in the core genome ($n_{\text{essential}}$).

Our next parameter is the fraction needed. We define fraction needed ($f$) as the probability that any particular gene at loci $j$ will be required by any particular environment. Such a parameter is important because it is unknown exactly how much of the accessory genome is needed by any particular environment. However, our model is sensitive to this: if the environment, $E$, requires a high number of ones, selection will favor individuals who maintain a higher number of ones. Additionally, If the environment changes, we would want the average number of genes required to remain constant across environmental changes. This allows us to observe how requiring different functions effects the long term genome diversity and complexity of the accessory genome.

Related to $f$ is the rate at which the environment changes. We define $\tau$ as the wait time between environmental changes (the change rate is technically $1/\tau$ but we will refer to $\tau$ as the change rate). When $\tau$ is small, the environment changes very quickly, and when $\tau$ is large, the environment changes very slowly.

### 2.2 The Simple Model

To build intuition, we start by examining the behavior of the model with only $n = 1$ locus.

#### 2.2.1 Equilibrium dynamics of 1 Loci

We looked at the fraction of ones at equilibrium for very small $\tau$. We define the fraction of 1’s in the population as

$$p_j = \frac{\sum_i x_{ij}}{N} \quad (6)$$

With this definition, we were able to write a deterministic description for the rate of the fraction of ones at locus $j=1$.

$$\frac{dp_j}{dt} = -\mu p_j + \mu_r (1 - p_j)$$

$$- f \ s_{\text{gains}} p_j - (1 - f) s_{\text{baggage}} p_j$$

$$+ (1 - f) s_{\text{baggage}} (1 - p_j) + f \ s_{\text{gains}} (1 - p_j) \quad (7)$$

This deterministic treatment of the model assumes that the environment changes rapidly. For very small $\tau$, the environment changes so quickly that any one individual experiences many environments. Consequently, the behaviour of the 1-loci model for small $\tau$ at equilibrium can be modelled with the fraction of ones at equilibrium

$$p_\infty = \frac{f \ s_{\text{gains}} - (1 - f) s_{\text{baggage}} + \mu_r}{\mu + \mu_r + f \ s_{\text{gains}}} \quad (8)$$
3 Results

3.1 Non-Equilibrium Behavior

First, we would like to understand the pre-equilibrium behaviour of the model. Understanding this behavior will provide intuition about the forces that impact our model and insight into the qualitative tendencies of the a population’s accessory genome. In particular, we would like to understand the pre-equilibrium dynamics of the functional genes in a population’s accessory genome. Because the most mutations are loss of function mutations, we expect the number of functional genes to decay with time. As the functional genes decays, the population becomes more specialized, able to tolerate less unique environments that require unique mutations. Consequently, we will call this metric the rate of specialization, $S^*$. We define the average rate of specialization, $\bar{S}^*$, over all replicates as

$$\bar{S}^* = \frac{1}{r} \sum_{i=1}^{r} \max \left\{ \frac{dp^*_i}{dt} \right\}$$

where $p^*$ is the fraction of functional genes across the entire population for a single replicate, and $r$ is the number of replicates.

There will be certain parameter regimes that increase $\bar{S}^*$. Regimes that promote faster specialization will favor accessory genomes that are closer to the functional genes required by the environment. These genomes have a selective advantage and will sweep the population, lowering the fraction of functional genes in the population. We define these genomes as specialist genomes (specialists). Specialists have the functional genes required by the current environment therefore they experience a selective advantage in these regimes. On the other hand, we define generalist genomes (generalists) as genomes that specialize slower, but are more aptly suited for surviving many different types of environments.
3.1 Non-Equilibrium Behavior

Figure 3: Large values of $\tau$ mark a regime where specialists are favored. The simulation parameters were $s_{\text{gains}} = 0.1$, $s_{\text{baggage}} = 0.01$, $f = 0.2$, $\mu_1 = 10^{-5}$.

When we increased $\tau$, $\bar{S}^*$ increased. Figure 3 shows this trend. For small $\tau$s, generalists are favored because their high functional gene content allows them to be better suited for a wider variety of environments. This drives the population’s fraction of functional genes up. Large $\tau$s favor specialists because, during the simulation time, their genomes converge faster to the functional genes required by the environment. This provides them with a selective advantage and allows them to sweep the population.

When we increased $s_{\text{gains}}$, $\bar{S}^*$ decreased (see Figure 11 in Supplemental Materials). Thinking about the model, we are able to gain intuition about this result. As the population evolves, individuals with too many loss of function mutations are selected against. Consequently, increasing $s_{\text{gains}}$ selects for individuals who maintain lots of functional genes in their genome. As the generalists survive and increase in frequency, their genomes prevent the decay of functional genes in the population, slowing the rate of specialization.

When we increased $s_{\text{baggage}}$, $\bar{S}^*$ increased (see Figure 12 in Supplemental Materials). In our model, $s_{\text{baggage}}$ is given when the individual loses a functional gene, and the environment does not require it. Specialists are more likely to have lots of zeros that match the environment; therefore, $s_{\text{baggage}}$ favors the specialists. Since specialists are favored and they tend to carry fewer functional genes, the fraction of functional genes in the population is driven down, increasing $\bar{S}^*$.

When we increased $f$, $\bar{S}^*$ increased (see Figure 13 in Supplemental Materials). $f$ is the fraction of time loci $i$ is needed. A larger $f$ corresponds to more genes being required in every environment. Consequently, the generalist is favored over the specialist for their genomes contain more functional genes. This
maintains a high fraction of functional genes in the population, slowing the rate of specialization.

Finally, when we increased $\mu_1$, $\bar{S}^*$ decreased (see Figure 14 in Supplementary Materials). As $\mu_1$ increases, more loss of function mutations occur. This promotes the creation of specialists. With more specialists, the fraction of functional genes in the population is driven down, increasing $\bar{S}^*$.

Table 1 below summarizes the results of the simulations.

Table 1: Summary of the effects of the model parameters on $\bar{S}^*$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Change</th>
<th>Effect on $\bar{S}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>$s_{gains}$</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>$s_{baggage}$</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>$f$</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

3.2 Equilibrium Behavior

3.2.1 Equilibrium Behavior of Genome Complexity

Figure 4: Genome Complexity at equilibrium is stabilized by fast-changing environments.

We looked at the number of functional genes at equilibrium to understand the effect of variable environments on pan-genomic complexity. Figure 4 shows the fraction of functional genes at equilibrium, $F_{eq}$. For large $\tau$ values, $F_{eq}$
Chibuikem Nwizu

3.2 Equilibrium Behavior

approaches a value close to the model parameter $f$. across parameter values, we observed stability in the convergence behaviour at large $\tau$s (see Figure 18 in Supplemental Materials). This observation suggests that in our model, for large $\tau$s (in fact, as $\tau \to \infty$), there is a fixed point in how complex the genome will be.

This behavior makes sense given what we know about the model. For large $\tau$s, the rate of specialization increases, indicating that in this regime specialist genomes have a selective advantage. Because this is a regime that favors the specialist, and specialists tend to maintain the functional genes required by environment, their complexity is limited to how complex their environment is. In our model, this value is $f$.

From Figure 4, we also observe that small $\tau$s maintain relatively high $F_{eq}$. We can also explain this behavior, for we know that in this regime the generalist is favored. Because the environment changes so frequently, complex genomes have a selective advantage. They contain functional genes that may be useful when the environment changes. This supports observations of complexity in naturally-occurring microbes.

![Stabilization of Genome Complexity at Large $\tau$](image)

Figure 5: Changing $\tau$ changes the selective pressures on generalists and specialists. The changing ratio of specialists to generalists gives rise to the curvature in the plot. In regime C, the specialist is heavily favored over the generalist. In regime D, the specialist lose this selective advantage.

But the curve in Figure 4 is non-monotonic. Its concavity may be explained by the specialization rate and $\tau$. In regime C (Figure 5), generalists begin to lose their selective advantage because the cost of maintaining their additional complexity is no longer greater than the gains received by seeing many environments. In the fast-changing environment, the additional complexity in the generalist’s genome is still “genomic baggage”, but the cost of maintaining this baggage is offset by the benefit of eventually encountering an environment that requires a functional gene. Moreover, the environment changes slow enough that the specialist gains a selective advantage because they have the time to
specialize.

In regime D (Figure 5), this advantage diminishes, resulting in the upward inflection of the curve. The specialists continue to have a selective advantage during moderate $\tau$s; however, these extended periods in a particular environment causes them to overspecialize on that particular environment. When the environment changes, it is likely they have lost the genes needed for the new environment. Therefore, moderate specialists with enough complexity (in fact, with the complexity closest to $f$) are selected for, raising the value of $F_{eq}$ to $f$.

### 3.2.2 Equilibrium Behavior of Genome Diversity

![Genomic Diversity](image)

Figure 6: Pan-genome diversity at equilibrium is maximized in moderately-changing environments.

Next, we looked the Shannon Index of the population at equilibrium to understand the effect of varying environments on the pan-genomic diversity. To quantify diversity, we used the Shannon Index:

$$H' = \sum_i h_i \ln h_i$$

where $h_i$ is the fraction of individuals with genome $i$. Figure 6 shows the Average Shannon Index, $\bar{H}'$, at equilibrium. The behavior of the plot at extremely large and extremely small values of $\tau$ show that diversity is very low at both extremes. This suggests that only one type group exists at both extremes: specialists, when $\tau$ is large, and generalists, when $\tau$ is small. Moreover, for large $\tau$s, $H'$ appears to converge to the same value for the same $f$, across independent simulation runs.
Based on our understanding of the model, the diversity behavior of pan-genomes at equilibrium appears to converge due to the dominance of specialists in this regime. We can explain why converges through the $S^*$ and $F_{eq}$. We know it is converging because constant environments favor genomic convergence to $f$ (Figure 3). Furthermore, we know they are converging to the same complexity value (Figure 4). Diversity shrinks because selection favors the specialist and forces genomes to converge to the environments complexity.

From Figure 6, we see that small $\tau$s also promote convergence. In this regime, the generalist dominate. Since, diversity is low we know that there are only a few unique genomes; however, only a few unique generalist genomes are needed to maintain the complexity observed in Figure 4. These genomes have the raw genomic material to cause a diversity explosion. Selection, however, prevents this. For large enough subpopulations, selection can not make a generalist lineage more diverse. The diversification only occurs through mutation. However, for low mutation rates, after generalist lineages split into a parent and child branch, the child branch is small in size and subject to drift. Therefore, selection keeps diversity low in rapidly changing environments.

In regime C (Figure 7), the diversity increases. This increase may be due to two factors that work in concert: (1) there are generalists to provide the genetic raw material for diversity and (2) selection begins to favor specialists more, allowing smaller lineages to grow and establish.

Likewise, the decrease in diversity in regime D (Figure 7), could be due to the lack of generalist: Specialization increases during this time and and degrades the genomic diversity. Another factor that could contribute to the decline is that in this regime, selection for specialists decreases. The true explanation is likely a combination of the both of the aforementioned hypotheses.
3.3 The Fate of a Mutator in variable Environments

We have defined the specialist by the content of their genome. In this section, we will expand the definition to include all types of genomes with an elevated mutation rate. These types of specialists are called mutators.

It is important to expand the definition of specialists to include mutators because much of the constant environment behavior noted in this report has been exhibited in mutators in controlled evolution experiments under laboratory conditions [33, 40]. Namely, many studies have observed that laboratory populations of prokaryotes evolve higher mutation rates. Unlike their laboratory counterparts, natural populations tend to evolve lower mutation rates [32]. We seek to understand how varying environments stabilize mutation rate using our model. Moreover, we would like to understand the fate of mutators in variable environments.

We are able to expand the definition of a specialist because most mutations break genes. We have seen that non-functional genes create a selection difference. We have observed that it is this difference in selection that changes the trajectory of the allelic composition of the population. With the present model we seek to understand the fate of mutators in variable environments, and the effect that the model parameters have on the mutator’s fate.

Figure 8: Mutators are have a selective advantage in slow-changing environments. The simulation parameters were $s_{\text{gains}} = 0.1$, $s_{\text{baggage}} = 0.01$, $f = 0.2$, and $\mu_2 = 10^{-5}$.

As $\tau$ increased, the mutator’s fixation probability increased. Figure 8 shows this trend. The dashed line is at $\frac{1}{N}$ and indicates the point at which the $\tau$ goes from deleterious to beneficial. The results of this simulation indicate a
very defined point of neutrality exists for that particular set of parameters. The point of neutrality appears to be between $\tau = 50$ and $\tau = 75$. As $\tau$ increases, the environment changes at a slower rate, meaning the mutator is allowed more time to specialize on a particular environment. This explanation is corroborated by experimental observation [33, 40].

The impact of the other parameters is noted in Table 2 below. When we increased $s_{\text{gains}}$, the fixation probability decreased (see Figure 15 in Supplemental Materials). Thinking about the model, we gain intuition about this result, mutators tend to lose functional mutations faster than the wildtype. In a changing environment, they may begin to lose functional genes not needed in one environment quickly, but when the environment changes and they encounter a new environment that may require this function, they are at a disadvantage exactly like a specialist would in the same scenario. As the cost of this disadvantage grows, the mutator selected against.

As we increased $s_{\text{baggage}}$, the fixation probability decreased (see Figure 16 in Supplemental Materials). The results of this simulation are more pronounced, and the effects of $s_{\text{baggage}}$ appears to have a large impact on $P_{fix}$. To gain intuition about this result, we note that when losing functions is very beneficial, the mutator will be favored for they tend to lose functional mutations faster than the wildtype. The results suggest that this benefit is strong enough to dominate even as the environment changes.

When $f$ increases, fixation probability increased (see Figure 17 in Supplemental Materials in Supplementals). Because $f$ is the fraction of time, loci $i$ is needed, larger $f$ mean that more genes are required in every environment. In these cases, the mutator disadvantaged because their tendency to lose functional genes means they are not able to maintain the fraction of ones need in any particular environment.
Figure 9: In fluctuating environments (τ = 100), mutation rate appears to stabilize at moderate values. The simulation parameters were \( s_{\text{gains}} = 0.1 \), \( s_{\text{baggage}} = 0.01 \), and \( f = 0.2 \).

Finally, as \( \mu_2 \) increases, \( P_{fix} \) behaves non-monotonic. Figure 9 shows this trend. The result suggests there is a non-monotonic relationship between \( \mu_2 \) and \( P_{fix} \). For extreme values of \( \mu_2 \), \( P_{fix} \) appears to be small. This suggests for this particular parameter set, there is an optimum mutation rate. Unlike any of the other parameters, \( \mu_2 \) appears to be deleterious for extreme values of \( \mu_2 \); however for median values of \( \mu_2 \) appear to be beneficial. The threshold is between \( 1 \times 10^{-6} < \mu_2 < 5 \times 10^{-6} \) and \( 5 \times 10^{-5} < \mu_2 < 5 \times 10^{-4} \). The presence of two thresholds suggests that \( \mu_2 \) is stabilized in variable environments. This could be because when \( \mu_2 \) is smaller than the wildtype’s mutation rate, \( \mu_1 \), it is unable to break genes as well as the wildtype and must pay the cost of carrying around genetic baggage. But when \( \mu_2 \) is much larger than the \( \mu_1 \), it will specialize too quickly on the environment, and pay the cost of not having the right function when the environment changes. Additionally, when \( \mu_2 \) is much larger than the \( \mu_1 \), it also overspecializes on the current environment, and break essential genes. Table 2 below summarizes the results of the simulations.

<table>
<thead>
<tr>
<th>Change</th>
<th>Effect on ( P_{fix} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>↑</td>
</tr>
<tr>
<td>( s_{\text{gains}} )</td>
<td>↑ ↓</td>
</tr>
<tr>
<td>( s_{\text{baggage}} )</td>
<td>↑</td>
</tr>
<tr>
<td>( f_{/n} )</td>
<td>↑ ↓</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>↑ ↓</td>
</tr>
</tbody>
</table>
4 Discussion & Conclusion

An important outcome of our study is that it provides a framework for understanding the diversity and complexity of microbial accessory genomes. This framework poses that two types of genomes exist, specialists and generalists. Furthermore, each has a particular regime in which they have a selective advantage over the other genomic type. This framework borrows heavily from classical ecology. In classical ecology, specialists are individuals in a community that occupy a narrow niche and only survive in a limited range of environmental conditions, while generalists are individuals in a community that occupy a broader niche and survive in a wider range of environmental conditions [10, 4]. The intuition behind the fate both types of organisms is similar to the intuitions that motivated our work: specialists tend to outperform generalists in the environments they are specialized on; however, generalists outperform specialists if the environment changes. Many have hypothesized that environmental heterogeneity is a plausible explanation of the expansive biodiversity in nature [20, 41, 26, 37, 9, 29, 16, 14], and here we provide support showing that it can explain some of the diversity we see.

Some studies report that fluctuating environments do not increase diversity [15]. Our model addresses why this may be the case. As seen in Figure 6, there are regimes were fluctuating environments decrease diversity. We suggest that observations of the decreased diversity is because experiments are within this regime. In ecology literature, this phenomenon is known as the intermediate hypothesis [5]. Furthermore, these contradictions, in light of our model, suggest that we start considering concepts such as the fraction of functional genes required by the environment and the change rate of the environment when discussing results of evolution experiments. More must be done to understand how these parameters affect the model as well as their real-world implications. We should also begin to figure out how to quantify and measure these parameters.

The present study invites more analytical modelling, in order to strengthen its results. A diffusion model of the allelic frequency could provide valuable insight into how selection and drift work promote or inhibit species diversification. We also invite the modifications of competition models of specialists and generalists to include the effects of environmental fluctuations [25, 22].
With our model, we see the qualitative behavior changes to the accessory genome. The results of the present study are illustrated in Figure 10. For extreme values of $\tau$, genomic complexity is favored by selection. In contrast, for intermediate values of $\tau$, genomic diversity is favored by selection. Natural extensions of this model include adding subpopulations and migration, to model more accurate microecologies [18]. We set out to understand the microbial accessory genome and the fate of mutators in variable environments. We were able to explain much of the experimental observations with our model and understand its mechanics using a borrowed framework from classical ecology. Further investigation of the model is needed to evaluate the extent of the overlap and strengthen the results presented in the paper.
5 Methods

We ran simulations using a combination of R2014a, Matlab 2015a, 2015b, and 2016a. Jobs were run on the Brown University CCV or on local machines. The genome of the organism was represented by a one-dimensional array of 1’s and 0’s where each index corresponds to a locus on the genome. A 1 corresponds to function at that particular locus. A 0 corresponds to no function at that particular locus. Each locus is independent of one another. The population was fed into a mutation function that, based on the mutation rate, performed loss of function mutations, gain of function mutations, and lethal mutations. A loss of function mutation refers to the conversion of a 1 to a 0 on the nonessential loci of the genome. A gain of function mutation refers to the conversion of a 0 to a 1 nonessential loci of the genome. A lethal mutation refers to the conversion of a 1 to a 0 on essential loci of the genome. A matrix of all of the individual bit string constituted a population. The environment was also represented by a one-dimensional array of 1’s and 0’s where each index corresponds to a locus on the genome. Then environment frequently based on the parameter \( \tau \).

We calculated the fitness of each organism in a population at a particular time step. To calculate the fitness, each row in the population matrix was compared to the environmental array. If the value at locus \( i \) agreed with the environment at locus, and the value was 1, the organism’s fitness score is added to by the amount \( 1 + s_{\text{gain}} \). If the value at locus \( i \) agreed with the environment at locus, and the value was 0, the organism’s fitness score is added to by the amount \( 1 + s_{\text{baggage}} \). The fitness score were feed into a Wright-Fisher function, that selected the individuals reproduce based on the fitness values. Termination of the simulation was dependent on the type of simulation: fixation simulation were terminated when either the mutant fixed or went extinct; all other simulations terminated after a defined number of generations.

6 Acknowledgements

I would like to thank my mentor and friend Christopher Graves for the work he has done in advising, guiding, and mentoring me through this research journey. Without Chris, I would not have been able to complete this project. I would also like to acknowledge the Weinreich lab (both its Principle Investigator, Dr. Daniel Weinreich, and its current/former memebers) for giving me a research home these past two years. I am very appriciative of the opportunities and experiences given to me, for the have allowed me to mature and grow intellectually.

References


7 Supplementary Materials

Figure 11: $\bar{S}^*$ as a function of $s_{gains}$. The simulation parameters were $s_{baggage} = 0.01$, $\tau = 50$, $f = 0.2$, and $\mu = 10^{-5}$.

Figure 12: $\bar{S}^*$ as a function of $s_{baggage}$. The simulation parameters were $s_{gains} = 0.1$, $\tau = 50$, $f = 0.2$, and $\mu_1 = 10^{-5}$.
Figure 13: $\bar{S}^*$ as a function of $f$. The simulation parameters were $s_{\text{gains}} = 0.1, s_{\text{baggage}} = 0.01, \tau = 50,$ and $\mu_1 = 10^{-5}$. 

Figure 14: $\bar{S}^*$ as a function of $\mu$. The simulation parameters were $s_{\text{gains}} = 0.1, s_{\text{baggage}} = 0.01, \tau = 50, \text{ and } f = 0.2.$
Figure 15: $P_{fix}$ decreases as a function of $s_{gains}$. The simulation parameters were $s_{baggage} = 0.01$, $\tau = 50$, $f = 0.2$, and $\mu_2 = 10^{-5}$.

Figure 16: $P_{fix}$ increases as a function of $s_{baggage}$. The simulation parameters were $s_{gains} = 0.1$, $\tau = 50$, $f = 0.2$, and $\mu_2 = 10^{-5}$.
Figure 17: $P_{fix}$ decreases as a function of $f$. The simulation parameters were $s_{gains} = 0.1$, $s_{baggage} = 0.01$, $\tau = 50$, and $\mu_2 = 10^{-5}$.

Figure 18: Genome Complexity for large $\tau$ converge to the $f = 0.20$