STRUCTURE AND EVOLUTIONARY DYNAMICS IN FITNESS LANDSCAPES

By

Anuraag R. Pakanati

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science—Doctor of Philosophy
Ecology, Evolutionary Biology, and Behavior—Dual Major

2015

ABSTRACT

STRUCTURE AND EVOLUTIONARY DYNAMICS IN FITNESS LANDSCAPES

By

Anuraag R. Pakanati

Evolution can be conceptualized as an optimization algorithm that allows populations to search through genotypes for those that produce high fitness solutions. This search process is commonly depicted as exploring a fitness landscape, which combines similarity relationships among genotypes with the concept of a genotype-fitness map. As populations adapt to their fitness landscape, they accumulate information about the fitness landscape in which they live. A greater understanding of evolution on fitness landscapes will help elucidate fundamental evolutionary processes.

I examine methods of estimating information acquisition in evolving populations and find that these techniques have largely ignored the effects of common descent. Since information is estimated by measuring conserved genomic regions across a population, common descent can create a severe bias by increasing similarities among unselected regions. I introduce a correction method to compensate for the effects of common descent on genomic information and empirically demonstrate its efficacy.

Next, I explore three instantiations of NK, Avida, and RNA fitness landscapes to better understand structural properties such as the distribution of peaks and the size of basins of attraction. I find that the fitness of peaks is correlated with the fitness of peaks within their neighborhood, and that the size of peaks' basins of attraction tends to be proportional to the heights of the peaks. Finally, I visualize local dynamics and perform a detailed comparison between the space of what evolutionary trajectories are technically possible from a single starting point and the results of actual evolving populations.

ACKNOWLEDGMENTS

I would like to acknowledge many people for their assistance in my journey here. My advisor, Dr. Charles Ofria, was the reason I came to Michigan State. He has provided persceptive guidance, detailed feedback, and an absolutely incredible environment of freedom in which I was able to pursue my research directions in my own peculiar and stubborn way. I also am grateful for my co-advisor, Dr. Chris Adami, who arrived at Michigan State at a fortuitious time for me, and who proved invaluable in helping develop and guide my research. Furthermore, I would like to thank my committee members, Dr. Erik Goodman and Dr. Robert Pennock, who read through drafts and provided insightful feedback that helped shape my work for the better, most notably exhorting me to go beyond studying Avida fitness landscapes.

I am grateful for the unique environment for interdisciplinary research that has existed at MSU embodied by both the EEBB program and the BEACON Center. I could not have performed this research without resources provided by the HPCC at MSU and a cluster funded by the NSF.

I thank current and past Devolab members, and those in the Ofria, Adami, Lenski and SENS labs, many of whom provided valuable feedback along the way. I worked with many people over the past few years, but work with Christopher Strelioff and Bjorn Østman particularly influenced many ideas explored in this work. I am also indebted to those who provided companionship and who proved generous with their hard-earned experience: Chad Byers, Andres Ramirez, Luis Zaman, Jory Schossau, and Brian Connelly, among others.

Finally, I would like to acknowledge my parents, Shraven Pakanati and Sreekala Pakanati, my sister, Sheena Pakanati, and Liz Duval for their unyielding support and encouragement without which I could never have completed my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The evolution of populations of organisms is a topic of great interest; long before Darwin described the finches of the Galapagos, scientists had observed how well organisms are adapted to their specific environments. Organisms incorporate information about the environment into their genetic blueprints to create adaptive physical and behavioral characteristics. A greater understanding of how populations of organisms explore their genetic landscapes and accumulate information can help us better understand the process of evolution in natural populations.

A deeper understanding of fitness landscapes and the processes that lead to effective solutions will also benefit evolutionary computation. Fitness sharing, novelty search, and other similar techniques are all built around finding better solutions by avoiding the classical pitfall of climbing a fitness gradient until permanently stuck in a local optimum. Understanding how solutions cluster and populations traverse paths between optima can help lead to improved search techniques.

1.1 Chapter Overview

Biological complexity has been described in [Adami, 2002] as a measure of how much information an organism stores about its environment in its genome. It is theorized that natural selection increases complexity in a static single niche environment. Sequence complexity is frequently measured by subtracting the summation of per-site entropies from the maximum possible entropy. Entropy is a measure of how 'random' a value is when drawn from a distribution as rigorously defined in [Shannon, 1948] and applied to genomes in [Schneider et al., 1986] and [Adami et al., 2000]. Sequence complexity tends to reflect other measures of complexity, such as structural and functional complexity, but has the advantage that it is straightforward to evaluate using accessible population measures.

Previous work in measuring complexity commonly overlooks the biases imposed by common descent. Coalescent theory predicts that an asexual population will often have a relatively recent common ancestor. As a consequence of this fact, many estimates of biological complexity often attribute a lack of variation due to common descent as being due to selective pressure. In Chapter 2, I explore some of the limits of existing biological information estimates, examine commonly used correction methods, show how they fail to correct for the problem, propose a correction that accounts for the effect of common descent, and demonstrate the utility of this correction experimentally.

Fitness landscapes have been widely used to conceptualize the relationship between the genotypes and their associated phenotypes. The structure of fitness landscapes in biological organisms is still largely unknown and varies immensely from environment to environment. Even so, these fitness landscapes carry important consequences for evolution. For example, in the context of protein folding, peaks tend to exhibit strong local clustering, suggesting that a stable configuration will often have other stable configurations in the neighborhood. This clustering would allow evolution to have an easier time to find new stable configurations than if these variants were randomly distributed throughout the genotype space. Thus, a

concrete understanding of peak distribution in the fitness landscape would be relevant to genetics and medicine since one might better predict when it is productive to restrict the search for fit sequences to the immediate area around already known configurations. Chapter 3 exhaustively examines peak distributions in three very different types of landscapes—NK, Avida, and RNA, each with length 18 genomes and four possible alleles at each site, which is a scale beyond any that have been systematically explored to date.

Peaks can attract evolutionary trajectories from nearby regions of the fitness landscape. When an evolutionary trajectory has reached a peak, it can be challenging to leave that peak again since it requires unlikely valley crossings or simultaneous mutations to escape. This gravitational quality of peaks is often characterized as each peak exerting a 'basin of attraction' on the nearby genotypes. [Kauffman, 1993] and [Ochoa et al., 2008] have examined basin size experimentally and find that basin size increases exponentially with the fitness of a peak in NK landscapes. The strength of this relationship is naturally dependent upon the amount of epistasis in the landscape; epistatic environments tend to be more rugged and the number of local maxima tends to increase with increasing epistasis. Basins may also exist on different evolutionary time scales; trivially, given an infinite amount of time and the possibility of simultaneous mutations, every evolutionary trajectory will eventually find the global maximum. In Chapter 4, I use my three model landscapes to experimentally investigate the nature of basins of attraction and their relationship to peaks.

When populations evolve, the fine structure of the landscape is important; the mutational neighborhood of a genotype naturally determines what innovations are easily reachable and which are not. Previous experimental work has often focused on a particular gene complex and the evolutionary trajectories to get there. For instance, [Poelwijk et al., 2007] looked at adaptation of the bacterial β-lactamase to cefotaxime, to which the bacteria had not previously had exposure. In this case, the five mutations conferring antibiotic resistance were already known, but they were able to measure intermediates and found that it was necessary for mutational combinations to traverse a valley to reach the resistant solution.

In Chapter 5, I focus on visualization techniques for fitness landscapes that allow us to look exhaustively at the neighborhood of points as well as repeated trajectories originating from these points. In each of the three study landscapes (NK, Avida, RNA), I select 2,000 neighborhoods to study, half around random points and half around peaks. I visualize each study neighborhood up to five mutations out and compare these results to the first five mutations that occur in real evolutionary trajectories. As such, I compare all possible routes with the realized evolutionary routes. In RNA and NK landscapes, I further compare these results to evolutionary trajectories where the strength of selection is increased dramatically, demonstrating a qualitatively different result.

1.2 Contributions

The contributions of this work include:

(1) **A demonstration of the role that common descent plays in biasing genomic estimates of entropy and complexity**, along with a correction method to compensate for common descent, which I validate experimentally.

(2) **An exhaustive exploration and structural analysis of instantiations of three very different landscapes: NK, Avida, and RNA**, that are each are of genome-length 18 with 4 alleles for a total of 68 billion genotypes, allowing us to identify both generalizations about these landscapes as well as distinctions. In all three landscapes, I demonstrate that peaks of higher fitness tend to have higher fitness peaks in their neighborhood. I also found that peaks are connected in the Avida and NK landscapes, whereas in the RNA landscape, peaks tend to be isolated. Similarly, in the Avida and NK landscapes, autocorrelation between peaks with regard to distance is high at low distances and eventually becomes negative. The RNA landscape has the interesting result that the autocorrelation becomes positive again at distance 18, which may point to the complement of the sequence being stable.

(3) **A novel structural method for measuring basins of attraction based on the**

Page Rank algorithm, allowing us to confirm prior results about NK landscapes in a more thorough way. I found that as in previous work, basin size correlated exponentially with fitness and that increasing K increased the size of the average basin while decreasing the size of the largest basins. I also applied my new method to NKp variants of NK landscapes with neutrality and detected that neutral networks can retain significant numbers of trajectories over long periods. Finally, I linked peaks with two-mutant valley crossings and showed that this increases the strength of the exponential relationship between fitness and basin size.

(4) **An identification of basin structure obtained by experimental evolution and analysis of the relationship between peaks and basins**. There was an exponential relationship between basin size and fitness in the NK and RNA landscapes, but not in the Avida landscape. I further show that the final fate of an evolutionary trajectory had very little to do with its starting point—trajectories in all three landscapes ended at a distance from the origin similar to the average distance of all genotypes. Additionally, I found that the concept of peaks play an important role in evolutionary trajectories—in all three landscapes, a significant proportion of trajectories were in the vicinity of a peak.

(5) **A new visualization technique to show landscape structure within a neighborhood** and a comparison of possible routes to realized routes in both random points and peaks.

In order to achieve the contributions above, these analyses were performed at a scale exceeding that previously systematically undertaken. The data alone encompasses over 250 GB, even with compression, hashing, and binary encoding schemes. High performance clusters and 100,000 hours of computing time went into collecting and analyzing this data.

Chapter 2

Estimating Complexity in Finite Populations

2.1 Motivation

Various methods have been proposed to measure biological complexity. Some complexity estimates focus on a mathematical definition of complexity, like Kolmogorov Complexity, which measures complexity by how much a sequence can be compressed. By this measure, a completely random sequence would have maximal complexity, as it could not be compressed at all. However, as in [Adami, 2002], from a biological perspective, I am interested primarily in the information the genome contains about the environment as a measure of complexity. Other methods have been proposed to measure biological complexity as a proxy for adaptation to an environment, as well as for understanding how many sites actively convey an adaptive benefit, versus those that are selectively neutral. I focus on methods that measure complexity without additional manipulations, as this is more tractable in real populations, but is subject to biases, described further below.

For the purposes of this work, I will be focusing on ideas that treat the genome as

an information channel, with noise produced by mutation and reduced by selection. These ideas operate on the assumption that natural selection in a single niche environment will fix adaptive mutations, while permitting non-informative sites to drift randomly, which allows us to search for sites that contain information by examining genomic commonalities within a population.

A flaw of the analysis is linkage—it is possible for neutral or even detrimental mutations to hitchhike along with other mutations that are sufficiently adaptive, that they cause a sweep. This is recognized for instance in [Adami et al., 2000] as adaptive sweeps cause temporary spikes in apparent complexity, but these spikes are clearly spurious. Asexual populations, ignoring mechanisms such as horizontal gene transfer, necessarily contain maximal linkage. Sweeps thus result in distortions during which the diversity of the population is reduced. Likewise, common descent would be expected to cause similar effects; any population will have a most recent common ancestor. Therefore, some of the variation in sites, and therefore entropy measured, can be attributed to this process. Estimates for biological information do not currently take into account neutral processes such as common descent. A large confounding factor in the use of such complexity estimates is that their theoretical assumption of infinite population does not hold for natural populations. If the populations are large enough, statistically, such an assumption might introduce relatively little error, but for those where the population size is greatly dwarfed by the genotypic space (as occurs in virtually every natural population), there will be some distortion driven by coalescence and neutral processes. At the core, this is my primary claim; that given an asexual population with a common fitness function, the population itself is biased relative to the fitness landscape. This means that not all of the traits expressed are necessarily of adaptive value. It is unknown, however, what that limit may be, and one of the goals of this work is to explore this phenomenon further. This phenomenon of complexity being overestimated because of common descent has not been generally examined in the literature. Primarily, complexity is not itself a directly useful metric to many practitioners, whereas mutual information can provide important cues as to

identifying functionally important or related sites, which can in turn yield useful targets for drugs or therapy. However, it is still important to those who want to quantify the information that an organism might hold about its environment.

Because of common descent, I would expect to see a lack of variation on many sites for a population with a large enough genome. Commonly-used information measures overestimate the amount of information in the genome, where the information content or lack of entropy is assumed to be driven by natural selection. But in fact, it may be constrained primarily by neutral processes. For a population of size 10000, there may be variation on only $\sqrt{10000} = 100$ sites at a time, following work by [Zhang et al., 1990] on the diffusion of random walkers in space. This does not imply that the remaining sites are being selected for or in any way contribute directly to the fitness or complexity of an organism. Naturally, this bias would be potentially problematic for complexity estimation. One of the goals of this work is to investigate the strength of this effect and to ultimately compensate for it.

2.2 Related Work & Mathematical Theory

In this section, I discuss related work beginning with early work in information theory and then introducing biological complexity concepts, before delving into common descent. These techniques are sensitive to finite sample sizes, and I discuss common correction mechanisms. Finally, I discuss concepts in neutral theory, linkage, coalescence, and models of evolution.

2.2.1 Information Theory & Complexity

Information theory was first developed by Shannon in [Shannon, 1948], in the context of communication theory. At the time, relatively little was known about the limits of communication channels with regards to transmitting messages across noisy channels, although there was some work by Nyquist and Hartley, both of whom heavily influenced Shannon.

Figure 2.1 shows Shannon's abstraction of a communication system. The information

Figure 2.1: Figure 1 from [Shannon, 1948] demonstrating a conceptualization of a communication system.

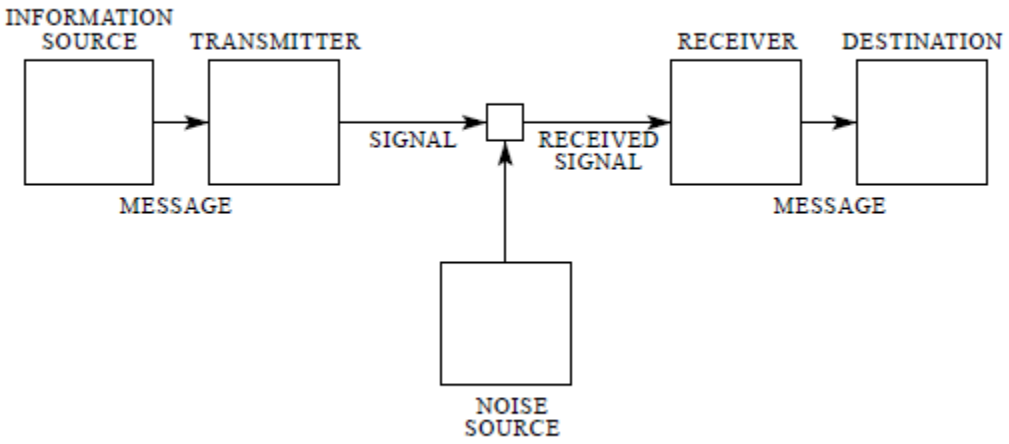


Fig. 1—Schematic diagram of a general communication system.

source contains a message, whether it be a sequence of characters as in a written message or a function encoding such as a function description or parameters. The transmitter is responsible for wrapping the information source into a transmittable signal. On the other end, the receiver decodes the transmitted signal into something meaningful for the destination which is the ultimate recipient of the message. In between the transmitter and the receiver is the channel, which can be any communication medium, including wires and radio signals. The channel is subject to noise and may cause part of the message to become corrupted.

Shannon represented discrete information sources as a Markov process, with the Markov property that each state is conditionally independent of all history except the immediate predecessor state. He looked at the uncertainty in predicting the next symbol from the point of view of the receiver. If the symbol is perfectly constrained by the current state, there should be no uncertainty; and therefore the transmission of the next symbol provides no 'surprise'. The surprise is highest when the chance of observing each symbol in the alphabet is equiprobable. This basic concept was used by Shannon to define a formula for entropy as in equation 2.1.

$$H(X) = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{2.1}$$

X is a random variable with N possible states in Equation 2.1. The use of base 2 assigns units of bits to the uncertainty; another common unit is mers which uses the same base as the alphabet size. This function also corresponds to the average number of yes-no questions that would be necessary to ascertain the state of a random variable.

Shannon then defined the joint entropy of a system with two random variables X and Y, denoted H(X,Y). This measures the uncertainty over each of the $N \cdot M$ possible states for $\{X, Y\}$. This can be seen in equation 2.2.

$$H(X,Y) = -\sum_{\{x,y\} \in \{1,\ldots,N\} \times \{1,\ldots,M\}} p_{\{x,y\}} \log p_{\{x,y\}} \tag{2.2}$$

The conditional entropy of X given Y then is a measure of the uncertainty of Y given the knowledge of X. This is equal to the difference of the joint entropy of X and Y and the entropy of Y. This is defined in equation 2.3.

$$H(X|Y) = H(X,Y) - H(Y) \tag{2.3}$$

From this came the concept of Shannon Information, which describes the information gained from knowing random variable X about the random variable Y. This is often expressed I(X;Y) and may be seen in equation 2.4.

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \tag{2.4}$$

2.2.2 Biological Complexity

Schneider in [Schneider et al., 1986] and others have suggested that the information content of a site can be quantified in aligned sequences as in equation 2.5.

$$I = -L \log A - \sum_{i=1}^{A} p_i \log p_i \tag{2.5}$$

The basic idea underlying this equation is that in an infinitely large population, a site with no selective advantage would be predicted to have frequencies of $\frac{1}{A}$ if there are A candidates for that site. Therefore any difference between the site's entropy and one determined by a completely uniform and random assumption should reflect a measure about the information contained in the population about the genetic landscape. [Adami and Cerf, 2000] characterize this assumption as 'invoking the principle of insufficient reason'. This approach is used by [Adami et al., 2000] to obtain estimates of complexity over whole genomes by summing the per-site information. It is also the approach I will take for this proposal.

It is also worth noting that the complexity measure provided by [Adami et al., 2000] is usually an underestimate of the true complexity. This understimation is caused because estimate subtracts off the entropy between each site individually without regard to the mutual information between sites. One correction, as provided by [Strelioff et al., 2010] and in [Adami, 2004] is to add the mutual information between sites back for an improved estimate. This 2nd-order correction of course still cannot compensate for the interactions between three or more sites, and may still contain error in either direction.

There also exist many alternative ways of estimating biological complexity; for instance, [Huang et al., 2004] examines a single genome and, by classifying the one- and two-point mutants around it as beneficial or deleterious, comes up with an estimate of the complexity of a single genome. This is robust to phenomena such as neutral hitchhiking which emerge during sweeps in population-based measures, but by its nature only involves a single genotype

and a potentially expensive analysis of the local landscape. This may be intractable in natural organisms, since even in computational systems this can become challenging. For instance, for an organism of length 100 with 28 possible instructions per site, there are ${}_{100}\mathrm{C}_2\ 27^2 = 3,608,550$ double mutants. There are also other measures of complexity—the functional information measure proposed by [Hazen et al., 2007], or the usage of the ev model to study the accumulation of information and estimate rates of evolution in the context of coevolving binding sites in [Schneider, 2000] or what [Weinberger, 2002] calls pragmatic information.

For this work, I focus on the simple definition in [Adami et al., 2000].

2.2.3 Entropy Correction

It is well known that when $N << L^A$, where N is the finite population size, L, the length of the genome, and A the number of alleles, that the entropy is poorly estimated since the population size cannot adequately cover the genotypic space. This limitation has spawned a number of estimators that seek to correct for the issue of biased sampling.

The most basic of these is the Maximum Likelihood Estimator (MLE) and can be seen in Equation 2.6.

$$\hat{H}(\theta) = -\sum_{p \in P} \theta_p \log \theta_p \tag{2.6}$$

where $\theta_p = \frac{p_n}{N}$

The MLE Estimator does poorly in the $N << L^A$ case since it will estimate many probabilities at zero as a population can cover only a tiny fraction of possible genotypes.

Another common family of approaches is Bayesian Estimators. [Minka, 2000] The basic idea is to estimate the probability of the entropy using a Bayesian prior, often the Dirichlet distribution. As with Bayesian approaches in general, the choice of an unbiased prior distri-

bution is nontrivial and a considerable number of different priors have been proposed, with varying results. This can be seen in Equation 2.7

$$\hat{H}(\theta) = -\sum_{p \in P} \theta_p \log \theta_p \tag{2.7}$$

where $\theta_p = \frac{p_n + \alpha}{N + \alpha|n|}$, where n is the number of categories and α is a parameter determining the prior.

One frequently used form is the Laplace Estimator: $\theta_p = \frac{p_n + 1}{N + n}$.

More advanced estimators have also been developed; for instance, [Nemenman et al., 2002] improves on methods that use prior Dirichlet uniforms and instead tries to estimate the characteristics of the distribution from looking at coincidental samples. [Hausser and Strimmer, 2009] also discusses the efficacies of various methods of entropy estimation, including the NSB estimator from [Nemenman et al., 2002], maximum likelihood, the Chao-Shen estimator proposed by [Chao and Shen, 2003], and the James-Stein estimator under a variety of distributions and sampling counts. [Paninski, 2003] also compiled several solutions to correct bias in entropy inherent in using MLE estimation methods.

There have also been attempts to correct for bias in sample selection, which assume there is some systematic differences between the sample to be studied and the population it is drawn from. Examples include [Dudk et al., 2005] who propose three approaches to correct bias, motivated partly by the problem of habitat estimation. [Dudk et al., 2005] points out that sampling distribution is driven by human factors driving ease of access such as roads, proximity of cities, and rivers, whereas species distribution is likely to be at least somewhat independent of those same factors. As such, these biases may cause species to be systematically oversampled in the case of species that prefer such environments or undersampled in the case of species that do not, which may easily lead to estimation errors for statistics such as population size.

I seek to show in the course of this work that the entropy distortion caused by common descent can, under many conditions, dwarf the error caused by finite sampling bias. This bias from common descent thus has implications for estimation of information in genomic contexts. Further, I demonstrate that this can be a proxy for estimating neutrality.

2.2.4 Neutral Theory & Evolution

In his seminal work on neutral theory, [Kimura, 1968] observed that the rate of molecular evolution seemed high compared to what would be predicted from the deleterious effects of most gene mutations. This led Kimura to propose that most molecular mutations may actually be neutral in nature and that the force of random genetic drift is more important than previously thought. The theory generated significant interest from researchers over following years and provided an influential null-model of evolution; focus was directed to whether neutral evolution could explain observed outcomes, with the goal of distinguishing functional parts of the genome and those under active selection from those under no selection, and understanding what types of selection may be at play. The relevance of this theory to complexity estimates is that neutral evolution allows multiple mutations to persist at a site, thereby increasing entropy at that site and decreasing information estimates. The null hypothesis for information estimates used by [Adami and Cerf, 2000] utilizes the assumption that given an infinite population, neutral evolution would cause a non-contributing site's fitness to be maximally divergent. For a given neutral site, one would expect 0 information, since $H_{max} = \log A = H$ and therefore the complexity estimate $C = H_{max} - H = 0$. I further investigate what this means for finite cases later in this work.

2.2.5 Coalescence

If one looks at the population as a whole and traces the phylogenetic tree back in time, eventually the population 'coalesces' at a common ancestral organism in asexual populations. This important fact underlies the foundation of coalescent theory. There are some well-known

results about the statistical properties of the distribution and timeline of how individuals are related. Understanding and incorporating these results will allow us to reexamine our complexity estimates, providing clues to how many sites are fixed by chance and relatedness as compared to sites that are fixed by selection.

[Derrida and Peliti, 1991] examined the process of evolution in flat fitness landscapes. Evolution on flat fitness landscapes is important because it helps shed light on variation caused by drift and coalescent processes. This better informs our understanding about the process of evolution in the presence of selection. Following their derivations, when the population size stays constant, each individual leaves one offspring on average, which implies a Poisson distribution of mean 1, at least where the population size $M >> 1$. The probability that an individual leaves behind m offspring is shown in Equation 2.8.

$$p_m = \frac{e^{-1}}{m!} \tag{2.8}$$

[Derrida and Bessis, 1988] showed that, given the definition that two individuals belong to the same family if they have a common ancestor within a specified time window, the distribution of the sizes of the families is equiprobable given the number of families that currently exist in flat fitness landscapes So, if there are K families, all possible ways to allocate the individuals of the population into those K families are equally likely.

2.2.6 Linkage Disequilibrium

Linkage disequilibrium (LD) refers to a non-independent association of alleles at different loci. LD may be driven by selection; for instance, a maladaptive mutation at a particular site may be consistently remedied by a compensatory change at another site. LD may also be driven by shared inheritance. It is frequently used in terms of studying sexual populations, since alleles that are physically close to each other on a genome are less likely to be broken up by recombination. In asexual populations, there is maximal linkage disequilibrium, since

there is no recombination to break up associations between alleles, which can result in the accumulation of neutral or even deleterious mutations via hitchhiking.

Following [Balding, 2006], there is no direct quantitative way to calculate LD, but three well-known proxy measures are D, D' and r^2.

The D statistic is a measure of the deviation of the observed frequencies of a genotype from the expected quantity. An example of such a calculation for a simple two-locus two-allele model can be found in equation 2.9, where X_{i_1,j_1} is the count of genomes with alleles i_1 and j_1 at loci i and j and p_{i_1} and p_{j_1} are the probabilities of the respective alleles appearing at those sites. The D' statistic rescales D by dividing it by its theoretical maximum. D' is again shown for the two locus, two allele case in equation 2.10. However, as [Balding, 2006] notes, these statistics are more likely to take on extreme values when allele frequency is low and are often not directly comparable between loci unless the frequency distribution is similar. The r^2 statistic is the correlation coefficient between the pair of loci and can be seen for instance in Equation 2.11.

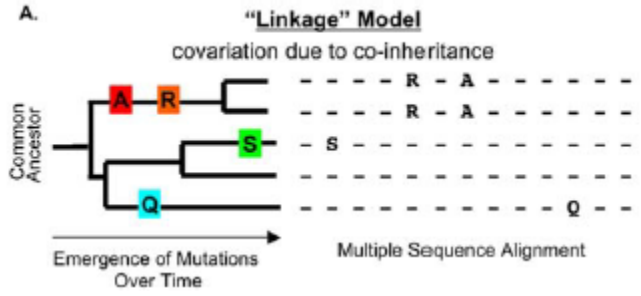
These measures all use two loci; they can be summarized for a region through simple averaging, or through more complex techniques such as the LD map proposed by [Maniatis et al., 2002] which uses an exponential decay function of the local distance to fit a summary LD estimate.

$$D = X_{i_1,j_1} - p_{i_1}p_{j_1} \tag{2.9}$$

$$D' = \begin{cases} \frac{D}{\min p_{i_1}p_{j_1}, p_{i_2}p_{j_2}} & \text{if } D > 0 \\ \frac{D}{\min p_{i_1}p_{j_2}, p_{i_2}p_{j_1}} & \text{if } D < 0 \end{cases} \tag{2.10}$$

$$r^2 = \frac{D}{\sqrt{p_{i_1}p_{i_2}p_{j_1}p_{j_2}}} \tag{2.11}$$

Figure 2.2: Figure 1A from [Wang and Lee, 2007] demonstrating background linkage disequilibrium. The covariation of mutation A and R in the sequence alignment is caused by common descent. This can be seen in the phylogenetic tree.



A related concept is background linkage disequilibrium (BLD), which can be thought of as linkage disequilibrium that is driven primarily by inheritance. [Wang and Lee, 2007] outlined this problem, describing it as the problem of distinguishing covariation due to selection from covariation due to common descent. Their formulation can be seen in Figure 2.2. They investigated synonymous and nonsynonymous mutations and measured a low rate of linkage disequilibrium in the HIV virus that decays with site distance due to recombination events.

[Korber et al., 1993] also examined mutual information between sites in the V3 loop of HIV and hypothesized that the covariance may be explained in part by a shared evolutionary heritage. [Bickel et al., 1996] continued the analysis of the data of [Korber et al., 1993] and proposed that it may be possible to compensate for the effects of shared ancestry by examining the similarity of sequences within suspected clades. [Service et al., 2001] estimated BLD in human genomes from a small subpopulation and found that although genetic distance strongly modulates BLD, its distribution is nonetheless nonuniform. This kind of analysis is potentially significant for researchers since it can help to identify functional sites that could be targets for drug development.

2.2.7 Models of Evolution

There are several models that have been developed in order to study evolutionary dynamics, including the Wright-Fisher model [Wright, 1931, Fisher, 1930] and the Moran model, among

others. However, for the purposes of this work, I focus on sequence evolution in asexual organisms. As such, I utilize the quasispecies model, also known as the Eigen model. I also discuss common assumptions used in mathematical simplifications and their use in deriving mathematical results about evolution in these domains.

2.2.7.1 Eigen Model

The discrete-time Eigen model, developed in [Eigen, 1971] and [Eigen and Schuster, 1977] and further applied in [Eigen et al., 1988] and [Eigen et al., 1989] is a discrete formulation of the quasispecies model devised in order to study sequence evolution. The quasispecies concept relates to the idea that given that a population is in equilibrium in the presence of mutation and selection, there may be stable molecules of lower fitness maintained by mutation, particularly if they are close to a local molecule of high fitness. The maintenance of such quasispecies would have an effect on information estimates utilizing sequence information. This equation is reproduced in Equation 2.12. Here Q_{ij} is the probability of mutating between genotype j and genotype i, A_i is the fitness-driven replication rate of genotype i, and x_i is the frequency of genotype i in the population. There has been some debate about whether the quasispecies models and more traditional population genetics models are distinct, but some, such as [Wilke, 2005] believe that they are in fact studying the same phenomena and are not contradictory. The Eigen model is the primary model I use for experimental evolution. Previous work such as [Adami et al., 1995] and [Bedau and Brown, 1999] have examined frequency distributions in instantiations of the quasispecies model. [van Nimwegen et al., 1999] explored the neutral evolution of mutational robustness in neutral plateaus.

$$\dot{x}_i(t) = [A_i Q_{ii} - E(t)]x_i(t) + \sum_{j \neq i} A_j Q_{ij} x_j(t) \tag{2.12}$$

2.2.7.2 Infinite Alleles Model

The infinite alleles model relies on the idea that genes tend to be composed of large segments of nucleotides [Hartl and Clark, 2006]. Thus, any new mutation is likely to create a new allele that does not exist in the population. [Ewens, 1972] derived the result that the expected number of alleles k is related to the sample size, n. This can be seen in Equation 2.13.

$$E[k] = 1 + \frac{\theta}{\theta + 1} + \frac{\theta}{\theta + 2} + ... \frac{\theta}{\theta + n - 1} \tag{2.13}$$

The relevance of this particular model to information estimates may be in analyzing real data where the maximum number of alleles is not known. This expected number of alleles then can be used as a proxy for the amount of information contained within a gene, since this provides us with an estimate of how many alleles there ought to be in the complete absence of selection. The main mathematical advantage in this assumption is simply that every mutation is necessarily novel, so accounting for repeated mutations or back mutations can be disregarded.

2.2.7.3 Infinite Sites Model

The infinite sites model considers an infinite sequence of loci. With an infinite-length genome, any novel mutation necessarily alters a different site. The infinite sites model is a popular choice for studying DNA sequence evolution at least partially because it offers certain mathematically attractive qualities. For instance, the probability of the same site being mutated twice is zero with infinite sites from which to pick, which means that any mutation may be treated as novel. There has been some relevant work focused on trying to find analytic formulations for the numbers of neutral sites.

[Watterson, 1975] was one of the many early attempts to distinguish between alleles experiencing neutrality and those undergoing active selection. The Watterson estimator

estimates the amount of segregating sites that would be expected given a certain mutation rate in the infinite sites model.

One of the applications of the Watterson estimator is to estimate the mutation rate given real data, as given in Equation 2.14. Here S_n is the number of segregating sites, θ is the population mutation rate, and h_n is the harmonic series. For instance, in [Durrett, 2008], an example is given with data from from Ward et al., 1991, in which 63 human mitochondrial sequences were sequenced, which yielded a count of 26 segregating sites. It was then possible to solve for θ in the equation, which works out to be $\frac{26}{h_6 3} = 5.5173$. Dividing this by 360 sequences gives a population mutation rate per base of 0.0153.

$$E[S_n] = \theta h_n, \text{where } h_n = \sum_{i=1}^{n-1} \frac{1}{i} \tag{2.14}$$

[Watterson, 1977] also examined distributions with the goal of trying to distinguish neutrality from heterosis. Watterson uses the order-statistics distribution. This defines the r most numerous allele frequencies, ranked in decreasing order. Specifically, Equation 2.1.12 from [Watterson, 1977] can be seen in Equation 2.15. He also uses the truncated distribution, which is the same as above, except with only frequencies above a certain threshold considered.

$$x_1 \geq x_2 \geq x_3 \geq ... \geq x_r \geq 0, \text{with } \sum_{i=1}^{r}(x_i) \geq 1 \tag{2.15}$$

Estimators such as the Watterson Estimator provide a compelling theoretical framework for assessing the expected number and variance of segregating sites in a population under the presence of no selection. These estimators provide a useful baseline for investigating fixation and distinguishing between neutral processes and the influence of natural selection. For example, suppose using the estimator, I predict that there should be X segregating sites in the population. Instead, there are Y $<$ X sites. It may be possible to conclude then within some degree of certainty that several sites are maintained by active pressure of selection.

However there are potential pitfalls here as well—for instance, the maintenance of two or more coexisting subtypes may cause systematic divergence across an array of sites. This divergence would have the effect of increasing the observed number of segregating sites. This scenario similarly confounds Shannon-Information-based complexity estimates, but may be detectable with an analysis of site frequency distributions.

[Tajima, 1989] built on the work by Watterson and suggested that the difference between the estimate of mutation rate derived from segregating sites and an estimate of the mutation rate, Θ, derived from the average number of pairwise mismatches, Π. At steady state, the average number of pairwise mismatches follows the direct formula: Equation 2.16

$$E[\Pi] = \Theta \tag{2.16}$$

This provides a D statistic in Equation 2.17

$$D = \frac{\Pi - \frac{S_n}{h_n}}{\sqrt{\mathrm{Var}(\Pi - \frac{S_n}{h_n})}} \tag{2.17}$$

[Fu and Li, 1993] also introduced a test statistic, G, that compares the difference of the two estimates of the mutation rate in order to measure the fit of a segment of genetic material to a neutral infinite-sites model at steady state. Here they utilize the structure of a genealogy without recombination, and divide the phylogenetic tree into internal and external branches, where external branches emerge from an internal node and terminate at a sequence, and internal branches connect internal nodes, which correspond to reconstructed ancestors. The expected time for the sum of the external branches is $4N_e$ where N_e is the effective population size. The significance of the external branches, following [Hartl and Clark, 2006] is that mutations along them result in a singleton nucleotide that is most likely not replicated anywhere else (by the infinite sites assumption). They produced the G statistic, which again

examines the difference of two estimates of Θ seen in Equation 2.18, where η_e is the number of mutations along external branches and η_i is the number of mutations along internal branches. [Fu, 1995] further investigated statistical properties of segregating sites by expanding on this idea of using the internal and external branches. He found, using the Wright-Fisher model, and that the expected number of mutations of size i,ζ_i can be given by Equation 2.19. This expectation holds true for the infinite alleles model and holds true asymptotically for the infinite sites model. In addition, he derived values for the variance of the size of the expected number of mutations and the covariance between expected numbers of mutations of different sizes.

$$\frac{\nu_e - \frac{\eta_i}{h_n - 1}}{\sqrt{\eta_e - \frac{\eta_i}{h_n - 1}}} \tag{2.18}$$

$$E[\zeta_i] = \frac{1}{i}\theta \tag{2.19}$$

[Zhang et al., 1990] examines populations as a collection of N random walkers (organisms) in a d-dimensional space (genetic lattice). They showed that over time, with a low mutation rate, the populations tended to congregate in small regions of genotypic space, even given an initial uniform distribution. They found that the population evolving on a d-dimensional lattice in equilibrium would have an actual dimension proportional to $\sqrt{N}$ where N is the size of the population. If $N << D$, then this is in fact a significant divergence from a uniform distribution; in other words, one expects to find convergence within a relatively small area, even with a complete absence of selection.

2.3 Approach

In order to illustrate the influence of coalescent processes on complexity estimates in finite populations, I conducted a series of experiments. I did not make any mathematical assump-

tions like the infinite sites / infinite alleles approaches outlined above, and instead rely on experimental data.

2.3.1 Methods

The primary experimental model I am using to evaluate the effects of neutral processes on information and complexity estimates is a string model based on the discrete Eigen model. In this model, a population with N individuals is maintained, each with L loci, each of which can take on one of A values. Every generation, N new individuals are sampled with replacement as parents from the population, proportionally to fitness, and mutation is applied to their offspring. This model can be used to study both genes with alleles abstracted at a single locus and more complex phenomena such as genes with fitness driven by contributions from multiple sites.

To isolate the baseline effects of population dynamics on calculated information, it is necessary to understand how these dynamics work in the absence of selection. Thus, for this work, I studied first flat fitness landscapes where mutations provide neither fitness benefit nor penalty, and all individuals have the same fitness and the same probability of producing offspring—$\frac{1}{N}$ where N is the population size. This landscape would be expected to produce the most extreme complexity estimate distortion due to shared descent.

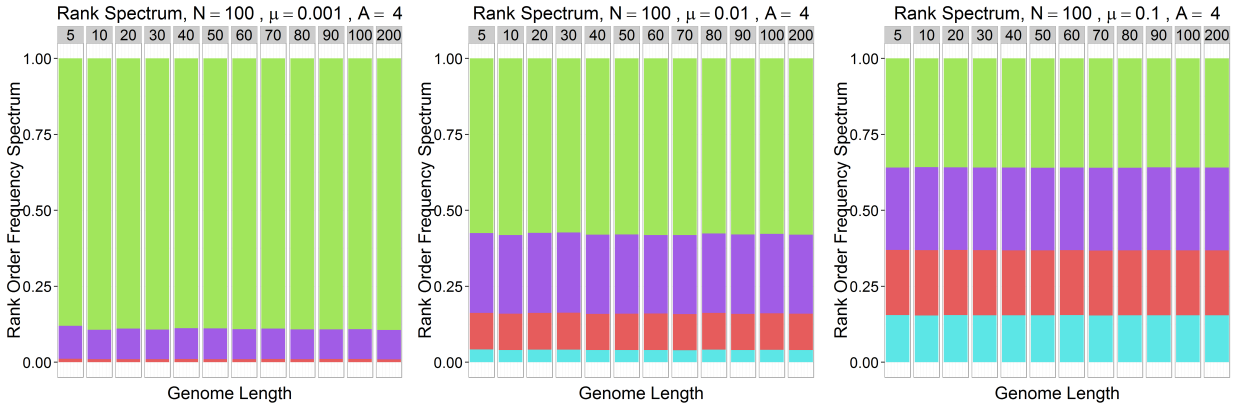
I used Gibbs sampling here, with an interval of 100 generations between samples for independence. The populations start from a common ancestor and 10,000 generations executed to allow the starting population to fan out from a single starting genotype with all zeros. The burn-in time is sufficient time to allow the population to no longer have any discernible bias towards zero, over other alleles. I sample the H_1 estimate of complexity as in [Adami and Cerf, 2000] at every time step. I did this for $A = 4$ and $A = 20$ alleles, consistent with DNA/RNA and amino acids in proteins.

2.3.2 Common Descent Biases Complexity Estimates

2.3.2.1 Rank Order Analysis

Similar to the analysis done in [Watterson, 1977], it is possible at any given time point at a site to rank the count of loci in numerical order of appearance. There appears to be a fairly stable distribution that is roughly length-independent, but that does vary with mutation rate and population size. For length 100, the rank-order distribution can be seen in Figure 2.3. The same plots can be seen for $N = 1000$ in Figure 2.4. These figures of rank-order distribution show what the *averaged* site frequency order statistics look like through time, although it is crucial to note that individual samples do in fact deviate from this distribution.

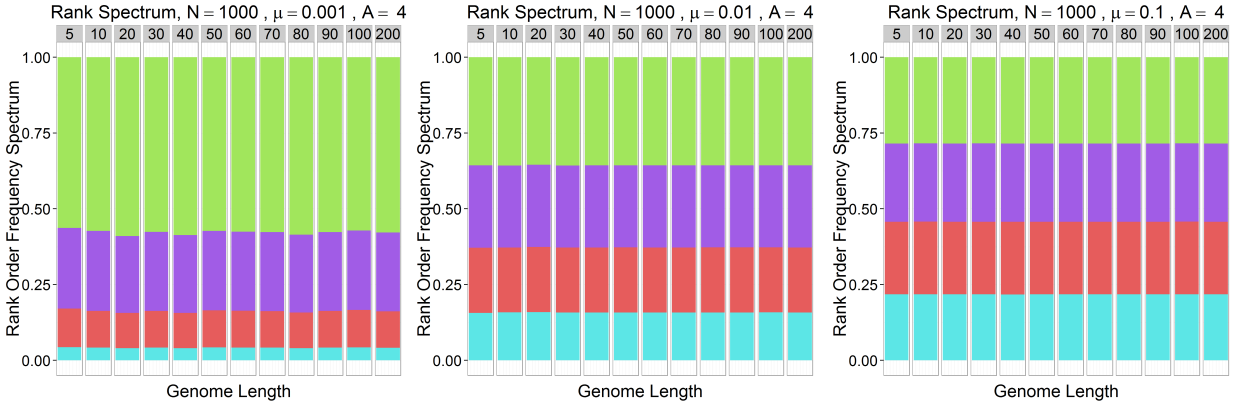
Figure 2.3: Frequency rank order for population size of 100. On the x-axis for each plot is length. At each sample in each run, allele frequency per site is ranked and averaged. This is then averaged across all samples per run to condense into a single figure per run. This shows that on average, even in the absence of selection, an imbalance exists on sites due to common descent. Colors come from rank ordering, e.g. in these plots the frequency of the most common allele appears in lime green at the top.



(a) Mutation rate 0.001 **(b)** Mutation rate 0.01 **(c)** Mutation rate 0.1

Both increasing population size and increasing mutation rate will, in turn, increase the diversity of the visible alleles at each site and this, in turn, increases the entropy. Although this aggregate distribution is stable, this is a summary result; sites are not actually in equilibrium on average as one would expect in a regime without selection. Measurements can and do vary from time slice to time slice and sometimes drastically.

Figure 2.4: Frequency rank order for population size of 1000. On the x-axis for each plot is length. At each sample in each run, allele frequency per site is ranked and averaged. This is then averaged across all samples per run to condense into a single figure per run. This shows that on average, even in the absence of selection, an imbalance exists on sites due to common descent. This effect is reduced, compared to the smaller population size of 100 in Figure 2.3

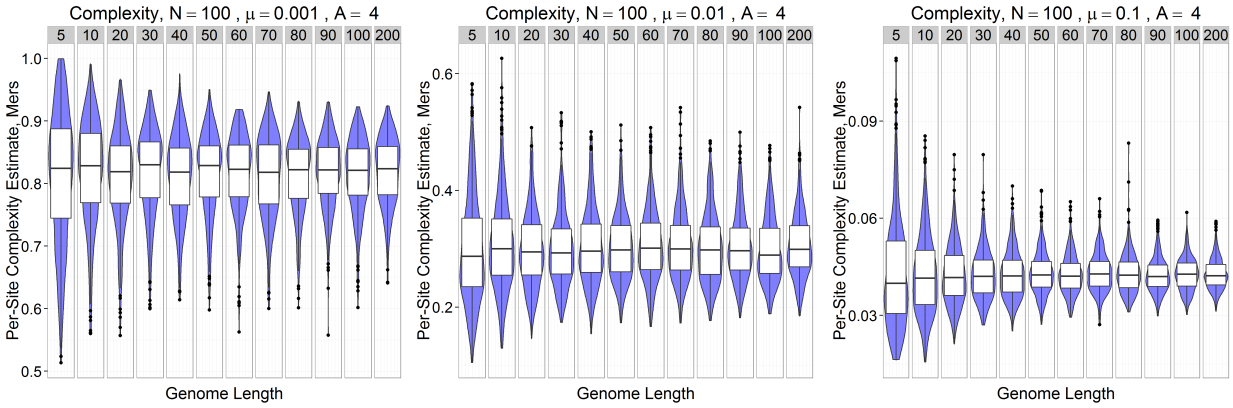


(a) Mutation rate 0.001 **(b)** Mutation rate 0.01 **(c)** Mutation rate 0.1

2.3.2.2 Complexity Estimates

My results show that in the absence of selection, this estimate still suggests that there is a considerable amount of complexity depending upon the experimental parameters, as might be expected from the divergence from uniform visible in the rank-order distributions. Figure 2.5 contains complexity distributions for three experimental conditions with increasing mutation rate from left to right. These experiments were conducted with four possible alleles per locus and fixed length, for many different values of length. Length plays a directly linear role in the complexity estimates, leading to a virtually constant per-site estimate, likely due to the fully asexual descent in these populations; in other words linkage is complete between any pair of sites.

Figure 2.5: Complexity Estimates for population size of 100 in the presence of no selection. On the x-axis is length and on the y-axis is information estimates. Each plot represents the distribution of information samples over a single run. The sample size increases with increasing length, leading to less uncertainty over the estimate.



(a) Mutation rate 0.001 **(b)** Mutation rate 0.01 **(c)** Mutation rate 0.1

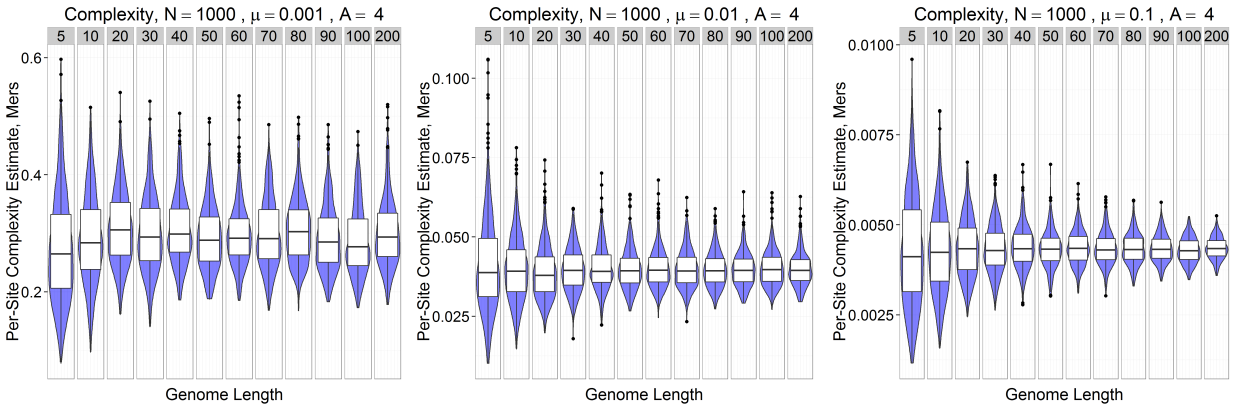
The complexity varies by mutation rate. This result is not surprising, since with a lower mutation rate, there is less of an opportunity for sites to diverge after the most recent coalescent event. The result is that many sites are fixed. As the mutation rate increases, there is more diversity at each site, which reduces the entropy, which increases the complexity estimate. The complexity estimate can be sizable; for instance, the maximum entropy for

length L is $\log_2 4 = 2$ bits per site, or $\log_4 4 = 1$ mer per site. Thus, with a mutation rate 0.001, nearly 80% of the maximum complexity estimate would appear, on average, even in the absence of any selective pressure. This 'spurious complexity' decreases as the mutation rate increases, with a mutation rate of 0.1 per site yielding a less than 5% total estimate of maximum complexity.

The effect also appears to be linear with respect to the L parameter within each condition. This is consistent with the theoretical literature, since the number of segregating sites is tied to θ, the whole genome mutation rate. Since there is a per-site contribution, the whole genome mutation rate is driven by the length of the genome.

It is also interesting to contrast the results shown in Figure 2.5 and Figure 2.6. The information estimates in the population size of 1000 are noticeably smaller compared with those with the population size of 100. In a larger population, the time from the most recent coalescence will be longer on average, and the population has thus had more time to diverge relative to this event. This increases the entropy and thereby decreases the complexity estimate, making it closer to the true complexity of 0 when no selection is present.

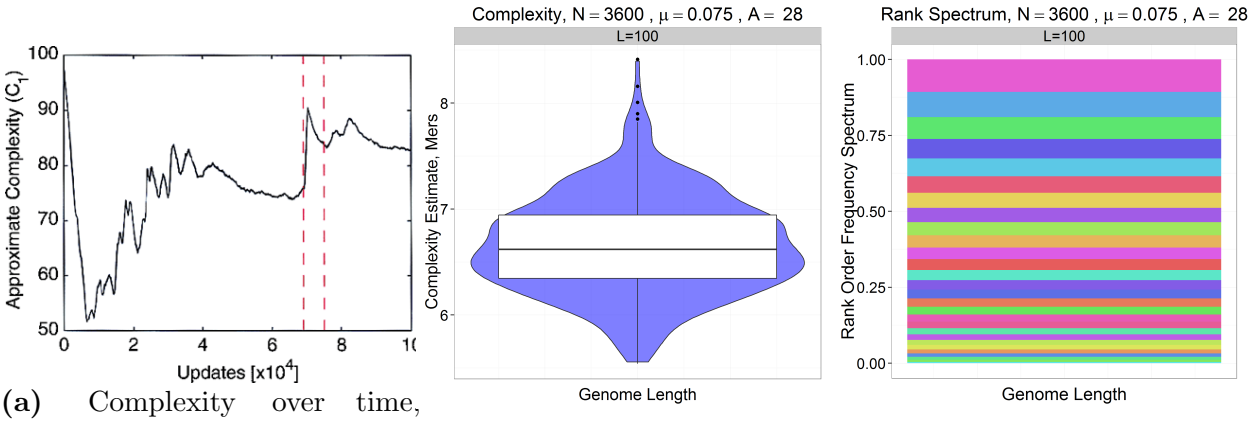
Figure 2.6: Complexity Estimates for population size of 1000 in the presence of no selection. On the x-axis is length and on the y-axis is information estimates. Each plot represents the distribution of information samples over a single run. The sample size increases with increasing length, leading to less uncertainty over the estimate.



(a) Mutation rate 0.001 **(b)** Mutation rate 0.01 **(c)** Mutation rate 0.1

The complexity graph from evolving a population of 3600 organisms [Adami et al., 2000] of length 100 can be seen in Figure 2.7. There are some differences between the models; in this work, simultaneous replacement per generation is used, whereas the Avida model used by [Adami et al., 2000] tends to replace individuals continuously.

Figure 2.7: Comparison of complexity estimates from [Adami et al., 2000] and expected complexity in mers from neutrality and resulting rank order frequency distribution for corresponding neutral landscapes of length 100 with 28 bases per site. In 2.7a, the complexity is measured over time, from [Adami et al., 2000]. In 2.7b is the distribution of information estimates that occur in a size 3600 population with 28 alleles with no selection. In 2.7c is the distribution of rank ordered alleles, which clearly deviates from the uniform $\frac{1}{28}$.



(a) Complexity over time, [Adami et al., 2000]

(b) complexity, neutral

(c) rank order, neutral

This result implies that a completely neutral distribution would register roughly 6.6 mers of information. The complexity estimates in [Adami et al., 2000] are well in excess of that figure, but nonetheless this could be as much as 10% of the final complexity estimate in an experimental model with a fairly robust population size. Thus, it does suggest that there is room for improvement in refining some of these complexity measures, even in relatively large populations.

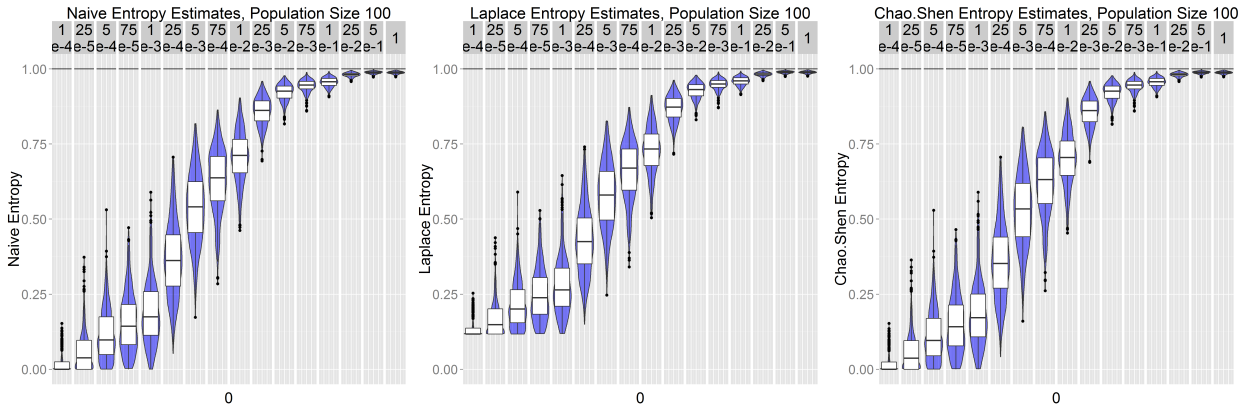
2.4 Finite Size Correction

As discussed previously in this chapter, there are a variety of entropy estimators that attempt to control for finite population size. I demonstrate in this section that the effect observed is

not simply due to choice of the naive estimator. The naive estimator is always an underestimate of the total amount of entropy present at each site [Schürmann and Grassberger, 2002]. Therefore, I look at the Laplace estimator (which is likely to be an overestimate) and the Chao-Shen estimator. The key takeaway here is that using a better estimator is not going to substantially change the necessity for correcting for common descent.

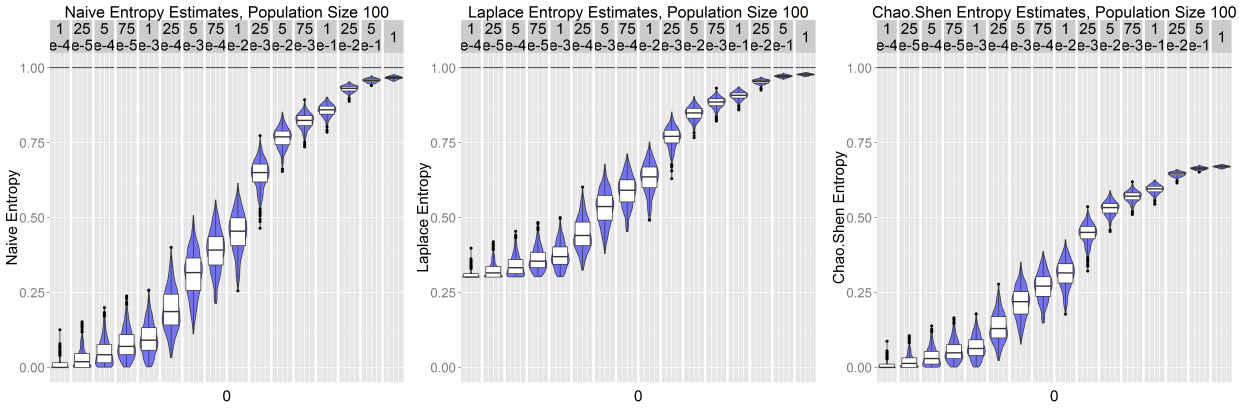
In order to demonstrate this problem, I first look at the complexity estimates for both four and twenty alleles to enable comparisons against RNA and Proteins respectively. Figure 2.8 and 2.9 show a cross section of the complexity estimates along with the true entropy (known *a priori* to be exactly 1 mer by the simple fact that there is no selection in the system).

Figure 2.8: Per-site complexity estimates for neutral populations without selection of size 100 with 4 alleles per site for three different estimators. On the x-axis is mutation rate and on the y-axis is information estimates. Each plot represents the distribution of information samples over a single run. The line represents the 'true' entropy from the underlying distribution which should be 1 since there is no selection. The mutation rates vary from 0.0001 on the left to 1 on the right, with semilog scaling.



(a) Naive Estimator **(b)** Laplace Estimator **(c)** Chao-Shen Estimator

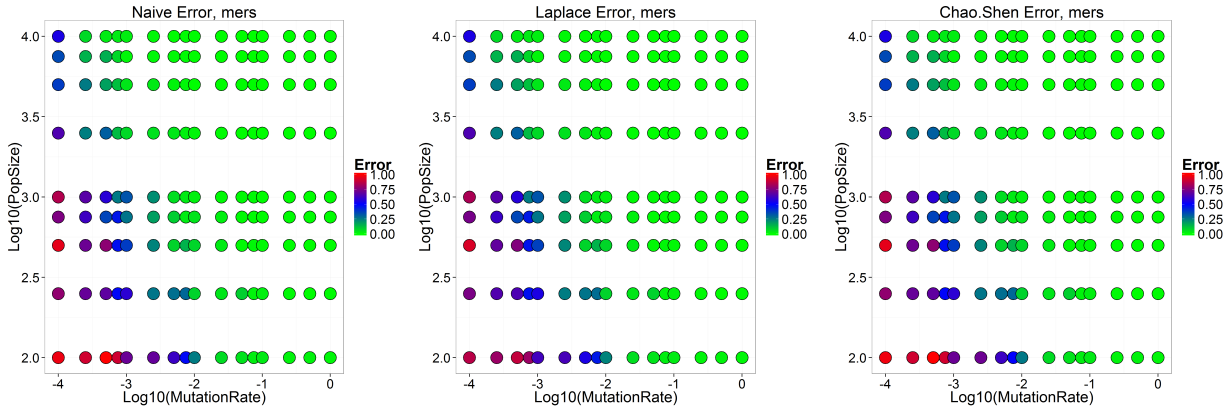
Figure 2.9: Per-site complexity estimates for neutral populations without selection of size 100 with 20 alleles per site for three different estimators. On the x-axis is mutation rate and on the y-axis is information estimates. Each plot represents the distribution of information samples over a single run. The line represents the 'true' entropy from the underlying distribution which should be 1 since there is no selection. The mutation rates vary from 0.0001 on the left to 1 on the right, with semilog scaling.



(a) Naive Estimator **(b)** Laplace Estimator **(c)** Chao-Shen Estimator

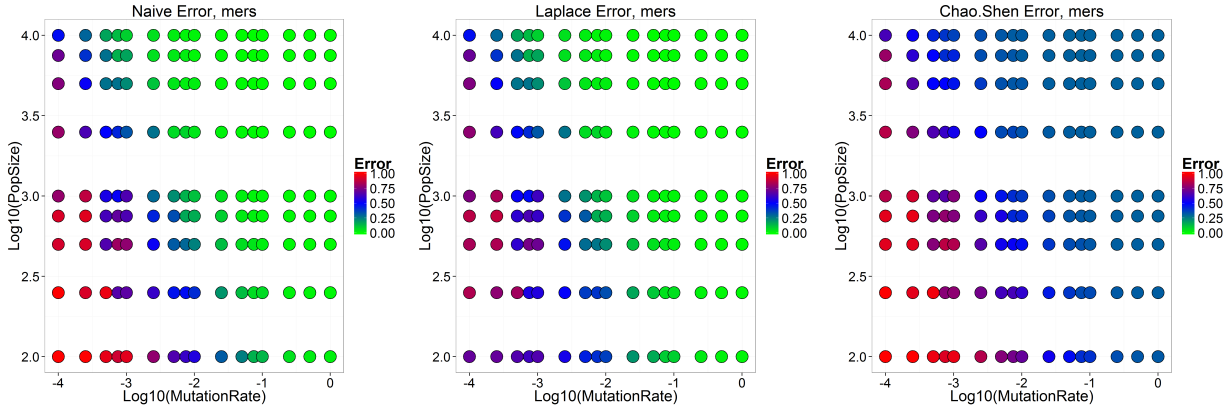
This pattern stays consistent across the entire sample set. Figures 2.10 and 2.11 demonstrate this. The error is less with the number of alleles

Figure 2.10: Relative error compared to true entropy for neutral populations for three different estimators. On the x-axis is length and on the y-axis is information estimates. Each point represents the distribution of information sample means over a single run.



(a) Naive Estimator **(b)** Laplace Estimator **(c)** Chao-Shen Estimator

Figure 2.11: Relative error compared to true entropy for neutral populations for three different estimators with 20 alleles. On the x-axis is length and on the y-axis is information estimates. Each point represents the distribution of information sample means over a single run.



(a) Naive Estimator **(b)** Laplace Estimator **(c)** Chao-Shen Estimator

The takeaway is that current entropy correction schemes do not correct for the issue of common descent. All three, unsurprisingly, do better as population size and mutation rate both increase. This means that the choice of entropy bias correction method is itself unlikely to help significantly in correcting for spurious complexity deriving from common descent and does not matter that much in this particular case as the vast majority of the error is not coming from entropy estimation error or simple biased sampling. The sample is always biased by selection.

2.5 Correcting for Common Descent

In the previous section, I raise the question of how common descent affects results, examine entropy correction methods, and demonstrate that existing methods do not correct for common descent. In this section, I seek to correct calculations of entropy and information to control for common descent given *a priori* knowledge about the fitness landscape. For any population-based calculations, I must assume a single niche environment and a population at equilibrium. In order to determine the idealized dynamics for maximum and minimum entropy at a site, I define two types of sites, informative sites, at which selection is acting, and neutral sites, which are not directly subject to selection. I define p as the proportion of the genome containing neutral sites. I further designate H_n as cumulative entropy from all non-informative (neutral) sites and H_s as entropy from informative (selected) sites. Earlier in this chapter, I have seen that H_n will be non-maximal because of coalescence. And in finite populations, H_s will be nonzero, due to mutation-selection balance.

I can then write total entropy as:

$$H_m = p(H_n) + (1 - p)(H_s) \tag{2.20}$$

For my correction, I first must estimate p and then use the formula to estimate the amount

of entropy H_c attributable solely to neutral sites.

$$H_c = p(H_n) \tag{2.21}$$

2.5.0.3 Methods

To test this correction, I used a royal road landscape where every allele provides a selective factor of twice the fitness. Parameter sweeps provided us with the H_s estimate for use in the formula above. I analyzed both 4-allele and 20-allele test conditions.

Next, I examined a landscape where precisely half of the sites were selected and half were completely neutral. This leads us to an expectation of $\frac{L}{2} \log_2 A$ bits of complexity for each genome, which in turn leads to an average per-site expected complexity of $\frac{\log_2 A}{2}$. For each estimate, I used Gibbs sampling, again measuring the entropy every 100 generations. For all these results, I focused on the naive entropy estimate, since I previously established that other corrections do not correct for common descent. The actual value gives us H_m which is our measured estimate for the half-selected landscape.

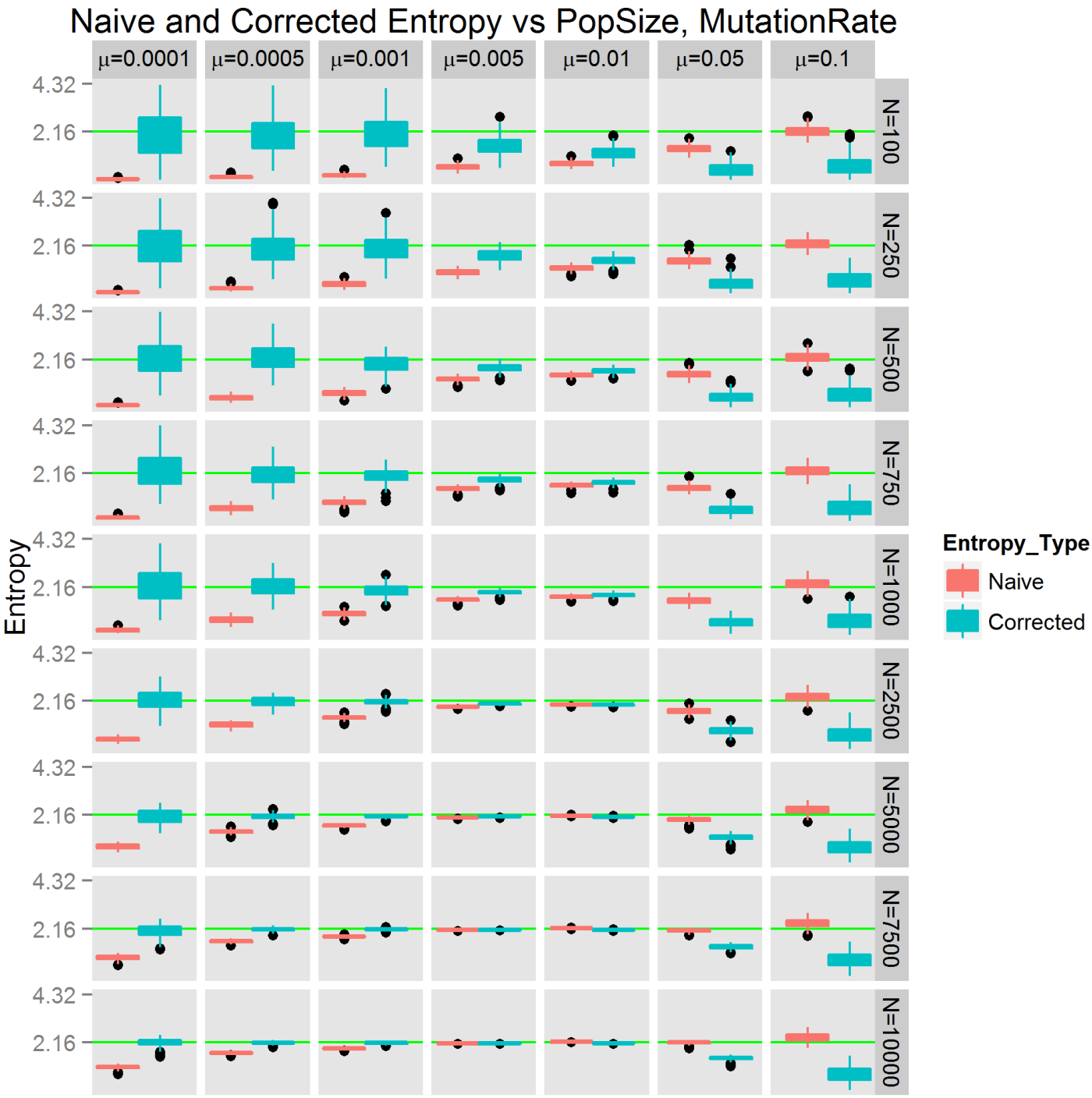
Finally, from my results before with completely flat landscapes, I have estimates for H_n. For each mutation rate & population size for both the 4-allele and 20-allele cases, there are three quantities: H_m, H_n, and H_s. I then apply Equation 2.20 to obtain an estimate for p and Equation 2.21 to obtain a new entropy.

2.5.1 Discussion

The results of our correction are summarized in Figure 2.12 and Figure 2.13.

Figure 2.12: Entropy, Naive & Corrected, 4 Alleles

Figure 2.13: Entropy, Naive & Corrected, 20 Alleles



For both allele counts that I tested, some common trends emerge. The correction performs very well at low to moderate mutation rate <= 0.01 and corrects for the effect of common descent. Beyond that threshold, however, the correction begins to perform poorly—even doing worse than the naive estimate. It is likely that the core assumption behind this correction, namely that sites that are actively selected do not hold entropy, is being violated. Selected sites begin to hold entropy because it is difficult for evolution to stay on a fitness peak as mutation rate increases past the error threshold, thereby maintaining both variance and entropy. It is likely possible to correct for this phenomenon, but we do not attempt to do so in this work.

In the real world, however, certain assumptions are not likely to hold—for instance a selected RNA site that may contain either a G or a C would constitute a selected site with nonzero entropy. Additionally, and perhaps more seriously, estimating H_s could be challenging since, of course, perfect fitness landscape information would not be available. It may also be possible to obtain an estimate for H_s using a known subset of the genome that is highly conserved. H_n, by contrast, can be obtained by virtue of simulation rather readily given allele count, mutation rate, and population size. Second, this particular analysis is specifically done in asexual populations—dynamics and numbers will be different in sexual populations, although a similar analytic approach should hold.

In conclusion, the contributions of this work are a demonstration that complexity from shared descent exists and can be substantial, a technique to compensate for the effects of common descent when calculating entropy or complexity, and a demonstration of its usage on an evolving population.

Chapter 3

Fitness Landscapes– Peaks, Ridges, and Plateaus

Fitness landscapes represent the underlying genetic structure on which populations evolve. As such, understanding this structure can help us better understand the process of evolution. More broadly, there is continued debate on the structure of fitness landscapes in biology, with consequences for how and if populations cross valleys. In this chapter, I present some of the related theory of fitness landscapes and demonstrate some preliminary data using the fitness landscapes from a theoretical NK landscape, a genetic programming Avida landscape, and an RNA folding landscape.

Before a meaningful discussion of fitness landscapes can take place, the term fitness must first be defined. Unsurprisingly, fitness can mean different things in different contexts, and to different people. I adopt the following definition from [Kauffman, 1993] for Chapter 3. In future chapters, I tie my fitness concept to the replication rate of organisms for the purpose of studying evolutionary dynamics.

> Before continuing, I must clarify what I mean by a fitness landscape. For an evolutionary biologist, "fitness" applies principally to an entire organism. It has

components of fecundity, fertility, and other factors, leading to reproductive success (Crow and Kimura 1965, 1970; Ewens 1973). These include complex issues such as the frequency of each genotype variant in a region, and even the entire ecosystem with which each organism interacts (S. A. Levin 1978). Therefore, in the general context, it is difficult to assign a fitness to a gene or even to a genotype, since all these factors depend upon the other organisms in the population.

For the purposes of the present chapter, I shall use the term "fitness landscape" in a much more restricted sense to refer to any well-defined property and its distribution across an ensemble. For example, the capacity of each protein in protein space to catalyze a specific reaction under specified conditions is, in principle, a well-specified property. The velocity of the reaction catalyzed by each protein can then be defined as the fitness of that protein. Then the distribution of velocities across the space of proteins constitutes the fitness landscape with respect to that defined function...

3.1 Fitness Landscapes

The concept of a fitness landscape was first elucidated by Wright in [Wright, 1931] and [Wright, 1932]. It refers to the idea that genetic adaptations may confer a fitness benefit (or penalty) in a reliable way, most often due to adjustment to the environment, and that the possible neighbors that a given population may explore are constrained by the current location in genotypic space that the population occupies. This limitation is for the simple reason that offspring are more likely to have mutations into neighboring regions of the landscape, at least for reasonably low mutation rates where evolution could be studied.

Fitness landscapes can be vast. As Wright points out in his seminal work, if there are 1000 genes, each with 10 possible alleles, there would be 10^{1000} possible combinations of genes. If each of these combinations solely determined fitness, the fitness landscape would

Figure 3.1: Figure 2 from [Wright, 1932] with original caption.

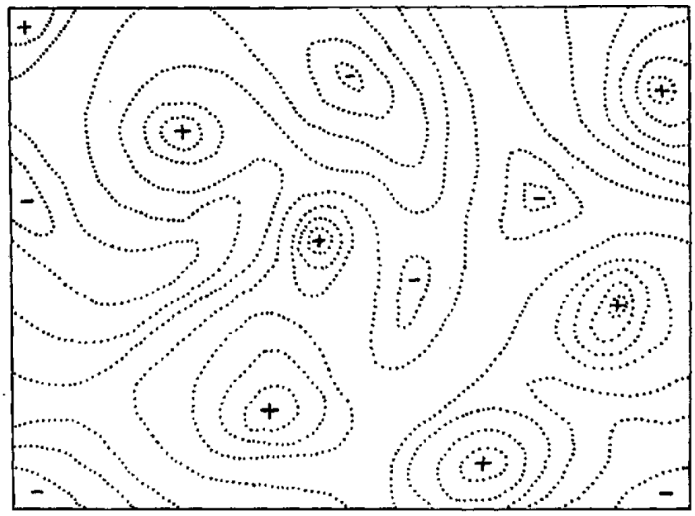


FIGURE 2.—Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.

then also consist of 10^{1000} mappings between genotypes and fitness. This number is well in excess of the 10^{80} estimated number of atoms in the universe. Estimates of the number of genes in the human genome further exceed this number; for instance [Roest et al., 2000] estimates the number at 28,000-34,000, although this exact number has been the focus of intense interest and revisions and more recently been estimated at 19,000 protein coding genes [Ezkurdia et al., 2014]. Regardless, the fitness landscape, even if it were purely determined by the genome, which it most assuredly is not, would be immense. Furthermore, even at very large population sizes, only a small number of possible genotypes can be sampled, when confronted with such vast landscapes. Wright visualized these combinations as representing a gene field in two dimensions. A 2D contour plot visualization of a fitness landscape from [Wright, 1932] is shown in Figure 3.1.

On the subject of the relationship between evolution and fitness landscapes, Wright wrote:

> ... With something like 10^{1000} possibilities, it may be taken as certain that there will be an enormous number of widely separated harmonious combinations. The chance that a random combination is as adaptive as those characteristic of the species may be as low as 10^{-100} and still leave room for 10^{800} separate peaks, each

Figure 3.2: Figure 4 from [Wright, 1932] with original caption. Here, 4NU refers to the genome wide mutation rate, whereas 4NS refers to the selective pressure. D and E are relevant for sexual populations and F corresponds to an island model. As Wright points out, differential selection on different islands in F may form the basis for or contribute to the process of speciation.

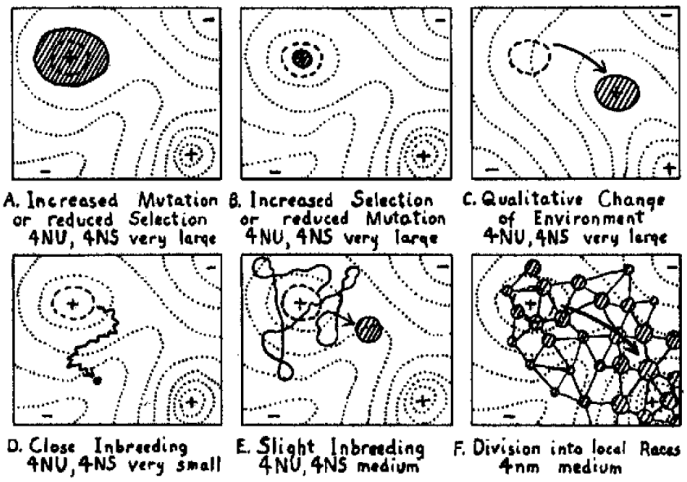


FIGURE 4.—Field of gene combinations occupied by a population within the general field of possible combinations. Type of history under specified conditions indicated by relation to initial field (heavy broken contour) and arrow.

> surrounded by 10^{100} more or less similar combinations. In a rugged field of this character, selection will easily carry the species to the nearest peak, but there may be innumerable other peaks which are higher but which are separated by "valleys". The problem of evolution as I see it is that of a mechanism by which the species may continually find its way from lower to higher peaks in such a field.

[Gavrilets, 2004] further distinguished between two conceptions of fitness landscape that often appear in the literature. The first is that the fitness landscape is determined by the fitness of gene combinations. This is the view described by Wright above as he explicitly defines fitness as determined by gene combinations. An alternative formulation that Gavrilets outlines and which often appears in the population genetics literature is that the fitness landscape is the relationship between gene frequencies and the mean fitness of a population. This formulation begins to suffer in multilocus systems where genes may have complex epistatic interactions that are not simply additive.

There are other conceptions of fitness landscapes that Gavrilets points out. For instance,

fitness landscapes may be visualized in terms of continuous quantitative traits. This can be useful in understanding the selection pressures on these traits, and is commonly used in quantitative biology to visualize selection on traits and investigate evolutionary tradeoffs.

Another area of broad interest in fitness landscapes is that of frequency dependence in which the fitness of an organism is not dependent mainly on the fitness lattice, but also on the distribution of other individuals within the population. Negative frequency dependent selection is often driven by predators or disease. An example would be a virus that becomes common and causes a host population to develop immunity–new variations to which the host was not immune would have higher fitness and thereby spread rapidly. Positive frequency dependent selection can be observed in the study of cooperators and cheaters, among other domains–a certain level of cooperation is required for cooperation to become a viable strategy.

For the purposes of this work, I focus on fitness landscapes as gene combinations. Epistasis plays an important role and is crucial in determining the landscape structure. However, I will ignore cross-organism interactions such as frequency dependence, which are much harder to represent in traditional landscapes, and move the fitness into an even higher dimension making it a function of both genotype and environment.

3.2 The Structure of Fitness Landscapes

The distribution of peaks in fitness landscapes is an open problem, with consequences including the likely distribution of viable protein sequences and the distribution of viable RNA/DNA sequences for viruses. Peak distribution is also important in the evolutionary computation space; the goal is to find the best solutions possible, which correspond to local optima or peaks, so improved understanding of evolutionary processes will also be useful.

Kauffman in [Kauffman, 1993] stated:

> ... Depending upon the distribution of the fitness values, the fitness landscape

> can be more or less mountainous. It may have many peaks of high fitness flanked by steep ridges and precipitous cliffs falling to profound valleys of very low fitness. Or it may be, like the gentle Normandy countryside, smoothly rolling with low hills and gentle valleys.

Kauffman introduced his NK model of fitness landscapes, which is explored in more detail in the next chapter. A significant finding from analyses of these landscapes was the realization that peaks were not distributed randomly in genotype space, but rather more closely packed than should be expected by chance. That is to say, there was global structure within the fitness landscape. Kauffman wrote:

> ... Like the alps, our landscape here possesses a kind of Massif Central, or high region, of genotype space where all the good optima are located...

Kaufmann analyzed the structure of the fitness landscape empirically by measuring random walks starting from random initial genotypes in the landscape. The correlation structure was measured by using a technique adapted from [Weinberger, 1991], which appears in equation 3.1. Here, f_t is the fitness at step t in the random walk, and f_{t+s} is the fitness s steps later.

$$R(t,s) = \frac{E(f_t \cdot f_{t+s}) - E(f_t) \cdot E(f_{t+s})}{\text{var } (f)} \tag{3.1}$$

As a result of this work, the claim that peaks tend to be clustered globally in the landscape is often referred to as the Massif Central Hypothesis. However, Kauffman describes what he calls a 'complexity catastrophe' that occurs in increasingly rugged landscapes. A depiction of the complexity catastrophe can be seen in Figure 3.3.

However, this view of fitness landscapes and their underlying structure is not undisputed. Gavrilets in [Gavrilets, 2004] had this to say at the beginning of his chapter on nearly neutral networks:

Figure 3.3: Figure 3.1 from [Kauffman, 1993]

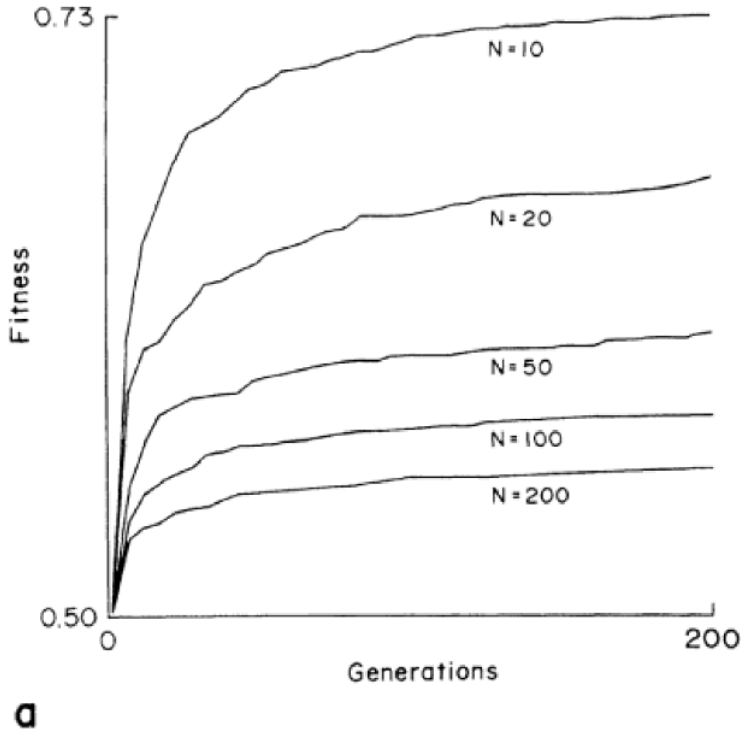


Figure 3.1 (*a*) The complexity catastrophe seen in long-jump adaptation on correlated $K = 2$ *NK* landscapes. "Generation" is the cumulative number of independent long-jump trials. Each curve is the mean of 100 walks. Fitness falls as N increases.

... Indeed, everybody knows from his or her own hiking experience that it is impossible to get from the top of one hill to another without having to descend to a kind of valley or depression between them. Our intuition tells us that things will stay the same in landscapes with many more dimensions than the three we are so well familiar with, and that extended ridges or chains of very shallow valleys connecting high-fitness genotypes are improbable. If the fitness landscape is constant (in space and time), then stochastic fluctuations in the population genetic structure must be a major mechanism for crossing the valleys. However, the main conclusion of the previous chapter is that there is no satisfactory solution to the problem of stochastic transitions across even moderately deep valleys in static landscapes. Only if the valleys are very shallow, can stochastic transitions happen on a time scale short enough to be of biological significance. But both speciation

> and diversification seem to require crossing deep valleys in static landscapes describing RI. A logical conclusion is that speciation and diversification on rugged landscapes are impossible. There must be some kind of an error or a weakness in the chain of arguments presented in the previous paragraph...

Gavrilets presented an alternative view of the structure between peaks; namely that peaks are connected by 'ridges' that are composed of neutral and nearly neutral genotypes. He called these 'holey landscapes'. His arguments for this view of fitness landscapes are motivated by percolation arguments; namely that at a certain threshold of connectedness, there is a tendency for all elements to be connected. A visual example of this from [Gavrilets, 2003] can be seen in Figure 3.4. He carries this same basic core argument into higher dimensions as justification for the idea of neutral networks.

Among the evidence [Gavrilets, 2004] cited for the existence of neutral networks is the NK family of models, although he relies primarily on variants like the NKp model by [Barnett, 1998] for formulating his arguments.

3.3 Evolutionary Implications of Fitness Landscapes

So far, two models of fitness landscapes have presented quite differing views on how fitness landscapes may be arranged. These have strong biological implications that affect how populations might be able to evolve on them.

In Kauffman's conception of rugged fitness landscapes, the complexity catastrophe implies that as an adaptive walk progresses, further improvements become harder and harder to obtain; in fact these innovations become exponentially harder to cross as fitness peaks are achieved since larger and larger valleys need to be crossed. In the competing view outlined by Gavrilets, continuous improvement is permitted via these incredibly long neutral networks, making it unnecessary for the process of evolution to cross deep valleys.

Figure 3.4: Figure 4 from [Gavrilets, 2003] with original caption.

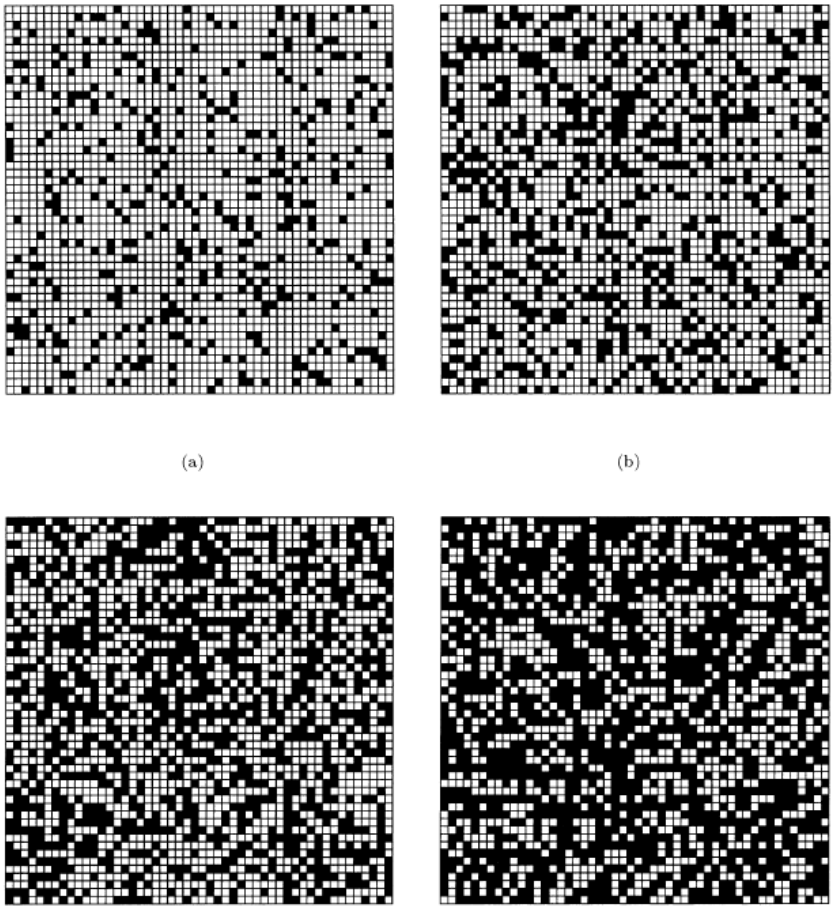


FIG. 4. Formation of clusters of viable genotypes (painted in black) in two dimensions for different values of the probability of being viable p. Inviable genotypes are painted in white. (a) $P = 0.15$, (b) $P = 0.30$, (c) $P = 0.45$, (d) $P = 0.60$.

There has been some work on evaluating the structure of fitness landscapes in a biological context. For instance, [Fontana et al., 1993] examines correlation and local structure in an RNA-folding landscape where the objective is to minimize free energy on a variety of alphabets and found that RNA landscapes correspond roughly to a K of 8 in the NK landscape, and also that there exist neutral networks in the vicinity of peaks, in contrast to NK landscapes. [Stadler et al., 2001] investigated the topology and evolutionary implications thereof in RNA landscapes.

[Kryazhimskiy et al., 2009] classifies fitness landscapes by looking at the rate of adaptation along fitness trajectories and tying this to real microbial data from long term evolutionary studies. There have also been limited fitness landscape mappings, such as the analysis done by [St Onge et al., 2007] and [Poelwijk et al., 2007] which have helped to provide real data on the nature of fitness landscapes for gene clusters and enzymes, respectively.

3.4 Model Systems

For my evaluation of fitness landscapes, I examine three different model systems in order to study structural properties of the landscapes. I perform exhaustive genotype-to-phenotype mappings for each of these environments. Each landscape examined was of length 18 with four possible alleles per site, constituting a fitness landscape with $4^{18} = 68,719,476,736$ genotypes. To my knowledge, no systematic and exhaustive analysis of landscapes this large has been conducted prior to this work. I choose a common length and number of alleles to eliminate confounding factors on evolution from structure and make the landscapes more easily comparable.

The three model systems are NK, which is a theoretical family of landscapes commonly used in the discussion of fitness landscapes, Avida, which provides us with a genetic programming landscape where genotypes map on to functional computer programs, and an RNA landscape where the fitness is of a genotype is determined by its free energy after folding. This

provides us with a theoretical landscape widely used to study evolution and optimization, a genetic programming language, and a biologically-derived landscape. The purpose of this work is to study some of the structural properties of these landscapes. It is also important to note that these are instantiations of each type of landscape: there are many variants of each, and it may not be possible to generalize. Nevertheless, they provide a useful starting point for the systematic investigation of landscapes, and I introduce them in more detail below.

3.4.1 NK

The NK fitness landscape, originally introduced in [Kauffman and Levin, 1987], is a biologically-inspired model of a fitness landscape, in which an organism's fitness is determined by N genes, each with K other interacting loci. This model has been used to study properties of adaptive walks and the influence of epistasis on rugged landscapes. This K parameter tunes the amount of epistasis in the landscape; as K increases, the amount of epistasis increases as does the number of local optima.

At K=0, there is a single peak in the landscape, and it is a trivial problem for evolution to solve since each site may be optimized independently. At K=N-1, the landscape is completely uncorrelated and any mutation effectively randomizes the organism's fitness. As in the original NK model, I calculate fitness by using the arithmetic mean of the fitness contributions from each of the N sites as in Equation 3.2. Furthermore, I also use the default nearest adjacency model of the NK landscape, where each gene is epistatic with K adjacent genes.

$$\begin{aligned} F(G) &= \frac{1}{N}\sum_{i}^{N} f(G_i) \\ \text{where } f(G_i) &= f(g_{i,0}, ..., g_{i,K}) \sim U(0,1) \end{aligned} \tag{3.2}$$

Here, $g_{i,j}$ denotes the j-th neighbor of site i.

For consistency with other landscapes studied, I use four alleles per site and I set the parameter of K = 4 to achieve a nontrivial amount of epistasis.

3.4.2 Avida

Avida is a software system that has been used to study biological evolution. [Ofria, 2004] Each "digital organism" in Avida is a program that is executed in a virtual computer, complete with its own registers and stacks. The programs may, among other things, execute logic tasks, for which they receive a reward that translates to an execution speed boost relative to their competitors. The relative execution speed is called 'merit' and corresponds to a metabolic rate. Fitness is calculated as merit divided by the gestation time. Fitter organisms are those that are able to reproduce themselves fastest, which equates to selective pressures to increase execution speed or improve the efficiency of replication.

Although no exhaustive mapping of this size has been done to study the fitness landscape of the Avida genome, there has been considerable previous work that informs about the nature of Avida fitness landscapes. [Lenski et al., 1999] sampled the landscape up to ten mutations away from simple and complex digital organisms and found that not only are complex organisms more robust to the generally deleterious effects of point mutations, but that there is significant epistasis in the mutational landscape as measured by comparing the fitness of mutational combinations with single point mutational fitness effects. There were also differences in the fragility of organisms and the fitness effects of mutations when comparing simple and complex organisms. [Lenski et al., 2003] suggested that rewarding intermediate logic functions is important to the evolution of EQU (the bitwise equals logic operator) with the implication that complex features often coopt simpler features in the process of evolution. This result is partly because higher-order logic tasks can be composed of lower order ones.

The Avida landscape I study is a reduced version of the default instruction set with only four instructions: IO, nand, nop-C, and swap. Organisms automatically reproduce at the end,

so all genotypes are technically viable. This model is a highly simplified compared to the full Avida model in many ways. The default Avida instruction set has 26 instructions and organisms must deal with the complexity of maintaining a copy loop in order to reproduce (the organism must copy over each instruction). This requirement of maintaining the copy loop has several consequences for the resulting fitness landscape. Primarily, the vast majority of random sequences are dead since they cannot produce viable offspring, whereas in the reduced instruction set, this is not the case. However, even in the reduced instruction set, the landscape is similar in that the vast majority of random sequences still are on the lowest fitness levels. Furthermore, even small genome sizes are difficult to exhaustively analyze with 26 possible instructions per site. A length 18 organism, which would still be a small organism by Avida standards, would have 26^{18} possible genotypes, which is roughly a landscape of size $2 \cdot 10^{25}$. Clearly, a landscape this immense would defy exhaustive analysis. And this is still a small genome at a much reduced scale compared to biological entities or proteins which often have sequence lengths well above the hundreds. Exhaustive enumeration of such entities may never be completely possible given the sheer exponential (with length) scale of the size of the fitness landscape. Indeed, the fitness landscapes that I am considering here with 68 billion states (four possible instructions at each of 18 sites) is already pushing the current limits of hardware in terms of both storage and computing power.

In this work, I use Avida not as the experimental evolution platform that it undoubtedly is, but specifically to evaluate the fitness of an input genomic string by executing the string and determining what logic tasks, if any, it can perform and what its gestation time would be. This genotype-to-phenotype mapping is what provides us with the fitness landscape. For the purposes of this work, I will be examining the repeated tasks environment, in which organisms may be rewarded for performing the same tasks multiple times.

Avida uses a concept of 'merit' to determine relative CPU speed. Merit is awarded by organisms successfully performing tasks; I use the default Avida logic-9 environment. In this environment, nine logic tasks are rewarded exponentially in order of rough complexity.

Specifically, NAND and NOT are rewarded by a factor of 2, AND and ORN (or-not) by a factor of 4, OR and ANDN (and-not) by a factor of 8, NOR and XOR by a factor of 16, and EQU (XNOR) by a factor of 32.

Organisms also have variable gestation times, since nop-C can be a modifying argument to an instruction. If this occurs, the nop-C itself is not executed by Avida and thus does not count towards the total execution length.

3.4.3 RNA

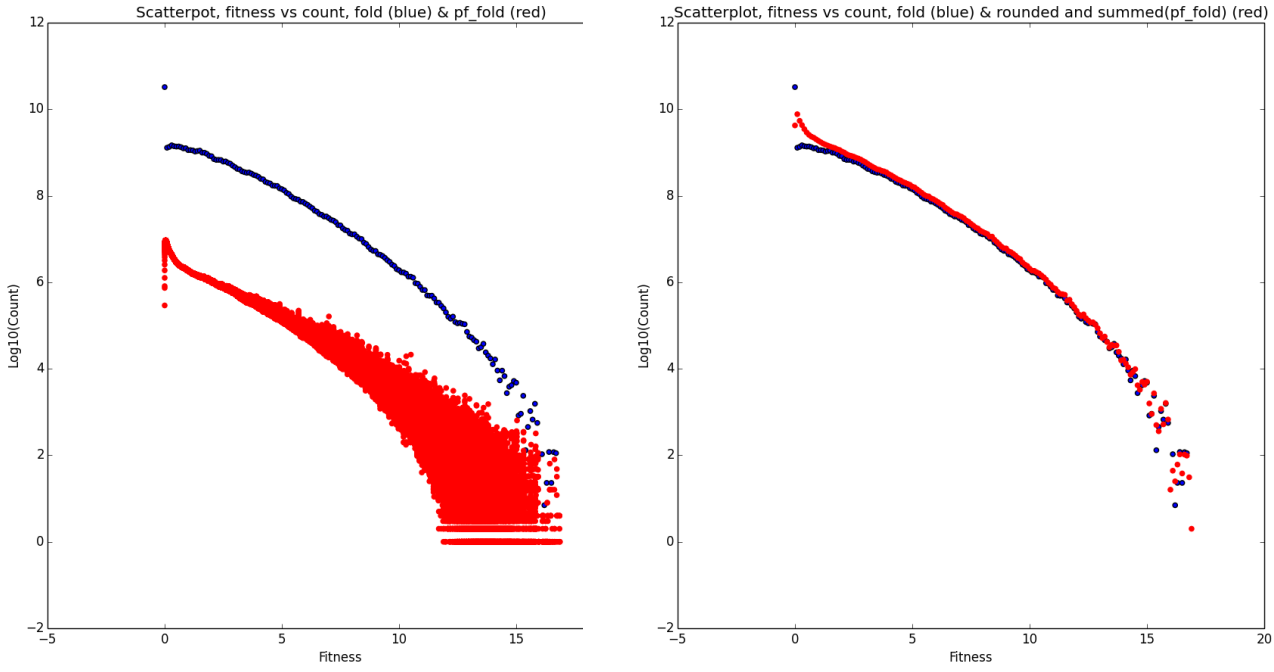
The third and final landscape I examine is the result of RNA folding of length 18 sequences using the Vienna RNA Package created by [Lorenz et al., 2011]. I examine two different RNA fitness landscapes generated by two functions: **fold**, which computes the minimum free energy of an RNA sequence, and **pf_fold**, which returns the Gibbs free energy of the folding ensemble.

Figure 3.5 shows a comparison of the two landscapes. Since pf_fold is based on the energy of the ensemble of possible folds, there is more variation in the fitness values than fold which only looks at the most likely candidate. Nonetheless, the discretization in Figure 3.5 implies that the landscapes are broadly similar. For the rest of this work, I use the results of pf_fold to denote the RNA landscape. This also has the side effect of decreasing the amount of neutrality in the landscape, as well as reducing the amount of extremely low fitness genomes while spreading the mass along the rest of the landscape, compared to the fold landscape. Previous work has found neutrality in RNA landscapes, but in this instantiation there is little–likely because of the decision to use the ensemble rather than the most likely sequence when calculating free energy. [Fontana et al., 1993]

Figure 3.5: Comparison of RNA pf_fold and fold landscapes. In Figure 3.5a, raw data for each landscape is plotted, with each fitness, count pair plotted separately. In Figure 3.5b, I discretize fitnesses of pf_fold by rounding fitness to the nearest tenth. The landscapes are highly similar as would be expected since they are both mappings of fold energy within the RNA landscape.

(a) Raw data

(b) Discretizing pf_fold



3.5 Landscape Summaries

3.5.1 Methods

In order to understand the structural properties of each landscape, I started by examining the distribution of fitnesses. With length 18 and four alleles per site, this works out to almost 69 billion possible genotypes. For each of the genotypes in each of the three landscapes, I look at the corresponding phenotype, which is tied simply to fitness in each landscape: an instantiation of an NK landscape with K=4, logical tasks performed in Avida, and the free energy of the ensemble of folding sequences in the RNA landscape.

As part of each landscape summary, I also examine properties of peaks within the landscape. Here, I present two definitions, that of a 'true peak' and a 'neutral peak'. A 'true peak' is a genotype with a fitness that is strictly greater than the fitness of each of its neighbors. This is a local optimum on the fitness landscape. I also define a 'neutral peak' as a genotype with an associated fitness that is greater than *or equal* to the fitness of each of its neighbors. Neutral peaks can be on plateaus or ridges in genotype space. Counterintuitive behavior can occur when the edge of a ridge or plateau adjoins a higher peak—the edge point would be considered neither a peak nor a true peak since there would be at least one genotype with higher fitness next to it, although it may neighbor other points with identical fitness that are still considered neutral peaks. The neutral peak dataset is a superset of the true peak dataset. This concept proves useful when looking at landscapes with large amounts of neutrality, such as the Avida landscape, since the highest points are rarely true peaks.

I exhaustively looked at each genotype to obtain the data in this chapter. I accomplished this by splitting the genotype space into several (usually 1024 or 4096) smaller chunks. Each chunk was then evaluated, and summary information generated for the chunk, as well as candidate 'true peaks' and 'neutral peaks'. These individual chunks were later combined to provide us with a list of all the peaks in as well as complete summary information for the

landscape. I used several techniques including binary encoding, hashing, and compression in order to deal with the dataset sizes.

3.5.2 NK

The NK landscape pictured in Figure 3.6 features a normal distribution, centered at a fitness of 9. This distribution is the consequence of the fact that each of the eighteen genes has a uniform distribution with an expectation of 0.5, and the result of this summation creates a normal distribution. The theoretical interval that fitnesses can take is [0,18], but I do not see values at either extreme, again due to the normal nature of the underlying distribution. The peaks also appear to follow a normal distribution, centered a little higher than 12 for this particular landscape instantiation. There is some, but very little, neutrality, and therefore few neutral peaks that were not peaks in this landscape. The neutrality itself is caused by the fact that I have only 32 bits of precision. Figure 3.7 shows the distribution of fitnesses for all genotypes and peaks in the landscape.

Figure 3.6: Phenotypic Summary, NK Landscape. X-axis is fitness, Y-axis is the count of the number of genotypes with that fitness. The neutral peaks line in green is overlaid by the peaks in blue since because there is so little neutrality in the landscape, the two sets are essentially identical.

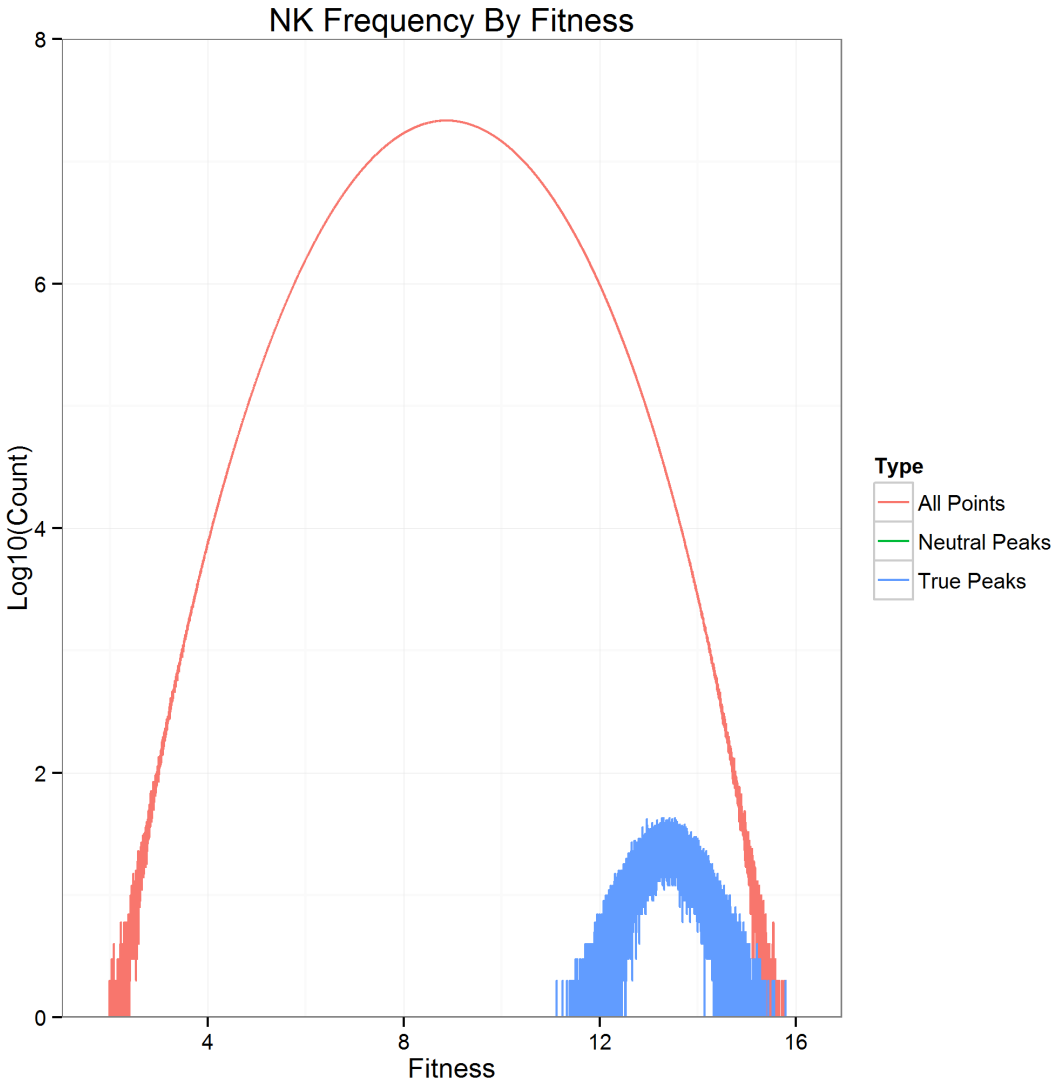
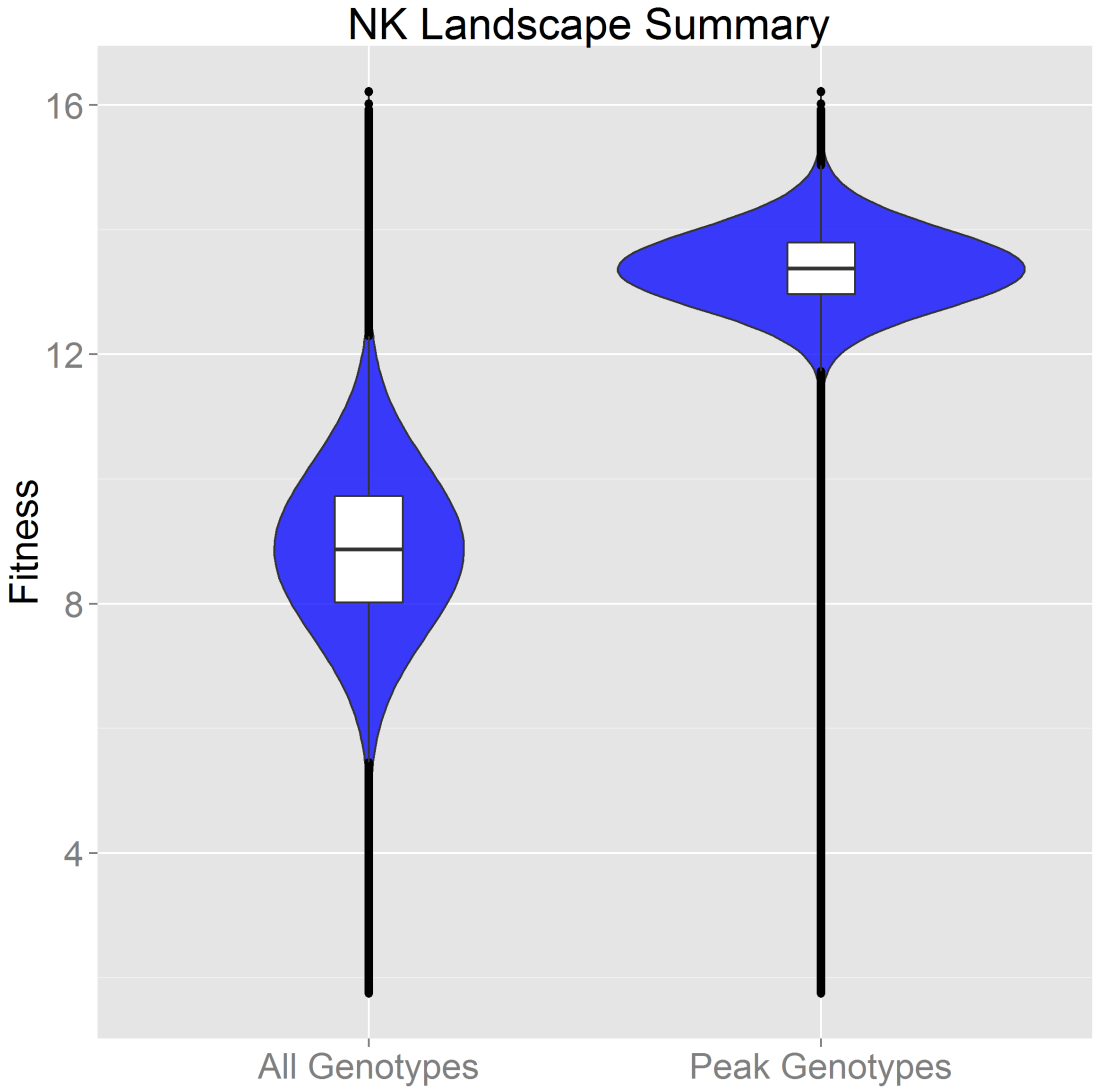


Figure 3.7: Box & violin plots of genotype fitness values on the NK landscape for all genotypes, peak genotypes.



3.5.3 Avida

Figure 3.8 is the summary of the reduced-instruction Avida landscape. The fitness structure of this landscape is exponential, so I take the logarithm of fitness when interacting with the Avida landscapes. There are only 1108 possible fitness levels in the Avida landscape, which is significantly fewer than the numbers observed in the RNA and NK landscapes which are more or less continuous. Figure 3.9 shows a summary of the Avida landscape. Another pattern that can be clearly seen is a wave structure. This appears to be connected to the structure of the landscape itself—in Figure 3.10, the phenotypic structure of the landscape is visible. Notable is the gestation structure—different amounts of optimization with the nop-C create a step-like structure. There are 98,120 true peaks in total and 86,100,995 peaks in the Avida data. I thresholded the peak data set by fitness to make it more tractable for the analyses used in the rest of the chapter, which resulted in 4.5 million peaks.

Figure 3.8: Phenotypic Summary, Avida Landscape. X-axis is fitness, Y-axis is the count of the number of genotypes with that fitness.

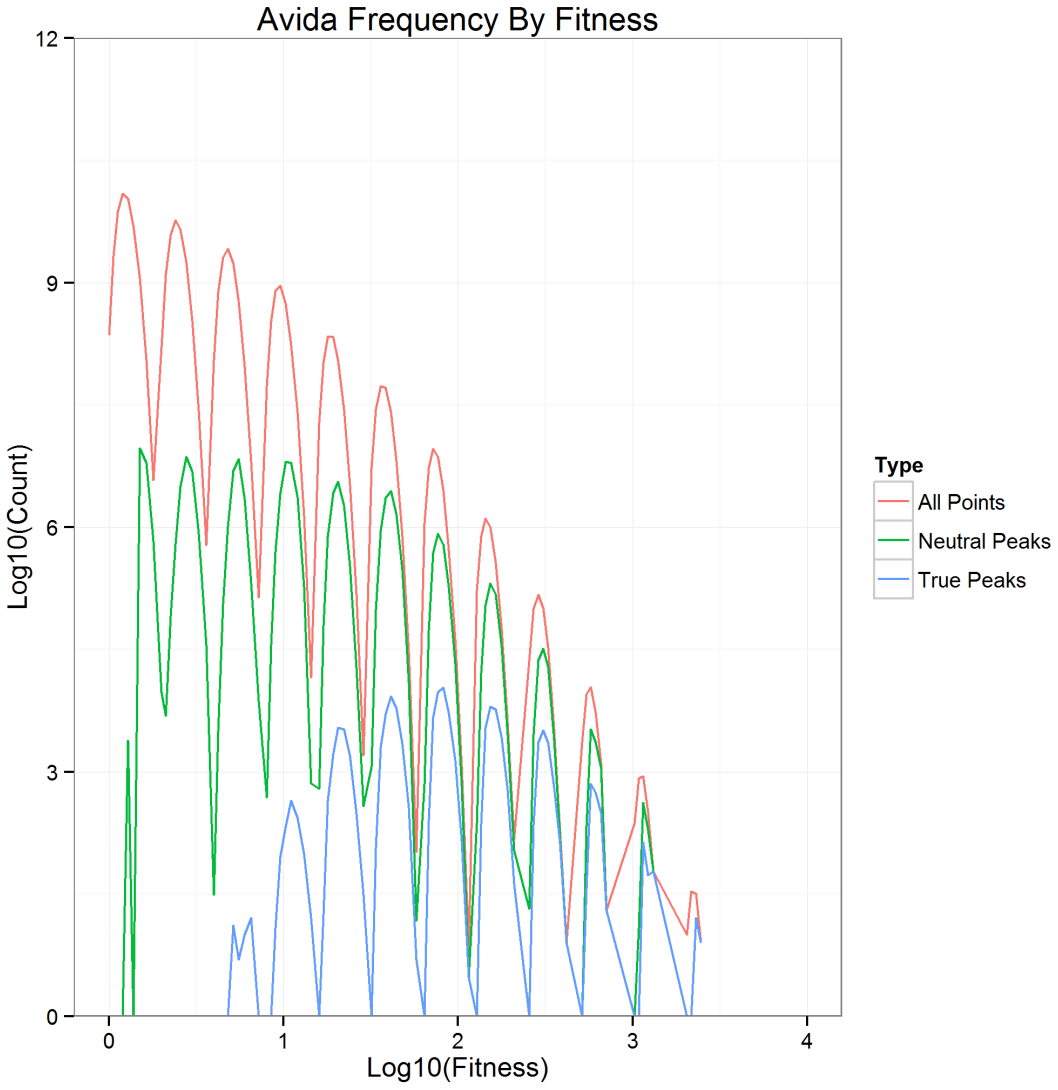


Figure 3.9: Box & violin plots of genotype fitness values on the Avida landscape for all genotypes, peak genotypes.

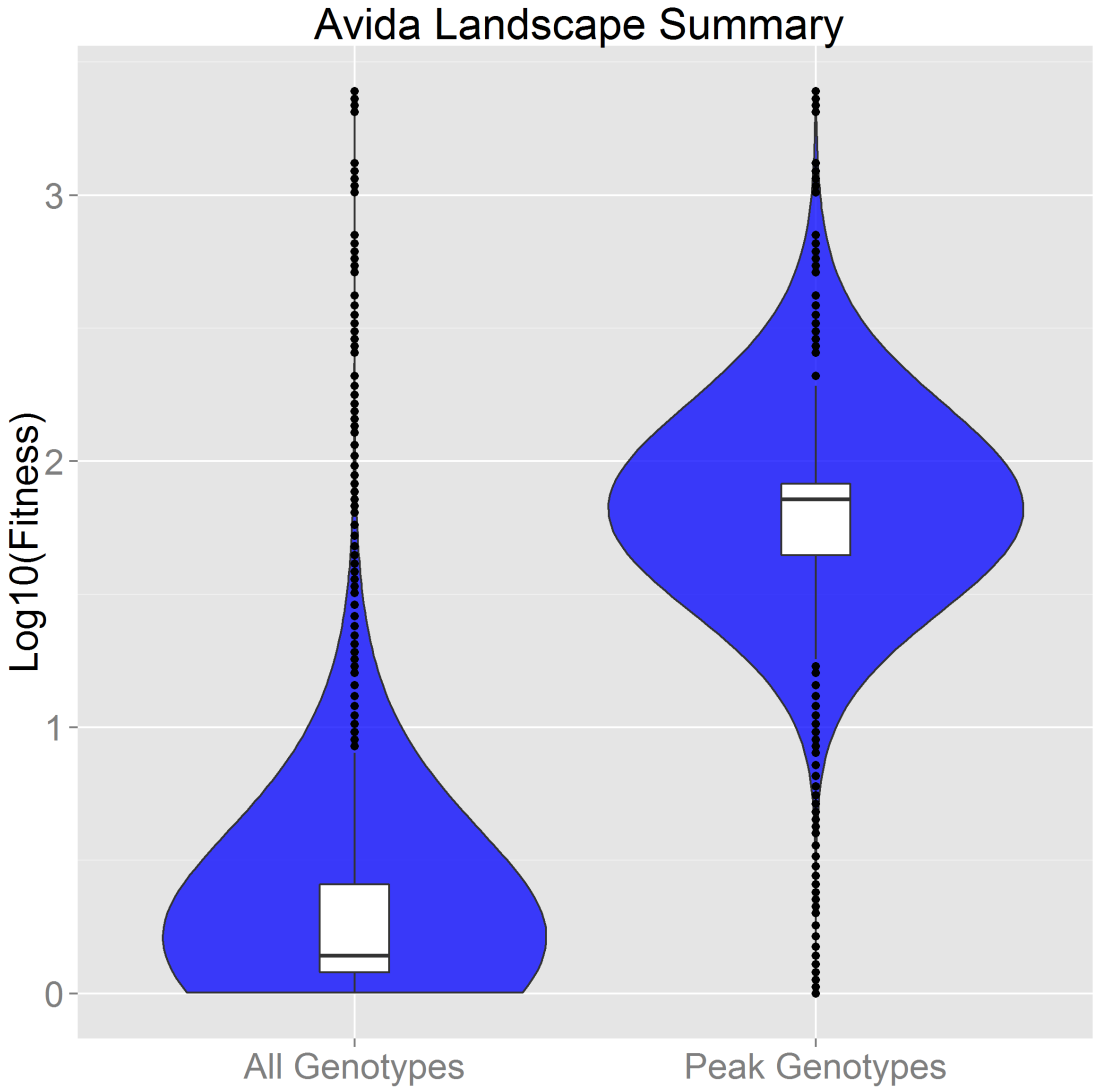
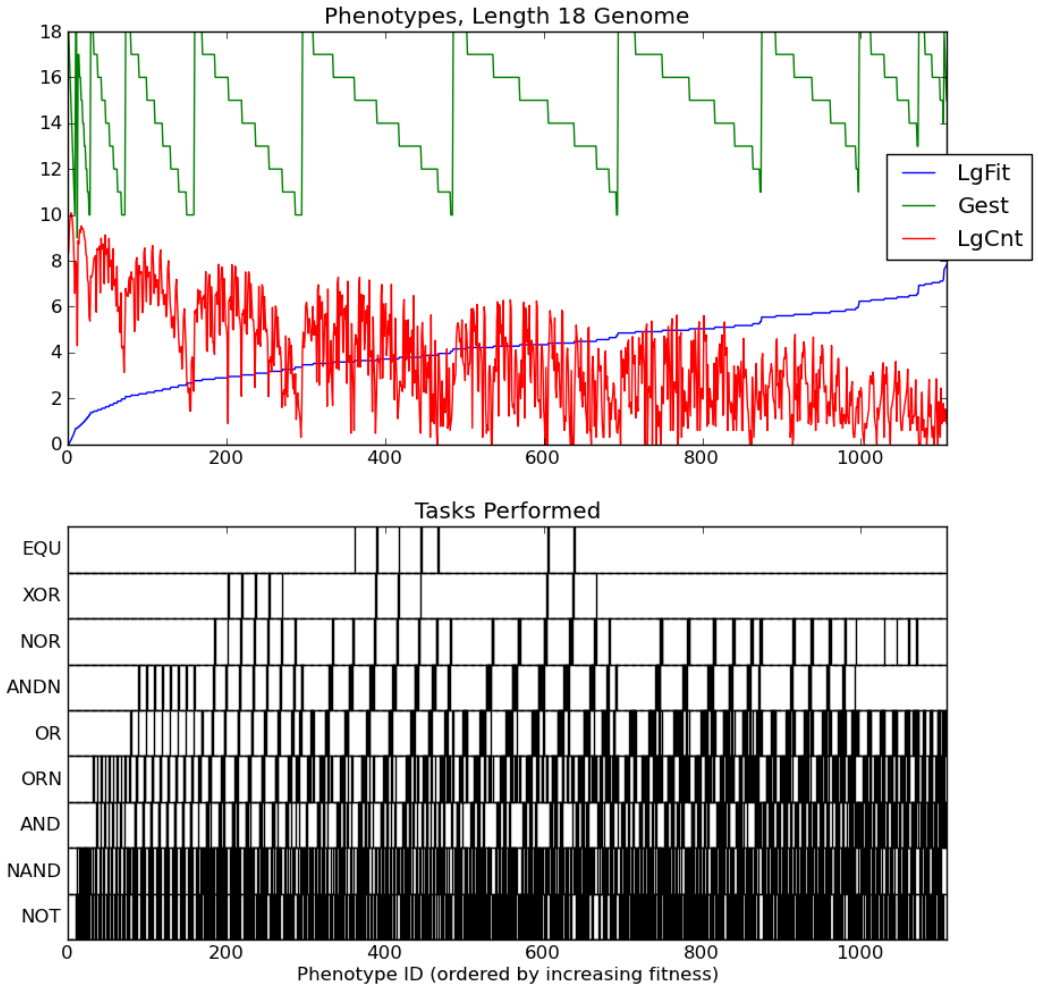


Figure 3.10: Phenotype of fitness values on the length 18, reduced instruction set landscape. Phenotypes are ordered by fitness. The blue line is log fitness of the phenotype. The green line is the gestation time. The red line is the number of genotypes. The general pattern observable is that for each level of merit, there are several optimizations using nop-C that reduce the execution time and thereby the gestation time, improving fitness. On the bottom, logic tasks performed by each ranked genotype are shown in the following order from bottom to top: NAND, NOT, AND, ORN, OR, ANDN, NOR, XOR, EQU. It can be seen that the most fit organisms perform several tasks rather than doing EQU.



3.5.4 RNA

The RNA landscape can be seen at a high level in Figures 3.11 and 3.12. As in the Avida landscape, there is a large preponderance of genotypes that do not fold particularly well. Also, similarly to the NK landscape, the fitness values range from 0 to 18. There are 184,020 neutral peaks and 58,024 peaks in this landscape, which indicates some clustering of the high regions in genotypic space, which will be explored further in following sections.

Figure 3.11: Phenotypic Summary, RNA Landscape. X-axis is fitness, Y-axis is the count of the number of genotypes with that fitness.

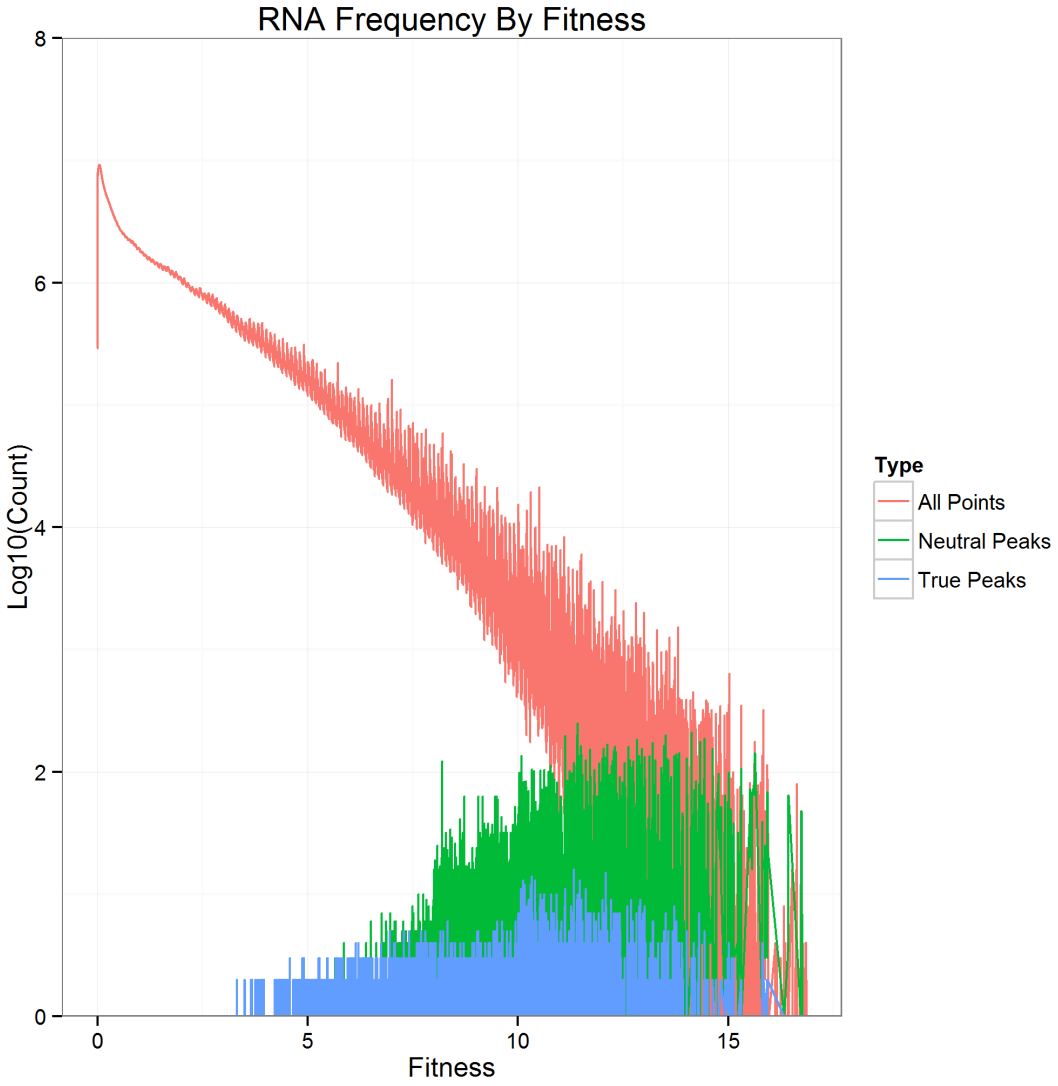
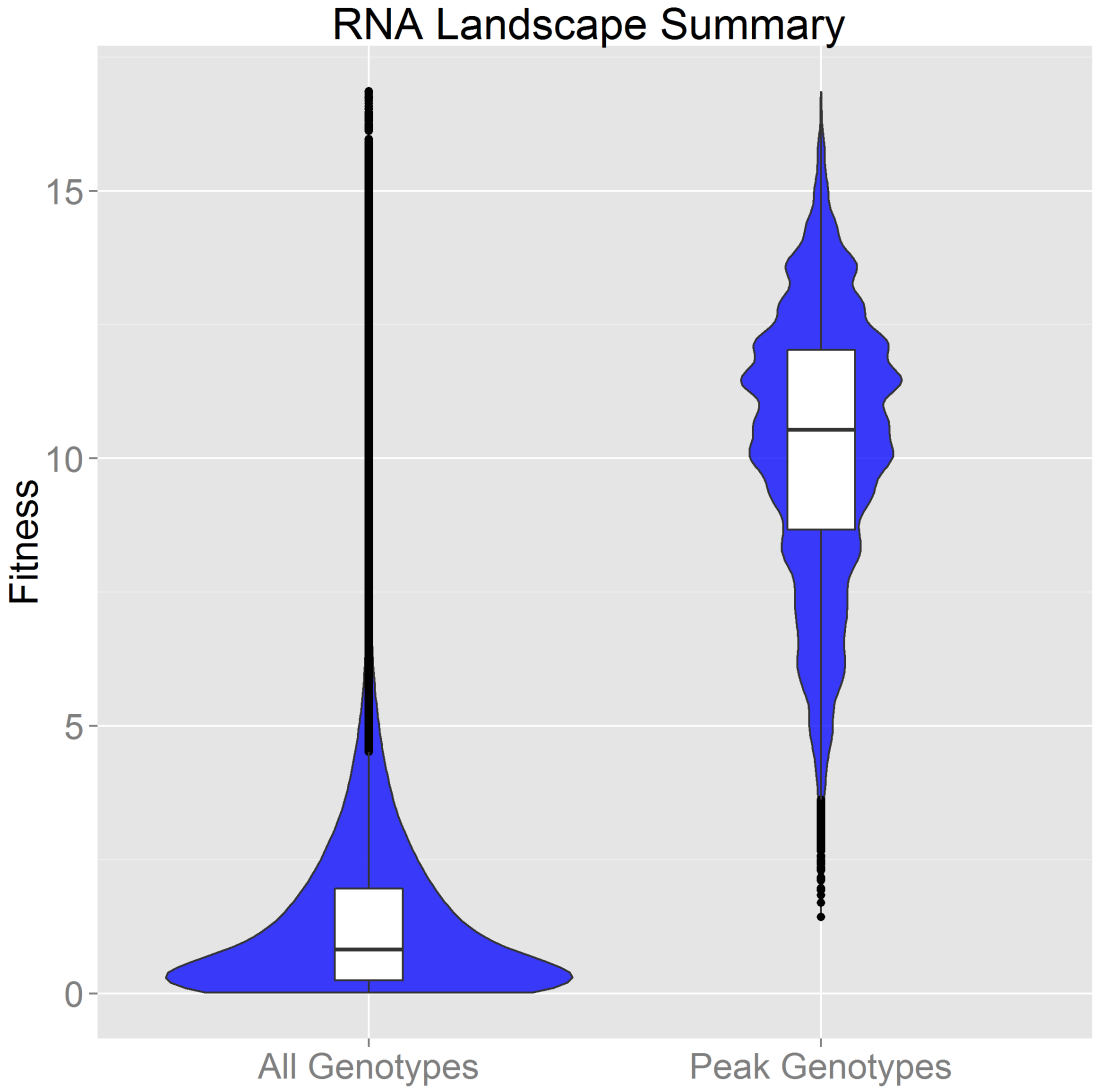


Figure 3.12: Box & violin plots of genotype fitness values on the RNA landscape for all genotypes, peak genotypes.



3.6 Peaks and Nearby Peaks

An analysis by [Østman et al., 2010] looked at the critical properties of true peaks in the NK landscape. One of the properties they measured was true peak fitness vs the average fitness of true peaks at a distance of two, which they call 'cluster mean fitness'. The key question here is whether more highly fit peaks correlate with more highly fit neighborhoods. [Østman et al., 2010] used a binary NK landscape for length 20 genomes and different Ks. They also performed a random assignment between genotypes and fitnesses as a control to ensure the relationship was not spurious. They found a correlation between higher fitness of a peak and higher cluster mean fitness as seen in Figure 3.13.

I perform a similar analysis here in each of the three model landscapes. I use the neutral peaks dataset to perform this analysis. One distinction from the methodology in [Østman et al., 2010] is that since my data set includes 'neutral peaks', I must count the neighborhood as being within a distance of two or less to include neighboring 'neutral peaks'.

Figure 3.13: Figure 4 from [Østman et al., 2010] with original caption

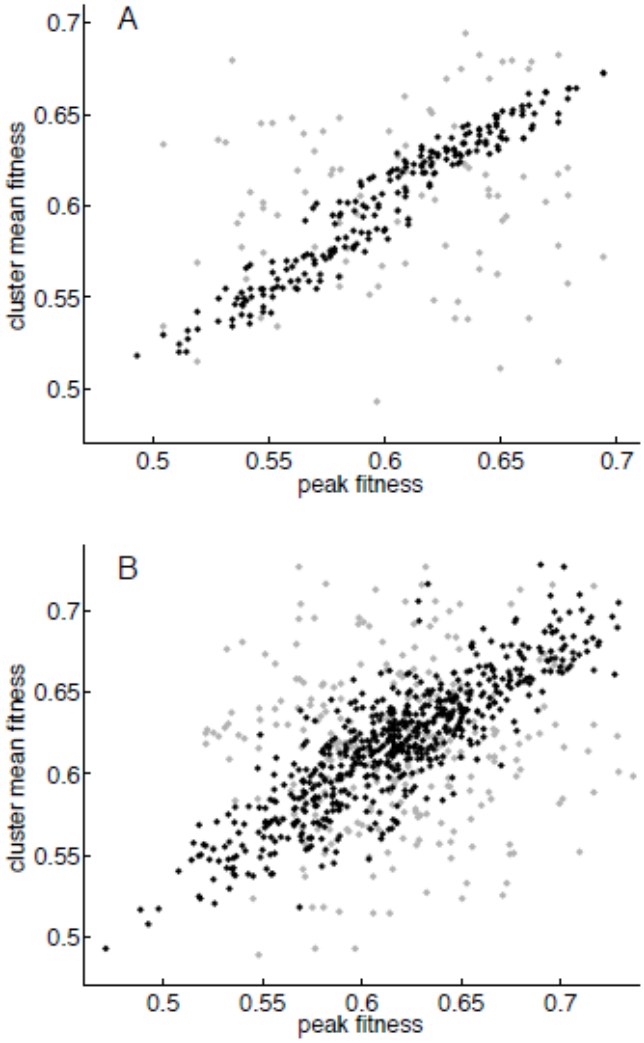


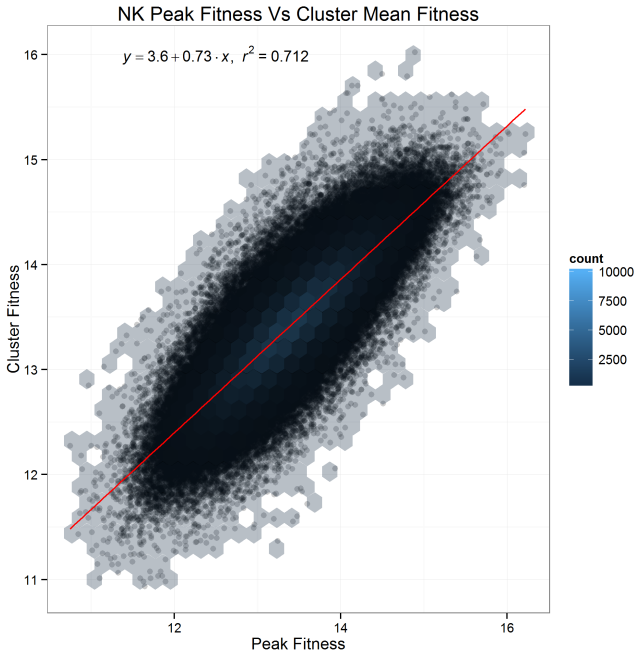
Figure 4: Mean fitness of peaks in circular clusters of radius $d = 2$ as a function of the fitness of the peak in the center of the cluster. (A) One landscape of $K = 2$ with 166 peaks (black dots). All landscapes show a strong correlation between cluster mean fitness and peak fitness, while the same analysis of assigning random genotypes to the peaks (but keeping the fitness) shows no such correlation (gray dots). The random data are from ten samplings. (B) One landscape of $K = 4$ with 679 peaks (black dots), and random genotypes (gray dots) obtained by sampling four times.

3.6.1 NK

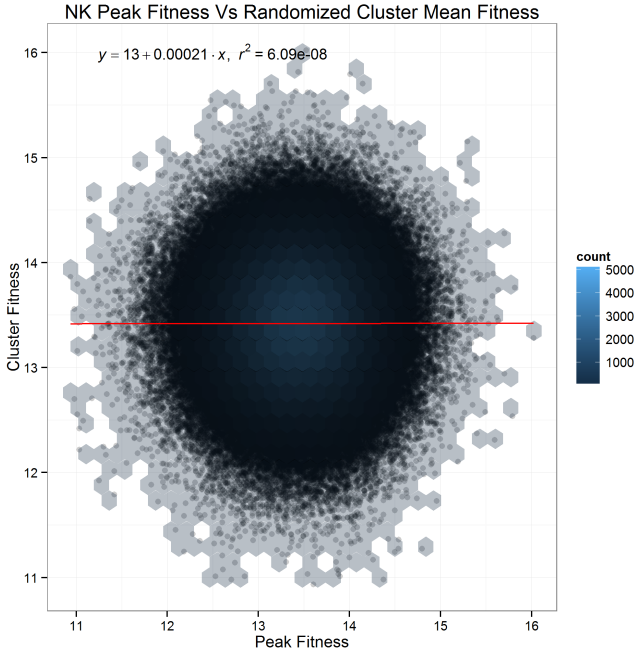
My NK landscape is qualitatively different from that in [Østman et al., 2010], it has length 18, but four alleles per site, rather than two alleles per site. The results can be seen in Figure 3.14 and show a strongly correlated peak fitness with cluster mean fitness.

Figure 3.14: NK data, Fitness vs Cluster mean fitness. Each neutral peak fitness along with its neighborhood fitness is shown as a dot. The colored hexagons overlay show density (count of points) and the red line shows the fit. The correlation and line fit can be seen at the top of the graph.

(a) Fitness vs Cluster mean fitness



(b) Control, with fitnesses randomly assigned to genotypes.

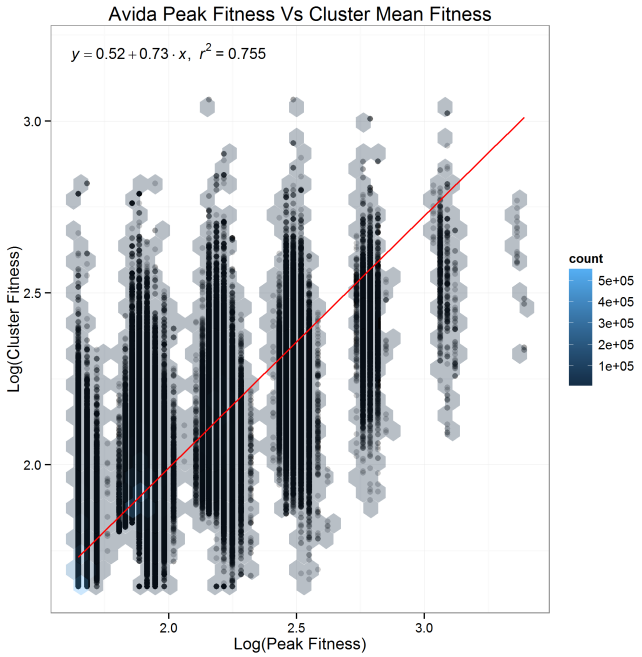


3.6.2 Avida

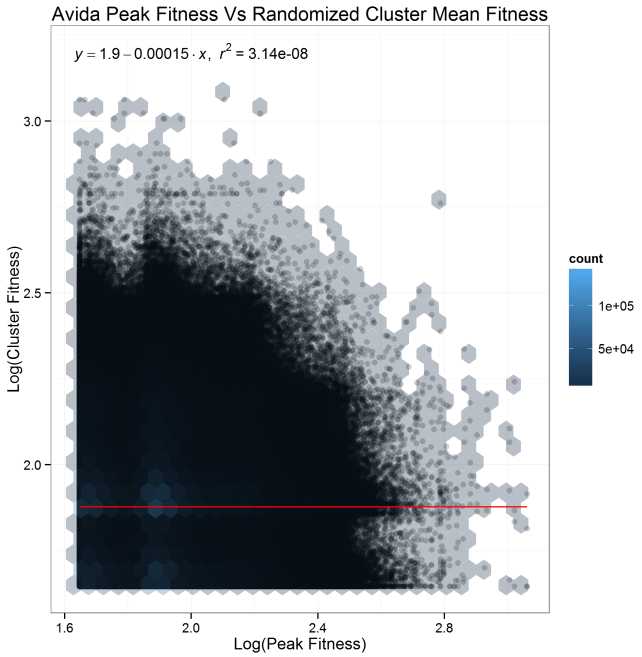
The results of the analysis on the Avida data can be found in Figure 3.15. Although a bit more visually confusing than the NK landscape due to the tiered structure of the Avida landscape, there is nonetheless, a strongly correlated positive relationship between the fitness of a peak and the height of any peaks in the neighborhood, and this relationship disappears completely when the genotypes are randomly assigned fitness.

Figure 3.15: Avida data, Fitness vs Cluster mean fitness. Each neutral peak fitness along with its neighborhood fitness is shown as a dot. The colored hexagons overlay show density (count of points) and the red line shows the fit. The correlation and line fit can be seen at the top of the graph.

(a) Fitness vs Cluster mean fitness



(b) Control, with fitnesses randomly assigned to genotypes.

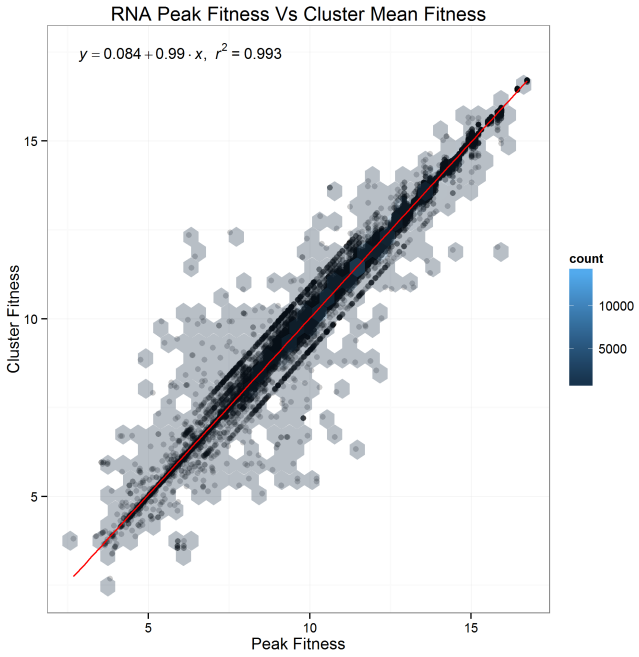


3.6.3 RNA

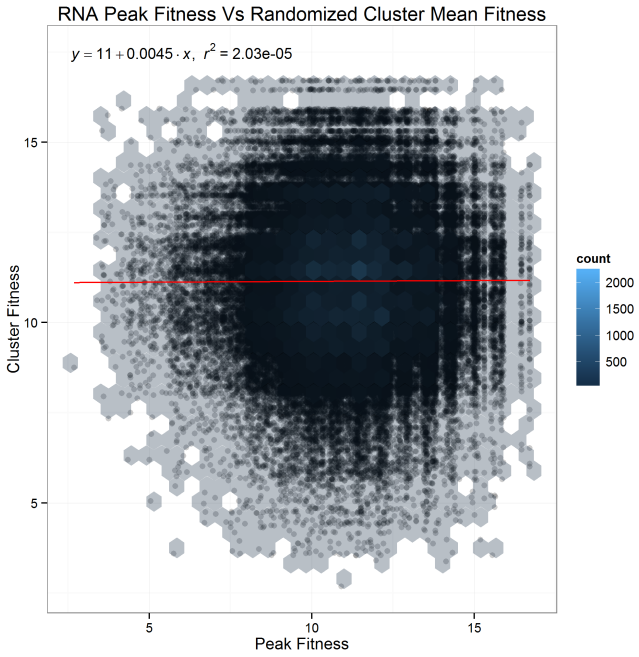
The results of the analysis on the RNA data can be found in Figure 3.16 and the same relationship holds; there is a similar positive and strong relationship between peak fitness and the fitness of peaks in their neighborhood.

Figure 3.16: RNA data, Fitness vs Cluster mean fitness. Each neutral peak fitness along with its neighborhood fitness is shown as a dot. The colored hexagons overlay show density (count of points) and the red line shows the fit. The correlation and line fit can be seen at the top of the graph.

(a) Fitness vs Cluster mean fitness



(b) Control, with fitnesses randomly assigned to genotypes.



3.7 Percolation

Another structural property of interest is percolation. Percolation theory is often used to model fluid flow and is particularly concerned with the connectedness of a graph. This is the key property that Gavrilets postulated might allow evolution to perform walks across a fitness landscape as in 3.4. The core argument is that as you add dimensions, the probability of there being an interconnected pathway may increase. If peaks are close to each other, even by small jumps, then it should be easier for evolution to walk from peak to peak, compared with the alternative where peaks might be interspersed by wide valleys.

Again, [Østman et al., 2010] explored percolation in the NK landscape, which can be seen in Figure 3.17. The basic approach is to add a fitness threshold on peaks and examine the spatial relationships of the remaining genomes. Peaks of distance no more than two apart are connected and belong to the same cluster. The same cluster may contain peaks of more than 2 distance apart due to a chain of connected peaks.

For this work, I follow a similar regime. I measure clusters and connectedness for increasing fitness thresholds for peaks. Here, I measure the size of the largest connected cluster. For the RNA and NK landscapes, which are continuous, I set threshold increments of 0.1. For Avida, since there are a finite number of phenotypes, I used the possible fitness levels as thresholds.

Figure 3.17: Figure 6 from [Østman et al., 2010] with original caption

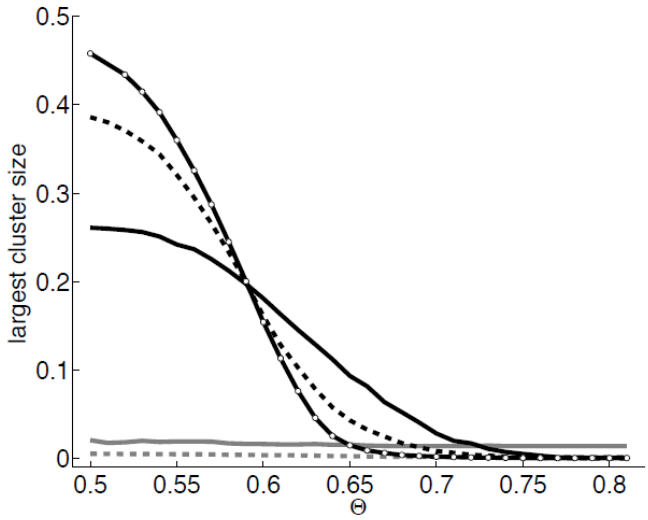


Figure 6: Size of the largest network cluster in the landscape averaged over 50 landscapes for each K as a function of fitness threshold, Θ. $K = 2$ (solid black line), $K = 4$ (dashed black line), and $K = 6$ (solid black line with white circles). The more rugged the landscapes are, the more abrupt the transition is from small network clusters to one cluster dominating the landscape. Random genotypes for $K = 2$ (solid gray line) and $K = 4$ (dashed gray line) show no increase in cluster size.

3.7.1 NK

The analysis for the NK data set can be seen in Figure 3.18. As in the [Østman et al., 2010] work, there appears to be an orderly phase transition from where most of the peaks are connected in the largest cluster, to where the largest cluster is significantly smaller.

3.7.2 Avida

The percolation graph for Avida can be seen in Figure 3.19. In the Avida landscape, the peaks start out almost completely connected. The horizontal leveling is tied to the phenotypic structure, since there are only 1108 possible fitness levels in the Avida landscape and these are distributed in a step-like fashion.

Figure 3.18: NK data, peak percolation, neighbors joined within distance 2. Visible are all points above the fitness threshold and the size of the largest connected cluster.

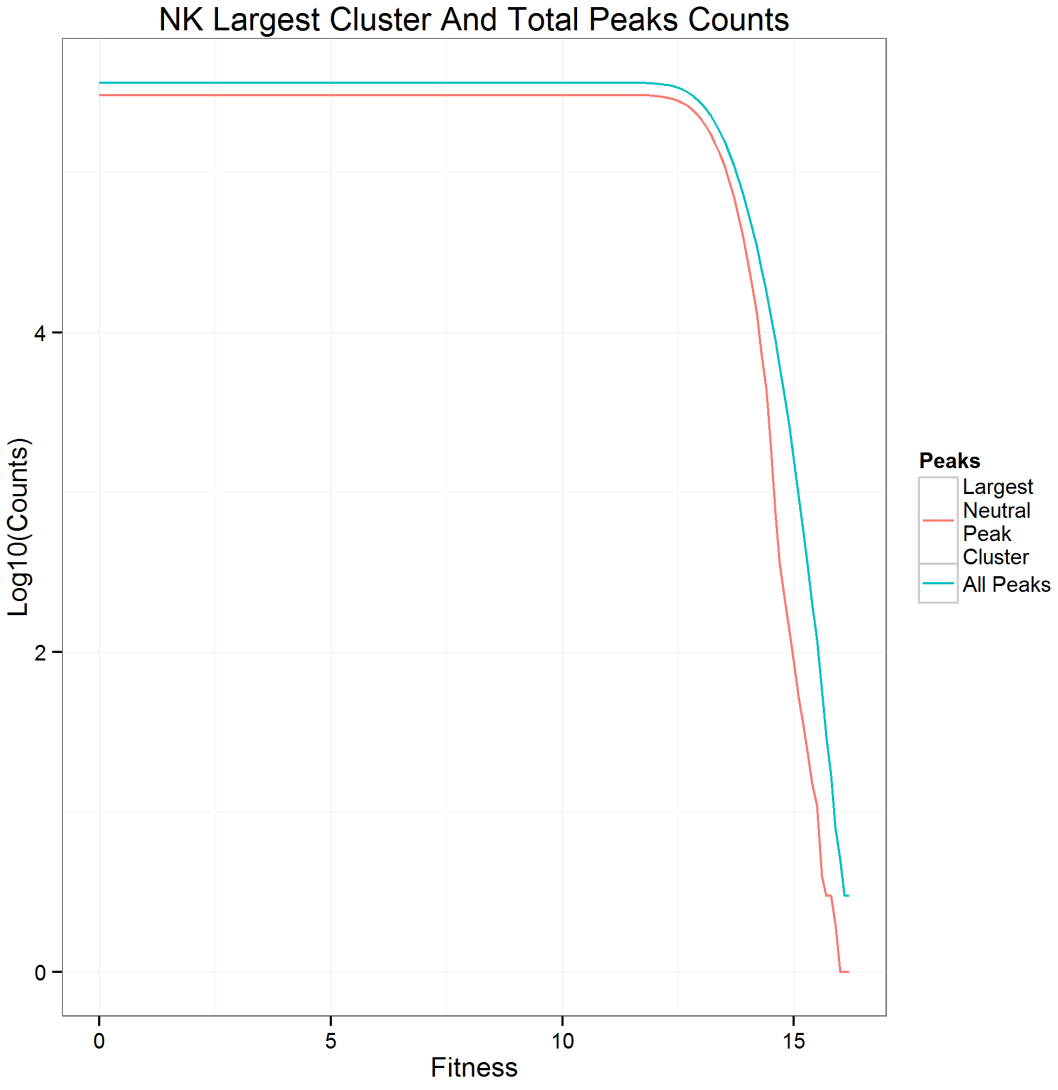
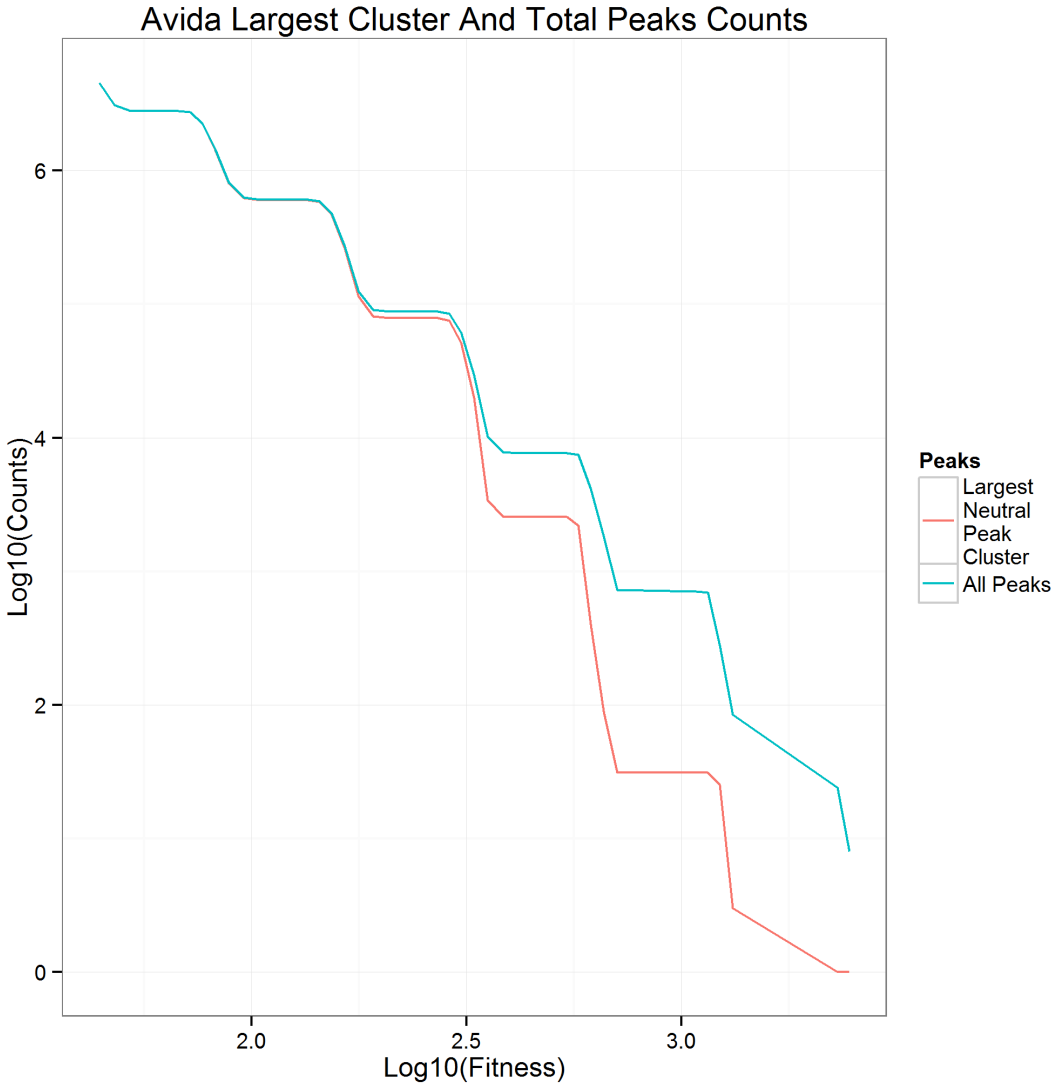


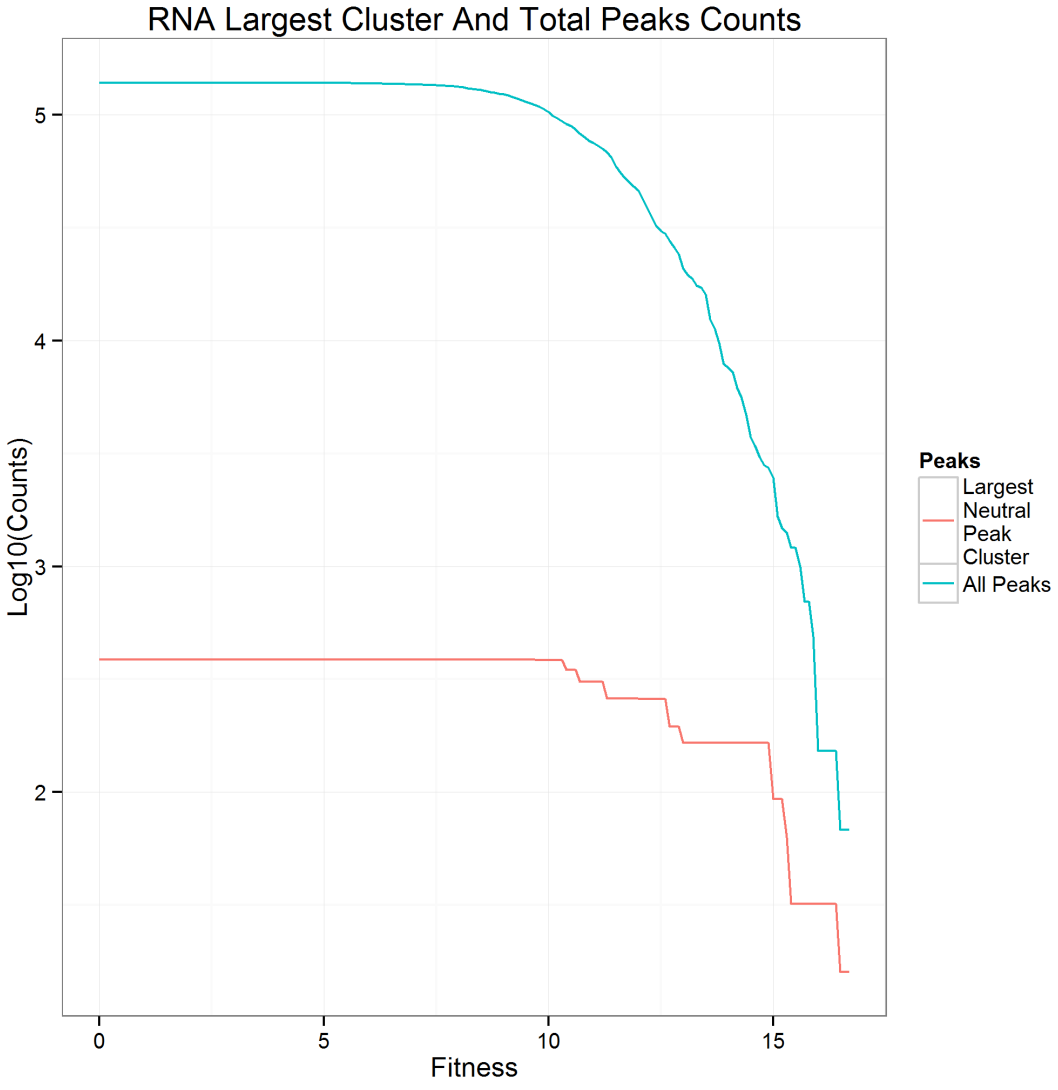
Figure 3.19: Avida data, Peak Percolation with neighbors joined within distance 2. Visible are all points above the fitness threshold and the size of the largest connected cluster.



3.7.3 RNA

The percolation for the RNA dataset can be found in Figure 3.20. Most neutral peaks are not connected; even at a fitness threshold of zero (allowing all peaks to be considered), I see that only a few hundred out of 184,020 neutral peaks are actually connected to begin with. The lack of connectedness in the graph suggests that, in the RNA data set, peaks are often isolated. The percolation pattern and connectedness is in sharp contrast to the pattern observed in the Avida and RNA datasets.

Figure 3.20: RNA data, peak percolation, neighbors joined within distance 2. Visible are all points above the fitness threshold and the size of the largest connected cluster.



3.8 Autocorrelation of Peaks

Finally, one last analysis I performed was similar to the autocorrelation analysis proposed by [Weinberger, 1991], except rather than measuring random walks, I measured a structural property of the landscape pertaining to peaks: namely peak fitness correlates with distance between peaks. In Equation 3.3, I reproduced equation 4 from [Barnett, 1998] and originally formulated by [Eigen et al., 1989]. Here, the autocorrelation at distance d is represented by $\rho(d)$, g and g' are two points within the landscape. $Q^N(d)$ represents points d distance apart, $\bar{f}$ is the average fitness, and ${\sigma_f}^2$ is the variance of the fitnesses of the points in the collection. In order to collect this data, I needed to look at all-pairs distances, which proved especially challenging when there were millions of data points as in the Avida set.

$$\rho(d) = \frac{1}{{\sigma_f}^2} \frac{1}{|Q^N(d)|} \sum_{g,g' \in Q^N(d)} (f(g) - \bar{f})(f(g') - \bar{f}) \tag{3.3}$$

3.8.1 NK

The autocorrelation in the NK landscape can be seen in Figure 3.21. There appears to be an exponential response of autocorrelation to distance.

Figure 3.21: NK peak autocorrelation

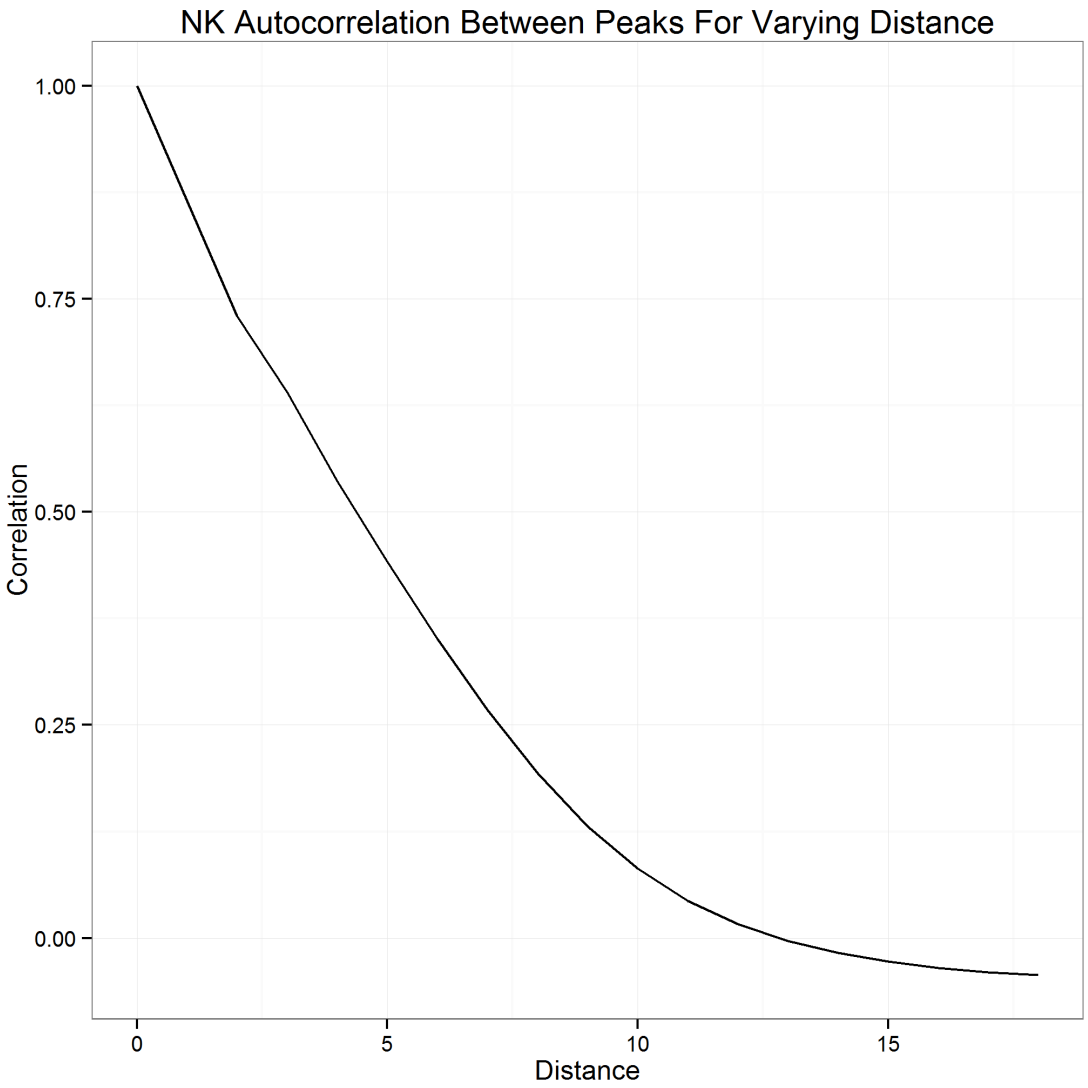
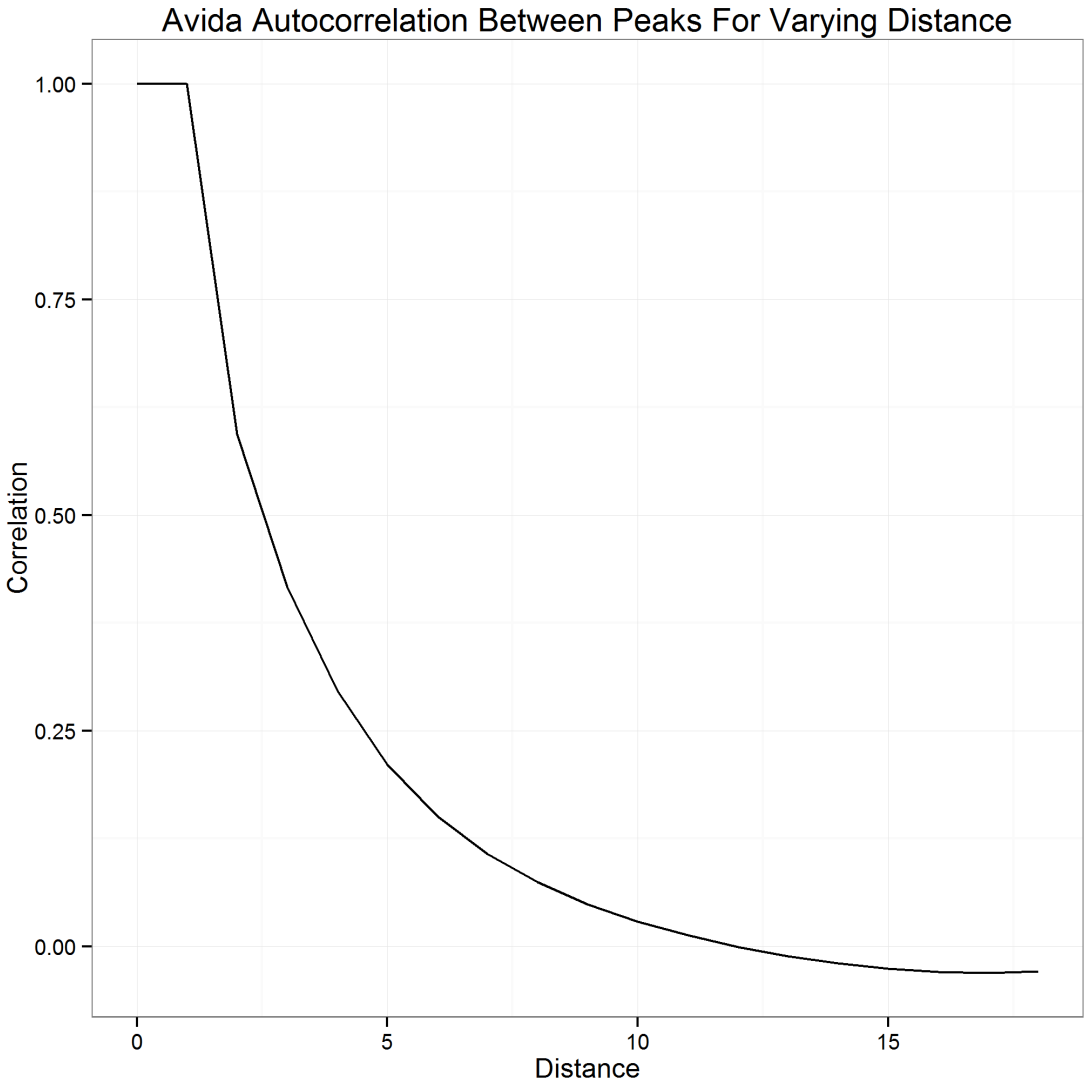


Figure 3.22: Avida thresholded peak autocorrelation



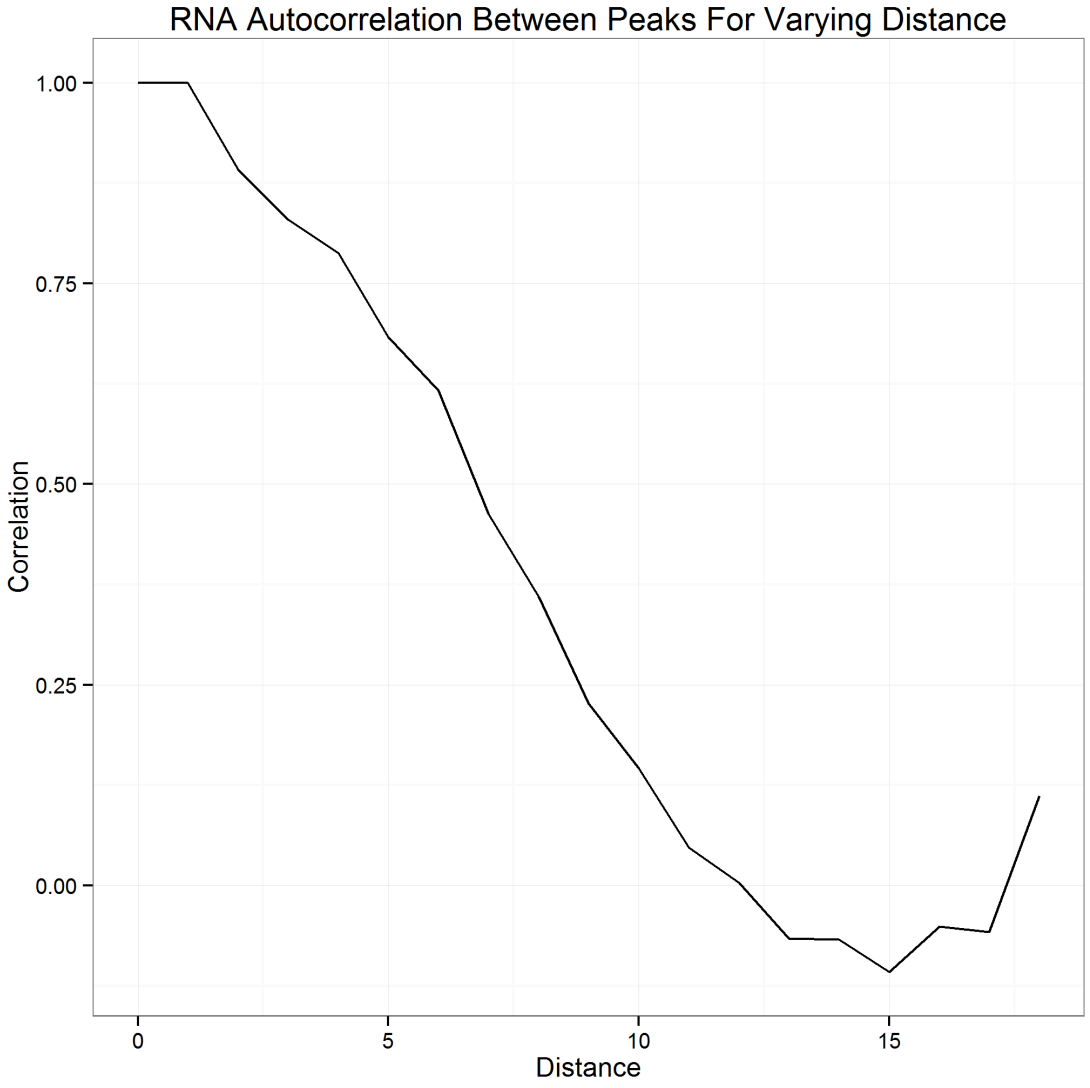
3.8.2 Avida

The resulting autocorrelation in Avida can be seen in Figure 3.22. At distance zero, there is full autocorrelation, since each peak's fitness correlates fully with itself. Similarly, at distance one, only neutral peaks neighbors may exist in the peak data set, which by definition have the same fitness, so the autocorrelation is one. After this, the autocorrelation drops off in what appears to be a very smooth exponential or power distribution.

3.8.3 RNA

Figure 3.23 shows the autocorrelation of peaks. Once again, there is full autocorrelation at distance one from the neutral peaks. As in the other landscapes, when peaks are close, the correlation in fitness is high. The correlation goes negative as distance increases and then finally becomes weakly positive at distance near 18. This last phenomenon may be due to the fact that if a given sequence is stable, its complement is also likely to be stable.

Figure 3.23: RNA peak autocorrelation



3.9 Discussion and Conclusions

I have presented a comprehensive investigation of several structural properties in length 18 genomes spanning NK, Avida, and RNA landscapes comprising 68,719,476,736 genotypes each. Landscapes of this size have not been previously studied so systematically. For each landscape instantiation, I did an exhaustive genotype to phenotype analysis. This investigation provides us with insight into theoretical landscapes often used to study evolution, genetic programming, and RNA sequences. The fitnesses found within the NK landscape follow a normal distribution, as do the peak fitnesses within the NK landscape. The RNA and Avida landscapes in contrast are mostly dead—that is most genotypes have minimal fitness due to the phenotypic nature of these landscapes. On the contrary, the expectation for a random NK landscape sequence is a fitness of 9—exactly at the 50th percentile. This is one important difference between NK landscapes and the landscapes I see in nature. The peaks in Avida and RNA landscapes skew towards the tail of higher fitnesses, but in RNA it is even more pronounced than in Avida due to the large prevalence of neutrality in Avida landscapes.

For all three instantiations of landscapes, peaks tend to cluster near other peaks; in Section 3.6, I found that there is a clear positive relationship between the fitness of a peak and the fitness of peaks within a neighborhood of two. This relationship held across all three types of landscapes. For NK, this effect has already been demonstrated in previous work, but the broader generality of this finding is novel with its expansion to larger NK landscapes as well as for both Avida and RNA landscapes. These results provide strong quantitative evidence that peaks are not distributed randomly throughout the landscape, but do indeed tend to cluster.

For both the NK and Avida landscapes, the autocorrelation starts high and drops rapidly, eventually going below zero as the distance approaches 18. The RNA landscape has a higher autocorrelation for longer, but also eventually declines and then, surprisingly, goes positive for large distances. This effect at high distances is potentially because my fitness function is

tied to the folding energy, and the complement of a sequence that is stable is also likely to be stable. Such a complement could only appear in regions at the maximum distance of 18. This result may indicate a certain amount of symmetry in the RNA landscape that does not exist in the other two model landscapes—namely if you locate a peak in RNA space, it is profitable to search regions around the complement.

Percolation (an indication of the ability for evolution to navigate to wide ranging peaks through intermediate peaks) is only present in the Avida and NK landscapes—the RNA landscape, in contrast, starts with only a small fraction of peaks connected. This result implies that nearly neutral pathways that connect clusters of peaks may not exist on RNA landscapes. Conversely, these pathways may exist on Avida and NK landscapes. This difference implies that evolution may work differently in these landscapes, a hypothesis I investigate further in future chapters.

Chapter 4

Basins of Attraction

4.1 Introduction

High fitness peaks dominate their local neighborhoods in genotypic space by allowing organisms to outcompete their neighbors, thus attracting and trapping evolutionary trajectories. These peaks are are often referred to as 'basins of attraction' because it is easier for evolution to permit a transition from a relatively unfit genotype to a relatively fit genotype, whereas the opposite transition is actively opposed by the process of natural selection. In general, the longer upward trajectories are that lead towards a peak, the larger the associated basin is likely to be.

In this chapter, I explore questions of predictability and evolvability within rugged adaptive landscapes by examining the size and distribution of basins of attraction. These properties of basins have implications for biological evolution since natural selection drives populations towards higher levels of fitness, so an improved understanding of how trajectories evolve on fixed fitness landscapes may help us understand how they evolve in nature.

4.2 Background

[Kauffman, 1993] was the first to examine basins of attraction in the context of NK landscapes, and he measured basin size by counting the number of random walks ending up at each optimum. Similarly, [Ochoa et al., 2008] examined NK fitness landscapes by exhaustively enumerating instantiations of these landscapes across a variety of K values for N = 16 and N = 18. Basins of attraction were determined for each point by starting a walk at each genotype, and picking the neighbor with the highest fitness until a local optimum was reached. If a tie for neighbor of highest fitness occurred (which may occur due to rounding error or when NKp landscapes with neutrality are used), the next step was chosen between them at random. Trajectories were then binned into genotypes which permitted them to make many conclusions about basins of attraction in NK landscapes. Among their conclusions, they found a positive exponential correlation between the fitness values of maxima and basin size. They also found that as K increased, the size of the largest fitness basin shrunk as a proportion of the fitness landscape.

[Ochoa et al., 2010] used the same technique, but instead of each walk favoring the most fit neighbor, it would choose any neighboring point with greater fitness. This technique resulted in a similar relationship between fitness and proportion as in [Ochoa et al., 2008], but randomly walking across neighbors tended to increase the size of smaller basins while decreasing the size of the largest basins. Additionally, they found that a genotype at N=16,K=4 may reach as much as 70% of the available basins, but at K=14, a solution may reach as little as 30%.

[Tomassini et al., 2008] measured transition probabilities between basins by examining the 'edges' of basins, defined as those points that belong to a basin that are one mutational step away from points that belong to other basins. The authors assembled a network upon the basins by examining the ratio of outgoing links to within-basin links.

[Vérel et al., 2011] extends the methodology of [Ochoa et al., 2008] to look at NKp neutral

networks. The hill-climbing algorithm for determining basins was altered by performing a random walk to points greater than or equal to the max fitness bordering any given point. The final points were then used to construct neutral networks by examining edges, with single peaks corresponding to a neutral network of size one.

4.3 Measuring Basins

In this section, I present a new method for measuring the size of basins, the Basin Flow Algorithm.

4.3.1 NK Fitness Landscape Variants

In the previous chapter, the standard form of the NK landscape family, with a genome of length N with K interacting genes, was introduced. A modified version of the NK landscape was proposed by [Østman et al., 2010], where instead of using the arithmetic mean, the geometric mean is preferred. The biological justification is that it simulates the effects of lethality—the expression of a single bad gene is enough to drop fitness substantially.

$$\left(f(G) = \prod_{1}^{N} f_i(G) \right)^{\frac{1}{N}} \tag{4.1}$$
$$\text{where } f(G_i) = f(g_{i,1}, ..., g_{i,k}) \sim U(0,1)$$

The NK landscape contains almost no neutrality since the probability of obtaining two random numbers of the same exact fitness is near zero since neutrality is possible due to finite precision. [Barnett, 1998] introduced the NKp landscape where there is a p probability that a given allele contribution will produce a gene with zero fitness, introducing neutrality. Similarly, I introduce a probability that a given allele contribution will produce a gene with a one fitness contribution. This probability has the secondary effect of increasing fitness values of the distribution due to the geometric mean.

4.3.2 Basin Flow Algorithm

I seek to measure basins of attraction by looking at where population centroids would 'flow towards', given a uniform initial distribution across the entire fitness landscape. I take broad inspiration for this idea from the page rank algorithm, [Page et al., 1999] originally employed by Google, and which seeks to establish the relative importance of a web page by examining the properties of its links. Internet surfers are modeled as random walkers who make decisions about where to navigate next based on outgoing links from the current page they are visiting. The page rank seeks to identify nodes where surfers spend the most time.

Similarly, here, I model the centroids of populations as random walkers on a fitness landscape and seek to identify which nodes they are drawn towards by examining one point mutational neighbors. By centroid, I mean the approximate center of the population on the genotypic lattice. Neighbors that are greater than or equal to are analogous to 'outgoing links' in the page rank algorithm. The accumulation of random walkers ultimately identifies nodes that are either local maxima or part of locally maximum neutral networks, or 'sinks'. Unlike the page rank algorithm, these are treated as true sinks.

This methodology differs from mutation-selection-balance population genetic models since rather than competing genotypes among populations or competing populations against each other, I am instead measuring random walks of population centroids. This method is a fixed structural measure of the fitness landscape. Previous work in this domain has focused on sampling trajectories, whereas my approach represents a belief of basin density over all genotypes in the fitness landscape. A genotype is then potentially simultaneously a member of multiple basins by distributing flow over all qualifying neighboring genotypes. I assume a uniform starting distribution, which admittedly is not the only prior belief one can have about populations upon the landscape, but is a straightforward choice. In reality, of course, finite populations necessarily cannot start with such a distribution and are already trapped in one small area of the landscape, because of both common descent (see Chapter 2) and

selection. With this method, I then explore the entire landscape without bias, which differs from sampling methods which are necessarily biased from the initial sample points.

Additionally, I seek to examine the effect that resulting network structure has on the process of evolution. To do this, I apply a secondary analysis to measure transience of basins by allowing jumps between genotypes two mutations apart. This secondary analysis provides insight into what the basin structure might look like on an evolutionary time scale when double mutations may be regularly explored.

In this work, I use an instantiation of the NK landscape described in the previous chapter, with length = 20 and two alleles per site. I use a modified version of the NKp landscape using the geometric mean as identified earlier in Equation 4.1 and where p, rather than a gene having a probability of adding 0 fitness, has a probability of multiplying by 1. As mentioned earlier, this NKP variant causes further skew to the right in NKp landscapes, as the geometric mean will be larger when a 1 is substituted rather than a random number between 0 and 1.

To measure basins, I assign a density value to each genotype. The density represents the proportion of population trajectories at each genotype in the landscape. I start with a uniform initial density of $\frac{1}{a^L}$, where a is the number of alleles and L is the length of the genome. I establish directional links between genotypes that may distribute density to neighboring genotypes, based on relative fitness and the local structure of the neighborhood. This transition probability represents the chance that a population centered at a given site may evolve to a neighboring site. There is some literature pertaining to competition in the context of clonal interference between beneficial mutations as in [Gerrish and Lenski, 1998], [Desai et al., 2007], and [Wilke, 2004], but I opt to use two simple treatments instead described below. There is no coherent view of what timescale an update corresponds to in this work, since selection strength would be varied, so I am more interested in the convergent dynamics than in the transient dynamics; after 1,000 updates, I measure the density in genotypes as an estimate of the size of basin of attraction.

In the ProbGE treatment, I constructed an outgoing edge $e(A, B)$ from genotype A to genotype B with weight $w(A, B)$ as follows:

$$w(A,B) = \begin{cases} 0 & \text{if } F(A) < F(B) \\ \frac{f(B)}{\sum_{n \in N_A \cup \{a\}} I(F(n) \geq F(A))F(n)} & \text{if } F(A) \geq F(B) \end{cases} \tag{4.2}$$

In this equation, $F(A)$ represents the fitness of genotype A, N_A is the neighborhood of one point mutations surrounding A, and I is the indicator function that enforces that these points have greater than or equal fitness relative to point A. The biological assumption underlying this treatment is that neighboring points have a chance of receiving a trajectory equivalent to their fitness divided by the total neighboring fitness of all neighboring candidates.

The second treatment is the FlatGE treatment, in which I constructed an outgoing edge from genotype A to genotype to B with weight $w(A, B)$ as follows:

$$w(A,B) = \begin{cases} 0 & \text{if } F(A) < F(B) \\ \frac{1}{\sum_{n \in N_A \cup \{a\}} I(F(n) \geq F(A))} & \text{if } F(A) \geq F(B) \end{cases} \tag{4.3}$$

In other words, all outgoing links have equal weight in the second treatment, whereas they are weighted by relative fitness in the first. After 1,000 updates, I examine the distribution of the density in points that are above a threshold of 0.00001. The key biological assumption under this treatment is that if mutations are rare enough, the population is small enough, or the selection coefficient is strong enough, sweeps may occur sufficiently fast relative to the appearance of new competitive types that the first beneficial or neutral mutation may sweep.

In both models, I make the following assumptions: (1) mutations within the neighborhood are equiprobable, (2) double mutations do not occur, (3) likewise, the population centroids, or centers of mass on the genotype space, may not take more than one step at once, and (4) natural selection prevents populations from taking downward steps. These assumptions

of course do not necessarily hold in natural populations, but they do make the model much more tractable, and I explore relaxation of the double mutation assumption next.

As I saw in the previous chapter, [Kauffman, 1993] described a feature common to the NK family of landscapes that he termed the 'Massif Central', which referred to the observation that peaks tend to cluster close to each other, as opposed to being randomly distributed across the landscape. [Østman et al., 2010] provides evidence that peaks tend to be clustered in the particular NK landscape I use. This idea, that peaks are not necessarily far from each other, lends itself to a secondary analysis, whereby I evaluate the idea that trajectories can and do escape local peaks during the process of evolution. This treatment relaxes the assumptions that double mutations do not occur and that natural selection prevents populations from taking downward steps. The basic assumption is that most peaks are themselves transient phenomena in an evolutionary trajectory, albeit those likely to have a longer waiting time to escape. In fact, since the density I measure represents population centroids, it would not be that difficult to cross a double mutation valley. Simultaneous mutations, compensatory mutations through a valley crossing, and even drift provide mechanisms whereby populations may evacuate regions of the fitness landscape altogether and accumulate in more highly fit regions.

In my secondary treatment, after running the first 1,000 steps as above, I run another 1,000 steps where in addition to percolation every update, I perform an additional calculation: I allow points with a density over threshold $\delta = 0.00001$ to lose mass to genotypes that are two steps out, again obeying the above equations, with the caveat that N_A instead represents the 2-step mutational neighborhood of genotype A. The density threshold permits us to ignore the vast majority of points underneath the threshold, which is a relatively useful optimization since there are $\frac{L(L-1)}{2}$ 2-step mutants for each genotype, and the vast majority of genotypes already contain negligible or no mass. This threshold cuts down on the processing requirement at the potential cost of sacrificing some accuracy.

Finally, I examine the effects of neutrality on basin size. The exact role of neutral networks is still far from understood; such regions may serve as conduits to higher regions of the landscape or as sinks where trajectories get stuck or both. The distribution of neutral regions also may play a significant role; it might be imagined that a large multidimensional neutral region surrounded primarily by deleterious genotypes may retain a lot of mass simply because trajectories are randomly walking around in this neutral region. There is some evidence for 'survival of the flattest' or the notion that even in the absence of fitness benefits, populations may evolve towards genotypes with fewer neutral neighbors as in [Wilke, 2001] and [Wilke et al., 2001].

4.3.3 Discussion

I perform both the ProbGE and FlatGE treatments in a length N=20 landscape for K = 2, 5, 8, 11, 14, 17, and 20 with p=0, p=0.2, p=0.5, and p=0.8. For each treatment, I generate five random landscapes. I store only genotypes above a threshold of 0.00001.

Figure 4.1 shows these results for the no neutrality case (p=0.0).

In order to examine the effects of fitness and treatment condition (ProbGE = 0, FlatGE = 1) on the basin size, I fit the model

$$\log_{10} B = \beta_0 + \beta_1(F) + \beta_2(T) + \beta_3(F\text{x}T) \tag{4.4}$$

Here, B is basin density, F is fitness, and T is treatment condition. I fit this model across each pair of fitness landscapes; i.e. for each K, the same five landscapes are used in both the ProbGE and FlatGE treatments. The reason I prefer this type of pairwise analysis is that the landscapes in question change significantly as K increases. In particular, increased epistasis permits more combinations, which increases the maximum fitness possible. The results of this can be seen in Table 4.1. To summarize, the log of the basin size is predicted by fitness ($p < 4.85 * 10^{-31}$) for all cases. Additionally, this relationship decreases with increasing

Table 4.1: Coefficients of $\log_{10} B = \beta_0 + \beta_1(F) + \beta_2(T) + \beta_3(F \text{ x } T)$. p-values are in parentheses.

K	Intercept	Fitness	Type	Fitness x Type
2	-2.482 (1.56e-252)	9.284 (4.85e-31)	0.06139 (0.301)	-0.9195 (0.391)
5	-4.629 (0)	6.841 (0)	0.07802 (8.39e-06)	-0.3776 (3.74e-05)
8	-5.819 (0)	6.215 (0)	0.08586 (1.86e-15)	-0.2649 (9.05e-14)
11	-6.464 (0)	5.62 (0)	0.09555 (2.01e-26)	-0.2349 (9.46e-25)
14	-6.608 (0)	4.634 (0)	0.1113 (2.77e-33)	-0.2375 (1.5e-32)
17	-6.265 (0)	3.201 (0)	0.08976 (1.61e-21)	-0.1687 (1.31e-20)
20	-6.004 (0)	2.365 (0)	0.05122 (1.73e-07)	-0.08538 (8.43e-07)

K, which is unsurprising since the additional epistasis caused by a higher K increases the number of optima in the landscape. Both the association with fitness size and basin size and the effect of increasing K on decreasing this association are consistent with the results of [Ochoa et al., 2008]. Secondly, I find that the FlatGE condition produces larger basin sizes than the ProbGE treatment. Furthermore, there is a negative interaction between fitness and treatment; which is to say that the maximum density of peaks in the FlatGE case was less than that in the ProbGE case. This interaction also seems analogous to previous work done by [Ochoa et al., 2010], who found a difference between taking the best neighboring point and taking the first neighboring point increased average basin size but decreased the largest basin sizes. The ProbGE treatment, by weighting fitness proportionately, does not exactly correspond to the best point, but it is likely a less biased version of the same effect; with equal weights, more trajectories are drawn towards local maxima. This effect also seems to decrease with increasing K, likely as a consequence of increasing ruggedness in the landscape.

Next, I examine the effects of the secondary network analysis on the basin size. Allowing valley crossings which permit basins to empty into higher nearby points results in a reduction of between 74% and 82% of all basin points above the threshold for the non-neutral case. This can be seen in Table 4.4. Table 4.2 and Table 4.3 show regression data for the ProbGE and FlatGE treatments, respectively. In both of these treatments, the slope of the fitness increases after the second order manipulation. The effects of the second order network can be seen in that it has an increasingly negative effect on basin size as K increases (when controlled for the range of fitnesses), but there is a large fitness x second treatment interaction term. In other words, higher fitness basins tend to get bigger and lower fitness basins are more likely to disappear via drainage.

Table 4.2: Coefficients of $\log_{10} B = \beta_0 + \beta_1(F) + \beta_2(S) + \beta_3(F \text{ x } S)$ for ProbGE treatment. p-values are in parentheses.

K	Intercept	Fitness	Second	Fitness x Second
2	-2.482 (7.21e-182)	9.284 (3.84e-28)	0.9298 (4.67e-15)	-4.544 (0.00205)
5	-4.629 (0)	6.841 (0)	0.3553 (9.47e-20)	0.04766 (0.775)
8	-5.819 (0)	6.215 (0)	0.05732 (0.015)	0.6512 (4.28e-22)
11	-6.464 (0)	5.62 (0)	-0.1768 (3.7e-21)	0.9685 (1.08e-112)
14	-6.608 (0)	4.634 (0)	-0.3357 (2.96e-74)	1.138 (2.32e-211)
17	-6.265 (0)	3.201 (0)	-0.5051 (5.79e-157)	1.375 (0)
20	-6.004 (0)	2.365 (0)	-0.7414 (2.18e-290)	1.691 (0)

Table 4.3: Coefficients of $\log_{10} B = \beta_0 + \beta_1(F) + \beta_2(S) + \beta_3(F \text{ x } S)$ for ProbGE treatment. p-values are in parentheses.

K	Intercept	Fitness	Second	Fitness x Second
2	-2.42 (3.48e-179)	8.365 (5.12e-24)	0.8822 (6.17e-14)	-3.773 (0.00993)
5	-4.551 (0)	6.463 (0)	0.2978 (1.43e-13)	0.3277 (0.0572)
8	-5.733 (0)	5.95 (0)	-0.007231 (0.765)	0.8528 (7.43e-35)
11	-6.368 (0)	5.385 (0)	-0.2472 (1.33e-37)	1.143 (3.35e-147)
14	-6.496 (0)	4.397 (0)	-0.4013 (6.49e-100)	1.282 (1.51e-253)
17	-6.175 (0)	3.033 (0)	-0.532 (4.51e-169)	1.43 (0)
20	-5.952 (0)	2.28 (0)	-0.7332 (5.29e-279)	1.678 (0)

Table 4.4: This table shows the number of basins before and after the secondary analysis is performed, as well as the mean. Each cell in the table has two entries—the first represents the value for the initial treatment with flow only to mutational neighbors, and the second represents the value of the secondary treatment with flow to genotypes two mutational steps away.

K	ProbGE-Points	ProbGE-Means	FlatGEPoints	FlatGEMeans
2	302 / 53	0.047 / 0.072	302 / 53	0.047 / 0.072
5	6581 / 1190	0.183 / 0.234	6581 / 1190	0.184 / 0.234
8	28290 / 5995	0.299 / 0.354	28298 / 5995	0.299 / 0.354
11	64672 / 15422	0.390 / 0.442	64811 / 15422	0.389 / 0.442
14	110800 / 29011	0.461 / 0.508	111482 / 28986	0.460 / 0.508
17	157246 / 41099	0.518 / 0.563	157671 / 40989	0.517 / 0.563
20	179852 / 44384	0.565 / 0.610	179377 / 44217	0.565 / 0.610

Figure 4.3 and Figure 4.4 show the single mutation and double mutation results for an NKp landscape of $p = 0.2$. Figure 4.5 and Figure 4.6 show the single mutation and double mutation results for a NKp landscape of $p = 0.5$. Figure 4.7 and Figure 4.8 show the single mutation and double mutation results for a NKp landscape of $p = 0.8$. For the $p = 0.5$ neutrality landscape, for the K=2 and the K=5 cases, there appear to have been several cases where the neutrality in the landscape did not successfully drain all the basins. This is visually apparent in the first four boxes of Figures 4.5 and 4.6. Several basins in the first analysis drop in proportion but do not disappear. The likely conclusion is simply that 1000 updates is not enough to allow trajectories to escape these regions. An alternative hypothesis might be that some of these are new points that did not appear previously due to being beneath the threshold. This phenomenon does appear to occur at a low level, reaching up to 15% of the points in the second order analysis in the K=20 case, where thresholding plays a larger role due to a higher incidence of peaks. However, at low K, this incidence was negligible, with, for example, two points for K=2 and no points for K=5 in the p=0.2 case, and two points each for K=2 and K=5 for the $p = 0.5$ case. The slow drainage of mass from certain basins thus appears to be a consequence of the neutrality in the landscape.

Figure 4.1: Log(basin proportion) as a function of fitness, $p = 0.0$, Single Mutations. Here density flows only to single mutant neighbors.

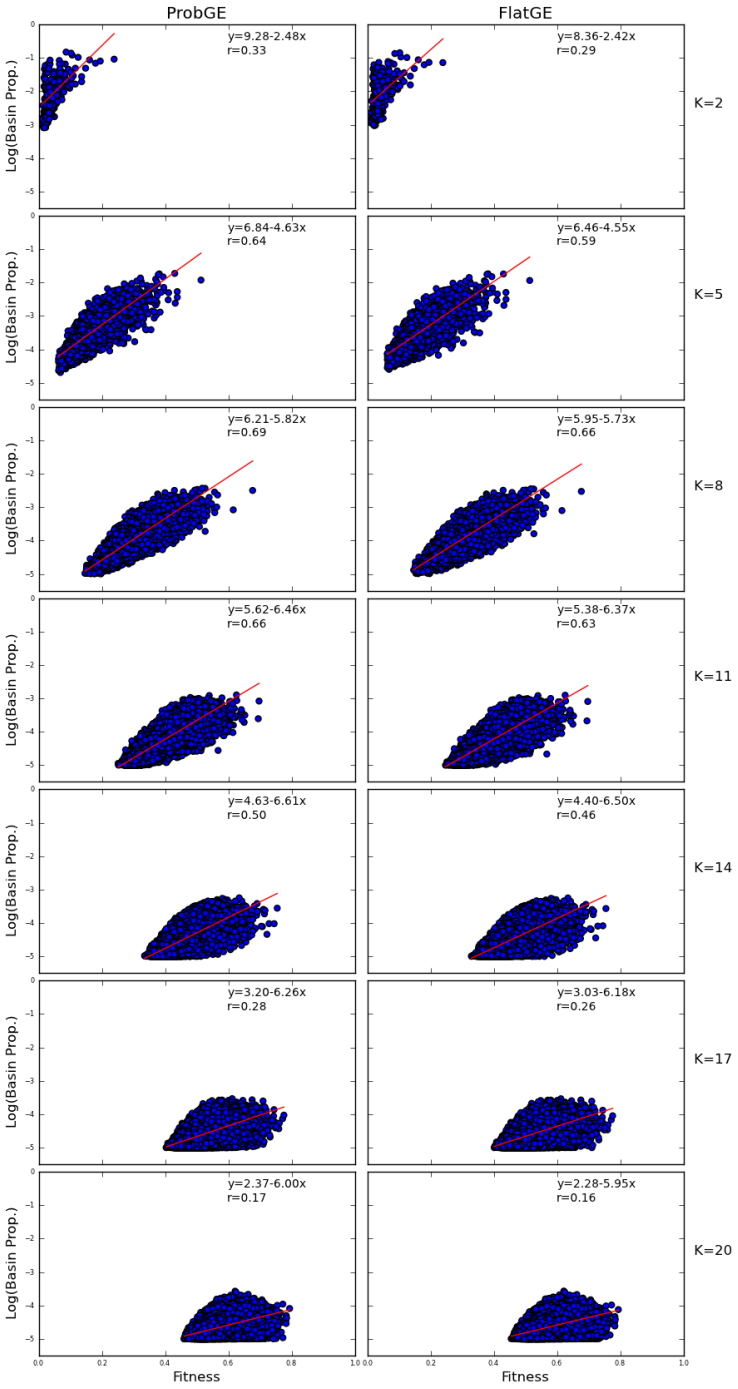


Figure 4.2: Log(basin proportion) as a function of fitness, $p = 0.0$, Double Mutations. Here density flows to single mutant neighbors as well as double mutants after 1000 updates.

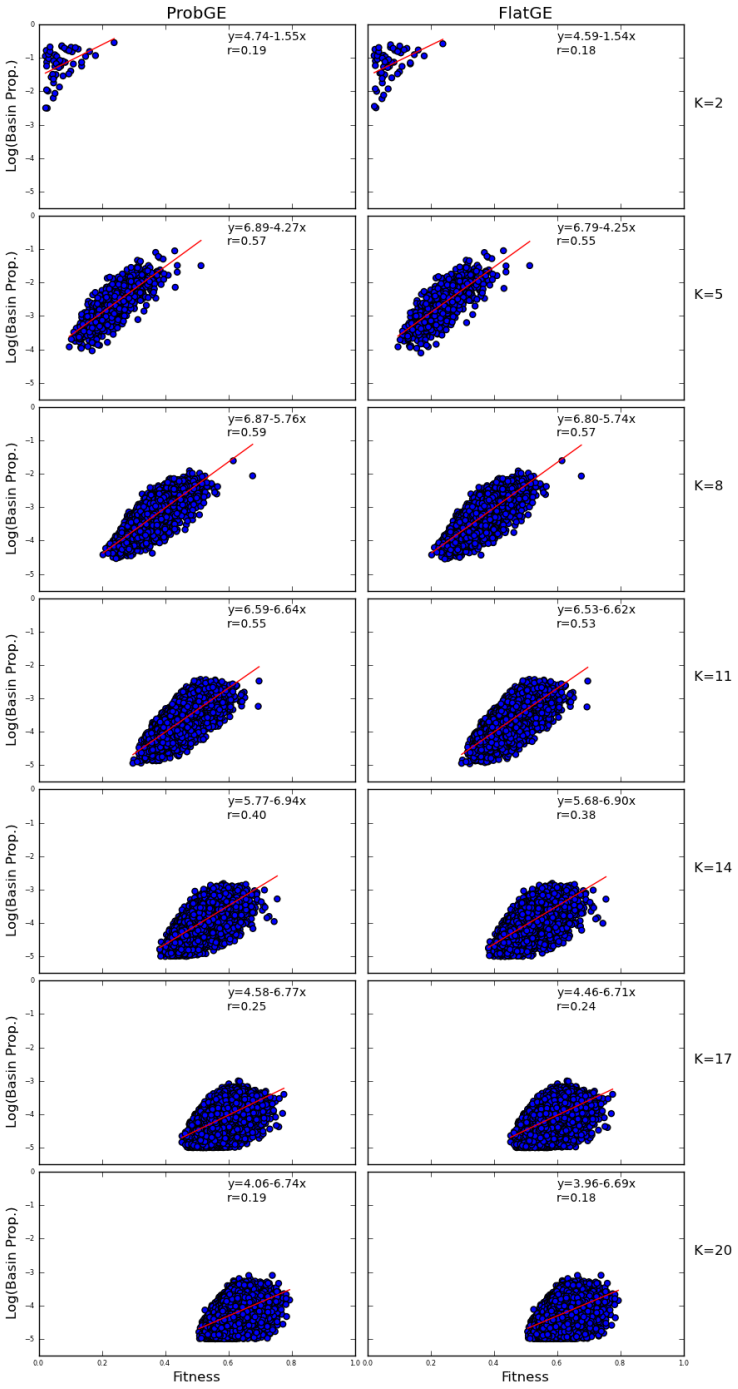


Figure 4.3: Log(basin proportion) as a function of fitness, $p = 0.2$, Single Mutations. Here density flows only to single mutant neighbors.

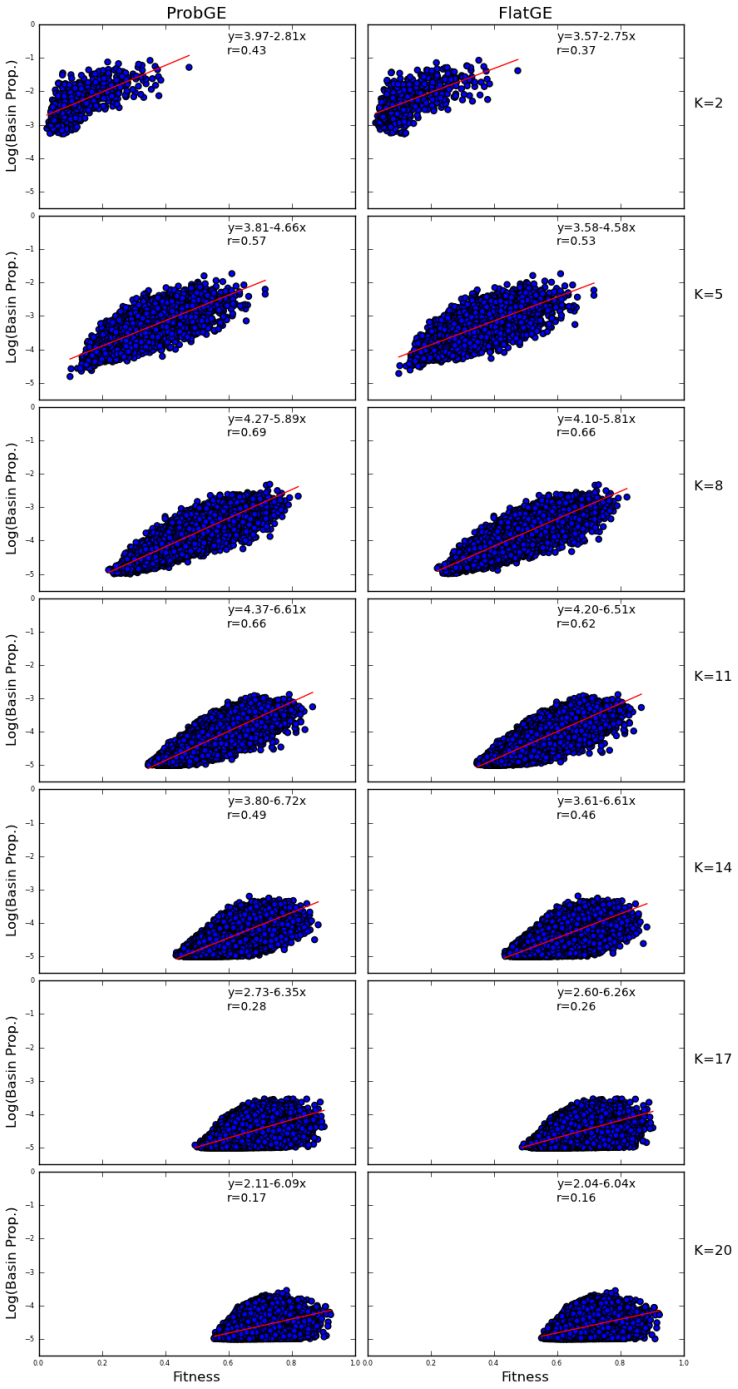


Figure 4.4: Log(basin proportion) as a function of fitness, $p = 0.2$, Double Mutations. Here density flows to single mutant neighbors as well as double mutants after 1000 updates.

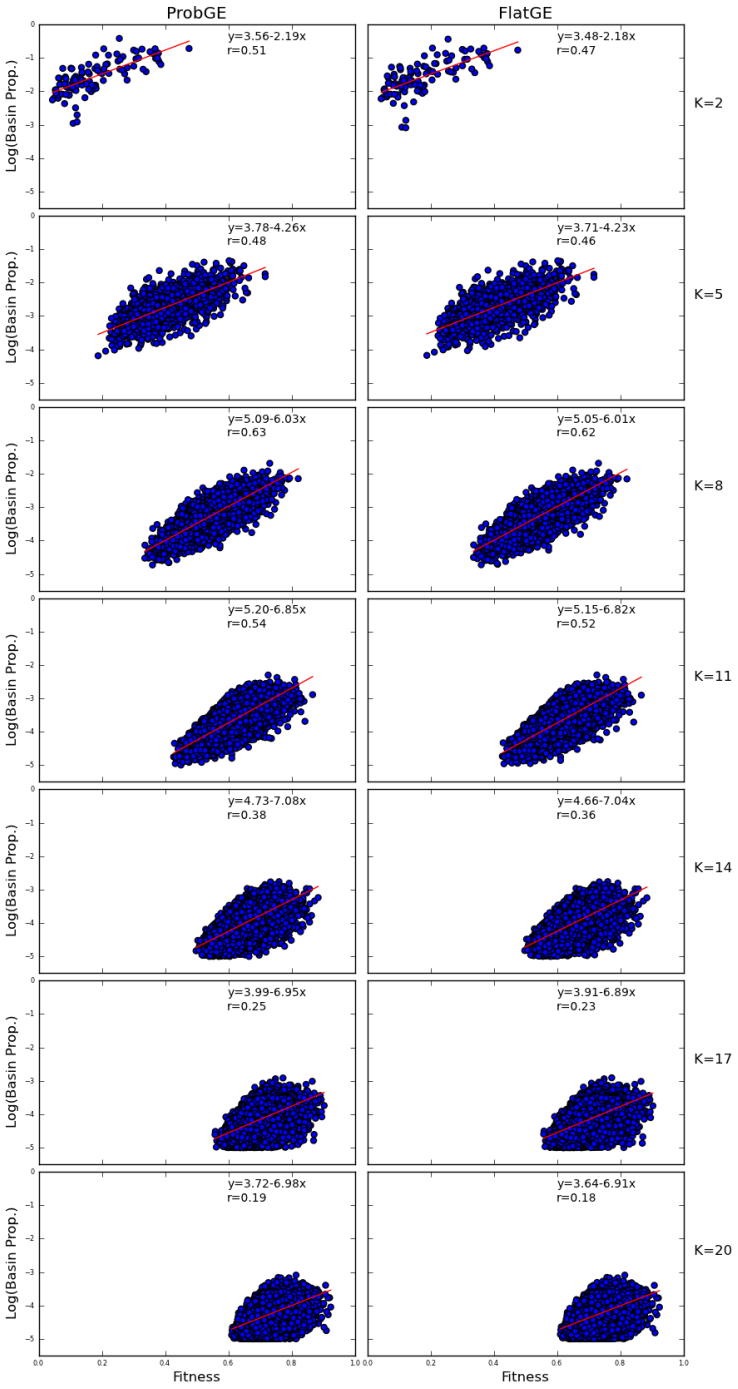


Figure 4.5: Log(Basin Proportion) as a function of fitness, $p = 0.5$, Single Mutations. Here density flows only to single mutant neighbors.

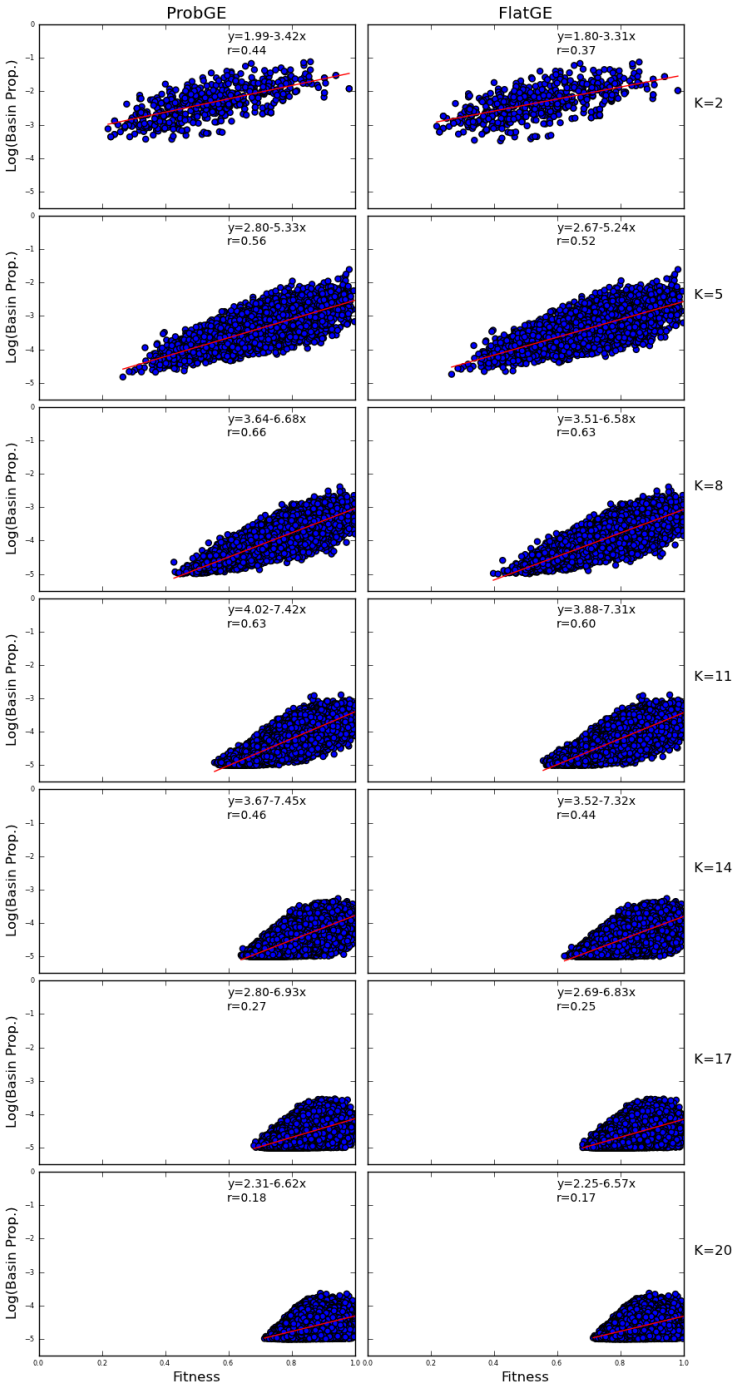


Figure 4.6: Log(Basin Proportion) as a function of fitness, $p = 0.5$, Double Mutations. Here density flows to single mutant neighbors as well as double mutants after 1000 updates.

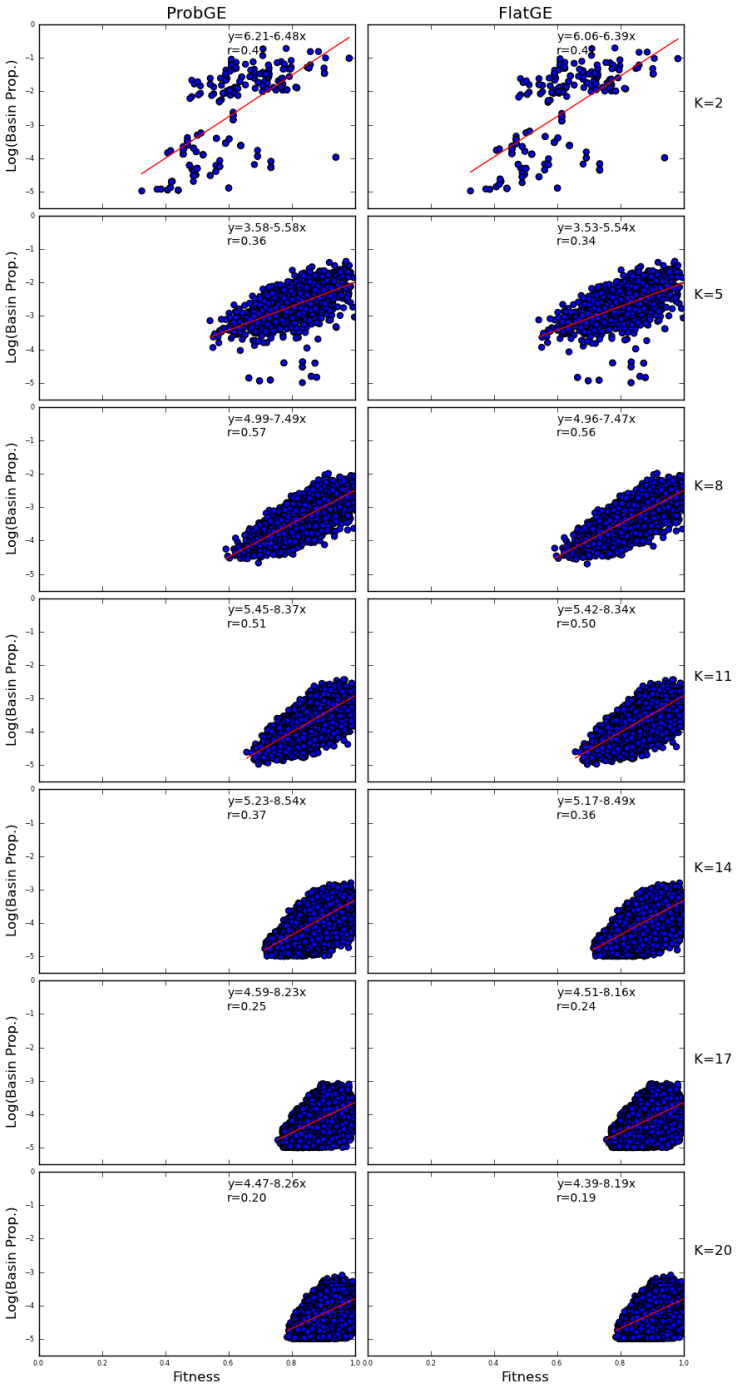


Figure 4.7: Log(basin proportion) as a function of fitness, p=0.8, Single Mutations. Here density flows only to single mutant neighbors.

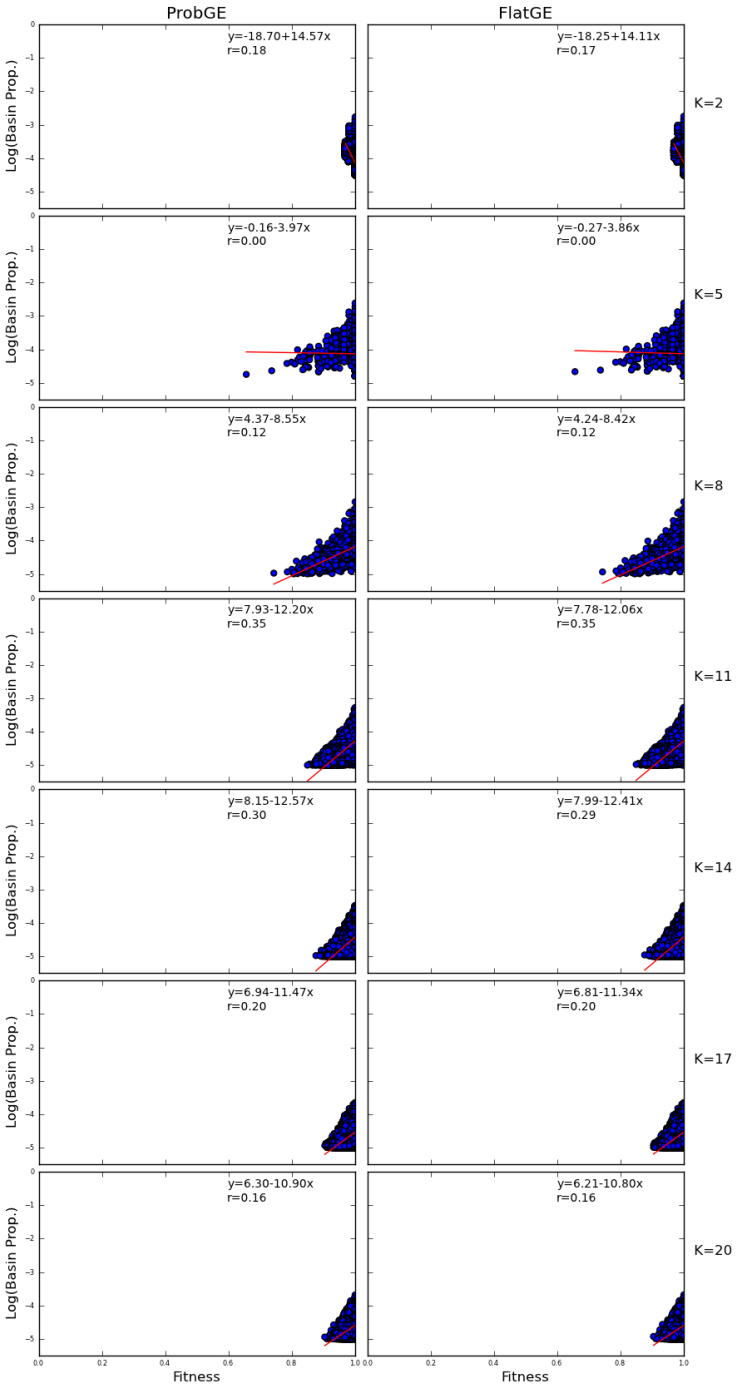
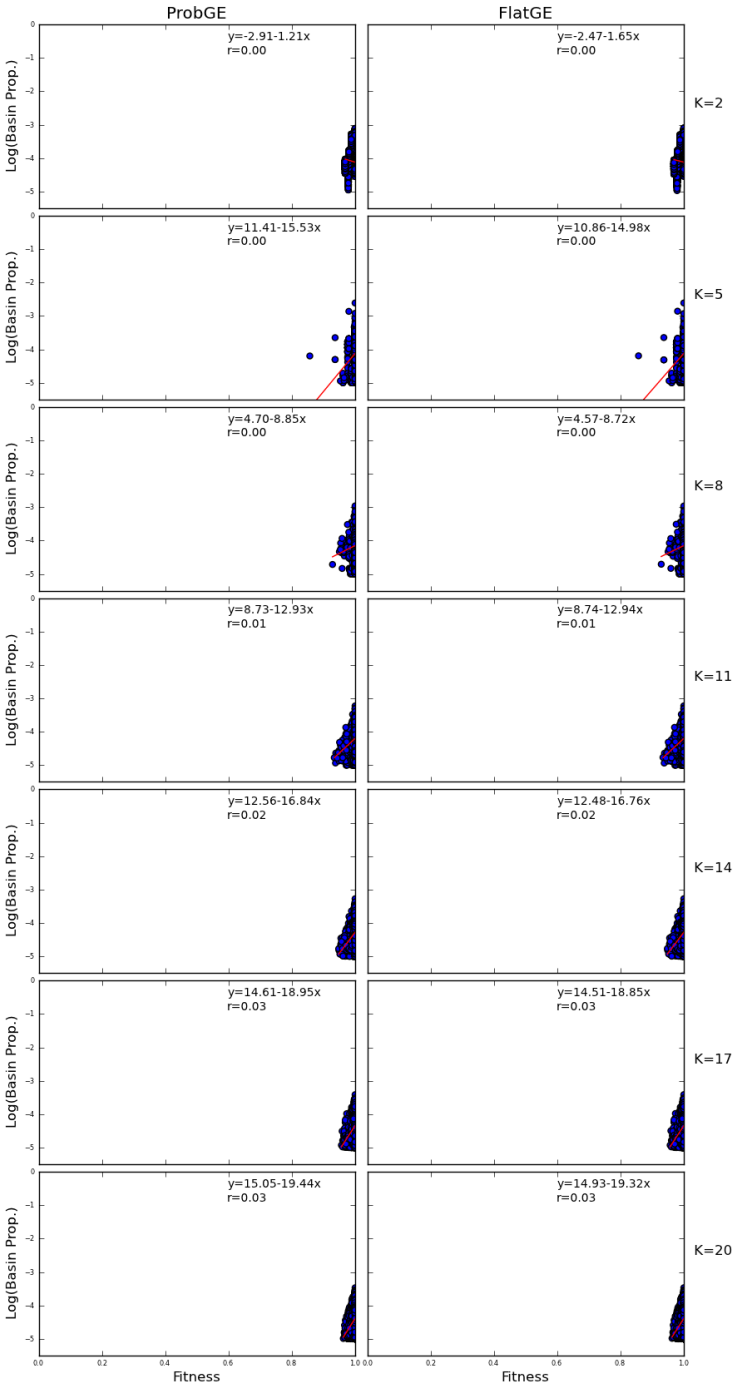


Figure 4.8: Log(basin proportion) as a function of fitness, p=0.8, Double Mutations. Here density flows to single mutant neighbors as well as double mutants after 1000 updates.



I have shown, using a method similar to page rank, that basin size in NK landscapes is exponentially correlated with fitness. This finding is consistent with previous work such as [Ochoa et al., 2008]. I set up two treatments, ProbGE in which genotypes are linked to neighboring genotypes of equal or greater fitness and distribute mass proportionately to fitness, and FlatGE in which genotypes are linked to neighboring genotypes of equal or greater fitness and distribute mass evenly without regard to fitness. I demonstrated that these treatments have an effect on the size of basins; the FlatGE treatment increases the average size of basins while decreasing the maximum size of basins relative to the ProbGE treatment. This result was reminiscent of that obtained in [Ochoa et al., 2010], comparing the accumulation of best vs. first improvement walks.

I also apply a secondary analysis where trajectories may escape local peaks by permitting them to move two mutational steps. The allowance of two mutant traversals has the effect of eliminating between 74% and 80% of peaks and increases the overall relationship between fitness and density of those that remained. This result shows that over evolutionary time, higher fitness basins tend to become even larger by cannibalizing nearby lower fitness basins.

Finally, I examined the effects of neutrality. I detected the presence of neutral networks that retain mass over large periods of evolutionary time, which functionally may act as sinks; even though there are exits from the network, they leak only slowly.

One confounding factor with this work would be the effect of the threshold on the interpretation of the data. Even in the K=20 case, only about 3.5% of the total genotypes appeared in the peaks data. It may be that the threshold was not low enough to capture all of the peaks in the landscape. This thresholding could result in very small basins being ignored for this analysis. This would likely skew the data reported in two ways: first, these basins would be of lower fitness and would not drain properly into nearby basins since the double mutant step was applied only above the threshold. If this skew does occur, when corrected, would only increase the reported effect of the secondary analysis of concentrating mass in larger

basins.

In this section, I have presented a novel method for evaluating fitness landscapes using the Basin Flow Algorithm and I demonstrated its usage and applicability on a simple NK landscape. This method produces comparable results to other methods of assessing basin size and is in line with previous work on NK landscapes. It also has the advantage of not being biased by starting location and provides a snapshot of 'true' relationships that might occur via transient dynamics.

4.4 Basins in Landscapes

In the previous chapter, I had explored complete landscapes of up to length 18 for the NK, Avida, and RNA landscapes. I sought to investigate the transient dynamics in these landscapes to see how trajectories accumulate in progressively higher peaks over time. Additionally, I was interested in measuring the amount of overlap between trajectories to get a sense for the size of basins in these landscapes.

I measured the size of fitness basins obtained by evolution by using a genetic algorithm to sample evolutionary trajectories I used a population size of 25 and 0.5 expected mutations on every division and sampled 100,000 runs, each originating from a different random genotype. Genotypes were sampled proportionately to fitness. Every update, in addition to mutation and selection steps, I record the genotype and fitness of the most common genome in the population, or 'dominant'.

I found that a population size of 25 was not sufficient to reliably overpower selection strength in the NK and RNA landscapes, which resulted in trajectories not accumulating in basins, as it was relatively difficult for them to stay in a peak. Therefore, I increased the strength of selection by applying the exponential transformation $f(x) = 2^x$ for the fitnesses in the RNA and NK landscapes to make these landscapes more readily comparable with the exponential Avida landscape and examine the basin distribution. This treatment preserves

the relative relationships of points—namely, peaks remain peaks, but the strength of selection is drastically increased. In this section, I will present both the untransformed and exponential for the NK and RNA landscapes.

4.4.1 Trajectories Over Time

I looked at the 100,000 collected trajectories at five update intervals: 1, 10, 100, 1000, and 10000 to see how the fitness of trajectories shifts over time in each of the three landscapes.

4.4.1.1 NK

NK basins over time can be observed in Figure 4.9. All of the distributions have a long tail—some trajectories continue to be trapped in local minima even 10,000 generations in. The resulting distributions appear to be normal or nearly-normal at any time, and as time elapses, the mean shifts upwards.

Figure 4.9: Distribution of fitnesses in evolutionary trajectories starting from 100,000 random points in the NK landscape at each of five exponential time points, as a measure of adaptation in basins over time.

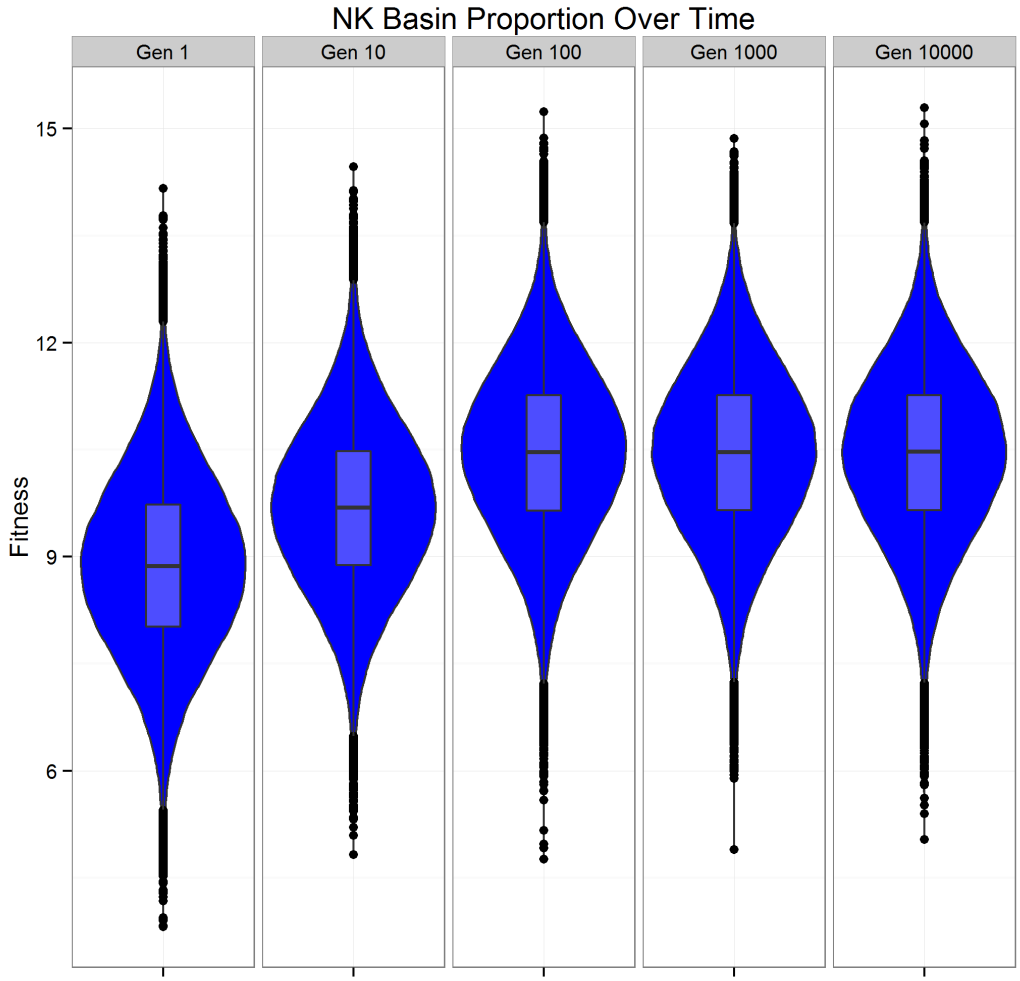
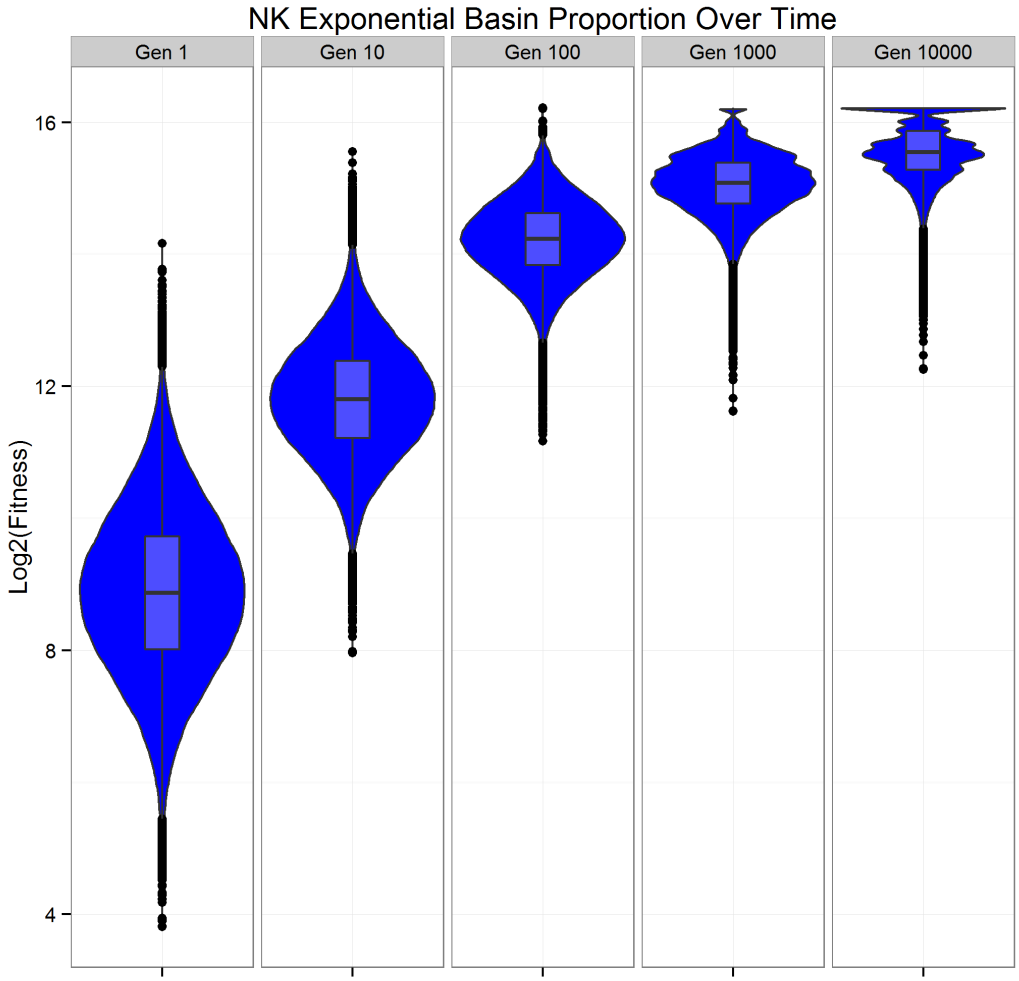


Figure 4.10: Distribution of fitnesses in evolutionary trajectories starting from 100,000 random points in the NK Exponential landscape at each of five exponential time points, as a measure of adaptation in basins over time.



4.4.1.2 Avida

In order to thoroughly test evolutionary trajectories in Avida I used two distinct evolutionary mechanisms: the same GA that I used for NK landscapes and the Avida digital evolution platform itself. The Avida system features continuous reproduction similar to a Moran model with overlapping generations, as opposed to the synchronized generations used in the GA. The two methods both produced strikingly similar results, and so for the purpose of this text, I present the Genetic Algorithm implementation for consistency with the other landscapes.

The results can be seen in Figure 4.11. Again, the distribution shifts clearly towards peaks with higher fitness as more updates elapse. The tiered availability of phenotypes is also highly visible, in contrast to the continuous nature of the other landscapes. This is also visible in the Avida landscape summary in Figure 3.9 in Chapter 3.

Figure 4.11: Distribution of fitnesses in evolutionary trajectories starting from 100,000 random points in the Avida landscape at each of five exponential time points, as a measure of adaptation in basins over time.

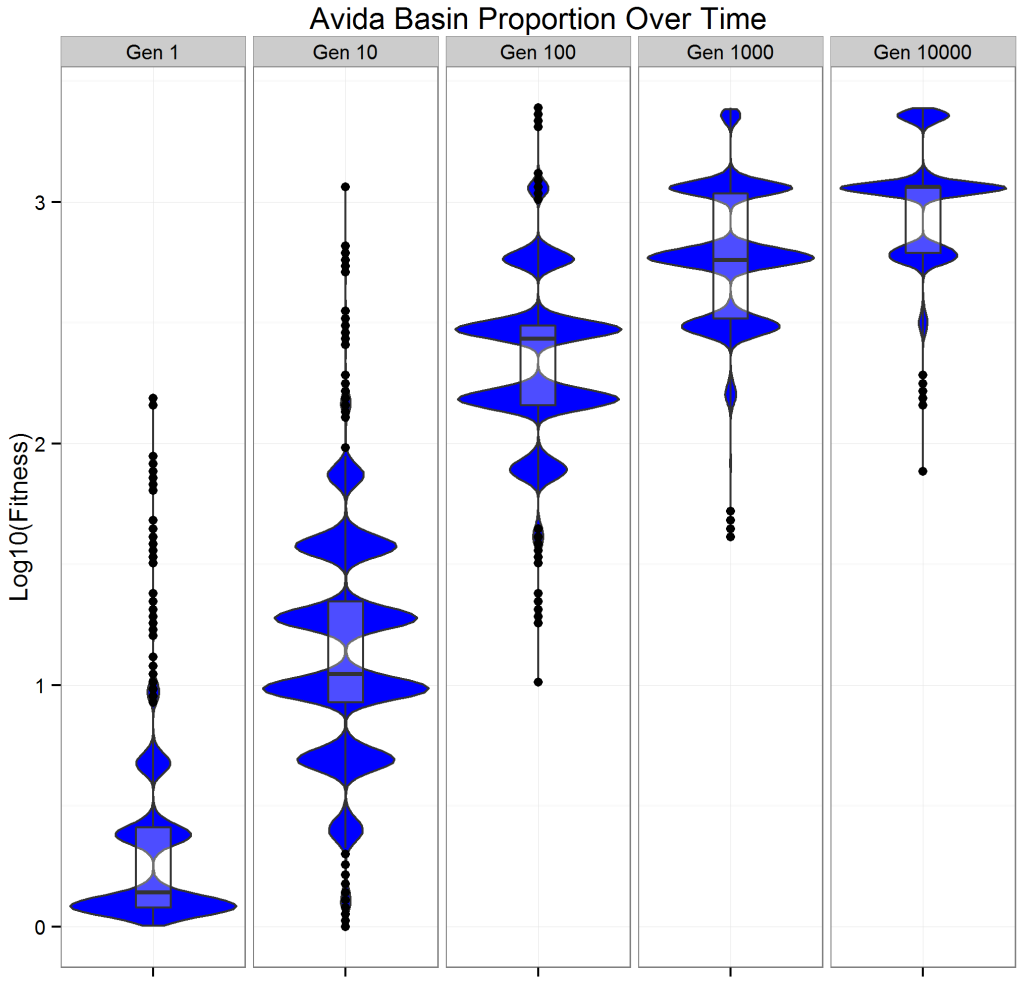
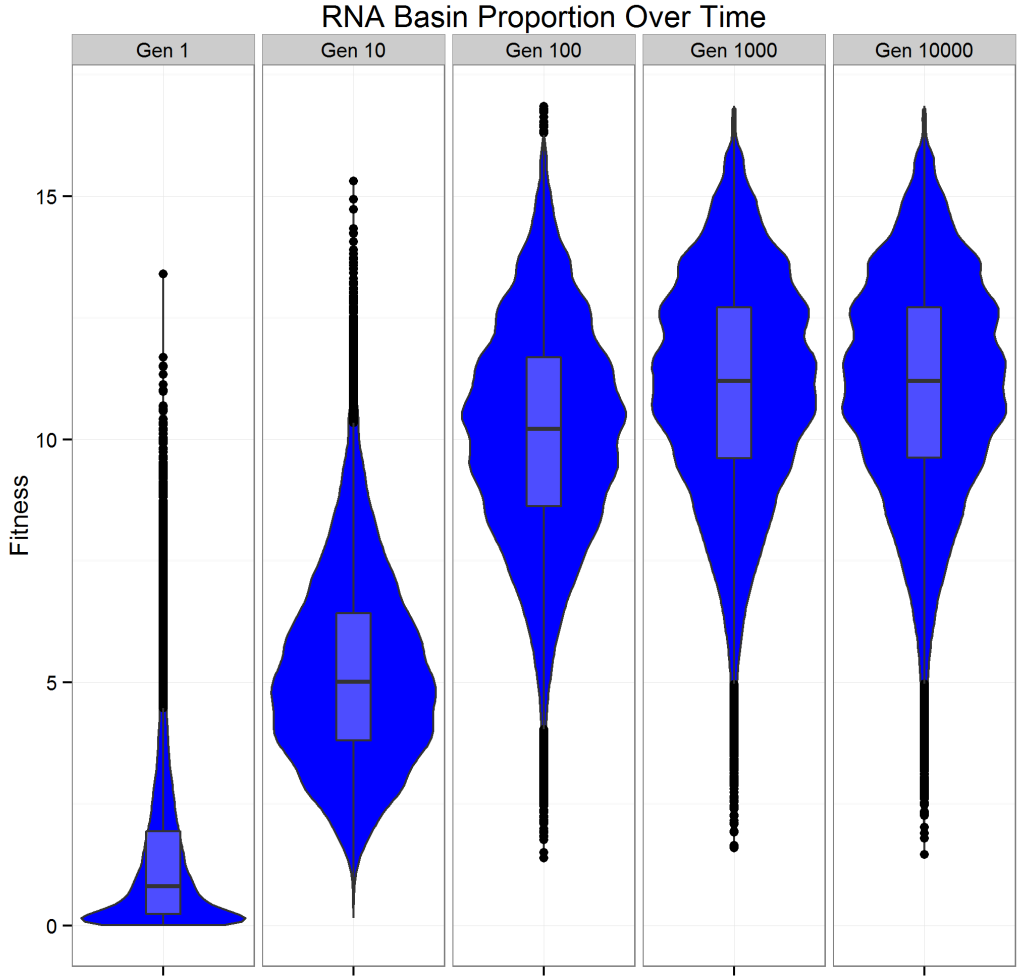


Figure 4.12: Distribution of fitnesses in evolutionary trajectories starting from 100,000 random points in the RNA landscape at each of five exponential time points, as a measure of adaptation in basins over time.



4.4.1.3 RNA

RNA basins over time can be observed in Figure 4.12. The basin fitness seems to stop improving between updates 1,000 and 10,000.

Figure 4.13: Distribution of fitnesses in evolutionary trajectories starting from 100,000 random points in the RNA Exponential landscape at each of five exponential time points, as a measure of adaptation in basins over time.

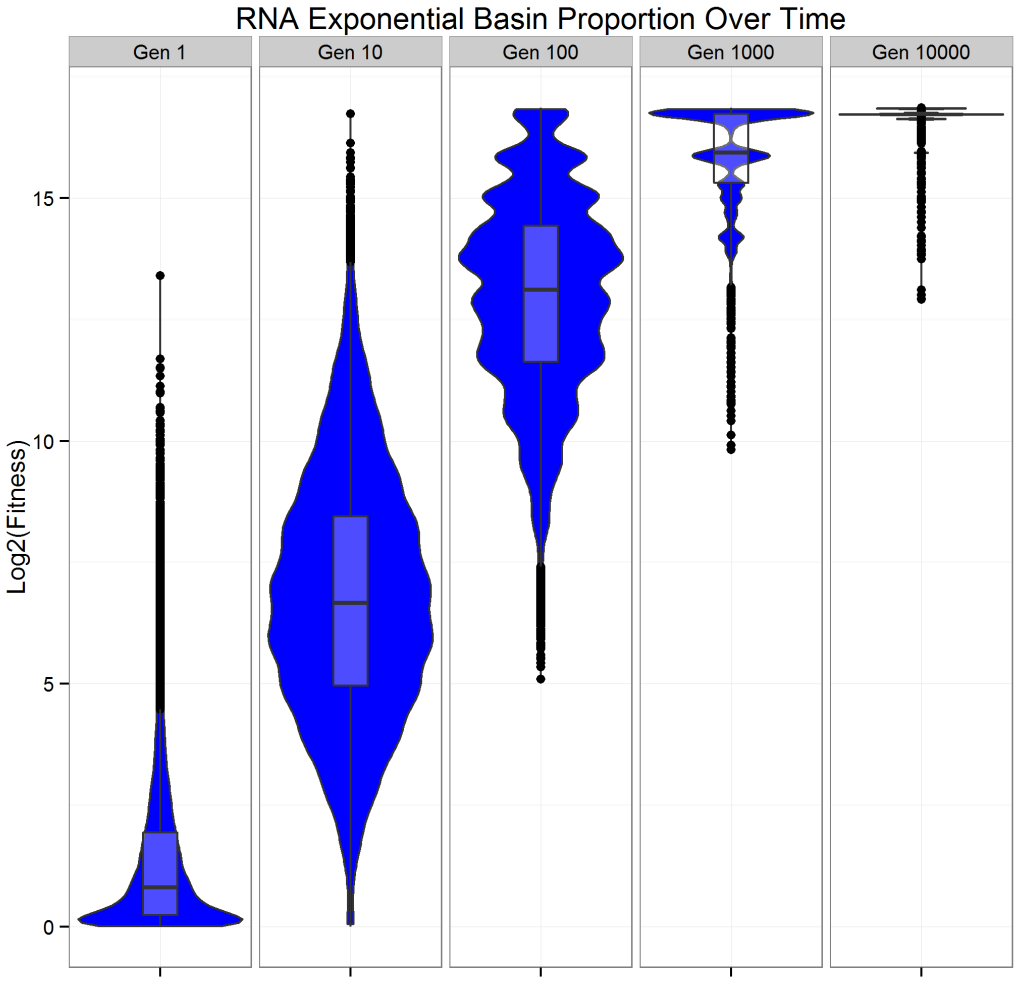
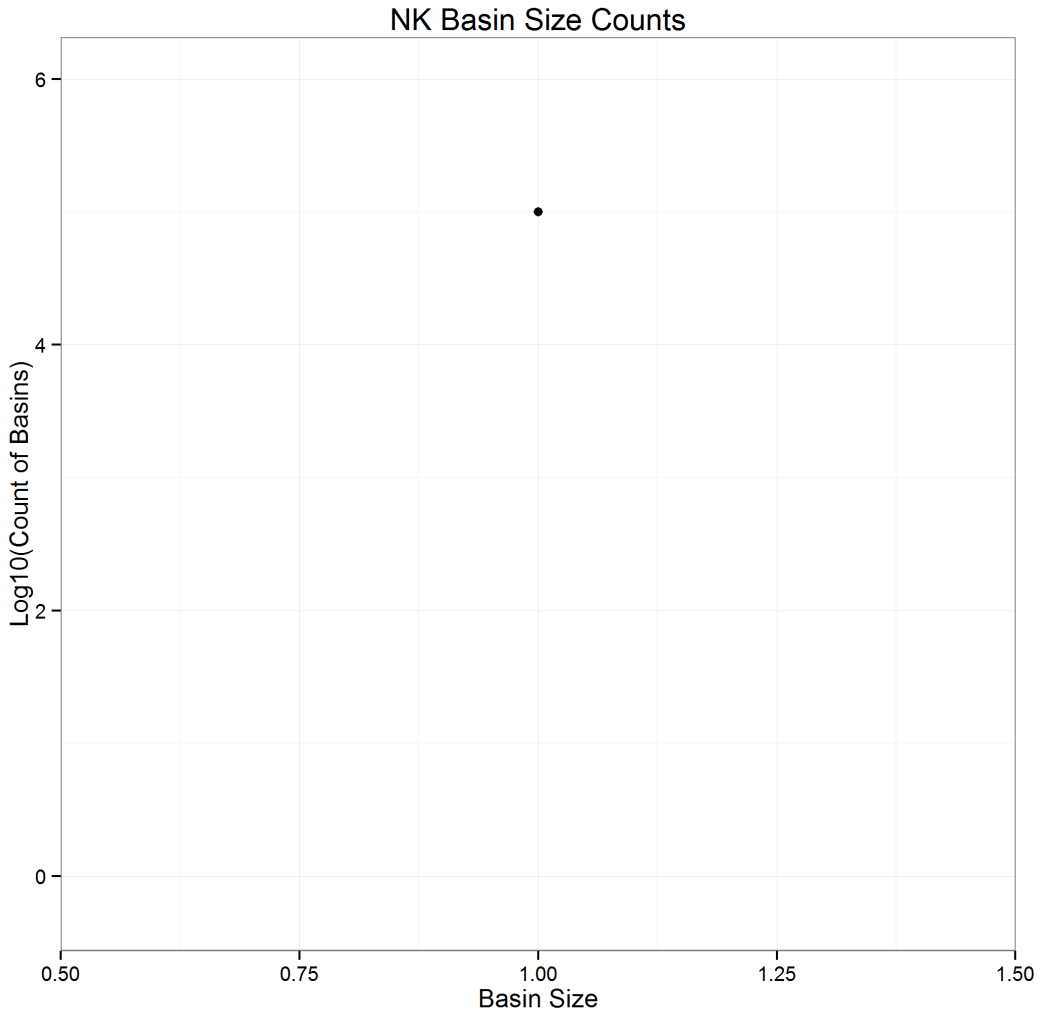


Figure 4.14: NK, 100,000 random points, basin size after 10,000 updates. This plot, as exciting as it appears, is not an accident. There are simply 100,000 unique basins as measured by the endpoints of the trajectories.



4.4.2 Basin Size

The next question I investigated was the basin size after 10,000 updates. The final genotype of the 100,000 trajectories after 10,000 updates serves as a proxy for basin size. This analysis gives us an idea about the distribution of the sizes of the basins of long term evolutionary attractors.

Figure 4.15: NK Exponential, 100,000 random points, basin size after 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked.

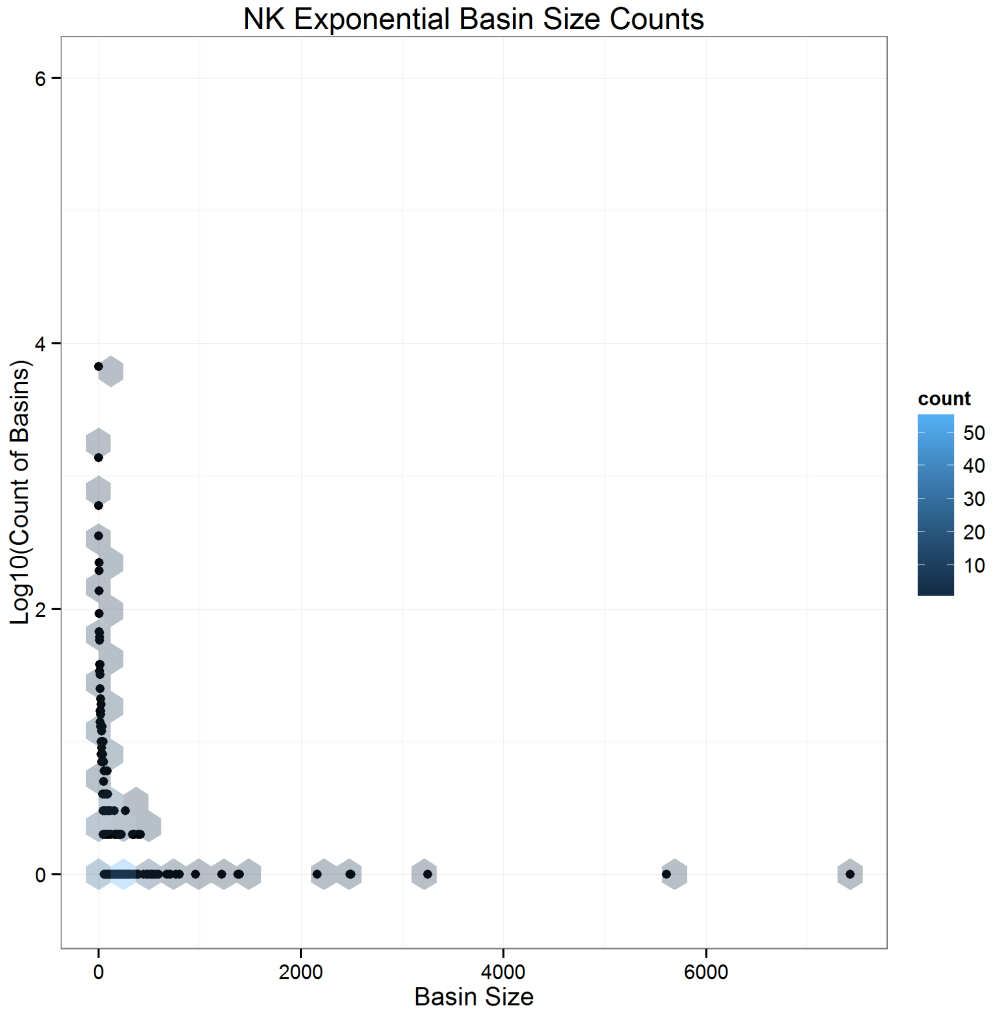


Figure 4.16: Avida, 100,000 random points, basin size after 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked.

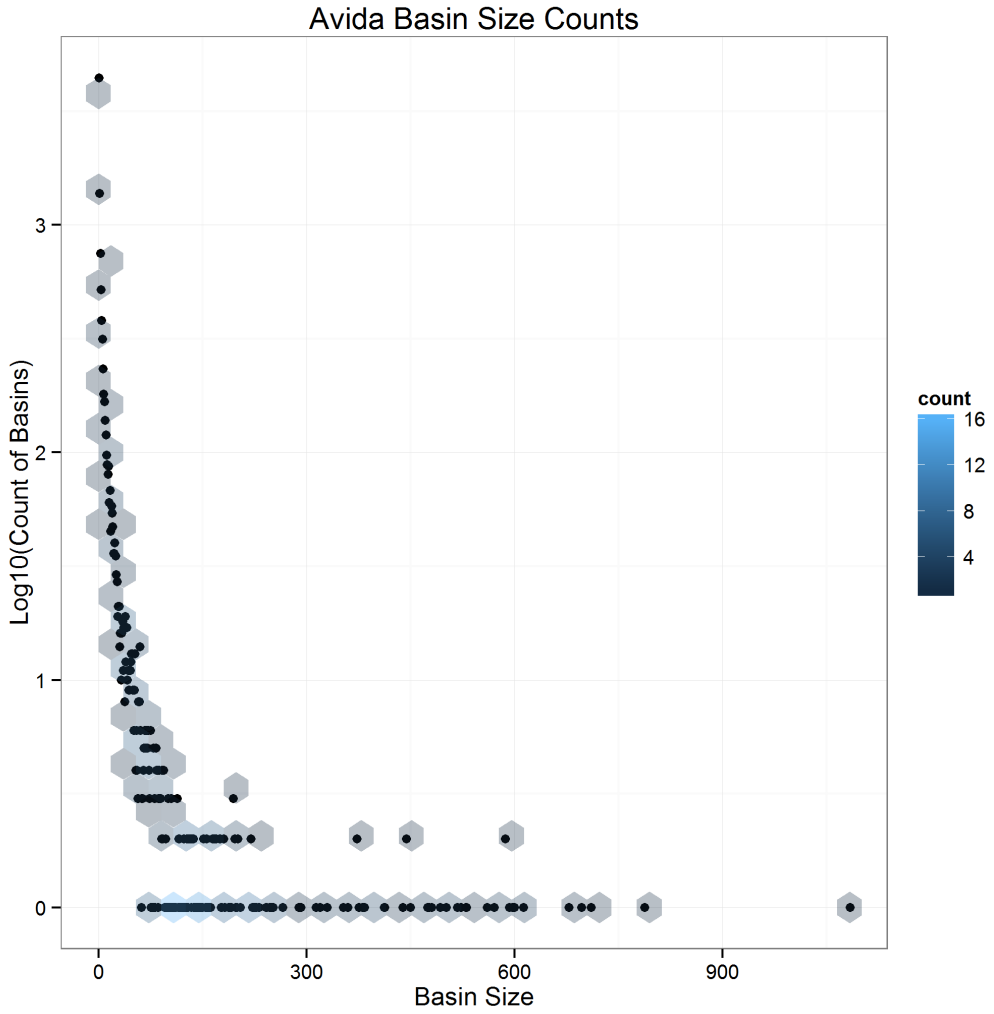


Figure 4.17: RNA, 100,000 random points, basin size after 10,000 updates. Similar to the NK landscape, most points are not accumulating in basins; over 98,000 trajectories ended in unique genotypes, with the rest accumulating in basins of size no more than four.

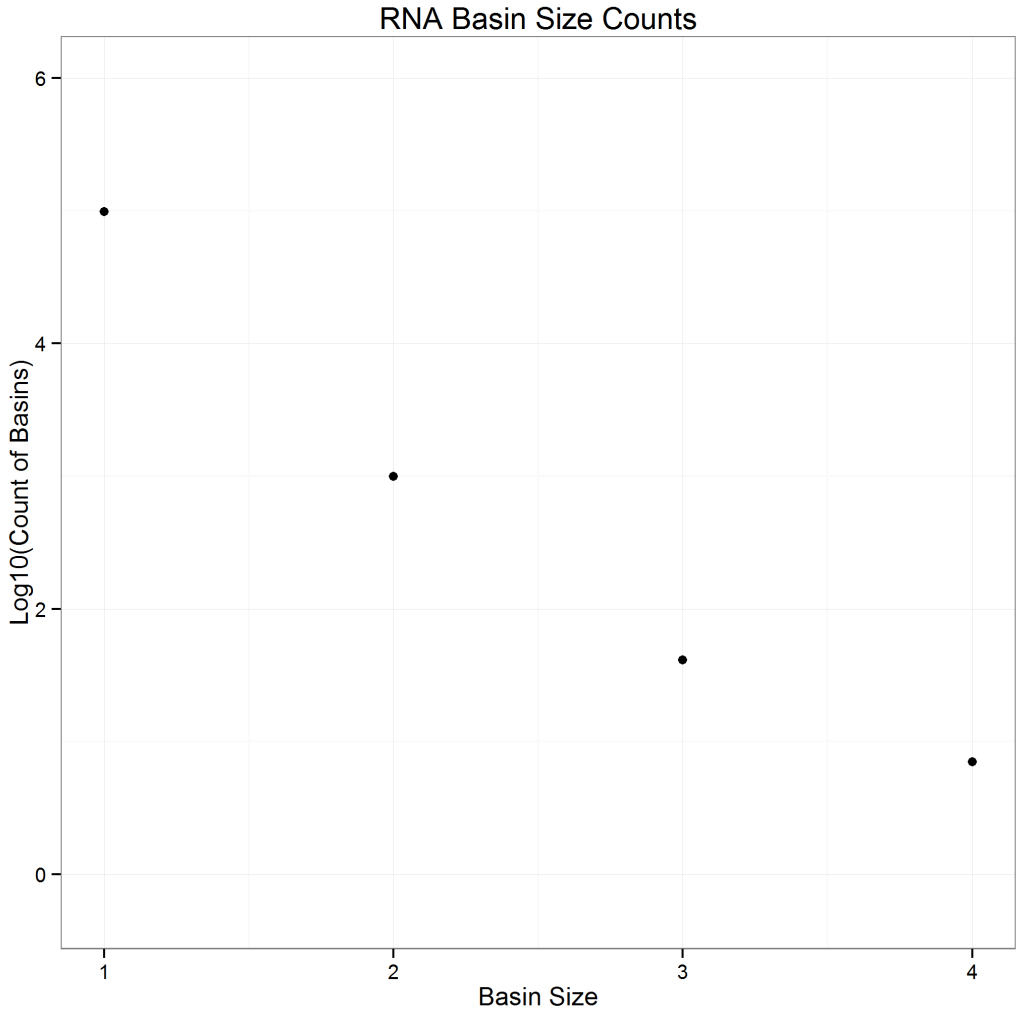
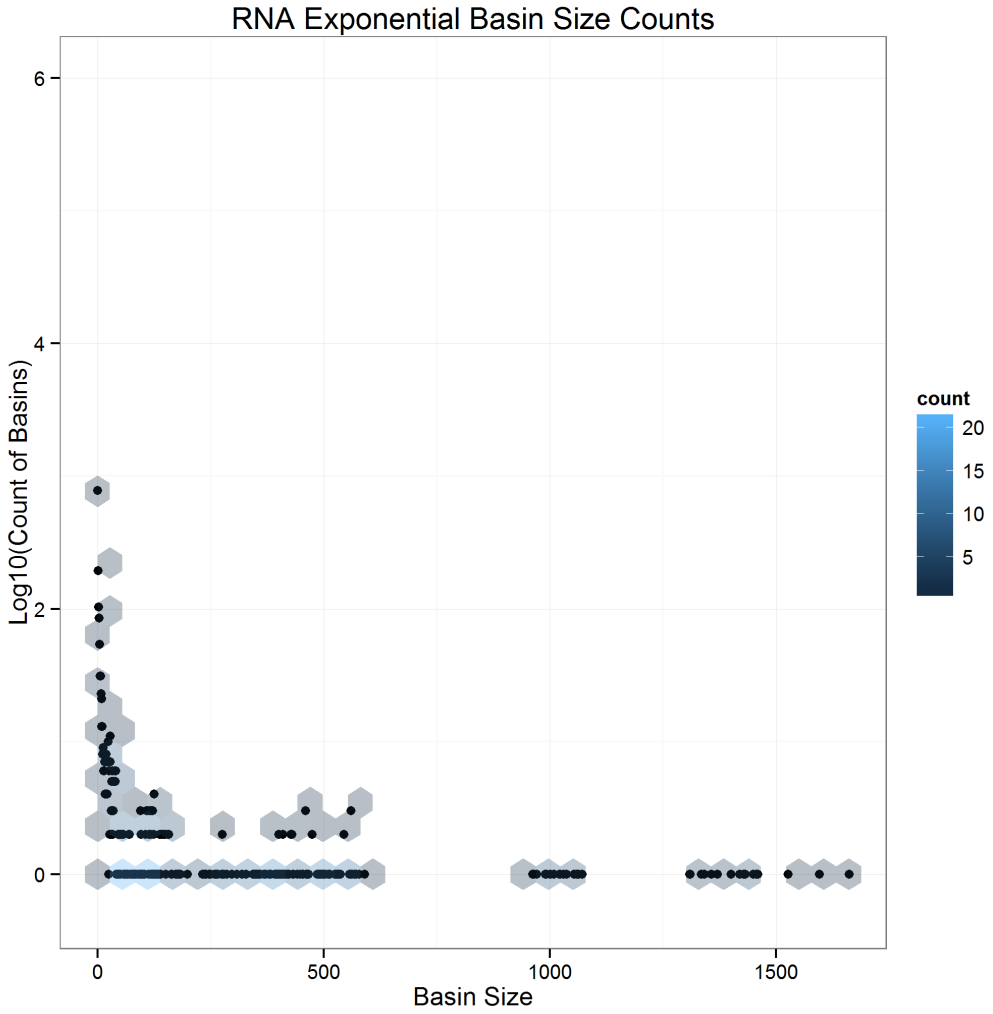


Figure 4.18: RNA Exponential, 100,000 random points, basin size after 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked.



4.4.3 Basin Size Relation With Fitness

Previous work on the NK landscape including that in 4.3 has shown an exponential relationship between basin fitness and size. We seek to investigate whether this relationship holds in larger landscapes when evolution is the main determinant of basins. Here I use the trajectory count as a proxy for basin size.

Figure 4.19: NK, 100,000 random points, basin size vs basin fitness at 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked. The red line and equation represent the exponential curve fit. Due to weak selection strength, there is not much relationship between the fitness and basin size.

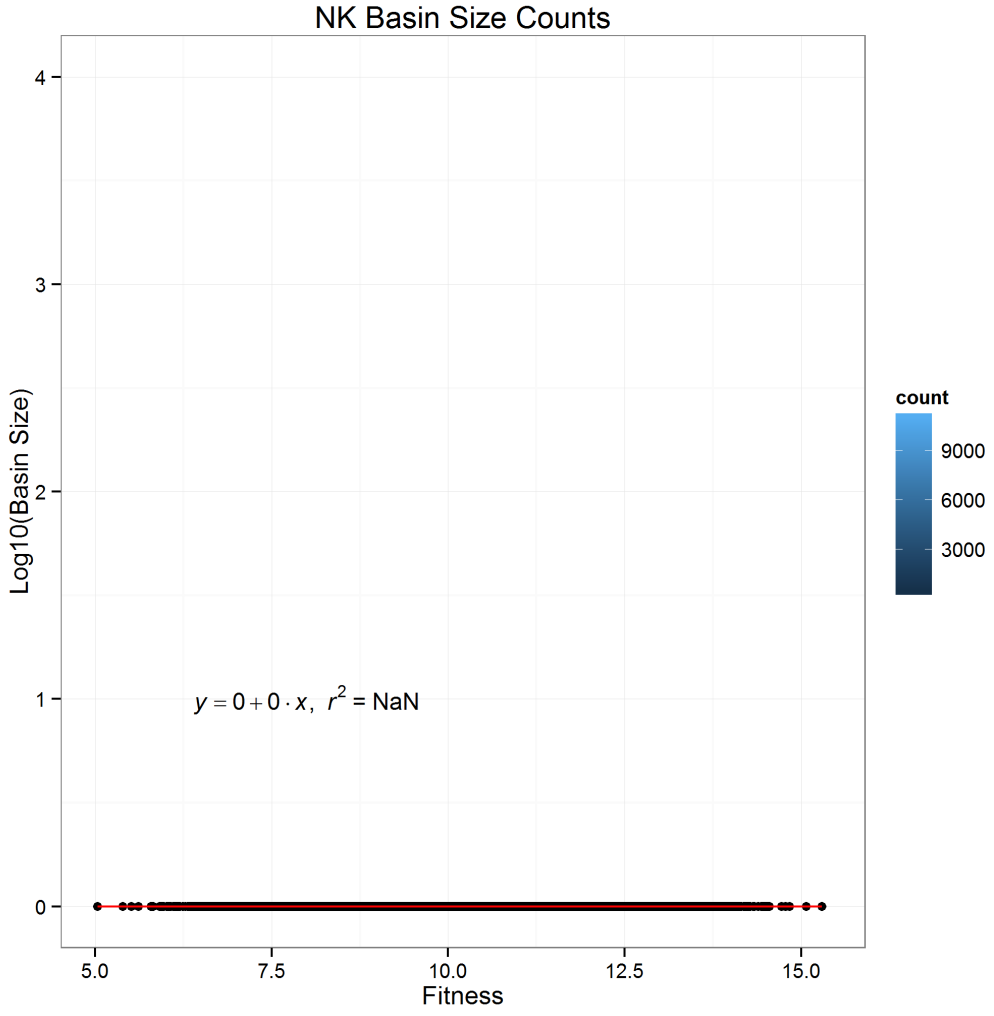


Figure 4.20: NK Exponential, 100,000 random points, basin size vs basin fitness at 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked. The red line and equation represent the exponential curve fit.

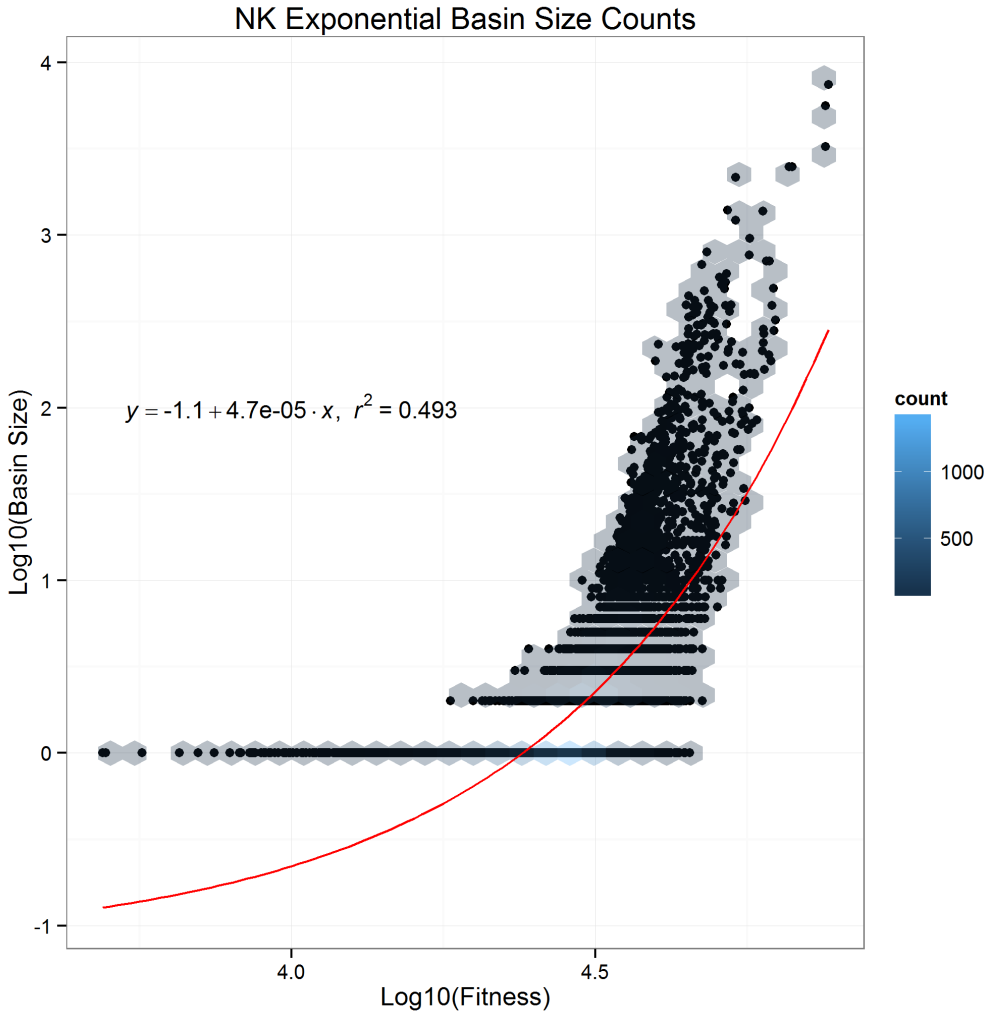


Figure 4.21: Avida, 100,000 random points, basin size vs basin fitness at 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked. The red line and equation represent the exponential curve fit. Avida seems to have a positive relationship with fitness but the fit is very poor.

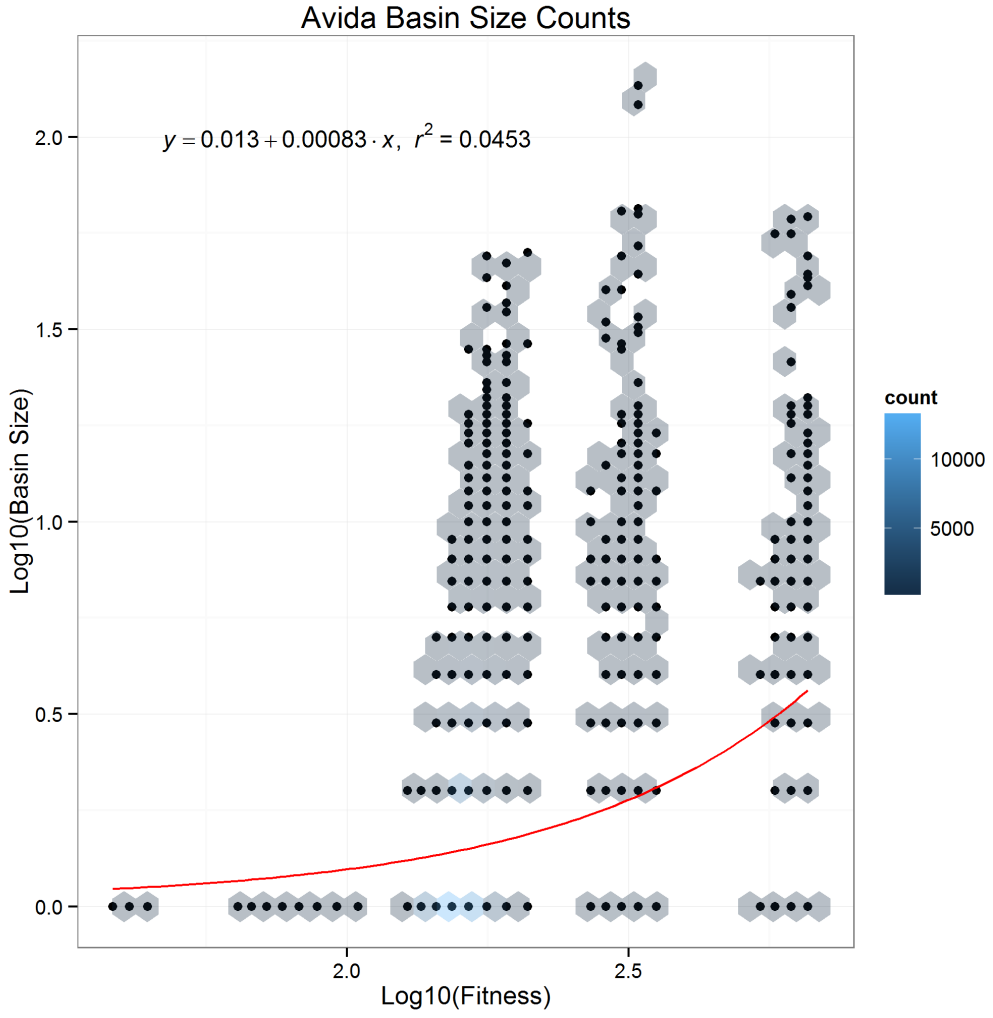


Figure 4.22: RNA, 100,000 random points, basin size vs basin fitness at 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked. The red line and equation represent the exponential curve fit. Due to weak selection strength, there is not much relationship between the fitness and basin size.

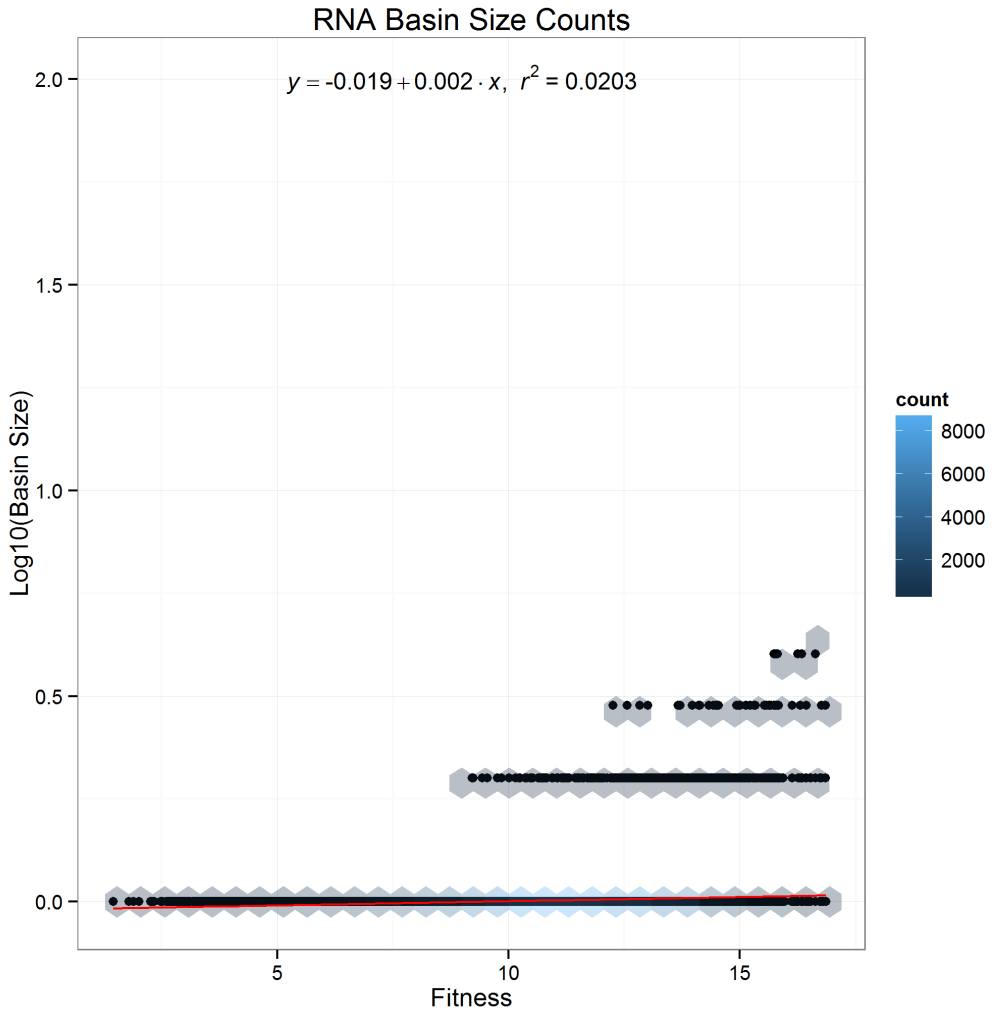


Figure 4.23: RNA Exponential, 100,000 random points, basin size vs basin fitness at 10,000 updates. The colored hexes represent binnings of points and are used to delineate areas of high density, where many points may be stacked. The red line and equation represent the exponential curve fit.

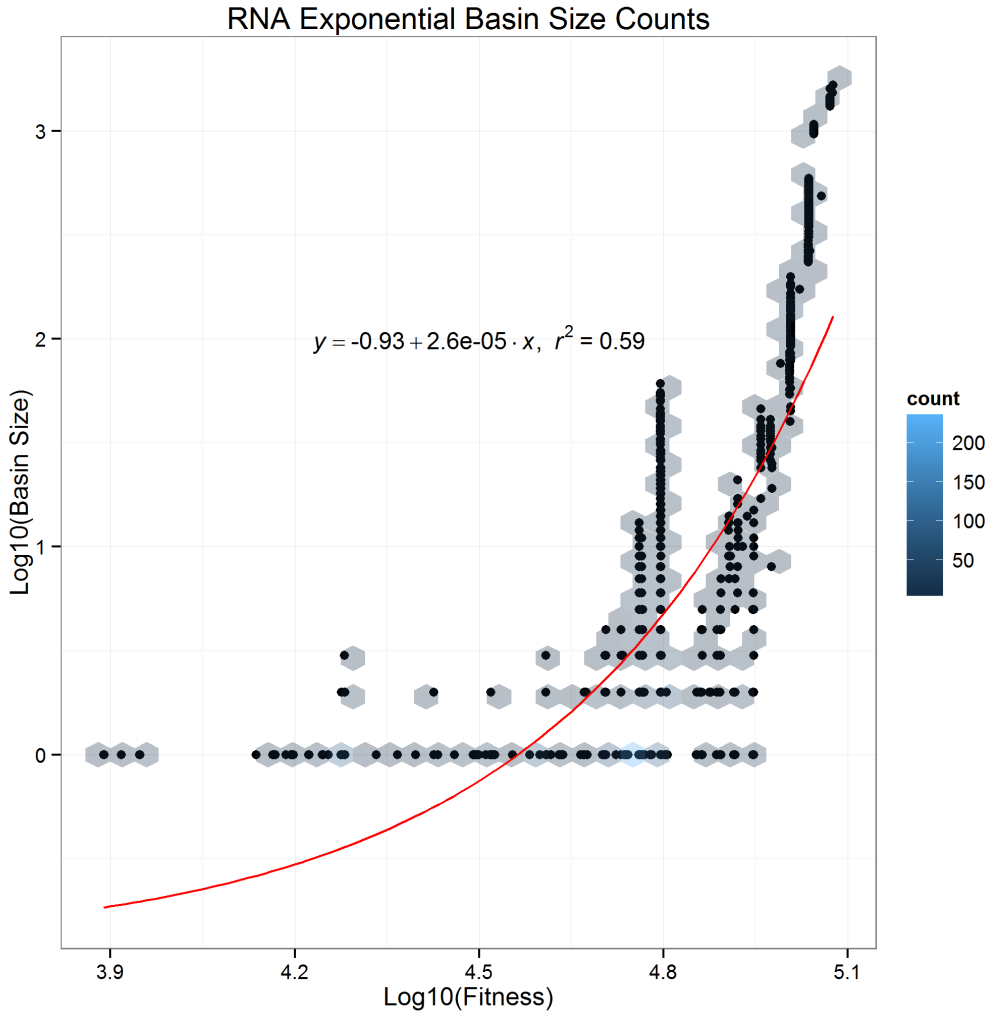
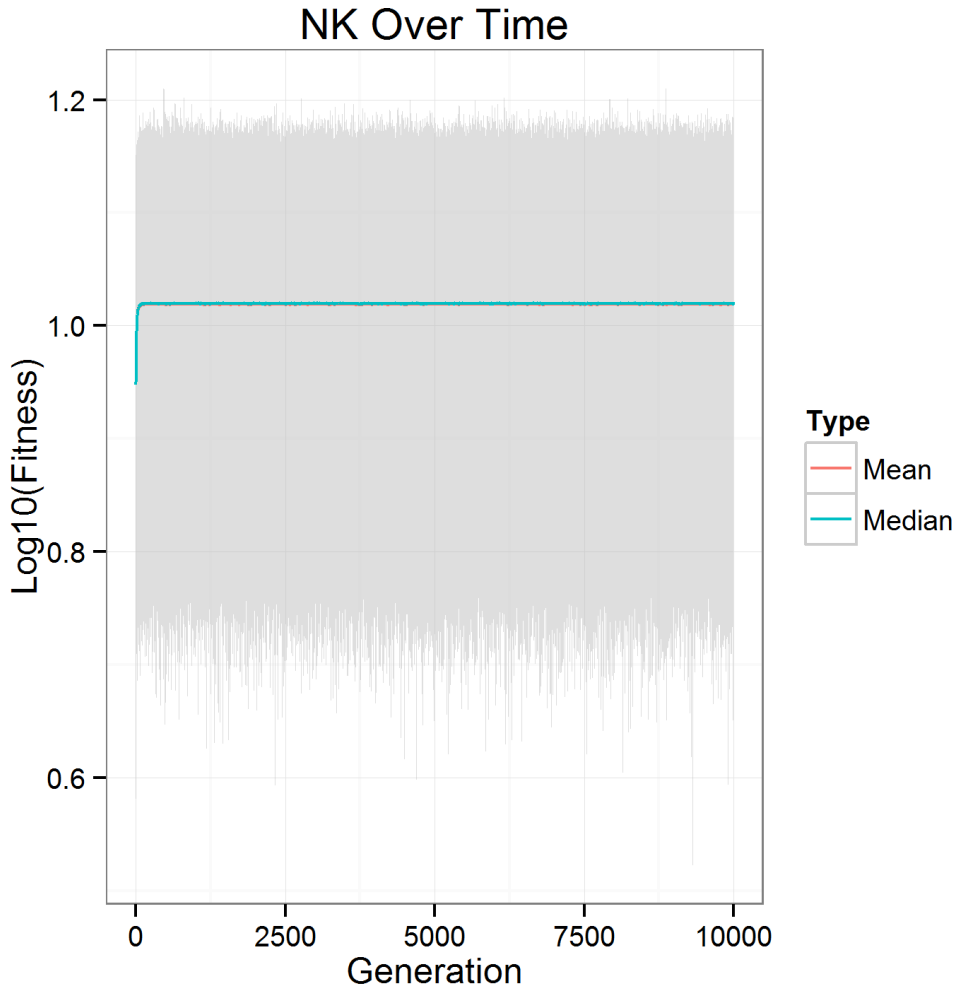


Figure 4.24: NK, fitness for trajectories from 100,000 starting random points, fitness over 10,000 updates. In gray is the area between the 5th and 95th quantiles.



4.4.4 Fitness Over Time

This section shows the fitness over time of 100,000 trajectories starting at random points in each of the environments. These are useful mostly to visualize the adaptation of the trajectories and understand on what time scales the trajectories experience improvement. Not surprisingly, fitness increases over time, but in the RNA and NK environments, it levels off and stops improving noticeably, whereas in the Avida, RNA Exponential, and NK Exponential environments, the period during which fitness increases in aggregate is larger.

Figure 4.25: NK Exponential, fitness for trajectories from 100,000 starting random points, fitness over 10,000 updates. In gray is the area between the 5th and 95th quantiles.

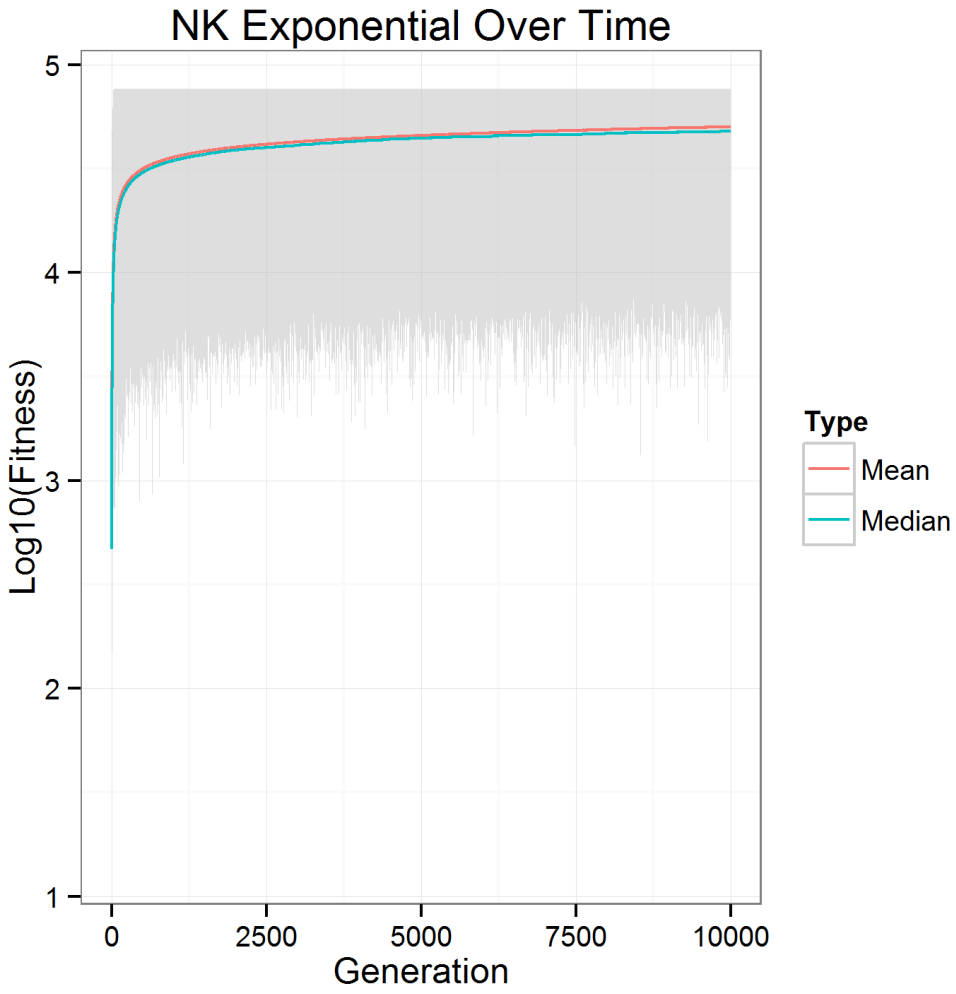


Figure 4.26: Avida, fitness for trajectories from 100,000 starting random points, fitness over 10,000 updates. In gray is the area between the 5th and 95th quantiles.

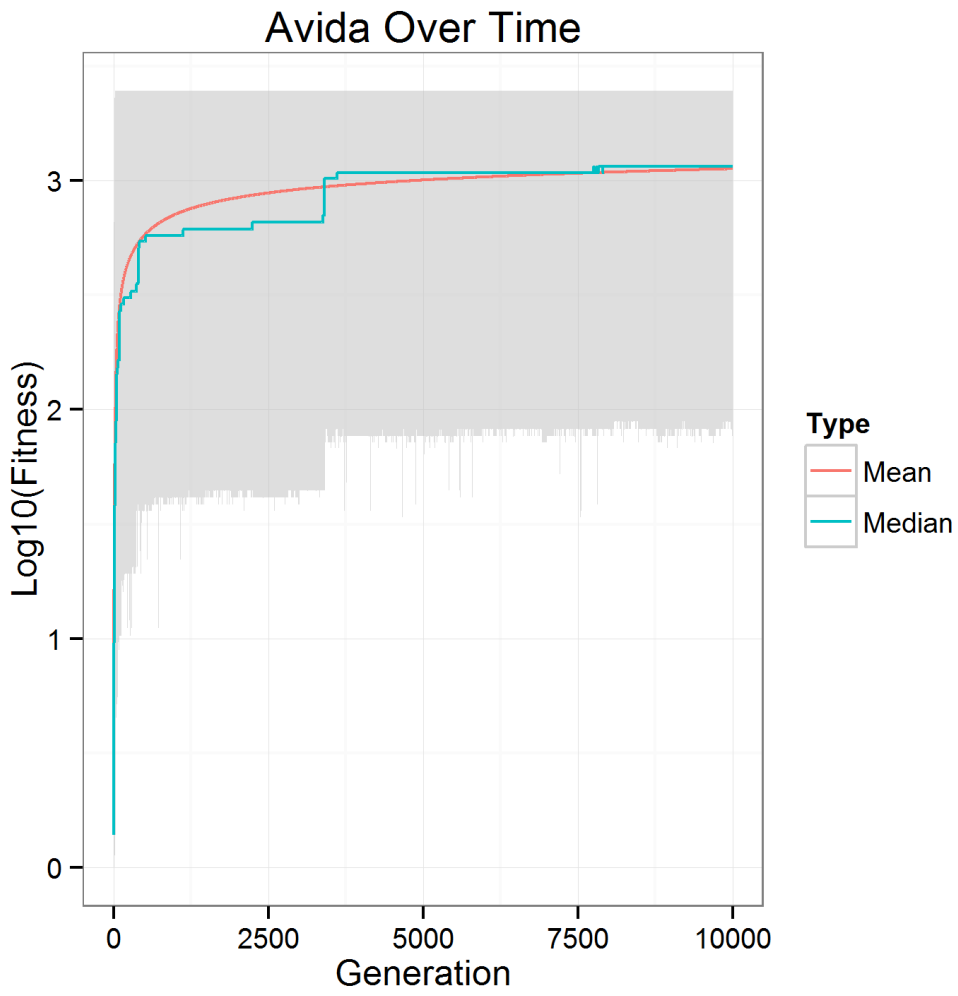


Figure 4.27: RNA, fitness for trajectories from 100,000 starting random points, fitness over 10,000 updates. In gray is the area between the 5th and 95th quantiles.

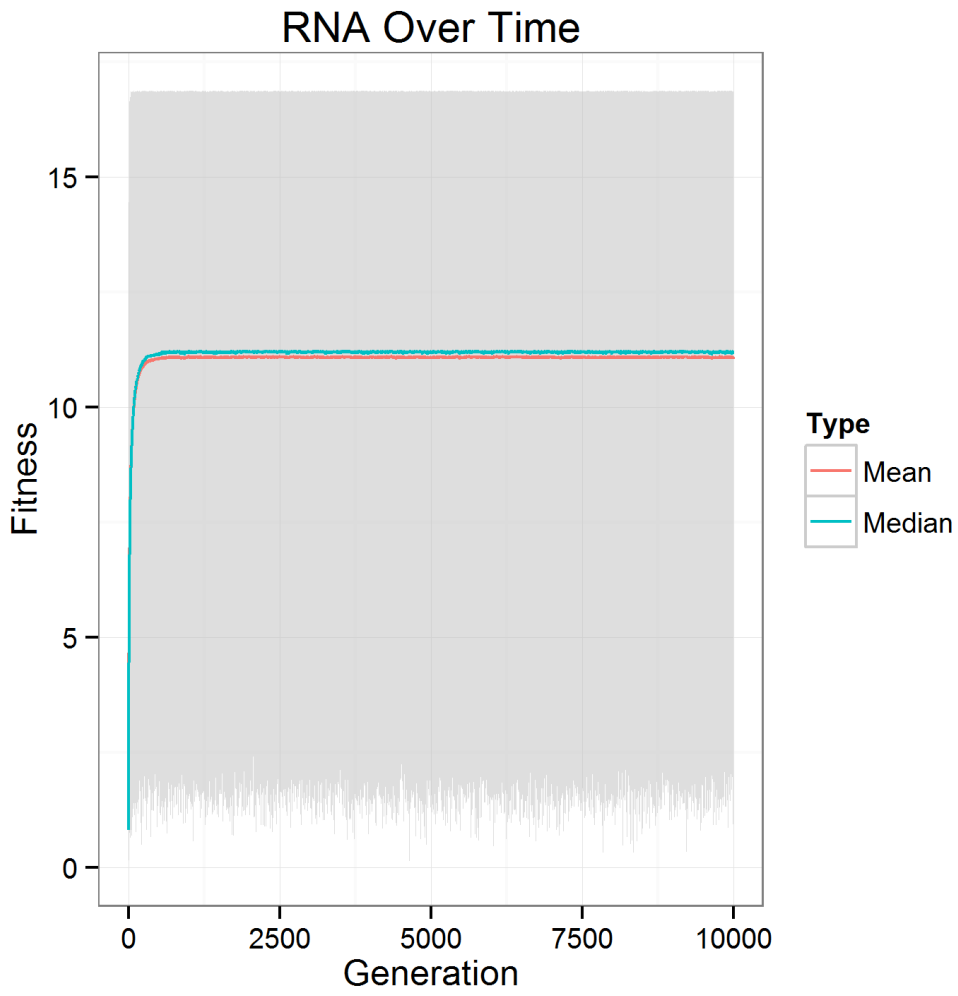


Figure 4.28: RNA, fitness for trajectories from 100,000 starting random points, fitness over 10,000 updates. In gray is the area between the 5th and 95th quantiles.

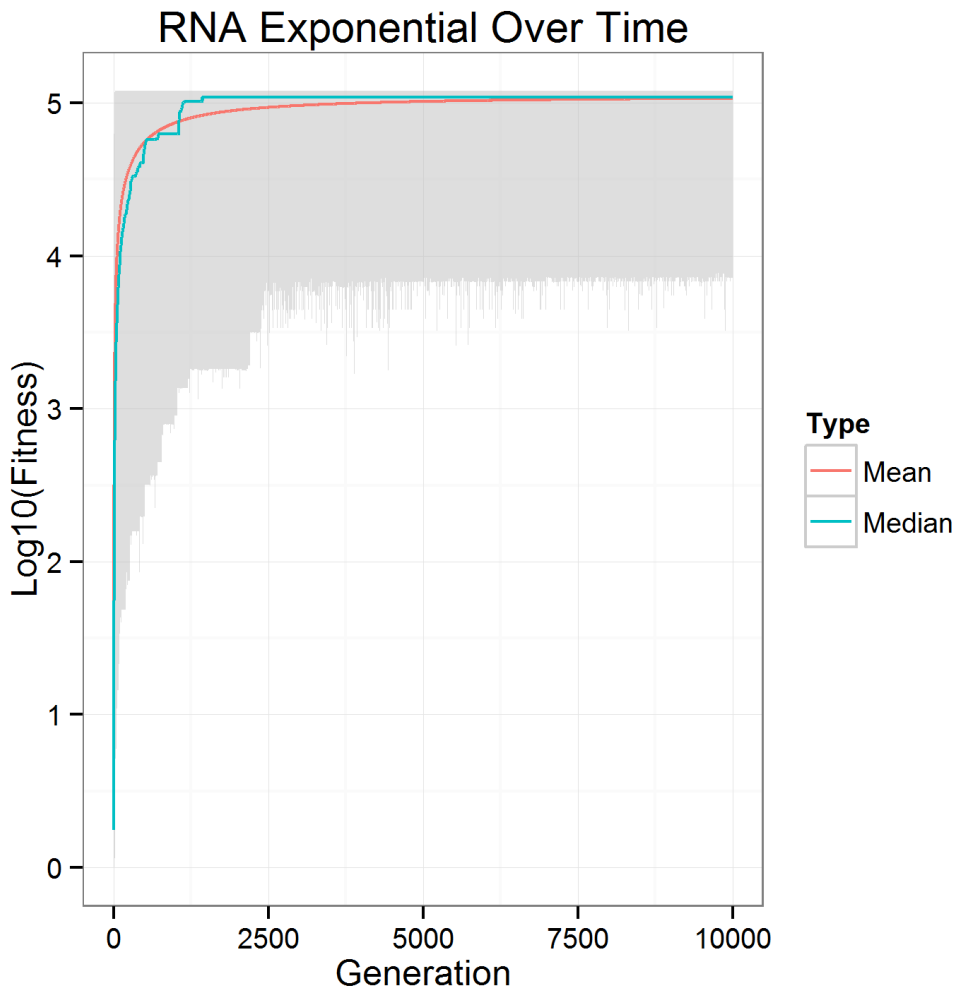
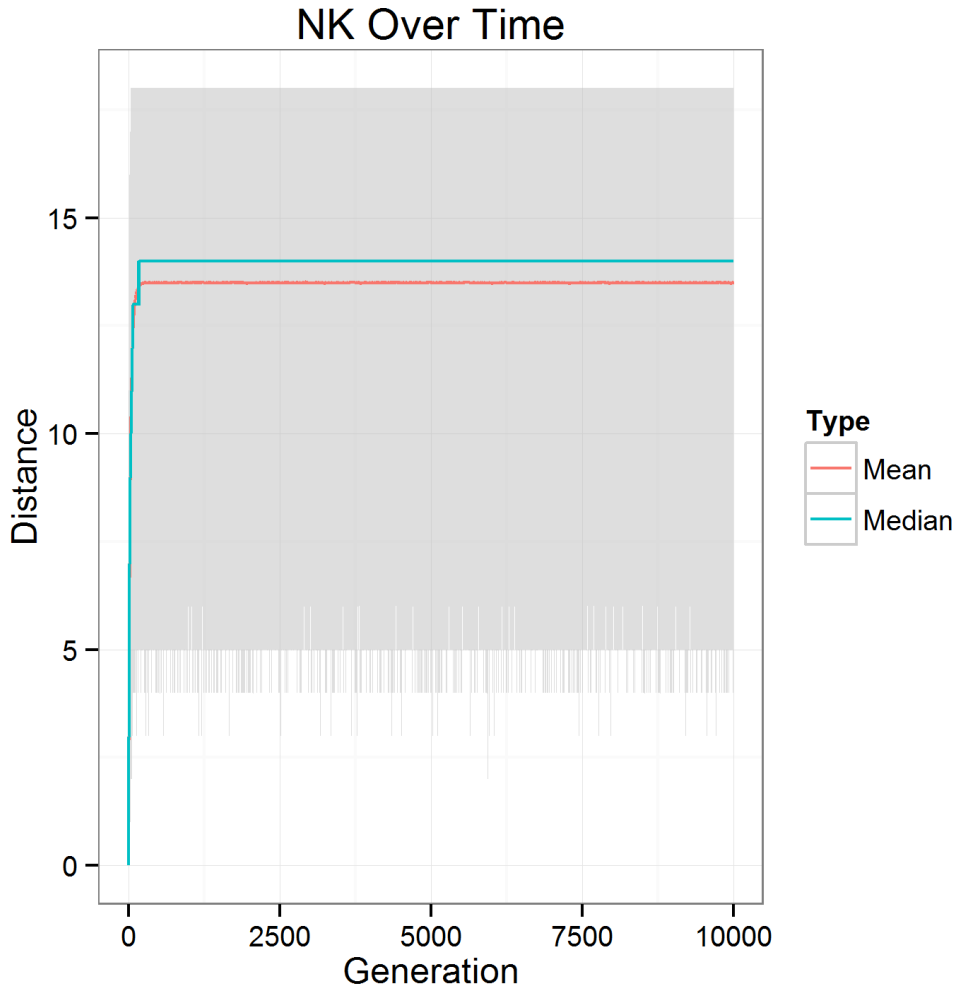


Figure 4.29: NK, distance from origin for trajectories from 100,000 starting random points. In gray is the area between the 5th and 95th quantiles.



4.4.5 Distance Traveled From Origin

This section shows distance traveled from the original genotype as time elapses. I measure this in order to understand how far points travel on average to their final destinations relative to their origin in the landscape. This is a proxy method for measuring the size and reach of basins.

Figure 4.30: NK Exponential, distance from origin for trajectories from 100,000 starting random points. In gray is the area between the 5th and 95th quantiles.

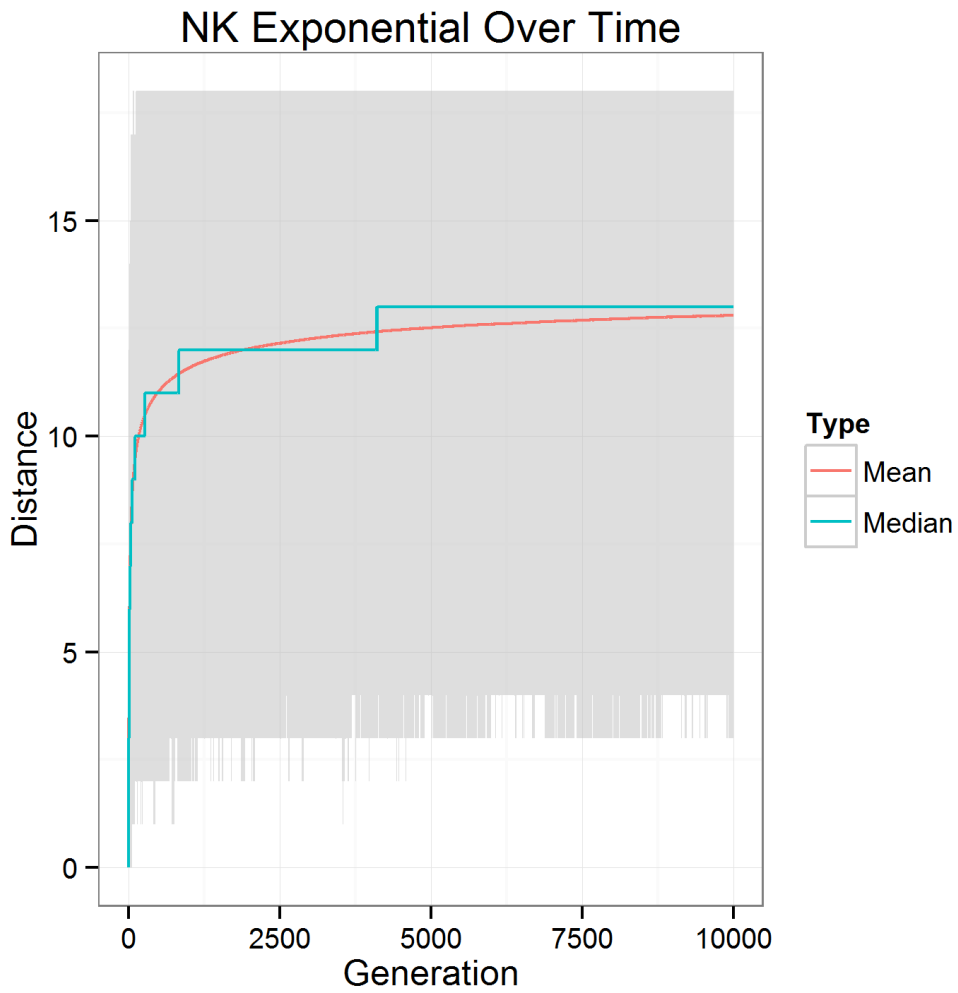


Figure 4.31: Avida, distance from origin for trajectories from 100,000 starting random points. In gray is the area between the 5th and 95th quantiles.

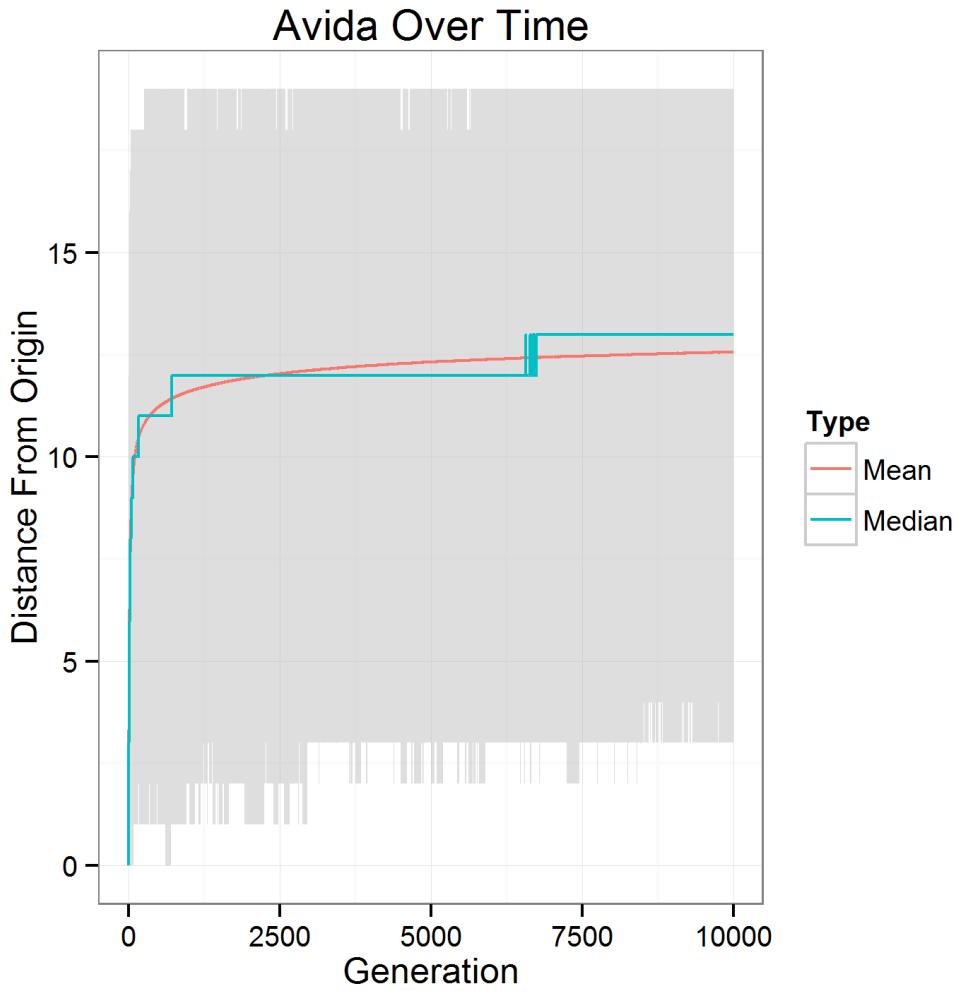


Figure 4.32: RNA, distance from origin for trajectories from 100,000 starting random points. In gray is the area between the 5th and 95th quantiles.

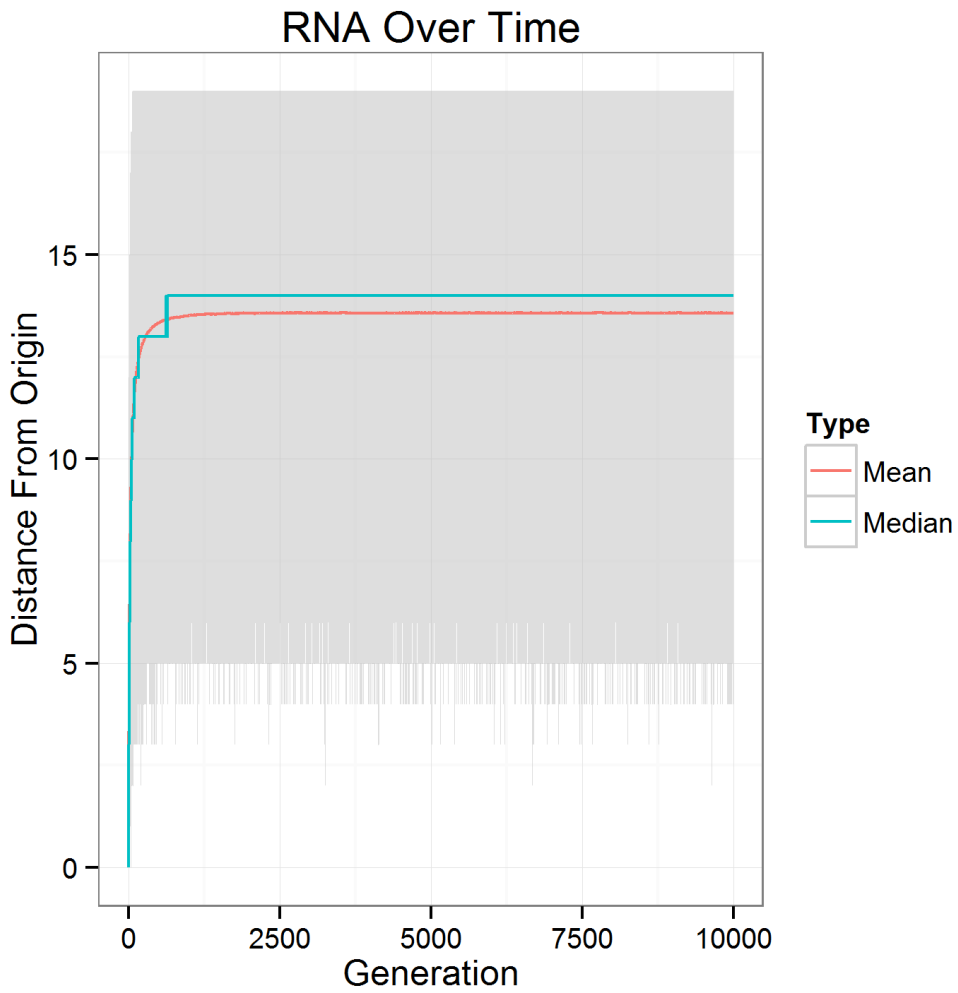
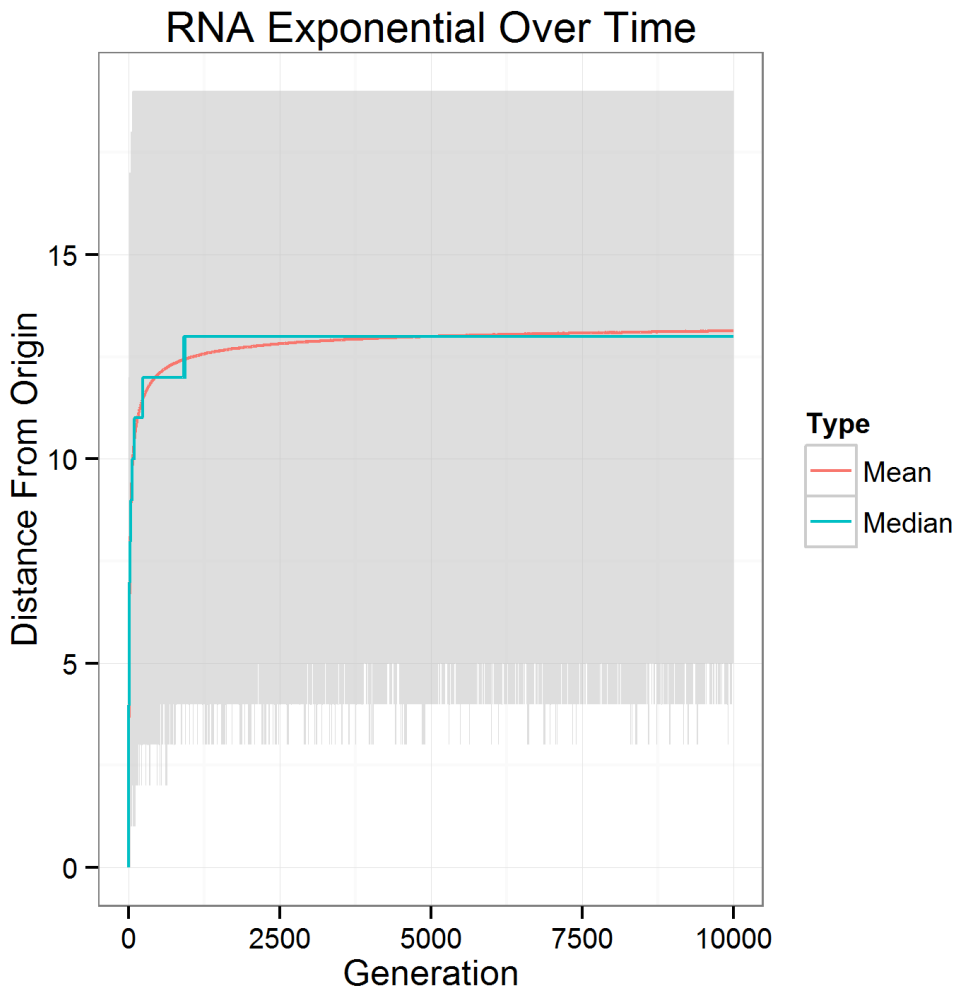


Figure 4.33: RNA Exponential, distance from origin for trajectories from 100,000 starting random points. In gray is the area between the 5th and 95th quantiles.



4.4.6 Peaks & Basins

In this section, I look at the proportion of each trajectory at each timestep that is a 'peak' as identified in the fitness landscapes in the last section. Note that for the exponential treatments, the peaks remain the same since the exponential treatment alters the strength of relative signal, but not the fitness valence relationship between neighboring points.

Figure 4.34: NK, proportion of trajectories from 100,000 starting random points currently in genotypes identified as peaks.

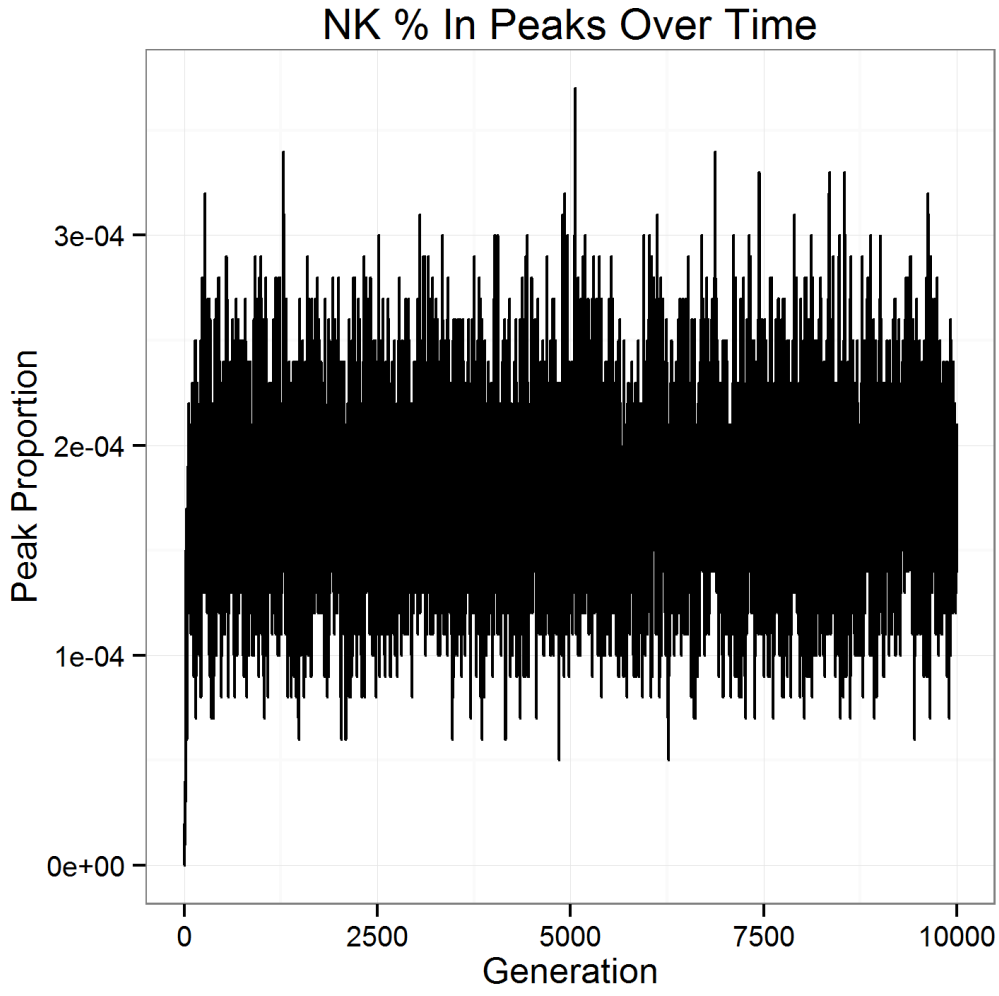


Figure 4.35: NK Exponential, proportion of trajectories from 100,000 starting random points currently in genotypes identified as peaks.

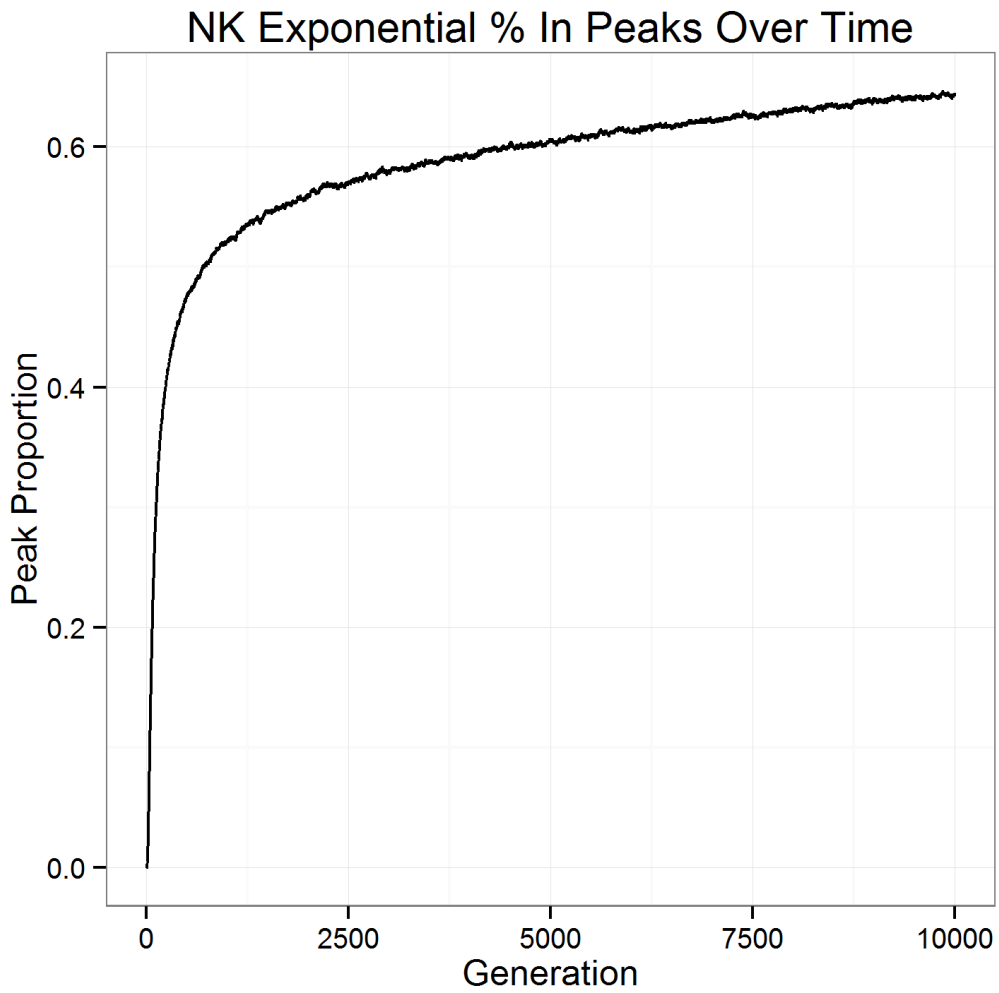


Figure 4.36: Avida, proportion of trajectories from 100,000 starting random points currently in genotypes identified as peaks.

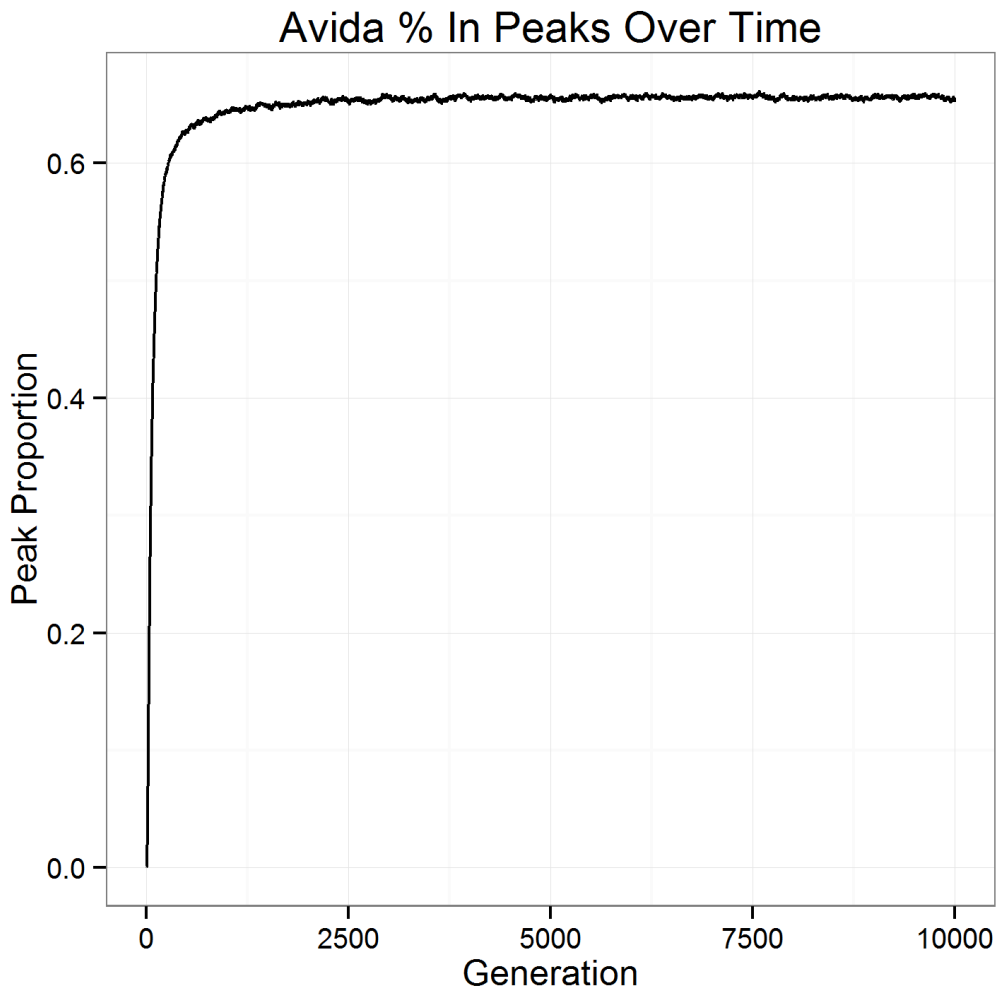


Figure 4.37: RNA, proportion of trajectories from 100,000 starting random points currently in genotypes identified as peaks.

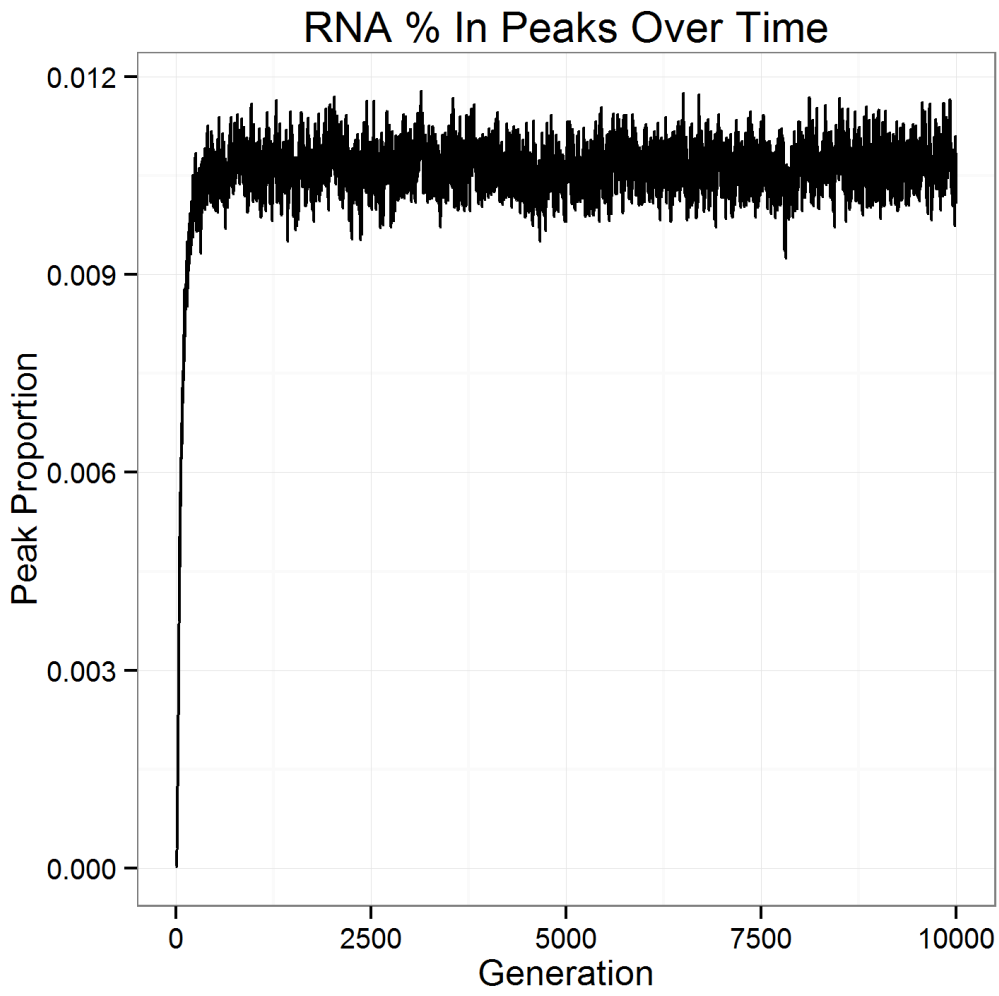
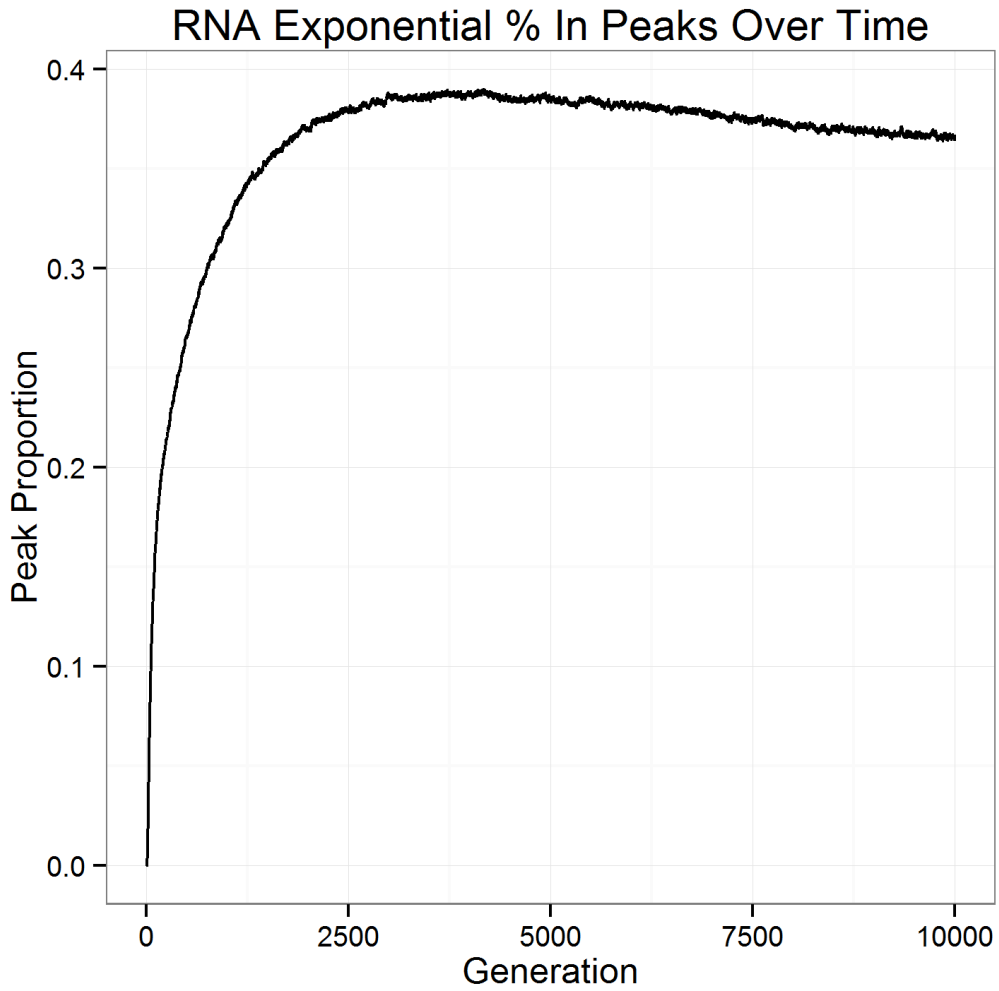


Figure 4.38: RNA Exponential, proportion of trajectories from 100,000 starting random points currently in genotypes identified as peaks.



4.4.7 Discussion

I have presented an investigation of basin size in the three model landscapes introduced in Chapter 3—the NK landscape, the Avida landscape, and the RNA landscape. For the RNA and the NK landscapes, the selection strength was insufficient to obtain a meaningful understanding of the landscape with a population size of 25. I therefore chose to increase the selection strength by remapping them with the fitness transformation $x = 2^x$. This simple transformation preserves the peak structure in the landscape and the relationship between nearby points and makes it easier for small populations to climb gradients and stay on peaks. I present both this as well as 100,000 trajectories on the original untransformed landscape. I did not perform this treatment for the Avida landscape since it already has an exponential form, and a snapshot of basin structure successfully emerged from my sampling of evolutionary trajectories.

All three landscapes exhibited increased fitness over time in the distribution of samples, as might be expected. Early improvements along the line of descent allowed for rapid increases in fitness, but eventually the rate of improvement slows down, although it never stops over the measured interval. The distribution of the endpoints of each run is landscape specific—the sparsity of phenotypes is visible in the Avida landscape, the normality of the fitnesses of peaks in the NK landscape is highly apparent and seems to be preserved through time. The exponential RNA and NK landscapes show continued improvement in marked contrast to the stagnation visible after roughly Update 100 in their non-exponential RNA and NK cousins. This difference in behavior is due to the strength of selection. I focus on the exponential landscapes for the rest of the discussion, but it is worth noting that transforming the landscape exponentially improved the solution quality for this particular parameter set quite drastically. In evolutionary computation applications, small populations and low mutation rates may benefit strongly from increased selection strength via transformation, such as my exponential treatment, in that there may be better final solution quality, but this may also lead to adverse

effects such as being trapped in local minima.

Next, I examined the frequency of each basin size. The NK and RNA landscapes both did poorly here—Figure 4.14 for instance shows the result that there were 100,000 basins of size one—i.e. every trajectory went its own way. The RNA landscape was very similar with over 98,000 basins of size one, and a handful of size two, three, and four basins. The exponential landscapes told a different story, however. In Avida, there were fewer than 10,000 unique basins of size one, which meant that the vast majority of measured basins had at least two trajectories and there was at least one with 1000 trajectories. Since these were unbiased random samples, these results indirectly point to genotypes in the landscape that have basins that can attract 1% of the genotypes on the landscape. Likewise in RNA and NK, the exponential treatments revealed similarly large attractors, with RNA having a point with about 1.5% of the trajectories and NK having a maximum point attracting roughly 6%.

I measured the interaction of basin fitness and basin size. Previous related work has found an exponential relationship in NK landscapes between basin fitness and basin size. While it is debatable whether the exponential relationship is the most appropriate one—it does not seem to apply strongly to the Avida landscape, but does do reasonably well in the exponential NK and RNA landscapes. Nevertheless, the more general assertion that higher peaks have bigger basins seems to be true.

Distance in all three landscapes was also of note. The average Hamming distance of a genotype to all other genotypes is 13.5, and the numbers in all three landscapes quickly approach that. In Avida, the mean distance from the origin at 10,000 updates was 12.565, with the 5% quantile being 3 and the 95% quantile at 19. In RNA Exponential, the mean distance was 13.135 after 10,000 updates, with the 5% quantile being 5 and the 95% quantile at 19. And finally in the NK Exponential, the mean distance was 12.805, with the 5% quantile at 3 and the 95% quantile at 18. These results seem to imply that, on average, the starting point in these landscapes has little relationship with the final destination—approaching average

distance over evolutionary time. Basins over evolutionary time are thus not the closest local optimum, as might be expected. One interpretation of these data is that a single genotype may go into several basins, with historical contingency determining which one is the final destination. This is the hypothesis I prefer since it does not impose other requirements on the structure of the environment. A competing hypothesis is that the basins themselves have long and narrow tendrils as opposed to being regular depressions with a certain radius.

I also looked at the fraction of trajectories currently in neutral peaks over time, where neutral peaks are defined as points without one mutant neighbors that have higher fitness. Throughout the 10,000 updates, the evolutionary trajectories in aggregate spent a large proportion of time in peaks, regardless of landscape. All three landscapes experienced a large boost in the first hundred updates, while presumably the process of evolution was locating easily accessible innovations. After 100 updates, however, things tended to level off, and there were qualitative differences between the landscapes in long term percentages in peaks; in the NK Exponential landscape, the proportion in peaks slowly increased over time, in Avida, it stayed more or less static, and finally in the RNA Exponential landscape, it decreased. The RNA landscape's behavior on this metric in particular is slightly puzzling—there may be other mechanisms at play such as survival of the flattest or it may be our definition of peaks that causes this unexpected slow decline over time. Nevertheless, it is clear that peaks as I defined them play an important role throughout evolutionary time as trajectories naturally find them and stay close to them when the selection strength is sufficient.

4.5 Conclusions

In this chapter I presented two broad investigations into the nature of basins of attraction in fitness landscapes.

In the first investigation, in Section 4.3, I presented a novel method, the Basin Flow Algorithm, for studying basins of attraction which was inspired by the Page Rank algorithm

used originally by Google for web page ranking. Previous attempts at measuring basin size had featured the use of trajectories and sampling to estimate the features of basins in NK landscapes with N=20 with varying choice of K. In line with previous work, I found that basin size was linked exponentially with fitness in NK landscapes, and that this relationship decreases with increasing K.

I also tested increasing levels of neutrality using the NKp family of landscapes. I found large neutral networks that occasionally were slow to drain, but it was difficult to compare directly with the NK landscapes due to the differences in fitness range naturally induced by these landscapes. I also investigated the effects of double mutations and demonstrated how allowing basins to drain via double mutations—in other words, basins with lower fitness do disappear and are absorbed into those of higher fitness, as might be expected.

In the second part of the chapter, I followed the three system approach introduced in Chapter 3 and sampled 100,000 evolutionary trajectories starting at random points. I measured properties of basins of attraction for each of the NK, Avida, and RNA landscapes. I treated the NK and RNA landscapes with exponentiation in order to increase selection strength. This treatment improved the solution quality in both landscapes and leads us to believe that similar sorts of treatments could be generally applicable in evolutionary optimization—namely, small population and a high mutation rate can be overcome by manipulating selection strength. Tournament selection, for instance, may work well precisely because it is not sensitive to the selection strength differential.

I also found that basins are strongly linked to our conception of peaks in Chapter 3, and that over evolutionary time, trajectories starting from random points end up at a distance from the origin that is, on average, close to the 13.5 average genotype Hamming distance in all three landscapes. The high distance implies that basins for evolutionary purposes are not at all localized, even though it is often a much shorter distance to the nearest peak. Historical contingency may thus play the most important role in determining the ultimate fate of a

population trajectory.

In the following chapter, I investigate the fate of trajectories starting at the same origin and shift focus from global structure to local structure.

Chapter 5

Visualization of Transient Dynamics

In the preceding chapters, I have examined evolutionary dynamics across three distinct and widely used landscapes. Here, I will use visualization techniques to further compare these landscapes, both in terms of local structure and the short-term evolutionary dynamics induced by their structures.

5.1 Related Work

5.1.1 Replaying the tape of life

Steven Jay Gould in [Gould, 1989] famously posited the thought experiment of 'replaying the tape of life', arguing that historical contingency and chance played a large role in evolution potentially even surpassing that of adaptation. This classical thought experiment asks that if it were somehow possible to rewind time and start over, how would things be different? The tension between adaptationist, i.e. that selection would produce the same result, views and those similar to Gould, favoring random chance, marked one of the important philosophical questions in evolutionary biology.

[Travisano et al., 1995] sought to better understand the relationship between historical

contingency, chance, and adaptation in a set of cleverly designed experiments. They took clones from a historical ancestor to seed experimental populations across a set of differing environments, before putting them back into the ancestral environment. At the conclusion, they were able to estimate using an ANOVA how much of the variance derived from history (the choice of ancestor), chance (differences given the same ancestor), and adaptation (the influence of a specific environment). Fitness was highly adaptive, as might be expected, but, another phenotypic metric, cell size, demonstrated strong effects of history and chance, since cell size is itself not adaptive. Related biological experiments following this experimental approach can be found as in dinoflagellates in [Flores-Moya et al., 2012]. [Wagenaar and Adami, 2004] performed similar experiments in Avida organisms, transplanting organisms from an environment they evolved in into a novel environment. In similar environments, they found a greater importance of shared history, whereas in more dissimilar environments, the effects of adaptation dominated the variance. [Kryazhimskiy et al., 2014] also conducted a hierarchical experiment in which they studied the effects of epistasis and historical contingency. They found broad fitness-level convergence while at the same time encountering a high level of randomness at the sequence level, with the surprising result that clones descended from the same founders were not more likely to share mutations than those not sharing the same founders.

5.1.2 Fitness Landscapes & Epistasis

There have been a number of investigations on fitness landscapes with regard to epistasis; [Tufts et al., 2014] examines the mutational neighborhood of hemoglobin via site-directed mutation and found sign epistasis in the mutations required for increased oxygen affinity in high-altitude pikas. Similarly, [Podgornaia and Laub, 2015] mapped 160,000 variants of PhoQ in *E. Coli* to identify 1659 functional variants, demonstrating significant epistasis in the resulting landscape and postulating that this epistasis has severely limited the number of orthologs observed. Similarly, [Poelwijk et al., 2007] examined adaptation on the bacterial β-

lactamase to cefotaxime, to which they had not previously had exposure. In this case, the five mutations conferring antibiotic resistance were already known, but they were able to measure intermediates and found that it was necessary for mutational combinations to traverse a valley to reach the resistant solution. Similarly, the fitness landscape in *E. Coli* has been analyzed in [Beerenwinkel et al., 2007], which found a strong correlation between epistasis and the amount of fitness loss inflicted by deleterious mutations. Similarly, [Covert et al., 2013] demonstrated the importance of deleterious mutations to adaptive evolution in digital Avida landscapes by comparing the achieved fitness of populations with normal adaptive evolution against those where deleterious mutants were reverted.

[Szendro et al., 2013] performed a computational analysis of trajectories on the experimentally-measured 8-locus fitness landscape from the fungus *Aspergillus niger*. They defined entropy measures both on endpoints and on trajectories and found that the entropy of both of these decrease with mutational supply and that larger populations become more greedy. Similarly, [Handel and Rozen, 2009] found that larger populations evolve more deterministically and can become trapped on local fitness peaks, whereas smaller populations may be able to reach higher peaks because of their ability to traverse the landscape more randomly. [Lobkovsky et al., 2011] proposed several landscape characteristics and linked them to trajectories in 39 protein folding landscapes of size up to 12,969.

[Otwinowski and Plotkin, 2014] illustrates the biases often present in methodologies involving inferring fitness landscape structure by sampling populations in both NK and RNA populations. These biases are especially prevalent in living organisms due to selection pressure.

5.2 Approach

My work seeks to investigate evolutionary trajectories against the backdrop of the fitness landscapes explored in previous chapters. However, unlike much previous work, where often

such an undertaking would be prohibitively expensive or impossible, I seek to exhaustively enumerate the possibility space before examining real evolutionary trajectories. It has been pointed out as in [Kogenaru et al., 2009] that mutational analyses restricted to only one area of the genome may be compensated for in other areas and may be therefore irrelevant—an exhaustive exploration avoids this potential issue.

In this work, I examine the mutational neighborhood of points out to five mutations out. There are $\binom{18}{5}3^5 = 2,082,024$ points at exactly a mutational distance of five, and each of these has 5! ways to achieve it, which yields 249,842,880 possible trajectories. Much previous work has focused on a particular gene complex and analyzed the fitness of paths by measuring each mutational combination along the way. My work differs in that I look at an unconstrained mutational neighborhood, exploring all possible combinations of steps, not just limited to specific target sequences.

The first segment of this work features an exploration of the local structure of several sampled points, and the second part is results of repeated evolutionary trials to see what actually happens as a result of evolution. It is worth noting that the evolutionary runs are highly parameter dependent—the same fitness landscape may experience radically different trajectories based on population size, mutation rate, and selection mechanism.

I visualize the potential trajectory space of 1,000 points for each of the three model landscapes introduced in Chapter 2—the NK landscape, the Avida landscape, and the RNA landscape. In order to accomplish this, for each sampled point, I calculated the fitness for all points within a mutational distance of five. I keep track of rings, defined as all points of the same Hamming distance from the source genotype—so for instance, all one-mutant neighbors is a ring, all two-mutant neighbors another, and so on. I also keep track of links between the origin genotype and one-step mutants, between one-step mutant and two-step mutants, and continuing up to the links between four-step mutants and five-step mutants.

My data was far too complex to easily plot; each point, each generating 250 million

trajectories is too much data for a traditional plotting program to visualize. Therefore, I preprocessed the data to accomplish this visualization. I started with a numeric matrix, with mutational distance on the x axis and fitness on the y axis. I used 100 pixels between each mutational step for the x axis. For each starting point, I assign a starting point on the left of the graph, with the y position determined by the fitness. I then find the possible endpoints; there are 54 one-mutation steps to one-mutant neighbors. These are assigned an x position of 200, and a fitness according to the fitness of the corresponding genotype. I use Bresenham's Line Algorithm, originally outlined in [Bresenham, 1987], to determine the pixels along a line and increase the density in the matrix for each of these. I then repeat the process linking one mutation neighbors with each two mutation neighbor it may link to, and so on until I finally connect the four and five-mutant points. After this process is completed, I renormalize between each step before visualizing the data via a density heatmap. This technique is used to produce the visualizations of both the local landscapes and trajectories out to five mutational steps.

5.3 Local Landscapes—Random Points

I show the visualization for 1,000 random points selected from each of these landscapes in this section, using the MLV method described above.

Figure 5.1: NK Random Starting Points, counts of beneficial/neutral/deleterious mutations relative to random starting genotype. Note that beneficial and deleterious mutations nearly overlap so only two lines appear. This overlap is because choosing a large enough sample of random points in the landscape should make beneficial and deleterious mutations nearly perfectly symmetric. At zero distance, the point itself is classified as neutral.

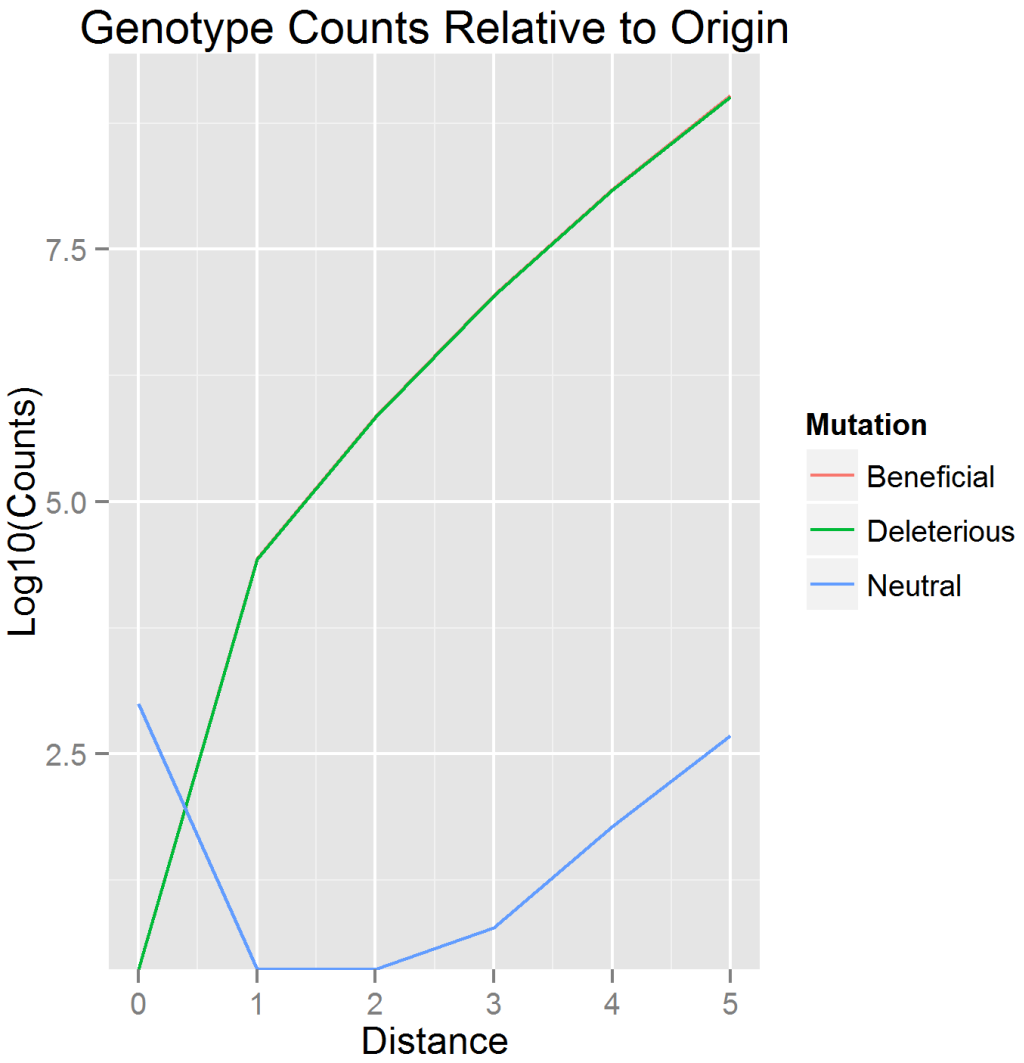


Figure 5.2: NK Random Starting Points, Path Counts Simplex. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations. Since this is the NK landscape, there is very little neutrality in this particular landscape.

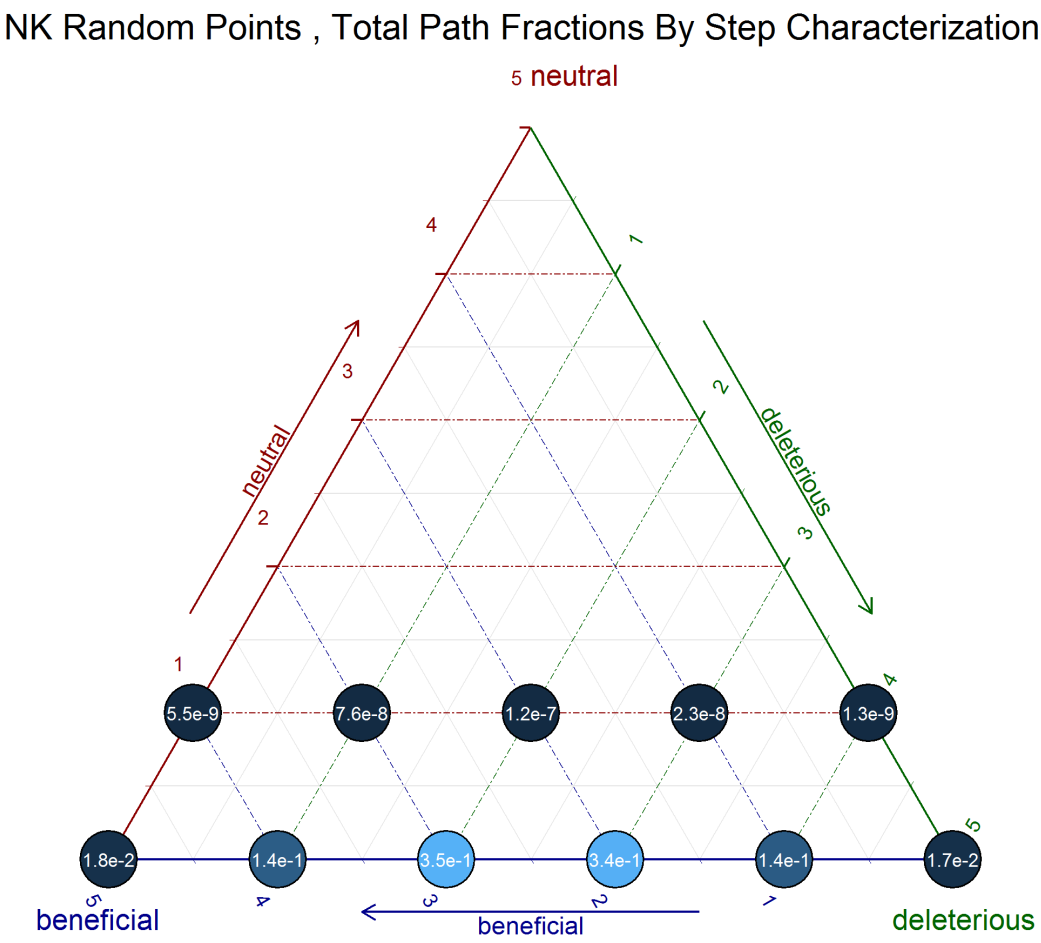


Figure 5.3: NK Random Starting Points, Landscape Visualization. This shows 1,000 NK landscapes starting at random points. The normality of the fitness distribution of the NK landscape is clearly visible and rare mutations in both directions appear at five mutations out.

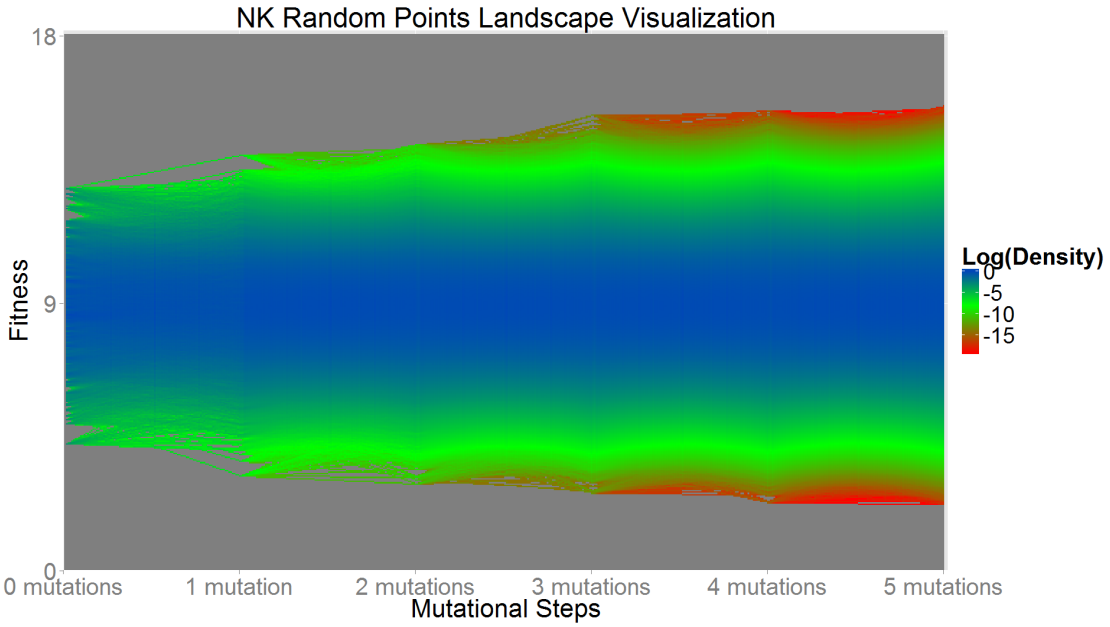


Figure 5.4: Avida Random Starting Points, counts of beneficial/neutral/deleterious mutations relative to randomly-selected starting genotype. Note that beneficial and deleterious mutations nearly overlap so only two lines appear. This overlap is because choosing a large enough sample of random points in the landscape should make beneficial and deleterious mutations nearly perfectly symmetric. At zero distance, the point itself is classified as neutral.

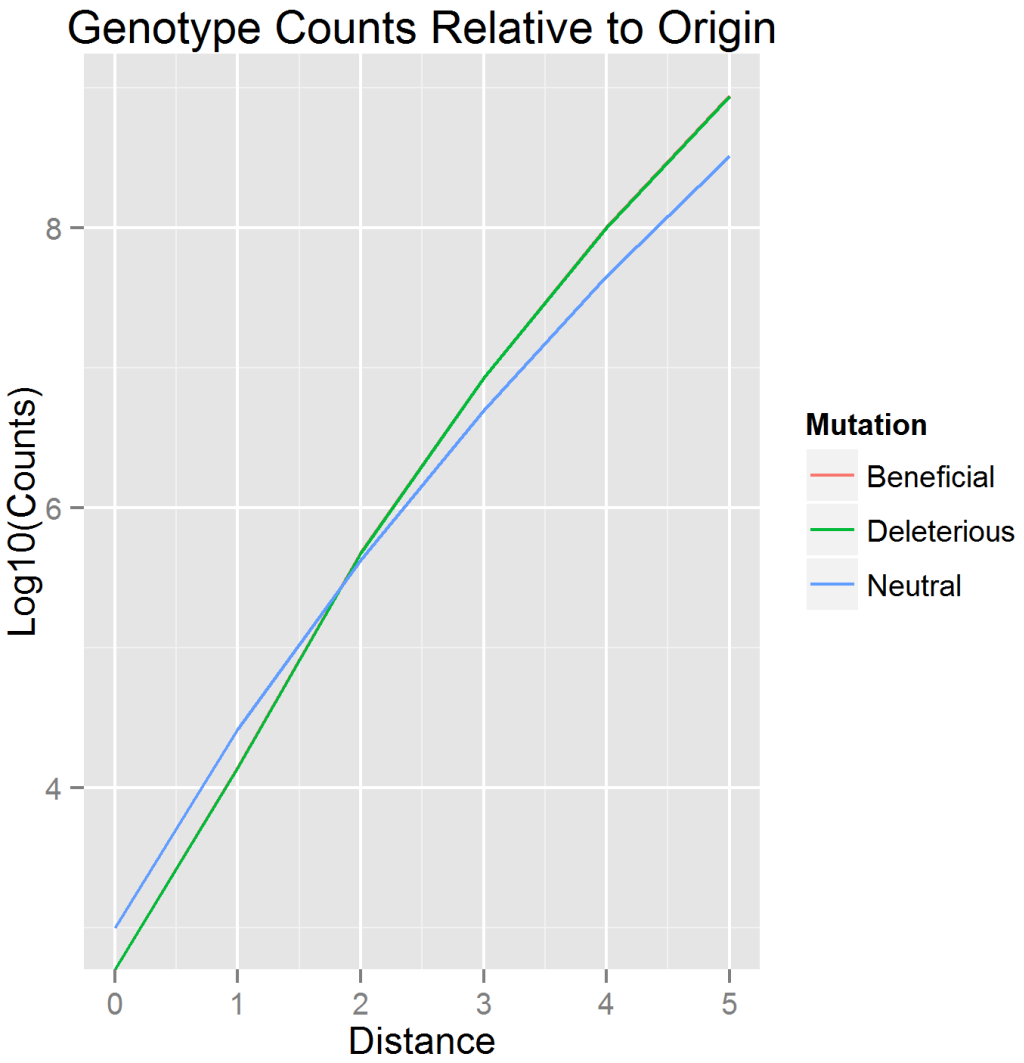


Figure 5.5: Avida Random Starting Points, Path Counts Simplex. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations.

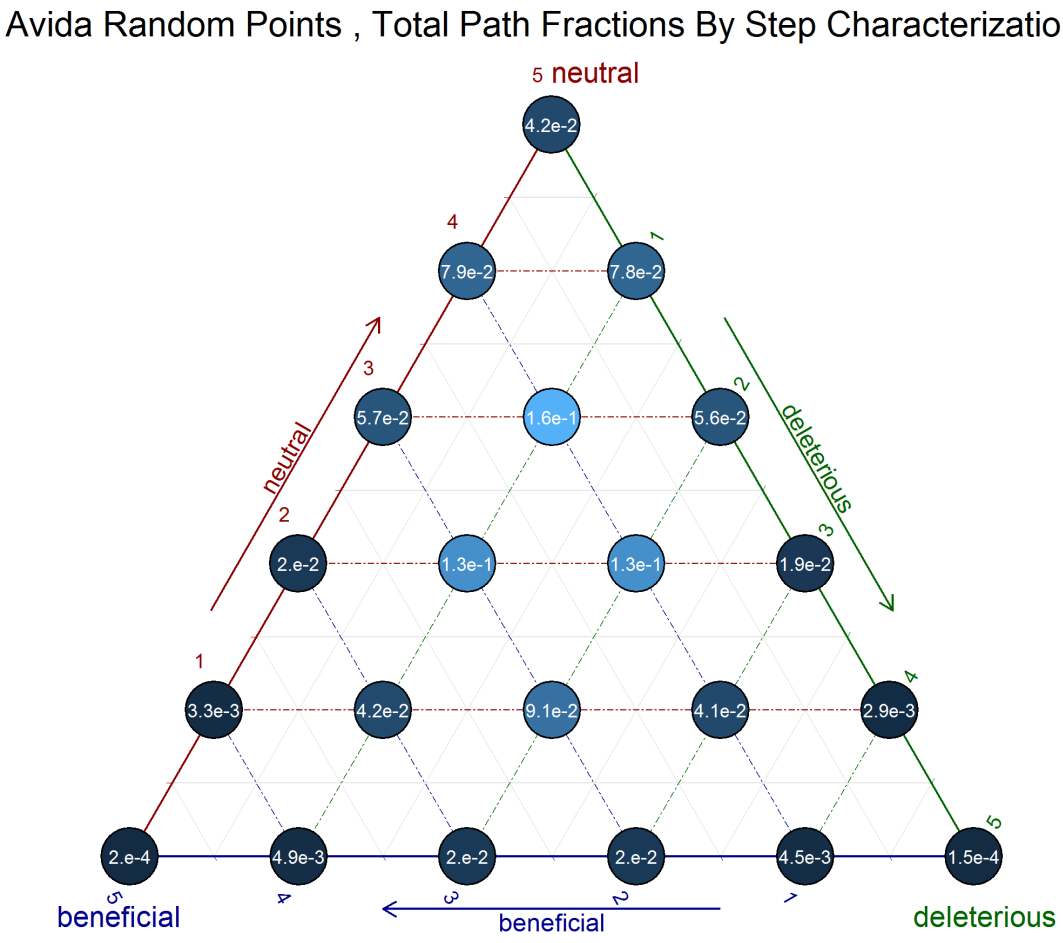


Figure 5.6: Avida Random Starting Points, Landscape Visualization. This shows 1,000 Avida landscapes starting at random points. Rare high fitness states are revealed at a distance of five mutations out.

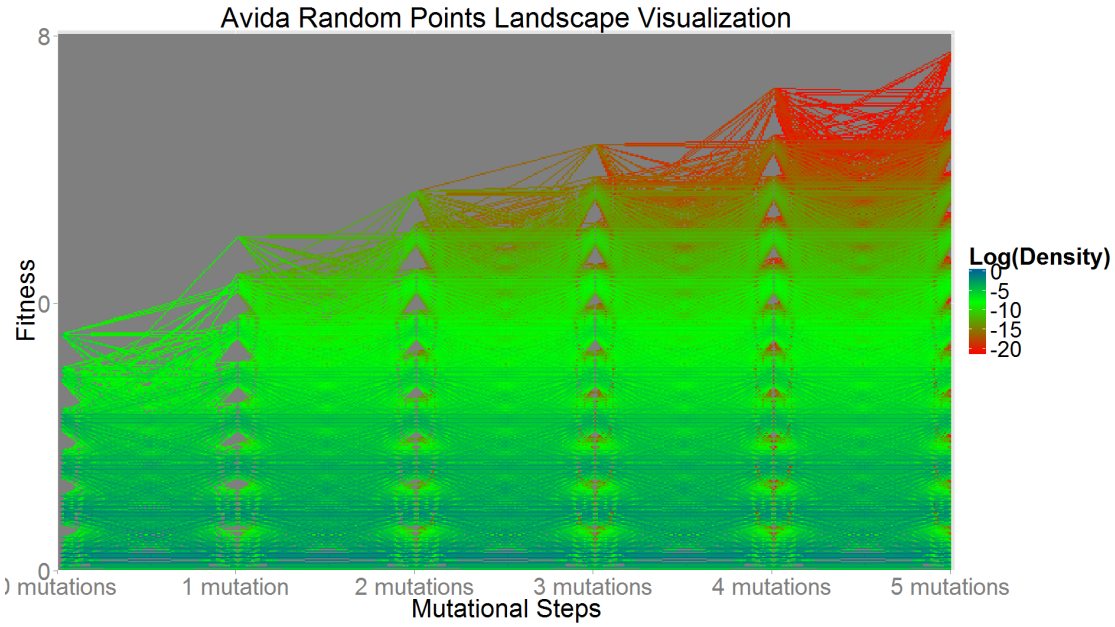


Figure 5.7: RNA Random Starting Points, counts of beneficial/neutral/deleterious mutations relative to randomly-selected starting genotype. Note that beneficial and deleterious mutations nearly overlap so only two lines appear. This overlap is because choosing a large enough sample of random points in the landscape should make beneficial and deleterious mutations nearly perfectly symmetric. At zero distance, the point itself is classified as neutral.

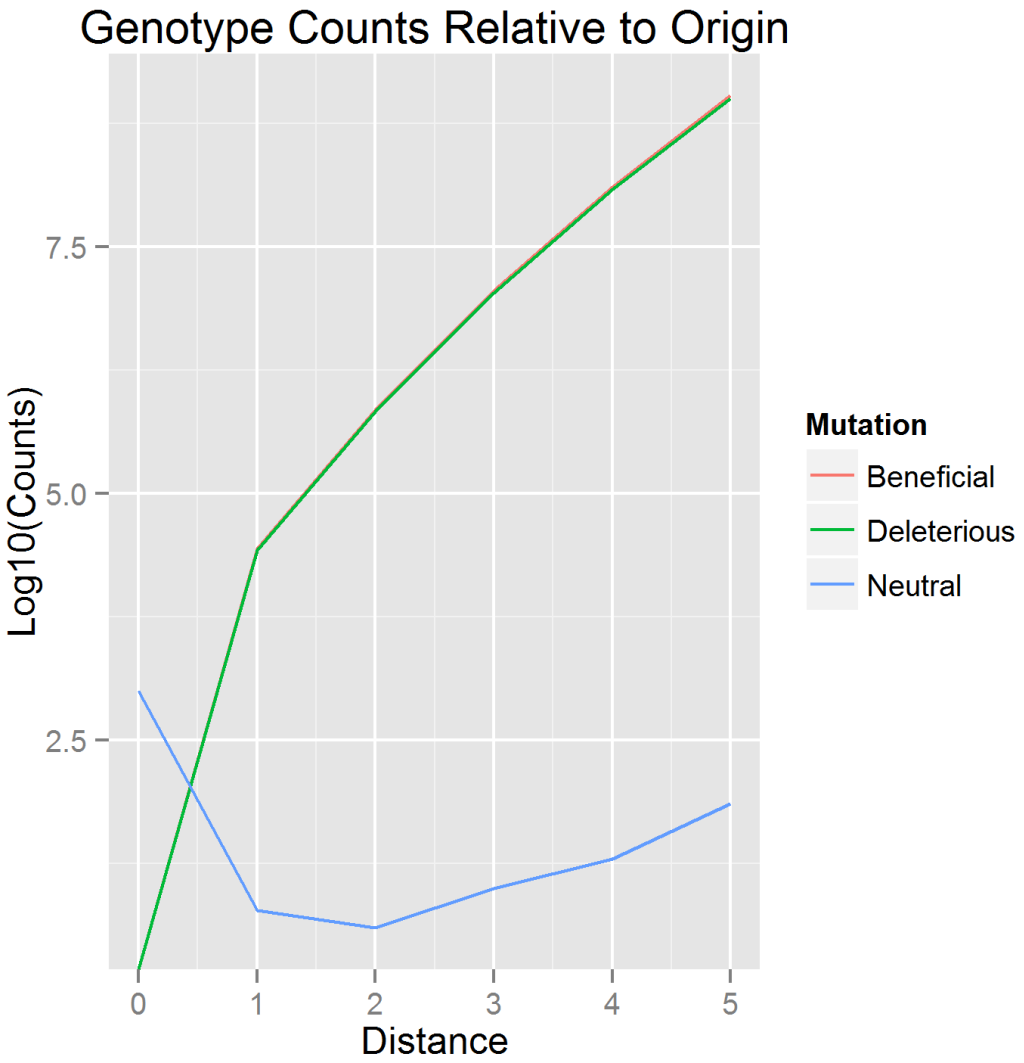


Figure 5.8: RNA Random Starting Points, Path Counts Simplex. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations.

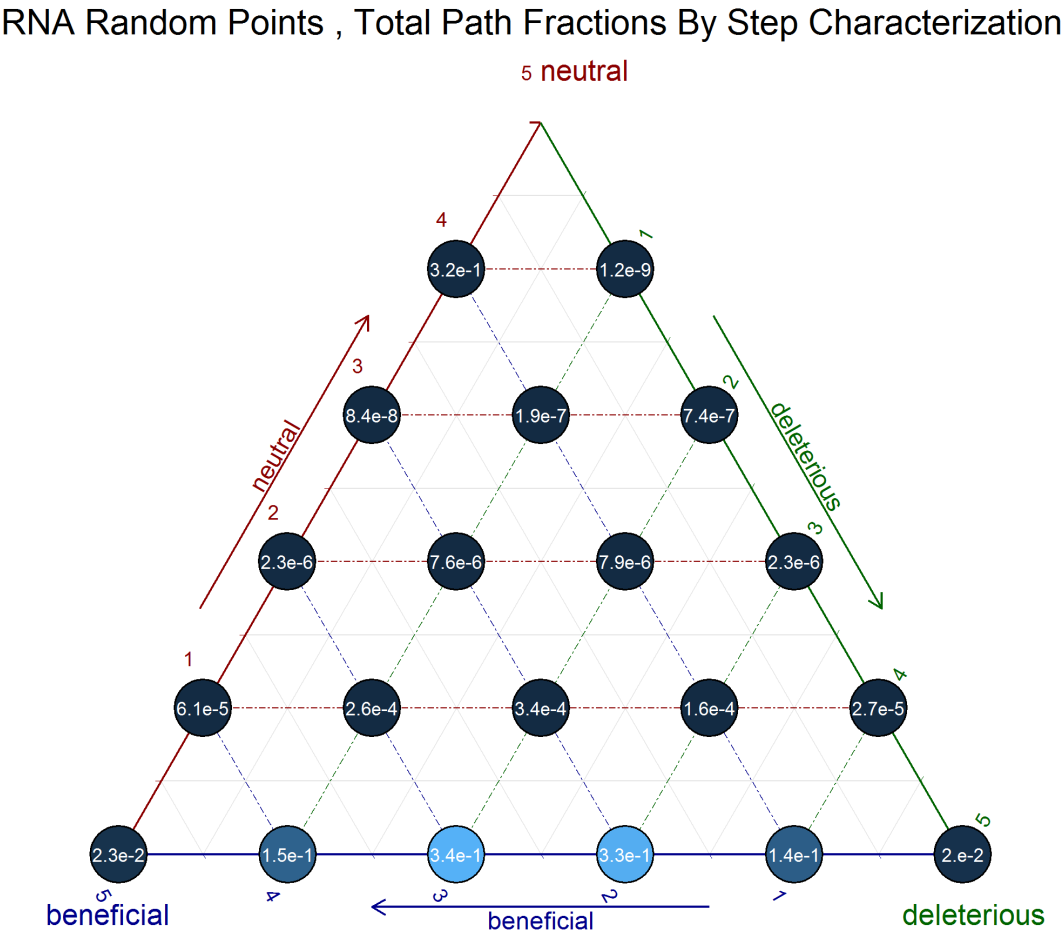
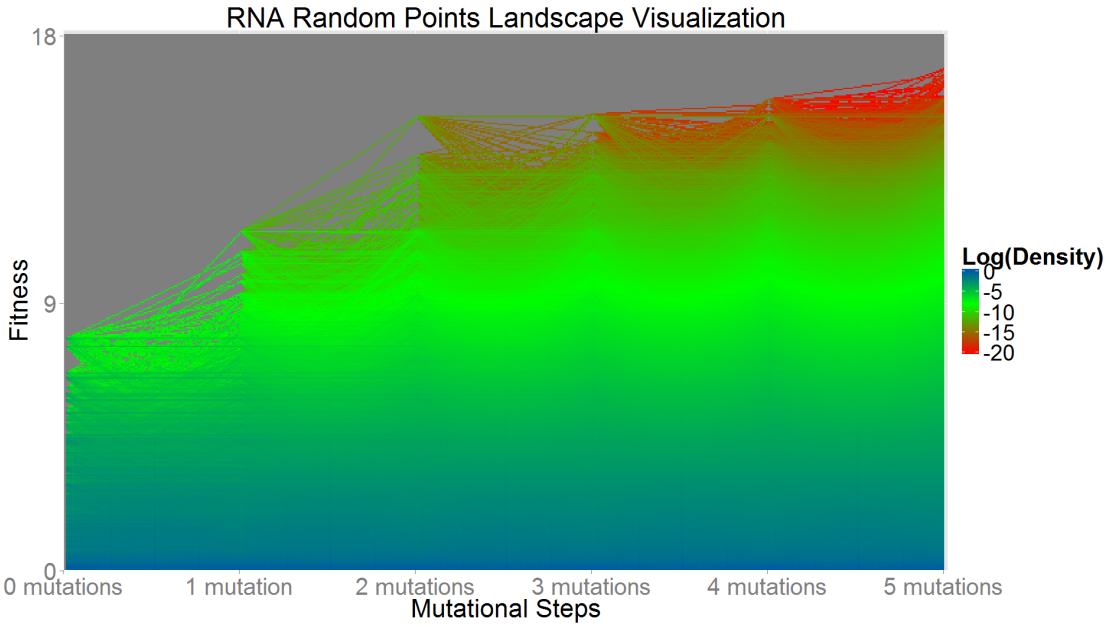


Figure 5.9: RNA Random Starting Points, Landscape Visualization. This shows 1,000 RNA landscapes starting at random points. Rare high fitness states are revealed at a distance of five mutations out.



5.4 Local Landscapes—Peaks

Next, I apply the same treatment to peaks. I select my subset of peaks to study by first sampling a random point, and then hill climb to a peak. This hill climbing is accomplished by iteratively examining the one-mutation neighbors and taking the most fit among them. Up to one hundred neutral steps are allowed, after which only positive steps are allowed, until none are available. By the end of this process, the current genotype is guaranteed to be a 'neutral peak' as defined earlier in Chapter 3.

Figure 5.10: NK Peaks, counts of beneficial/neutral/deleterious mutations relative to starting genotype for each distance. At zero distance, the point itself is classified as neutral.

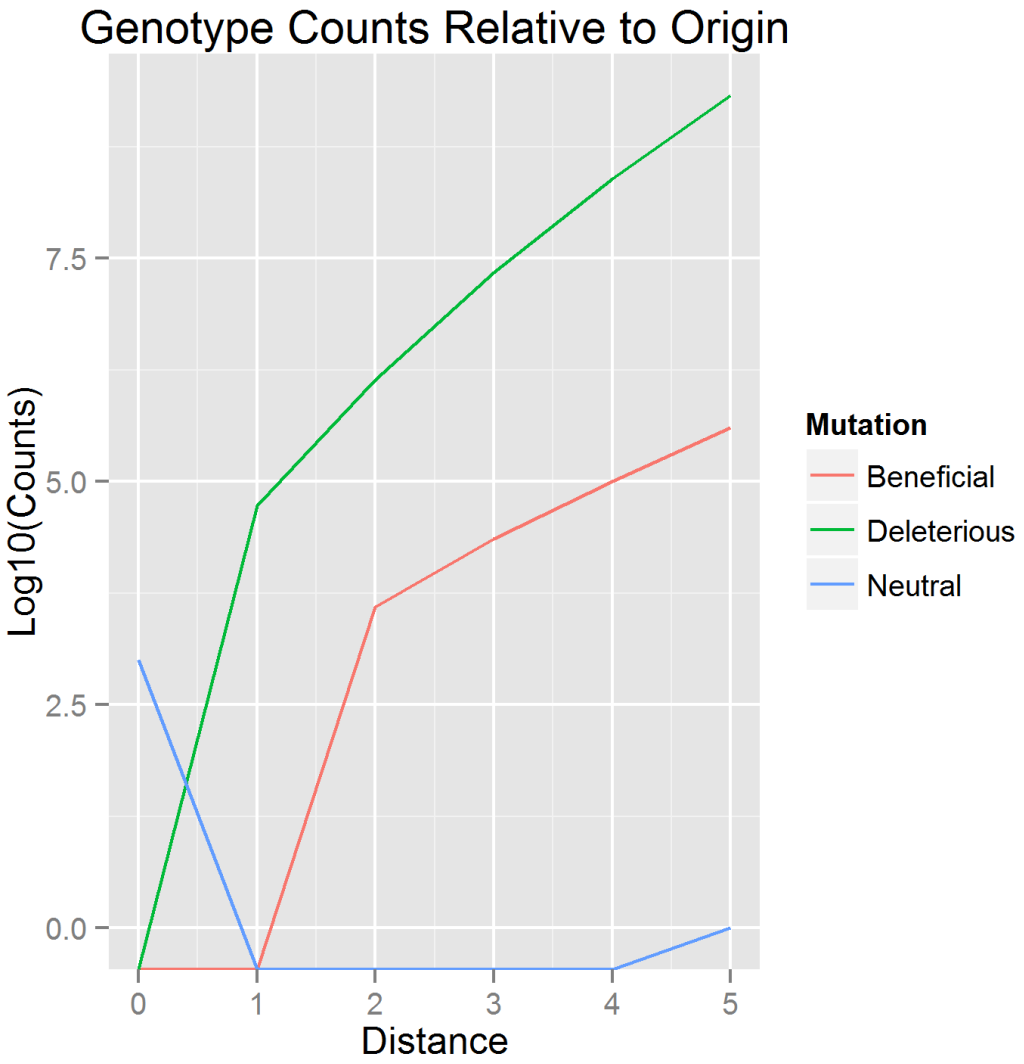


Figure 5.11: NK Peaks, Path Counts Simplex. This is a summary of all the possible five-mutation trajectories starting at each of 1,000 sampled peaks. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations.

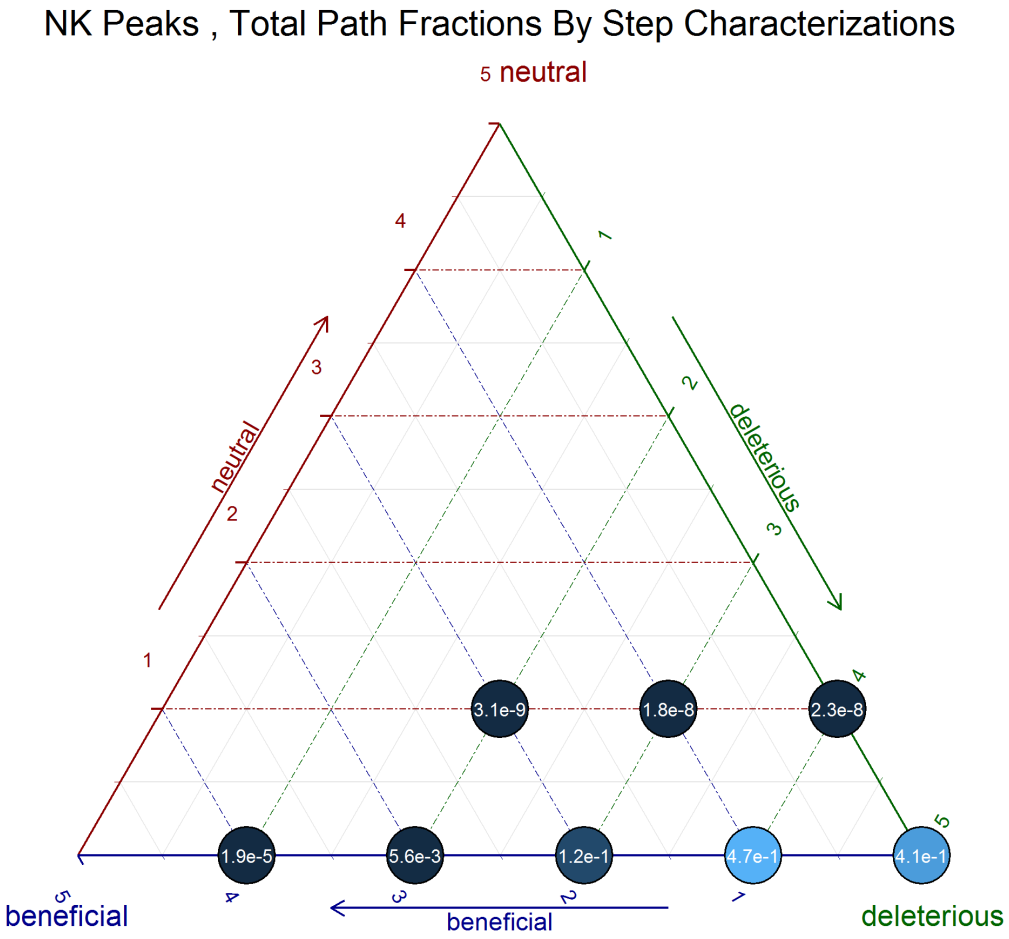


Figure 5.12: NK Peaks, Landscape Visualization. This shows 1,000 NK landscapes starting at peaks. A very few paths leading to higher fitness peaks can be seen.

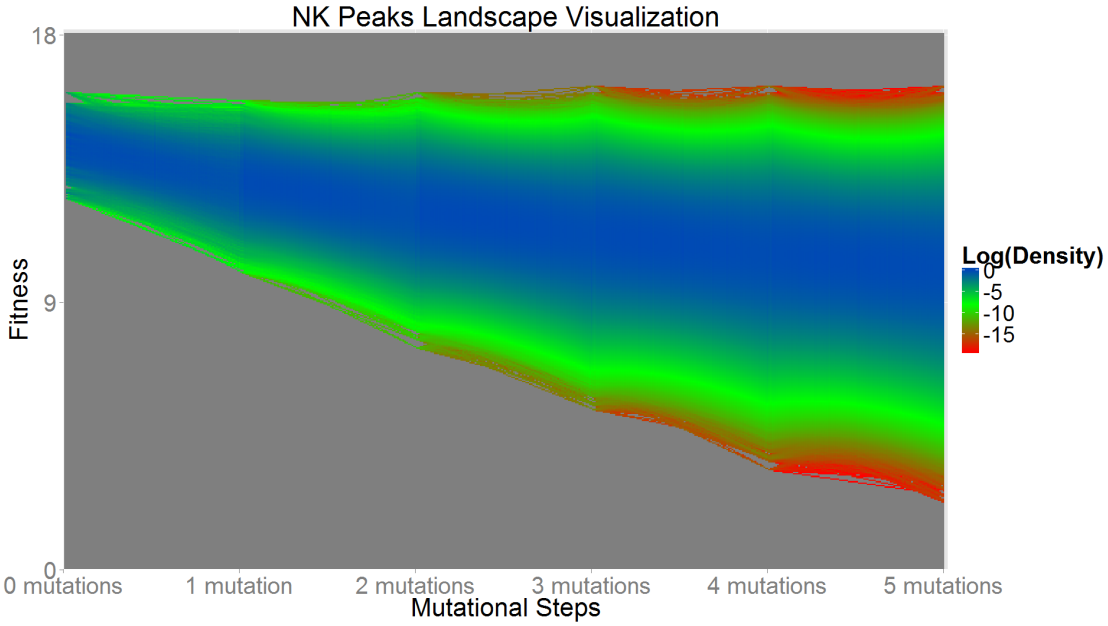


Figure 5.13: Avida Peaks, counts of beneficial/neutral/deleterious mutations relative to starting genotype for each distance. At zero distance, the point itself is classified as neutral.

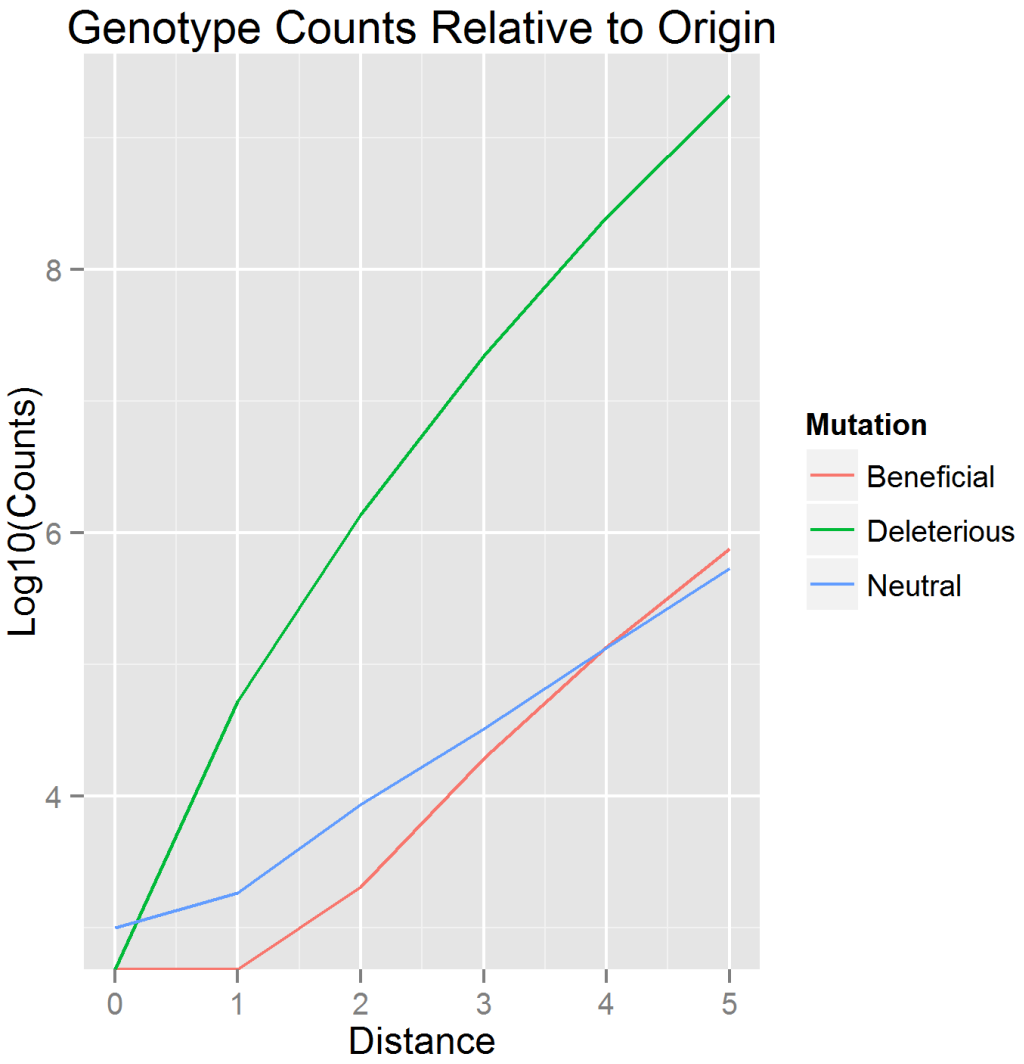


Figure 5.14: Avida Peaks Path Counts Simplex. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations.

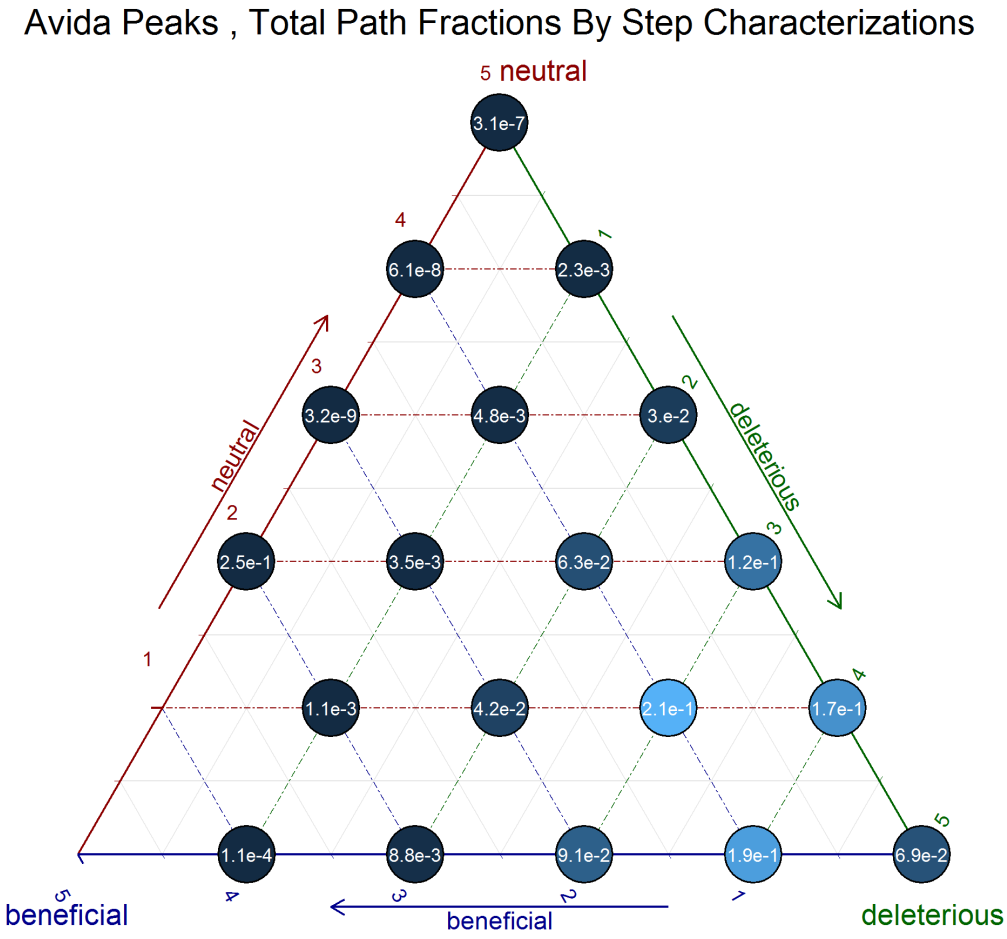


Figure 5.15: Avida Peaks Landscape Visualization. This shows 1,000 Avida landscapes starting at peaks.

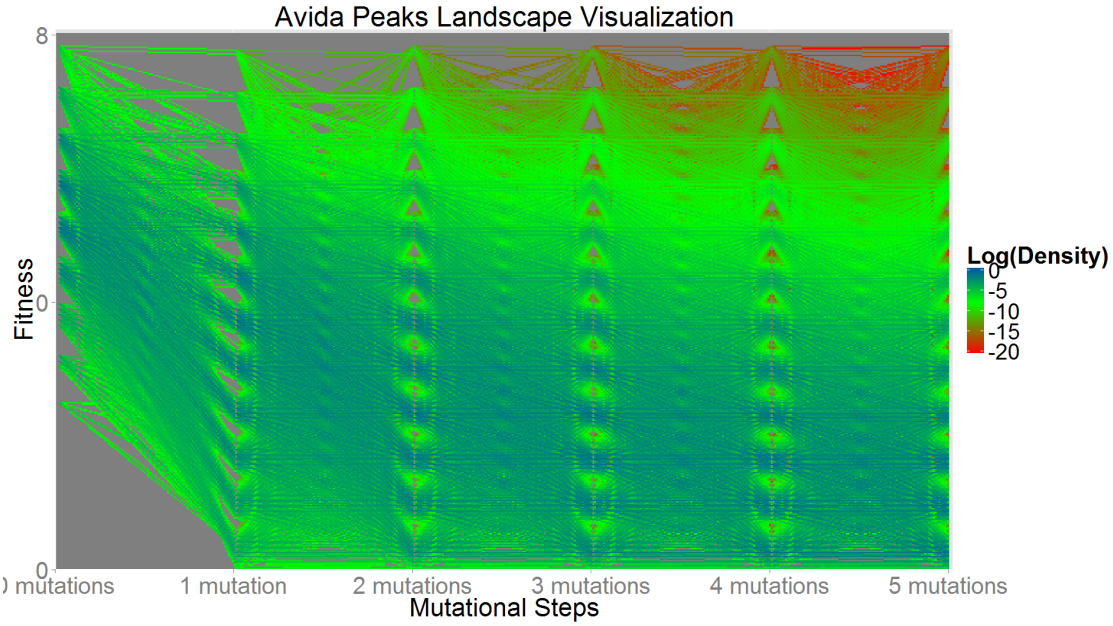


Figure 5.16: RNA Peaks, counts of beneficial/neutral/deleterious mutations relative to starting genotype for each distance. At zero distance, the point itself is classified as neutral.

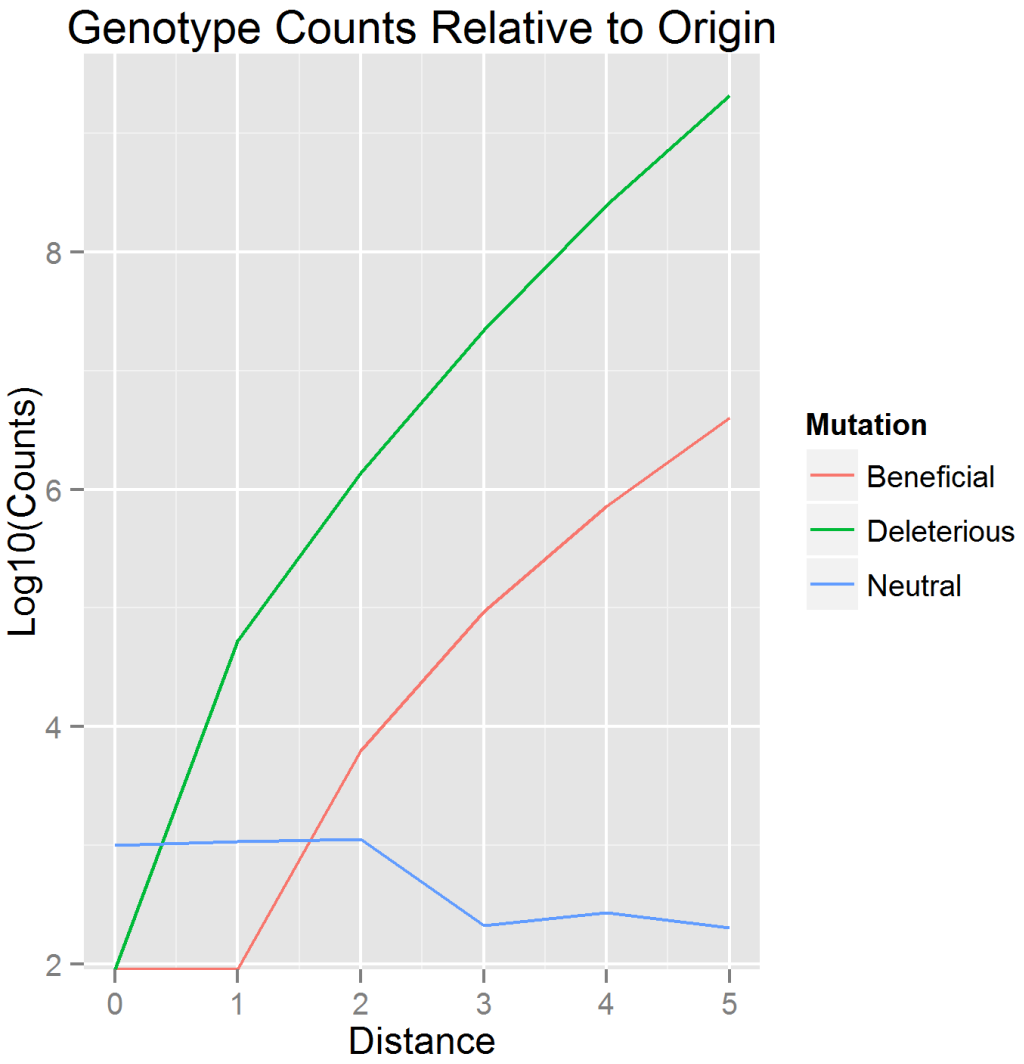


Figure 5.17: RNA Peaks Path Counts Simplex. This diagram is a summary of all of the possible five-mutation trajectories. To interpret the meaning of a point on this graph, go left to find the number of neutral mutations, up and right to find the number of beneficial mutations, and down and right to get the number of deleterious mutations.

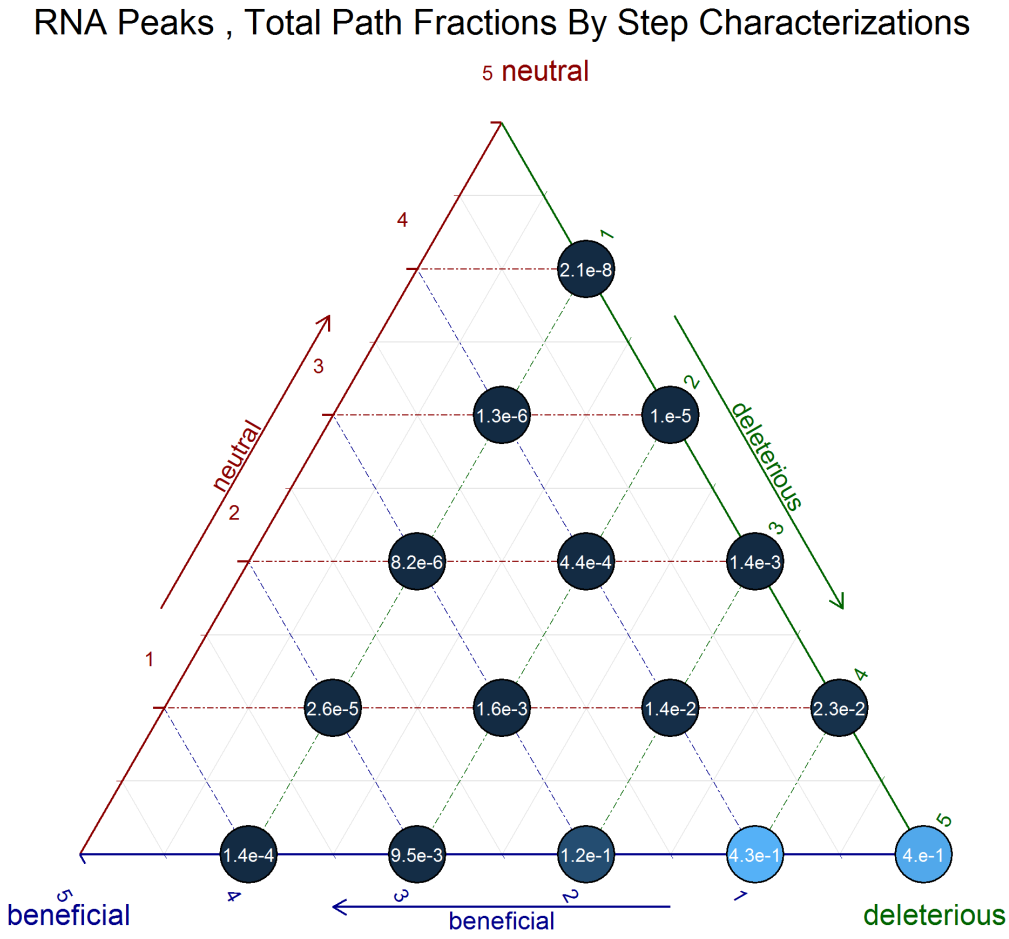
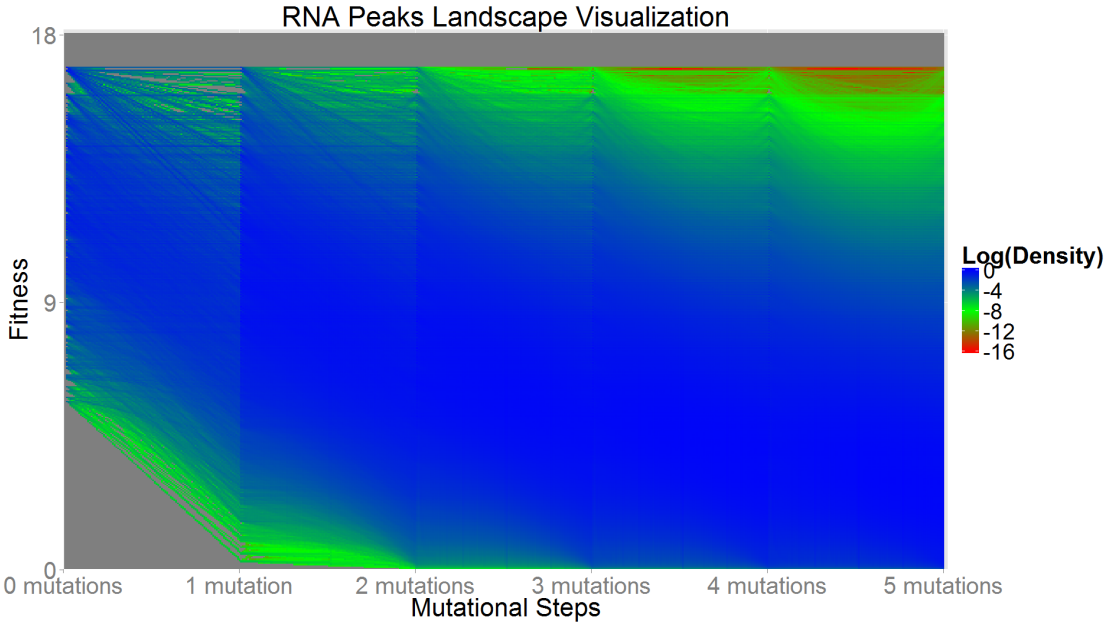


Figure 5.18: RNA Peaks Landscape Visualization. This shows 1,000 RNA landscapes starting at peaks.



5.5 Repeated Trajectories—Random Points Visualization

For each of the 1,000 points previously landscaped, I now began a further experiment to investigate what course evolution actually takes. To this end, I sample, 10,000 trajectories from *each* starting point for 100 generations and look at where the trajectories leave the five mutation radius. I examine the line of descent of the dominant genotype after 100 generations, and trace the unique genotypes along the line. I stop at five mutations out, for effective comparison against the complete landscapes. To enforce this five-mutation limit, I discard trajectories that skip a five distance mutant in the lineage through the simultaneous acquisition of two or more mutations.

The matrices are arranged with mutational distance on the x axis and fitness on the y axis. For each mutational step along the line of descent, I draw a line connecting the origin fitness and the origin distance with the destination line and distance. I start with a numeric matrix of zeros, and once again I use Bresenham's Line Algorithm to determine the pixels along the line and increase the density in the matrix. I normalize the entire data set at once by the number of trajectories, unlike in the landscape analysis where it was necessary to normalize between steps.

I show the visualization of trajectories for 1000 points for each of the landscapes.

Figure 5.19: NK Random Points Evolutionary Trajectory Visualization. This shows 1,000 NK collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

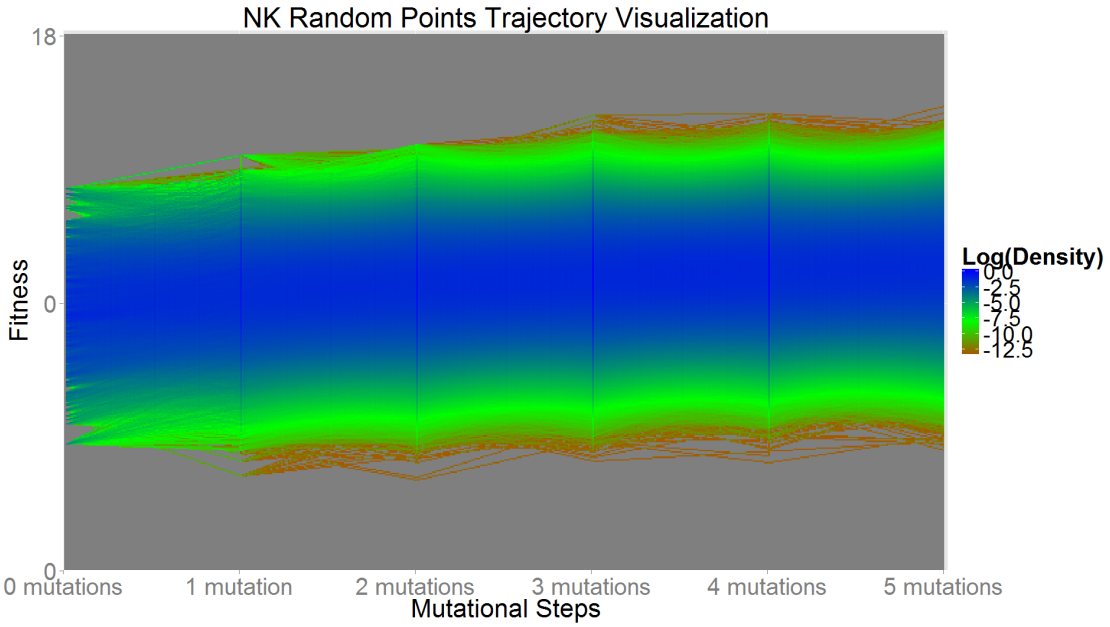


Figure 5.20: NK Random Points Exponential Evolutionary Trajectory Visualization. This shows 1,000 NK collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

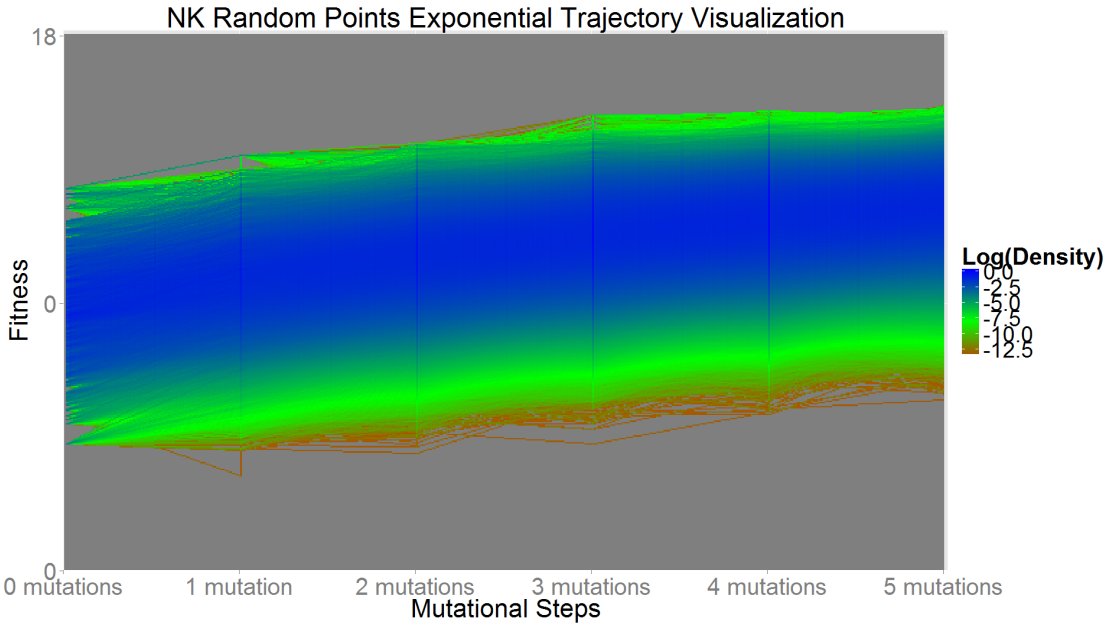


Figure 5.21: Avida Random Points Evolutionary Trajectory Visualization. This shows 1,000 Avida collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

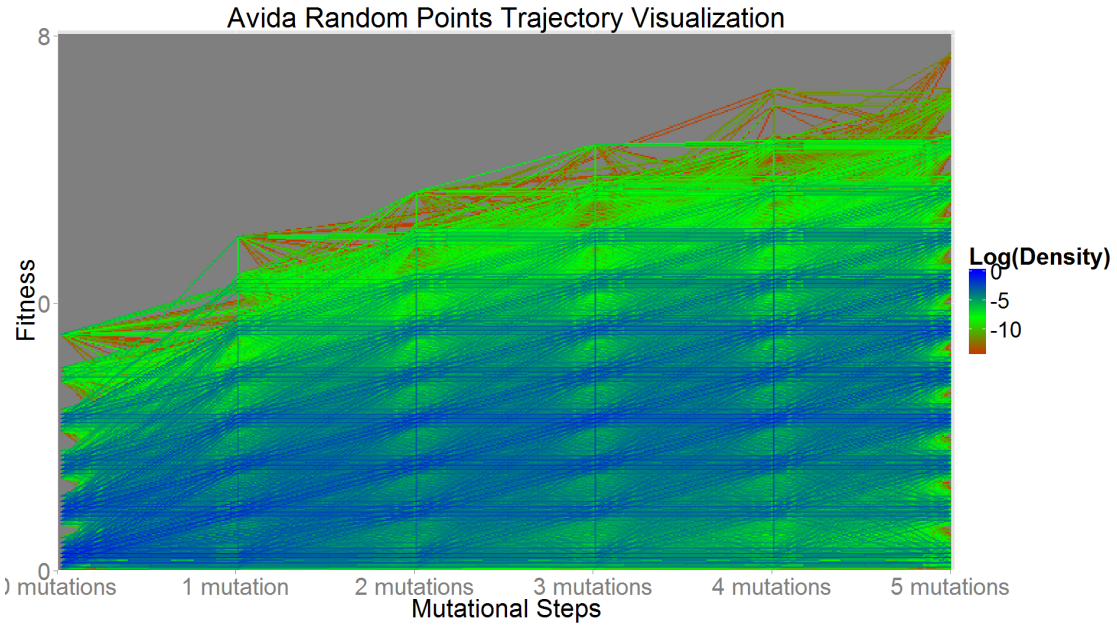


Figure 5.22: RNA Random Points Evolutionary Trajectory Visualization. This shows 1,000 RNA collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

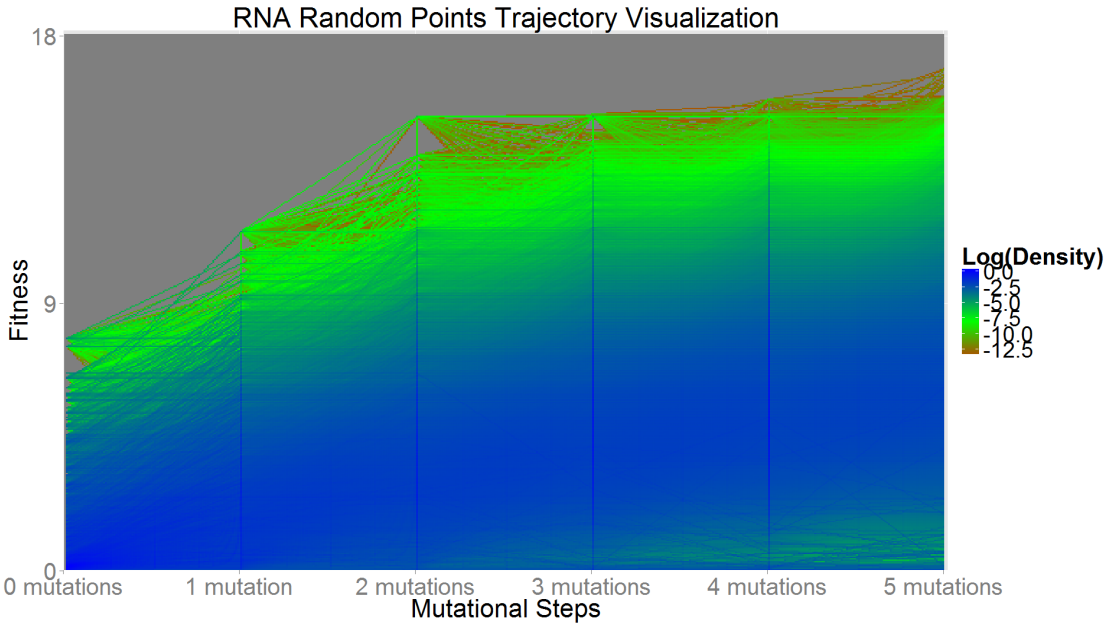
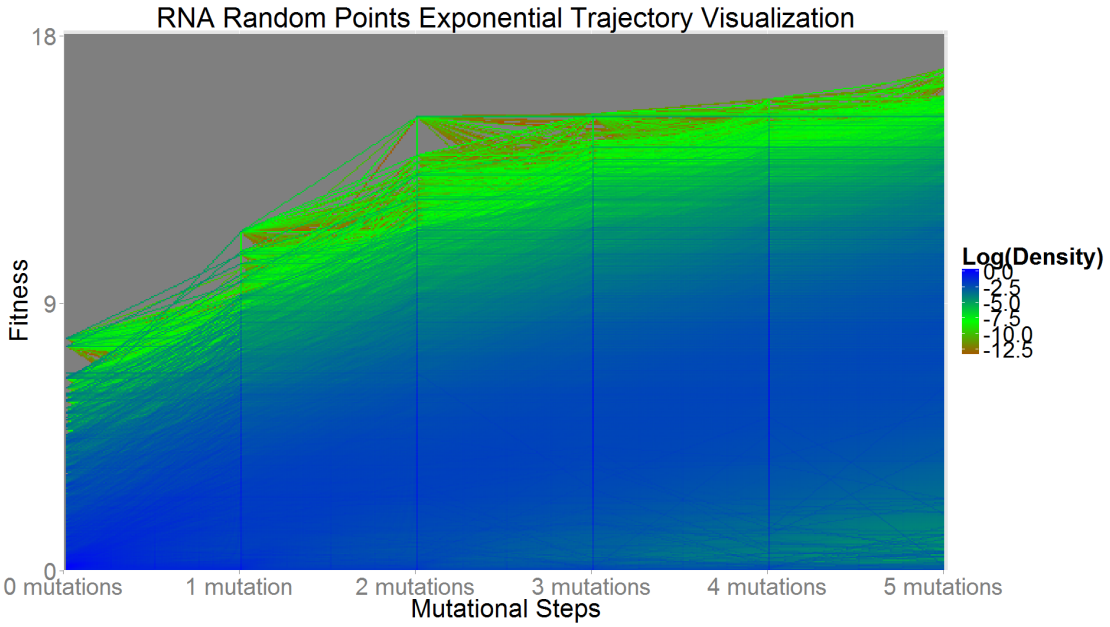


Figure 5.23: RNA Random Points Exponential Evolutionary Trajectory Visualization. This shows 1,000 RNA collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape



5.6 Repeated Trajectories—Peaks Visualization

I sample trajectories in a similar method to the previous section, with the difference being that these trajectories begin at the same peak points also used earlier in the landscaping.

Figure 5.24: NK Peaks Evolutionary Trajectory Visualization. This shows 1,000 NK collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

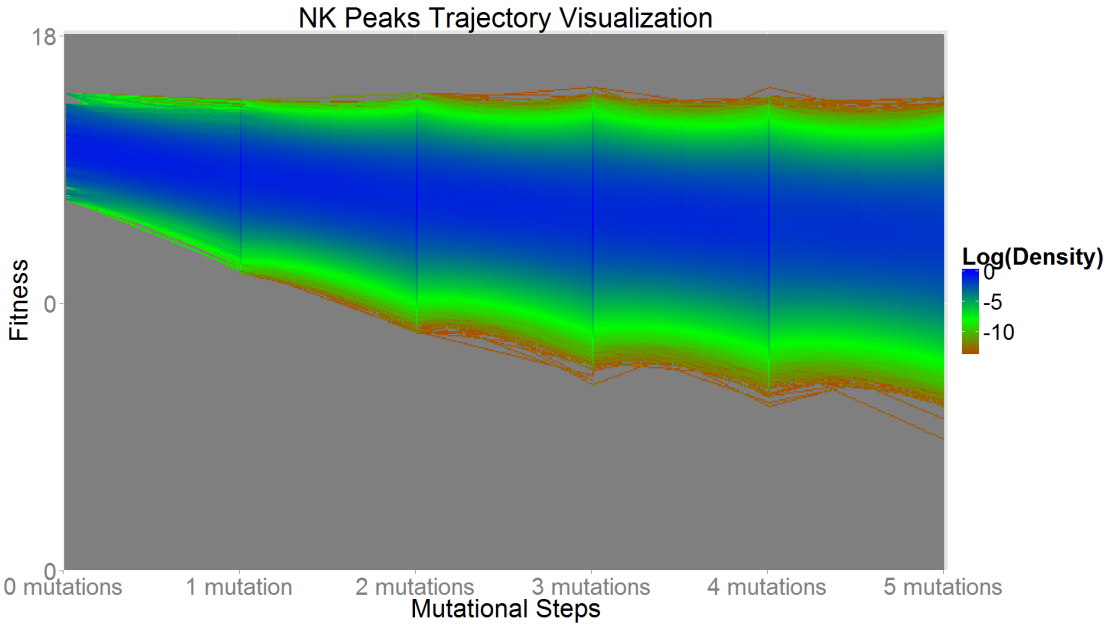


Figure 5.25: NK Peaks Exponential Evolutionary Trajectory Visualization. This shows 1,000 NK collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

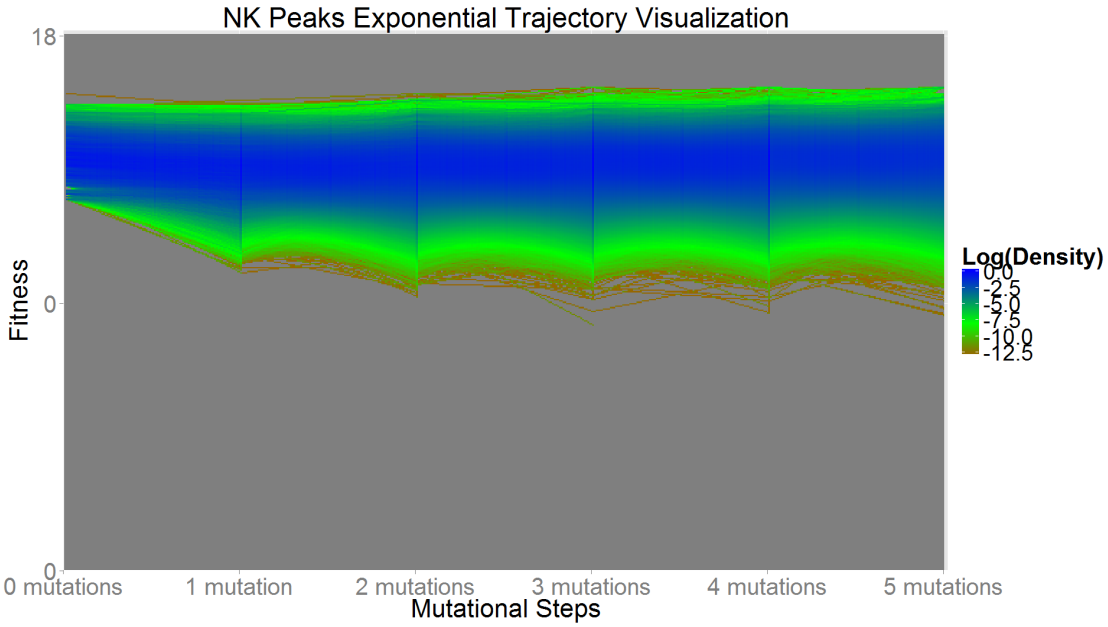


Figure 5.26: Avida Peaks Evolutionary Trajectory Visualization. This shows 1,000 Avida collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

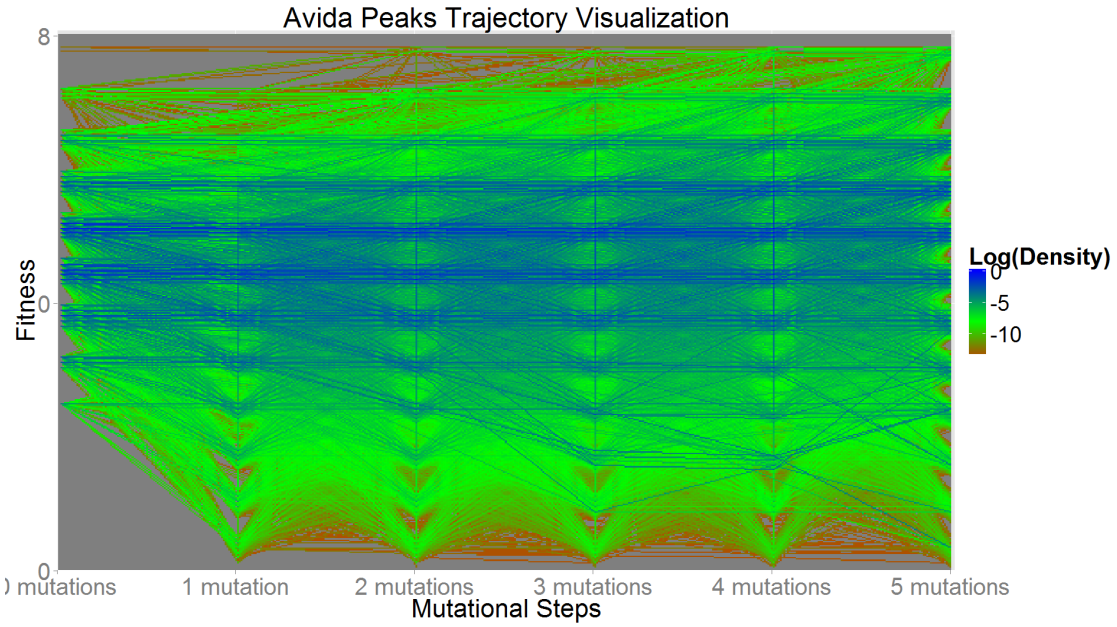


Figure 5.27: RNA Peaks Evolutionary Trajectory Visualization. This shows 1,000 RNA collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape

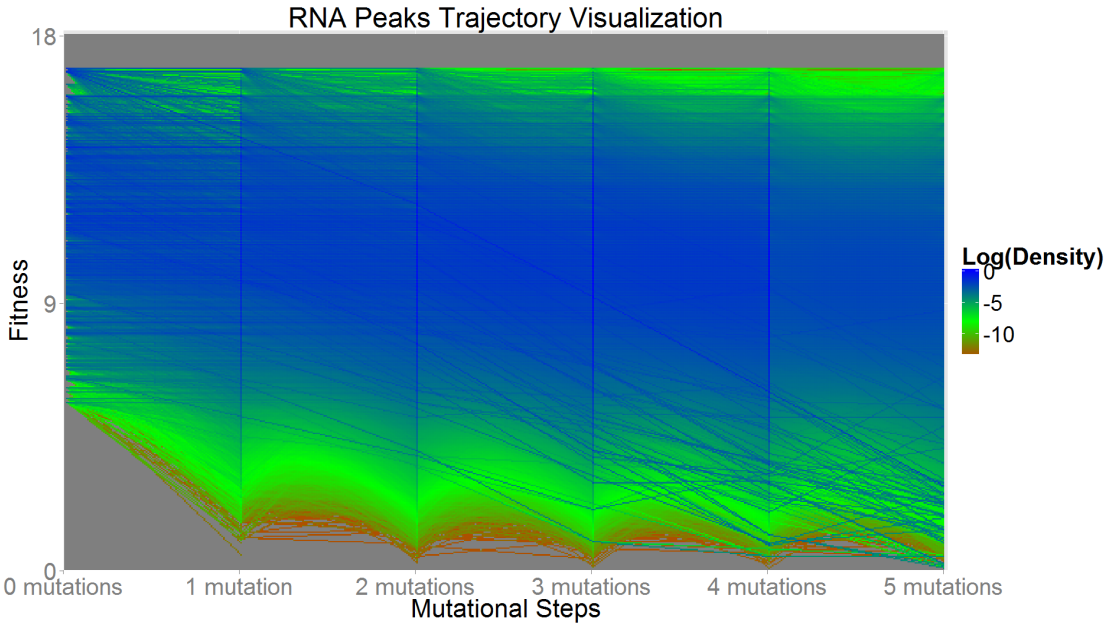
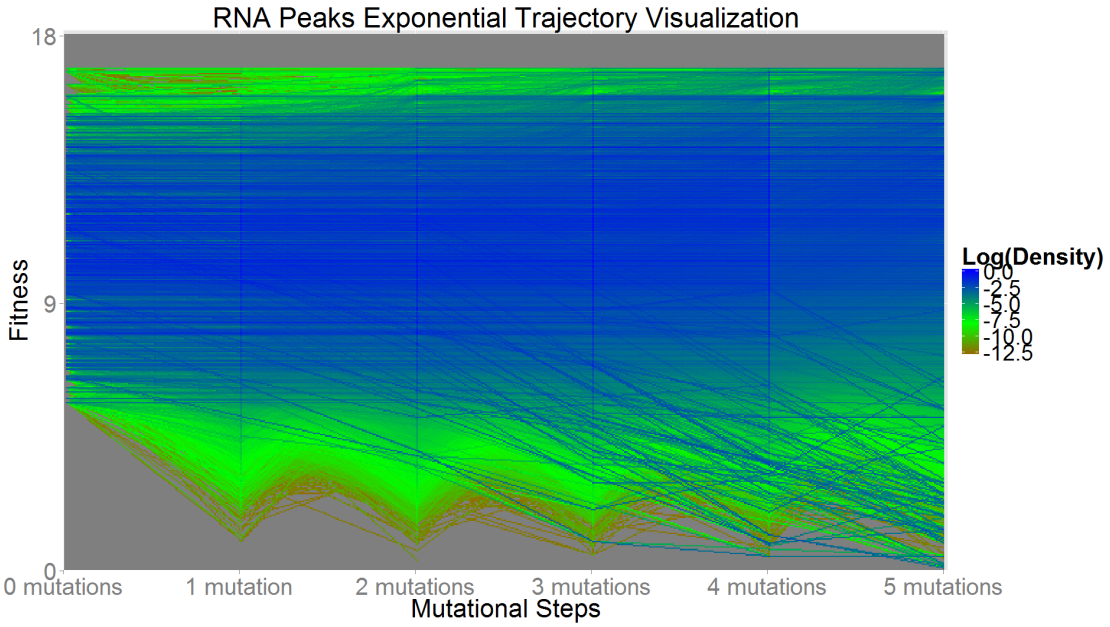


Figure 5.28: RNA Peaks Exponential Evolutionary Trajectory Visualization. This shows 1,000 RNA collections of up to 10,000 evolutionary trajectories each, starting at the same points as the corresponding landscape



5.7 Repeated Trajectories Random Points—The fate of trajectories

I examine the differences between the evolutionary trajectories obtained and the possibilities in the landscape, within a five-mutation threshold. To do this, I examined for each mutational step, the fitness of all points in the landscape at that mutational step. For each step, I record the median, minimum, maximum, 5th quantile, and the 95th quantile in the landscape. I combine the data from the 1,000 points by taking the median of each of these data points, yielding the median-median, median-minimum, median-maximum, median-5th quantile, and median-95th quantile. I similarly classify trajectories into points by distance, up to five, and do a similar collection—first recording the same five statistics for individual trajectories, and then combining the results by taking medians.

Figure 5.29: NK Random Points Landscape and Trajectory Fitness. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

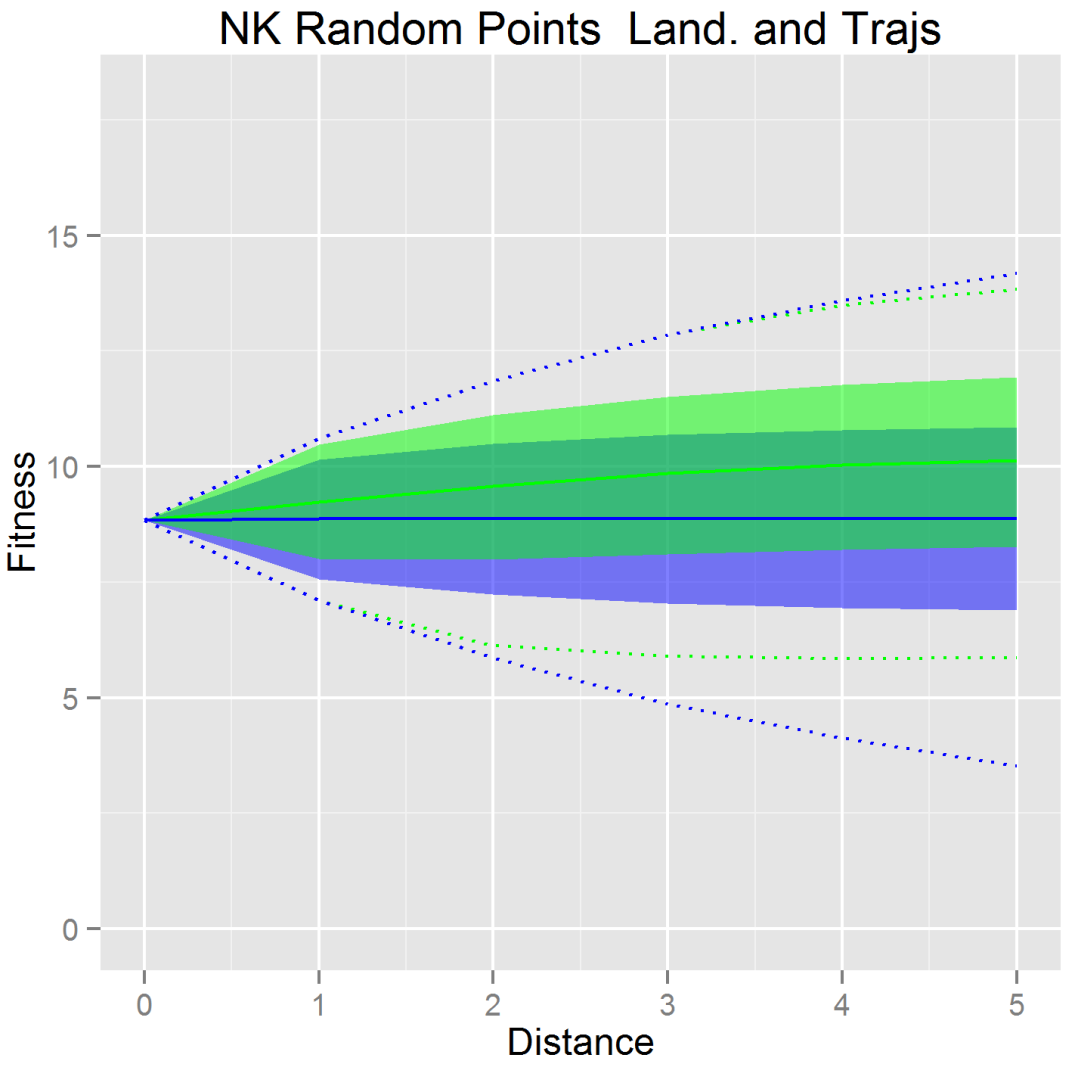


Figure 5.30: NK Random Points Exponential Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid. Here, selection strength results in a stronger improvement bias in the trajectories relative to the NK Points Trajectory Visualization.

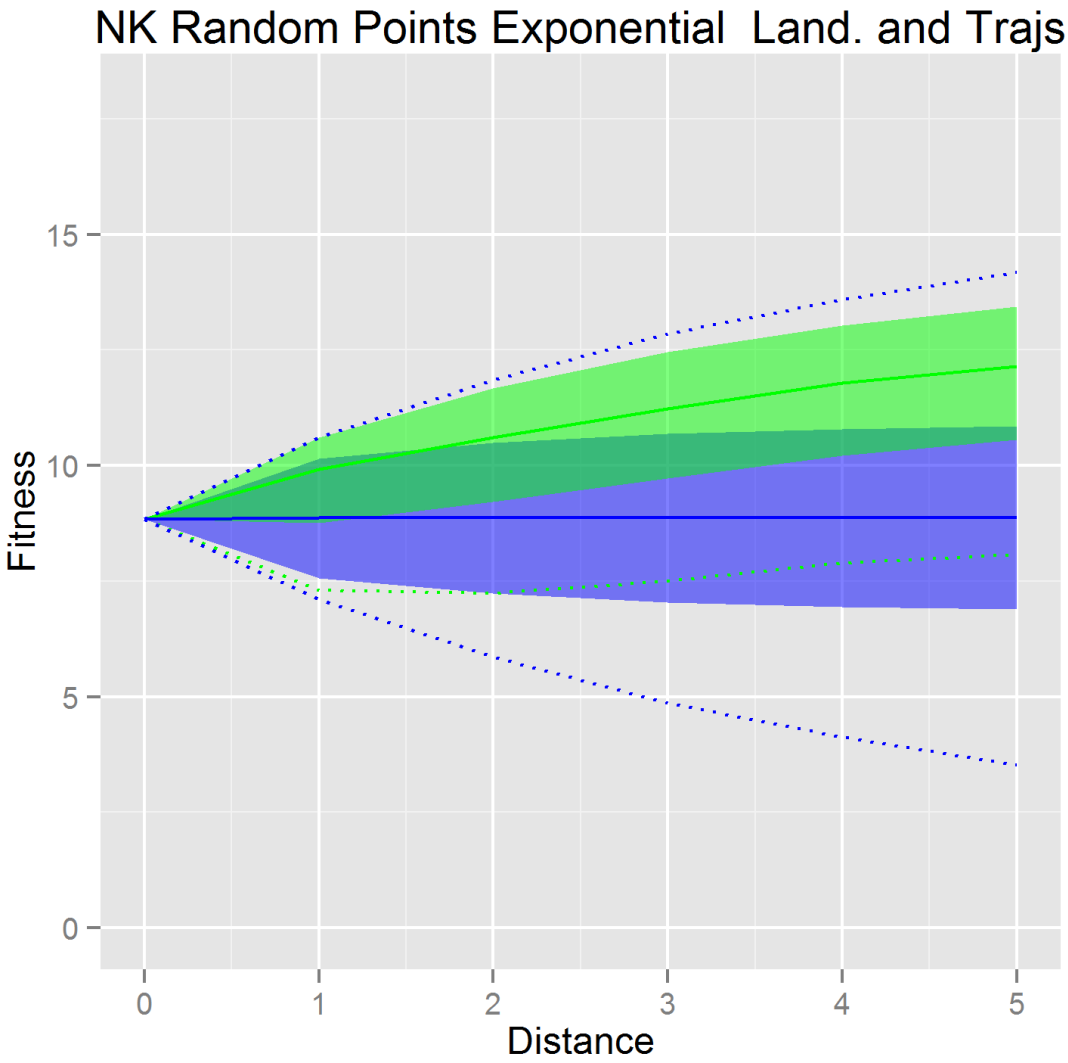


Figure 5.31: Avida Random Points Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

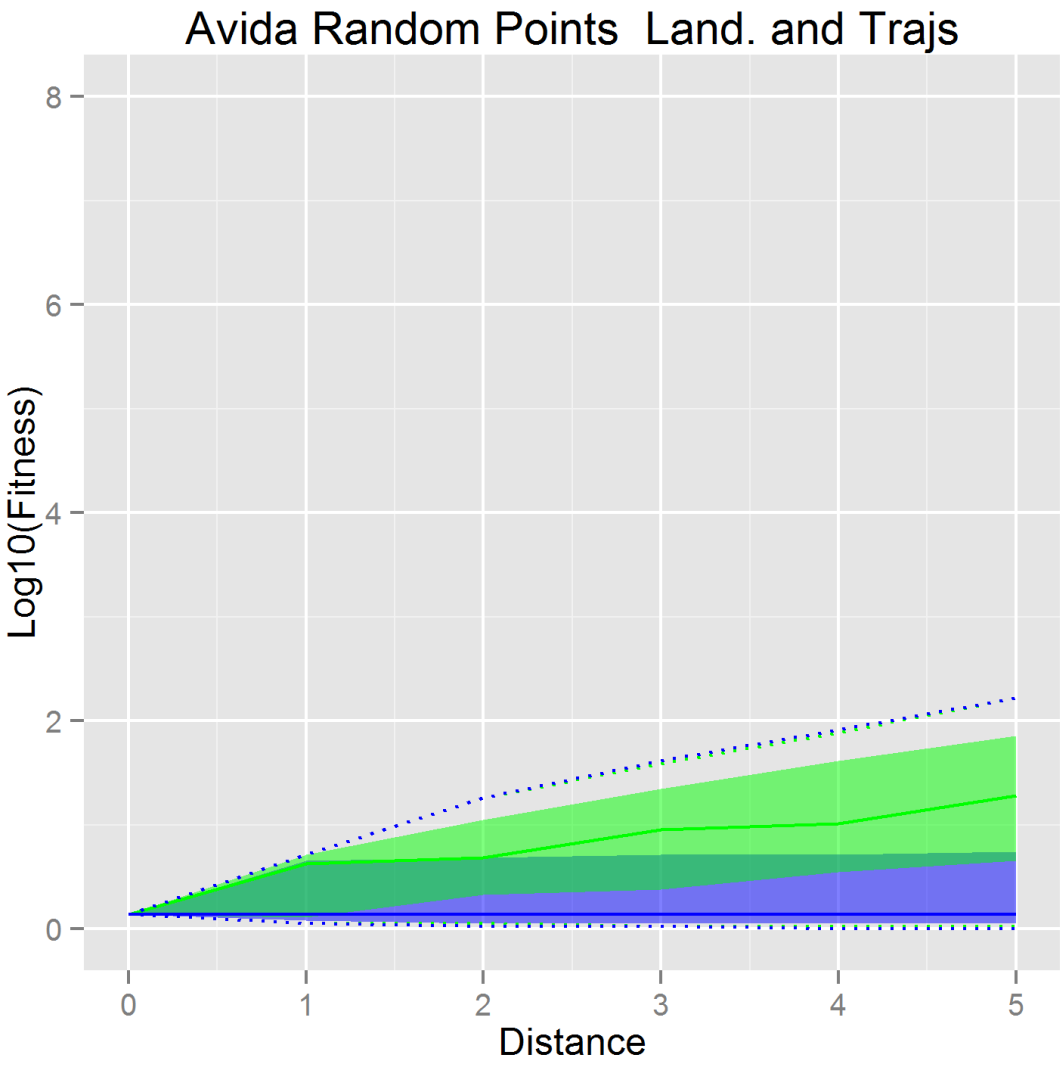


Figure 5.32: RNA Random Points Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

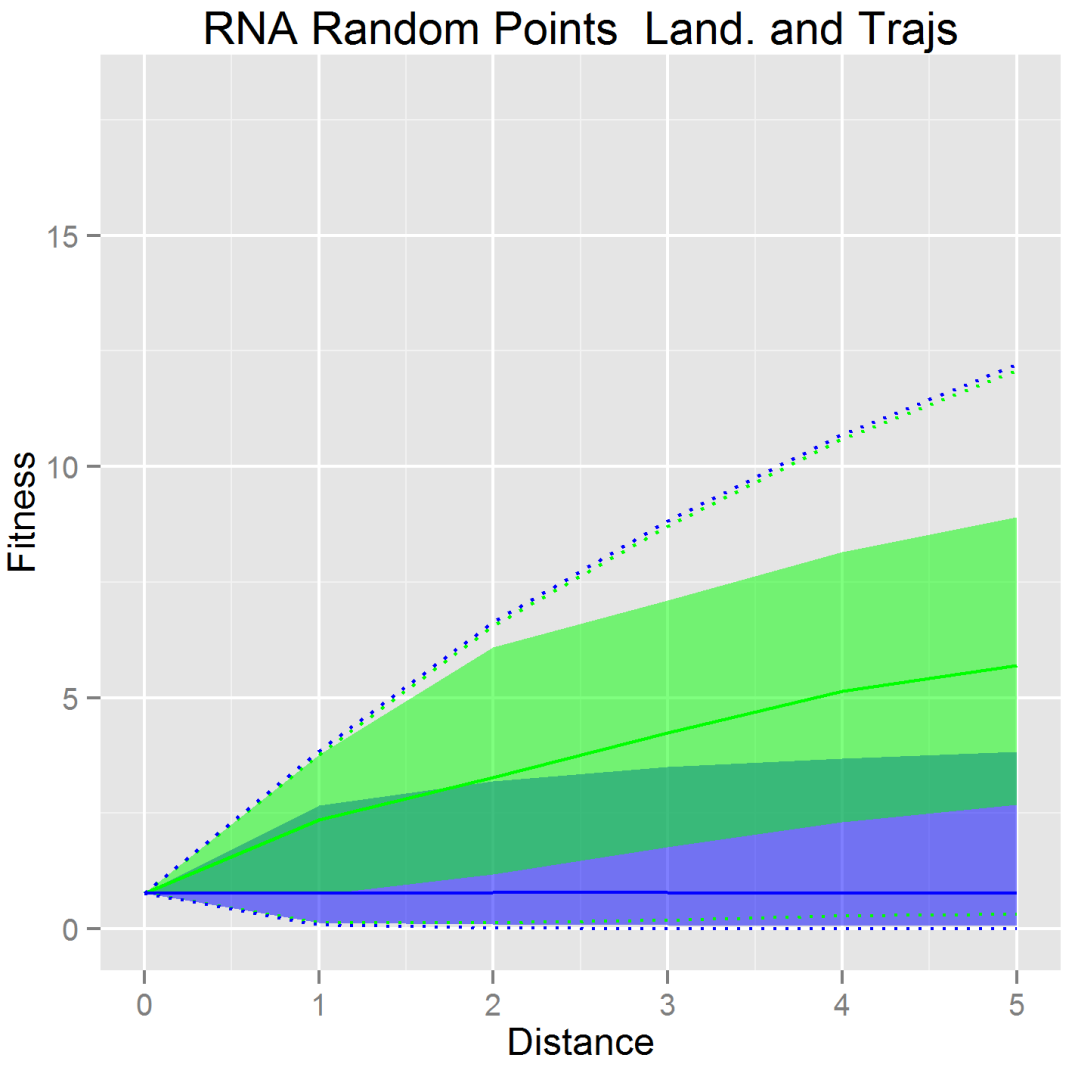
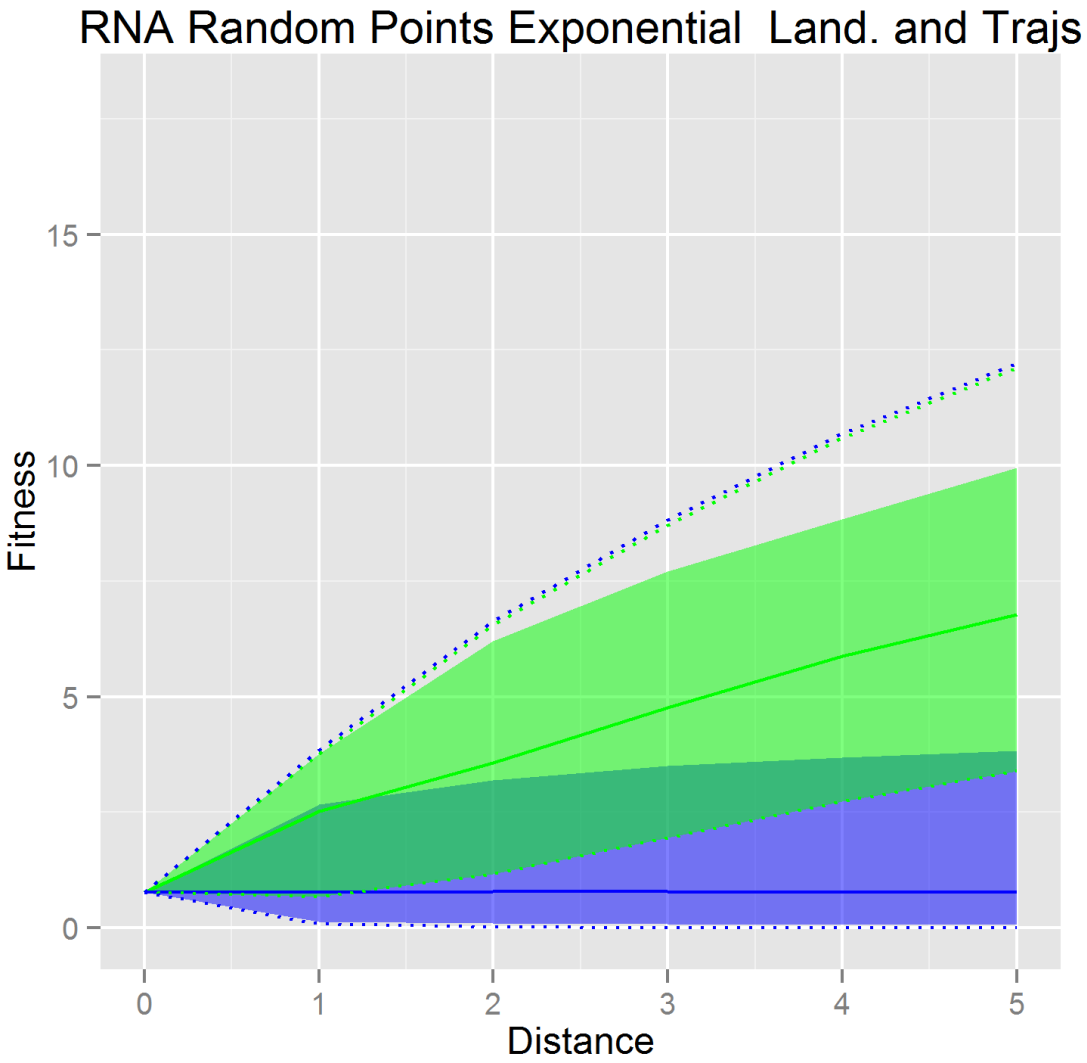


Figure 5.33: RNA Random Points Exponential Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid. Here, the higher selection strength results in a stronger improvement bias in the trajectories relative to the RNA Points Trajectory Visualization.



5.8 Repeated Trajectories Peaks—The fate of trajectories

I repeat the trajectory analysis, but this time for the analyses starting at peaks. Once again, these graphs depict the possibility space compared to what the evolutionary trajectories actually achieve.

Figure 5.34: NK Peaks Landscape and Trajectory Fitness. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

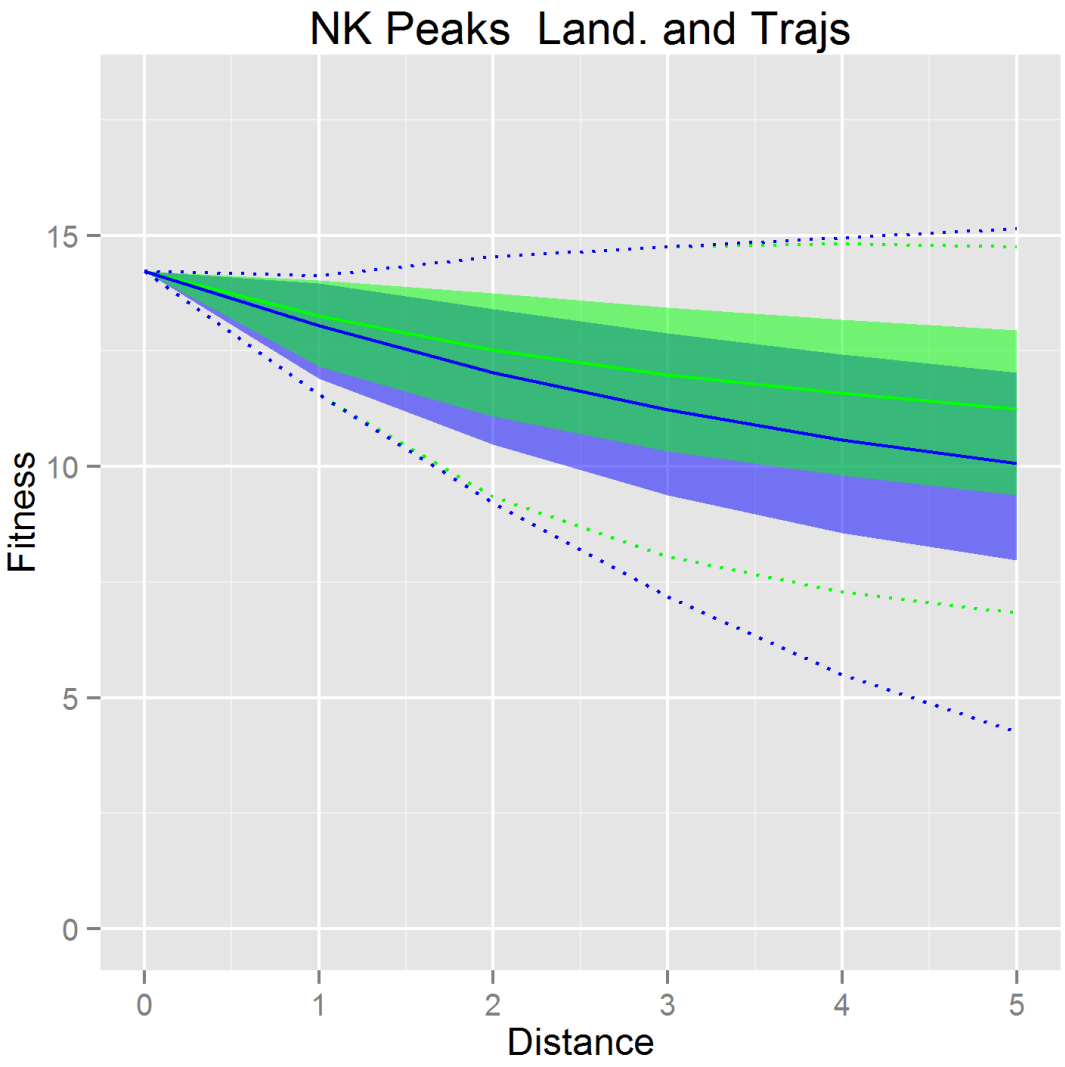


Figure 5.35: NK Peaks Exponential Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

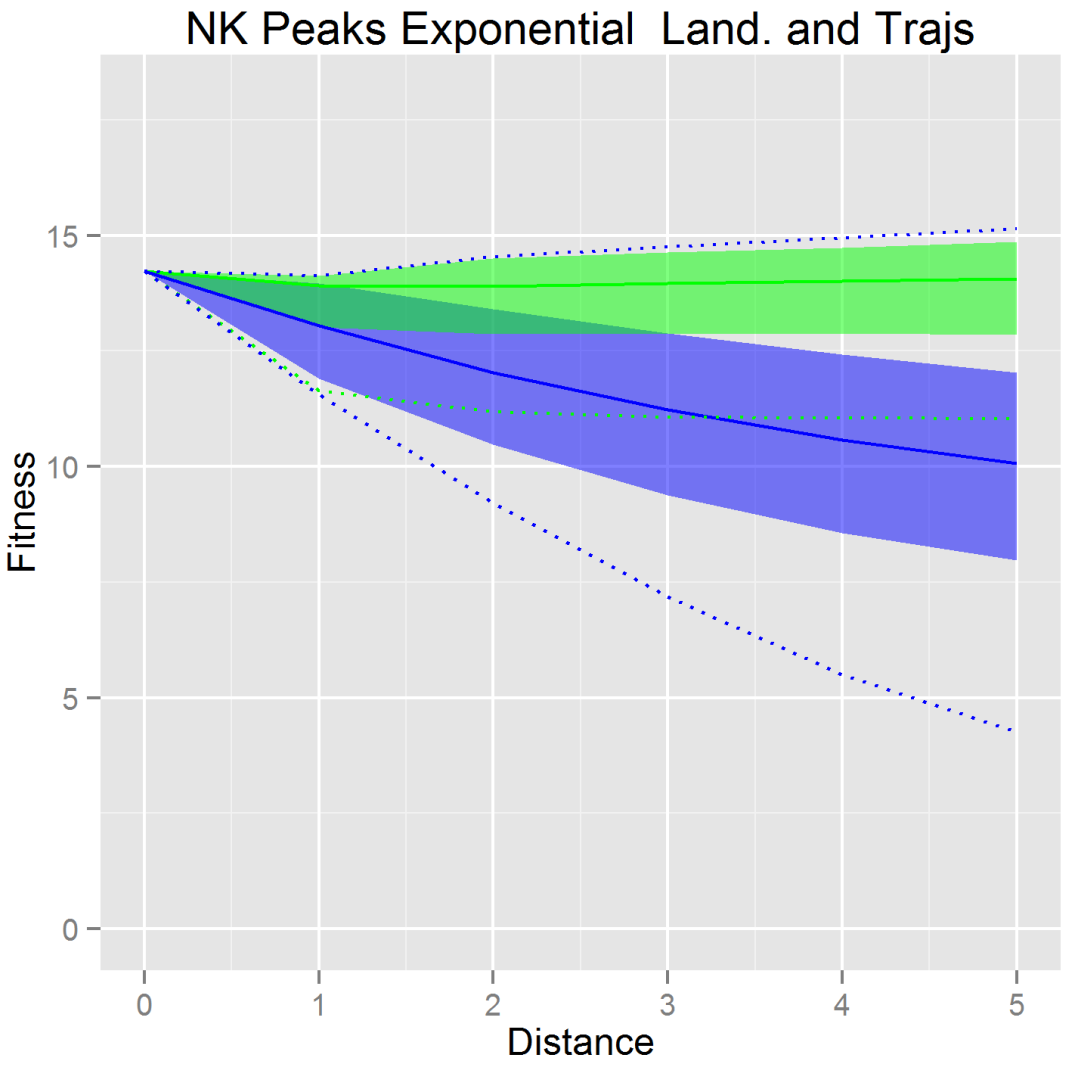


Figure 5.36: Avida Peaks Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

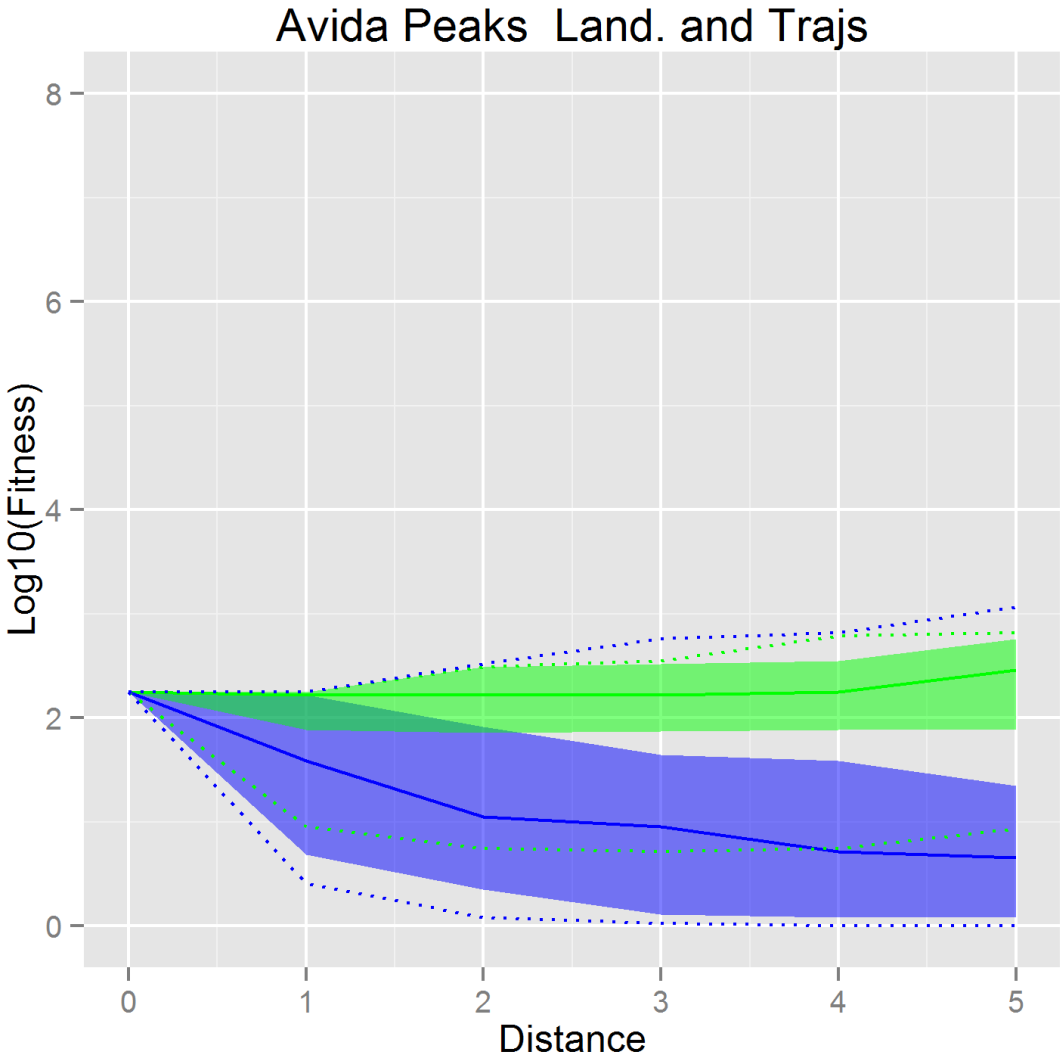


Figure 5.37: RNA Peaks Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.

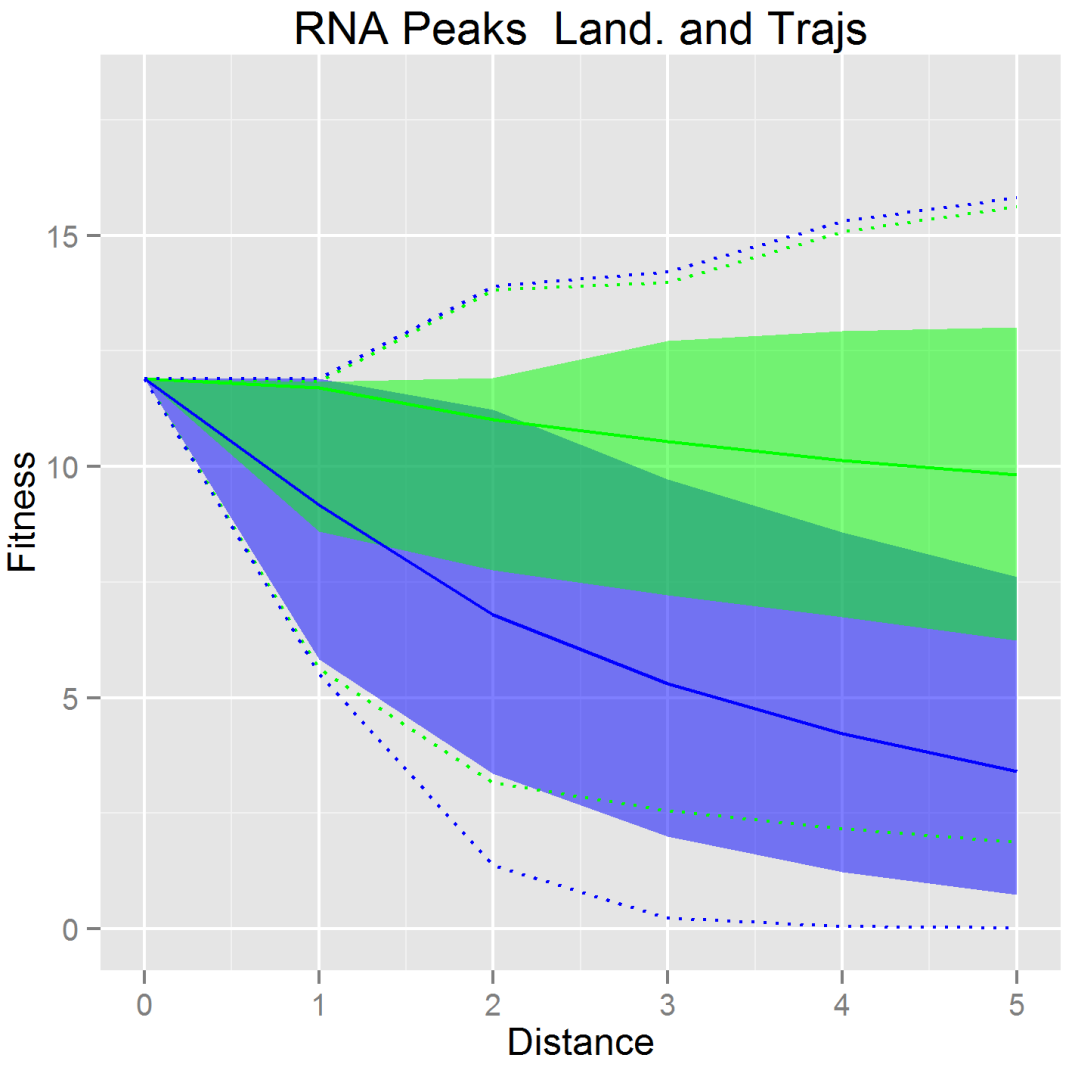
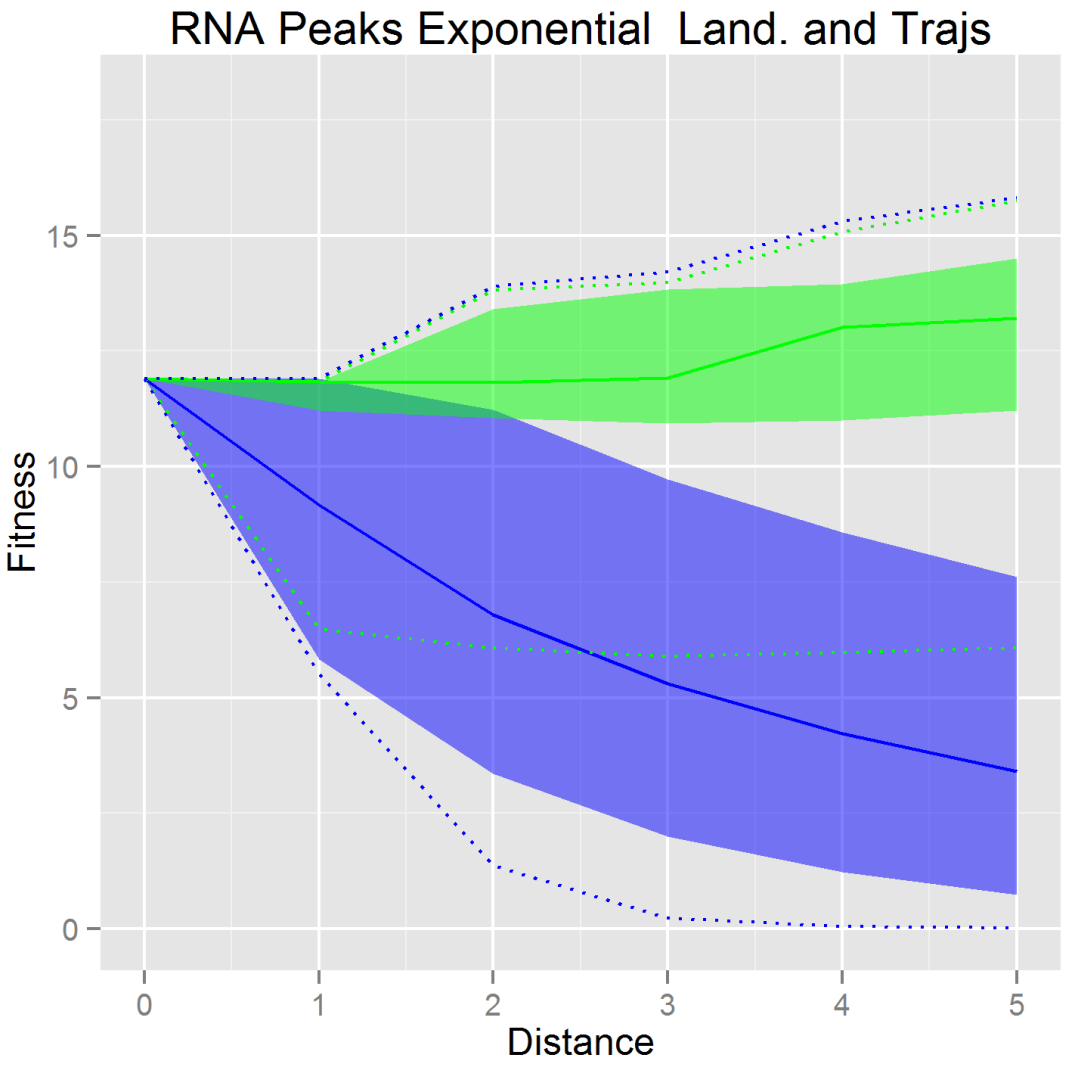


Figure 5.38: RNA Peaks Exponential Trajectory Visualization. The landscape is in blue and the trajectories are in green. The dotted lines are the median-minimum and median-maximum and the shaded area for each is the region between the median-5th and median-95th quantiles. The median-median is solid.



5.9 Discussion

In this chapter, I examined the local dynamics of two datasets—random points and peaks for each of the three model landscapes and two exponential variants with stronger selection. I provide a technique for exhaustive landscape and trajectory visualization via the use of Bresenham's Line algorithm, which allows us to visualize these complex multidimensional landscapes in a new way, which may be useful for building intuition about complex landscapes.

Features such as the normal distribution of fitnesses of the NK landscape in Figure 5.1 stand in stark relief to the Avida landscape in Figure 5.4 and the RNA landscape in Figure 5.7, both of which feature a preponderance of low-fitness genotypes in the randomly selected points. This result is not surprising—as seen in Chapter 3 and previous work that the most common fitnesses in both Avida and RNA are extremely low, whereas in the NK landscape, the most common fitnesses are also the median fitness (extremely low fitnesses in the NK landscape are just as rare as extremely high fitnesses). These visualizations also show that most potential trajectories emanating from random points in the Avida and RNA landscapes lead to equally low fitness, whereas in the NK landscape, the most common potential trajectories lead from median fitness to median fitness. The main differences among these random-starting-point potential trajectories are obvious when the more extreme cases are examined; the NK landscape contains equally rare potential trajectories to higher paths and lower paths. Despite its prevalent use in studying epistasis due to the easy tuning parameter in K, this result suggests that the NK landscape may not be the best choice to represent biological systems. This problem is most severe when trying to represent organisms with phenotypic complexity of any sort, simply because in such landscapes, most symbol sequences are of low fitness by chance.

The degree of neutrality also distinguished the potential trajectory classifications among the three landscapes. The most common potential trajectory classification in the Avida landscape was three neutral, one beneficial, and one deleterious step. In contrast, as neutrality

was rare in the NK and RNA landscapes, three beneficial and two deleterious and two beneficial and three deleterious step pathways were about equally likely. Since these are random points, and we would expect, *a priori*, no particular bias in either direction. This prediction is borne out by the virtually identical beneficial and deleterious counts in Figure 5.1, Figure 5.4, and Figure 5.7.

Starting from peaks the potential trajectories in all three landscapes revert to the mean. For the NK landscape, this reversion is to the expected average fitness of 9, and for the RNA and Avida landscapes, this reversion is towards the lowest fitness values. There are an extremely small number of pathways to equal or higher fitness in each of the three studied landscapes. From peaks, the most common potential trajectory type in the NK landscape was five deleterious mutations, followed by four deleterious and one beneficial (again, neutral mutations are unlikely). In Avida, the most common characterization was three deleterious, one neutral, and one beneficial, followed by four deleterious and one beneficial. Finally, in RNA, the most common potential trajectory configuration was four deleterious and one beneficial, followed by five deleterious.

Next, I examined 10,000 evolutionary trajectories from random starting points in each of the three aforementioned landscapes, as well as the exponential variants for the NK and RNA landscapes as in Chapter 4. The results are visually striking and often distinguishable from the full range of potential trajectories possible in each landscape. For the Avida and RNA trajectories, the paths upward driven by evolution are clearly visible. This result contrasts with the strong horizontal bias in the landscape visualizations. In evolutionary trajectories on NK landscapes, the selective pressures skews the normality of the potential trajectories in a distinctly upwards trajectory. This pattern is more strongly biased in the NK Exponential and RNA Exponential as compared to the NK and RNA regimes, respectively.

Next, I generated an equivalent set of 10,000 evolutionary trajectories, this time starting from peaks instead of random points. The weak selection in the NK and RNA landscapes

is clearly evident as the fitness actually appears to decline over time in Figure 5.19 and Figure 5.22 in the trajectory graphs. This effect largely disappears in the exponential variants (Figure 5.25 and Figure 5.28). The Avida landscape, in contrast, has strong horizontal lines with a slight bias upwards.

As should be expected, the trajectory visualizations are consistent with the combined landscape-and-trajectory summaries; for instance in Figure 5.34; the NK low selection regime trajectories do decline in fitness, whereas the NK exponential in Figure 5.35 does significantly better—performing above the 95th quantile of the landscape. In general, also consistent with expectations, evolution was favoring the top half of trajectories. Even with weak selection, evolution seems to find improvements for random points, outperforming the landscape average. However, in the case of peaks, strong selection was needed to avoid degradation of trajectories along with the landscape.

In summary, I have presented new visualization techniques which provide insight into the nature and structure of different landscapes and how evolutionary trajectories actually navigate them.

Chapter 6

Conclusions

Populations evolve in predictable ways because of common descent and the nature of fitness landscapes. As populations adapt, they accumulate information about the fitness landscape which they inhabit. In this work, I have used computational techniques to investigate the process of evolution by studying the acquisition of information, the structure of fitness landscapes, and the transient dynamics of evolving populations on these landscapes.

In Chapter 2, I established that error from common descent in complexity estimates can be sizable, present a correction to compensate for its effects, and test the correction empirically. I showed that the correction performed well in the presence of low mutation rates, but once mutation rates increased past a per-site rate of 0.01 it started to do poorly, likely because of the underlying assumption that informative sites have no entropy. This assumption breaks as the mutation rate increases and it becomes increasingly hard for a population to stay on a peak. A natural future direction for this work is to further refine the model so that it is less sensitive to this condition, but the real prize would be in a closed form characterization of the magnitude of this effect. A further important step would be to test this technique in other systems as a prelude to demonstrating its utility in real biological systems.

In Chapter 3, I investigated instantiations of large NK, reduced-instruction Avida, and

RNA fitness landscapes of 68 billion genotypes each and studied the landscape structure, specifically peak structure and distribution. In all three landscapes, I found that peaks of higher fitness tended to have higher peaks in their neighborhood. I also found that peaks are connected in the NK and Avida landscapes but are not in the RNA landscape. Finally, I found that while peak fitness is negatively autocorrelated at high distance in NK and Avida landscapes, peaks in the RNA landscape at maximum distance are positively correlated. This result may point to the fact that the complement of some stable RNA sequences may be disproportionately likely to be stable themselves. In the future, I would like to extend my analyses to more than just peak structure—other structural properties such as local roughness have been proposed [Lobkovsky et al., 2011]. Also, a non-binary definition of peak may be useful for structural understanding especially when there is substantial neutrality or near-neutrality.

In Chapter 4, I studied the concept of basins of attraction. First, I presented a novel structural method of measuring basins based on the Page Rank algorithm, which I applied to NK landscapes. With this method, I confirmed previous findings that basin size increases exponentially with fitness, and that as the K parameter increases, the average basin size increases while the largest basins decrease. I also examined NKp variants of NK landscapes with neutrality and detected neutral networks that can retain mass over time, and found that linking double mutants with my method increased the strength of the exponential relationship between fitness and basin size. In the second section, I experimentally derived basins by using 100,000 evolutionary runs starting at random points on each landscape. I again found an exponential correlation between the peak fitness and basin size in both the NK and RNA landscapes, but not in the Avida landscape. I also demonstrated that in all three landscapes, the distance from the origin point approached the average distance of 13.5 between all pairs of points in the landscape, which suggests either overlapping basins or that basins may have long and narrow tendrils. Further investigation is needed to tease this particular distinction apart. Finally, I showed that in all three model landscapes, evolutionary trajectories spend

a disproportionate amount of their time in peaks. For this work, it is also natural to want to adjust the evolutionary parameters of population size and mutation rate and understand how that affects landscape traversal, the accumulation in basins over time, and final fitness, which would have important consequences for both theoretical and applied work.

In Chapter 5, I looked at the local dynamics of fitness landscapes. Starting from both random points and peaks, I exhaustively studied mutants up to five steps out from each source genotype and then examined the results of 10,000 evolutionary runs originating from each studied genotype. I provided a novel way to visualize both the roughly 250 million potential trajectories out to length five, for each point in each treatment, and also applied this visualization to evolutionary trajectories. There are many future directions for this particular work; in the work presented here, I have focused on techniques and visualization, but the next logical step would be to investigate epistasis and its relationship to the trajectories realized by evolution. A further goal is to gain a more in-depth understanding about how these different landscapes are both similar and different, and how sampling and full-structural analyses might agree and disagree. There are likely some landscapes that defy exhaustive analysis, or which are prohibitively expensive to test; so a broader understanding of how these might contribute to mapping the global landscape would be useful. Finally, I hope to understand more clearly how evolutionary pathways traverse from peak to peak. Peaks play an important and consistent role in evolution through time: in Chapter 4, I saw that fitness increases over time, which means that evolution must be finding pathways between successively higher peaks.

To accomplish this work, I have relied heavily on computational techniques to produce exhaustive analyses of the structure and consequences of fitness landscapes. As computational power grows it should be possible to map increasingly large and realistic fitness landscapes and provide more insight into the nature and workings of these landscapes and their relevance for evolution.

BIBLIOGRAPHY

BIBLIOGRAPHY

[Adami, 2002] Adami, C. (2002). What is complexity? *Bioessays*, 24(12):1085–1094.

[Adami, 2004] Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, 1:3.

[Adami et al., 1995] Adami, C., Brown, C. T., and Haggerty, M. R. (1995). Abundance-distributions in artificial life and stochastic models: "age and area" revisited. In Morn, F., Moreno, A., Guervs, J. J. M., and Chacn, P., editors, *ECAL*, volume 929 of *Lecture Notes in Computer Science*, pages 503–514. Springer.

[Adami and Cerf, 2000] Adami, C. and Cerf, N. (2000). Physical complexity of symbolic sequences. *Physica D: Nonlinear Phenomena*, 137(12):62 – 69.

[Adami et al., 2000] Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity. In *Proceedings of the National Academy of Sciences*, pages 4463–4468.

[Balding, 2006] Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–791.

[Barnett, 1998] Barnett, L. (1998). Ruggedness and neutrality - the nkp family of fitness landscapes. In *Alive VI: Sixth International Conference on Articial Life*, pages 18–27. MIT Press.

[Bedau and Brown, 1999] Bedau, M. A. and Brown, C. T. (1999). Visualizing evolutionary activity of genotypes. *Artificial Life*, 5(1):17–35.

[Beerenwinkel et al., 2007] Beerenwinkel, N., Pachter, L., Sturmfels, B., Elena, S., and Lenski, R. (2007). Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology*, 7(1):60.

[Bickel et al., 1996] Bickel, P. J., Cosman, P. C., Olshen, R. A., Spector, P. C., Rodrigo, A. G., and Mullins, J. I. (1996). Covariability of V3 Loop Amino Acids. *AIDS Research and Human Retroviruses*, 12(15):1401–1411.

[Bresenham, 1987] Bresenham, J. (1987). Ambiguities in incremental line rastering. *Computer Graphics and Applications, IEEE*, 7(5):31–43.

[Chao and Shen, 2003] Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10:429–443.

[Covert et al., 2013] Covert, A. W., Lenski, R. E., Wilke, C. O., and Ofria, C. (2013). Ex-

periments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences*, 110(34):E3171–E3178.

[Derrida and Bessis, 1988] Derrida, B. and Bessis, D. (1988). Statistical properties of valleys in the annealed random map model. *Journal of Physics A: Mathematical and General*, 21(9):L509.

[Derrida and Peliti, 1991] Derrida, B. and Peliti, L. (1991). Evolution in a flat fitness landscape. *Bulletin of Mathematical Biology*, 53:355–382. 10.1007/BF02460723.

[Desai et al., 2007] Desai, M., Fisher, D., and Murray, A. (2007). The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17:385–394.

[Dudk et al., 2005] Dudk, M., Schapire, R. E., and Phillips, S. J. (2005). Correcting sample selection bias in maximum entropy density estimation. In *In Advances in Neural Information Processing Systems*.

[Durrett, 2008] Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Probability and Its Applications. Springer.

[Eigen, 1971] Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523. 10.1007/BF00623322.

[Eigen et al., 1988] Eigen, M., McCaskill, J., and Schuster, P. (1988). Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891.

[Eigen et al., 1989] Eigen, M., McCaskill, J., and Schuster, P. (1989). *The Molecular Quasi-Species*, pages 149–263. John Wiley & Sons, Inc.

[Eigen and Schuster, 1977] Eigen, M. and Schuster, P. (1977). The Hypercycle: a Principle of Natural Self-Organisation, Part A. *Naturwissenschaften*, 64(11):541–565.

[Ewens, 1972] Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87 – 112.

[Ezkurdia et al., 2014] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*.

[Fisher, 1930] Fisher, R. (1930). *The genetical theory of natural selection*. Clarendon Press, Oxford.

[Flores-Moya et al., 2012] Flores-Moya, A., Rouco, M., Garca-Snchez, M. J., Garca-Balboa, C., Gonzlez, R., Costas, E., and Lpez-Rodas, V. (2012). Effects of adaptation, chance, and

history on the evolution of the toxic dinoflagellate alexandrium minutum under selection of increased temperature and acidification. *Ecology and Evolution*, 2(6):1251–1259.

[Fontana et al., 1993] Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D., and Schuster, P. (1993). Rna folding and combinatory landscapes. *Phys. Rev. E*, 47:2083–2099.

[Fu, 1995] Fu, Y. X. (1995). Statistical Properties of Segregating Sites. *Theoretical Population Biology*, 48(2):172–197.

[Fu and Li, 1993] Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709.

[Gavrilets, 2003] Gavrilets, S. (2003). Perspective: Models of speciation: What have we learned in 40 years? *Evolution*, 57(10):pp. 2197–2215.

[Gavrilets, 2004] Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton University Press.

[Gerrish and Lenski, 1998] Gerrish, P. and Lenski, R. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103:127–144.

[Gould, 1989] Gould, S. J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton and Company, New York.

[Handel and Rozen, 2009] Handel, A. and Rozen, D. (2009). The impact of population size on the evolution of asexual microbes on smooth versus rugged fitness landscapes. *BMC Evolutionary Biology*, 9(1):236.

[Hartl and Clark, 2006] Hartl, D. L. and Clark, A. G. (2006). *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc., 4th edition.

[Hausser and Strimmer, 2009] Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, 10:1469–1484.

[Hazen et al., 2007] Hazen, R. M., Griffin, P. L., Carothers, J. M., and Szostak, J. W. (2007). Functional information and the emergence of biocomplexity. *Proceedings of the National Academy of Sciences*, 104(Suppl 1):8574–8581.

[Huang et al., 2004] Huang, W., C., O., and Torng, E. (2004). Measuring biological complexity in digital organisms. In *Artificial Life IX*.

[Kauffman, 1993] Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, Oxford.

[Kauffman and Levin, 1987] Kauffman, S. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11 – 45.

[Kimura, 1968] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(129):624–6.

[Kogenaru et al., 2009] Kogenaru, M., de Vos, M. G. J., and Tans, S. J. (2009). Revealing evolutionary pathways by fitness landscape reconstruction. *Critical Reviews in Biochemistry and Molecular Biology*, 44(4):169–174. PMID: 19552615.

[Korber et al., 1993] Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180.

[Kryazhimskiy et al., 2014] Kryazhimskiy, S., Rice, D. P., Jerison, E. R., and Desai, M. M. (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522.

[Kryazhimskiy et al., 2009] Kryazhimskiy, S., Tkacik, G., and Plotkin, J. B. (2009). The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences*, 106(44):18638–18643.

[Lenski et al., 2003] Lenski, R., Ofria, C., Pennock, R., and Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

[Lenski et al., 1999] Lenski, R. E., Ofria, C., Collier, T. C., and Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400(6745):661–4.

[Lobkovsky et al., 2011] Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2011). Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol*, 7(12):e1002302.

[Lorenz et al., 2011] Lorenz, R., Bernhart, S., Honer zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P., and Hofacker, I. (2011). Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

[Maniatis et al., 2002] Maniatis, N., Collins, A., Xu, C. F., Mccarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X., and Morton, N. E. (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A*, 99(4):2228–33.

[Minka, 2000] Minka, T. P. (2000). Bayesian inference, entropy, and the multinomial distribution.

[Nemenman et al., 2002] Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. In *Advances in Neural Information Processing Systems 14*. MIT Press.

[Ochoa et al., 2008] Ochoa, G., Tomassini, M., Vérel, S., and Darabos, C. (2008). A study of nk landscapes' basins and local optima networks. *CoRR*, abs/0810.3484.

[Ochoa et al., 2010] Ochoa, G., Vérel, S., and Tomassini, M. (2010). First-improvement vs. best-improvement local optima networks of nk landscapes. In *PPSN (1)*, pages 104–113.

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

[Østman et al., 2010] Østman, B., Hintze, A., and Adami, C. (2010). Critical properties of complex fitness landscapes. In *Proc. 12th Intern. Conference on Artificial Life*.

[Otwinowski and Plotkin, 2014] Otwinowski, J. and Plotkin, J. B. (2014). Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences*, 111(22):E2301–E2309.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[Paninski, 2003] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253.

[Podgornaia and Laub, 2015] Podgornaia, A. I. and Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677.

[Poelwijk et al., 2007] Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., and Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386.

[Roest et al., 2000] Roest, C. H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W., and Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.*, 25(2):235–238.

[Schneider, 2000] Schneider, T. D. (2000). Evolution of biological information. *Nucleic Acids Research*, 28(14):2794–2799.

[Schneider et al., 1986] Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431.

[Schürmann and Grassberger, 2002] Schürmann, T. and Grassberger, P. (2002). Entropy

estimation of symbol sequences. *Arxiv preprint cond-mat/0203436*.

[Service et al., 2001] Service, S. K., Ophoff, R. A., and Freimer, N. B. (2001). The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet*, 10(5):545–51.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[St Onge et al., 2007] St Onge, R., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R., Nislow, C., Roth, F., and Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet*, 39(2):199–206.

[Stadler et al., 2001] Stadler, B. M., Stadler, P. F., Wagner, G. P., and Fontana, W. (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213(2):241 – 274.

[Strelioff et al., 2010] Strelioff, C. C., Lenski, R. E., and Ofria, C. (2010). Evolutionary dynamics, epistatic interactions, and biological information. *Journal of Theoretical Biology*, 266(4):584 – 594.

[Szendro et al., 2013] Szendro, I. G., Franke, J., de Visser, J. A. G. M., and Krug, J. (2013). Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences*, 110(2):571–576.

[Tajima, 1989] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA Polymorphism. *Genetics Society of America*, 123:585–595.

[Tomassini et al., 2008] Tomassini, M., Vérel, S., and Ochoa, G. (2008). Complex-network analysis of combinatorial spaces: The *nk* landscape case. *Phys. Rev. E*, 78:066114.

[Travisano et al., 1995] Travisano, M., Mongold, J., Bennett, A., and Lenski, R. (1995). Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194):87–90.

[Tufts et al., 2014] Tufts, D. M., Natarajan, C., Revsbech, I. G., Projecto-Garcia, J., Hoffmann, F. G., Weber, R. E., Fago, A., Moriyama, H., and Storz, J. F. (2014). Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Molecular Biology and Evolution*.

[van Nimwegen et al., 1999] van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, 96(17):9716–9720.

[Vérel et al., 2011] Vérel, S., Ochoa, G., and Tomassini, M. (2011). Local optima networks

of nk landscapes with neutrality. *CoRR*, abs/1107.4162.

[Wagenaar and Adami, 2004] Wagenaar, D. A. and Adami, C. (2004). Influence of chance, history, and adaptation on digital evolution. *Artificial Life*.

[Wang and Lee, 2007] Wang, Q. and Lee, C. (2007). Distinguishing functional amino acid covariation from background linkage disequilibrium in hiv protease and reverse transcriptase. *PLoS ONE*, 2(8):e814.

[Watterson, 1975] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256 – 276.

[Watterson, 1977] Watterson, G. A. (1977). Heterosis or neutrality? *Genetics*, 85:789–814.

[Weinberger, 1991] Weinberger, E. D. (1991). Local properties of kauffman's N - k model: A tunably rugged energy landscape. *Phys. Rev. A*, 44:6399–6413.

[Weinberger, 2002] Weinberger, E. D. (2002). A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems*, 66(3):105 – 119.

[Wilke, 2001] Wilke, C. (2001). Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology*, 63:715–730. 10.1006/bulm.2001.0244.

[Wilke, 2005] Wilke, C. (2005). Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5(1):44.

[Wilke, 2004] Wilke, C. O. (2004). The speed of adaptation in large asexual populations. *Genetics*, 167(4):2045–2053.

[Wilke et al., 2001] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412:331–333.

[Wright, 1931] Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

[Wright, 1932] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc.Int.Cong.Gen.*, 1:356–366.

[Zhang et al., 1990] Zhang, Y. C., Serva, M., and Polikarpov, M. (1990). Diffusion reproduction processes. *Journal of Statistical Physics*, 58:849–861. 10.1007/BF01026554.