ON THE CONSTRUCTIVE POWER OF ECOLOGY IN OPEN-ENDED EVOLVING
SYSTEMS

By

Emily L. Dolson

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

2019

ABSTRACT

ON THE CONSTRUCTIVE POWER OF ECOLOGY IN OPEN-ENDED EVOLVING
SYSTEMS

By

Emily L. Dolson

Ecology is a powerful force for unlocking the full potential of evolution. Ecological interactions create feedback loops that promote diversification, both in natural ecosystems and in more applied evolutionary computation frameworks. However, we currently lack a strong theoretical framework that predicts how a given ecological community will evolve. Such a framework would allow us to better understand and anticipate change in evolving systems and facilitate the harnessing of ecology as a tool for guiding evolution. In this work, I develop tools to begin the development of such theory in computational systems. Using these tools, I show that different mechanisms for creating ecology (e.g. spatial structure and varied competition schemes) produce radically different community structures and evolutionary outcomes. I explore the implications of these differences in the context of evolutionary computation and biology.

This dissertation is dedicated to my parents, who made this work possible by fostering my love of science for the past almost 28 years.

ACKNOWLEDGEMENTS

Department (particularly Lisa Meeden).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Over the past decade, there has been an increasing awareness of the profound impact that ecological and evolutionary dynamics have on each other (Schoener, 2011; Hendry, 2016). Ecological interactions shape the underlying fitness landscape that is traversed by evolving populations. At the same time, evolutionary forces continually adjust the composition of organisms that make up that ecosystem. As such, ever-shifting fitness landscapes and fluctuating ecosystems continually fuel each other's change.

At the heart of these dynamics is diversity. Evolution requires diversity within a group for selection to act on; low-diversity populations typically have low evolutionary potential (Lavergne and Molofsky, 2007; McDonald and Linde, 2002; Walker and Ofria, 2012). Ecological communities are, by definition, diverse groups of species and one of the fundamental questions in ecology is how so many species can stably coexist; communities that are more diverse tend to be more stable, productive, and resilient (Hooper et al., 2005; Tilman et al., 2014). In both ecology and evolution, there are some processes that should clearly reduce diversity and some that should promote – or at least preserve – it. Improved understanding of these processes is critical to creating theory to predict how biological systems will respond to perturbations, particularly as humans continue to drastically change the environment around us, and our attempts at remediation often backfire due to ill-formed theoretical foundations.

Diversity is also of critical importance to evolutionary computation, a machine learning technique that draws inspiration from biology with the goal of finding diverse and effective solutions to challenging computational problems. The core idea behind evolutionary algorithms is that a population of potential solutions is maintained, where the highest quality ones are replicated (replacing lower-quality solution) and varied (typically through mutation or crossover). As a result, better solutions evolve. However, non-diverse populations can get stuck on good, but suboptimal solutions (local fitness peaks) from which all paths to the global optimum require deleterious mutations, a phenomenon is called premature convergence. Generating and maintaining diversity, in essence, is seen as a technique for increasing the amount of exploration in a system and avoiding premature convergence. While evolutionary computation researchers have developed a wide range of techniques for promoting diversity, the field currently lacks a theoretical framework for understanding how they relate to each other or predicting which would be most appropriate in a given scenario.

Criss-crossing biological and computational areas of research is the work of scientists studying artificial life. In this field, researchers create artificial lifelike constructs in the form of computer programs, robots, or synthetic organisms. These systems make it possible to address questions that would be intractable in natural systems. In particular, artificial systems make it possible to perform experiments at spatiotemporal scales that would otherwise be infeasible in the lab or field. As a result, scientists have used artificial life to understand the hallmarks of living systems in nature, such as long-term ecosystem development, open-ended evolutionary dynamics, and the origins of complex traits and behaviors. Complex ecologies appear to be a critical driver of each of these phenomena, helping to generate novelty that gets incorporated back into the system.

Untangling eco-evolutionary dynamics has proven challenging across all of these fields. These dynamics are governed by a web of feedback loops, stochastic events, and emergent properties. Often simply adding a property, such as spatial structure, to a system will completely change the outcome. To add further complications, these processes often play out

over vast spatiotemporal scales, making them difficult to manipulate in a controlled experiment. As such, digital systems are a valuable tool for understanding the behavior of these complex systems (Adami, 2002). They allow us to untangle feedback loops by performing experiments in which we disable some processes. For instance, sometimes it would be useful to separate the effects of ecological and evolutionary dynamics from each other so that we can understand what the exact drivers of a result are; such understanding is important for predicting how our results map to other systems, in which the relative rates of these two processes may be different. In a wet lab or field system, such an experiment is effectively impossible to carry out cleanly. Another obstacle to understanding eco-evolutionary dynamics is that evolution experiments often require thousands of generations. Even for rapidly-reproducing organisms like bacteria, it takes most of a year just to reach 2000 generations. Similarly, collecting ecological data at a high enough resolution to answer many questions is often incredibly labor-intensive.

These problems can be alleviated by performing parallel digital evolution experiments. Since digital systems give us perfect control over our experimental design, tasks such as isolating ecological and evolutionary dynamics are possible; we can simply disable mutations and prevent any evolution from occurring, leaving only the purely ecological effects. Such controls make it possible for us to delve into mechanisms that cannot be addressed in other systems. Digital systems also generally enable evolution to happen substantially faster than natural systems. In Avida (discussed below), for example, 2,000 generations take less than 24 hours. This speed makes digital evolution a particularly valuable tool for performing exploratory digital experiments to develop theory that can then be tested in a natural system once we have determined the precise questions that it is worth investing the time to study in the lab or field.

1.1 Why use digital evolution to understand eco-evolutionary dynamics in biology?

A wide range of digital systems for studying evolution exist. Perhaps the most popular of these systems, and the one that I will use here, is the Avida Digital Evolution Platform (Ofria and Wilke, 2004). Avida has all of the properties that I will argue are critical for the study of open-ended eco-evolutionary dynamics. Each organism in Avida is a self-replicating computer program, with a genome made up of computer instructions. Experiments generally begin by seeding the population with a single program containing a sequence of code that copies each instruction in its genome into its offspring. Every time an instruction is copied, there is a small probability that it will mutate to be a different instruction. Organisms in Avida inhabit a lattice of cells, with each cell containing at most one organism. When an organism is done copying its program, the resulting offspring is placed in an adjacent cell, overwriting any organism previously occupying that cell. This dynamic creates competition for space, which incentivizes organisms to copy themselves as fast as possible to reduce the chance of being overwritten before they reproduce. Organisms can increase their reproductive rate either by evolving to be more efficient or by evolving to perform computational tasks that the experimenter has associated with a bonus to CPU speed. When organisms successfully complete these tasks, they receive bonus CPU cycles each update relative to the other members of the population, meaning that they are able to execute more code per unit time. Since the Avida system has inheritance, variation, and selection (some organisms may have more offspring in their lifetime than others), these populations of digital organisms can evolve by natural selection.

Note that the conditions that are necessary and sufficient for evolution are not sufficient for the formation of ecological communities. Such communities require that there be multiple potential strategies for survival (i.e. niches) and that it be possible for groups of organisms using different strategies to coexist. There are a vast number of ways to add this property

to a system, many of which will be discussed later in this chapter.

We argue that digital evolution systems have a number of properties that make them well-suited to the task of studying eco-evolutionary dynamics:

1.1.1 Digital Evolution Systems are Individual-based

Because the digital evolution systems that I am working with are instantiations of evolution (as opposed to merely simulating evolutionary dynamics), they are necessarily individual-based. Each member of the population is an individual with some amount of agency, allowing for a broad range of survival traits to evolve. Experiments are conducted by modifying the rules governing the behavior of individuals or their interactions with each other and the environment. This approach is in contrast to systems that represent the population at a more abstract level, such as a Gaussian distribution of trait values. In these non-individual-based systems, experiments are more often performed by changing the equations that govern how the population changes over time. These two approaches generally target different levels of abstraction. Because digital evolution systems try to avoid introducing any preconceptions about the range of potential outcomes that could result from a given set of starting conditions, individual-based models are more appropriate in cases where the cumulative impact of individual behavior can lead to complex and nuanced population-level effects. Most open questions in eco-evolutionary dynamics fall squarely into this category, as they largely concern the large-scale trends that result from simple rules governing interactions between a large number of individuals.

1.1.2 Digital Evolution Systems can be Generalized

The underlying genetic representation of organisms in most digital evolution systems are computer programs, making them fundamentally different from DNA-based organisms. While this fact means that digital evolution is generally not a good fit for modeling any specific system on earth, it also provides a valuable framework for producing insights that

generalize across systems (Maynard Smith, 1992; Kawecki et al., 2012; Ofria and Wilke, 2004; MacPherson and Gras, 2016). Traditionally, general inferences about biology are drawn by observing the same behavior across multiple model systems (e.g. *Drosophila*, *E. coli*, and *Arabadopsis*). The more distinct these systems are from each other the better, as any differences reduce the chance that the observed results were caused by an idiosyncrasy of a specific subset of the phylogeny of life on earth. However, all of life on earth shares some common ancestry (Maynard Smith, 1992). Thus, even results that can be replicated across many model systems are not necessarily unavoidable outcomes of evolution that will always be true. If we observe a pattern in both digital and more natural systems, however, that is strong evidence that it will be true across most evolving systems.

The problem of generalizability has historically been even more challenging to overcome in ecology than in evolutionary biology, as evolution's set of three necessary and sufficient conditions provides a useful starting point for building ground-up conceptual models. As a result, ecologists are often forced to take a more top-down approach, trying to understand the complex ecosystems that exist on earth. This work is critical, but the process of finding general trends across ecosystems is arduous, and attempting to understand the mechanisms driving those trends is even more so. We can speed this process up by starting with comparisons to a digital system. This technique will allow us to more quickly determine whether results are likely to generalize before putting in the substantial work that would be required to conduct an comparable field study. More importantly, if we do not observe the same results in biology as we do in digital evolution, that is an indication that there is an important factor at play that we have not noticed. Such disparities will help us identify mistaken preconceptions and either help us build more realistic digital environments or point us toward critical areas where ecological theory may need further development.

1.1.3 Digital Evolution Systems have Realistic, yet Tractable Fitness Landscapes

Since evolution is a key component of eco-evolutionary dynamics, it is important to consider the adaptive landscape in which that evolution is occurring. Fitness landscapes, a common idea in evolutionary biology, map the fitnesses associated with various genotypes and the ease of mutating between those genotypes (Wright, 1932). When visualized in three dimensions, values of two traits are generally arranged along the x- and y-axes, with the fitness corresponding to each of these values shown as a surface on the z-axis. This surface generally has peaks that we expect the population to climb over time, and valleys that we expect will be challenging to cross. In reality, fitness landscapes exist in many more than three dimensions, which may impact the validity of the intuitions they provide (Gavrilets and Vose, 2005), but they are generally believed to be a useful mental model nonetheless. The fitness landscape on which a population is evolving plays a large role in determining the population's adaptive trajectory, often in relatively complex ways. Having a well-understood fitness landscape is critical, or at least helpful, for most research on evolutionary theory. The fitness landscape determines which mutations will be beneficial, how easy it is to arrive at the most fit phenotype, and, to some extent, the strength of selection – factors that are often critical to evolution research. When we study eco-evolutionary dynamics, we essentially layer the complexity of ecology on top of this landscape. Thus, we can maximize our chances of understanding why we observe the results that we do by studying eco-evolutionary dynamics in systems that are already well-understood in the absence of ecology. For instance, we may want to study which ecotypes are best at invading a given niche. If we already know that a given ecotype is a common stepping stone on the way to the adaptation associated with the niche, we have a much better sense of what results we should be expecting to see. Thus, if we observe different results, we will be able to recognize that there is something unexpected going on there and explore it in more depth.

It would be easy to create fitness landscapes that we understand well because they are

so simple. However, fitness landscapes in biology tend to be relatively complex, and some evolutionary dynamics are driven by this complexity. Thus, in order to investigate these dynamics, we need a somewhat complex fitness landscape. There are a variety of techniques for creating complex fitness landscapes for computational systems. Many of these techniques have been studied enough that we have a reasonably good intuition for the type of fitness landscapes that they create. NK models are a prime example (Kauffman and Levin, 1987; Ochoa et al., 2008). However, these fitness landscapes are generally either static, or being changed by some abiotic, exogenous factor. Thus, by definition, there is no ecology. Most of ecology can be thought of as a fitness landscape that is largely based on the other members of the population. As such, fitness landscapes in systems with ecology are subject to continuous endogenous change. This dynamic nature can make such fitness landscapes rather unwieldy. Every action that any organism takes changes the fitness landscape in which all other organisms are evolving. These changes last for an indefinite length of time. As a result, attempting to mathematically characterize these fitness landscapes would be incredibly challenging. Nevertheless, these ever-shifting fitness landscapes are a useful mathematical construct for the purposes of conceptualizing the complexity of adaptive dynamics in these ecosystems. In research on eco-evolutionary dynamics, this biotically-mediated fitness landscape is often layered on top of a static fitness landscape representing abiotic conditions.

The static underlying fitness landscape of Avida has been studied extensively under various configurations; this is true of other digital evolution systems as well, but for simplicity the rest of this section will focus on Avida. For research that requires a more complex fitness landscape, Avida includes the option to reward organisms with extra CPU cycles for completing various computational tasks. The most widely-used tasks are the set of all eight two-input Boolean logic functions in addition to the logical *not* function (hereafter referred to as logic-9). The fitness landscape of the logic-9 environment is well-studied, as it has been used in a large number of experiments (many of which are described later in this chapter). Moreover, it has some important structural similarities to fitness landscapes

in which biological organisms are evolving, in that the simpler tasks are somewhat related to the more complex tasks. While evolving the ability to do the simpler tasks is useful in its own right, the simpler tasks can also serve as building blocks for the more complex tasks (Lenski et al., 2003). Thus, Avida meets the criteria of having a complex-but-well-studied static fitness landscape making it well-suited for use in studying eco-evolutionary dynamics in a digital system.

Avida also makes it easy to incorporate ecology by layering on a fitness landscape based on biotic interactions. This layering is simplest in the case of competitive interactions. As previously mentioned, digital organisms in Avida can evolve to perform various computational tasks. Akin to evolving the ability to metabolize a new resource in biology, evolving a new task in Avida allows for the exploitation of a novel resource. By default, the resources associated with tasks in Avida are unlimited. However, we can introduce competitive interactions between organisms by limiting the quantity of these resources. This scarcity creates negative frequency-dependent selection as organisms compete for the resources. Any time an organism uses a resource it consumes a portion of it, making the ability to use that resource less valuable or all organisms. Avida also includes a number of other mechanisms by which organisms can influence each other's fitness landscapes, such as parasitism, predation, and cooperation.

1.2 Previous eco-evolutionary dynamics research using digital evolution

1.2.1 Competition

While the first major discoveries to come out of digital evolution research fell squarely into the category of pure evolutionary theory(Wilke et al., 2001; Lenski et al., 2003), simple ecological dynamics were soon explored. Cooper and Ofria demonstrated that limiting the

quantity of resources available was sufficient to evolve a stable and diverse ecosystem(Cooper and Ofria, 2003). They set up this experiment by associating each of the tasks from the logic-9 environment with a limited resource. Every time an organism performed the task and received the reward, it used up some of the resource, resulting in a situation where digital organisms performing tasks is effectively equivalent to biological organisms metabolizing chemical resources. This arrangement created negative frequency dependence; tasks being performed by fewer organisms were more valuable, just as inhabiting any underutilized niche in an ecosystem is advantageous. In prior experiments, where resources were unlimited, some phenotypes in the population were simply more fit than the others; the fitnesses did not change over time. Layering the shifting biotic fitness landscape on top of the static abiotic landscape created pressure for adaptive radiation and subsequent stabilizing selection on the different ecotypes.

Chow et al. expanded on this work, exploring the continuum from environments with very scarce resources to environments with resources so plentiful as to be effectively unlimited(Chow et al., 2004). They found that ecotypic diversity peaks at intermediate levels of resource availability. Resource availability, in this context, is roughly analogous to primary productivity; it determines the carrying capacity of each niche. Previous field research had found that species richness tends to be highest in environments with intermediate productivity, but there had been a lack of consensus over which properties of environments were driving this effect. Chow et al. replicated this finding in Avida, purely by varying the amounts of each resource that were present in the environment. This result demonstrated that limited resources alone are sufficient to explain the productivity-diversity relationship. This ability to isolate potential drivers of a phenomenon and determine which ones are sufficient to produce a given result is a major strength of individual-based digital systems.

These initial digital eco-evolutionary dynamics papers tell a compelling story about how evolution guides the long-term assembly of ecosystems. However, ecology also effects long-term evolutionary trends. Walker and Ofria explored how continued evolutionary innovation

plays out differently within ecological communities assembled at different levels of productivity (Walker and Ofria, 2012). Previous research on the connection between population diversity and evolutionary potential suggested that there is generally a positive correlation between these two variables. This finding makes sense, as a more diverse population is necessarily more spread out across the fitness landscape, increasing the probability that some members of the population will be close to an adaptive peak. However, Walker and Ofria found that the relationship between diversity and evolutionary potential is more complicated; evolutionary potential peaked at a different level of productivity than diversity did. This observation is an example of the difficulty of predicting eco-evolutionary feedbacks *a priori*. Digital systems give us the fine-grained control necessary to find such deviations from our expectations. In chapter 3, I present some follow-up work to these findings.

1.2.2 Facilitation

So far, I have discussed ecosystems assembled purely on the basis of competitive interactions. While these model ecosystems are useful for gaining basic insights into eco-evolutionary dynamics, ecosystems in nature involve a wide variety of different types of interactions among organisms. These include various forms of facilitation, that is species making it easier for other species to survive.

Perhaps the simplest of facilitation interactions is cross-feeding (also called syntrophy), wherein organisms of one ecotype feed off of byproducts generated by organisms of a different ecotype. Yet, even such a simple interaction can have unpredictable eco-evolutionary dynamics. Johnson and Wilke built an incredibly simple cross-feeding ecosystem in Avida by providing two resources, each of which created the other as a byproduct when it was metabolized (Johnson and Wilke, 2004). This type of environment produces an even stronger form of stabilizing selection than occurs in the purely competitive interactions in the logic-9 environment, as increased competition for one resource is combined with a decrease in the production of that resource (assuming a fixed population size). The result is a pair of

ecotypes that exhibit either Lotka-Volterra-like oscillatory dynamics or a more stable equilibrium. Which one of these regimes occurs appears to be largely determined by stochastic factors. More complex hierarchical cross-feeding food webs have also been constructed in Avida (Yedid et al., 2008, 2009, 2012).

Cross-feeding is a relatively simple form of facilitation as it is not susceptible to cheating. Cooperation, in which cheating is a possibility, poses a more challenging evolutionary problem. Nevertheless, cooperative species interactions play a critical role in many ecosystems (Goudard and Loreau, 2008; Frederickson et al., 2005). Understanding what evolutionary pressures can lead to such scenarios is critical to understanding this component of ecosystem assembly, but is, again, challenging to address in vitro or in vivo. Digital evolution has, in many cases, facilitated a deeper exploration of the intricate mechanisms that allow for the evolution of cooperative behaviors.

For example, based on inclusive fitness theory we would expect that genes that code for altruistic behavior should be more evolutionarily successful if they more effectively direct that altruistic behavior at other individuals with the same gene. Clune et al. tested this hypothesis by providing various altruistic instructions that could be included in organisms' genomes by mutation (Clune et al., 2010). When organisms executed these instructions, they would lose some of their CPU cycles and a neighboring organism would receive some number of CPU cycles. The neighbor that received this benefit was determined by the specific altruistic instruction that was used. The available instructions fell along a continuum of how accurately they could target the benefit at other altruistic organisms. The least accurate (other than the controls) was kin targeting, which donated the CPU cycles to a parent or offspring. Slightly more accurate was similarity targeting, in which the organism receiving the CPU cycles had to have a genome that was similar to the donating organism (edit distance below a certain threshold). Lastly, there was greenbeard targeting, named for the idea that altruistic organisms can display visual markers (like a green beard) that is tied to their altruism and thus provides a reliable signal to other organisms (Hamilton, 1964; Dawkins,

1976). With this method, only other organisms that have previously executed the greenbeard altruism instruction can recieve benefit from an organism executing it. Theoretically, this greenbeard approach should be the most accurate form of altruism targeting, as, in the Avida implementation, it literally allows organisms to help only those that express the same gene. Surprisingly, however, while similarity targeting was favored over kin targeting, greenbeard targeting was not selected over the other forms. Further experimentation revealed that the key difference was that greenbeard targeting did not provide a way to distinguish mildly altruistic organisms from highly altruistic organisms (those that executed the instruction repeatedly). This result may explain the fact that most greenbeard altruism observed in nature tends to relate to binary rather than continuous forms of altruism. Indeed, when the greenbeard instruction would only benefit other organisms exhibiting the same level of altruism, it did become the preferred mechanism.

Some traits, such as releasing a chemical into the surrounding environment, do not have a specific other organism being targeted. For these traits to be evolutionarily favored, there must be some mechanism for ensuring that their benefit primarily goes to kin. In many cases, this guarantee can come from living in a sufficiently spatially-structured environment, such that the majority of neighboring organisms are kin. Another option is quorum-sensing, wherein altruistic organisms signal their presence to each other, allowing them to base their actions on the number of kin that are present (Diggle et al., 2007). Beckmann and McKinley demonstrated that digital organisms in Avida are capable of evolving quorum sensing if allowed to pass messages to each other (Beckmann and McKinley, 2009). In collaboration with biologists, they were then able to use this system to understand how bacteria might evolve resistance to efforts to interfere with their quorum sensing (Beckmann et al., 2012). Vostinar (née Johnson) et al. expanded on this research, finding that not only is the evolution of quorum sensing in Avida possible, it also increases the range of conditions in which altruism will evolve and the extent to which it is beneficial (Johnson et al., 2014). This work provided the basis for them to build a more precisely targeted digital system to directly model quorum

sensing in a species of bacteria (Vostinar et al., 2018).

1.2.3 Predation

Predation is a fundamentally different interaction than any that I have discussed so far, and a critical component of many ecological communities in nature. There is an asymmetry in predator-prey relationships that is not present with competition or cross-feeding (one species consumes the other, not the other way around), which results in a system with fundamentally different behavior (Johnson and Wilke, 2004). Studying predation is complicated by the fact that most predators need to be able to move in order to reach enough prey to sustain themselves. As such, predation research in Avida generally allows organisms to execute instructions that cause them to move around the world as well as instructions that give them sensory information about the landscape, making for far more complex (and thus slower) experimental worlds. Prey organisms consume abiotic resources from their cells in the environment and store them (initial research on prey foraging behavior in the absence of predators was conducted by Walker (Walker, 2011)). Predators can attack prey and receive a portion of their stored resources by executing a specific instruction while facing them.

Because of the huge impact that allowing movement has on the strategies evolved by digital organisms, a lot of predation research in Avida thus far deals with the evolution of movement strategies and sensory behaviors. As it turns out, these behaviors profoundly affect and are affected by the presence of predators. Wagner, et al. compared the behavior of prey evolved with and without a co-evolving population of predators and found that predators force the prey to explore a much wider space of strategies (Wagner et al., 2013). They also were able to evolve to make much better use of their sensory capabilities, which in turn made them better at foraging as well. I hypothesize that the fact that ecosystems often include a variety of related-but-different survival pressures, many of which consistently become more challenging as a population adapts to them, makes them particularly effective at promoting the evolution of more complex organisms.

O'Donnell et al. expanded on some of these questions by investigating properties of prey communities that make them better at evolving to avoid novel predators (O'Donnell et al., 2014). They looked at two different factors: the amount of standing genetic variation in a prey population, and the evolutionary history of that population (i.e. whether the lineages had been exposed to different predators in the past). Surprisingly, they found that standing genetic variation had very little impact on the extent to which the prey population was able to evolve resistance to predators; it had a slight positive effect among populations without predators in their evolutionary history. Evolutionary history, on the other hand, had a dramatic effect. Despite the fact that the predators used in the experimental phase were from a completely different population than the ones that the prey had co-evolved with, prey populations that had historically been exposed to predators were substantially more likely to evolve resistance to novel predators.

Further research on the specific techniques used to avoid predators has addressed the evolution of mimicry and prey grouping. Mimicry requires a fascinating and complex set of ecological interactions: a poisonous prey species that signals its toxicity to potential predators, a predator species that responds to this signal, and an additional non-poisonous prey species that mimics the signal so that predators will avoid it as well. This set-up creates an interesting form of density-dependence, where the efficacy of the mimicry decreases as the relative abundance of the mimic species increases. Lehmann et al. created such an ecosystem in Avida, and studied the scenarios under which it is stable (Lehmann et al., 2014). They found that the degree to which the toxin harms predators and the accuracy of the mimicry are both important variables in the evolution of a stable population of mimics. The more harmful the toxin, the greater the density of mimics the system can support.

Prey grouping is a more complex predator avoidance strategy, as it requires some degree of coordination among prey organisms. Biswas et al. used Avida to untangle the mechanisms that drive the evolution of such behavior (Biswas et al., 2014). Two different potential mechanisms had previously been proposed: dilution and predator confusion. Dilution is the

simple observation that being in a group can lower an individual's chances of being selected by predators for attack, while predator confusion is the idea that predators may have trouble visually perceiving individual prey to hunt if those prey are in a group. To study this phenomenon, the authors observed the extent to which prey organisms evolved grouping behavior under various conditions. While the presence of predators did significantly increase prey grouping behavior, the effect was no different when predators' visual sensors were inhibited. This result suggests that the dilution effect is sufficient to evolve prey grouping behavior, and that the more complex explanation of predator confusion is unnecessary.

1.2.4 Parasitism

There is increasing evidence that parasites play a critical role in many ecological communities (Lafferty et al., 2006). Conveniently, parasites fit neatly into Avida. In fact, parasites evolved naturally and unexpectedly in Avida's precursor, Tierra (Ray, 1994). Since the presence of parasites in Tierra led to various side effects that made for much more complex dynamics, Avida implements parasites more intentionally. Like normal organisms in Avida, parasites are a sequence of simple computer code instructions. Unlike normal organisms, parasites cannot survive in a cell on their own. Instead, they inhabit the same cell as a normal organism and steal its CPU cycles (i.e. when the host organism should have been allowed to execute an instruction, the parasite instead executes one from its own genome). For most parasite experiments in Avida, the normal organisms (hosts) are required to perform at least one task in order to replicate. Parasites are able to infect hosts that perform the same computational tasks as them. Thus, there is pressure for hosts to evolve the ability to perform new tasks and lose the ability to perform the old tasks. This set-up promotes the generation and maintenance of diversity (Zaman et al., 2011), and ultimately leads the hosts to evolve to be more complex than they would otherwise have been (Zaman et al., 2014).

The parasitism-mutualism continuum has been addressed in other digital systems as well. For example, Symbulation is a platform that is designed to simplify the study of

these topics by reducing the complexity introduced in other systems (Vostinar and Ofria, 2019). Notably, Symbulation was recently used to demonstrate that spatial structure has an entirely different impact on host-parasite co-evolution than it does on other forms of potentially-cooperative co-evolution (Vostinar and Ofria, 2019).

1.2.5 Perturbations

Most of our most pressing concerns about ecosystems on Earth stem from the fact that many of them are currently facing novel perturbations, often human-driven. Biological data tend to be too limited to allow us to fully understand how an ecological system will be affected by particular perturbations. Digital systems like Avida, on the other hand, cannot easy model a specific, real-world ecosystem. In conjunction, however, they can help us dramatically improve our expectations about the results of a particular perturbation in a natural environment.

The most dramatic types of perturbations an ecosystem can experience are mass extinction events. Mass extinctions are divided into two categories: press and pulse (Erwin, 1998). Press extinctions are those that are induced by a prolonged stressor, to which adaptation is theoretically possible. In Avida, press extinctions can be induced by lowering resources to low levels (Yedid et al., 2008). Pulse extinctions, on the other hand, are those that occur too quickly and too intensely for populations to adapt to. In Avida, pulse extinctions can be induced by killing a high percentage of the population (Yedid et al., 2009).

Part of the reason that extinctions can be so damaging to ecosystems is that most ecological communities are highly interdependent (Roopnarine et al., 2007). While communities evolved via adaptive radiation in the logic-9 environment would almost certainly change if they lost a member, they lack the level of interdependence that a food web creates. As such, understanding the impact of mass extinctions on an ecological community requires us to assemble a more complex community. Most of the research on mass extinctions in Avida has used a complex cross-feeding environment that functions similarly to a trophic pyra-

mid (Yedid et al., 2008, 2009, 2012). In this environment, simpler tasks produce byproduct resources that can be metabolized by performing more complex tasks. Thus, as in most ecological communities, species at higher trophic levels (which are often more complex) are dependent on species at lower trophic levels.

In a press extinction, this dependency means that the most complex ecotypes are generally lost from the population (Yedid et al., 2008). If the amount of resources entering the lowest trophic levels is decreased, less of the byproduct resources can by produced, with a cascading effect across all levels. This deficit percolates up through the trophic pyramid, resulting in minimal resource available for the highest level. What are the long-term of effects of such episode on a community? Yedid et al. studied the process of regaining complex traits after the press event is over, and found that ecological communities that contained ecotypes with the ability to perform the most complex trait were more likely to re-evolve it than communities that never had such functionality (Yedid et al., 2008). While the mechanisms behind this dynamic were complex and varied, many of them are also present in biological ecosystems, suggesting that this trend might be general.

However, a follow-up study by Yedid et al. showed that recovery from press extinctions still takes a long time, particularly at higher trophic levels. Recovery from pulse extinctions, on the other hand, occurs much more rapidly (Yedid et al., 2009). This tendency is believed to be the result of press extinctions providing an alternative selection regime that encourages the loss of more complex traits in favor of simpler traits that will more reliably yield a reward. The authors also observed some parallels between the results of this experiment and observations from the fossil record, further supporting the idea that these same dynamics may occur in biological ecosystems. Further analysis of the phylogenies of populations subjected to these perturbations revealed that stronger perturbations, particularly press events, results in a substantial loss of phylogenetic diversity (Yedid et al., 2012).

Similar research conducted in the standard logic-9 environment (with only a single trophic level) suggests that the use of multiple trophic levels in Yedid et al's work were

important to their results. Communities evolved in logic-9 and then subjected to extreme bottlenecks (effectively strong pulse extinctions) were able to regain their original levels of consumption of each resource relatively quickly, in contrast to the delayed recovery of the most complex task in the cross-feeding set-up (Olson et al., 2013). Interestingly, although the post-bottleneck communities fulfilled all of the same functions as the original communities, the ecotypes present often partitioned resources very differently. This result reflects the fact that there are a wide variety of sets of ecotypes that are able to stably co-exist in the logic-9 environment.

A much more mild form of perturbation is a localized disruption of some regions within the environment, but not others. Such a scenario is commonly observed in nature in the form of anthropogenically fragmented ecosystems. Often, governments will set aside some regions as biological reserves, where the native ecosystem will, theoretically, be undisturbed. Making such choices determines the spatial structure that populations in this environment will have going forward, a property known to have a dramatic effect on evolutionary dynamics (Tomassini, 2005). While this topic had received a great deal of attention from ecologists, there is a relative lack of knowledge of how evolutionary considerations should inform such decision-making. In chapter 3, I use Avida to preliminarily investigate this question (Dolson et al., 2016b).

1.2.6 Interaction networks

All of the ecological forces mentioned above can, of course, be combined. Such studies can quickly become too complex to understand the results of, so it is important to take care in experimental design. However, a number of interesting directions for research on eco-evolutionary dynamics concern the network of interspecific interactions that defines an ecological community. For such lines of research, combining different drivers of ecology in a single experiment is necessary. A number of ideas for approaching such questions have been laid out by Fortuna et al. (Fortuna et al., 2013), and some are already being put into

practice, with promising results (Strona and Lafferty, 2016).

1.3 Eco-evolutionary dynamics in evolutionary computation

Evolutionary computation researchers have devoted substantial attention to understanding how to promote coexistence among lineages exploring different regions of a fitness landscape (Goldberg et al., 1987; Mahfoud, 1995; Mouret and Doncieux, 2009; Pugh et al., 2015). Promoting diversity in an evolving population is important for EC because it reduces premature convergence on suboptimal fitness peaks while still encouraging both exploration and exploitation. However, some types of diversity facilitate finding global optima better than other types. For example, a high mutation rate generates more new genotypes, but this increased exploration tends to be localized to a single region of the fitness landscape and sacrifices fine-tuned exploitation of promising prospective solutions; the population plows ahead exploring rather than refining the solutions already found.

Even amongst more advanced diversity-promoting approaches, some produce forms of diversity that are spread across the fitness landscape in ways that appear to be more or less conducive to solving a given problem. For example, Mouret and Doncieux found that diversity in agent behavior is much more predictive of success in evolutionary computation than diversity in neural network architecture (Mouret and Doncieux, 2009). Similarly, Walker and Ofria observed different evolutionary potential at equivalent diversity levels under different environmental configurations in Avida (Walker and Ofria, 2012). Although we have a high-level idea of why many techniques promote diversity, the mechanistic details by which these populations spread across the fitness landscape are less well understood. For example, fitness sharing promotes diversity via negative density dependence (Goldberg et al., 1987). Does this diversity represent qualitatively different portions of the fitness landscape than, for example, lexicase selection (Spector, 2012)? The answer to this question depends on subtle

differences in the evolutionary pressures that these different selection schemes place on the evolving population.

Because techniques for promoting diversity rely on creating fitness interactions between individuals in the population (beyond basic competition for space in the next generation), they are, by definition, creating simple ecologies. Ecologists have developed rigorous theory to predict how ecological communities change over time and are exploring their interactions with evolutionary dynamics. A particularly popular area of study concerns the conditions under which long-term stable coexistence between different species is possible (Pacala and Tilman, 1994; Chesson, 2000; Chase and Leibold, 2003; Letten et al., 2017). Such a theoretical framework can be applied to understand stable coexistence in EC populations as well, ultimately allowing us to determine dynamic properties of the system. Of particular interest, such theory should facilitate the prediction of which pathways can simultaneously be traversed along a given fitness landscape. By understanding this interaction between diversity maintenance strategies and the fitness landscape, we can predict which algorithmic techniques will be the most effective *a priori*. These insights should apply both to choosing an appropriate diversity maintenance technique and to setting its parameters. Additionally, an improved mechanistic understanding of why existing algorithms work should facilitate building more effective variations of those algorithms.

A number of highly effective evolutionary-computation techniques owe their success to ecological dynamics. In particular, I will focus on **fitness sharing**, **Eco-EA**, and **lexicase selection**.

Fitness sharing, one of the first ecology-based diversity maintenance approaches, is by far the simplest of these techniques (Goldberg et al., 1987). Under this regime, the fitness of each individual is reduced in proportion to the number of similar individuals in the population. As such, constant pressure to diversify is imposed.

Eco-EA is an approach, originally built on to of Avida, to solving complex problems by associating limited resources with simpler challenges that can be used as stepping stones

(Goings and Ofria, 2009; Goings, 2010; Goings et al., 2012). These challenges can be sub-problems of the larger problem, individual test cases, or other related tasks that may be used as components of a global solution. Resources remain plentiful until a solution to the corresponding sub-problem is discovered. When a population first solves a sub-problem, the large amount of resource dramatically boosts the fitness of individuals that use it. As the number of individuals using the resource increases, it become less valuable. As long as there is a cost to attempting to use each resource, negative frequency dependence fosters small, but stable sub-populations capable of solving each sub-problem.

Lexicase selection (Spector, 2012) has proven extraordinarily successful in genetic programming (Helmuth et al., 2015; Helmuth and Spector, 2015). In lexicase selection, a large number of test cases are used as criteria for evaluation. Each time an organism has to be selected to propagate into the next generation, the test cases are applied in a random order, successively filtering out all but the most fit candidates. Once a single candidate remains (or all of the test cases have been applied and a random individual is chosen from the final set) it is replicated into the next generation and the process is repeated. The fact the the ordering of these test cases continually changes means that solutions successful at different subsets of tasks are all able to co-exist. Although lexicase selection has traditionally been used with test cases, there should be no reason it cannot be used with multiple fitness functions instead to provide ecologically-mediated hints. In this way, any hint that could be given to a system like Eco-EA could just as easily be provided to lexicase selection as well.

From an ecological perspective, lexicase selection creates pressure for the population to diversify into different niches that are based on building blocks to a more complex problem. In order for an organism to be evolutionarily successful, it must be one of the best at at least one of the test cases/fitness functions. In most cases, it must be among the best at many of them (probably a related group). Thus, organisms compete with others within the niches created by each test case/fitness function. Eco-EA also promotes intra-niche competition, but in a subtly different way. Limited resources are used up as organisms receive fitness

bonuses, meaning that they compete with other organisms benefiting from the same bonus. Thus, those with higher fitness will competitively exclude less fit organisms that depend too much on the same resources (Chesson, 2000). Natural systems with limited resources also promote intra-niche competition, but in a subtly different way. Limited resources are used up as organisms consume them, meaning that competition occurs per-resource. Thus, those more capable of consuming multiple resources will competitively exclude less fit organisms that depend too much on those same resources (Chesson, 2000).

This distinction between competition in lexicase selection and in systems with limited resources has a few implications. First, limited resources allow for generalists that do not excel at any tasks but are decent at many. It is unclear whether this flexibility opens up additional pathways through the fitness landscape. Second, lexicase selection always exerts selective pressure to improve on all test cases/fitness functions, whereas the pressure exerted by limited resources can be very weak (or even non-existent) if a niche is already occupied. This dynamic may make lexicase selection less robust to receiving bad advice about how to solve a problem than other mechanisms for providing ecologically-mediated hints. Third, lexicase selection automatically incentivizes organisms that excel at uncommon combinations of test cases/fitness functions. Ideally, we will be able to find a hybrid approach to providing ecologically-mediated hints that combines limited resources with lexicase selection to achieve the best of both worlds.

MAP-Elites is another successful algorithm that leverages ecological dynamics and user input (Mouret and Clune, 2015). In MAP-Elites, the user chooses axes along which they believe variation will be important. MAP-Elites then breaks the hyperspace defined by these axes into discrete regions. Each region can be occupied by at most one solution. When a new candidate solution is created, it is assessed on each axis and the corresponding bin is located. If the new candidate solution has a higher fitness than the previous occupant of that bin, it replaces that occupant. Otherwise, it is tossed out. Evolution proceeds by choosing occupants of various bins in the search space and mutating them to create new candidate

solutions.

The sub-division of axes for MAP-Elites explicitly partitions the search space into niches in a way that has clear parallels to choosing tests to provide to lexicase selection or hints to associate with limited resources. The lack of directionality in MAP-Elites, however, makes it distinct from lexicase selection. At face value, it might seem to make it different from Eco-EA as well. Certainly this intuition is true to some extent - hints provided as limited resources do have directionality to them. However, the foundation of limited resources is negative frequency-dependence which, in addition to reducing the risk of the entire population being led astray, has the important effect of promoting diversity along the axis of the hint. This diversity goes beyond merely allowing a portion of the population to drift; niche partitioning should force the population to spread out along the axis in question to produce meaningful diversity that may be helpful in solving the overall problem.

There are a wide variety of other ecologically-inspired strategies that have proven to be effective for maintaining diversity in evolutionary algorithms. These largely fall into four categories: niching/speciation (e.g. (Goldberg et al., 1987; Stanley and Miikkulainen, 2004)), parent selection (e.g. (De Jong, 1975; Mahfoud, 1992)), dividing the population into subpopulations (e.g. (Hornby, 2006; Hu et al., 2005)), and adjusting the objective function to favor diversity and/or novelty (e.g. (Mouret and Doncieux, 2009)). Of these approaches, only niching/speciation consistently promotes stable coexistence of different types of strategy. Even among niching/speciation strategies, most emphasize a diversity of phenotypes rather than the diversity of evolutionary building blocks that Eco-EA and Lexicase Selection promote. While solutions built on different building blocks may often exhibit different phenotypes, they may also arrive at similar phenotypes despite taking very different paths through the fitness landscape. These different paths will likely result in underlying differences in genetic architecture that may influence which behaviors are easy for a lineage to evolve next. As such, I suspect that diversity of building blocks will promote greater evolutionary potential than other kinds of population diversity.

1.3.1 Limited resources and Eco-EA

The original formulation of limited resources that led to ecologically-mediated hints arose in work by Cooper and Ofria where they demonstrated that limited resources are sufficient to evolve stable branching of different ecotypes (Cooper and Ofria, 2003). The resulting ecological community was incredibly simple, with competition being the only form of interaction between organisms. However, I hypothesize that these simple ecologies (with meaningful differences between niches) were better positioned to solve complex problems than populations in which diversity was promoted by other means.

Goings and Ofria initially tested a more applied form of limited resources in Eco-EA, using a toy bitstring matching problem, with the goal of maintaining a population of diverse solutions (Goings and Ofria, 2009). In this experiment, resources (either limited or unlimited) were associated with different bitstrings that could be matched to varying extents by members of the population. When resources were unlimited, the population converged on matching a single one of these strings. Limiting the resources, however, produced consistent subpopulations specialized on each string. The negative frequency dependent selection imposed by the limited resources caused the different bitstrings to stably coexist within the population. These results held when the task was to match a more general pattern of bits, rather than an exact string.

Following this initial conceptual test, Goings et al. applied Eco-EA to a complex real-world problem: generating models for the behavior of sensor nodes in a flood warning network (Goings et al., 2012). Not only did Eco-EA successfully evolve a diversity of models that satisfied the constraints of the problem, but the models evolved by Eco-EA were better than a comparison algorithm at continuing to diversify when transferred to an environment without limited resources. These results demonstrate the strength of Eco-EA at evolving a diversity of solutions to complex problems.

Eco-EA is just one example of the benefits that making more thoughtful and intentional use of ecology can provide. Here, I define ecological dynamics as interactions between mem-

bers of the population that affect fitness. I propose that such dynamics provide a variety of benefits, some of which have begun to be put in to practice and some of which have not. The most obvious of these effects is the benefit to diversity alluded to above.

1.3.2 Complexifying environments

In nature, the network of interactions within a community grows increasingly complex over evolutionary time, as the evolution of new species creates new niches for other species to evolve into. In turn, new types of interaction emerge (such as parasitism, mutualism, and predation). This gradual complexification has two implications that may be important for evolutionary computation: 1) The gradual increase in complexity of the biotic environment facilitates the continual production of more complex niches and thus the evolution of more complex traits, and 2) Ecological communities as a whole can often perform functions that no single species could perform alone.

The niches inhabited by various members of an ecological community usually bear commonalities. While they are not identical, the ability to occupy one niche is often a building block for the ability to occupy a different one (a clear example of this effect in nature is metabolic pathways for metabolizing various resources). When new niches appear, they are closely enough related to existing niches that it's plausible that existing populations will evolve to inhabit them. Without the gradual feedback loop of increasing complexity that eco-evolutionary feedbacks enable, evolving the ability to exploit newly established niches would be incredibly improbable.

Community-level functions represent a different mechanism for solving complex problems. For example, a forest, collectively, is able to store solar energy in sugar molecules, fix nitrogen, take in various nutrients from the soil, and use these nutrients and energy to power mobile, decision-making agents. Of course, these functions are each performed by different species in the community. While it may be possible to evolve a single species that carries out all of these tasks, it is worth considering the possibility that it is easier to evolve

26

a community that carries them out collectively. The fact that such communities seem to be so common in nature is certainly suggestive of this possibility. Such an approach would have a variety of potential benefits. For example, it would promote more functional modularity, a trait that is believed to be critical for evolutionary potential. Moreover, the closely interdependent species found in a complex ecological community are a likely precursor to egalitarian major transitions, which are believed by many to be a critical step towards evolving complex species. Lastly, as ecological communities can effectively be thought of as a set of subroutines running in parallel, they lend themselves easily to evolving parallel programs, somewhat akin to Holland's bucket brigade algorithm (Holland, 1985) and more recent advances in learning classifier systems.

1.4 Thesis Statement

The central thesis of this dissertation is that ecology plays a critical role in determining the outcome of evolution and is necessary to unlock its full constructive power. Thus, we can use ecology as a tool to intentionally guide evolution.

1.5 Contributions

This dissertation is organized into four parts:

1.5.1 Part 1: Introduction

In this chapter (1), I provide a high level introduction and literature review of concepts that are relevant across chapters. Chapter 2 describes all study systems and evolutionary computation approaches used across this dissertation.

1.5.2   Part 2: Understanding eco-evolutionary dynamics in biology

Nature exhibits a striking range of eco-evolutionary dynamics. To understand how they might be harnessed in other evolutionary contexts, it is important to understand the forces the generate and maintain diverse ecological communities. To this end, I begin by furthering the current understanding of eco-evolutionary dynamics in biology. Specifically, the two chapters in this section deal with the impact of spatially heterogeneous environments on eco-evolutionary dynamics. Spatial heterogeneity is an important driver of eco-evolutionary dynamics that is relatively easily to manipulate. Thus, from the perspective of using ecology as a tool for manipulating evolutionary outcomes, spatial heterogeneity is a logical entry point. Within this section, chapter 3 examines the ecological communities formed by spatially heterogeneous environments, showing that certain kinds of heterogeneity promote diversity and increased probability of evolving complex traits. Chapter 4 then delves into the specifics of this later finding; I show that environments contain local regions where evolving a given complex trait is more likely. Although I do not isolate a precise mechanism, this work provides strong support for the idea that we can be build landscapes to promote desirable outcomes.

1.5.3   Part 3: Harnessing eco-evolutionary dynamics for evolutionary computation

In this section I show that many of the best breakthroughs in evolutionary computation have come from adding ecology to existing algorithms and demonstrate the power of further incorporating ecology into evolutionary algorithms. This work functions as a case study on the value of ecology for promoting evolutionary innovation. Moreover, it lays the groundwork for future use of evolutionary computation as a model system for experimenting with approaches to controlling eco-evolutionary dynamics.

Specifically, in chapter 5, I demonstrate that theory and techniques from ecology can be

used to improve our understanding of and predictive capacity for evolutionary computation. I show that that these two systems are governed by the same underlying laws (and are even in some cases mathematically identical). Using this new theory, I explore the functional differences between similar-seeming algorithms, generating hypotheses to explain why they behave differently.

I extend this work in chapter 6 with a proposal to use ecology to facilitate a new approach to evolutionary computation. For the same reasons that humans are notoriously bad at creating fitness functions, attempts to provide evolutionary algorithms with "hints" (e.g. domain knowledge) about how to solve a given problem tend to be ineffective. Intentionally creating corresponding ecological niches, however, may be a more robustly successful approach to providing such guidance. In this chapter, I demonstrate that Eco-EA can be used in this way; it is able to make use of good advice without being harmed by bad advice. This property is in contrast to other selection schemes. Thus, in addition to being a promising direction for algorithm development, this work furthers our understanding of the practical differences between the ecological communities created by different selection schemes.

1.5.4 Part 4: Developing tools to identify and understand complex digital ecologies

Making the case that ecology is critical to evolution requires comparing systems that have more or less ecology. In chapter 7, I present an approach to identifying the presence of ecology and quantifying the number of occupied niches in a system. This work is part of a larger project to develop a suite of metrics which aid in identifying a range of hallmarks for "open-ended evolution" (i.e. evolution that displays the same large-scale properties as evolution in nature). Of course, simply being able to identify the presence or absence of various dynamics is not sufficient for understanding the mechanisms that produced them. Thus, in chapter 8, I present an additional suite of metrics and visualizations aimed at quantifying the underlying evolutionary dynamics that are occurring in a system. Together,

these metrics open up a world of possibilities for advancing the study of eco-evolutionary dynamics in digital systems.

Chapter 2

Methods

2.1 Study systems

To understand how the role of ecology generalizes across evolutionary substrates, I use a range of experimental systems: NK landscapes, real-valued function optimization, and two versions of linear genetic programming. Across all systems, reproduction was asexual.

2.1.1 NK Landscapes

An NK model uses two parameters, N and K, to randomly generate a fitness landscape (Kauffman and Levin, 1987). N specifies the number of sites in the genome, each of which is a 0 or a 1. The fitness landscape specifies the effect of a given value at a given site on the fitness of the bit-string organism. This fitness effect depends on the values at the K subsequent adjacent sites. As such, K tunes the ruggedness of the landscape; low values of K produce smooth landscapes with few peaks, whereas high values produce landscapes with many peaks. I use NK models in chapter 7 to test my intuitions about evolutionary processes, because they are a well-understood system for studying general questions about evolutionary dynamics.

2.1.2 Real-valued function optimization

In real-valued function optimization, a population of vectors of floating-point numbers is evolved to find the optimum of a function. These problems have easy-to-visualize fitness landscapes, making them another useful domain for testing intuitions about evolutionary processes. I used two types of problem in this context:

2.1.2.1 The N-dimensional box problem

The N-dimensional box problem is a toy problem that I use in chapter 6 to explore the possibility of using ecological niches to provide "hints" to an evolutionary algorithm about how to solve a problem. In the version used here, solutions exist within a 10-dimensional box. All sides of the box have a length of 1. Candidate solutions in the population are sequences of 10 floating point numbers between 0 and 1, representing a point within the box. The goal is to find the origin (i.e. the sequence $[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$). This problem could be trivially solved by using the inverse Euclidean distance between a point within the box and the origin as the fitness function. To simulate a more challenging problem, I use the following fitness function:

$$
fitness = \begin{cases} .01 & \sqrt{\sum_{i=1}^{10} x_i^2} > .1 \\ \frac{1}{\sqrt{\sum_{i=1}^{10} x_i^2}} & \sqrt{\sum_{i=1}^{10} x_i^2} \leq .1 \end{cases}
\tag{2.1}
$$

In this function, inverse Euclidean distance from the origin is the fitness only when Euclidean distance from the origin is less than 0.1 (which, given the high dimensionality of the space, represents 2.5×10^{-13} of the possible positions). For all points in the box that are farther away from the origin, fitness is 0.01. I used 0.01 rather than 0 as the base fitness to ensure the Eco-EA's fitness multipliers would have an effect. The cut-off of 0.1 was chosen such that it is unlikely for evolution to solve the problem without hints. Using this fitness function creates a quintessential "needle in a haystack" problem, where the fitness landscape

is flat except for one incredibly tall and thin peak. Such problems are generally considered to be among the most challenging for evolutionary computation to solve, as they do not allow for incremental improvement.

2.1.2.2 Niching Competition Benchmark Problems

The GECCO Competition on Niching Methods is an annual competition that evaluates the performance of evolutionary algorithms on multi-niche problems (i.e. those with multiple optima). Entrants compete on a range of real-valued optimization functions (Li et al., 2013). I select a subset of these functions for use here: Himmelblau, Shubert, Composition Function 2, and Six-Humped Camel Back. For each test problem, the X and Y coordinates offered by a given organism are translated by the function into a fitness value. Because of their low dimensionality, it is possible to fully visualize each problem's actual fitness landscape, a property I take advantage of in chapter 8. I use the implementations of these problems at `https://github.com/mikeagn/CEC2013` (C++ for fitness calculations during evolution, Python for post-hoc analysis). Figure 2.1 shows the fitness landscapes defined by each of my four chosen test problems. Each organism's genome consisted of two floating point numbers that defined its position in the fitness landscape.

2.1.3 Linear genetic programming

In linear genetic programming, an individual is a computer program composed of a sequence of instructions. The program receives input and must produce a corresponding output. In this work, I use two different linear genetic programming systems: one (ScopeGP) developed with the goal of solving problems via evolutionary computation and one (Avida) developed with the goal of being a model system for better understanding biology.
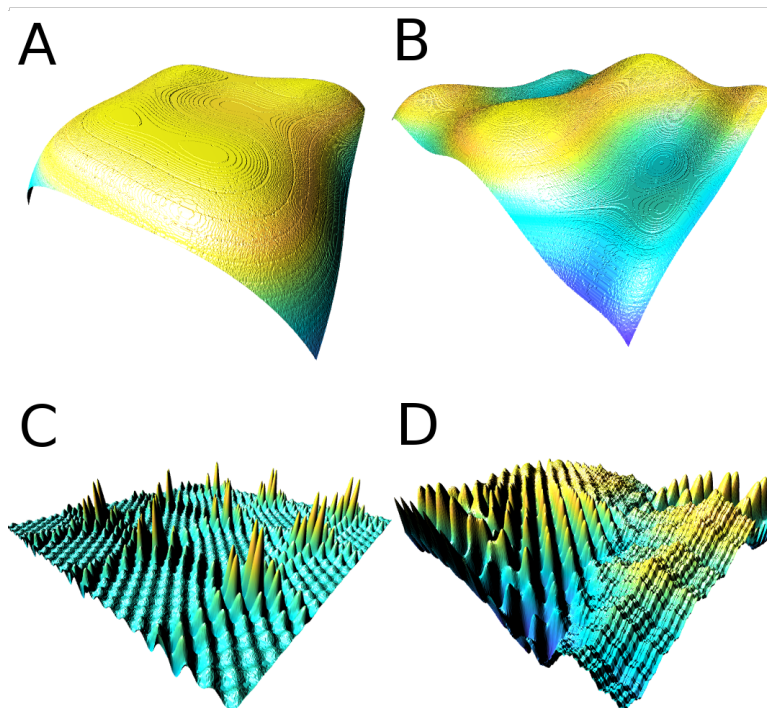
Figure 2.1: **The fitness landscapes used in this experiment**: A) Himmelblau, B) Six-humped Camel Back, C) Shubert, and D) Composition Function 2. Interactive versions available at `https://emilydolson.github.io/fitness_landscape_visualizations`.

2.1.3.1 ScopeGP

ScopeGP is a linear genetic programming representation designed around the use of modular subroutines called "scopes." Making efficient use of modular subroutines has long been thought to be important to facilitating the evolution of genetic programs that solve complex problems. ScopeGP is designed to facilitate the evolution of modular and reusable code while minimally affecting the way in which traditional linear genetic programs are organized (*i.e.*, linear sequences of instructions). The instruction set includes a number of instructions that automatically create programming structures like loops and subroutines. ScopeGP is described in more detail in (Dolson et al., 2018c).

I use ScopeGP in the context of solving three different problems:

- **Squares:** in this task, which was designed to be quickly-solvable, programs must take a number as input and return its square as output.

- **Collatz:** a problem taken from Helmuth et. al's general program synthesis benchmark suite (Helmuth and Spector, 2015) that is known to be challenging. Programs receive a number as input and must return the number of iterations that it takes the a Collatz sequence starting from that number to converge.

- **Dow chemical challenge:** a real-world symbolic regression problem White et al. (2013) in which programs must predict details about the outcome of an industrial chemical synthesis process based on a wide range of measurements.

For all of these problems, programs are evaluated based on their performance on a suite of test-cases.

2.1.3.2 Avida

The Avida Digital Evolution Platform (Ofria and Wilke, 2004) is a well-established artificial life system that has been used to study a wide range of evolutionary dynamics

(Lenski et al., 2003; Goldsby et al., 2012; Zaman et al., 2014; Dolson et al., 2016b). Avida's track record and popularity make it a good platform both for testing intuitions about new analytical techniques and for conducting experiments aimed at improving our understanding of biology.

Avida is a world of self-replicating linear computer programs competing for space in a 60 x 60 toroidal lattice. Organisms in Avida replicate asexually by copying themselves instruction-by-instruction and dividing; copy and divide operations, however, are not perfect and can result in mutated offspring. When an organism successfully replicates, the resulting offspring is placed randomly within the parent's Moore neighborhood (the surrounding 8 cells), replacing that location's former occupant. Thus, there is selection pressure for organisms to replicate quickly before being copied over by others.

Improvements in fitness (i.e. reproductive output) can either come from optimizations to the code of the genome (increasing its efficiency) or from gaining the ability to take in resources by performing boolean logic operations. Resources give programs extra CPU cycles, allowing them to copy themselves faster than competitors. Each of the nine two-input Boolean logic functions allows organisms to use a different resource, with more complex functions corresponding to more valuable resources. The presence and value of these resources can be adjusted to create a wide range of environments.

I use the following environments in this work:

- **Minimal:** This environment contains no resources. As such, selection pressure is entirely focused on optimizing the efficiency at which organisms can self-replicate.

- **Logic-9:** The environment contains the nine resources associated with non-trivial one- and two-input Boolean logic functions: NAND, NOT, OR-NOT, AND, OR, AND-NOT, NOR, XOR, and EQUALS (for more information on these logic functions in Avida see Lenski et al., 2003).

- **Limited resources:** This environment contains the same resources as Logic-9. How-

ever, in this environment each resource is limited in quantity. When organisms use resources, they deplete them, creating pressure to use resources that are not used by other members of the population. Unlike the previous two configurations, the limited resources environment contains multiple niches and support stable ecologies (Cooper and Ofria, 2003; Chow et al., 2004).

- **Changing:** In this environment, the rewarded functions cycle between two opposing states: ENV-NAND (where NAND is rewarded and NOT is punished) and ENV-NOT (where NOT is rewarded and NAND is punished). ENV-NAND and ENV-NOT are configured such that no phenotype can be optimal across both environment states. To achieve maximal fitness in either ENV-NAND or ENV-NOT, organisms must perform *only* the focal function.

- **Spatially heterogenous:** In this class of environments, resources are geographically limited, creating an environment in which different traits are advantageous in different spatial regions. I create these environments by placing circular patches of different resources within the world. By allowing these patches to overlap, I can insure the presence of a wide variety of niches.

2.2 Selection schemes

For all experiments performed outside of Avida (in which selection is implicit), a means of selecting individuals to reproduce is required. In much of this work, I explore the differences amongst four selection schemes: tournament selection (the control, which does not create ecology), fitness sharing (Goldberg et al., 1987), lexicase selection (Spector, 2012), and Eco-EA (Goings and Ofria, 2009).

2.2.1 Tournament selection

As tournament selection is one of the most popular selection techniques and does not create an ecology, I use it as a control. In its simplest form, tournament selection has one distinct parameter, T, the number of individuals to randomly pick from the population (with replacement) for each tournament. The fittest individual in the tournament is then chosen to reproduce. An independent tournament is run to fill each position in the population. T controls strength of selection where higher tournament sizes correspond to strong selection and lower tournament sizes correspond to weak selection (Blickle and Thiele, 1995). A tournament size of one is equivalent to no selection pressure (*i.e.*, every organism in the population has an equal chance of being selected to reproduce).

2.2.2 Fitness sharing

Fitness sharing was an early use of ecology to promote diversity in EC (Goldberg et al., 1987). In fitness sharing, the fitness of every member of the population is reduced in proportion to the density of similar individuals (i.e. they "share" their fitness). Specifically, the extent to which each organism needs to share its fitness is calculated by summing the sharing equation, $sh(d)$, over the population:

$$sh(d) = \begin{cases} 1 - (\frac{d}{\sigma_{\text{share}}})^{\alpha} & d < \sigma_{\text{share}} \\ 0 & d \geq \sigma_{\text{share}} \end{cases} \qquad (2.2)$$

In this equation, d is the distance (defined in terms of genotype or phenotype) between the individual that is having its fitness calculated and the other member of the population that it is being compared to. α is a parameter that tunes the relationship between the similarity of individuals and how much they compete with each other. σ_{share} (the sharing threshold) specifies how similar individuals need to be to compete with each other at all. For each individual, $Sh(d)$ is summed across the population and that individual's fitness is

divided by the sum.

2.2.3 Lexicase selection

In lexicase selection, solutions are evaluated on a large number of criteria (Spector, 2012). Traditionally, these criteria are test cases, but other types of fitness functions are also effective (Dolson et al., 2018a). To choose an individual to reproduce, the selection criteria are applied to the population in a random order. Only the individuals that perform best on each criterion move on to be evaluated on the next, until only one individual remains. In case of a perfect tie, the winner is selected randomly from the remaining options.

Lexicase selection is highly effective at solving challenging problems in genetic programming (Helmuth and Spector, 2015; Helmuth et al., 2015) while maintaining diverse populations (Helmuth et al., 2016a). Prior work on the population dynamics of lexicase selection has noted the surprising prevalence of "hyperselection events," occasions in which a single individual is selected as a parent for the vast majority of the next generation (McPhee et al., 2016b). Given lexicase selection's success at maintaining diversity, the fact that such events occurred at all was unexpected. Further research, however, has suggested that these events are not important to lexicase selection's success at solving challenging problems and maintaining diversity (Helmuth et al., 2016b). Additional case studies of phylogenetic trees generated by lexicase selection suggest that lexicase selection promotes the existence of sub-populations that are focused on a specific selection criterion (McPhee et al., 2016a).

2.2.4 Eco-EA

Eco-EA (Goings and Ofria, 2009; Goings, 2010; Goings et al., 2012) is an ecology-inspired approach to maintaining diversity with respect to various fitness criteria. Each criterion is associated with a limited resource, that only confers fitness benefit in proportion to the amount remaining. Each resource flows into the environment at some rate, I, and flows out at some rate, O. Here, we use an inflow rate, I, of 100, and an outflow rate, O, of .01. Thus,

100 units of each resource enter the environment over the course of each generation, and 1% of the total quantity of each resource exits the environment at the end of each generation. Resources enter the environment at a constant rate as fitnesses are evaluated, minimizing stochastic effects from the order in which the population is evaluated.

A few other parameters are necessary for Eco-EA. First, there is C_f, the consumption fraction. Just as no organism in an ecosystem in nature is capable of consuming all of a resource in that environment, no solution in Eco-EA is. C_f specifies what fraction of the total quantity of resource any individual solution consumes. Next, there is m, the maximum amount of resource an individual is capable of consuming at any one time. For all experiments, I use $C_f = .0025$ and $m = 5$ to maintain consistency with previous Eco-EA research (Goings, 2010). Additionally, there is the c, the cost of attempting to use a hint. In order to create negative frequency dependence, there must be a cost to attempting to use a hint when too many other members of the population are also attempting to use it. In most scenarios, there is an implicit cost to attempting to use a hint, stemming from the trade-offs inherent in choosing to take one action over another. In a more traditional evolutionary computation setting, however, there is no such implicit cost. Without an explicit cost, there will not be negative frequency-dependence. Since negative-frequency dependence is a core element of Eco-EA, I impose a cost of $c = 1$. Lastly, if there is a cost to attempting to use a hint, then we also must define the range within which a solution is considered to be attempting to use that hint. This range is the niche width, n, which is set to 0.2 for all of these experiments. Since the maximum score for any hint function is 1, this means that solutions must score a minimum of 0.8 on a hint function in order to potentially get a reward or pay a cost from it.

In order to calculate the fitness impact of a resource, we need two more pieces of information. The first is R, the current amount of the relevant resource present. The second is s, the score on the associated fitness criterion, which is squared to incentivize even small increases in performance on the criterion. The amount of resource successfully used, A, is

calculated with the equation:

$$A = \begin{cases} 0 & s < 1 - n \\ min((s^2 * C_f * R) - c, m) & s \geq 1 - n \end{cases} \tag{2.3}$$

A is subtracted from the current amount of resource in the environment, and is used to update the base fitness of the current organism with the equation:

$$fitness = fitness * 2^A \tag{2.4}$$

Thus, successfully using more than c resource will multiply the solution's fitness by a number greater than one, whereas using less than c resource will multiply the solution's fitness by a number less than one.

These calculations are performed for all resources for all members of the population to determine resource-adjusted fitness for all candidate solutions. Tournament selection with a tournament size of two is then performed on the population based on these fitness values. To ensure that the impact of hints doesn't wash out small fitness gains on the main fitness function, every generation created with Eco-EA also contains a copy of the individual from the previous generation with the highest base fitness.

2.3 Statistical methods

Except where otherwise noted, all statistics are performed using the R statistical computing language (R Core Team, 2017) and plots are made with ggplot2 package (Wickham, 2016). Statistical significance for differences among groups is assessed using Kruskal-Wallis tests followed by post-hoc pairwise Wilcoxon rank-sum tests accompanied by Bonferroni corrections for multiple comparisons.

2.4   Code Availability

Avida is open source and freely available at https://github.com/devosoft/avida. All systems other than Avida were built using the Empirical library (Ofria et al., 2018), and the metric suites developed in chapters 8 and 7 are implemented there. The library is header only, and designed to be as easy to integrate into existing systems as possible. Code reliability is ensured with a suite of unit tests automatically run when code is added. As a proof of this concept, the Avida experiments presented in chapters 8 and 7 are carried out using a lightly modified version of Avida that incorporated this implementation of the metrics (Bryson et al., 2018).

Chapter 3

Spatial resource heterogeneity increases diversity and evolutionary potential

This chapter is adapted from a pre-print on *bioarxiv* (Dolson et al., 2017).

3.1 Introduction

A central problem in ecology is understanding the mechanisms that promote species coexistence and richness in natural ecosystems (Hutchinson, 1961; Chesson, 2000). There is mounting evidence that spatial heterogeneity (the lack of homogeneity across space) is an important driver of biological diversity (Stein et al., 2014). A common hypothesis, here termed the positive heterogeneity-diversity hypothesis, states that even seemingly small amounts of spatial heterogeneity promote diversity by creating new spatially-restricted niches for specialist species to inhabit, thereby reducing competition for limited resources (Pianka, 1966a,b; Stein et al., 2014). This hypothesis is supported by many cases in which spatial heterogeneity clearly influences the dynamics of species coexistence, allowing for greater species diversity than in homogeneous environments (Pacala and Tilman, 1994; Jankowski et al., 2014; Barnett and Beisner, 2007; Tilman, 1977; Harpole and Tilman, 2007).

Contrary to the positive heterogeneity-diversity hypothesis, some recent studies have

found evidence that diversity should peak at intermediate levels of spatial heterogeneity (Allouche and Kadmon, 2009; Allouche et al., 2012; Haller et al., 2013). These results are driven by two related factors: 1) the area-heterogeneity tradeoff and 2) decreasing contiguity of any given habitat.

The area-heterogeneity tradeoff is an observation that increasing the number of unique regions within an environment must decrease the average area spanned by each, if total geographic area is held constant (Allouche et al., 2012). In thinking through the effects of such a decrease, it is important to consider whether these regions represent: A) fundamental niches - the full range in which a species is capable of surviving, or B) preferred habitat - a set of conditions to which a species may be particularly well-optimized to outcompete others. Research suggesting that the area-heterogeneity tradeoff leads to a peaked heterogeneity-diversity relationship has so far only considered heterogeneity in fundamental niches for species of interest (Allouche and Kadmon, 2009; Allouche et al., 2012). Such a scenario is akin to dynamics theorized to occur in harsh climates, in which environmental gradients impose strict constraints on where a species can and cannot survive. The amount of space that is within a species' fundamental niche places a cap on its maximum population size. Since smaller populations have an increased risk of stochastic extinction, species whose fundamental niches occupy less area (relative to the ideal amount of area they need) are at greater risk of being lost from the community. These extinctions are believed to drive decreased diversity at high levels of heterogeneity. However this effect should not hold if we instead focus on heterogeneity of preferred habitats. If all species are capable of surviving anywhere in the environment, any stochastic extinction will allow for subsequent range expansion by other species and a corresponding drop in their risk of stochastic extinction. This process leaves an unexploited niche for future speciation. These factors should balance out such that, at worst, diversity remains constant as heterogeneity increases (this hypothesis is supported by the exploration of niche width in (Allouche et al., 2012)). This scenario is more akin to dynamics observed in the tropics, where species' realized niches are thought

44

to be constrained primarily by competition. While these two scenarios are both important to understand, and likely represent extrema of a continuum, here we will focus on preferred habitat heterogeneity.

Some metrics of environmental heterogeneity, such as edge-density, define more fragmented environments as more heterogeneous. In a system that defines heterogeneity along only a single environmental gradient, high values of these metrics can be achieved only if the environment is highly fragmented (as in (Haller et al., 2013)). Similar to the area-heterogeneity tradeoff, this decreased patch contiguity increases the likelihood of stochastic extinctions, reducing population diversity at high levels of heterogeneity. However, when niches are defined along more than one dimension, it is possible to increase heterogeneity without meaningfully increasing fragmentation. Fragmentation is an important issue to study, but it is only one component of heterogeneity. For simplicity, here we will assume that patches of any one resource are large enough to support a stable population, although intersections where multiple patches overlap may not be.

The development of a general theory of the impact of spatial heterogeneity on diversity has been inhibited by the fact that the causal mechanism, populations evolutionarily diversifying into new spatial niches, occurs over a long time scale. As evolution is the origin of biological diversity, a thorough understanding of the effects of spatial heterogeneity on diversity requires consideration of the associated evolutionary dynamics. Evolution occurs on a broad temporal scale, while the relevant ecological dynamics occur on a broad spatial scale. Dealing with either of these scales in a biological experiment is challenging, and their combination is effectively intractable. For this reason, we turn to a Avida to perform *in silico* experiments with digital organisms. In digital evolution, rapid generation times and transparent data allow us to perform large-scale experiments while perfectly controlling resource distributions and measuring organismal phenotypes.

Previous studies in Avida have found that introducing negative frequency dependence via resource competition is sufficient to generate and maintain diverse phenotypes through

adaptive radiation into multiple niches (Chow et al., 2004; Cooper and Ofria, 2003). In this study, we explore the effect of introducing a spatial component to this resource competition. Each organism in Avida occupies one cell in a two-dimensional grid. We create spatial heterogeneity by varying the set of resources available in each grid cell. Specifically, we place circular patches of resources in the environment. Since many patches will overlap, this placement technique creates many new spatial niches. We expect that subsets of the population will adapt to be able to use locally available resources. This process is closely related to ecological speciation (Schluter, 2009), but, as speciation is challenging to define for large and complex populations of asexual organisms such as ours, we will focus instead on the diversity of distinct phenotypes. Prior research on speciation in spatially heterogenous environments has demonstrated that some types of heterogeneity can indeed promote evolutionary branching (Haller et al., 2013); that study found a peak in diversity at intermediate levels of heterogeneity, which we expect is the result of focusing on heterogeneity along a single environmental gradient.

3.1.1 Evolutionary potential

Here focus on short-term evolutionary potential, which we define as the probability that a mutated offspring possesses new beneficial adaptations. It should be clear that the diversity of a community will be entangled with the combined evolutionary potential of its constituent populations: On one hand, diversity generally increases evolutionary potential (Lavergne and Molofsky, 2007; McDonald and Linde, 2002), since a community that is more spread out across a genotypic or phenotypic space is more likely to contain an organism that is a short mutational distance away from a given feature. There is also substantial empirical evidence from the field of evolutionary computation that diversity is important to finding globally optimal solutions when using evolution as a problem-solving tool (De Jong, 1975; Goldberg et al., 1987; Lehman and Stanley, 2008; Mouret and Doncieux, 2009). Evolutionary computation, as a context, allows us to sensibly talk about how "useful" diversity is as a whole

and which types of diversity are most effective toward achieving the desired goal (Mouret and Doncieux, 2009). This concept can also be translated back to biological evolution in contexts where we are interested in the evolution of a specific feature or set of features (such as increased complexity, or recovery of community traits lost due to a severe disturbance). Thus, understanding how to promote the evolutionary generation of diversity has important practical applications. Conversely, greater evolutionary potential facilitates the evolutionary generation of diversity, as populations that are able to adapt rapidly can take advantage of new niches more quickly.

A previous study in Avida on the relationship between the availability of resources and the resulting phenotypic diversity found that evolutionary potential peaks when all resources are intermediately plentiful. Interestingly, diversity peaks at lower resource levels than evolutionary potential (Walker and Ofria, 2012). These results are inconclusive, however, as to whether a direct relationship exists between diversity and evolutionary potential, and highlight the complexity of the relationship between diversity, resource availability, and evolutionary potential. Here, we attempt to clarify the causal relationships between these factors.

3.1.2 Quantifying environmental heterogeneity

Many metrics have been proposed to quantify spatial heterogeneity, but most fall into two categories: composition and spatial configuration (McGarigal, 2006). Composition metrics ignore the relative positions of resources, and instead deal with what types of spatial niches are present in an ecosystem. For example, commonly used composition metrics include *richness* (a count of how many distinct niches exist), *evenness* (do all niches cover the same amount of space?), and composites of the two such as *Shannon entropy*. For the purposes of this study, we use Shannon entropy of the number of cells in which each set of resources is present as a metric of environmental composition. This metric balances both richness and evenness and is a generally accepted metric of spatial heterogeneity (Forman

and Godron, 1986).

Spatial configuration metrics, on the other hand, deal with arrangements of niches in space and include *patch shape complexity*, *core area*, *contrast*, and *isolation*. There is no single metric that can accurately summarize spatial configuration—for any given study it is important to identify which variables are meaningful. Here, we track the extent to which single-resource patches overlap to form multi-resource spatial niches. Because we define an organism's phenotype as the set of resources that it is able to consume, a spatial niche containing more resource types will be hospitable to a wider variety of phenotypes. Moreover, because of the tendency of simpler functions in Avida to facilitate the evolution of more complex ones, niches with more types of resources will be easier for populations to adapt to. Thus, spatial resource overlap is the primary configuration metric we track here. This overlap is captured by the distribution created by counting the number of resources in each cell of an environment. We will discuss overlap in terms of the mean, variance, skew, and kurtosis of this distribution.

3.2   Results and Discussion

3.2.1   Niche stability

Our goal is to isolate the effects of heterogeneity from the effects of fragmentation. To do so, we must first determine the minimum patch size that would be required to stably support populations specialized to live in those regions.

Likewise, we need to identify the minimum overlap between two patches to ensure the resulting spatial niche is similarly stable. We measured these threshold sizes empirically by carrying out two series of experiments in which the world contained only two circular resource patches. In the first set of experiments, we tested all possible pairs of resources at three levels of overlap: complete, half-way, or none (Fig 3.1). For each region, we defined the optimal phenotype as one that performed a set of functions that perfectly matched the
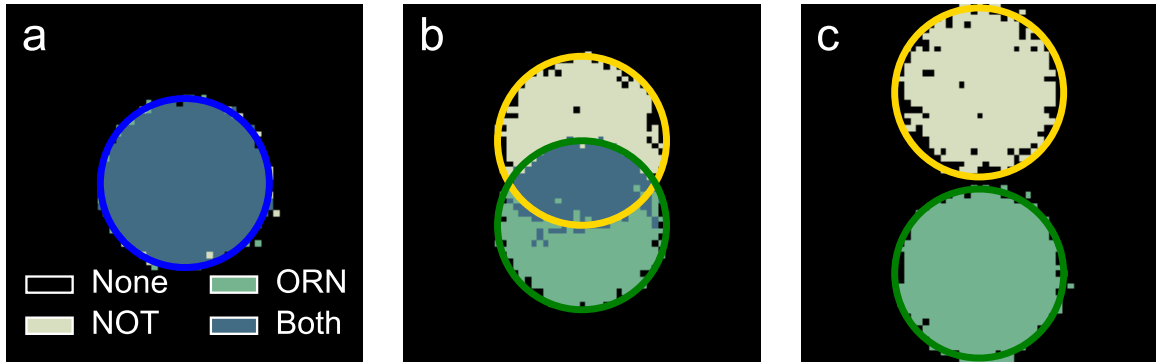
Figure 3.1: **Spatial arrangement of phenotypes at the end of each experiment in worlds containing patches of two different resources**, NOT and ORN, each radius 14. Patch locations are indicated by circles - gold represents the NOT resource, green represents the ORN resource, and blue represents both resources overlapping. The color of each cell in the grid indicates the most common phenotype in that cell across the 30 replicates. **A**: Patch centers are 0 cells apart, creating complete overlap. **B**: Patch centers are 14 cells apart, leading to three coexisting phenotypes. **C**: Patch centers are 30 cells apart, resulting in complete patch separation and no phenotype capable of both functions.

available resources.

The optimal phenotype for the spatial niche at the intersection of the two resource patches evolved consistently when the organisms needed to perform sufficiently simple functions (Fig 3.1). For example, the optimal phenotype evolved 66% of the time across all 30 replicates within all pairwise combinations of the resource associated with the simplest function and the resources associated with the five next simplest functions. The probability of evolving the optimal phenotype at the intersection within a given resource pair varied from a low of 30% to a high of 100%, suggesting that it is easier for organisms to evolve some pairs of functions than others.

This outcome is likely the result of one of the valuable functions serving as a building block for the other function (Lenski et al., 2003). Such interaction effects illustrate the complex evolutionary dynamics that spatially heterogeneous environments can cause.

In our second round of experiments, we chose a single pair of resources, but tested all possible amounts of overlap. If resource patches have sufficiently small overlap, spatial niches
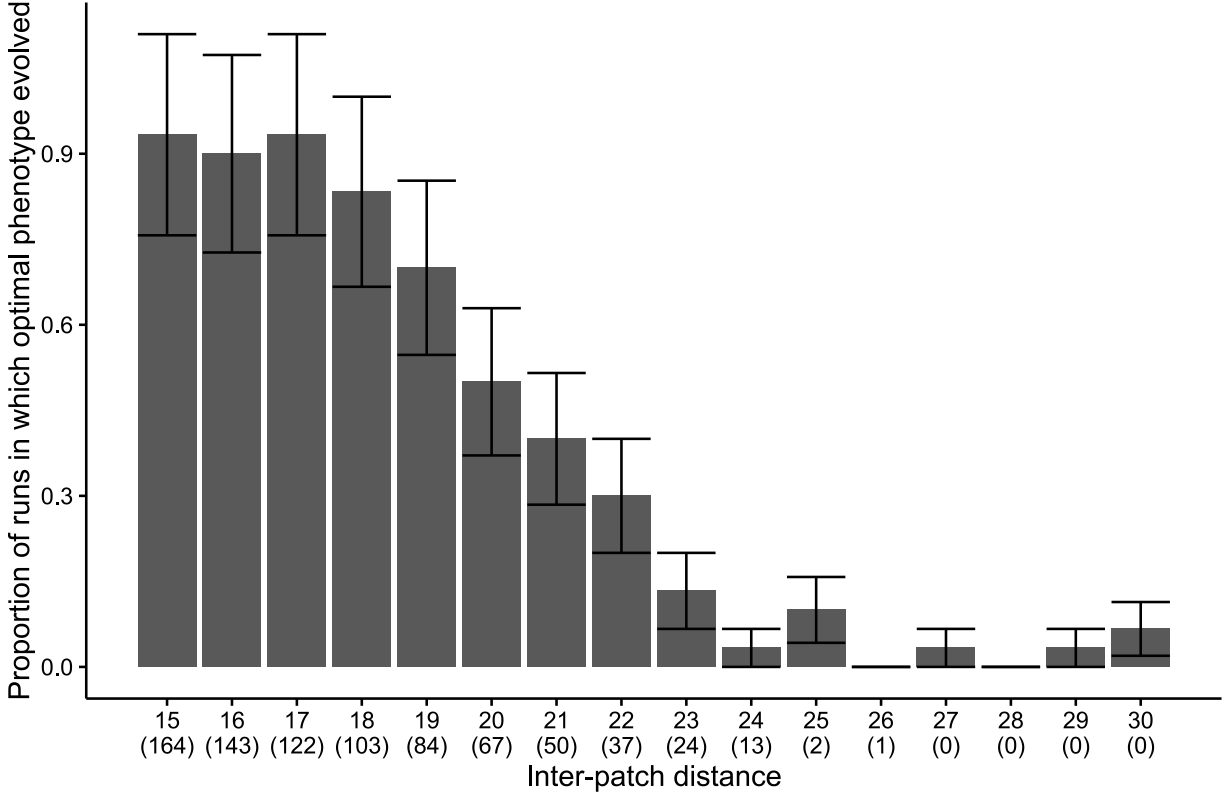
Figure 3.2: **Proportion of populations (out of 30) at each inter-patch distance that contained organisms performing both functions** (NOT and ORN) after 200,000 updates (approximately 4,000 generations). Error bars represent estimates of the standard deviation under the assumption that the underlying pattern resembles a Poisson distribution. Number of cells in the area where the patches overlap is specified in parenthesis for each distance.

should not be able to maintain a phenotypes capable of both functions. While phenotypes can extend beyond their target niche, their fitness will decrease in neighboring niches. As spatial niche size decreases, the probability that the matching phenotype will be lost from the population due to drift increases. As expected, the probability of the two-resource niche being occupied by organisms that could perform both valuable functions decreased with niche size. The configuration with 21 spaces between the centers of the circles, translating to a niche size of 50 cells, was the smallest at which this probability was greater than 50% (Fig. 3.2).

3.2.2   Effect of spatial heterogeneity on phenotypic diversity

Having established that sufficiently large overlaps between patches create new niches, we can now move on to more complex environments with more patches and more heterogeneity. We used three different techniques to arrange patches in the environment. In the first, we anchored patches in fixed positions and controlled spatial heterogeneity by adjusting patch size (Fig. 3.3). In the second, we gave all patches the same radius and controlled heterogeneity by setting the distance between their center points (Fig. 3.4).

In the third, we randomly selected the number, size, and location of patches in the environment, resulting in a variety of spatial heterogeneity levels (Fig. 3.5) (for details, see methods). The resource associated with the most complex logic function, EQU, was consistently available. We tracked the evolution of EQU as a proxy for the evolutionary potential of the population. However, the evolution of EQU should diminish the importance of spatial heterogeneity. To account for this confounding factor, we measured the diversity for each replicate as the phenotypic diversity of viable organisms in the population at the last time point prior to the evolution of EQU. For replicates in which EQU never evolved, we used phenotypic diversity from the final time point. To ensure that the presence of these two types of replicates did not bias the results, we repeated all analyses using data from the final time point for all replicates, and did not find any substantive difference.

Average phenotypic Shannon diversity differed significantly across resource patch sizes and proximities (Size: Kruskal-Wallis test, Chi-squared = 237.05, p< 2.2E-16, Fig. 3.3; Distance: Kruskal-Wallis test, Chi-squared = 104.0322, p < 2.2E-16, Fig. 3.4). Both size and distance experiments showed trends in phenotypic diversity qualitatively similar to the trends in spatial heterogeneity. The locations of the peaks in diversity overlapped with the peaks in environmental entropy (see Fig. 3.3 and Fig. 3.4).

The third experimental treatment involved continuous rather than categorical variation in spatial heterogeneity, and we used a simple linear model regressing phenotypic Shannon diversity on environmental entropy to analyze the results (Table 1, 3.5). We also generated
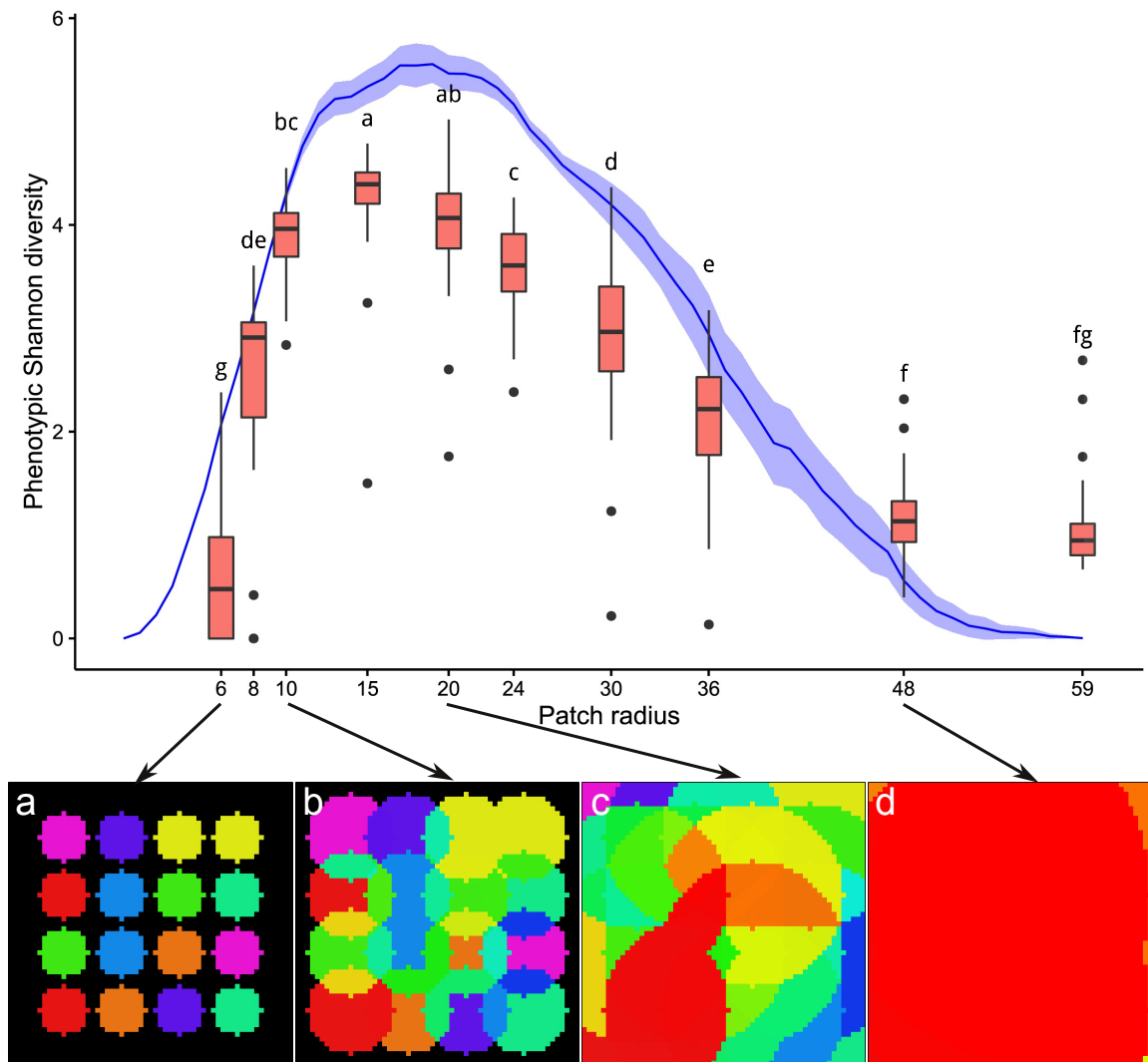
Figure 3.3: **Shannon entropy (representing both phenotypic Shannon diversity and environmental entropy) across resource patch radii**. The blue line shows environmental entropy at each radius (mean and standard deviation, calculated from 100 randomly generated environments at each radius). Environmental entropy peaks at patch radius 18. In environments with patch radii less than this peak (see sub-figures a and b), resources are more spatially isolated; conversely, in environments with patch radii greater than this peak, resources increasingly overlap (see sub-figure c). The high-radius extreme represents a homogeneous control environment (see sub-figure d), as all patches overlap and completely fill the world. Final phenotypic diversity at each radius condition tested in Avida is shown by the boxplots (n=30 for each radius; boxes show 25%, 50%, and 75% quantiles, with whiskers extending up to 1.5*IQR). Radii that do not share a letter have significantly different distributions of phenotypic diversity (pairwise two-sided Wilcoxon tests with Bonferroni adjustment, p<0.05). Sub-figures a-d show representative environments for the radius sizes indicated. Phenotypic diversity appears to peak at a patch radius somewhere between 13 and 23, an interval which includes the peak in environmental entropy.
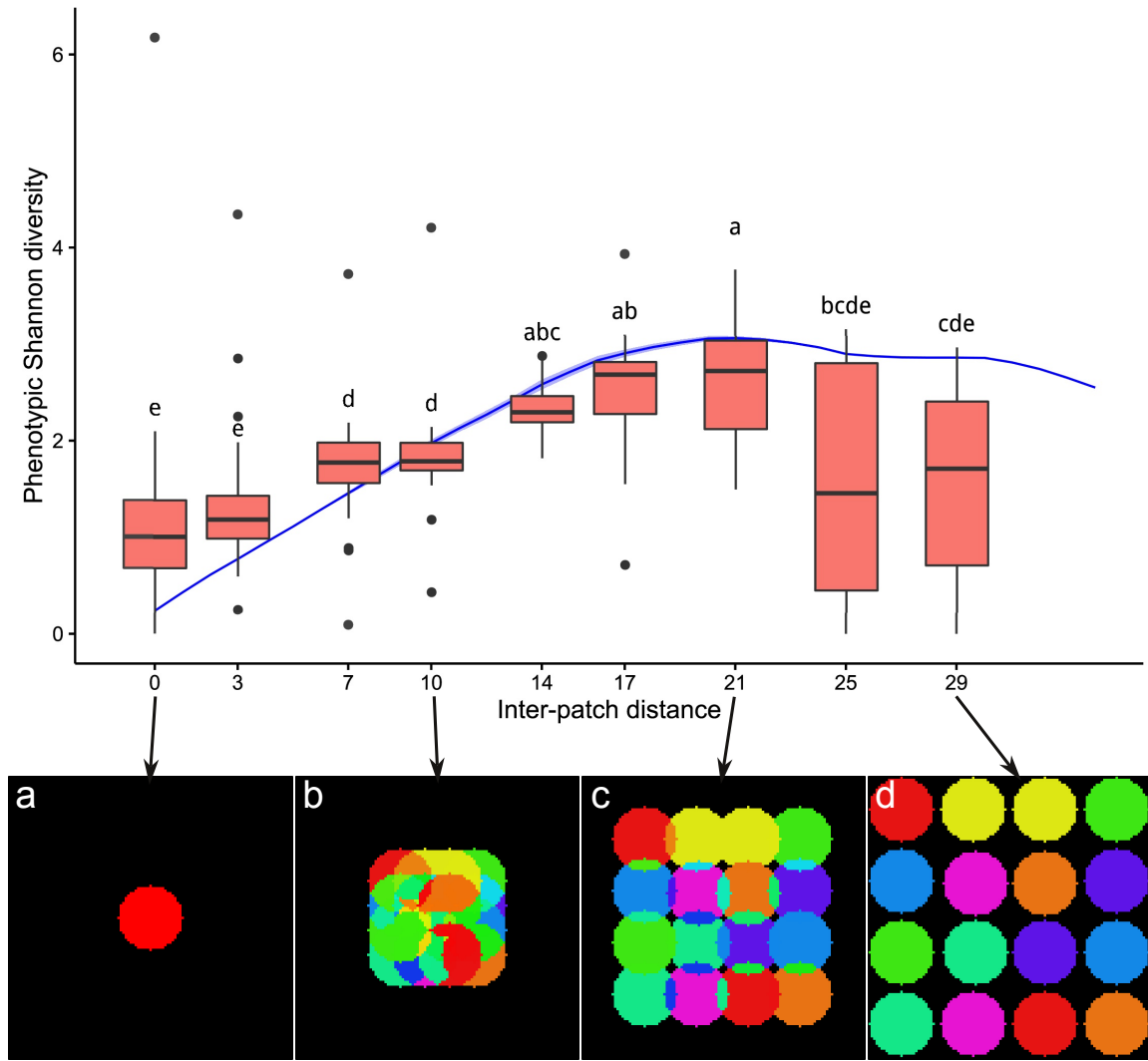
Figure 3.4: **Shannon entropy (representing both phenotypic Shannon diversity and environmental entropy) across different resource inter-patch distances.** The blue line shows environmental entropy at each distance (mean and standard deviation, calculated from 100 randomly generated environments at each distance; note that the standard deviation is small and hard to see). Environmental entropy peaks at inter-patch distance 21 (sub-figure c). In environments with inter-patch distances greater than this peak, resources are more spatially isolated (sub-figure d); conversely, at inter-patch distances lower than this peak, resources increasingly overlap (sub-figures a and b). Distance 0 represents a homogeneous control environment, as all resources overlap completely. Boxplots show final phenotypic diversity at each level tested in Avida (n=30 for each treatment; boxes show 25%, 50%, and 75% quantiles, with whiskers extending up to 1.5*IQR). Distances that do not share a letter have significantly different distributions of phenotypic diversity (pairwise two-sided Wilcoxon tests with Bonferroni adjustment, p<0.05). Sub-figures a-d show representative environments for the inter-patch distances indicated. Phenotypic diversity appears to peak at an inter-patch distance somewhere between 14 and 21, an interval which includes the peak in environmental entropy.

y = Phenotypic Diversity			
Variable	**Estimate**	**SE**	**CI**
Intercept	1.53 ***	0.07	(1.39, 1.67)
Environmental Entropy	0.32 ***	0.02	(0.28, 0.37)
Residual standard error: 0.71 on 498 DF			
$R^2 = 0.28$, F-statistic = 196.3 on 1 and 498 DF			

Table 3.1: **Regression table for model predicting phenotypic diversity**, measured as Shannon entropy of viable organisms on the last measured update before EQU evolved (or the last update if EQU never evolved), from spatial heterogeneity, as measured by Shannon entropy of niches in the environment.

models including patch overlap as a metric of spatial configuration, but found no support for retaining the additional terms. Based on this model, we estimate that phenotypic diversity will increase by approximately one third of a unit, on average, for every unit that spatial heterogeneity increases - a fairly substantial effect.

Across the board, populations that evolved in spatially heterogeneous environments were more diverse than those evolved in more homogeneous environments, regardless of how that heterogeneity was generated. Increasing spatial heterogeneity partitions resources into spatially structured areas, thereby creating new, unique niches into which organisms can diversify and partially escape competition. This trend is in contrast to the peak in diversity found at intermediate heterogeneity by some other experiments (Haller et al., 2013; Allouche et al., 2012). We suspect that this result occurred because, in our experiments, increased heterogeneity does not necessarily come at the cost of decreased fundamental niche size. Our results agree with the many biological studies that have found a positive relationship between phenotypic diversity and spatial heterogeneity (Stein et al., 2014).

3.2.3 Spatial heterogeneity and evolutionary potential

The proportion of replicates in which the most complex task (EQU) evolved had a positive saturating response to increasing patch radius. As patch radius increased from low
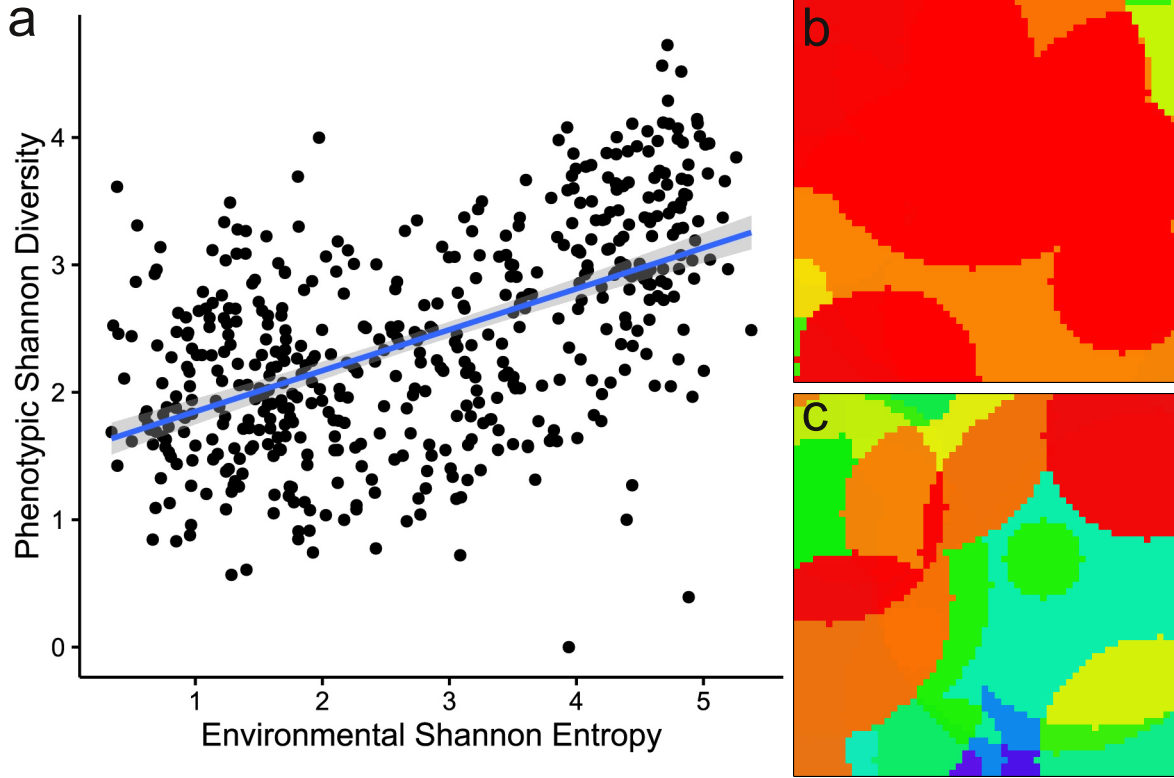
Figure 3.5: **Results from randomly generated environments**. **A**) Scatterplot of environmental entropy vs phenotypic Shannon diversity across different randomly generated environments (Method 3) (n=500). Line represents the linear model shown in Table 1, and shaded area represents 95% confidence interval for the slope. **B**) and **C**) are two examples of randomly generated environments.

(6-10) to intermediate values (15-24), populations evolved the EQU function significantly more frequently. (Fig. 3.6). Evolutionary potential reached a maximum saturating response such that we were not able to detect a significant difference between the probabilities of EQU evolving in trials with medium and high patch radii (> 24) after a Bonferroni correction (Fig. 3.6).

The inter-patch distance experiment highlighted the importance of patch overlap relative to spatial entropy in evolving EQU. At low distances between patch centers (0-14), resource patch overlaps and evolutionary potential were at their highest levels (Fig. 3.7). At intermediate to large inter-patch distances, evolutionary potential decreased, as did the intersection between patches. Larger overlaps have multiple resources available, allowing populations to build up to more complicated functions with greater ease than than in conditions in which patches overlap less (Lenski et al., 2003).

The common trend from the first two experiments is that the amount of overlap between resource patches (i.e. the number of different resources present within a spatial niche) has a stronger impact on evolutionary potential than diversity. To disentangle the effect of resource overlap (our spatial configuration metric) and spatial heterogeneity (our spatial composition metric) on evolutionary potential, we built a linear model using the overlap as a predictor. Recall from our discussion of Quantifying Heterogeneity that we measured overlap by looking at the distribution of the number of resource types per cell in an environment, and taking its mean, variance, skew, and kurtosis. Ultimately, these statistics capture similar information about the environment to that captured by spatial entropy, as demonstrated by the high R^2 value of the linear model predicting environmental entropy from the mean and skew of number of resources available per cell (see Table 2). Variance and kurtosis were dropped from the final model because of the high colinearity between mean and variance and between skew and kurtosis. Thus, we did not include spatial entropy in the overall model of evolutionary potential, as it would have added largely redundant information.

We used a logistic regression to predict the probability of evolving the EQU function
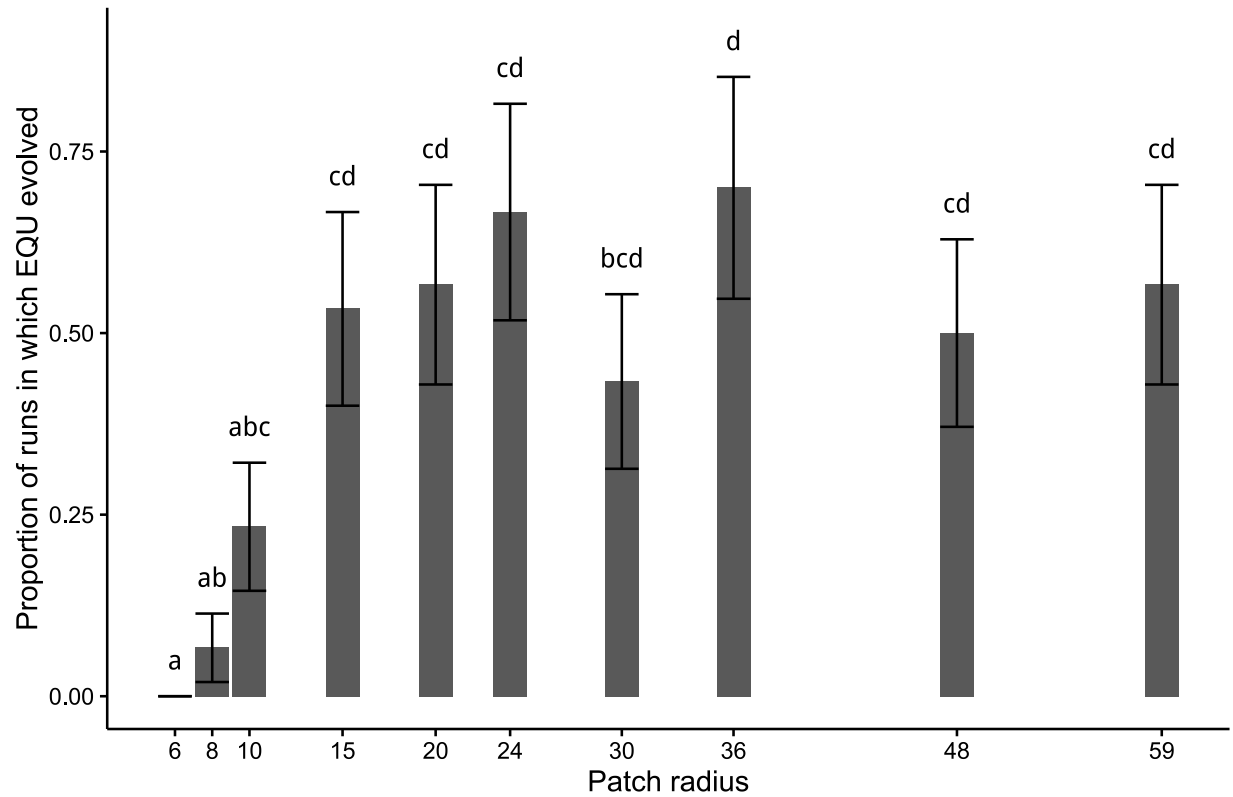
Figure 3.6: **Proportion of populations that evolved EQU across different patch radius sizes** (n=30 per treatment). Error bars are estimates of standard deviation based on the square root of the count of populations evolving EQU for each condition. Radii that do not share a letter have significantly different probabilities of evolving EQU (pairwise two-sided Fisher's exact tests with Bonferroni adjustment, $p<0.05$).
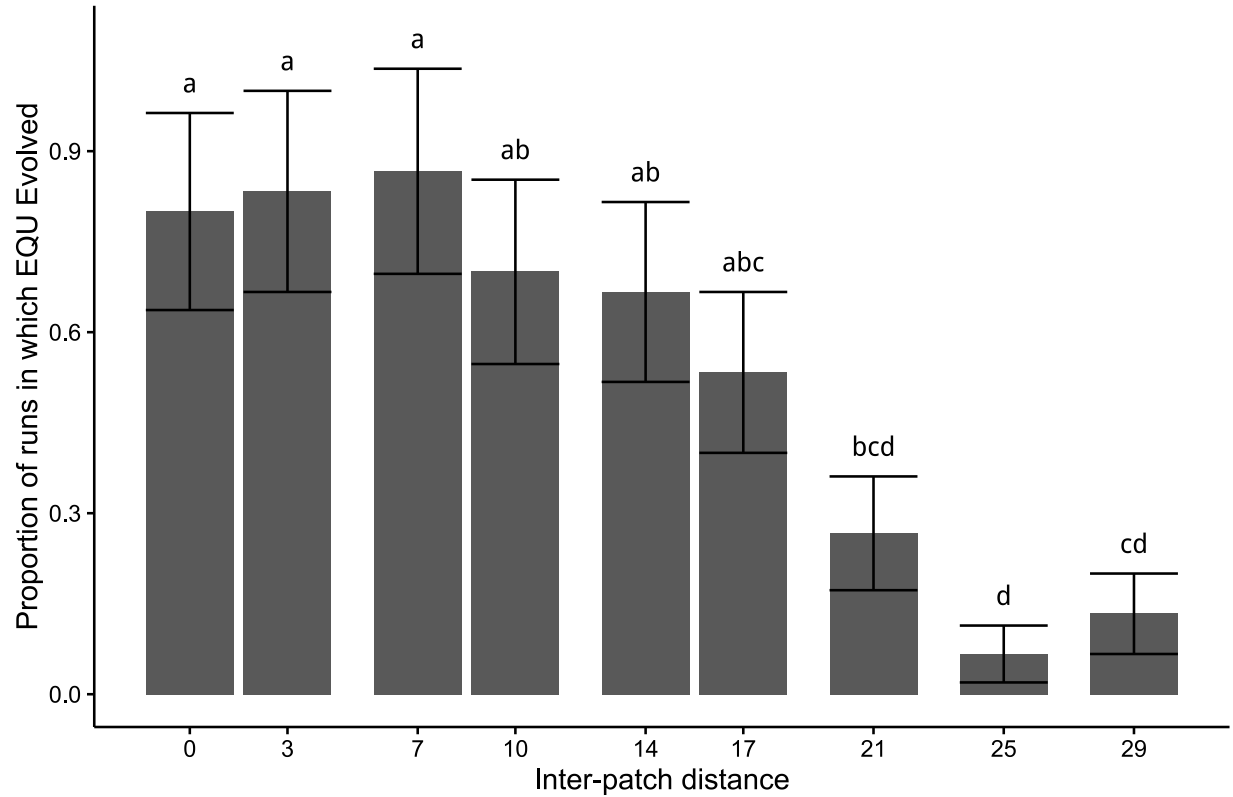
Figure 3.7: **Proportion of populations that evolved EQU across different inter-patch distances** (n=30 per treatment). Error bars are estimates of standard deviation based on the square root of the count of populations evolving EQU for each condition. Inter-patch distances that do not share a letter have significantly different probabilities of evolving EQU (pairwise two-sided Fisher's exact tests with Bonferroni adjustment, p<0.05).

y = Spatial Entropy			
Variable	**Estimate**	**SE**	**CI**
Intercept	4.27***	0.07	(4.12, 4.41)
Mean resources/cell	1.75***	0.19	(1.37, 2.12)
Skew resources/cell	0.24***	0.02	(0.19, 0.29)
Mean * Skew	-2.62***	0.08	(-2.77, -2.47)

Residual standard error: 0.36 on 496 DF
Adjusted R^2 = 0.93, F-statistic = 2320 on 3 and 496 DF

Table 3.2: **Regression table for linear model predicting entropy of the environment** from the mean and skew (and their interaction) of the number of resource types per cell.

given phenotypic diversity and environmental heterogeneity (Table 3). As environmental descriptors, we used mean number of resources per cell, skew of the number of resources per cell, and their interaction. We included phenotypic diversity (before EQU evolved) as the final predictor in the model of evolutionary potential, due to its positive association with spatial heterogeneity. As expected, this model suggests that diversity is important to evolutionary potential. Every additional unit of phenotypic diversity increases the odds of EQU evolving by approximately 4.3 times. However, this effect is dwarfed by the effect of mean resources per cell (i.e. overlap). Increasing the average number of resources per cell by 1 (e.g. adding an additional resource patch that covers the entire grid) would increase the odds of EQU evolving by approximately 148.8 times. The interaction between diversity and mean resources per cell has an effect of a similar magnitude. This interaction suggests that either A) environments with more overlap amplify the benefits of diversity to evolutionary potential, B) environments with more overlap promote the evolution of diversity that is more useful to evolving EQU than the diversity promoted by other environments, or C) both.

Ultimately, the amount of overlap between single-resource patches is the strongest driver of the evolutionary potential of EQU, with global phenotypic diversity playing an important secondary role. This impact of diversity may be responsible for the saturating evolutionary potential curves seen in the first two experiments; during the plateaus in these curves, overlap is decreasing while diversity is increasing. Given the strength of the synergies between

y = EQU Evolution			
Variable	**Estimate**	**SE**	**CI**
Intercept	1.30***	0.17	(0.97, 1.66)
Phenotypic Diversity	1.46***	0.22	(1.05, 1.90)
Mean resources/cell	5.00***	0.84	(3.40, 6.71)
Skew resources/cell	0.74***	0.21	(0.32, 1.13)
Diversity * Mean	4.38***	0.93	(2.60, 6.25)
Diversity * Skew	0.81**	0.26	(0.30, 1.32)
Mean * Skew	1.70	0.81	(-0.00, 3.21)
Diversity * Mean * Skew	2.65**	1.01	(0.64, 4.62)

Null deviance: 651.08, 499 DF, Residual deviance: 549.88, 492 DF

Table 3.3: **Logistic regression table for linear model predicting whether or not EQU evolved** from phenotypic diversity, mean resource types per cell, and skew of resource types per cell, as well as all interactions between these variables.

functions in Avida, it is possible that the effect of overlap in this system is greater than it would be in other contexts. These results support the idea that it is possible to select for diversity that is more or less useful to evolving a specific feature.

3.2.4 Conclusion

We have presented evidence supporting the positive diversity-heterogeneity hypothesis. We found that population diversity increases in response to higher levels of spatial heterogeneity, rather than peaking at an intermediate level. Further, we have shown that this increase in diversity leads to higher evolutionary potential of a complex function. However, the spatial configuration of the landscape in which a population is evolving can have a much stronger effect on evolutionary potential than diversity alone. The presence of a synergistic interaction between diversity and spatial configuration provides preliminary evidence that landscape configuration can promote diversity that is more or less useful to evolving a given trait. These findings help to solidify the growing consensus that there is a causal relationship between spatial heterogeneity and population diversity. Moreover, by exploring these changes in diversity in an evolutionary context, we push toward a more mechanistic understanding of spatial heterogeneity and diversity.

However, a number of open questions remain. Most notably, we only considered diversity as phenotypic Shannon diversity within individual populations. It is possible that other measures of diversity, such as beta-diversity across populations, would have different relationships to spatial heterogeneity. Similarly, the only metric of spatial heterogeneity that we explored was Shannon entropy. There are many alternative metrics, particularly metrics of spatial composition (see Quantifying Heterogeneity), that may have interesting subtleties in their relationship to diversity. Additionally, allowing sexual recombination, motility, or the evolution of dispersal distance could produce substantially different results.

Our results have implications beyond ecology; for example these same dynamics can be harnessed to improve computational problem solving with evolutionary algorithms. A common approach to solving hard problems in these systems is to first evolve solutions to easier related problems and use them as building blocks (Lenski et al., 2003; Bongard and Hornby, 2010; Grabowski and Magaña, 2014). Spatially heterogeneous environments can be created by rewarding population members that solve a range of simpler problems in different parts of a spatially structured population. Recent research on an evolutionary algorithm that creates spatial heterogeneity via host-parasite co-evolution supports the idea that this approach may be useful (Harper, 2012). Thus, it may be possible to use the results of our study to create environments that are particularly conducive to solving problems with evolutionary computation. Such an approach would synergize well with previously discovered positive impacts of spatial structure on diversity maintenance in evolutionary computation (Whitley et al., 1998; Cantú-Paz and Goldberg, 2003).

Understanding drivers of diversity, particularly with regard to their impact on evolutionary potential, also has important practical implications for preserving biodiversity in natural systems. Landscapes world-wide are experiencing new disturbance regimes and there is currently little theory predicting how ecosystems will respond (Arnell, 2004; Galloway et al., 2004). Anthropogenic disturbances tend to be a homogenizing force (McKinney, 2006), but this effect can be mitigated by urban planning if we understand the evolutionary and

ecological dynamics at play. Given the theorized importance of biodiversity to ecosystem stability and function, knowing how spatial heterogeneity can affect diversity across different ecosystems will be critical to achieving such goals (Craine et al., 2013; Wilmers and Getz, 2005). Thus, going forward, it will be important to solidify our understanding of the factors affecting the shape of the relationship between spatial heterogeneity and diversity.

3.3 Experimental Design

Specific details for experimental treatments are described below. For each, we performed 30 replicates and initialized all configuration settings to Avida defaults (unless otherwise specified) with the exception of copy mutations, which we set to a probability of 0.0025 per instruction copied. We allowed each replicate to evolve for 100,000 updates. An update is the amount of time it take for the average organism in the population to execute 30 instructions; 100,000 updates translates to approximately 30,000 generations on average.

We recorded phenotypic Shannon diversity of organisms and the number of populations that had at least one organism that performed the most complex logic function at the end of the experiment. We define an organism's phenotype as the set of logic functions it performs. We performed all statistics in R version 3.0.1 (Wickham, 2016; R Core Development Team, 2013). For each experimental condition, we ran a large number of replicates relative to the amount of variation in results, ensuring sufficient power to detect reasonably-sized effects. Because each replicate had a unique random number generator seed, all data points were independent and, as such, the assumptions for all statistical tests were met.

3.3.1 Niche stability

For the niche stability experiments, each circle had a radius of 14 cells. We set up the patches to be perfectly on top of each other in the complete-overlap condition (0 cell distance between centers), 14 cells between centers for halfway overlap, and 30 cells for no overlap.

We tested these conditions in a full-factorial design with 30 replicates of all potential pairwise combinations of resources, for a total of 192 treatments.

We used the two simplest logic functions in the functional niche-size assessment experiments because both functions consistently evolved in all of the two-circle overlap experiments where they were present. These patches also had a radius of 14, and between 0 and 30 cells between their centers. To ensure that niches created in this experiment were ecologically stable, we allowed these replicates to continue for an additional 100,000 update "ecology phase" with no mutations. We measured niche stability as the frequency with which intermediate phenotypes evolved and persisted until the end of the experiment.

3.3.2   Fixed layout experiments

To systematically measure the effect of spatial heterogeneity on diversity, we ran two sets of experiments with 16 resource patches (two for each resource type) placed at pre-determined locations in the environment. In the *patch radius experiment*, we varied the size of each circular patch, while in the *inter-patch distance experiment*, we varied the spacing between fixed-sized patches. To select treatment values for each of these variables, we wrote a simulation in Python that created random environments with the specified parameters and measured their heterogeneity (available at https://github.com/emilydolson/resource-heterogeneity). The results of these runs produced the data for the curves in figures 3 and 4. We used these curves to select values of each variable that would produce a range of levels of environmental heterogeneity. Across both sets of experiments, the most complex logic function was valuable everywhere to ensure that it would be maintained in the population long enough to be measured if it evolved. We designed many treatments within each experiment (detailed below) and ran 30 replicates for each treatment. While the positions of each resource patch were consistent within a treatment, we randomized the type of resource in each patch across replicates.

In the patch-radius experiment, we manipulated spatial resource heterogeneity by chang-

ing the radius sizes of the single-resource patches. We placed patch centers 12 spaces apart. Based on the results of the Python simulation, we used patch radii of 6, 8, 10, 15, 20, 24, 30, 36, 48, and 59 as our experimental conditions (see Fig 3.3). Changing patch radii introduces a potentially confounding variable: the overall amount of resources in the environment. To control for this additional factor, we ran another experiment in which we manipulated spatial resource heterogeneity by changing the distances between the resource patches while holding patch radius constant. At each inter-patch distance, we set resource patches to a constant size (radius 14) in a 115 x 115 non-toroidal world. We manipulated distances between patch centers across treatments (using distances 0, 3, 7, 10, 14, 17, 21, 24, or 29) (see Fig 3.4). This condition had a constant quantity of each resource, but introduced its own constraints on the types of environments that could exist at each level of heterogeneity. In combination with each other, however, these two sets of experiments address the effect of heterogeneity in a more comprehensive way.

We used Kruskal-Wallis tests to determine if the distribution of phenotypic Shannon diversity varied significantly across treatments. Subsequently, we determined where average diversity peaked using post-hoc pairwise Wilcoxon tests with a Bonferroni correction for multiple comparisons. We used Fisher's exact tests to determine if the number of populations that performed the the most complex logic function by the end of the experiment changed across distance treatments. We performed additional Fisher's exact tests to determine where evolutionary potential peaked.

3.3.3   Randomized patch experiments

In the two prior experiments, it is hard to disentangle the specific effects of patterned resource overlap from the more general effects of heterogeneity, as these variables covaried. To sort out these factors, we placed a random number of patches of random sizes in random locations in the environment. Using the Python simulation from the previous section to determine parameter ranges that would produce the widest variety of levels of spatial

heterogeneity, we tested 500 environmental configurations, consisting of 4-124 patches with radii between 6 and 35. Ultimately, the total number of resource patches in an environment was the primary determiner of its heterogeneity; varying patch sizes and placements achieved variation in patch overlap within a given level of heterogeneity.

Because these experiments produced continuous rather than categorical data, we used linear regression models in R to determine the factors influencing phenotypic diversity and spatial heterogeneity. For the logistic regression predicting evolution of EQU, we used a generalized linear model. We narrowed down potential models using AIC; we then evaluated and trimmed factors with high collinearity until we found the minimally adequate model. We also fit a model that included total sum of resources across all cells, but there was a curvilinear relationship between total sum of resources and spatial entropy that rendered the factors largely redundant and added little explanatory power. We identified a substantial outlier, in the form of a single run in which the population went extinct, but removing it had negligible effect on our statistical analyses, so it was included.

3.3.4 Code and Data Availability

These experiments were carried out using Avida version 2.12.4. Data and scripts used to generate different environmental conditions and analyze data are available via Zenodo with DOI 10.5281/zenodo.162981 (Dolson et al., 2016a).

Chapter 4

Spatial resource heterogeneity creates local hotspots of evolutionary potential

This chapter is adapted from a paper that appeared in the 2018 Artificial Life Conference (Dolson and Ofria, 2017).

4.1 Introduction

Understanding the evolutionary processes that facilitate the evolution of complex traits is at the heart of many open questions in artificial life, evolutionary computation, and evolutionary biology. Numerous studies have suggested that spatially complex environments can drive the evolution of complex traits (Doebeli and Dieckmann, 2003; Wagner et al., 2013; Dolson et al., 2017), but it is unknown whether this result is a global pattern that is only visible at the scale of the entire environment, or whether it is driven by local spatial dynamics. If sub-regions exist that disproportionately promote the evolution of specific adaptive traits (*evolutionary hotspots*), they would be a valuable phenomenon to understand. Understanding what factors produce evolutionary hotspots should help us design environments that are more likely to evolve complex traits. This ability would also enable us to build more interesting artificial life systems, more effective evolutionary algorithms, and biological

environments that promote desirable evolutionary outcomes.

Where should we expect to find evolutionary hotspots? Instinctively, we might expect a trait to arise only in regions where it is beneficial. However, selection can only favor traits once they appear, not promote that initial appearance, so other aspects of the environment must play a critical role. At the other extreme, hotspots seem unlikely in a homogeneous environment with no spatial structure. However, few environments in nature are truly homogeneous, and almost all have some form of spatial structure. Even among digital systems, once we consider the fact that any kind of interaction between neighboring individuals introduces spatial heterogeneity, few are truly spatially homogeneous. Thus, understanding if and how evolutionary potential varies across space could be important for understanding evolutionary dynamics in a variety of systems.

Why should we expect evolutionary hotspots to exist? To better understand this idea, it is useful to consider Wright's fitness landscape metaphor (Wright, 1932). At its simplest, a fitness landscape is a three-dimensional surface representing two traits (x and y) with varying values, and the fitness of an organism having a given combination of those trait values (z). When there are gradients in this space, we expect populations to climb to a fitness peak over time. Typically, populations will then get stuck at the peak of a local gradient, regardless of whether it is the highest location in the landscape. As such, the topography of the local fitness landscape plays a pivotal role in determining what traits will evolve in a population. A wide variety of factors can influence what part of a fitness landscape a population ends up in at a given point in time (Wilke et al., 2001). I predict that spatial variation across the environment is one such factor. A spatially heterogeneous environment can be conceptualized as a network of different fitness landscapes, one for each location in space. However, in a dynamic, interdependent ecological community, these fitness landscapes shift over time. Some local landscapes will tend to guide portions of the population toward regions of mutational space from which new traits are more easily accessible.

Of course, this network of fitness landscapes may not guide evolving populations in a

consistent direction. While it is possible that landscapes could align to perfectly guide a lineage through the most challenging steps in the evolution of a complex trait, this condition seems unlikely. We might just as easily expect that the lack of a constant landscape will only serve to increase stochasticity, rather than having a predictable effect. Perhaps the fitness landscapes in adjacent locations even work at cross-purposes, with each guiding the population in opposite directions. Given this uncertainty, the primary goal of this chapter will be to determine whether hotspots actually occur with sufficient frequency and strength to be worthy of further consideration. Addressing this question requires vast numbers of repetitions of experimental evolution in complex spatial environments. Conducting such an experiment in a wet lab system would not be worth this immense effort until additional theory has been developed. Once we have generated concrete hypotheses about what we would expect to see in biology, more targeted follow-up experiments in a wet lab system may be worthwhile. Here, I lay the groundwork for such theory by performing experiments in a digital system, where I will be able to precisely control the spatial distribution of environmental conditions and measure the drivers behind any spatial variation in evolutionary potential that I detect.

Here, I focus on a simple environment to avoid adding unnecessary complications to an already complex question. To this end, I use sessile, asexual organisms in an unchanging abiotic environment where different sets of traits are advantageous in different regions. Importantly, these traits are related to each other. This set-up is roughly analogous to a wide variety of scenarios within biology, ranging from bacteria surviving on a surface treated with a range of antibiotics (similar to (Baym et al., 2016)) to plants adapting to survive on different soil types within a region.

4.2 Biological Background

Biologists have traditionally studied patterns of diversification rather than patterns of novel trait evolution (but see (Baym et al., 2016)). These processes have important

distinctions from each other; diversification usually requires either physical separation of sub-populations or conflicting selection pressures, whereas novel trait evolution does not. Moreover, they are maximized by different conditions (Walker and Ofria, 2012). Nevertheless, diversification and novel trait evolution are closely related, as diversity is thought to increase evolutionary potential (Rouzic and Carlborg, 2008). The most well-studied biological pattern pointing to spatial variation in evolutionary processes is the latitudinal diversity gradient: a consistent pattern of high biodiversity near the tropics, and progressively less toward the poles (Hillebrand, 2004). While it remains impossible to perform controlled evolution experiments on a planetary-scale, progress has been made on untangling the drivers of this pattern. Most importantly for the purposes of this chapter, evidence is mounting that diversification rates are higher in the tropics (Mittelbach et al., 2007). The drivers of this variation are still debated, but many hypothesized drivers are related to spatial dynamics within the environment, most notably the idea that the tropics may contain more environmental gradients, barriers to gene flow, and other spatial properties that promote speciation (Moritz et al., 2000; Doebeli and Dieckmann, 2003). Although these dynamics are more complex than those addressed here, particularly given that many rely on sexual reproduction, they suggest that spatial variation in evolutionary processes has broad relevance.

In related work, conservation biologists have sought to locate regions with an elevated intensity of evolutionary processes. This research stemmed from an increasing recognition that preserving the processes that generate diversity is at least as important as preserving existing diversity (Moritz, 2002; Ferrière et al., 2004). In accordance with this motivation, these studies tend to focus on small regions of land, making them more directly comparable to the research presented here. At a global scale, certain regions simply lack the genetic background for a given trait to have a chance of evolving (e.g. a novel heat tolerance mechanism seems unlikely to evolve in the arctic). Our question centers on more localized regions that are spatially heterogeneous with regard to the extent to which various related traits are useful. The results from conservation biology suggest that variation in evolutionary

potential on such a local scale is plausible; even within relatively small regions, these studies have demonstrated substantial spatial variation in phylogenetic diversity (Forest et al., 2007; Vandergast et al., 2008).

Most of the processes that are hypothesized to be involved in the formation of hotspots of phylogenetic diversity are based on insights from simpler systems, most notably island groups. Islands represent clearly delineated spatial patterns and processes, making them an excellent context in which to build up insights about how eco-evolutionary dynamics will play out across space. For example, islands are known to promote diversification via adaptive radiation (Losos, 2010), and the formation of new islands within a system has substantial impacts on the assembly of ecological communities over evolutionary time (Gillespie, 2004). Island systems are also a particularly pronounced example of variation in spatial structure. Spatial population structure has a profound effect on how evolution will proceed; more structured populations (i.e. those with more constraints on where offspring end up relative to their parents) are generally more diverse and more likely to reach the highest peak in a fitness landscape (Tomassini, 2005; Nahum et al., 2015). The theory of island biogeography makes various shorter timescale predictions (MacArthur and Wilson, 1967). More recently, biologists have begun to translate theory developed for island systems to more continuous land masses, often by treating regions of a species' preferred habitat as "islands" (also called patches) within a matrix of less suitable habitat (Forman, 1995). Many have argued this approach is likely an oversimplification (McGarigal et al., 2009; Franklin and Lindenmayer, 2009). A better understanding of the relationship between the dynamics observed in island systems and the dynamics that occur in more continuously varying environments will likely be important to understanding the spatial evolutionary dynamics that create evolutionary hotspots.

4.3 Methods

4.3.1 Experimental design

I arbitrarily selected eight of the 500 randomized environments from 3 for further study without examining them beforehand. In each environment, I performed 100 replicate runs of Avida. Each replicate was run for 100,000 updates (approximately 2,000 generations). I also performed 100 replicate runs of a homogeneous control condition where all resources were globally available. From these results, I extracted the coordinates of the grid cell in which each trait first appeared within each run. While a trait may evolve independently multiple times within a run of Avida, including these subsequent evolutions would introduce various biases into the dataset (especially when traits were lost and then re-evolved). Thus, I consider only the first organism ever to display a given trait within a replicate, before selection has an opportunity to play a role.

These extracted (x,y) coordinates translate into 81 separate point patterns (one for each of the 9 traits across each of the 8 environments plus the control) (see Figure 4.1A): one for each of the nine different traits in each of the eight environments. Each pattern contains a maximum of 100 points. Many patterns contain fewer points, as it is not guaranteed that every trait will evolve in every replicate. Because each replicate is completely independent of each other replicate, the only relationship between points in the same pattern is that they evolved in the same environment.

4.3.2 Statistical Approach

To determine whether or not different regions have different evolutionary potential, I need to determine whether the point pattern is significantly different from complete spatial randomness. To make it possible to appropriately correct our statistics for multiple comparisons, I performed the test for spatial randomness in two steps: 1) a test to see whether each pattern, as a whole, was random, and 2) a follow-up test to determine which regions of
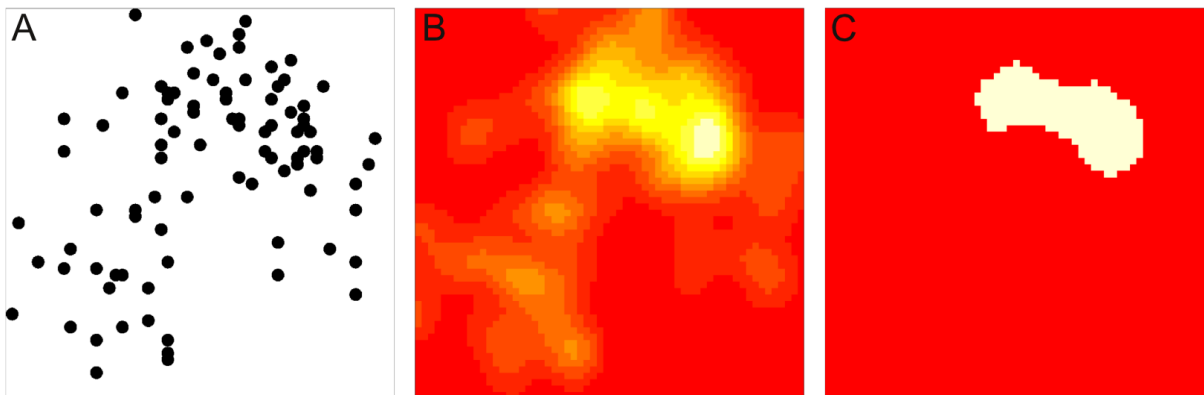
Figure 4.1: **Hotspot location technique.** A) An example point pattern, generated from 100 runs of Avida within a single environment. This pattern corresponds to the most complex trait, EQU. Each point represents the location in space of the first organism within one of the 100 runs that was able to perform EQU. B) Kernel intensity map generated from the point pattern in part A. C) Regions within that kernel intensity map that are substantially more intense than observed in the Monte Carlo simulations (i.e. the hotspot).

non-random patterns had significantly more points than expected by chance.

For the overall randomness test, I calculated a test statistic for various distances, r. For each point in the pattern, I counted the number of other points falling within r distance units of it. I then averaged this number across all points to get our test statistic for that value of r. For each point pattern, I generated a range of expected results under the null hypothesis of complete spatial randomness by running 100 Monte Carlo simulations where I randomly placed the number of points in the actual data across an area of the same size. If the value of the test statistic calculated from our observed data was greater than the highest value from the Monte Carlo simulations for more than 5% of the r values, I concluded that the observed pattern was substantially more clumped than I would expect to see by chance (see Figure 4.2). A test statistic below the range of those observed for the Monte Carlo simulations would suggest that the observed data was more uniform than I would expect by chance; such a result would be unexpected for our experiments. Note for those familiar with spatial statistics: this statistical test is functionally equivalent to Ripley's K function (Ripley, 1981). However, since Ripley's K function assumes first-order spatial stationarity
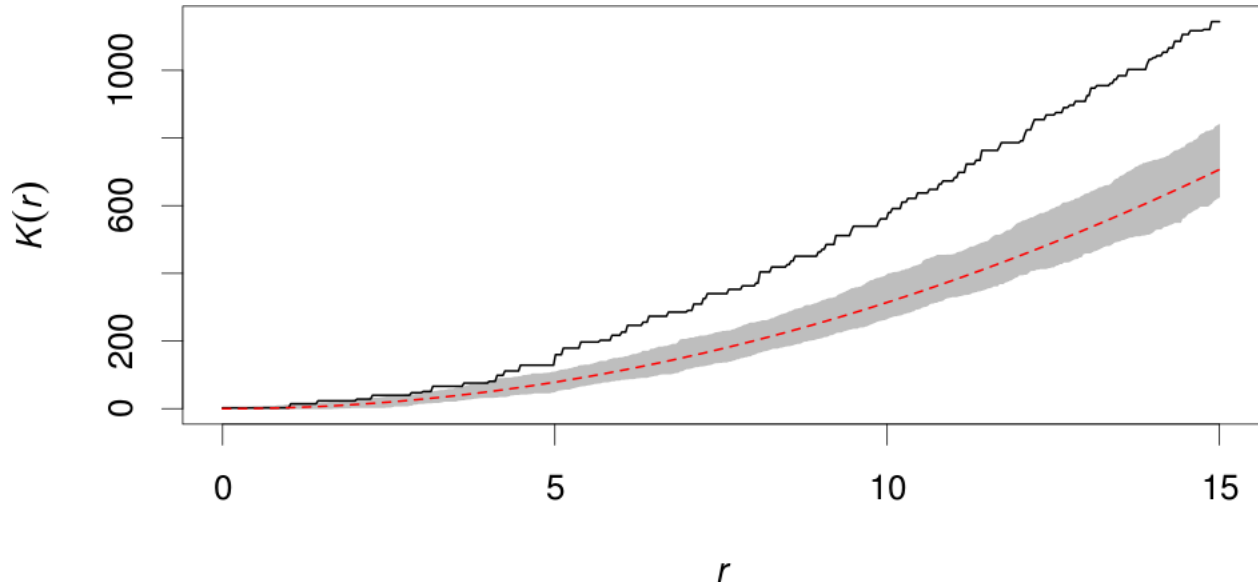
Figure 4.2: **Results of performing the test for spatial clumping on the point pattern from Figure 4.1A**. Observed values of the test statistic (black line) are well above the values observed in the Monte Carlo simulations (shaded area), indicating a clumping in this pattern. The red line indicates the theoretically derived expected value for a random pattern.

.

and the data is exclusively the result of first-order trends, I do not expect to see the same patterns traditionally observed in Ripley's K function.

For those point patterns that were significantly more clumped than I would expect by chance, I performed a follow-up test to determine the location of the hot spot. I calculated a kernel intensity surface (essentially a heat map of point density across the environment) for each of these point patterns using the spatstat library (Baddeley et al., 2015) (see Figure 4.1B). Then, to determine whether the patterns within the data were significantly more intense than I expect from a random process, I performed a Monte Carlo hypothesis test with 100,000 simulations. I simulated patterns generated under the assumption of complete spatial randomness (our null hypothesis) by randomly selecting points from a uniform 60x60 lattice grid approximating the lattice of cells in Avida. For each point pattern, I selected the number of points observed in the original pattern. Kernel intensity values were created for each of these simulations. For each cell in the grid, I compared the experimentally-derived
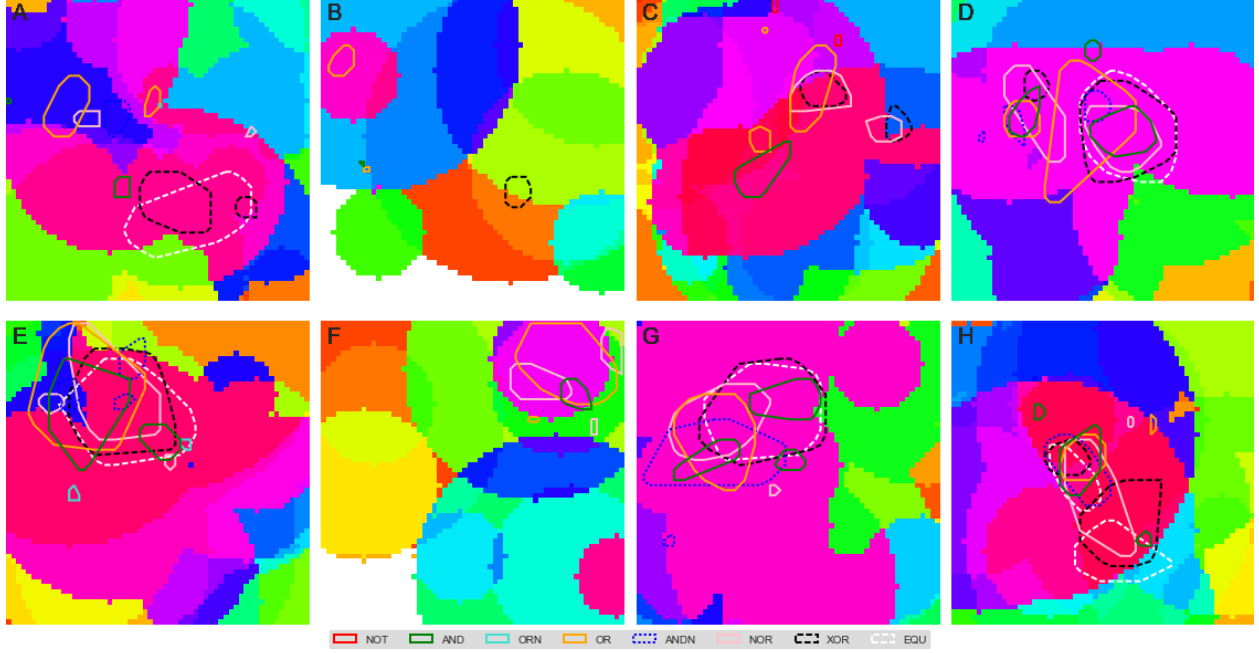
Figure 4.3: **Hotspots across all traits and all environments**. Each panel represents a different environment. Background colors represent the different combinations of resources that are present in each location; each color represents a different combination, with colors assigned such that higher hue values roughly correspond to more complex environments (either in number of resources or complexity of traits associated with resources). Each polygon color represents the outline of hotspots for the trait specified in the legend. Note that only hotspots significant at an alpha-level of .00001 are shown. This conservative value was necessary to be certain that hotspots are present despite the large number of hypothesis tests performed. However, in practice, it likely translates to an underestimation of the true size of many of the hotspots.

intensity to the distribution of simulated intensity values. Cells in which our observed value was higher than 99.999% of the simulated values were labeled as hotspots of evolutionary potential for a given trait. This significance threshold ($p < .00001$) was selected to ensure all results are significant after a sequential Bonferroni correction for multiple comparisons across all 3600 grid cells.

4.3.3 Code and data availability

Scripts used to generate environment files are available via Zenodo with DOI 10.5281/zenodo.162981. Statistical analyses were performed using the spatstat and qqtest

R libraries (Baddeley et al., 2015; Oldford, 2016). Data and scripts used in this project are available at https://github.com/emilydolson/evo_in_space.

4.4 Results and Discussion

4.4.1 Existence of evolutionary hotspots

Out of the 81 point patterns, 43 differed significantly from random. All 43 of these non-random patterns showed significant clumping of points. For 40 of these 43 non-random patterns, the kernel density map had at least one region of higher intensity than I would expect to see by chance (hotspot) (see Figure 4.3). It is possible that the three patterns without hotspots instead had areas of significantly decreased evolutionary potential. Alternatively, these patterns may contain hotspots that were too weak to be detected by the highly conservative threshold for significance. Every test environment had hotspots for at least two traits, while none of the point patterns from the spatially homogeneous control displayed significant clumping, suggesting that these hotspots are a product of the spatially heterogeneous environment. These results provide clear evidence that hotspots of evolutionary potential not only exist but are relatively common; some regions of the environment promote the evolution of new traits to a greater extent than others.

In most environments, there appears to be strong overlap in hotspots for the evolution of different traits (see Figure 4.3). However, there are some environments in which this is not the case, and even the environments with the neatest overlap have some traits with non-overlapping hotspots. Some amount of overlap is to be expected, for three reasons: 1) If any trait is easier to evolve in a given region, then lineages living in that region are more likely to be successful as a result of having evolved the trait, giving them an older evolutionary history, 2) many traits serve as building blocks for other traits, so possessing some traits is an indicator that a lineage is in a part of the fitness landscape from which other traits are easier to reach, and 3) some traits are close to each other within the fitness landscape

75

(for example, the two most complex traits, XOR and EQU, which often have hotspots in nearly identical locations). The fact that the overlap is not universal suggests that the local environment has effects on the likelihood of evolving a trait that go above and beyond the ability of lineages to perform other traits.

4.4.2 Potential drivers of hotspots

Now that I have established that a spatial pattern exists in the locations where traits first evolve, the obvious next question is: what causes this pattern? Since all traits have potential to serve as building blocks for each other, the simplest explanation would be that locations with more resources have higher evolutionary potential. Note that because I am considering only the first mutation to confer each trait, the resource for that trait cannot exert selective pressure.

To test this explanation, I performed a multi-variate linear regression attempting to predict the value of each position in the kernel intensity map from a series of boolean variables indicating which resources were present in a cell. I tried such regressions on various subsets of the data: a single trait within a single environment, a single trait across environments, and all traits across all environments. Models built from data on a single trait in a single environment varied wildly in the percent of variation that they explained, but many explained a high percentage (R^2 values ranged from .11 to .95). However, little (if any) of this explanatory power generalizes across environments. Resources that increase evolutionary potential for a given trait in one environment (as indicated by a regression coefficient significantly greater than 0) may decrease it in a different environment (as indication by a regression coefficient significantly less than 0). Linear models calculated across environments were unable to reconcile these differing patterns, explaining minimal variation in evolutionary potential (R^2 values ranged from .03 to .13). A similar set of models in which the predictor variables were the distance to each resource was similarly ineffective. While high degrees of spatial autocorrelation in the kernel intensity maps mean that the assumption of independence is

violated for these linear models, such a reduction in degrees of freedom should only inflate the apparent variation explained. The failure of this simple mechanism to explain our observed results suggests that more complex processes are driving these effects. Note: Interestingly, there does appear to be a correlation between hotspot location and the pink background color in Figure 4.3. This suggests that the color selection algorithm may be making useful generalizations that are worth further exploration. Nevertheless, it does not appear to be a complete explanation (not all pink regions are hotspots and not all hotspots are in pink regions) and further exploration will be necessary to determine what factors are driving the correlation.

Many hotspots appear to overlap with borders between different combinations of resources. Such regions likely have elevated local diversity, as different phenotypes will be better adapted to the different regions. As discussed above (see Background), there is reason to believe that diversity leads to increased evolutionary potential. Thus, variation in local diversity is another possible driver of evolutionary hotspots. To test this hypothesis, I calculated local diversity across the environment for time points immediately before a novel trait first appeared. Specifically, I measured local diversity as the Shannon entropy of phenotypes within the 25 cells in the 5x5 patch around a given focal cell. I performed this calculation for each cell in the environment at a given time point, producing a map of local diversities. Within each of these maps, I calculated a percentile for each of the 3600 local diversity values, based on the proportion of sites that had a lower diversity.

Next, I located the cell in which a novel trait was about to evolve and recorded its percentile. If diversity has no impact on the probability of a new trait evolving, I would expect all percentile values to be equally likely (i.e. follow a uniform distribution). If, on the other hand, new traits are more likely to evolve in areas with elevated local diversity, I would expect high percentiles to be overrepresented.

To test this hypothesis, I compared the distribution of observed percentiles to a uniform distribution. For the simplest traits, low percentiles were overrepresented due to the presence
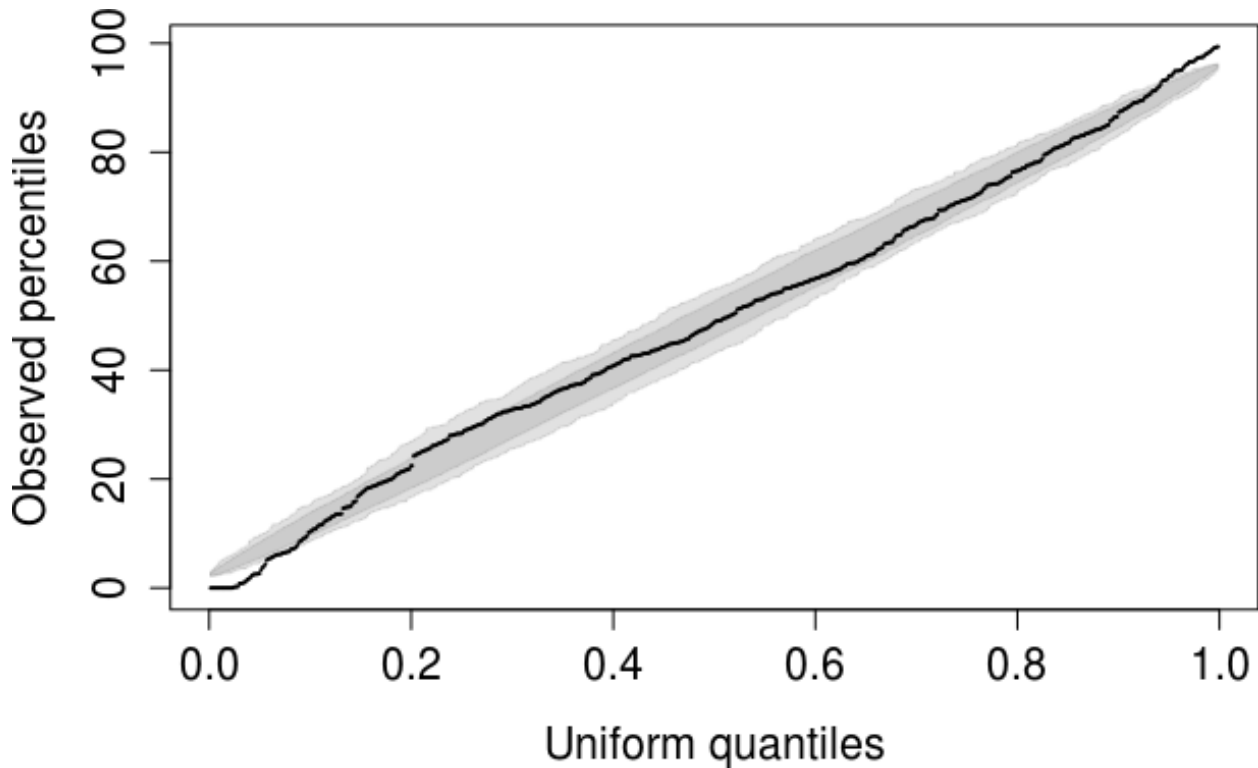
Figure 4.4: **Quantile-quantile plot comparing the distribution observed diversity percentiles for the most complex trait (EQU) to a uniform distribution** (Oldford, 2016). The shaded area represents the 95% confidence interval for where I would expect this line to fall if these distributions were identical. Since the line is mostly within this area, I conclude that the distribution of ranks does not deviate substantially from a uniform distribution. This finding suggests that there is nothing special about the level of diversity in the regions where traits are more likely to evolve. Results for other complex traits are similar.

of populations that contained large swaths of area populated by minimally evolved organisms with none of the traits. However, among the five most complex traits, I found no substantial deviation from a uniform distribution (see Figure 4.4). This result suggests that traits did not evolve more frequently in more diverse regions. While it is possible that this lack of an effect is the result of the specific diversity metric used, it strongly suggests that local diversity is not an important driver of hotspot location. This result may also cast doubt on the idea that hotspots of diversity are hotspots of all evolutionary processes, but I would need to examine phylogenetic diversity to fully address that point.

A third potential driver of evolutionary hotspots is the temporal series of local conditions

that lineages experience over evolutionary time as each offspring is born in a slightly different location than its parent. Perhaps some sequences of environments are more conducive to the evolution of certain traits than others. To explore this idea, I located the first organism with each trait across all replicates and traced the spatial path of their lineages. I then overlaid these paths on the corresponding environment (see Figure 4.5). While a thorough statistical analysis of the variation in paths is beyond the scope of this chapter, I do see some qualitative patterns. Indeed, some traits display consistent patterns that are not seen in paths generated from the control data. These patterns include the middle portions of paths, indicating that not just the ending conditions are likely to matter. For instance, in the three environments in which most of the hotspots are overlapping and in the center (D, E, and G in Figure 4.5), it is rare for any lineage that has ever gone through the lower right corner to be the first one to evolve the most complex task (EQU). This pattern is almost certainly related to the fact that all replicates are seeded with a single organism in the upper left corner; in these environments, successful lineages go directly to the hotspots and don't leave. In other environments, however, successful lineages appear to travel through different sequences of regions (see Figure 4.5F, as an example). While more research is required to understand the mechanisms behind such patterns, this effect is the most promising explanation for the existence of evolutionary hotspots.

4.5 Conclusions

I have demonstrated the existence of spatial variation in evolutionary potential across a heterogeneous environment. Research on what specific aspects of a local region are responsible for elevated evolutionary potential is ongoing. The failure of obvious explanations, such as resource presence/absence and local diversity, to predict the location of hotspots suggests that the actual drivers are complex and nuanced. Preliminary investigations of the sequence of locations that a lineage passes through suggest that this avenue is promising for further

exploration; there appear to be broad commonalities among paths of lineages that evolved the same trait in the same environment.

Understanding the drivers of spatial variation in evolutionary potential would give us a valuable tool for controlling evolutionary trajectories. The most immediate applications are in evolutionary computation. Many evolutionary algorithms vary the fitness landscape over time, starting with easy problems and working up to more challenging problems (Hornby, 2006; Ovaska et al., 2009). This approach often requires hand-tuning of the different problems and when they appear in the environment. Varying the fitness landscape across space instead could achieve the same outcome. If we can purposefully build spatial networks of fitness landscapes to promote the evolution of a solution to the problem, that would make such an approach even more powerful.

In the longer term, once our findings have been replicated in a biological system, I hope to use these same concepts to promote the evolution of biological traits of interest. For example, we may be able to engineer habitats that increase the chances of species adapting to climate change through evolutionary rescue. Conversely, we may also be able to arrange environments to inhibit the evolution of undesirable traits, such as antibiotic resistance in bacteria. Overall, this line of research presents a variety of promising applications.
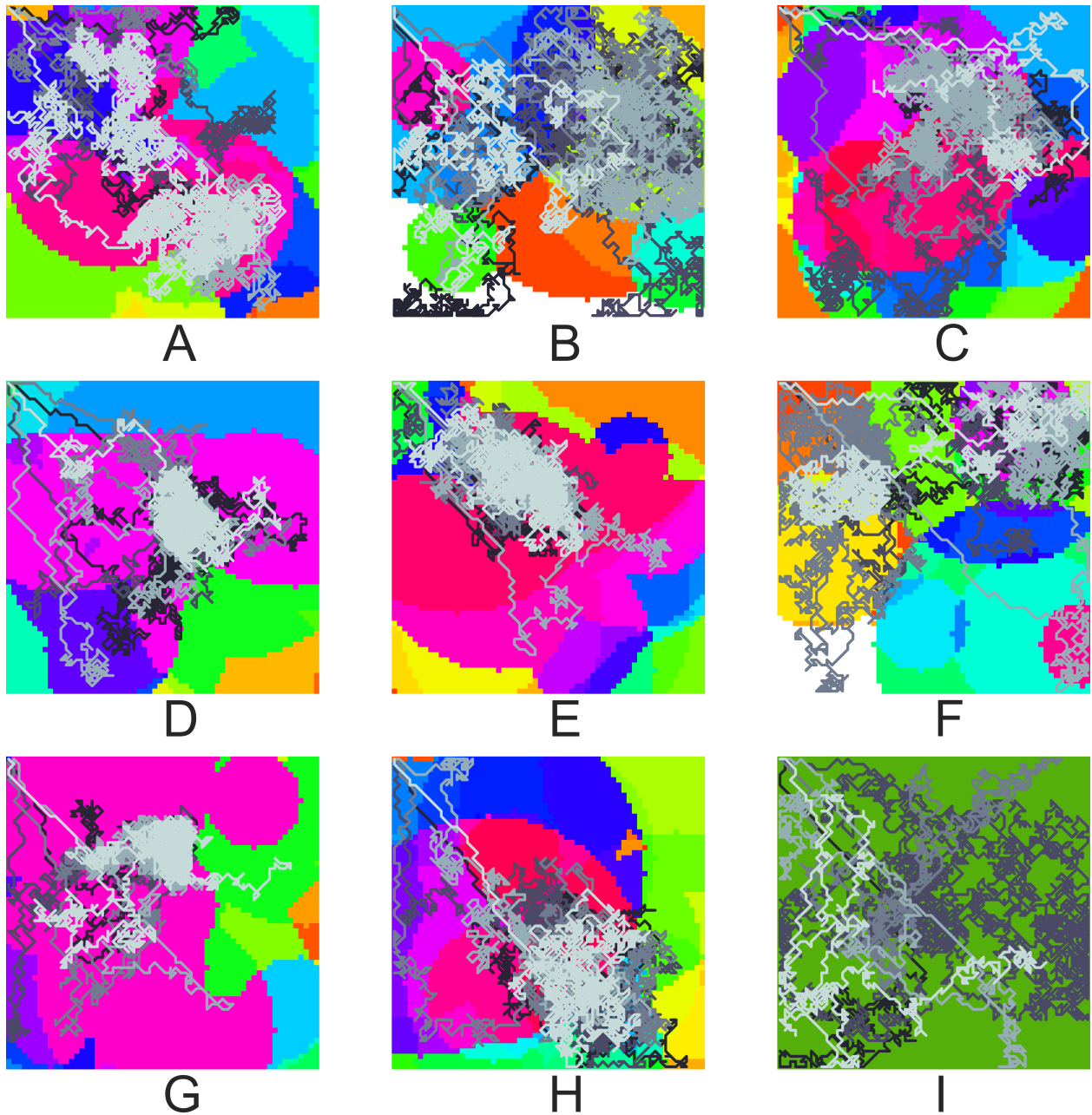
Figure 4.5: **Spatial paths taken by the first lineages to evolve the most complex trait (EQU) from five arbitrarily-selected replicates from each environment**. The path from each replicate is colored a different shade of gray. As in Figure 4.3, background colors represent the set of resources in each part of the environment. I) shows the homogeneous control environment.

Chapter 5

Ecological Theory Provides Insights into Evolutionary Computation

This chapter is adapted from a pre-print on *PeerJ Computer Science* (Dolson et al., 2018d).

5.1 Introduction

Evolution and ecology are fully entwined in nature (Kokko et al., 2017). As such, evolutionary theory and ecological theory need to build upon each other to realize their full potential. Here, I argue that ecological theory is similarly important in the context of the theory of evolutionary algorithms. We can better understand much of what happens in evolutionary algorithms if we reformulate our analyses with ecological theory in mind. In particular, concepts such as diversity maintenance, species co-existence, niches, and spatial effects are suddenly at the forefront of concerns in such a perspective. While the No-Free-Lunch theorem states that there is no single best algorithm for solving all possible optimization problems (Wolpert and Macready, 1997), progress is possible because real-world problems have patterns that can be exploited, which tend to cluster into groups with similar properties. Algorithm can solve problems from across these classes by simultaneously exploring qual-

itatively different paths through the solution space. Such a technique effectively produces multiple "species" of solutions that coexist in an ecological community. Thus, by leveraging ecological theory, we should be able to use first principles to design evolutionary algorithms that take maximal advantage of this property and mitigate the practical ramifications of the No-Free-Lunch theorem.

There are two main sets of tools from ecology that I expect can be helpful for analyzing evolving EC communities: mathematical theory and empirical techniques to evaluate that theory. Ecologists use mathematical theory to make *a priori* predictions about the fate of natural communities. However, translating these predictions to work with EC systems requires careful attention to implicit assumptions about nature that may or may not carry over. For example, ecologists can safely assume that organisms inhabit a finite three-dimensional Euclidean space. A more substantial problem is that most ecological theory assumes that evolution is too slow to be relevant and attempting to introduce evolution can create messy feedback loops (although eco-evolutionary dynamics research is bridging this gap (Hendry, 2016)). Another important limitation is that equations in ecology generally calculate the average or expected behavior of a system. Such results can be misleading in highly contingent processes, such as evolution, where rare mutations can redirect a population into a new region of the fitness landscape. Nonetheless, ecological theory is a useful starting point for predictions. As such, I will lay groundwork here for using it to guide development of EC systems.

Empirical measurements complement ecological theory, allowing biologists to assess predictions and refine theoretical frameworks. Furthermore, empirical measurements allow us to uncover general patterns even in situations where theory is lacking (Dolson et al., 2018b). To this end, ecologists have developed a toolbox of techniques to evaluate the diversity of a community. Some of these metrics, such as richness (number of unique species) and Shannon diversity (entropy) are already used in EC. Other valuable metrics have not yet been adopted in EC. Phylogenetic diversity (Winter et al., 2013), for example, can assess the extent to

which an algorithm is maintaining independent subpopulations that are exploring distinct regions of the fitness landscape vs. organisms that only recently diverged from a common ancestor (Dolson et al., 2018b). This distinction can impact how useful the population diversity is likely to be for adaptive evolution. We can also use empirical ecological measurements to assess hypothesized mechanisms for different diversity maintenance techniques. For example, ecologists often build graphs representing the pairwise interactions between community members; the topology of these graphs signals how the community is likely to change over time (Fontaine et al., 2011). We can do the same for interacting communities in EC.

In the rest of this Introduction, I will provide additional background on ecological theory and the selection schemes that I will be examining. In the rest of this contribution, I will examine four illustrative selection schemes in the light of ecological theory. Our goal is to develop unifying principles for understanding the ecological dynamics inherent in an EC system and predicting its resulting adaptive potential. I intend for this theoretical framework to provide powerful insights about the dynamics in EC systems and lay the groundwork for easier application of ecological ideas in the future.

Box 1: Glossary

Biological fitness (W): An individual's expected number of offspring.

Equalizing forces: Forces that increase the similarity of fitnesses in a population.

Interaction network: A weighted directed graph describing interactions between members of a population.

Phylogenetic diversity: The amount of diversity in ancestry represented within a population.

Phylogeny: The graph of parent-child relationships in a population.

R^*: The lowest quantity of a limiting resource for which a given species' population growth rate is positive. In an EC context, can be thought of as the lowest quantity of a

---

limiting resource for which attempting to use that resource increases fitness.

Stable coexistence: A scenario in which a group of species are expected to coexist indefinitely.

Stabilizing forces: Forces that increase the fitness of rare species.

5.1.1 Ecological theory

One way that ecological theory can inform EC is to identify conditions under which different types of organisms can or cannot **stably coexist**. The ability to predict coexistence dynamics will inform us about types of lineages that can simultaneously explore the fitness landscape. As such, determining coexistence criteria is a primary focus of this chapter.

Early theory on coexistence in ecology focused on competition for resources that are both *limited* in quantity and *limiting* of the growth rate of species that rely on them. The most important value for determining coexistence in this context is a species' $\boldsymbol{R^*}$, the resource availability level at which the species' population growth rate is 0. If the current resource availability is less than a species' R^*, that population will shrink, reducing the utilization of the resource and (in the absence of other species) increasing its availability. Conversely, if the current resource availability is greater than the R^* for a species, its population will grow. Assuming no other ecological factors are affecting a species abundance, its population size will stabilize with a resource abundance equal to R^*.

In the simplest case of species competing for a single resource, the species with the lowest R^* should out-compete the others (Grover, 1997). That species' population will continue to increase, depleting the resource until it reaches the species' R^*; meanwhile, the resource availability will dip below the other species' R^*, causing their populations to decrease until they finally go extinct, having been out-competed. Adding additional resource types introduces the potential for stable coexistence among multiple species if each species is a better competitor for a different resource and consumes more of that resource (Chase

and Leibold, 2003) (summarized in (Letten et al., 2017)). Note that I use the word "species" here to be consistent with ecology; the same insights apply to other taxonomic units, such as phenotypes or genotypes. In most cases, these are more appropriate taxonomic units in the context of EC.

This resource-mediated coexistence effect is an instance of a broader rule: species can coexist if individuals of each species compete with each other more than they compete with individuals of other species (Chesson, 2000). This rule works because it forces species to be self-limiting, creating negative frequency dependent dynamics where each additional member of a species reduces the competitive ability of other individuals of the same species (Adler et al., 2007). The magnitude of difference between interspecific (between species) and intraspecific (within species) competition that is required to enable long-term stable coexistence is determined by the difference in the fitness of the two species. If one species is dramatically more fit, it can drive the other species to extinction even with limited competition (Chesson, 2000). Ecologists draw a distinction between "**stabilizing**" dynamics, which alter the ratio of interspecific competition and intraspecific competition, and "**equalizing**" dynamics which alter the difference in fitness between two species (Adler et al., 2007). In the absence of any stabilizing dynamics, equalizing dynamics will lead to an unstable equilibrium; even if two species have identical fitness, one should eventually drift to extinction. Stabilizing dynamics, on the other hand, actively correct any deviation from equilibrium. Note that in biology, "**fitness**" is strictly a measurement of reproductive output. This definition is in contrast to the externally-defined "fitness functions" used in EC to determine competitive advantage.

In more heterogeneous environments, there is another mechanism for species coexistence: spatial segregation. Such segregation can arise due to a variety of factors. The simplest of these factors is the presence of physical barriers that inhibit movement between regions. In biology, physical barriers play an important role in facilitating diversification (Faith, 1992; Dolson et al., 2016b), and these benefits carry over to EC in the context of island

models (Tomassini, 2005). Spatial segregation can also be brought about by environmental conditions. For example, one region might require tolerance to extreme heat, while another might require tolerance to acidic soil. In this case, species would only compete when both could inhabit the same regions. The resulting traversal (over evolutionary time) of multiple environments has a profound effect on the traversal of the fitness landscape (Dolson and Ofria, 2017). An even weaker form of spatial segregation occurs when species are capable of surviving in range of environments but are best optimized to a specific region (Dolson et al., 2017).

What do these coexistence dynamics mean for EC? First, when choosing parameters, we should consider the circumstances under which they promote coexistence. When two independent lineages are traversing the fitness landscape, what genotypic or phenotypic differences are required for them to both stably persist? Second, when designing selection schemes, we should determine how many lineages we want to co-exist and how frequently we want lineages to turn over. If it is costly to stochastically lose established lineages, we should consider including stabilizing dynamics in addition to or in place of equalizing dynamics. Finally, we should give careful thought to the metaphors that we use to compare EC to biological populations; determining whether a given selection scheme is more akin to competition for resources, spatially segregated habitats, or something else all-together will make it much easier to draw parallels and gain a deeper understanding. If we can show that a system in EC is isomorphic to a system in biology, we can rapidly import the vast wealth of insights from biological research into how that system behaves.

5.1.2 Empirical ecological techniques

Empirical ecologists seek to find general patterns in complex, messy data. While data from EC may be less noisy than field data, the interconnections within any ecological community are still complex. Ecologists have developed many tools for extracting meaning from this complexity, which may be equally helpful in trying to understand the fine-scale dynam-
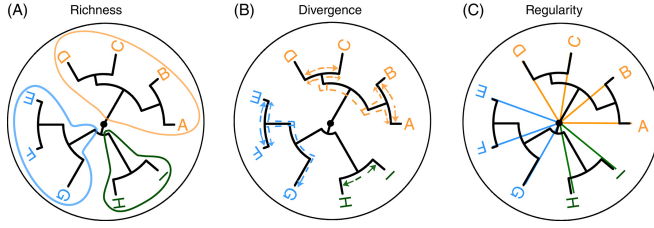
Figure 5.1: **Conceptual illustration of phylogenetic diversity metrics.** The letters around the outside represent the full set of extant taxa and the circle in the middle indicates their most recent common ancestor. Branching points indicate the locations of intermediate taxa (note that in biology these have to be inferred, but in EC we have perfect information). The three panels indicate the three different facets of phylogenetic diversity: richness, divergence, and regularity. Reproduced from (Tucker et al., 2017).

ics of EC systems. Here I focus on two approaches: phylogenetic analysis and interaction networks.

5.1.2.1 Phylogenetic analysis

Phylogenetic analysis refers to a suite of metrics that are used to quantify the topology of a populations' **phylogeny** (ancestry tree) (Winter et al., 2013). In biology, these trees generally need to be inferred from extant species, but in EC we can record the actual tree as it forms. Many of these methods measure the amount of evolutionary history discernible in a population. For example, evolving sub-populations that have stably coexisted for a long time will reflect deeper evolutionary history than a single population that is all descended from a recent common ancestor. Even if the latter population contains many unique phenotypes, they will all be relatively close to each other on the fitness landscape. As such, the population is likely exploring only one basin of attraction within the fitness landscape at a time, limiting the rate of adaptation. **Phylogenetic diversity** provides more direct evidence about the efficacy of diversity maintenance techniques than the more commonly used measures of genotypic and phenotypic diversity.

Techniques for analyzing the topology of phylogenies can be split into three categories (Tucker et al., 2017) (see Figure 5.1 for an illustration of each):

1. Richness: the total quantity of evolutionary history represented

2. Divergence: how spread out the population is in phylogenetic space

3. Regularity: how evenly the population is divided across evolutionary space

Here, I focus on richness and divergence, as our goal is to analyze selection schemes, which regularity does not offer clear insight into. A number of metrics quantify phylogenetic richness and phylogenetic divergence (summarized in (Winter et al., 2013; Tucker et al., 2017)). Here, I measure phylogenetic richness with the original phylogenetic diversity metric (simply referred to as "phylogenetic diversity" (Faith, 1992)), as the count of nodes in the minimal spanning tree connecting each extant genotype to the most recent common ancestor of all extant genotypes. I measure divergence as the mean pairwise distance between genotypes in the population (Webb and Losos, 2000).

Note that in most cases biologists implicitly assume that trees are rooted at the most recent common ancestor of all taxonomic units being compared. This assumption is unavoidable in the context of biology, because phylogeny reconstruction techniques cannot make inferences about anything preceding the most recent common ancestor. In EC, we have the full history. However, including it would not add additional information; phylogenetic diversity would increase by a constant for each member of the population and mean pairwise distance would not change. Note that in this chapter I calculate these metrics on a per-genotype basis, but they can also be calculated per-individual or per-phenotype.

5.1.2.2 Interaction networks

Since ecological communities are collections of interacting organisms, drawing a graph representing the network of pairwise interactions is a useful technique for understanding them (Fontaine et al., 2011). The topology of this graph illustrates the selective pressures that organisms place on each other (see Figure 5.2). Ecologists often generate these networks from knowledge of the species involved such as the fact that two species use the same limited resource and thus must compete with each other. Of course, in EC we can directly measure

the fitness of each member of the population with and without each other member present to identify interactions.

5.2   Ecology in EC systems

In this section, I discuss how each of my chosen EC selection schemes can be intuitively mapped onto ecology. I believe that these metaphors will facilitate a two-way flow of ideas between fields. With the aid of these metaphors, I consider how to calculate two ecologically-important factors: an individual's expected number of offspring (i.e. biological fitness, W) and the conditions under which stable coexistence is possible. The former is critical for calculating interaction networks within ecological communities (see Section 3.1) and the latter provides insight into not only the amount of diversity that a selection scheme promotes, but how the population is situated across the fitness landscape.

5.2.1   Tournament selection

Individuals in tournament selection have no interaction beyond competition for space in the next population. Since this competition affects all individuals equally, tournament selection approximates a condition where there is no ecological interaction. Nature lacks comparable situations, as ecology is nearly ubiquitous outside of carefully controlled ex-
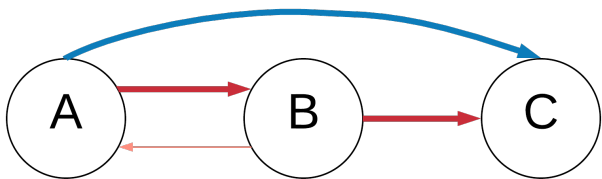
Figure 5.2: **Example interaction network with three nodes.** Red arrows denote harmful interactions and blue arrows denote beneficial interactions. Line width and color darkness indicate magnitude of effect. In this example, A reduces B's fitness a lot, B reduces A's fitness a little, and B reduces C's fitness a lot. A increases C's fitness a lot. A plausible mechanism for this positive interaction is that, by harming B, A reduces B's harmful effect on C.

perimental environments (although evolutionary theorists do often make this simplifying assumption). The absence of ecology in tournament selection means that stable coexistence is impossible in the long term.

Calculating the biological fitness, W, of an individual in tournament selection requires determining the proportion of the population with a lower fitness score than that individual's, p_{less}. Based on this number, the probability of the focal individual winning a tournament can be calculated. We must also account for ties based on p_{equal}, the proportion of the entire population with a fitness score equal to that of the focal individual (including that individual itself, since I select tournament members with replacement). The fitness of an individual can be calculated with the following equation:

$$S \times \left(\left(p_{\text{equal}} + p_{\text{less}} \right)^T - p_{\text{less}}^T \right) \times \frac{p_{\text{self}}}{p_{\text{equal}}} \tag{5.1}$$

This equation finds the probability of an individual with fitness equal to that of the focal individual winning a tournament by finding the probability of a selecting a tournament that contains no individuals with a higher fitness. By multiplying this value by the probability that the winning individual is the focal individual and the number of selection events per generation (S), the focal individual's expected number of offspring can be calculated. T is the tournament size and p_{self} is the probability of selecting the focal individual from the population (generally $\frac{1}{\text{population size}}$).

5.2.2 Fitness sharing

Fitness sharing operates on the basic ecological assumption that individuals compete more against others that they are more similar to. The most intuitive ecological scenario to compare fitness sharing to is one where a population consumes a single, continuously-varying resource. In cases where the distance function is calculated over a discrete number of dimensions (e.g. Euclidean distance between two vectors of equal length), the resource

can be thought of as varying over the same number of dimensions. An analogous situation in ecology is that of Darwin's finches (Schluter et al., 1985). Many of these finches eat seeds, which vary along dimensions such as size and shape. Each species of finch has a beak that is best adapted to eat seeds near a target size and shape. Over evolutionary time, the finch species have partitioned the space of possible beak morphologies into stable niches specialized on different seed types. In the context of fitness sharing, I expect the population to partition the space of possible phenotypes into stable niches. The theory of limiting similarity suggests that these niches should be somewhat separated from each other in genotypic/phenotypic space (Pacala and Tilman, 1994). Fitness sharing results suggest that these niches are often associated with peaks in the fitness landscape (Goldberg et al., 1987).

Deb and Goldberg have already established the criteria for stable coexistence in fitness sharing (Deb and Goldberg, 1989), which I find to be mathematically identical to Chesson's predictions under modern coexistence theory (Chesson, 2000; Chesson and Kuang, 2008). For the algebra demonstrating this equivalence, see supplemental information. This coexistence criterion can be summarized by the following equation:

$$\rho \leq \frac{k_1}{k_2} \leq \frac{1}{\rho} \tag{5.2}$$

where k_1 and k_2 are the fitness function scores of the genotypes being evaluated and ρ is the niche count between these two genotypes, as calculated by Equation 2.2. I chose these symbols to be consistent with Chesson's framework (Chesson, 2000). Note that in Chesson's framework, ρ generalizes to be the amount of niche overlap between the two individuals. The intensity of stabilizing dynamics (i.e. negative frequency dependence) can be calculated as $1 - \rho$. The smaller ρ is, the greater the difference in fitness values it is capable of stabilizing.

The biological fitness, W, of individuals in fitness sharing can be calculated by modifying fitness scores based on Equation 2.2 and then applying Equation 5.1 based on the adjusted scores.
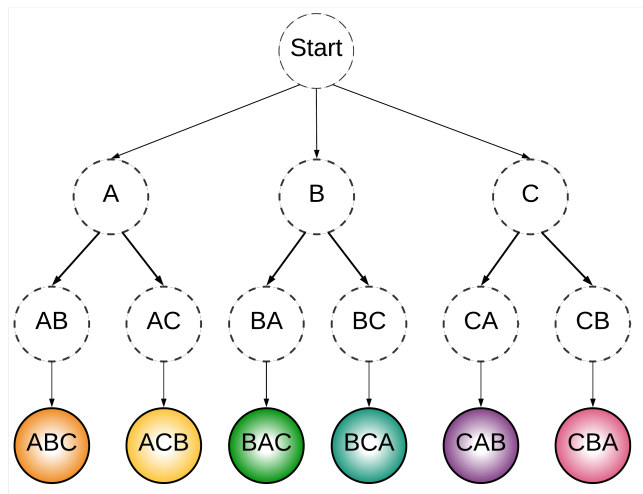
Figure 5.3: **Depiction of a Lexicase decision tree.** A, B, and C are test-cases, and the internal nodes represent the order in which they are picked. Each possible path through the tree leads to a different leaf node ("island"). The individuals on that "island" are the ones that have the potential to be selected for reproduction.

5.2.3 Lexicase selection

In ecological terms, lexicase selection creates a vast number of niches nested within each other. While it is tempting to use a resource metaphor to understand the competition within these niches, the concept of resources does not clearly map onto lexicase selection. Whereas using a resource harms all other individuals that use that resource, improving on a selection criterion in lexicase selection only harms a (usually small) subset of the population. Instead, I argue that population structure is a more apt metaphor.

Imagine $N!$ islands of equal size, where N is the number of selection criteria. Each island corresponds to a single potential ordering of selection criteria and can only be inhabited by the individuals that are best at that ordering (see Figure 5.3). This arrangement is analogous to situations in nature where niches are defined exclusively by an organism's ability to survive a set of harsh abiotic conditions. Being better able to survive these conditions increases the number of offspring an individual can have. Over time, genotypes that are better at surviving in a given set of conditions competitively exclude those that are worse at surviving there. This competitive exclusion happens more rapidly in lexicase selection than it usually would

in biology, but the principle is the same.

While multiple genotypes can theoretically inhabit the same island if they are pheno-typically identical, this coexistence will be unstable due to the lack of stabilizing dynamics to increase the populations of rare genotypes on an island. Eventually, all but one should stochastically go extinct. Furthermore, much as an island in nature can experience a ran-dom catastrophic event, (e.g., a volcanic eruption), the stochastic nature of lexicase selection means that in every generation there is a chance that an island will get unlucky and not be selected. The probability of such an event progressively increases as the proportion of the population living on each island ($\frac{\text{population size}}{N!}$) decreases. When the total population size is less than the number of islands, genotypes must inhabit multiple islands to survive. Thus, stable coexistence in lexicase selection depends on whether a genotype's range (i.e. the proportion of islands it can occupy) is large enough to survive in the long term.

The proportion of islands a genotype can occupy is equivalent to its probability of selection. Just as species in nature face greater risk of stochastic extinction if their range (the area they inhabit) is too small, so too do genotypes in lexicase selection. Details for calculating this quantity are described in equation 4 of (La Cava et al., 2018). To convert the probability of selecting an individual under lexicase selection to biological fitness, it simply needs to be multiplied by the number of selection events that will occur:

$$W = S \times P \tag{5.3}$$

where S is population size (the number of selection events per generation) and P is the proportion of islands occupied. What does this condition mean for stable coexistence in lexicase selection? The chances of a genotype with fitness W surviving for G generations in a population of size S are given by the equation:

$$P(\text{survival}) = (1 - (1 - W)^S)^G \tag{5.4}$$
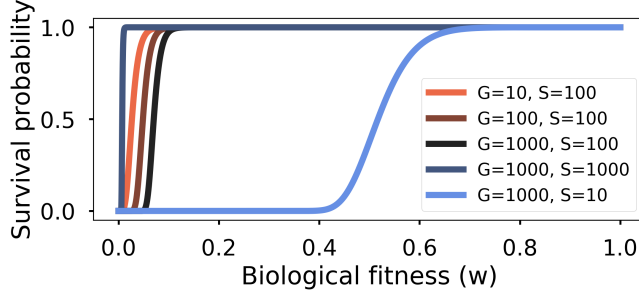
Figure 5.4: **The probability of long-term survival under lexicase selection across different population sizes (S) and lengths of evolutionary time (G).**

Plotting this function for various parameter values, we see that there is generally a cut-off where the probability of survival rises abruptly from approximately 0 to approximately 1 (see Figure 5.4). The value at which the transition occurs is effectively the minimal percentage of islands that a genotype must occupy in order to be expected to survive in the long term. This value is closely related to the concept of a minimum viable population in biology, and is useful to consider when making decisions about how large of a population size to use.

5.2.4 Eco-EA

Eco-EA is analogous to a traditional resource competition scenario in ecology. All individuals occupy the same region of space and compete with each other by targeting the same resources. The threshold amount of a resource in the environment needed by each individual to benefit from using it can be calculated:

$$R^* = \frac{cost}{C_f * score} \tag{5.5}$$

where *cost* is the cost of performing the corresponding task, C_f is the fraction of the available resource consumed, and *score* is how well the individual performed the task (normalized to 1). In terms of resource competition theory, this value is effectively that individual's R^* for that resource.[1] Note that a higher *score* reduces R^*, meaning that the individual can

---

[1]This definition is subtly different from the traditional definition of R^*, which states that R^* is the lowest value of resource for which a species' per capita population growth is not negative (Chase and Leibold, 2003).

95

benefit from the task even if less of the resource is available, making it a better competitor for that resource. Thus, as in natural ecosystems, the individual with the lowest R^* can out-compete individuals with a higher R^*.

If a group of individuals use completely different resource types from each other, there will be a high degree of stabilization, meaning they should coexist under a wide range of fitness differences. The more complex coexistence scenario occurs when two individuals compete for the same resources. Ecologists distinguish resources that can be used interchangeably with each other ("substitutable resources") from resources that cannot be used interchangeably ("essential resources"). Resources in Eco-EA are substitutable, because an equivalent fitness gain could be achieved using either one (up to a point); one just requires more resource to do so. In this scenario, two individuals can stably coexist as long as they have lower R^*s for opposite resources, and each one consumes more of the resource it has a lower R^* for. Because both R^* and the amount of resource consumed are determined by the individual's score on the task associated with the resource, the latter criterion will always be met. Thus, as long as neither individual has the lowest R^* for every resource, coexistence should be possible. In EC terms, this requirement boils down to each individual being Pareto dominant. As before, coexistence also requires that the individuals have somewhat similar fitnesses. The amount of stabilization (and thus the maximum fitness difference) can be calculated based on the resource consumption of each species (Letten et al., 2017).

Biological fitness, W, can be determined in the same way as in fitness sharing: first, calculate the adjusted fitnesses based on resource use, then use Equation 5.1.

5.2.5 Summary

As we would expect, selection schemes that successfully maintain diversity allow for long-term stable coexistence. For fitness sharing and Eco-EA, the criteria are described by Chesson's coexistence theory (Chesson, 2000; Letten et al., 2017); to stably coexist, each

I frame R^* in terms of benefit rather than population growth because population growth in EC is generally restricted by competition for space in the population.

species/genotype/phenotype must limit itself more than it limits others (as formalized in Equation 5.2). In lexicase selection, biological fitness (i.e. the proportion of "islands" the species dominates) must be above a cutoff (described in Figure 5.4).

Although Eco-EA and fitness sharing have the same coexistence criteria, they have an important difference from each other: in Eco-EA, competition occurs along multiple, potentially orthogonal dimensions, whereas in fitness sharing competition is mediated through a single function that summarizes all aspects of an individual. This difference should result in fewer, more intense pairwise competitive interaction in Eco-EA than in fitness sharing (although still not as few, or as intense, as those in lexicase selection). I predict that the more focused interactions in Eco-EA and lexicase selection will promote forms of evolutionary divergence that are more useful for adaptive evolution. The requirement in Eco-EA that coexisting individuals must be non-dominated suggests that individuals in Eco-EA should fall roughly along a pareto front, similarly to individuals in lexicase selection. An important distinction between these two selection schemes, however, is that lexicase selection's rigid population structure produces strong pressure for specialists, whereas generalists can be successful in Eco-EA.

5.3 Empirical Methods

To confirm and extend the intuition I developed in the previous section, I now empirically investigate these selection mechanisms in the context of actual evolving populations.

5.3.1 Interaction networks

In Section 5.2, I predicted the competitive pressures that different selection schemes exert on populations. Here, I assess these predictions by drawing interaction networks for real populations. For simplicity, I use a population containing 10 individuals with 5 integer traits selected from a geometric distribution. Each trait represents a niche that these individuals are

competing to occupy, with higher numbers corresponding to higher competitive ability. For lexicase selection, each trait is a selection criterion and the individual with the highest value there wins. In Eco-EA, I added a resource associated with each trait and an individual's value for that trait defines its ability to use that resource. In fitness sharing, distance is measured as euclidean distance between sets of traits. The fitness landscape is otherwise flat. I compare the same population across all selection schemes. To calculate the effect that a given individual, A has on another individual, B, I first calculate the fitness of B in the presence of the whole population. Then I remove A from the population and recalculate the fitness of B. The difference between these two fitnesses is the effect of A on B. Note that these fitness values are biological fitness, i.e. the expected number of offspring, rather than the fitness produced by the fitness function. For fitness sharing and Eco-EA, I assume a tournament size of 2 when making this calculation.

5.3.2   Phylogenetic analysis of evolved populations

What is the long-term effect of these different interaction network topologies? I arrive at a first-order approximation by analyzing the phylogenies of populations evolved under each selection scheme. Across selection schemes, the behavior of a population depends on the part of the fitness landscape that it is currently exploring. If the entire population is climbing a steep hill, all forms of diversity should be low, due to frequent selective sweeps. To assess the effect of such differences in fitness landscapes, I compare phylogenetic diversity across three different genetic programming problems believed to have qualitatively different fitness land-scapes: squares, Collatz, and the Dow chemical regresssion problem (see chapter 2). Using a simple linear genetic programming representation (ScopeGP, descibed in chapter 2), I evolve linear genetic programs for 1000 generations. I use 11 test cases for the squares problem and 100 test cases each for the Collatz problem and Dow chemical challenge. Phenotypes are the vectors of inverse error for each test case (so higher scores are better). Fitness is calculated as the sum of these vectors. For lexicase selection and Eco-EA, each test case corresponds

to a selection criterion or resource.

I evolved 30 populations for each selection scheme and each problem. Because of the profound effect of the sharing threshold parameter on the behavior of fitness sharing, I also performed 30 runs each of fitness sharing with five different sharing thresholds. Statistics across selection schemes were calculated using the sharing threshold with the highest phylogentic diversity. For each run, I calculated a variety of metrics, including phenotypic diversity (measured as Shannon entropy), genotypic diversity, phylogenetic diversity (Faith, 1992), and mean pairwise distance of genotypes in the phylogeny (Webb and Losos, 2000). Across all runs, I used a genome length of 200 and a population size of 500. I applied mutations to every new individual by randomly changing the code at up to three sites in the genome (the specific number was selected from a uniform distribution). To simplify calculation of the phylogenetic metrics, I seeded each population with a single individual that served as the common ancestor to all others. For Eco-EA, I used a resource inflow rate of 100, a cost of 1, and a consumption fraction for 0.0025.

5.3.3 Code availability

Data for the interaction network analysis was generated using a simulation written with Python 3.6.3. Graphs were visualized using the networkx package (Hagberg et al., 2008). All code for generating and analyzing the data presented in this chapter is open source and available (Dolson, 2018a).

5.4 Results and Discussion

5.4.1 Interaction networks

Different selection schemes create strikingly different interaction networks (see Figure 5.5). Notably, lexicase selection's rigid population structure leads to far fewer interactions

(a) Representative example networks.
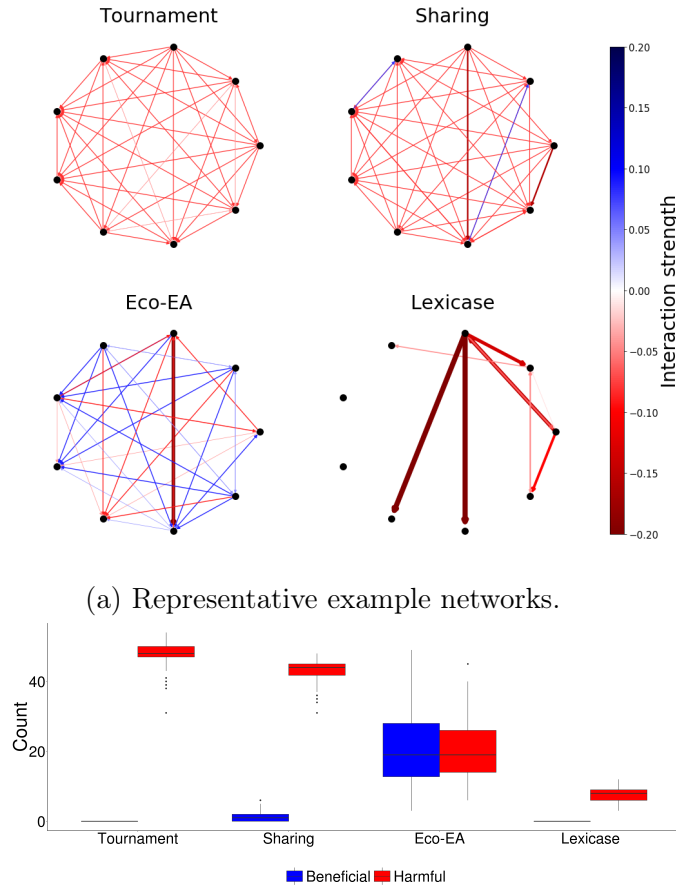
(b) Summary over 100 randomly generated populations.

Figure 5.5: **Interaction networks for the same community under four different selection schemes:** tournament selection, fitness sharing, Eco-EA, and lexicase selection. a) Shows the networks for a single population. Red edges indicate harmful interactions, blue edges indicate beneficial interactions. Edge width denotes interaction strength. An interactive version of this chapter is available (Dolson, 2018a). b) Shows boxplots summarizing the number of beneficial and harmful interactions observed under each selection scheme across 100 randomly generated populations.
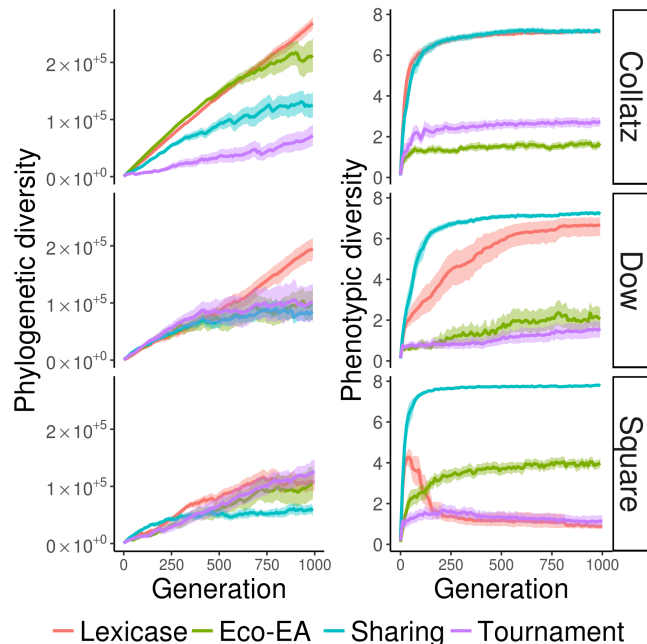
Figure 5.6: **Phenotypic and phylogenetic diversity over time for each problem.** Shaded area is the bootstrapped 95% confidence interval around the mean.

of both types[2] than under other schemes (pairwise Wilcoxon rank-sum test, $p < .0001$), and ones that are predominantly negative. Eco-EA creates far more interactions, because individuals can use different resources to different extents. Importantly, there are many beneficial interactions; if A competes with B and B competes with C, A can help C by suppressing B. These interactions imply a stabilizing dynamic between A and C, because they indicate that A is competing with C less than with itself. Additionally, the Eco-EA community has many weak interactions, which can promote stabilizing dynamics community-wide (McCann, 2000). Lastly, in fitness sharing, most individuals harm each other approximately the same amount. The sharing threshold subtly affects interaction strength.

5.4.2   Phylogenetic analysis

As hypothesized, phylogenetic diversity is low for all selection schemes on the square problem (see Figure 5.6). This result is presumably due to fitness differences so large that

---

[2]With one exception: I observed no beneficial interactions in lexicase or tournament selection, so these groups were not different from each other.

stabilizing effects are insufficient to overcome them. In the context of such a steep evolutionary gradient, such behavior is expected and often desirable. Due to negative frequency dependence, phenotypic diversity remains high for fitness sharing and Eco-EA, even after most populations solved the problem (generally between 200 and 500 generations in). For lexicase selection, on the other hand, it initially increases and then drops rapidly as the populations converge on the solution.

Results from the Collatz problem support our hypothesis that selection schemes with more restrictions on which individuals compete promote phylogenetic diversity (see Figure 5.6). Lexicase selection and Eco-EA did not have significantly different final phylogenetic diversity (Wilcoxon rank-sum test, p=0.31), but all other pairs of selection schemes did (Wilcoxon rank-sum tests, p < .05). Results for mean pairwise distance were similar, suggesting that lexicase selection and Eco-EA (and to a lesser extent fitness sharing) do promote the coexistence of divergent branches.

Interestingly, phenotypic diversity does not correlate especially closely with phylogenetic diversity (see Figure 5.6). In particular, Eco-EA has relatively low phenotypic diversity, despite its high phylogenetic diversity. Conversely, fitness sharing has relatively high phenotypic diversity despite its mid-range phylogenetic diversity. This result suggests that whereas lexicase selection and fitness sharing allow similar phenotypes to coexist, Eco-EA forces them to converge. A potential explanation for this difference is that Eco-EA rewards generalists substantially more than lexicase selection.

This discrepancy is even more pronounced in the context of selecting a sharing threshold for fitness sharing (see Figure 5.7). Phenotypic diversity is maximized at $\sigma_{\text{share}} = 1$ (Wilcoxon rank-sum tests, p < .005), whereas mean pairwise distance and phylogenetic diversity are highest at $\sigma_{\text{share}} = 10$ (Wilcoxon rank-sum tests, p < .005). This result emphasizes the fact that phylogenetic diversity is meaningfully different from phenotypic and genotypic diversity in ways that can affect choice of parameter values.

Results from the Dow problem (see Figure 5.6) illustrate the strong effect that the
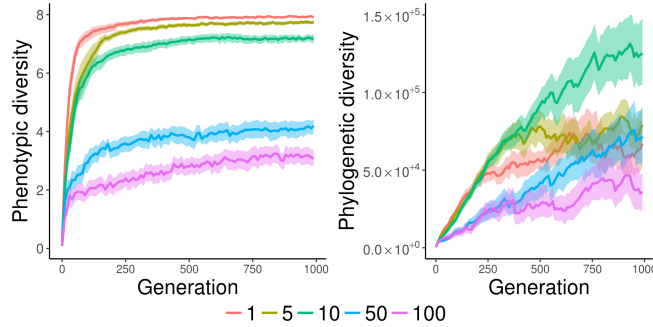
Figure 5.7: **Phenotypic and phylogenetic diversities in fitness sharing** across various values of sharing threshold (σ_{share}) in the context of the Collatz problem. Shaded areas are bootstrapped 95% confidence intervals around the mean.

topology of the fitness landscape has on which techniques are most effective at maintaining phylogenetic diversity. In contrast to its high efficacy on the Collatz problem, Eco-EA maintains no more phylogenetic diversity than fitness sharing and tournament selection on the Dow chemical problem (Wilcoxon rank-sum tests, p = 1 for all). Lexicase selection, on the other hand, maintains phylogenetic diversity on both of these problems (Wilcoxon rank-sum tests, p < .005). Understanding what properties of the fitness landscape account for this difference is an important area of future research.

5.5   Conclusions

In this chapter, I have taken a high-level tour of tools that ecology can contribute to EC. Using metaphors and mathematical theory from ecology, I have identified important differences across four selection schemes. Importantly, I have shown mathematical equivalence between rules governing coexistence in ecology and in EC. This framework should allow us to directly translate insights between these fields.

Building on this conceptual understanding, I empirically measured the interactions that describe the communities created by different selection schemes. The topology of these interaction networks supports our theoretical predictions, and provides insight into the mechanisms by which these techniques maintain diversity. A more systematic study of interaction

networks under various selection schemes will further improve our understanding. I gained further insight by exploring the long-term effects of these selection schemes with phylogenetic metrics. Notably, I demonstrated that phylogenetic analysis metrics describe aspects of the underlying dynamics of diversification that are distinct from those described by genotypic and phenotypic diversity. This result holds across different parameter choices for the same selection scheme, and across different selection schemes. Analyzing the phylogenies of populations evolving on a wider range of fitness landscapes has the potential to help predict which parameters and selection schemes are most appropriate for which problems.

From our analysis thus far, I can conclude that lexicase selection and Eco-EA rely on distinct underlying mechanisms to promote the evolution and maintenance of solutions that are evolutionarily divergent. Lexicase selection creates a small number of intense competitive interactions, whereas Eco-EA creates a larger number of weaker interactions that can be either competitive or facilitative. As a result, lexicase selection generates many specialist phenotypes, while Eco-EA supports fewer, more generalist phenotypes. Both of these selection schemes produce populations representing a wider diversity of evolutionary history than those produced by fitness sharing.

Ecological techniques have the power to revolutionize our understanding of diversity in EC, and our work here has barely started. In the future, I plan to delve deeper into these approaches, to import more ideas and statistical techniques from ecology, and to apply them to an even wider range of selection schemes.

Chapter 6

Applying ecological principles to genetic programming

This chapter is adapted from a chapter in the proceedings of *Genetic Programming in Theory and Practice XV* (Dolson et al., 2018a).

6.1   Introduction

Natural evolution produces effective solutions to complex problems, often well beyond the ability of human engineers to duplicate. If we are to harness these natural evolutionary dynamics, we must understand the full depth of how they function and why they are so effective. In this chapter, I explore how ecological factors promote more open-ended evolutionary systems that have a greater potential to produce complex, dynamic, and practical solutions to targeted problems. First I discuss why I believe that this approach can help solve difficult AI problems, then I review the current scientific understanding of ecological dynamics of interest. Next, I present experiments that I have performed using Eco-EA and lexicase selection to test the potential of ecologically-mediated hints, before discussing the implications of these results, and finally laying out the next steps I plan to take in this line of research.

6.1.1 Motivation

Many problems in machine learning center around creating artificial intelligence systems that can resolve challenging problems that are traditionally solved by humans. Often, these programs are written with the goal of mimicking the strategies employed by skilled humans. When human strategies can be clearly articulated as a set of well-defined rules, the process of writing the AI can be straight-forward. However, humans tend to rely on intuition when solving many types of problems. While human intuition is often effective, it is challenging to abstract into an algorithm that an AI can follow without missing important nuances. Attempts to do so often produce rigid AIs that fail to appropriately adapt to situations that are subtly different than those that were originally expected. Such issues are exacerbated for problems where early decisions determine what scenarios the AI will later encounter. For example, in many board games, minor mistakes early on can drastically reduce success and alter what options are available for the middle and end of the game.

This property—whereby early decisions shape what strategies are possible later—can pose even more substantial obstacles to evolving AIs. If a good strategy is critical early on, an AI without one would consistently lose. However if a good early strategy is not sufficient for overall success, the selection pressure in favor of it will be weak (at best) relative to its importance. Specifically, in any problem where many tasks must all be performed reasonably well for fitness to be non-zero, it is nearly impossible for evolution to get enough initial traction to be successful. Taken, from another perspective, the fitness landscape for such a problem will be almost entirely flat, with a small rugged region featuring steep peaks that are challenging to even find, let alone navigate. Previous attempts to make this landscape more easily navigable by giving the evolving AIs "hints" based on strategies successfully employed by humans have generally been ineffective due to the aforementioned difficulties with abstracting human intuition.

I hypothesize that we can apply ecological dynamics to evolve models of human decision making and reliably create human-competitive AIs. Here, I present our concepts in terms

of playing games, as this is an intuitive set of problems to think about, but the techniques I propose should generalize to other categories of problems. The key is to have evolution be responsible for the process of abstracting general strategy from human intuition. I believe that we can achieve this goal by supplying data on human decisions from a wide variety of scenarios and selecting for AIs that are capable of *predicting* the move that a given human made in any given situation. This approach will allow selection to operate evenly across early-, mid-, and late-game strategy, removing the issue of temporally compounded mistakes. Furthermore, it eliminates the need for humans to be able to codify their strategies into a concrete rule set.

Selecting for AIs that can predict human moves has a second, more fundamental benefit. Often, problems that are too complex for evolutionary computation to solve in isolation can be solved if there is also a fitness benefit for solving simpler, related problems. These simpler problems, often referred to as "building blocks", cause genomes within the population to accumulate information that is relatively easy to repurpose into solving the actual problem (Lenski et al., 2003), a technique that has proven to be effective in evolutionary computation (Bongard and Hornby, 2010). I hypothesize that AIs evolving to predict a human's choices will often do so by recreating the underlying building blocks that comprise that human's strategy—even if the human does not recognize it themself.

An obvious problem with this approach is that some humans may be poor at a given problem. For example, predicting their moves in a game may be impossible because they are effectively random, or worse employing actively poor strategy that will lead the evolving population down a maladaptive path. Interestingly, some preliminary work on Eco-EA suggests that it might be robust to bad advice (Goings and Ofria, 2009; Goings, 2010; Goings et al., 2012). Resources associated with useful sub-tasks should produce evolutionary building-blocks that get incorporated as parts of a more complex strategy and used in the overall solution. Unhelpful or misleading sub-tasks, however, should be used by only a small fraction of the population, resulting in a trivial slowdown as compared to running the

107

algorithm without the resources present.

The paradigm of supplying hints to an evolving algorithm without worrying about whether they will be misleading opens up the potential to evolve solutions to highly complex problem. For example, if we wanted to evolve a controller for a robot that would perform search-and-rescue, we would never get there if we started with random controllers and only rewarded robots that successfully rescued people. We could, however, use ecologically-mediated hints by adding a set of limited resources that each rewarded some component, such as (1) moving to a target location, (2) exploring, (3) fully scanning an area, (4) identifying dangers, (5) navigating around obstacles, (6) identifying trapped people, (7) moving toward trapped people, (8) freeing trapped people, (9) moving with people, and (10) finding your way back to safety. With all of these helpers as building blocks, it is easy to imagine that a controller would eventually evolve that could occasionally find and return people safely, which would allow it to start getting rewarded by the primary (unlimited) fitness function. This approach uses ecologically-mediated hints to push the population to solve the problem from multiple directions at once; before the whole problem is solved, some individuals might be able to find people, but then ignore them. Some will explore and then find their own way back. Others might be able to free people they find, but not know what to do with them next. Some evolutionary trajectories might be more likely than others to continue to evolve toward the final goal. Of course, some of the "hints" might be associated with counter-productive tasks. If there were a limited resource associated with avoiding all dangers (because a misguided programmer thought this would be helpful), the result might be to make it less likely for the robot to find trapped people. With ecologically-mediated hints, however, only a small portion of the population (a single niche) would be rewarded for this ability, so others would still plow ahead and complete the mission.

A key component of this approach is that in an ecology, niches drive the types of diverse solutions that appear. If an organism is the first to occupy a new niche, it must have some traits associated with that niche, but is otherwise free from direct competition, allowing

it to sustain substantial loss of unrelated traits. This dynamic allows for many different pathways to coexist in a population, any of which can be followed to solve the high-level problem, with the most successful strategies dominating multiple niches. In essence, these co-existing niches facilitate the creation of a variety of building blocks leading to successively more complex strategies.

Another benefit to an eco-evolutionary approach is that ecological communities are, by definition, at least somewhat diverse. While promoting diversity in evolutionary computation has long been recognized as critical to avoiding the problem of premature convergence, most existing mechanisms to promote diversity in evolutionary computation select for solutions that are distinct from each other, regardless of other qualities (Goldberg et al., 1987). In natural systems, however, diversity arises due to organisms filling niches, each requiring specific phenotypic traits for success. Thus there is pressure for a diversity of functional traits. Furthermore, new niches are continuously created in nature as organisms interact with each other and modify both their physical and social environment. In problem domains such as playing complex games, this diversity of solutions becomes even more important, as there is no single, deterministic, best strategy. Instead, there are various strategies that are effective in different situations and against different opponents, just as there would be in an ecological community.

6.2 Methods

I expect that ecologically-mediated hints provided via Eco-EA and lexicase selection will display similar set of strengths. Both maintain a diverse population with respect to the hints that are provided. However, as discussed above, I expect the use of limited resources to be robust to hints that turn out to be counter-productive, while lexicase selection will have difficulty escaping uncontested bad advice. To assess the accuracy of these hypotheses, I evaluate them in the context of a proof-of-concept problem, the N-dimensional box, where

we can easily manipulate the quantity and quality of the hints.

The N-dimensional box problem is designed to be nearly impossible to solve without hints. As such, we provide hints about the optimal value for each dimension. Since the goal is to minimize all dimensions, a good (i.e. informative) hint would be to minimize an individual dimension. Conversely, maximizing an individual dimension would be a bad (i.e. misleading) hint, leading a population away from the goal. We provided various combinations of good and bad hints to Eco-EA (as resources) and to lexicase selection (as test cases). As an additional control, I ran an equivalent number of trials using unaltered tournament selection.

6.2.1 Configuration details

For all experiments, I evolved a population of 5000 vectors containing 10 floating-point numbers between 0.0 and 1.0 (inclusive) for 50,000 generations. The next population for each generation was chosen using one of the three selection schemes being compared: lexicase selection, Eco-EA, or tournament selection. To avoid giving Eco-EA an unfair advantage due to its use of elitism (in a problem domain where that could only be beneficial), I always preserved a copy of the individual in the population with the highest fitness. For selection schemes requiring a tournament size (tournament selection and Eco-EA), I used a tournament size of two.

I placed the individuals selected by each iteration of the selection scheme into the next generation. Each site in each genome was mutated by adding a value randomly selected from a Gaussian distribution centered at 0 with a standard deviation of .05. Subsequently, I recombined each vector with a random other vector, using one-point crossover.

6.2.2 Statistical methods

To determine the effects of good advice, bad advice, and selection scheme on the probability of solving the problem, I performed a logistic regression. The predictor variables were the number of good hints, the number of bad hints, a boolean indicating whether lexicase

110

selection was used, and a boolean indicating whether Eco-EA was used. Tournament selection, being the control, was the base case to which all other conditions were compared. The regression coefficients for the interactions between good or bad hints and the selection type were used as the test statistic for all statements about the effect of a hint type on a selection scheme.

Since some combinations of variables were able to perfectly separate successes from failures, which can bias the results of logistic regression, I used Firth's bias reduction technique, as implemented in the R package brglm (Kosmidis, 2017).

6.2.3   Code availability

The full source code for the experiments and analysis in this chapter is freely available at `https://github.com/emilydolson/eco-ea-box`.

6.3   Results and Discussion

Both Eco-EA and lexicase selection make effective use of good advice (see Figure 6.1). Eco-EA solved the problem consistently when given at least four good hints. Lexicase selection did even better, solving the problem most of the time when given as few as two good hints. Good hints enable both lexicase selection and Eco-EA to significantly outperform tournament selection (logistic regression, $\beta = 1.3334$ for lexicase, $\beta = 1.8363$ for Eco-EA, $p < .0001$ for both).

These data strongly support our hypothesis that Eco-EA is essentially unaffected by receiving bad hints; the regression coefficient for the interaction between Eco-EA and bad hints is not significantly different from 0 (logistic regression, $\beta = 0.0633$, $p = .66$). Lexicase selection, on the other hand, is harmed dramatically by bad hints (logistic regression, $\beta = -10.6841$, $p < .0001$).

Overall, our results illustrate the power of ecological approaches as a vehicle for providing

hints to evolutionary algorithms. I have provided a proof-of-concept that this technique can make it possible for evolution to solve problems that would otherwise have been out of reach. Moreover, I have clarified the instances in which two specific ecological approaches, Eco-EA and lexicase selection, are most appropriate; lexicase is best when all of the hints are accurate, whereas Eco-EA is robust in scenarios where there may be some misleading hints.

6.4   Conclusions and future work

Thus far, I have tested ecologically-mediated hints on only simple model problems as a proof-of-concept, but I expect this approach to excel on complex problems where fitness functions do not provide a clear path from random starting conditions to meaningful solutions. Artificial Intelligence is a perfect example of this type of problem. Problem-solving strategies may require many different components to coordinate to formulate plans. Board-game-playing agents are a particularly accessible problem that has these properties: they are complex, while still being experimentally tractable, and are often well-studied. They involve clear measures of success, while allowing for multiple co-existing strategies. Most importantly, they are intuitive to humans who can provide "suggestions" for limited resources to produce building blocks for complex strategies.

The next step will be to more thoroughly explore the parameter space in which providing hints via Eco-EA and lexicase selection is effective. In the process, I hope to gain insight that will allow us to develop an approach that includes the best properties of both. Thus far, I have explored the impact of good and bad hints that are orthogonal to each other. However, most hints in real world problems will not be independent, and we expect that this connection may influence the way Eco-EA and lexicase selection respond to them. Similarly, I have not explored the impact of completely neutral hints, which may be harmful if they are present in excessive quantities. That said, a resource is helpful only when it produces a

building block that turns out to be useful for a more complex goal, after which it is no longer needed. As such, many resources would be helpful if they were around just long enough for the associated niche to be filled as a stepping-stone to more complex niches.

If good advice is helpful and other advice is harmless, it should be possible to bootstrap the solving of a problem by generating random hints. Each hint, once used, will remain for a limited amount of time (providing an opportunity to be used as a building-block) before it is removed and replaced by a new hint associated with a new randomly-determined behavior. These random behaviors will usually be harmful and thus ignored, but any that turn out to be helpful building blocks should be incorporated into the successful players.

Finally, I will use human problem-solving patterns to create niches for evolving "crowd-sourced" AIs. While we believe transient resources will have some utility, most of those building blocks are likely to be useless, merely using up computational time. A benefit of working with board games, however, is that I can collect a wealth of information about how human players make decisions across a variety of situations. Rather than reward building blocks that I identify from those logs, I can instead create limited resources that reward AI players for consistently predicting the next move made by a human player. In other words, we do not need to understand why a human player made a decision in order to reward an AI for following the same type of strategy. If that player plays well, the building blocks produced to mimic them should be generally useful, even for some other strategy types. Ultimately, the biggest rewards will still come from winning games, so mimicking poor players should have a minimal negative impact (as is usually the case with Eco-EA).
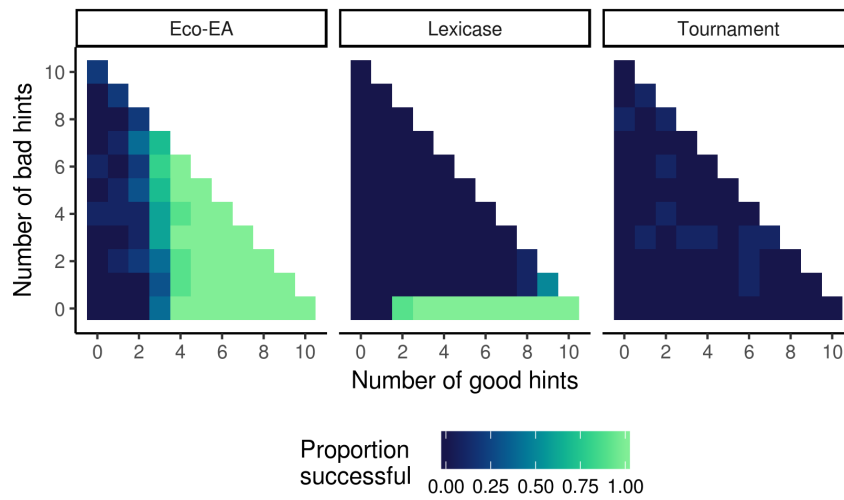
Figure 6.1: **Impact of good and bad advice on Eco-EA and Lexicase.** Heat maps for each algorithm show the success rate of that algorithm in the presence of varying quantities of good and bad advice. Note that tournament selection is not actually capable of receiving hints; it is presented here as a control. Each cell in the heat maps represents the proportion of runs (out of 10) that successfully found the optimal solution to the 10-dimensional box problem.

Chapter 7

The MODES Toolbox: Measurements of Open-ended Dynamics in Evolving Systems

This chapter is adapted from a paper in the *Artificial Life Journal* (Dolson et al., 2019b).

7.1 Introduction

A central goal of the field of artificial life is to build evolving systems that capture the full range of dynamics found in natural systems. Such systems should be capable of producing evolutionary outcomes such as sophisticated navigation behaviors, novel cooperative strategies, complex ecosystems, and major evolutionary transitions, to name but a few. Researchers seek such "open-ended" systems for a number of reasons:

1. For biologists, access to systems exhibiting complex and nuanced evolutionary processes allows rapid experimentation and facilitates developing a deep intuition for underlying mechanisms (Tenaillon et al., 2016).

2. For evolutionary computation researchers, insights from open-ended evolving systems

will allow researchers to break complexity barriers, expanding the classes of engineering problems that evolutionary algorithms can solve (Hara and Nagao, 1999; Potter and De Jong, 2000) and producing more general forms of evolved intelligence.

3. For artificial life researchers, it is concerning that there may be dynamics of fundamental importance to biology that artificial life systems do not exhibit. The existence of such dynamics suggests that we are not building evolving systems as innovative as those found in nature, be it due to limited memory, limited time, or simply an insufficient understanding of the necessary components. Identifying these missing factors should allow us to better understand life as it is and to better explore life as it could be.

While various artificial life systems have reproduced individual dynamics – such as the evolution of complex traits (Lenski et al., 2003), cooperative behaviors (Goldsby et al., 2012), and coexistence of diverse ecotypes (Cooper and Ofria, 2003) – these accomplishments have been in highly controlled circumstances. The overarching goal of open-ended evolution research is to create a system where all of these dynamics emerge more organically, as in nature. Additionally, replicating this process would provide substantial insights into our own origins, including the evolution of human brains. Indeed, harnessing a more open-ended set of evolutionary dynamics could help us spur breakthroughs in the evolution of general artificial intelligence.

Open-ended evolution is a many-faceted concept. A number of patterns are considered to be hallmarks of open-ended evolution (Taylor et al., 2016), most notably the continual production of novelty (Lehman and Stanley, 2011; Banzhaf et al., 2016), unconstrained increases in diversity (Bedau and Bahm, 1994), ongoing increases in complexity (Lenski et al., 2003; Korb and Dorin, 2011), and shifts in individuality such as those often associated with major transitions in evolution (Maynard Smith and Szathmary, 1997). There is a growing consensus in the field that all of these dynamics are important pieces of the open-ended evolution puzzle (Taylor et al., 2016). In addition, I, along with colleagues, have
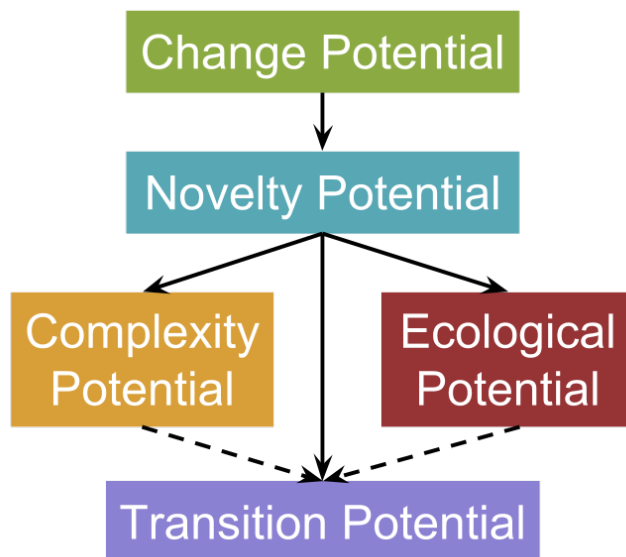
Figure 7.1: **Relationships between the metrics.** Originally published in (Dolson et al., 2015). Solid lines with arrows indicate metrics which are prerequisites for other metrics.

previously suggested that there is a fifth necessary and even simpler dynamic: continuous change in the information content of components in the population (Dolson et al., 2015).

These five properties of a system fit into a hierarchy, as shown in Figure 7.1. For novelty to exist, there must be some degree of change in the information within a population. While this observation is trivially true, many evolutionary algorithms suffer from premature convergence, the absence of non-trivial change. Thus, it remains an important prerequisite to define and explicitly address. Similarly, complexity and diversity can only increase indefinitely if novel members of the population continue to be generated. Finally, transitions in individuality typically involve multiple organisms coming together into a single individual, building from complex and diverse progenitors. All of these dynamics capture different subsets of interesting behavior that evolving systems might exhibit and I propose they are all necessary (but perhaps not sufficient) in a fully open-ended system.

To draw conclusions about what factors of a system promote or inhibit these dynamics, we need methods for measuring the extent to which each dynamic is present. Importantly, these methods must be applicable across a wide variety of systems. Some progress has been

made toward this end with evolutionary activity statistics (Bedau et al., 1998; Channon, 2001), an approach to isolating and quantifying the adaptive component of an evolving system, separating out the non-adaptive dynamics. Evolutionary activity statistics require that the user decide on two things ahead of time: a definition for "components" (meaningful individual pieces of a system) and a way of filtering noise out of the system (typically by contrasting with a shadow population that evolves with selective pressures turned off).

Thus far, components have needed to be defined for each system on a case-by-case basis. In artificial life systems, alleles or genotypes are typically used as components, while in the fossil record, whole species were used as components (Bedau et al., 1998). This flexibility to choose different components is valuable, as it allows for the study of open-ended evolution at different scales of organization. However, it also means that care must be taken when comparing evolutionary activity statistics across systems. Here, I suggest a component definition that should work for any system in which genomes are composed of elements that collectively determine fitness (see Identifying Meaningful Sites in the Genome). Note that, although this component definition will not work with every system, I will suggest other approaches with broader compatibility.

Due to the critical role of stochasticity in evolution, most evolving systems are noisy. In order to make behavioral generalizations, we need a way to distinguish evolutionary signal from this noise. In the original description of evolutionary activity statistics, a specific method was proposed for doing so: for each run of a system, there should be a corresponding "shadow" run in which any outcome of selection is replaced with a random choice. Dynamics observed in the shadow run can then be subtracted out from those in the main run. While this control can be highly informative, it is challenging to implement in many systems and requires researchers to be able to isolate all selective events in the system. For example, when evolutionary activity statistics were applied to the fossil record, a different filter had to be used: the assumption that any species that was successful enough to have made it into the fossil record was probably evolutionarily successful for a substantial amount of time. In

this chapter, I build on this idea to propose a filter for evolutionary activity that can be more easily implemented across a variety of systems (see Filtering Out Noise).

Evolutionary activity statistics classify evolving systems based on how open-ended they are. However, it is relatively easy to create a system that falls into the most open-ended class while still failing to further our goals for open-ended evolution research or to match our subjective understanding of what we would expect from a truly open-ended system (Maley, 1999). Indeed, there is debate over whether "open-endedness" is even quantifiable (Stanley and Soros, 2016). Moreover, it is my opinion that most efforts to define systems as either open-ended or not have largely been unproductive; open-endedness is likely better thought of as a continuum than as a binary. While there is much debate over what would constitute a fully open-ended system, there is consensus in the field that we are not particularly close to building such a system[1]. My goal in this chapter is to extend evolutionary activity statistics into easy-to-use diagnostic criteria that quantitatively measure key hallmarks of open-ended evolution. I want researchers to be able to isolate the effects of experimental settings on these hallmarks, keeping such results relevant across experimental platforms. As such, I hope to spur a more consistent and comparable march toward true open-endedness, adding new metrics to this toolbox as the community reaches a consensus on the features that we should promote.

In the rest of this chapter, I will introduce the Measurements of Open-ended Dynamics in Evolving Systems (MODES) toolbox and explore the behavior of the metrics it contains in the context of two evolving systems: NK-Landscapes (Kauffman and Levin, 1987) and the Avida Digital Evolution Platform (Ofria and Wilke, 2004).

---

[1]This line of thought originally lead us to conceptualize the metrics described here in terms of possible barriers a system might encounter that would prevent it from being open-ended (Dolson et al., 2015). However, my attempts to measure these barriers align closely with dynamics that have since been identified as hallmarks of open-ended evolution. Ultimately, these perspectives are two sides of the same coin and both are useful frames through which to view open-endedness. For simplicity, I phrase this chapter in terms of hallmarks rather than barriers.

7.2 Background

7.2.1 Evolutionary Activity

Evolutionary activity statistics attempt to quantify the extent to which adaptive dynamics are occurring in a population. In most applications, evolutionary activity has been measured as the length of time that components exist in the population beyond what would be expected in the absence of selection (Bedau et al., 1997, 1998; Channon, 2003). This value was chosen because it translates easily across systems and represents a universal measure of evolutionary "success". In earlier work, a measure of selective sweeps was used instead of component existence time (Bedau and Packard, 1992), but this metric could not be easily generalized across systems.

Multiple facets of evolutionary activity are used in the interpretation of evolutionary activity statistics: the activity of new components (A_{new}), the mean (or median) cumulative activity of components in the population ($\bar{A}_{cum}$), and the diversity of components in the population (D). Based on the long-term behavior of these values, systems that exhibit qualitatively similar dynamics have been grouped together into a class of evolutionary dynamics. Initially, three possible classes were described: no evolutionary activity, bounded evolutionary activity, and unbounded evolutionary activity. Over time, additional classes have been added to more precisely reflect the types of systems observed. For ease of referring to these classes, table 7.1 merges together all prior additions to the original classification system of which we are aware.

According to the original formulation of evolutionary activity statistics, in order for a system to be categorized among the most open-ended systems (originally class 3, now class 4) it must exhibit unbounded growth in summed evolutionary activity across all components in the population (Bedau et al., 1997) (see Table 7.1). Technically, this growth could happen either because of an unbounded increase in the number of components (diversity) or because of an unbounded increase in the average evolutionary activity of components in the popu-

lation. The latter case was originally thought to not occur (Bedau et al., 1998), however when such a case was observed, Channon suggested that class 3 open-ended dynamics should be broken up into three subcategories. These subcategories depend on whether the growth in evolutionary activity was driven by diversity, per-component evolutionary activity, or a combination of both (Channon, 2001) (see Table 7.1).

In parallel, Skusa and Bedau refined the classification in a different way (Skusa and Bedau, 2003), inserting a new second class in which evolutionary activity was unbounded but no novel components came into being (see Table 7.1). Such a situation involves purely ecological dynamics. This observation may seem surprising at first - shouldn't unbounded evolutionary activity involve adaptation? However, when evolutionary activity is measured as the existence time of a component, evolutionary activity statistics draw no distinction between stabilizing selection and selection favoring changes to the status quo (Channon, 2003). Thus, pressure for multiple eco-types to persist in their current form (i.e., ecology) will show up as as evolutionary activity above and beyond that observed in the shadow run.

In fact, the presence of a single component under stabilizing selection will trivially cause the mean evolutionary activity to increase indefinitely; such a component will sit in the population, increasing the population's activity counter despite being quickly lost from the shadow population. This behavior casts doubt on how we should interpret class 4b, as well. To remedy this concern, Channon suggested that we should look for unbounded growth in median (rather than mean) per-component evolutionary activity (Channon, 2003). This adjustment is a drastic improvement, but it still does not eliminate the possibility that systems exhibiting class 4b evolutionary dynamics are not doing quite what we would expect. If at least 51% of the components in the population are under stabilizing selection – as would be expected in an ecological system – the rest of the population could still be behaving like a class 3 system. While such a system would still be interesting for ecological studies, our understanding of it would not be well-served by conflating it with systems that are exhibiting open-ended adaptive evolution.

121

Class	Median[2] Evolutionary Activity ($\tilde{A}_{cum}$)	Change	Novelty A_{new}	Diversity (or Ecology) (D)	Complexity	Type of evolutionary dynamics	Described in
1	zero	?	zero	bounded	bounded	None	Bedau et al, 1997
2	unbounded	?	zero	bounded	bounded	Uncreative	Skusa and Bedau, 2003
3	bounded	positive	positive	bounded	bounded	Bounded	Bedau et al, 1997
4a	bounded	positive	positive	unbounded	bounded	Unbounded	Bedau et al, 1997
4b	unbounded	positive	positive	bounded	?	Unbounded	Channon, 2001
4c	unbounded	positive	positive	unbouned	?	Unbounded	Channon, 2001

Table 7.1: **Table combining all previously described classes of evolutionary dynamics as measured with evolutionary activity statistics.** For each class, I show the response of all quantities measured for evolutionary activity statistics and in these proposed metrics (novelty and diversity should behave equivalently between the two systems). Note that I expect bounded evolutionary activity to imply bounded complexity, as any scenario in which complexity is growing without bound should imply that evolutionary activity is too. Question marks indicate that the value of a given metric is not specified in the description for a class of evolutionary activity. Higher-numbered classes are generally believed to fall further along the continuum of open-endedness than lower-numbered classes. In principle, classes 4b and 4c could each be further split into sub-classes based on whether complexity is bounded or unbounded. Likewise, classes 1 and 2 could be further sub-divided based on whether change is 0 or positive. In the absence of further data on the behavior of real-world systems, it is unclear how helpful such increased precision would be.

How can we know whether evolutionary activity is driven by stabilizing selection rather than more interesting dynamics? If every component is experiencing directional (as opposed to stabilizing) selection, the change metric I propose here should theoretically be comparable to the number of components[3]. In contrast, if most of the population is under stabilizing selection, the change metric should be very low.

Ultimately, the change metric (described in the next section) is in keeping with the original evolutionary activity measurement, which sought to quantify the acquisition of new genetic information (Bedau and Packard, 1992). For this reason, in this suite of metrics, I replace the concept of evolutionary activity with change. I believe that this framing will be easier to measure and interpret with little loss of information (although of course I encourage the use of other measures of evolutionary activity where appropriate). My change metric does have the downside of not being possible to usefully classify as bounded or unbounded. Because I seek only to compare systems and identify progress toward higher levels of open-endedness, this limitation should not be a problem.

7.2.2 Prior work using MODES

Soros (Soros, 2018) used a preliminary version of this framework (Dolson et al., 2015) to study open-ended evolution in the artificial life system Chromaria. Agents in Chromaria are colorful circles controlled by CPPNs that must find a region of the world that matches their color in order to reproduce. These agents can be classified into species based on their patterns of coloration, and change and novelty can be assessed by measuring the emergence of new species. Ecological interactions in Chromaria occur as a result of individuals planting themselves in the world, which alters the color environment that subsequent agents must navigate. Thus, Soros was able to measure the ecology of Chromaria through a series of

---

[2]The original formulation of evolutionary activity statistics used mean rather than median, but Channon Channon (2003) makes a compelling argument for using median instead. Using median rather than mean does not change any of the intuitions for how we expect this metric to behave and reduces the risk of non-intuitive behavior due to outliers.

[3]However, the change metric may often be lower than the number of components because not every component will change during every measurement period.

visual snapshots of the world, as well as by measuring the number of species that occur over the course of a run. Lastly, she measured complexity in terms of the number of elements in the CPPNs controlling the agents.

Using these MODES-inspired metrics, Soros investigated three hypothesized necessary conditions for open-ended evolution: 1) some sort of minimal criterion (Soros and Stanley, 2014) must be met before reproduction, 2) when new types of individuals evolve, it should create new ways to satisfy the minimal criterion, and 3) individuals should be responsible for making decisions about how they interact with the world. By measuring hallmarks of open-ended evolution under various controls that removed these conditions, Soros found strong evidence that all of the conditions are indeed necessary for change and novelty (let alone ecology and complexity) in Chromaria. These experiments perfectly illustrate the kind of hypothesis-driven research that I hope a further formalization of these metrics will enable. Additionally, they serve as an example of the range of approaches that can be taken to translate these concepts between systems.

7.2.3   Applying MODES to biology

Since many hypotheses about open-ended evolution involve comparisons to the biosphere, it is critical that MODES metrics are applicable not only to digital systems, but are also relevant to experimental biological systems. To confirm that they are, consider how we would apply them to a well-studied wet-lab experiment. The Long-Term Evolution Experiment (LTEE) (Lenski et al., 1991) is an exemplar of experimental evolution, consisting of 12 populations of the bacteria *E. coli*, which have been evolving independently for more than 60,000 generations (Good et al., 2017). As detailed in (Taylor et al., 2016), the LTEE exhibits many hallmarks of open-ended evolution, including the criteria I propose here. Because fitness within the LTEE is best described by an unbounded power law function (Wiser et al., 2013; Lenski et al., 2015), the system meets the change metric: populations continue to change in non-trivial ways over time. Further, studies of individual populations

within the LTEE have shown numerous examples of the generation of novelty, including exploration of new areas of the fitness landscape (Tenaillon et al., 2016), repeated selective sweeps (Maddamsetti et al., 2015), and new diversity arising after such sweeps (Blount et al., 2012). Toward the ecology metric, several populations within the LTEE demonstrate frequency-dependent fitness dynamics (Ribeck and Lenski, 2015; Rozen and Lenski, 2000; Le Gac et al., 2012; Maddamsetti et al., 2015), which are necessarily cases of ecological interactions. Included in these cases of frequency dependence is a special case (Blount et al., 2008, 2012; Turner et al., 2015a) driven by cross-feeding and specialization on different resources (Turner et al., 2015b). Because each population in the LTEE descends from a single ancestor present at the start of the experiment, all ecological complexity in any population must have arisen during the course of the experiment, and thus satisfies the ecological metric. The complexity metric is inherently harder to quantify in a biological system than in a computational one, but recent large-scale genome sequencing from the LTEE (Tenaillon et al., 2016) offers the promise of being able to measure complexity at the genome level over the course of the experiment. Because these metrics can theoretically be applied to a well-studied and open-ended biological system, they can be used to compare dynamics in a broad range of systems and enable the field of artificial life to move forward in quantifiable steps to open-ended evolution.

7.3 Metrics

Like the original evolutionary activity statistics (Bedau et al., 1997), the metrics I present here can operate on a wide range of units. Often, these units will be genotypes or phenotypes but in other cases they may by higher-level taxonomic groups, such as species. To highlight this agnosticism, Bedau *et. al* referred to these units as "components". I will use the same terminology where applicable.
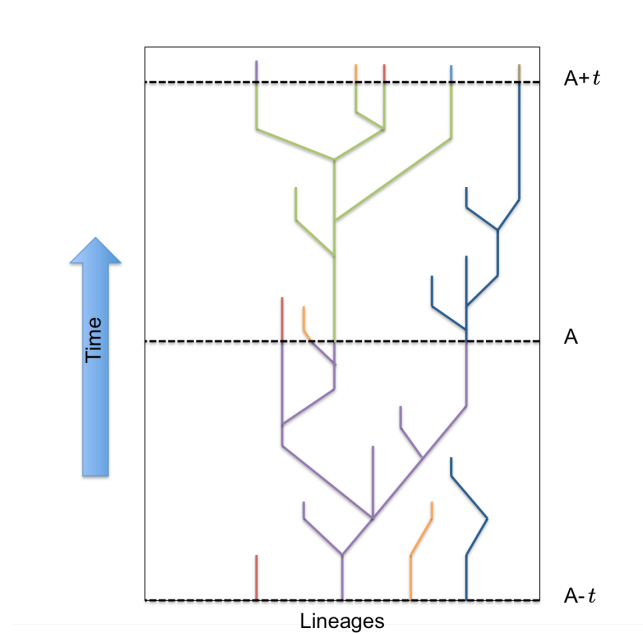
Figure 7.2: **An illustrative example of how I filter components for persistent lineages.** At time point A, the purple lineage has proven to be persistent and therefore the original component from A-t will be considered meaningful. Similarly, the green and blue lineages persist to time point A+t and so the original green and blue components will be considered meaningful as they existed at time point A.

7.3.1 Overarching techniques

I use two broad techniques to ensure that these metrics can focus on the most relevant and meaningful information in an evolving population. Additionally, I describe a technique for determining whether a metric is bounded or unbounded.

7.3.1.1 Filtering out noise

In any evolving population, mutations continually produce new maladapted components that are then purged from the population via natural selection. All of the MODES metrics assume that some form of filtering has been applied before-hand, to prevent maladapted components from overwhelming the hallmarks that I am measuring. Here, I describe a persistence filter that I use in these experiments. However, shadow runs are also a viable filter option, and there are likely further useful filtering techniques that have not yet been invented.

To focus only on the adaptive products of evolution, I limit my analysis to those components whose descendants persist for a substantial number of generations. I refer to this technique as a persistence filter. I mark each organism with a lineage ID at a given time point A, as demonstrated in Figure 7.2 (where color indicates lineage ID). The lineage IDs are passed on to offspring for the next t generations, where t is a pre-determined number of generations indicating the length of the filtering process (hereafter referred to as filter length). At time point $A + t$, I determine which components from the population at A have descendants at $A + t$. At this point, those components are considered **persistent**; in the example in Figure 7.2 the individuals at the bases of the green and blue lineages are considered persistent at time point $A + t$. These are the individuals that would be evaluated in the MODES metrics. This filtering leads to a delay in counting a component in a metric until t generations later, but enables us to avoid an apparent increase in metrics due to drift via mutation. For example, the red and orange components from time point A would not be considered in these metrics because their lineages do not persist to time point $A + t$.

How large should t be? The correct value depends on our goals. If we are interested in evolution on a shorter time scale, we may only want to filter out deleterious mutants, which will likely only survive a few generations. In this case, a relatively small value of t should suffice. Indeed, prior open-ended evolution research has used what is effectively a persistence filter with $t = 1$ as a supplemental filtering technique (Channon, 2003). If we are interested in a broader time scale, however, we may want to filter out neutral mutants too and measure only adaptive evolution. In this case, coalescence theory can inform the choice of t. In an asexual population without diversity-preserving forces, the population will periodically "coalesce", i.e. neutral clades will die out resulting in a new most-recent-common ancestor of the current population. If we take a snapshot of such a population at any given point in time and let the population continue evolving for long enough, a single individual from the snapshot will eventually be a common ancestor of the entire extant population. I refer to coalescence time here as the amount of time that this process takes, although it should be

127

noted that coalescence time is more commonly thought of retrospectively.

If we want to filter out all neutral mutants that do not go on to play a critical role in evolution, it would be ideal to choose a value for t that falls above the expected distribution for coalescence times. If we did so, then we could be confident that any individuals that made it past the filter represented a meaningful part of the evolutionary history of the population. If only a single individual makes it through the filter, that individual must be along the line of descent for the entire population. Multiple individuals making it through the filter would be evidence of ecological dynamics promoting their coexistence.

The median coalescence time for a well-mixed asexual population of N haploid individuals under no selective pressure is $2N$ generations (Fu and Li, 1999). Unfortunately, the expected distribution of coalescence time is exponential, meaning that we would have to choose a potentially impractically large value for t if we want to guarantee that it is rare to get through the filter by chance. However, the presence of selective pressure dramatically reduces expected coalescence time. Since most systems in which people study open-ended evolution do have selective pressure of some form, in practice relatively low values of t are still effective filters.

For a meaningfully comparison across populations, we must filter them using consistent values of t. We always expect filters with lower t values to let more individuals through, and it is challenging to separate this effect from changes in the underlying dynamics of the system. Additionally, t must be measured in generations to ensure consistency in the amount of filtering that occurs. Researchers studying systems that use a different time scale need only calculate the average generations within the population to measure t.

In evaluating results, the field should strive to use consistent values of t relative to population size and be aware that, all else being equal, increasing selective pressure will reduce the number of taxa that get through the filter.

This effect brings up an important distinction between this filtering technique and the shadow run traditionally used with evolutionary activity statistics. Whereas shadow runs

filter out the effect of neutral processes, the persistence filter does not entirely. I view this reduced filtering primarily as an advantage – drift can be an important part of the evolutionary process – but there may also be situations where it is undesirable. These metrics are unable to distinguish between class 1 and 2 dynamics or between class 3 and 4b dynamics (see Table 7.1), although they are able to distinguish between useful subcategories within those classes (as discussed in the Change Metric section below).

7.3.1.2 Identifying meaningful sites in genomes

Because genomes are such a common unit of taxonomic organization to use as components in open-ended evolution research, I present a technique for filtering noise out of genomes. Although this step is not necessary for using the MODES metrics, it will improve the signal-to-noise ratio in a variety of common use cases and simplify the calculation of complexity. I recommend its use where applicable.

While a genome may have descendants in t generations, if t is relatively small this persistent genome may not be phenotypically different from another persistent genome in the population. To ensure that we are not separately counting genomes that differ only in non-coding regions, we use an additional filter in which we determine which sites in the genome contain information about the environment. In calculating all of the following metrics, we first reduce the genome to its meaningful sites.

This approach can easily be extended to any system in which the genome is made up of a set of elements that collectively determine fitness. Whether or not a genomic position is meaningful can be approximated by measuring the overall fitness[4] effect of either removing it or changing it to a null alternative that is known to not contribute information. If removing the site resulted in a lower overall fitness, that implies that the site contained information (*i.e.* was meaningful). Conversely, if removing the site increased or had no effect on fitness we can conclude that it is most likely not meaningful. We can then define a component as

---

[4]As defined in the system being studied. If the system does not have a fitness definition, average lifetime reproductive output can be used.

the sequence of meaningful sites in a genome rather than the whole genome. By doing so, we avoid treating functionally identical components as distinct.

When should we remove sites and when should we replace them with a null alternative? A null alternative should be used in cases where changing the structure of the genome changes the meaning of other sites. For example, in Avida it is critical that we replace instructions with nulls rather than completely removing them because information can be encoded in the number of instructions between two other instructions. A more accurate technique would be to examine the fitness effect of substituting all possible alternative elements and calculate the potential entropy at that site. When null substitutions are not possible, this technique is an effective method.

A caveat to this technique is that, in practice, there are interactions between sites. By only knocking out a single site at a time, we miss these interactions. How to best remedy this situation is an open question, as measuring all possible combined effects is computationally intractable. In many cases, measuring pairwise interactions is possible and may be worthwhile. This issue will reduce the efficacy of this approach at reducing noise, because some functionally equivalent genomes will be classified as different. When used for calculating the complexity metric, it may cause fragile genomes to appear more complex than robust ones.

Note that, although identifying informative sites can be computationally intensive, we would need to do so anyway to calculate the complexity metric. As such, this additional layer of filtering is effectively free.

7.3.1.3 Determining boundedness

In the design of these metrics, I have primarily focused on determining the effect that small changes to a system have on the extent to which that system exhibits hallmarks. However, they can also be used to classify systems in much the same way that evolutionary activity statistics do. As described in Table 7.1, this classification requires determining

whether diversity is increasing without bound. In addition, it would be informative to determine whether complexity is growing without bound. In previous work, the definition of boundedness in this context has been stated in terms of the limit of the suprememum of diversity as time goes to infinity (Bedau et al., 1998). While this is an excellent theoretical definition, taking limits of empirical data as time goes to infinity is generally not practical. Previous applications of evolutionary activity statistics seem to determine boundedness based on whether or not a line on a graph appears to be plateauing. This technique has the potential to be misleading (Wiser et al., 2018).

Instead, I advocate the use of statistics to determine what mathematical model best fits the observed data. The pattern can then be classified as bounded or unbounded based on the limit of the best-fitting mathematical model. Such an approach has previously been used to demonstrate that fitness is following an unbounded growth pattern in a long-term wet lab evolution experiment with *E. coli* (Wiser et al., 2013; Lenski et al., 2015).

7.3.2 MODES Metrics

7.3.2.1 Change Metric

Our first metric focuses on whether the genetic makeup of the population is changing in a non-trivial way. This metric will be above zero during adaptive evolution, including situations where the population is returning to previous states, perhaps due to environmental cycling. In the work presented here, I use a persistence filter (explained above) to ensure that I mark a component as new only if its lineage full t generations. However, a different filtering technique (e.g. a shadow run) could be used instead. For this comparison, I first find the components from persistent lineages from generation A by determining which components have descendants in generation A $+ t$. In the example shown in Figure 7.2, the components at the roots of the green and blue lineages would count as persistent. I then compare these components to those found to have been from persistent lineages in the previous time point (e.g. I would compare the roots of the blue and green lineages to the root of the purple lineage

in Figure 7.2). In this way, I create a sliding window to observe change in the population. Note that the example in the figure assumes the resolution at which data are collected (i.e. the number of generations between time points) is equivalent to the value of t, but this does not need to be the case. It may be desirable to have a very long length of t but still gather data frequently. In such a case, each time point is individually filtered by looking ahead t generations, but change is calculated by comparing the set of persistent taxa in the current time point to the set of persistent taxa in the previous time point.

For those who find it helpful, the change metric can be formalized with the following equation:

$$[h]\text{change} = \sum_{c \in F} \begin{cases} 0 & \text{if } c \in F' \\ 1 & \text{otherwise} \end{cases} \tag{7.1}$$

where F' is the set of components from the previous time point that passed the filter, F is the set of components from the current time point that passed the filter, and c is a component in F.

While there is no change metric in the original conception of evolutionary activity statistics, I expect that it will provide similar information to cumulative evolutionary activity (Bedau et al., 1997). Change must be positive in systems exhibiting class 3 or higher evolutionary dynamics, as these systems must all exhibit positive novelty. Class 1 systems may or may not exhibit change; an evolving system that stagnates (e.g. many genetic algorithms) would have zero change, whereas a completely neutral system where all change was caused by drift would sometimes have a non-zero amount of change (depending on the value of t). Class 2 systems would have non-zero change if they were cycling between fixed states, but not if they were purely the result of stabilizing selection.

7.3.2.2 Novelty Metric

The novelty metric measures how many components have evolved in the population that have never been seen previously in the experiment. For this metric, I again filter out components that do not have descendants in t generations, enabling us to focus on meaningful novelty. As with change, I could have used a different filtering technique instead. To measure novelty, I simply count how many components from persistent lineages have never been in a previous time point's persistent component pool. It is possible with this metric for a component to evolve, but not persist, and therefore not be recorded in the permanent history, but then evolve and persist at a later point and be counted as novel. Once a component has been counted as novel, however, it is part of the permanent history and will never be counted in the novelty metric again. Thus, while a component could be delayed in being counted as novel, or not counted if it never persists, it will not be counted twice. The novelty metric is functionally equivalent to A_{new} in evolutionary activity statistics (Bedau et al., 1997).

The novelty metric can be formalized with the following equation:

$$\text{novelty} = \sum_{c \in F} \begin{cases} 0 & \text{if } c \in S \\ 1 & \text{otherwise} \end{cases} \tag{7.2}$$

where S is the set of all components that have ever passed the filter, F is the set of components from the current time point that passed the filter, and c is a component in F.

7.3.2.3 Complexity Metric

The complexity metric measures the maximum complexity of any component found in the entire population. There is still debate over how to best measure complexity and not all approaches are viable for all component types. Here, I recommend an information-theoretic approach which works well with the components described in the section on "Identifying meaningful sites in genomes" above. Once the meaningful sites have been identified, they

can simply be counted to get a measurement of complexity; the value of the complexity metric at a given time point is the highest observed count of informative sites across all components in the population that make it through the filter. This metric is somewhat crude and can be improved by using more advanced information-theoretic techniques where all possible mutations are considered at each site. Ideally epistatic interactions between sites would also be considered by measuring the fitness effects of knocking out combinations of genes. Unfortunately, doing so is often not viable in practice.

There is no equivalent to the complexity metric in evolutionary activity statistics. However, as many believe growth in complexity to be an important hallmark of open-ended evolution (Taylor et al., 2016), I feel it is a critical addition. In particular, it would be interesting to find non-trivial systems that exhibit unbounded growth in complexity. I suspect that such growth could only occur in systems exhibiting class 4b or 4c evolutionary dynamics, as bounded evolutionary activity should imply bounded complexity (although the converse is not true).

7.3.2.4 Ecological Metric

The ecological metric measures the amount of information in the population as a whole. While components may not individually contain increasing amounts of information (as measured by the complexity metric), they could still be increasingly diverse and therefore contain increased information collectively in the population. Ideally, we would measure this collective information by tracking the origin of each piece of information across all components in the population and summing up the number of unique pieces of information. Unfortunately, this approach is not computationally practical for many systems. As a proxy, we can look at the diversity of post-filter genotypes (reduced to informative sites, where possible). Complex ecologies in which multiple subsets of the population are using different information about the environment to survive are likely to be characterized by a relatively balanced distribution of individuals across the various successful phenotypes. Thus, I use Shannon entropy

(Shannon, 1948), a popular metric of diversity that also measures evenness, to measure the diversity of the persistent genotypes and calculate the ecological metric. This metric is equivalent to D in evolutionary activity statistics (Bedau et al., 1997).

The equation for Shannon entropy is:

$$\text{ecology} = -\sum_{c \in F} P(c) \log_2(P(c)) \tag{7.3}$$

where c is a component in F, the set of components at the current time step that passed the filter. $P(c)$ is the proportion of F occupied by component c. This value gets higher when the number of components in F increases and when the components occupy more equal proportions of the population.

7.4   Experimental Design

I used two radically different experimental systems in order to ensure both that these metrics can be broadly applied and that they produce meaningfully consistent results. For both systems, I used genomes as components.

7.4.1   NK Landscape

Our basic treatment used $N = 20$ (i.e., 20 bits in an individual) and $K = 3$ (the fitness contribution of each bit was influenced by three other bits). I used a population size of 200 and a per site mutation rate of .05, with tournament selection and a tournament size of 2. In addition to this baseline treatment, I tested the effects of eight experimental treatments: **High K** tests the effect of a highly rugged landscape (K=10) where fitness is effectively randomized whenever a mutation occurs. **High N** tests the effect of longer bit-string genomes (N=100; mutation rate was adjusted to 0.01 to keep the whole-genome mutation rate consistent with the base condition), allowing for a higher potential complexity. **Low Mutation** and **High Mutation** test the effects of more extreme mutation rates (0.005

and 0.1 respectively); I expect the mutation rate to be important for finding new areas of the fitness landscape and thus the novelty metric. **Small Pop** and **Large Pop** vary the population size (to 20 and 1000 respectively); in small populations I expect more drift in the population, allowing more change, while in a large population I expect stronger selection and consequently that a higher percentage of changes along the line of descent are beneficial. Finally, I included two special treatments: in **Oscillating Environment**, the fitness function was toggled between two pre-defined NK Landscapes every 500 generations, allowing us to see the effect of changing selective pressures where the populations was not able to stay on a single peak. In **Fitness Sharing** organisms that were too similar to each other detracted from each other's fitness, creating a pressure to explore multiple portions of the landscape at the same time and, ideally, maintain a high diversity. I used the fitness sharing equations described by Goldberg and Richardson Goldberg et al. (1987), with a sharing threshold of 50 and an α of 1. For all experiments, I used a filter length (t) equal to the population size.

7.4.2   Avida

To understand how MODES metrics will behave in a full-featured artificial life system, I tested them in Avida under a variety of scenarios. For all experiments, I used a well-mixed population in order to speed up the expected rate of coalescence. All other parameters in Avida were left at their default values. I ran experiments in both the **minimal** and **Logic-9** environments.

Artificial life systems necessarily have constraints on the amount of time and memory we can give them. It is important in open-ended evolution research to determine whether these constraints are imposing practical limitations on the dynamics the system exhibits (Zaman, 2018). To do so, I ran experiments in each environment at three different population sizes: 500, 1000, and 2000. In each condition, I ran 30 replicate runs of Avida.

Additionally, to understand how sensitive these metrics are to the choice of the filter length (t) I conducted some additional experiments in the empty environment in which we

varied t. In general, since the Avida runs are so long, I aim to filter neutral mutants out with the persistence filter, rather than just deleterious mutants. At each of the three population sizes, we tried t values of 500, 1000, and 2000. To ensure that I always have data from a filter length larger than population size, I also included a condition with a population size of 2000 and a t of 4000.

7.4.3 Implementation Details

If not implemented with care, these metrics can become computationally intractable in the context of the long experiments that open-ended evolution research often entails. In particular, RAM requirements can become prohibitive. I provide a few high-level approaches to mitigating these issues.

The largest memory cost is imposed by the novelty metric's requirement that we keep track of every taxon that has ever passed the persistence filter. Because we only need to know when we encounter a repeat taxon (rather than storing an archive of all taxa we have encountered), this cost can be dramatically reduced by using a Bloom filter (Bloom, 1970). Although this approach does introduce a (tunable) risk of false negatives (i.e. miss-classifying a novel taxon as not novel), this risk only makes the metrics more conservative.

The next largest cost is imposed by needing to keep track of the phylogeny over time. In addition to standard phylogenetic pruning techniques (such as removing all taxa that do not have extant descendants), all taxa that died before the current generation minus t can safely be removed[5]. This optimization prevents the tree from growing without bound over the course of the experiment.

Lastly, it is helpful to be aware that increasing t will reduce computational demands by increasing the percentage of taxa that will be filtered out. With these optimizations, MODES metrics can be implemented with minimal overhead.

---

[5]An important caveat is that this approach will only work with a strictly increasing unit of time. In many systems (including Avida) the average generation is not guaranteed to consistently increase. To support such systems, the implementation of the metrics allows for time to be tracked using two units at once, one corresponding to generation, and one that is guaranteed to be strictly increasing.

7.4.4 Statistical Methods

Effect size measurements determine whether a treatment has a meaningful impact on a variable. Because standard deviations varied wildly among conditions, I used Glass's Δ as a measure of effect size (Hedges, 1981). As a general guideline for interpreting Glass's Δ, a value of 0.2 is generally thought to be low while a value of 0.8 is generally thought to be high (although this guideline is context dependent). Distributions of final metric values are visualized using rain-cloud plots (Allen et al., 2018). Statistical code and supplemental statistical information is freely available (Dolson, 2018b).

7.4.5 Code Availability

All code used in this chapter is open source and freely available (Dolson, 2018b; Bryson et al., 2018; Ofria et al., 2018).

7.5 Results and Discussion

To ensure that these metrics are capturing the dynamics that we want them to, I tested them on a range of variants of the basic NK model and a range of conditions in Avida. The preliminary results for each metric are presented here.

7.5.1 Change Metric

In the baseline and low mutation rate conditions for the NK Landscape, change is close to 0 (see Figures 7.3 and 7.4), indicating that these metrics are capable of detecting the stagnation typical of many genetic algorithms. As shown in Figure 7.3, several environmental changes increase the amount of meaningful change found in the NK Landscape populations over time. When organisms are forced to share fitness between others with the same genotype, the amount of change increases and remains higher than the baseline over time.
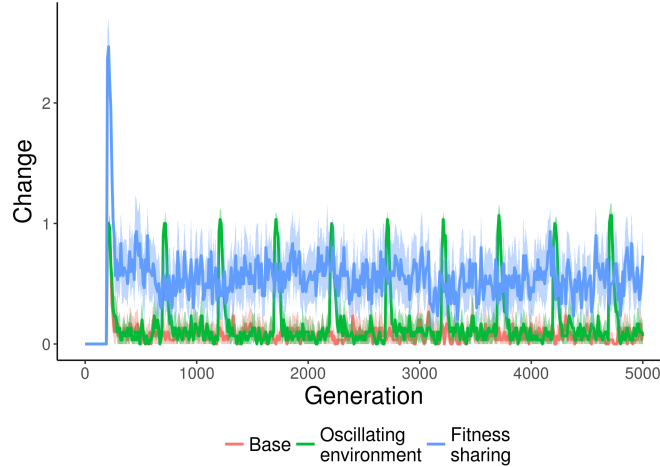
Figure 7.3: **Amount of change at over time in varying NK Landscape environments.** Fitness sharing increases the amount of change in the population over time. Conversely, a routinely changing environment leads to spikes in change that quickly drop as the population converges again. Shaded region represents a bootstrapped 95% confidence interval around the mean.

Conversely, when the environment changes frequently, there is an initial spike of increased change that quickly drops back down to the baseline value.

The majority of environmental conditions I tested in the NK Landscape system produced dynamics over time qualitatively similar to the baseline treatment. In Figure 7.4, I show the amount of meaningful change in populations at the final time point in more environmental conditions. A higher mutation rate leads to increased meaningful change ($p < 0.0001$, Wilcoxon test; Glass's $\Delta = 0.80$) because mutations are necessary to create any meaningful change in this system. A smaller population size produces more meaningful change ($p = 0.004$, Wilcoxon test; Glass's $\Delta = 0.33$) because a small population cannot hold as many different genomes at one time and therefore there are more genomes that can arise that are different than what is in the previous population. Finally, fitness sharing produces increased meaningful change ($p < 0.0001$, Wilcoxon test; Glass's $\Delta = 0.63$) because it creates a constant pressure for the population to adapt away from whatever is the current dominant genotype.

In Avida, there is always at least a little change (see Figure 7.5. This observation is
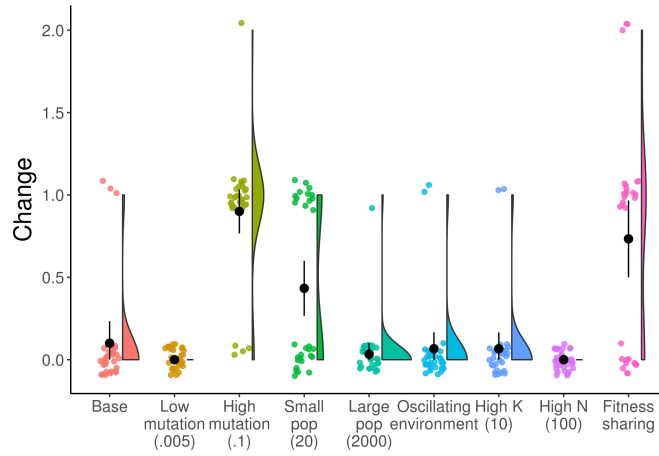
Figure 7.4: **Rain-cloud plot of change at final generation across NK Landscape treatments.** Environmental conditions that increase the amount of change at the final time point include: increasing the mutation rate, decreasing the population size, and implementing fitness sharing. Black circle and line indicate mean and bootstrapped 95% confidence interval.
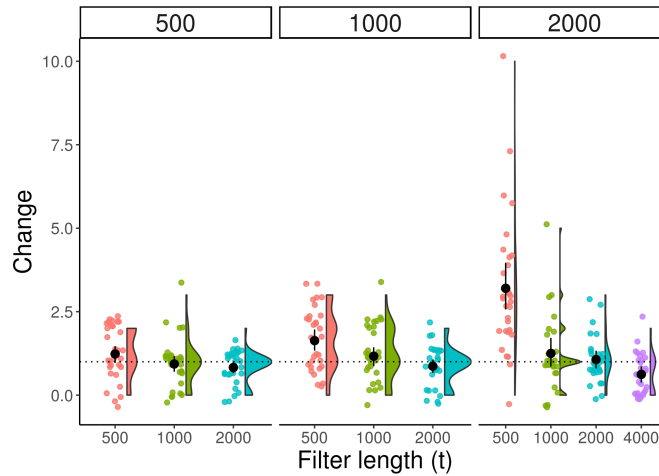
Figure 7.5: **Rain-cloud plot of change at final generation across multiple population sizes and filter lengths in Avida in the empty environment.** Labels along the top indicate population size. Black circle and lines indicate mean and bootstrapped 95% confidence interval. Horizontal bar indicates change = 1, the expected average change in the empty environment.

consistent with previous findings that fitness in Avida increases indefinitely (Wiser, 2015), as an increase in fitness implies both change and novelty. Based on coalescence theory, I would expect change in the empty environment to usually be less than or equal to one, because it is a single-niche environment. During each interval, either a fitter genotype will arise and sweep the population or the current fittest genotype will remain dominant. Because the value of t is not higher than the maximum expected coalescence interval, we should also expect to see the occasional time point with change greater than one. These data are roughly consistent with this expectation. In addition, there is a subtle downward trend in the change data, likely due to the progressively increasing difficulty of finding beneficial mutations.

As expected, increasing the filter length, t, decreases the amount of change observed because fewer taxa are able to get through the filter (see Figure 7.5). In general, using a value of t equal to population size seems to be an adequate filter. The confidence interval for the mean of these conditions always overlaps 1, indicating that a substantial amount of filtering is occurring. Using lower values of t begins to lead to substantial increases in the variance of observed change. Using a higher value of t does further reduce noise, but with diminishing returns. In the empty environment, population size does not appear to have much impact on change, implying (unsurprisingly) that population size does not exert pressure on evolutionary dynamics in this environment.

In the logic-9 environment, however, there is a slight increase in change as population size increases, particularly early in the experiment (see Figure 7.6). Additionally, change is generally a little higher in the logic-9 environment than the empty environment. This distinction is an unexpected benefit of using a value of t too low to guarantee coalescence. Logic-9 is a single-niche environment, so if we chose a large enough value of t, we would expect change to always be less than or equal to 1. However, logic-9 is a more complex environment than the empty environment, which increases the odds that multiple lineages will be able to keep evolutionary pace with other for a substantial amount of time. As such, at the values of t I used, the change metric is able to reflect the fact that more is going on

Figure 7.6: **Change over time across different environments and population sizes in Avida.** Note that y axes have different scales. In general, change is much higher in the Logic-9 environment. Filter length, t, is equal to population size.

in the logic-9 environment than the empty environment.

While change is a metric often not considered in discussions of open-ended evolution, these results show that the amount of meaningful change can reflect differences in the environment and evolution of the populations and is likely a necessary dynamic for open-ended evolution. The change metric responds in intuitive ways to variations in parameter settings, suggesting that it is a reliable indicator of the dynamics I designed it to capture.

7.5.2   Novelty Metric

As shown in Fig 7.7, a higher mutation rate increases the amount of novelty generated by the NK Landscape system. This result is to be expected, because more mutations make it easier to cross fitness valleys and otherwise traverse the fitness landscape. Even at a high mutation rate, novelty does start to decrease over time as the search space is explored.

I again found that the majority of treatments had a qualitatively similar trajectory over time and therefore in Figure 7.8 I show only the final novelty value. As predicted, the baseline treatment has, with the exception of 1 replicate, stopped producing meaningful novelty by the final time point. Indeed, the only environment still reliably producing meaningful novelty
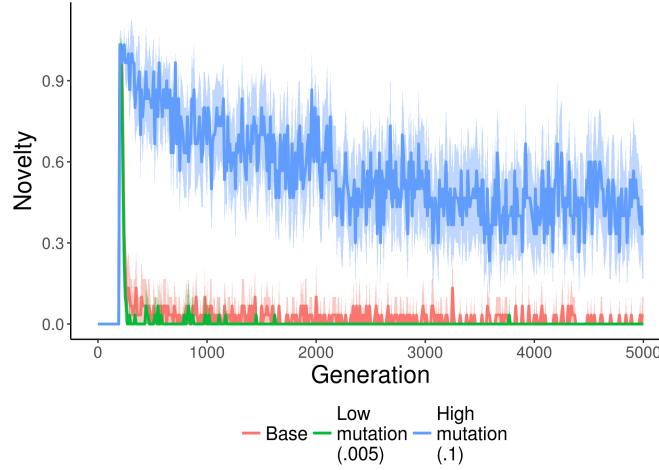
Figure 7.7: **Amount of novelty over time in NK Landscape with varying mutation rate.** The novelty metric measures the number of completely new meaningful genomes that have lineages that persisted since the previous time point. As the mutation rate increases, more novelty is continuously produced. However, at all mutation rates the novelty decreases over time. Shaded region represents a bootstrapped 95% confidence interval around the mean.

Figure 7.8: **Rain-cloud plot of novelty at final generation across NK Landscape treatments.** . Black circle and lines indicate mean and bootstrapped 95% confidence interval. At the final time point, little meaningful novelty is found in the baseline populations. Only increasing the mutation rate increases novelty, implying that other conditions with high change are simply promoting cycling between a fixed set of genotypes.

143

is the high mutation rate condition ($p = 0.003$, Wilcoxon test; Glass's $\Delta = 0.3$). High mutation rates produce ongoing novelty by shifting the mutation-selection balance such that drift is able to preserve novel lineages for longer than t. Notably, fitness sharing does not increase the final novelty ($p = 0.333$, Wilcoxon test), presumably because it is promoting cycling among previously-discovered genomes.

Like NK Lanscapes, Avida shows gradually declining novelty over time (although novelty in Avida declines more slowly) (Dolson, 2018b). This trend presumably reflects the declining availability of beneficial mutations. Interestingly, the novelty and change graphs from Avida look almost identical, implying that there is little cycling among previously discovered solutions. This result, too, is consistent with the previously observed indefinite fitness increases in Avida (Wiser, 2015) - if the population is always getting fitter, genotypes from earlier in the run would be unlikely to survive in the present.

These results highlight the power of the novelty metric to identify environments and populations that have the potential to be open-ended due to the high number of new genotypes being consistently discovered. Novelty is likely necessary, but not sufficient, for open-ended evolution because if nothing new is being produced by a population, neither the complexity nor the ecological metric can be non-zero.

7.5.3 Complexity Metric

In an NK Landscape populations, organisms cannot evolve to be more complex than N. As a result, the complexity increases over time and then saturates. High mutation rate ($p < 0.0001$, Wilcoxon test; Glass's $\Delta = -2.6$) and small population size ($p < 0.0001$, Wilcoxon test; Glass's $\Delta = -.23$) reduce complexity because they shift mutation-selection balance to make staying on a fitness peak more challenging (see Figure 7.9). All other treatments were able to achieve maximal complexity fairly reliably. Note that, due to the rough nature of the complexity metric, this merely indicates the presence of an individual on a fitness peak. I cannot distinguish between higher and lower peaks. The high N treatment
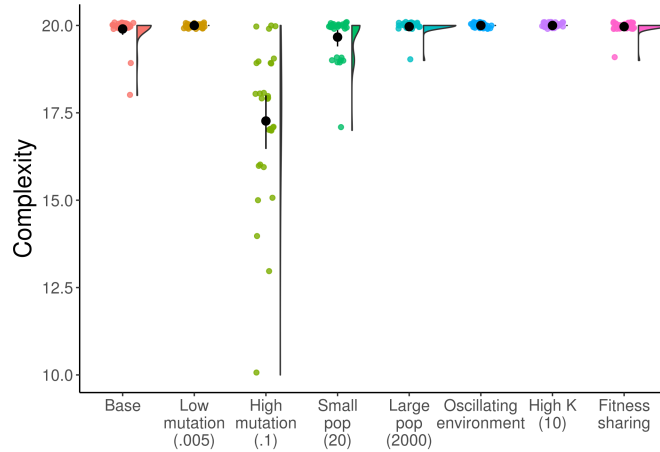
Figure 7.9: **Rain-cloud plot of complexity at final generation across NK Landscape treatments.** Black circle and lines indicate mean and bootstrapped 95% confidence interval. Note that I have excluded the high N condition from this graph because it throws off the axes. Most of the treatments reach the maximum complexity allowed by the genome length (20 or 100) and cannot continue to increase. High mutation rate and smaller populations decrease the final complexity achieved by the populations on average.

was, unsurprisingly, able to attain a drastically higher complexity due to its increased upper bound on complexity.

The NK Landscape results demonstrate that the complexity metric correctly identifies a system that is not able to continuously produce more complex solutions. Once the maximum complexity allowed by the genome length is reached, no higher value is possible.

In Avida, the complexity metric reveals a stark difference between the empty and logic-9 environments. In the empty environment, there is a rapid rise in complexity followed by a decrease and leveling out. This behavior is likely due to the strong pressure in Avida to become a more efficient self-replicator by optimizing code. Over time, simpler solutions are selected for, all else being equal. In the logic-9 environment, on the other hand, there is an ongoing upward trajectory in complexity. While logic-9 still rewards efficiency, algorithms that can make maximally efficient use of the tasks are complex. These results are consistent with other measurements of complexity in Avida over time (Adami et al., 2000).

In the biosphere, complexity appears to be growing without bound (Korb and Dorin, 2011), although there is debate over the mechanisms behind this process. As such, building a
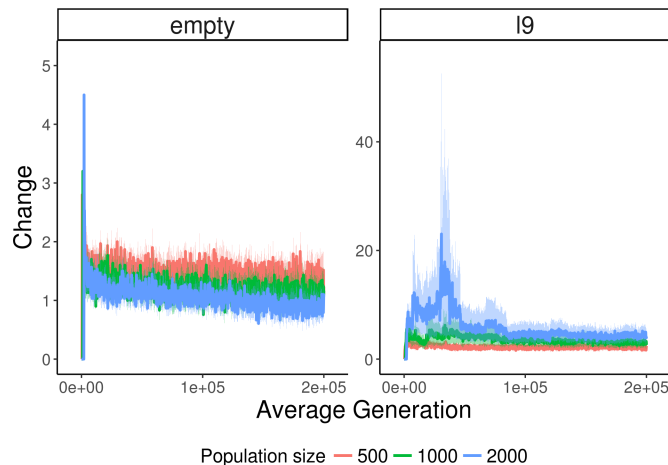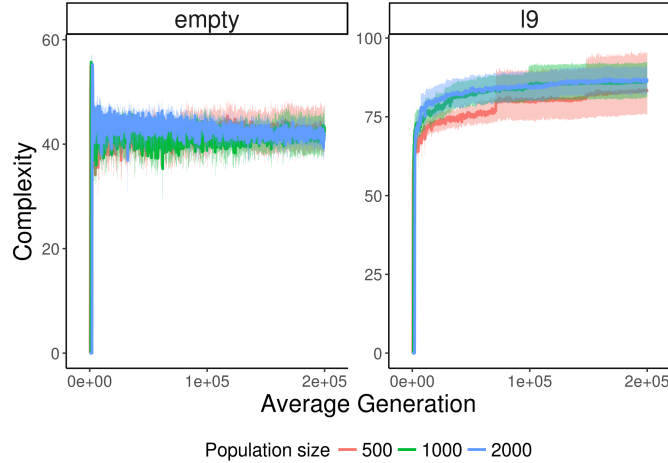
Figure 7.10: **Complexity in Avida over time across different environments and population sizes.** Note that y axes have different scales. In general, complexity appears to continue increasing in the Logic-9 environment, whereas it drops and then stabilizes in the empty environment. Filter length, t, is equal to population size.

non-trivial system that exhibits such behavior is a worthwhile goal for open-ended evolution research. Unbounded growth in complexity is only possible in a system with a sufficiently complex environment such that there is always new information to be integrated into the genome.

7.5.4 Ecology Metric

Across both systems, the only condition that creates a multi-niche environment is the fitness sharing condition in NK Landscapes. Accordingly, that is the only condition in which I observe ecology significantly above the baseline ($p < 0.0001$, Wilcoxon test; Glass's $\Delta = 0.91$) (see Figure 7.11). Because fitness sharing specifically rewards organisms with less common genotypes, it promotes a stably high ecology value over time. This result demonstrates the trade-off inherent in fitness sharing because it leads to higher ecology at the expense of lower complexity. Interestingly, in the Avida data, I see a consistent low level of ecology across all conditions (Dolson, 2018b). I hypothesize that this slight increase over NK landscapes is due to the noise introduced by Avida's variable generation times.

Ecology is an incredibly powerful force in nature, leading to feedback cycles of ever-

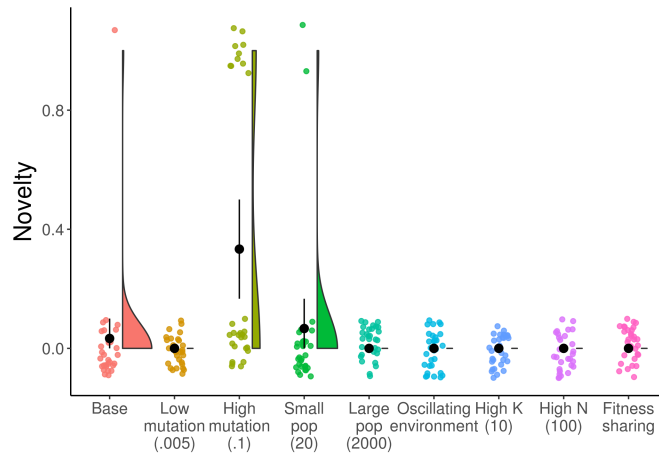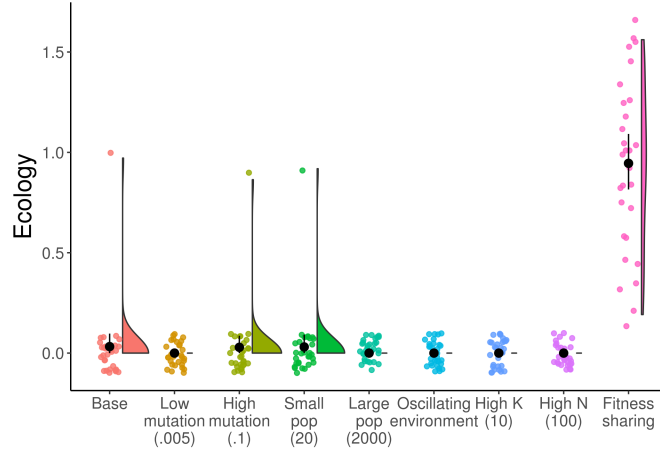Figure 7.11: **Rain-cloud plot of ecology at final generation across NK Landscape treatments.** Black circle and lines indicate mean and bootstrapped 95% confidence interval. Fitness sharing is the only condition that reliably produces ecology.

increasing diversity. Like complexity, diversity seems to be growing without bound in the biosphere (Harmon and Harrison, 2015). Thus, it will be important to see what mechanisms are important to promoting it in artificial life systems, too.

7.6 Conclusions

I have proposed a suite of metrics that quantify the presence of four generally-accepted hallmarks of evolution. These metrics build on prior work with evolutionary activity statistics and are largely compatible with them. Additionally, I have proposed techniques for reducing noise in these statistics. By testing them on two very different well-understood evolutionary systems, I have demonstrated that these metrics respond in an intuitive way to the dynamics these systems exhibit. Thus, these metrics should also be useful in understanding the extent to which novel systems exhibit hallmarks of open-ended evolution. Moreover, we can use them to understand the impact of incremental changes to a system. By breaking the seemingly monolithic problem of designing an open-ended evolutionary system into smaller, measurable pieces, we facilitate improved use of the scientific method. One of the primary goals in building an open-ended evolutionary system is to understand the

underlying components that are necessary to do so. By measuring the effects that controlled changes to a system have on this suite of metrics, we can more productively work toward these goals.

Going forward, it will be interesting to see how a wider variety of artificial life systems respond to the metrics in the MODES toolbox. In particular, further investigation on the differences between using a shadow run as a filter and using the persistence filter described here would be worthwhile. Ultimately, these two techniques capture sufficiently different information that it may be valuable to use each in turn.

For a long time, the field of open-ended evolution has been plagued by a lack of data that can be meaningfully compared across systems. I believe that the MODES toolbox will help remedy this problem by making useful metrics easily accessible. As new hallmarks of open-ended evolution are identified and new techniques of reducing noise are developed, I encourage contributions to the toolbox. It may also be worthwhile to complement these high-level metrics that screen for the presence of hallmarks of open-ended evolution with a suite of lower-level metrics that provide more mechanistic insight into why hallmarks are or are not being observed (Dolson et al., 2018b). By working together as a community of users, researchers, and developers, we can dramatically increase the rate at which open-ended evolution research progresses.

Chapter 8

Interpreting the Tape of Life: Ancestry-based Analyses Provide Insights and Intuition about Evolutionary Dynamics

This chapter is adapted from a paper in the *Artificial Life Journal* (Dolson et al., 2019a).

8.1 Introduction

Evolution is a collective effect of many smaller events such as replication, variation, and competition that occur on a fine-grained temporal scale. While evolution's emergent nature can be fascinating, it also presents challenges to studying the short-term mechanisms that, in aggregate, govern long-term results. In computational evolutionary systems, we can theoretically collect data to help untangle these mechanisms. In practice, however, the sheer number of constituent events produce an overwhelming quantity of data. In response, I have developed a standardized suite of diagnostic metrics to summarize short-term evolutionary

dynamics within a population by measuring lineages and phylogenies. Here, I describe these metrics and provide experimental results to develop an intuition for what they can tell us about evolution.

A lineage describes a continuous line of descent, linking parents and offspring in an unbroken chain from an original ancestor. A complete lineage can provide a post-hoc, step-by-step guide to the evolution of an extant organism where each step involves replication and inherited variation. Indeed, lineage analyses are a powerful tool for disentangling evolutionary dynamics in both natural and digital systems; digital systems, however, allow for perfect lineage tracking at a level of granularity that is impossible in modern wet lab experiments. These data allow us to replay the tape of life in precise detail and to tease apart the evolutionary recipe for any phenomenon of interest (McPhee et al., 2016b). In one notable example, Lenski *et al.* used the lineage of an evolved digital organism in Avida to tease apart, step by step, how a complex feature (the capacity to perform the *equals* logical operation) emerged (Lenski et al., 2003).

Yet, tracking the full details of a single lineage, much less a population of lineages, can be computationally expensive and will inevitably generate an unwieldy amount of data that can be challenging to visualize or interpret (McPhee et al., 2016a). Summary statistics can help alleviate these issues by enabling the user to focus on aggregate trends across a population rather than needing to examine each individual's lineage. The question is how to effectively summarize a path through fitness space. One useful abstraction is to treat the path as a sequence of states. Here, we primarily use phenotypes and genotypes as the states in the sequence, but we could just as easily use some other descriptor of the lineage at a given point in time. With this abstraction in hand, a few metrics are easily formalized: the number of unique states, the number of transitions between states, and the amount of time spent in each state. Additionally, we may care about how the transitions between states happened. What mutations led to them? Were those mutations beneficial, deleterious, or neutral at the time? These mutations are particularly notable because they did not simply

appear briefly, but stood the test of time, leaving descendants in the final population. Here, I explore a subset of these metrics that I expect will be broadly useful.

Whereas a lineage recounts the evolutionary history of a single individual, a phylogeny details the evolutionary history of an entire population. Measurements that summarize phylogenies can provide useful insight into population-level evolutionary dynamics, such as diversification and co-existence among different clades. A variety of useful phylogeny measurements have already been developed by biologists (Tucker et al., 2017). These measurements tend to treat the phylogeny as a graph and make calculations about its topology. Tucker *et al.* group them into three broad categories: assessments of the quantity of evolutionary history represented by a population, assessments of the amount of divergence within that evolutionary history, and assessments of the topological regularity of the phylogenetic tree. Such measurements can help quantify the behavior of the population as a whole, providing insight into interactions between its members. Thus, they are useful indicators of the presence of various types of eco-evolutionary dynamics.

Here, I present three types of lineage measurements and suggest adopting four phylogeny measurements from biology; these are lineage length, mutation accumulation, phenotypic volatility, depth of the most-recent common ancestor, phylogenetic richness, phylogenetic divergence, and phylogenetic regularity. For each metric, I discuss its application and my expectation for what it can tell us about evolution. I evaluate my intuition in two computational contexts: first, on a set of two-dimensional, real-valued optimization problems under a range of mutation rates and selection strengths and, second, on four qualitatively different environments in the Avida Digital Evolution Platform (a minimal control environment, an environment that rewards solving nine Boolean logic functions, an environment with limited resources, and a simple changing environment). For simplicity, I restrict my attention to asexually reproducing populations; however, I suggest techniques for extending these metrics to sexual populations.

There are many visual analysis tools that operate on lineages and phylogenies. These

data visualizations can provide further insight into evolutionary dynamics within a run. In this chapter, I discuss some approaches to visualizing lineages and phylogenies; I use these tools to build intuition about the behavior of these metrics.

In addition to demonstrating a range of analysis tools that are useful to digital evolution research, I intend for this work to begin a conversation within the artificial life community about how we quantify, interpret, and compare observed evolutionary histories. There have been extensive efforts to improve our ability to represent and visualize both lineages and phylogenies (Standish and Galloway, 2002; Burlacu et al., 2013; McPhee et al., 2016b,a; Lalejini and Ofria, 2016), which are indispensable for building intuitions and qualitatively understanding the dynamics embedded in a population's evolutionary history. However, I am unaware of efforts to formalize a suite of quantitative lineage and phylogeny-based metrics for computational evolution. Lineage and phylogeny-based analyses in artificial life are often a component of a larger set of experiment-specific analyses or are limited to qualitative descriptions and/or visualizations. Well-defined metrics not only provide valuable tools for teasing apart evolutionary dynamics, they also allow us to move away from exclusively qualitative descriptions of results toward a deeper quantitative understanding. This understanding, in turn, facilitates rigorous statistical analyses and hypothesis testing as well as comparison of evolutionary dynamics across different digital evolution systems.

8.2   Metrics

8.2.1   Lineage Metrics

Each of the three lineage metrics that I discuss — lineage length, mutation accumulation, and phenotypic volatility — reduces a lineage to a linear sequence of states where each state represents an individual or sequence of individuals that share a common genotypic or phenotypic characteristic of interest; Figure 8.1 is given as a toy example to help guide discussion of these metrics. While I limit my focus to three lineage metrics, this abstraction
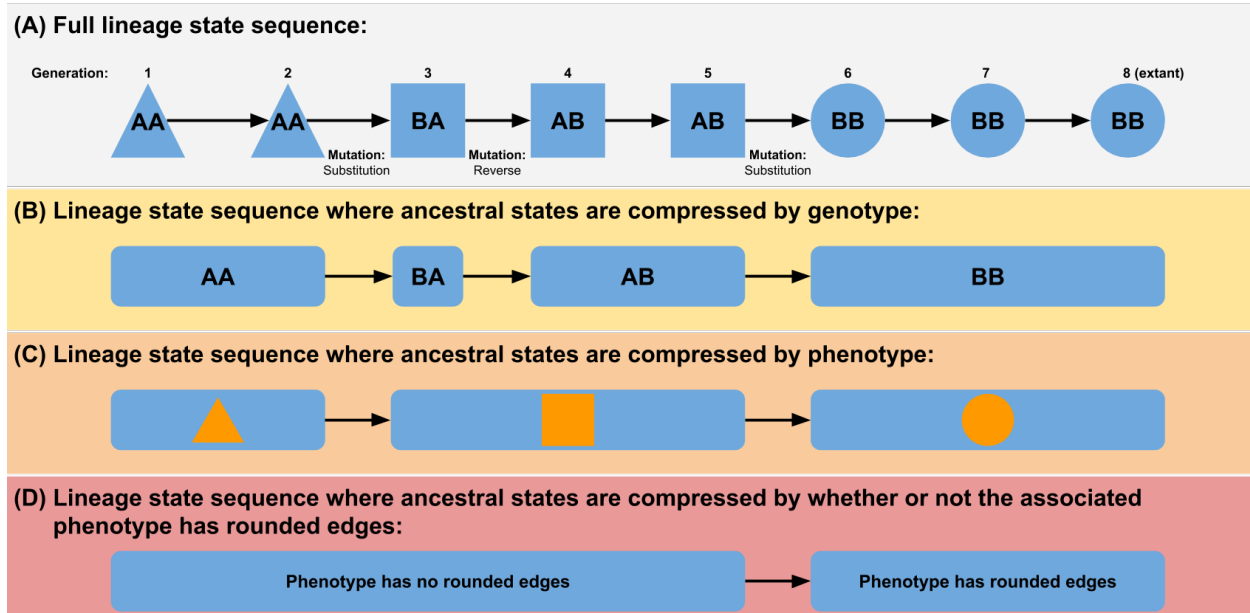
Figure 8.1: **Four methods of representing a lineage.** This example lineage has accumulated three mutations (one reverse mutation and two substitutions) and gone through three distinct phenotypes. In (A), each state along the lineage represents a single individual; lineage length is the number of generations spanned by the lineage (eight). In (B), states represent the sequence of genotypes along the lineage, reducing lineage length to four. In (C) states represent the sequence of phenotypes along the lineage; lineage length is the number of times a different phenotype is expressed (three). In (D), states are a particular phenotypic characteristic; here, lineage length is two.

places lineages in a form suitable for a wide range of measurements, including the direct application of many data mining techniques designed to operate over sequences such as sequential pattern mining, trend analysis, *et cetera* (Han et al., 2011).

Only asexual lineages where genetic material is exclusively vertically transmitted can be directly abstracted as a *linear sequence* of states. Sexual reproduction (and any form of horizontal gene transfer) complicates matters significantly as such lineages are more appropriately represented by trees rooted at the extant organism, branching for each contributor of genetic material. For the metrics I present here, I limit discussion to asexual populations; however, I suggest three approaches for generalizing these lineage metrics to sexual populations:

1. **Build lineages based on sites in a genome**. For genetic representations composed

of multiple constituent parts that are inherited atomically (*i.e.* sites in a genome), the complication of sexual lineages can be avoided by tracking the lineages of these parts. Genomic sites will only ever have a single parent, so their lineages are effectively asexual. An organism could then be viewed as a collection of lineages rather than a single one.

2. **Apply a lossy compression to reduce sexual lineages into linear sequences**. By modeling sexual reproduction events as asexual reproduction events, it is possible to compress sexual lineages into linear sequences of states. One parent is designated to be a part of the lineage and the genetic contributions of other parents are considered to be sources of genetic variation (mutations). The primary downside to this approach is its lossy-ness (*i.e.*, the fact that it discards potentially important parentage information).

3. **Extend the metrics to operate over non-linear (tree) structured lineages**. Alternatively, we can extend these metrics to operate over the more complex state sequences that constitute the lineages of sexually reproducing organisms. One such approach would be to consider all possible ancestor paths for an extant individual, calculate a given metric for each of them and then average the resulting values together.

Assessing the efficacy of these and potentially other approaches would be a useful line of research to pursue in the future.

8.2.1.1 Lineage Length

Lineage length describes the number of states traversed by a lineage. If a state is defined as a single individual, lineage length is a count of the number of generations. Generation count is most useful in systems where generational turnover is not fixed, but instead determined by the life history strategies of organisms. For lineages that span equal lengths of time, more generations imply faster replication rates (*e.g.*, r-selected lineage) while fewer generations imply slower replication rates (*e.g.*, K-selected lineage).

Lineage length becomes a more flexible and informative metric if we consider more abstract definitions of states along a lineage. We might measure lineage length where a state represents a sequence of individuals that share a particular phenotypic or genotypic characteristic. In these cases, lineage length only increases when the characteristic of interest changes from parent to offspring. For example, in an environment where organisms must perform functions to be successful, we might define state as the set of functions performed by an individual. In this scenario, lineage length would only increase when the set of functions performed by an ancestor changes; sequential ancestors that perform the identical sets of functions would be compressed into a single state in the sequence, even if other traits differ.

8.2.1.2   Mutation Accumulation

Mutation accumulation defines a set of measurements that track mutational changes across a lineage. These changes can be measured as the magnitude of the change (for real-valued genomes) or as the total count of changes (for discrete-valued genomes). Mutation effects can also be tracked to gain insights about their distribution along a given lineage. Measures of mutation accumulation along the lineages of successful individuals can help tease apart the relative importance of different types of mutational events when compared to what is expected by chance.

In conjunction with collected fitness information, the class of a mutation (*e.g.*, beneficial, deleterious, or neutral) can also be tracked. Different evolutionary conditions are expected to cause different distributions of mutations along a lineage (Barrick and Lenski, 2013); deviations revealed by measures of mutation accumulation can act as a barometer for unexpected evolutionary dynamics. The number and magnitude of deleterious mutations along a lineage can tell us both about the ruggedness of the fitness landscape, and about a lineage's ability to cross fitness valleys (Covert et al., 2013). Similarly, an elevated measure of neutral mutations relative to beneficial or deleterious mutations can suggest that the fitness landscape has neutral space that the lineage is spending most of its time drifting around.

155

8.2.1.3 Phenotypic Volatility

Phenotypic volatility addresses the rate at which phenotype changes as you move down a lineage (although the same concept can be applied to specific phenotypic traits or other types of state). In systems with discrete/categorical phenotypes, this can be measured by summing the number of times the phenotype changes. A related but subtly different measurement in such systems is the number of unique phenotypes on a lineage. In most cases, these values will be similar; a discrepancy would suggest that the lineage was cycling through a set of phenotypes. Such behavior could, for example, be indicative of some form of evolutionary bet-hedging (Beaumont et al., 2009).

In systems with continuous-valued phenotypes, a subtly different approach is needed to measure phenotypic volatility, because there are no discrete state transitions. Instead, we can measure the overall variance in phenotype along a lineage. In some cases, it may be desirable to smooth out the noise inherent in a real-valued phenotype. We can do so by instead taking the variance of the moving average of fitness, to more closely approximate the idea of measuring phase transitions.

8.2.1.4 Summary Statistics

Each of these metrics can be calculated for each member of the population at each time step. Doing so, however, would produce an amount of data so large that it would be difficult to make sense of. Instead, we need to come up with ways to generate useful summaries. There are two main approaches to doing so: 1) choose a small number of representative lineages from a given time point, or 2) collect summary statistics about the distribution of metric values across the population.

A single lineage can be chosen by selecting the lineage of a representative organism (either the most fit or the most numerous). In populations where diverse strategies coexist, this approach can be uninformative as any one lineage is unlikely to be representative of all successful lineages. One alternative is to filter out lineages that do not have offspring some

predetermined number of generations later as such lineages were likely not representative of an important subset of the population. Still, any approach based on measuring only a subset of lineages can be challenging to interpret when the current dominant lineage (or lineages) is replaced with a different one; such changes can introduce a discontinuity if the value is being measured over time. If graceful responses to changes in which lineage is dominant are required, it can be advantageous to instead measure summary statistics (*e.g.*, mean, variance, and range) across the entire population.

In scenarios with frequent selective sweeps, the dominant lineage will likely be similar to the average lineage, as most of the population will be closely related. When the population contains more phylogenetic diversity, however, the dominant lineage may differ from the mean. Of course, the nature of such differences is likely informative about the evolutionary dynamics occurring in the population.

8.2.2   Phylogeny Metrics

These metrics operate on entire phylogenies rather than single lineages within a population, eliminating the need to identify a representative organism or lineage. Because they use data from the entire population, phylogeny metrics can be more computationally expensive to calculate than single lineage metrics. On the other hand, because most lineages tend to share substantial history, phylogeny metrics can usually be calculated more rapidly than full-population lineage metrics. Note that phylogenies can be constructed with regard to any taxonomic level of organization, be it individual, genotype, phenotype, *et cetera*. Thus, when I refer generally to items in a phylogeny, I will use the term *taxa*.

A standard technique for saving memory and time when working with phylogenies in computational systems is to "prune" them, removing dead (extinct) branches. Since all of the phylogeny metrics I discuss here are borrowed from natural systems (where we do not have information about taxa without offspring), they all are designed to work on pruned phylogenies. Thus, for the remainder of this chapter, I will assume we are working with

pruned phylogenies.

In populations without ecological forces promoting coexistence, phylogenies should coalesce periodically, resulting in pruned lineages that mostly consist of a single path. When there is strong selection, this coalescence should happen even more rapidly. Thus, phylogenies with topologies that deviate from that expectation are an indication of ecological interactions within the population. The metrics discussed here can provide insight into the nature of those interactions and their long-term evolutionary effects (Dolson et al., 2018d). As a result, they are often referred to as phylogenetic diversity metrics (Tucker et al., 2017).

An important distinction between phylogenies in natural versus computational systems is that natural phylogenies are generally inferred from extant taxa, whereas computational phylogenies are directly recorded. Inferred phylogenies do not contain internal nodes except at branch points. They also do not contain history prior to the most recent common ancestor (MRCA) of all extant organisms. For consistency, I exclude pre-MRCA taxa from these analyses. However, I will not remove non-branching internal nodes, as these only serve to make the phylogenies more informative.

Here I provide a high-level summary of phylogeny metrics that I expect will be particularly useful. For more metrics and more detail on all of these metrics, see (Winter et al., 2013; Tucker et al., 2017).

8.2.2.1 Depth of Most-Recent Common Ancestor

The depth of the MRCA (*i.e.*, the number of steps it is from the original ancestor) is an informative metric and is easy to calculate. A recent MRCA implies frequent selective sweeps and less long-term stable coexistence between clades. Measuring the frequency with which the MRCA changes (*i.e.*, the number of coalescence events) can also be informative, as some conditions can inflate the length of the lineage relative to other conditions without actually increasing the frequency of selective sweeps. This scenario is particularly likely when the population size is changing over time. A downside to the depth of MRCA as a metric is

158

that any population that does have a stable ecology will likely never change its MRCA after the very beginning of evolution (which at least allows us to detect stable coexistence in the population).

8.2.2.2 Phylogenetic Richness

Measurements of phylogenetic richness quantify the total amount of evolutionary history contained in a set of taxa. The most traditional metric of phylogenetic richness is "Phylogenetic Diversity", which is calculated as the number of nodes in the minimum spanning tree from the MRCA to all extant taxa (Faith, 1992). Another approach is to calculate the pairwise distances between all taxa and sum them (Tucker et al., 2017). A third approach is to sum evolutionary distinctiveness, a measurement of a taxon's evolutionary uniqueness (Isaac et al., 2007), across all extant taxa (Tucker et al., 2017).

8.2.2.3 Phylogenetic Divergence

Measurements of phylogenetic divergence quantify how distinct the taxa in the population are from each other and are often averaged across individual taxa. For example, one option is to average the pairwise distances across all taxa in the population (Webb and Losos, 2000). Similarly, phylogenetic divergence can be calculated by averaging the evolutionary distinctiveness across each taxon in the population.

8.2.2.4 Phylogenetic Regularity

Measurements of phylogenetic regularity quantify how balanced the branches are in a phylogeny and are often the variances of values calculated for individual taxa. Just as the mean of the pairwise distances between all taxa in the population is a measurement of phylogenetic divergence, taking their variance is a measurement of phylogenetic regularity. The same is true of the variance of evolutionary distinctiveness across the population.

159
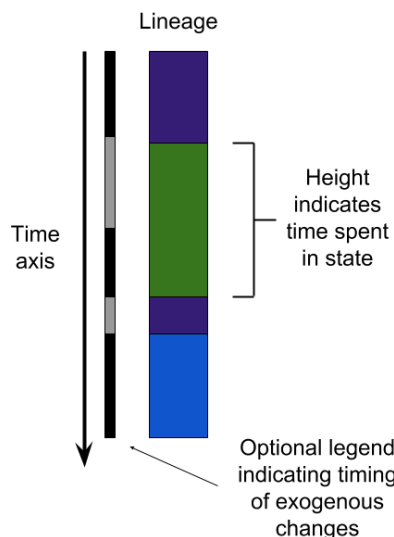
Figure 8.2: **Schematic representation of state sequence visualization.** Colors indicate different states. The optional legend along the side can be used to indicate any information relevant to understanding the drivers of the state changes.

8.3   Visualizations

Data visualization is a critical component of data analysis. Summary measurements can be misleading (Anscombe, 1973), but we often have too much data to digest from purely numerical values. Visualizations allow us to structure data in a way that is more intuitive to the viewer. Through this lens, we can confirm that our intuitions about the drivers of a system's behavior are correct. In this section, I will present some visualization techniques that are useful for understanding lineages and phylogenies. These visualizations have the power to substantially strengthen our intuitions about evolutionary systems. Later, I will use these visualizations to ensure that these metrics are capturing the information intended.

The idea that any lineage can be abstracted into a sequence of states (see Figure 8.1) is important for visualization as well as quantitative measurements. We must be careful about choosing the right level of abstraction; using too low a level can obscure broader patterns, whereas too high a level loses the pattern entirely. Furthermore, generating some of these visualizations at all becomes intractable with too little abstraction.
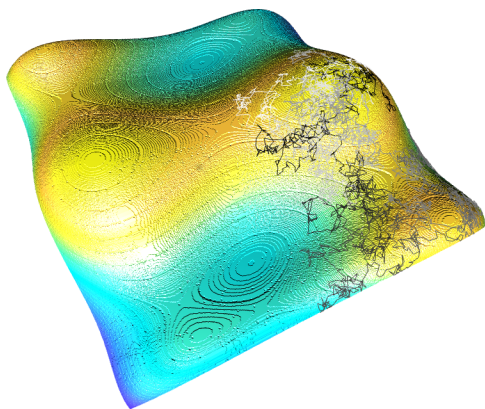
Figure 8.3: **A lineage drawn on top of a fitness landscape**

8.3.1 State Sequence Visualizations

A direct consequence of the fact that lineages are just sequences of states is the fact that we can visualize these sequences (Lalejini and Ofria, 2016). Each sequence can be represented as a series of stacked rectangles. The colors of these rectangles indicate which state they represent and their heights represent how long the lineage stayed in that state (see Figure 8.2). Placed side-by-side, these sequences provide an aggregate visual representation of the behavior of lineages from a given run or experiment.

8.3.2 Fitness Landscape Overlays

In order to understand how a particular evolutionary algorithm performs on a given problem, we can observe how that algorithm moves a population across the problem's fitness landscape (Kauffman and Levin, 1987). A fitness landscape can be thought of as a map from solution representations into the quality of that solution (its fitness) (Wright, 1932). Traditionally, fitness landscapes have most commonly been visualized as three-dimensional surface plots illustrating the effects of two continuously-varying traits on fitness (*e.g.*, Li et al., 2013, or Figure 2.1 of this chapter). One of these traits is the x-axis, one is the y-axis, and the fitness conferred by that combination of values is the z-axis. Although I will focus here on fitness landscapes that can be fully depicted in three dimensions, it is important

to recognize that most real-world fitness landscapes have far more dimensions. Various dimensionality-reduction techniques can be applied to reduce them to three dimensions, but doing so may be misleading (Gavrilets, 2010).

We can visualize the path that each lineage takes through the fitness landscape, mapping the x, y, and z (fitness) coordinates of each ancestor of each member of the population (Virgo et al., 2017). In some cases, it may also be helpful to overlay lineages on top of other spaces (Dolson and Ofria, 2017; Mouret and Clune, 2015). Creating such a visualization entails condensing a large quantity of information into a limited space. When projected onto two dimensions, lineages can obscure and be obscured by parts of the fitness landscape (and each other).

Fortunately, we are entering an era where we no longer need to squeeze this information into two dimensions, confined to a computer screen or piece of paper. Over the past few years, virtual reality technology has advanced to the point where it is a viable tool for data visualization (van Dam et al., 2002; Olshannikova et al., 2015). To this end, I used the A-Frame framework (A-Frame authors, 2018) to build a three-dimensional data visualization that can be viewed in virtual reality (see Figure 8.6) (Dolson and Ofria, 2018).

A-Frame supports building 3D scenes and rendering them to a variety of platforms. In the simplest case, the visualization is rendered in WebGL, allowing it to be viewed in a standard web browser. Mouse interactions such as rotation make it possible to view the visualization from all angles, and WebGL's use of the graphics card allows it to render data-rich visualizations. A-Frame also supports rendering the page with WebVR, allowing it to be viewed using various virtual reality headsets. These platforms allow the user to explore the data in three dimensions. For the data interpretation in this chapter, I used an Oculus Rift to provide myself with fine-grained control of which part of the visualization I was looking at.

Scenes in A-Frame are built by layering components on top of each other. The base component of this visualization is a three-dimensional surface plot of the fitness landscape.

On top of this layer, I overlay a three-dimensional path from an individual that existed at the end of an evolutionary run all the way back to its earliest ancestor, passing through the locations of all intermediate ancestors along the way. A path can be added for each extant member of a final population to produce an entire overlaid phylogeny.

To make visualizations more interpretable, I introduced a color gradient along the lineage; in the visualization presented here, I use a grayscale lineage drawn onto a colorful landscape (see Figure 8.3). Lineages transition from white to black as evolutionary time progresses, indicating when each portion of the lineage existed.

Our full visualization, complete with data, can be viewed on the web or using a virtual reality headset at `https://emilydolson.github.io/fitness_landscape_visualizations`.

8.3.3 Phylogenetic Trees

Phylogenetic trees are a time-honored data visualization in biology. They depict the parent-child relationships of taxa over the course of evolution (see Figure 8.4A). As with phylogeny metrics, memory usage is a primary concern in constructing these visualizations, as phylogenies can be very large. Applying pruning and choosing more abstract taxonomic units (as discussed in the phylogeny metrics section) are the best ways to mitigate this problem.

While simply drawing the phylogenetic tree is valuable in its own right, a lot of additional information can be added to these visualizations. Nodes in the tree can be colored based on features of the corresponding taxon, such as phenotype or fitness. Edges can be colored based on information about the transition between taxa, such as the types of mutations that occurred or the direction of fitness change. Note, though, that in large trees coloring nodes and edges the same way often makes patterns easier to see.

Another way to convey additional information is in the placement of nodes. In biology, the length of edge between nodes is generally used to indicate the amount of time that elapsed between the origin of taxa. This convention also results in contemporaneous nodes
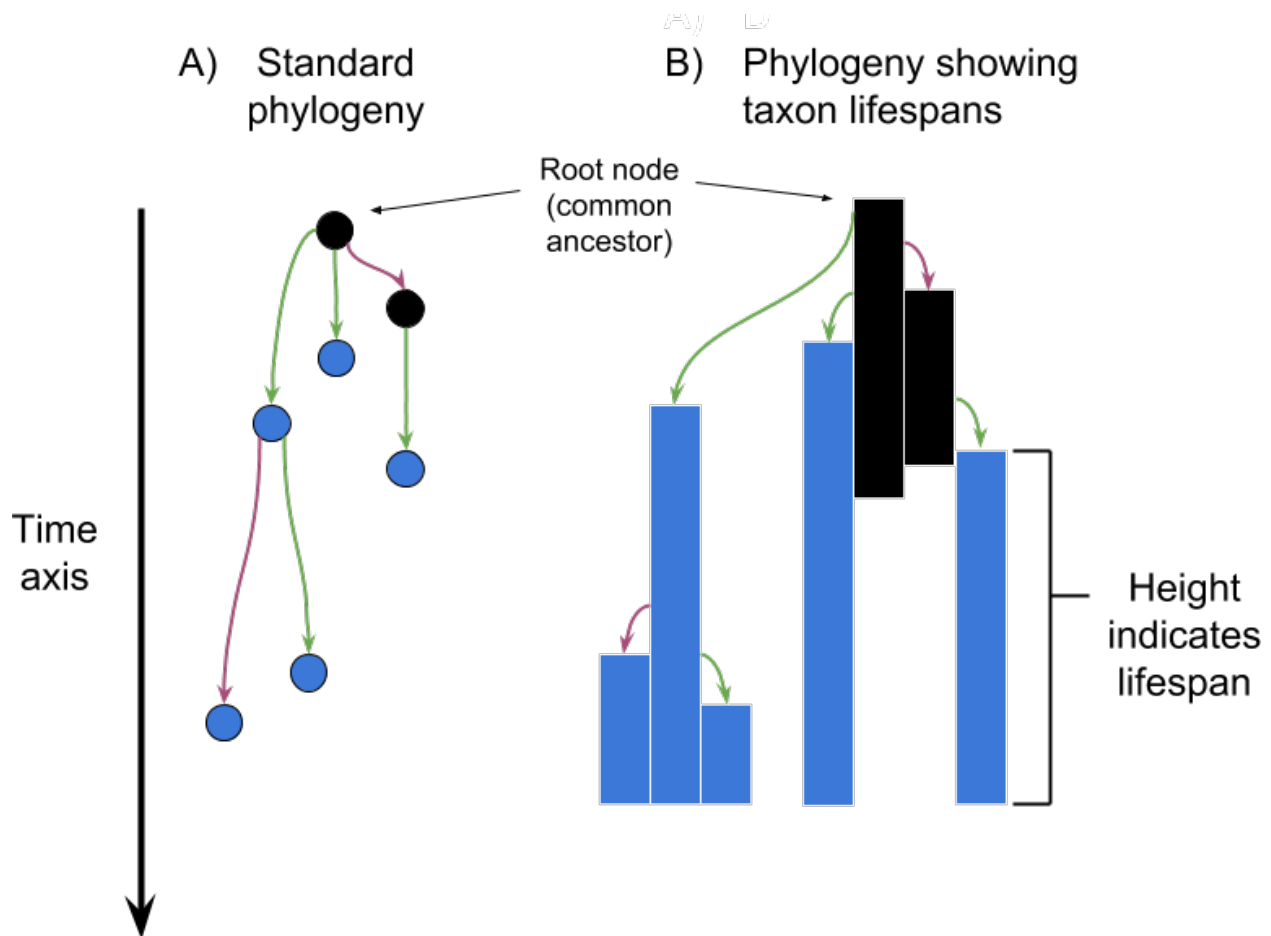
Figure 8.4: **The same phylogeny depicted in two different ways.** A) Depicts a standard phylogeny, where nodes represent taxa and edges indicate parent-child relationships. In this phylogeny, node position along the time axis represents time of birth. Nodes are colored to distinguish taxa that are extant in the present (blue) from extinct taxa (black). Note that, because this is a pruned phylogeny, all leaf nodes are currently extant. Edges can be colored to convey whatever information is most useful. B) Depicts the same phylogenetic tree but with boxes indicating taxon lifespans in place of circular nodes.

being placed at the same level of the tree; in essence, the tree has a time axis running from top to bottom. In Artificial Life systems, we can take this idea a step further. Since we know the complete time span that a taxon existed for, we can depict it on the phylogenetic tree by plotting rectangles in place of circular nodes (see Figure 8.4B). The tops and bottoms of these rectangles correspond to the birth and death times of the taxon along the time axis. While this approach runs the risk of producing an overly-complicated plot, the boxes can yield wildly different intuitions than circles would.

8.3.4 Muller Plots

Muller plots show the abundance and ancestry of taxa in a population over time (Muller, 1932; Maddamsetti et al., 2015). The percentage of a population occupied by each taxon at a given time point is depicted via a stacked-area plot – the vertical space of the plot is occupied by a stack of colored polygons, each with height proportional to the relative abundance of a corresponding taxon (see Figure 8.5). These polygons are connected over time (as in a streamgraph), showing the changes in each taxon's abundance. Additionally, each taxon's polygon is depicted as emanating from its ancestor's polygon. Muller plots in this chapter were made using the ggmuller package (Noble, 2017) for the R Statistical Computing Language (R Core Team, 2017).

When should you use a Muller plot versus a phylogenetic tree? In general, Muller plots are more useful when you care about the population dynamics of a relatively small number of fairly abstract taxonomic units. Phylogenetic trees can accommodate a larger number of less abstract taxonomic units and allow for a greater flexibility in depicting evolutionary patterns.

8.4 Experimental Design

I applied the metrics defined above to four benchmark functions from the GECCO Competition on Niching Methods (Li et al., 2013) and to four qualitatively different environments in Avida. Both of these evolutionary contexts are well understood and studied, making them particularly well suited for building intuition about what this proposed suite of ancestry-based metrics can tell us about evolutionary dynamics.

8.4.1 Niching Competition Benchmark Problems

For each test problem, I evolved populations of 1000 organisms under a range of mutation rates and selection strengths for 5000 generations. I initialized populations by randomly
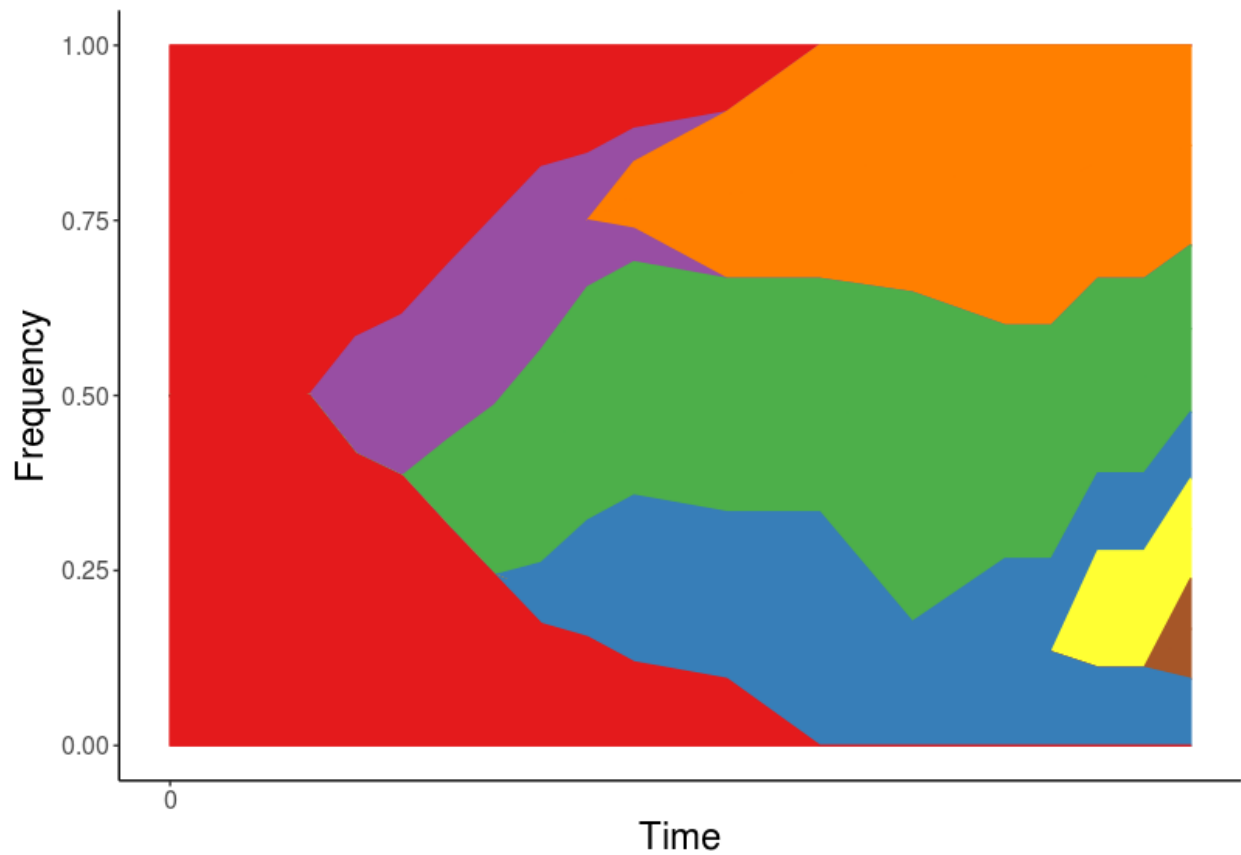
Figure 8.5: **A Muller plot depicting the same phylogeny as the visualizations in Figure 8.4.** The red region represents the root node. As it gradually goes extinct, the proportion of the figure it takes up gradually diminishes. Its three offspring are shown in purple, green, and blue.

generating a number of organisms equal to the population size. I evolved populations under tournament selection with five different tournament sizes: one, two, four, eight, and sixteen. Organisms selected to reproduce did so asexually. Values in an offspring's genome were mutated by adding noise given by a normal distribution with a mean of 0; the 'mutation rate' of a treatment defined the standard deviation used to define this normal distribution and was given as a proportion of the test problem's domain. I prevented mutations from causing a value to exceed the valid domain of the given problem. For each problem and tournament size, I evolved populations at eight mutation rates: 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02, and 1e-01.

I also ran a second set of experiments using these benchmark problems to explore the impact of ecological dynamics on this suite of ancestry-based metrics. For these experiments, I generated a stable ecology using the Eco-EA algorithm as a selection technique (Goings et al., 2012). In these test problems, I created niches associated with spatial locations across the fitness landscape. For all experimental conditions, I ran ten replicates, each with a unique random number seed. This experiment is implemented using the Empirical library; the implementation (and analysis scripts) is included in the supplemental material for the paper this chapter is based on (Lalejini et al., 2018).

8.5 Results and Discussion

8.5.1 Niching Competition Benchmark Problems

Overall, my results were consistent with evolutionary theory. As mutation rate increases, coalescence takes longer, as evidenced by the fact that the MRCA is farther back in time at higher mutation rates (see Figure 8.7). Consequently, phylogenetic richness (as measured by phylogenetic diversity) is higher at high mutation rates. Phylogenetic divergence, measured here as mean pairwise distance between taxa, is similarly higher at high mutation rates. Evolutionary distinctiveness, being another measurement of phylogenetic divergence,
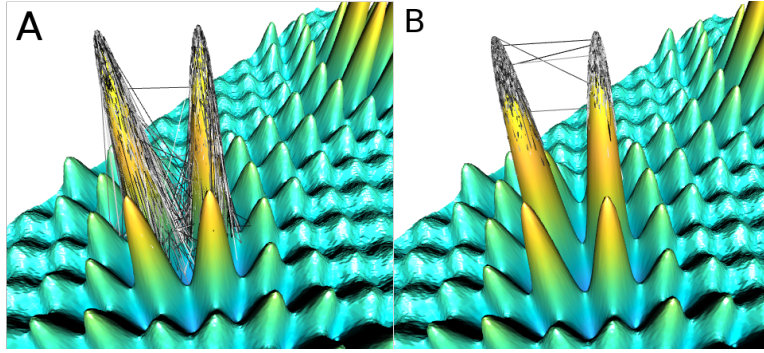
Figure 8.6: **A close-up on two adjacent peaks in the Shubert function fitness landscape.** Lineages are depicted as paths fading from white to black over evolutionary time. The lineages shown here evolved under a mutation rate of 0.01. A) Was evolved using a tournament size of 2, whereas B) was evolved using a tournament size of 16. These figures neatly illustrate how increased tournament size keeps the lineage near the tops of the peaks.

behaved almost identically (Lalejini et al., 2018). Variance of evolutionary distinctiveness and pairwise distance between taxa (phylogenetic regularity metrics) behaved similarly to the phylogenetic divergence metrics. This pattern makes sense, as most phylogenetic divergence on these landscapes will produce unbalanced phylogenetic trees. If there were stable coexistence between multiple clades, we would expect to see a reduced correlation between the phylogenetic divergence metrics and the phylogenetic regularity metrics. Increased mutation rate also increases the number of deleterious steps taken, a logical consequence of increasing mutation relative to strength of selection.

Similarly, increasing tournament size generally increases the rate of coalescence, as higher tournament sizes correspond to stronger selection (see Figure 8.8). As a result, all of the measurements of phylogenetic richness and divergence decrease as tournament size increases. MRCA depth, on the other hand, increases, directly reflecting the increased frequency of selective sweeps.

Surprisingly, there is no clear effect of tournament size on the count of deleterious steps along the dominant lineage (as evidenced by the fact that the confidence intervals all overlap). Values for all selection schemes and tournament sizes hover near 2500, meaning that a deleterious step is taken in roughly half of the 5000 generations. This result is partially
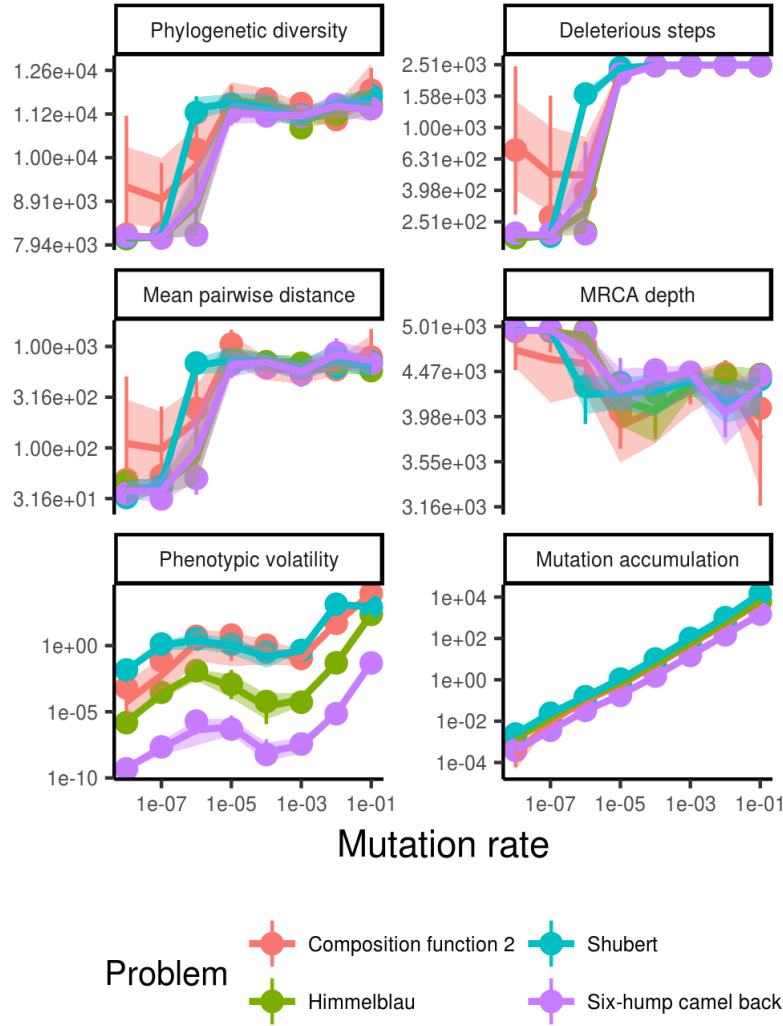
Figure 8.7: **Values of example metrics across different mutation rates for each of the four problems.** All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All experiments shown here used a tournament size of 4. Circles are medians, vertical lines show inter-quartile range, and shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.
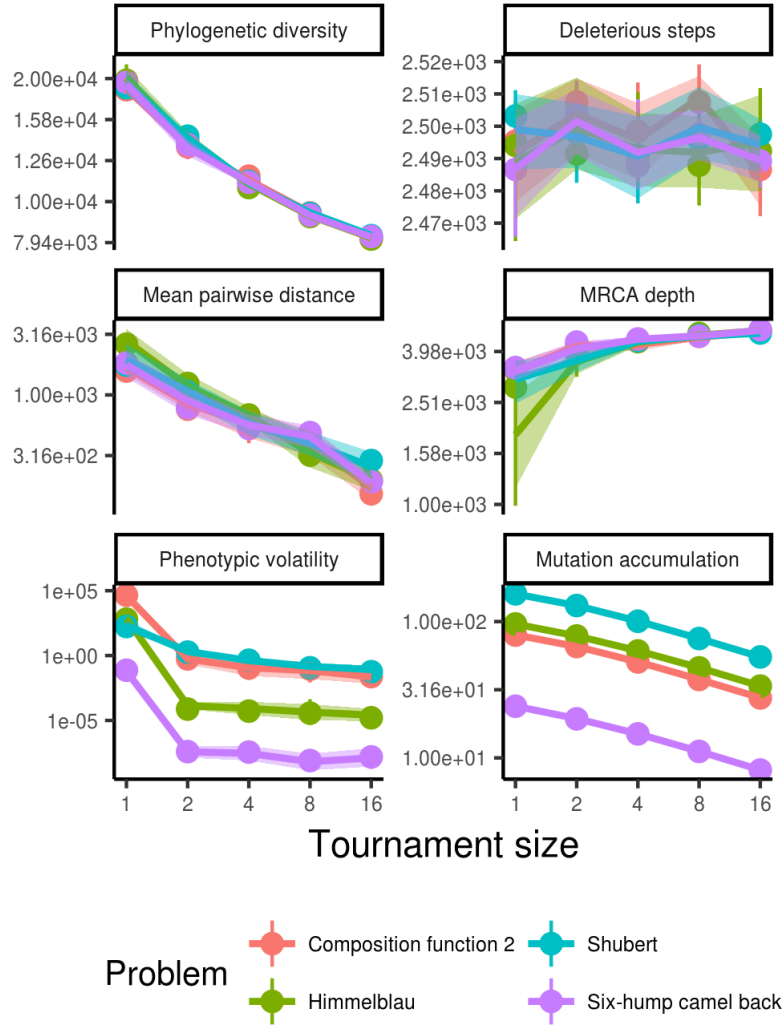
Figure 8.8: **Values of example metrics across different tournament sizes for each of the four problems.** All experiments shown here used a mutation rate of 0.001. All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. Circles are medians, vertical lines show inter-quartile range, and shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.
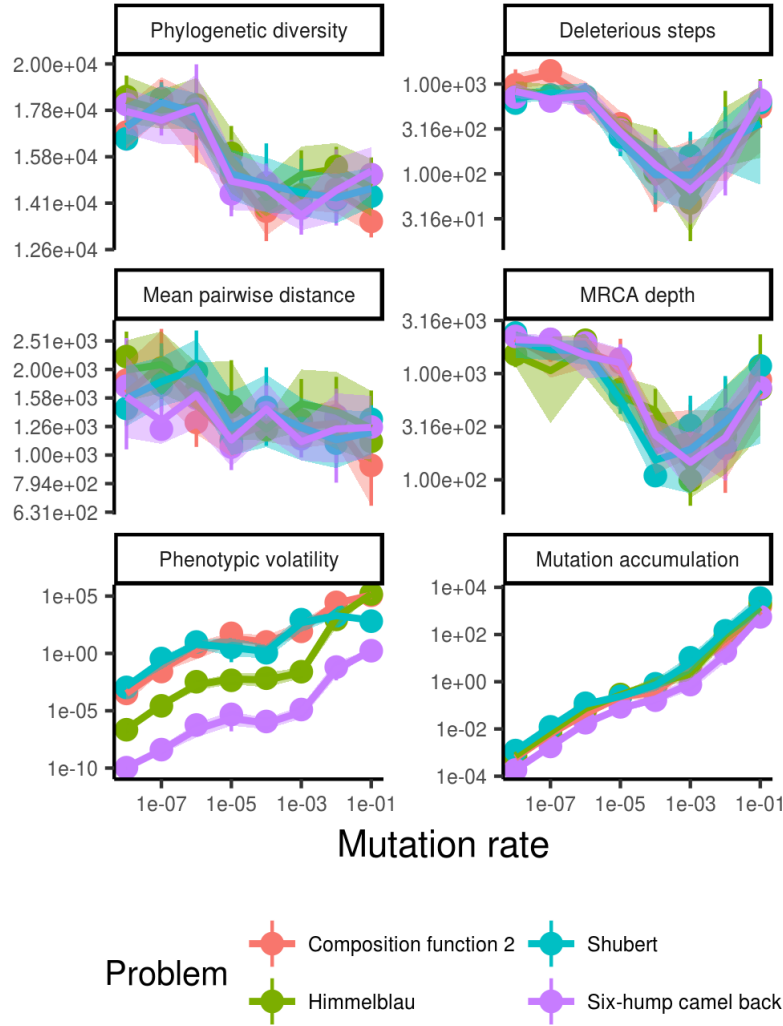
Figure 8.9: **Values of example metrics across different mutation rates for each of the four problems under a diversity-preserving selection regime, Eco-EA.** All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All experiments shown here used a tournament size of 4. Circles are medians, vertical lines show inter-quartile range, and shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.

an effect of mutation rate; at the lowest mutation rate, there is a clear trend toward fewer deleterious steps as tournament sizes increase (Lalejini et al., 2018). However, the effect of mutation rate on the relationship between tournament size and dominant deleterious steps is complex, particularly for Composition Function 2 (Lalejini et al., 2018). These trends likely share a common cause with the thresholding effect evident in Figure 8.7, where the number of deleterious steps along the dominant lineage abruptly climbs between mutation rates of 10^{-7} and 10^{-5} and remains relatively flat over other mutation rates. Based on an inspection of the 3D fitness landscape visualizations, we can see that this is not an effect of lineages moving from peak-to-peak; at most mutation rates, they tend to remain on a single peak (with the exception of adjacent peaks in the Shubert function; see Figure 8.6). Thus, we can infer that this effect is the result of a drift-like phenomenon where, at sufficiently high mutation rates, all members of the population are constantly somewhat displaced from their local fitness peak.

Having reinforced our intuition about these metrics in a simple system, we can now expand them to a slightly more complex system. A large proportion of interesting short-term evolutionary dynamics relate to interaction between individuals in the population (*i.e.*, ecological dynamics). In particular, such interactions often promote the stable coexistence of clades occupying different niches. As such, it is important to establish a baseline for how these metrics respond to ecological coexistence.

Indeed, the presence of stabilizing ecological dynamics substantially changes the values I observe for most metrics (see Figure 8.9). Perhaps the least surprising of these is MRCA depth is far lower than it was for tournament selection, reflecting the rarity of coalescence events under these conditions. Consequently, phylogenetic diversity is higher, as the extant population represents a greater amount of evolutionary history. Relatedly, mean pairwise distance among extant taxa is higher in the presence of ecology, as clades in different niches continue to diverge. Interestingly, the relationship of many metrics (*e.g.*, deleterious steps and phylogenetic diversity) to mutation rate is reversed in the presence of ecology. Explaining
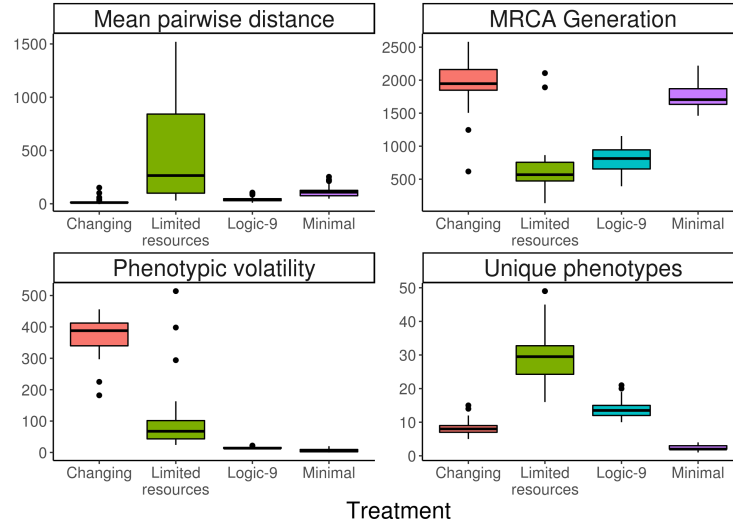
Figure 8.10: **A sampling of informative lineage metrics from Avida** calculated after approximately 10,000 generations. Mean pairwise distance is a measurement of phylogenetic divergence; higher values imply greater phylogenetic diversity. MRCA generation is the generation at which the current most recent common ancestor occurred (lower numbers mean it was longer ago). Phenotypic volatility is the number of changes in phenotype along the dominant (most plentiful) lineage. Unique phenotypes is the count of unique phenotypes that occurred along the dominant lineage.

the underlying mechanisms behind these distinctions is beyond the scope of this chapter, but the ease with which the metrics identified their presence clearly indicates their power.

8.5.2 Avida

The data from Avida are also consistent with my expectations and additionally point towards some interesting directions for future investigation. As predicted, the only environment that preserved phylogenetic diversity was the limited resources environment. Lacking diversity-preserving dynamics, the other environments all repeatedly lose phylogenetic diversity due to selective sweeps (see Figure 8.10).

Looking at phenotypic volatility, I see that the changing environment condition has dramatically higher volatility than the others. This result is precisely what I hypothesized, as the phenotype should change approximately every time the environment does (see Figure 8.10). Indeed, from Figure 8.13, we can see that this predicted mechanism is accurate. The limited
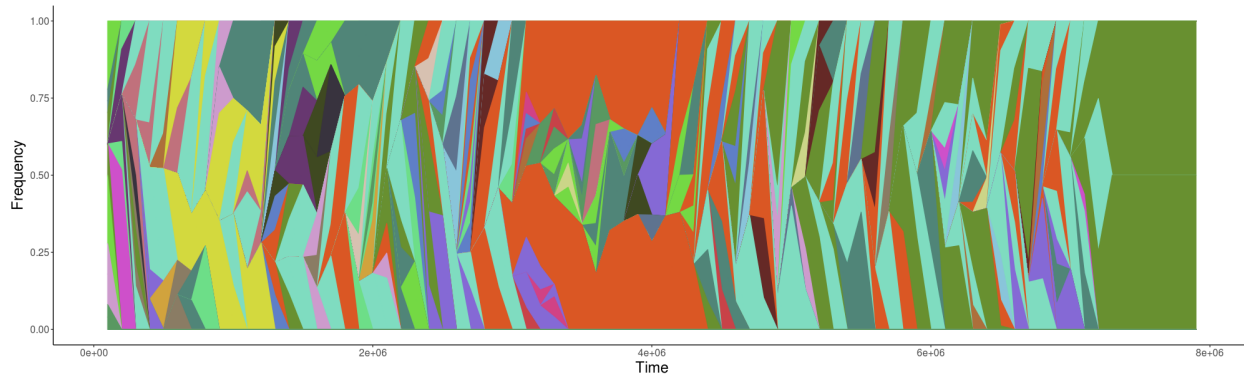
Figure 8.11: **Muller plot from a representative run of Avida in the limited resources environment.** This plot shows the population dynamics of different phenotypes over 8,000,000 updates (a unit of time in Avida that is proportional to the number of CPU cycles used thus far). Only phenotypes that make up at least 5% of the population at at least one time point are shown. Note the region in the middle of the plot where we can see repeated selective sweeps that are contained to a single niche.

resources environment has lower phenotypic volatility than the changing environment, but higher phenotypic volatility than the other two environments. This observation is consistent with the fact that limited resources is also a changing environment; competitive pressures shift with the composition of the population. This interpretation is corroborated by Muller plots from the limited resources condition, which show frequent shifts in the abundance of various phenotypes (see Figure 8.11). We can gain further insight by comparing phenotypic volatility to the count of unique phenotypes along a lineage. Despite its high phenotypic volatility, the changing environment condition produces lineages with relatively few unique phenotypes. This finding makes sense, as the changing environment is only creating pressure to toggle back and forth between two phenotypes. Limited resources, however, has a high count of unique states relative to its phenotypic volatility. From this comparison, I can conclude that the competition in this environment is creating pressure for the population to explore new phenotypic states, rather than simply cycling among a fixed set.

From the mutation accumulation data, we can begin to glean information about the precise dynamics of adaptation (see Figure 8.12). Across the board, the limited resources environment has more mutations than any other condition, which is consistent with the

174
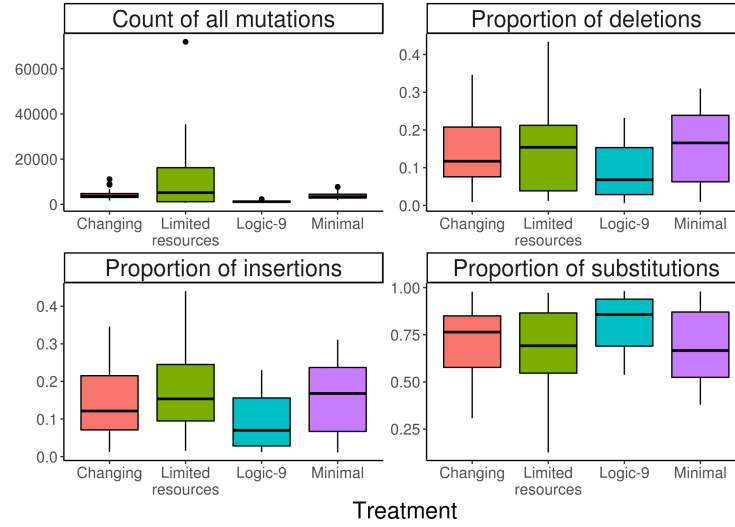
Figure 8.12: **Mutations along the dominant lineage across treatments in Avida,** calculated after approximately 10,000 generations. The upper left shows the total count of mutations of all types. The other plots show the quantity of each mutation type as a proportion of the total.

greater number of unique phenotypes explored. The changing environment has the second greatest mutation accumulation, presumably due to the mutations that are required to change phenotype as the environment changes. If we break down mutation accumulation by type, we can see that approximately 75% of mutations along dominant lineages are substitution mutations. This value is dramatically higher than the approximately 7% that we would expect to see by chance. Thus, this result suggests that substitution mutations are more likely to prove helpful in these contexts over evolutionary time.

8.6   Conclusions

I have demonstrated that these analysis techniques conform to my intuitions in well-understood systems. In more complex environments, they provide a quick way to identify interesting behavior to study further. We believe they are now ready to be adopted more broadly.

Our goals for this work are two-fold: 1) to suggest a set of analyses that will improve
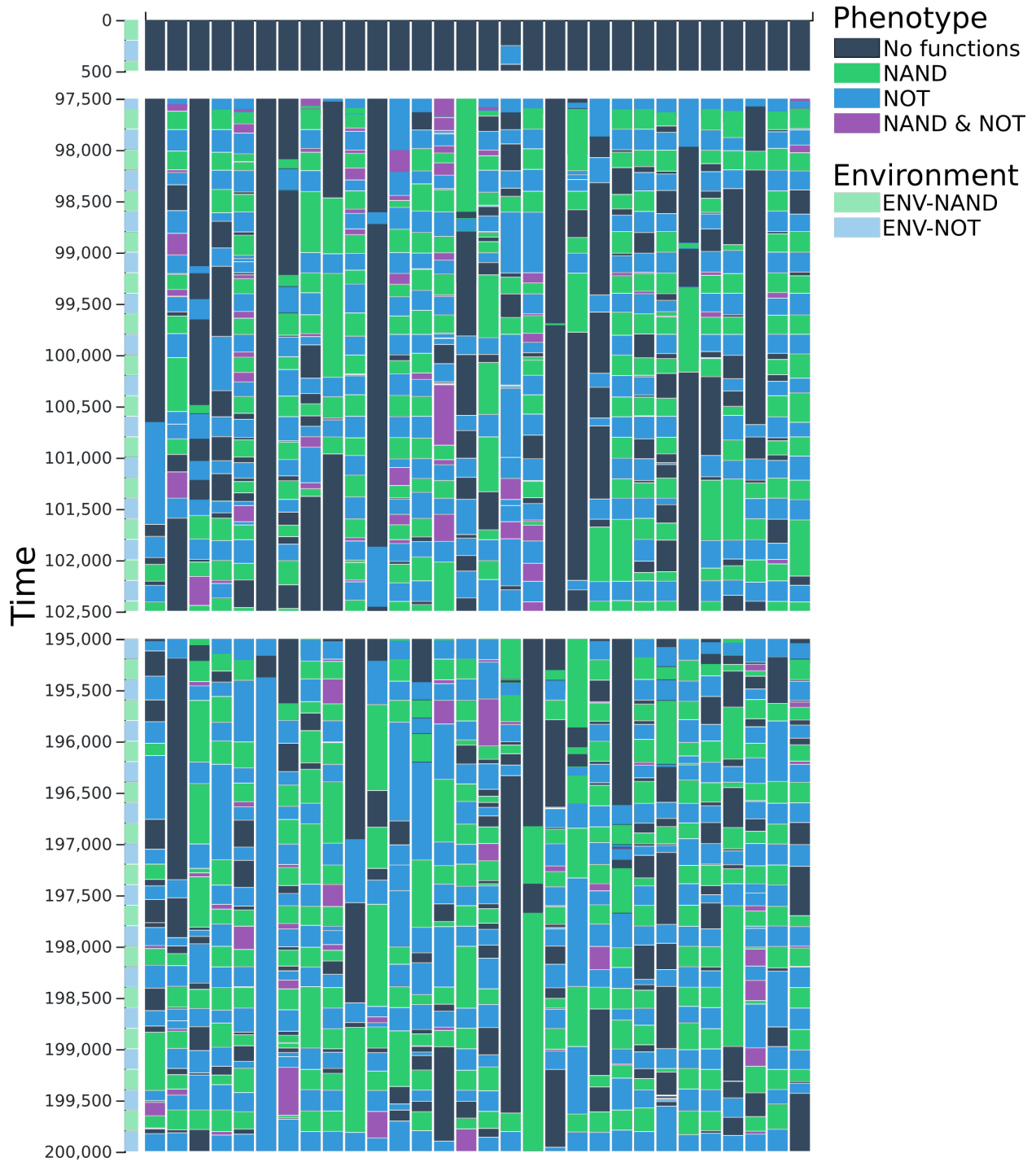
Figure 8.13: **State sequence visualization of phenotypes along the dominant lineage in the changing environment condition.** The key along the left indicates the environment's state over time. Each column is one replicate. From this visualization, we can see that most successful lineages are able to adapt to a change in the environment relatively quickly. We can also see that there is a fair amount of variation in the precise timings of these changes and whether they occur for every change of the environment.

our capacity to quantitatively understand evolutionary histories in digital evolution experiments, and 2) to spark a conversation in the computational evolution community about how to quantify, interpret, and compare observed evolutionary histories. With feedback from the community, I will expand this suite of lineage and phylogeny metrics, compiling accessible descriptions and examples of each metric. More generally, I suggest that artificial life researchers should keep phylogeny- and lineage-based analyses in mind when trying to answer specialized questions as well.

Chapter 9

Conclusions

Just as evolution is a general purpose algorithm that applies across substrates, so is ecology. I have illustrated this point by comparing ecological dynamics across a range of digital systems. In the process, I have developed tools for further exploring, understanding, and controlling eco-evolutionary dynamics. At the same time, I have addressed fundamental questions across the fields of biology, evolutionary computation, and artificial life.

9.1 Contributions

This dissertation makes the following contributions:

- In chapter 3, I demonstrated that some types of spatial heterogeneity promote the evolution of diversity and produce populations with elevated evolutionary potential. I have proposed an explanation for why this finding differs from the results of similar studies.

- In chapter 4, I showed that the spatial configuration of niches has a profound impact on the way a population traverses a fitness landscape. As a result, certain region of a spatially heterogeneous environment can be hotspots of evolutionary potential. This finding opens up the possibility of manipulating evolutionary trajectories via the

spatial configuration of the abiotic environment.

- In chapter 5, I illustrated the power of importing ecological theory into evolutionary computation. Indeed, I discovered that the same co-existence equation has been independently derived in both fields, lending credence to the idea that it generalizes across all ecological systems. Moreover, I used ecological techniques to clarify the relationship between evolutionary algorithms that seem similar at face value but turn out to behave very differently on a mechanistic level.

- In chapter 6, I presented a new approach to evolutionary computation that is made possible by ecology-inspired techniques: ecologically-mediated hints. By encoding hints in the form of niches, the user of an algorithm can create a fitness gradient towards a global solution without many of the pitfalls traditionally associated with attempts to do so. Additionally, I explored the specific properties of Eco-EA that make it a good candidate for this approach.

- In chapter 7, I developed a suite of measurements for identifying the presence of open-ended evolutionary dynamics in an evolving system. Chief among these measurements was an approach to recognizing the presence of ecology.

- In chapter 8, I described an additional suite of measurements that I hope to use in the future to provide insight into the lower-level mechanics of the evolutionary process.

9.2 Future Directions

Thus far, I have transferred techniques amongst the fields of biology, evolutionary computation, and artificial life to address open questions within these fields. However, I have barely scratched the surface of what we can learn from work at this interface. The next step is to go beyond transferring techniques between these fields and begin transferring questions. Evolutionary computation has fantastic potential as a model system for studying

179

eco-evolutionary dynamics in the way theoretical biologists traditionally do. Furthering our understanding of how to build complex evolving systems, as artificial life researchers often attempt to do, could expand the classes of problems that are solvable via evolutionary computation. Although research on the drivers of diversity has historically been a question for biologists rather than artificial life researchers, these drivers are likely a critical component of creating open-ended evolution. Cross-cutting all of these fields is the question of how to control evolutionary trajectories of entire ecological communities. This question is relevant in artificial life (where it helps us build life-like systems), biology (where we often benefit from ecological communities being in certain target states), and evolutionary computation (where it helps us solve problems).

9.2.1 Evolutionary Computation

My work has set the stage for a rapid bi-directional transfer of ideas between ecology and evolutionary computation. The first step in this process will be to continue importing ecological theory as I have already done for modern coexistance theory in the context of fitness sharing, resource competition theory in the context of Eco-EA, and spatial structure in the . Research in evolutionary computation often gets hung up on the efficacy of an approach at the expense of understanding why it works. Now that **I have shown diversity maintenance techniques in evolutionary computation to be isomorphic to ecological dynamics** (Dolson et al., 2018d), I will use the decades of work that ecologists have put into developing ecological theory to understand why these techniques work in evolutionary computation.

In this phase, I will take stock of the current landscape of diversity maintenance techniques as viewed through this new ecological lens, locating each in relation to the others. Critically, I hypothesize that ecologically-similar diversity maintenance techniques will behave similarly to each other. I will test this hypothesis by **observing the behavior of different diversity maintenance techniques on an array of benchmark problems**

180

(Helmuth and Spector, 2015; White et al., 2013; Li et al., 2013). To characterize the behavior of diversity maintenance techniques, I will use a combination of interaction networks and evolutionary summary statistics that directly evaluate the underlying evolutionary behavior of the population (Dolson et al., 2018b). Over the course of this phase, I will establish a meaningful taxonomy of diversity maintenance techniques cross-referenced with their performance on various problem classes.

I will also characterize properties of the benchmark problems using a pre-existing suite of metrics (Bosman et al., 2018). Once I have both of these characterizations in place, correlations between them will be apparent. The presence of such correlations would suggest that different ecological communities are better suited to solving different problems. At this point, practitioners of evolutionary computation will, for the first time, have **the ability to choose the best diversity maintenance technique for a specific problem** *a priori*. This ability will tangibly improve the efficacy of all branches of evolutionary computation. In subsequent work, I will delve into the mechanisms that make different techniques more effective on different problems. As this work will be a direct attempt to predict the future evolutionary state of an entire community, it will synergize closely with Aim 1.

9.2.2 Biology

This line of research also has important consequences for biological systems. There are a wide variety of evolving ecological communities that it would be advantageous for us to control the evolutionary trajectories of. In the long run, I hope to apply my findings to cancer, the gut microbiome, and macro-scale ecosystems.

My immediate future plans are to extend my research thus far to the context of cancer, as it is the most direct and easily-studied analog to my digital work. Because cancer readily evolves resistance to new treatments, it would be incredibly valuable to control its evolutionary trajectory. Although the goal in this line of research is to prevent the evolution of complex traits (whereas my goal in other systems is to promote it), achieving this goal

requires answering the same underlying questions.

The abiotic environment of tumors is spatially heterogenous, resulting in the formation of various micro-environments to which cancer cells can adapt. Low-oxygen micro-environments create a distinct ecological niche, the presence of which is associated with the evolution of treatment resistance and other negative outcomes. The spatial configuration of these regions appears to play an important role in determining when treatment will and will not be effective (Scott et al., 2016). I will be extending my work on spatial resource heterogeneity in chapters 3 and 4 and my work on spatial mortality heterogeneity (Dolson et al., 2016b) to understand this phenomenon better. In the long run, this understanding will help us design treatments that prevent cancer from evolving resistance. At the same time, working with a non-digital system will allow me to confirm that my results generalize across evolutionary substrates.

Studying eco-evolutionary dynamics in the gut microbiome (and other microbiomes) is likely more challenging than it is in cancer due because the ecological interactions are more plentiful and difficult to measure. Nevertheless, it is another ecological community evolving in a spatially heterogeneous environment which it would be advantageous to control; there is increasing evidence that certain gut microbial communities lead to better health outcomes for humans. Thus, it represents a promising future direction. Once the general principles behind harnessing the power of ecology to control evolution have been thoroughly untangled and tested across a variety of more tractable experimental systems, they may be a valuable tool for mitigating the collapse of larger-scale ecological communities as result of climate change.

BIBLIOGRAPHY

BIBLIOGRAPHY

A-Frame authors (2018). A-Frame: a web framework for building virtual reality experiences. https://github.com/aframevr/aframe.

Adami, C. (2002). Ab initio modeling of ecosystems with artificial life. *Natural Resource Modeling*, 15(1):133–145.

Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences*, 97(9):4463–4468.

Adler, P. B., HilleRisLambers, J., and Levine, J. M. (2007). A niche for neutrality. *Ecology Letters*, 10(2):95–104.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., and Kievit, R. (2018). Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ Preprints*, 6:e27137v1.

Allouche, O. and Kadmon, R. (2009). A general framework for neutral models of community dynamics. *Ecology letters*, 12(12):1287–1297.

Allouche, O., Kalyuzhny, M., Moreno-Rueda, G., Pizarro, M., and Kadmon, R. (2012). Area–heterogeneity tradeoff and the diversity of ecological communities. *Proceedings of the National Academy of Sciences*, 109(43):17495–17500.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, 27(1):17–21.

Arnell, N. W. (2004). Climate change and global water resources: SRES emissions and socio-economic scenarios. *Global Environmental Change*, 14:31–52.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Banzhaf, W., Baumgaertner, B., Beslon, G., Doursat, R., Foster, J. A., McMullin, B., Melo, V. V. d., Miconi, T., Spector, L., Stepney, S., and White, R. (2016). Defining and simulating open-ended novelty: requirements, guidelines, and challenges. *Theory in Biosciences*, 135(3):131–161.

Barnett, A. and Beisner, B. E. (2007). Zooplankton biodiversity and lake trophic state: explanations invoking resource abundance and distribution. *Ecology*, 88(7):1675–1686.

Barrick, J. E. and Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827.

Baym, M., Lieberman, T. D., Kelsic, E. D., Chait, R., Gross, R., Yelin, I., and Kishony, R. (2016). Spatiotemporal microbial evolution on antibiotic landscapes. *Science*, 353(6304):1147–1151.

Beaumont, H. J., Gallie, J., Kost, C., Ferguson, G. C., and Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature*, 462(7269):90.

Beckmann, B. E., Knoester, D. B., Connelly, B. D., Waters, C. M., and McKinley, P. K. (2012). Evolution of resistance to quorum quenching in digital organisms. *Artificial life*, 18(3):291–310.

Beckmann, B. E. and McKinley, P. K. (2009). Evolving quorum sensing in digital organisms. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 97–104. ACM.

Bedau, M. A. and Bahm, A. (1994). Bifurcation structure in diversity dynamics. In Brooks, R. and Maes, P., editors, *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pages 258–268, Cambridge, MA. MIT Press.

Bedau, M. A. and Packard, N. H. (1992). Measurement of evolutionary activity, teleology, and life. In Langton, I. C., Taylor, C., Farmer, D., and Rasmussen, S., editors, *Artificial Life II*, volume X of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 431–461. Addison-Wesley, Redwood City, CA.

Bedau, M. A., Snyder, E., Brown, C. T., and Packard, N. H. (1997). A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In Husbands, P. and Harvey, I., editors, *Proceedings of the Fourth European Conference on Artificial Life, ECAL97*, pages 125–134, Cambridge, MA. MIT Press.

Bedau, M. A., Snyder, E., and Packard, N. H. (1998). A classification of long-term evolutionary dynamics. In Adami, C., Belew, R. K., Kitano, H., and Taylor, C. E., editors, *Artificial Life VI: Proceedings of the Sixth International Conference on Artificial Life*, pages 228–237, Cambridge, MA. MIT Press.

Biswas, R., Bryson, D., Ofria, C., and Wagner, A. (2014). Causes vs Benefits in the Evolution of Prey Grouping. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 641–648.

Blickle, T. and Thiele, L. (1995). A mathematical analysis of tournament selection. In *ICGA*, pages 9–16. Citeseer.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Blount, Z. D., Barrick, J. E., Davidson, C. J., and Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, 489(7417):513–518.

Blount, Z. D., Borland, C. Z., and Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 105(23):7899–7906.

Bongard, J. C. and Hornby, G. S. (2010). Guarding Against Premature Convergence While Accelerating Evolutionary Search. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, GECCO '10, pages 111–118, New York, NY, USA. ACM.

Bosman, A. S., Engelbrecht, A. P., and Helbig, M. (2018). Progressive gradient walk for neural network fitness landscape analysis. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '18*, pages 1473–1480, Kyoto, Japan. ACM Press.

Bryson, D., Baer, B., Ofria, C., Barrick, J., Vostinar, A., Goldsby, H., Zaman, L., Chandler, C., Goings, S., Dolson, E., Vogel, M., Covert, A., Wright, G., Nahum, J., Misevic, D., Wagner, A., Rupp, M., Yilmaz, E., Pakanati, A., and Biswas, R. (2018). Code for using modes metrics in avida. Available at https://doi.org/10.5281/zenodo.1439479.

Burlacu, B., Affenzeller, M., Kommenda, M., Winkler, S., and Kronberger, G. (2013). Visualization of Genetic Lineages and Inheritance Information in Genetic Programming. In *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '13 Companion, pages 1351–1358, New York, NY, USA. ACM. event-place: Amsterdam, The Netherlands.

Cantú-Paz, E. and Goldberg, D. E. (2003). Are Multiple Runs of Genetic Algorithms Better than One? In Cantú-Paz, E., Foster, J. A., Deb, K., Davis, L. D., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Standish, R., Kendall, G., Wilson, S., Harman, M., Wegener, J., Dasgupta, D., Potter, M. A., Schultz, A. C., Dowsland, K. A., Jonoska, N., and Miller, J., editors, *Genetic and Evolutionary Computation — GECCO 2003*, number 2723 in Lecture Notes in Computer Science, pages 801–812. Springer Berlin Heidelberg.

Channon, A. (2001). Passing the ALife test: activity statistics classify evolution in Geb as unbounded. In Kelemen, J. and Sosík, P., editors, *Advances in Artificial Life*, pages 417–426, Berlin, Heidelberg. Springer Berlin Heidelberg.

Channon, A. (2003). Improving and Still Passing the ALife Test: Component-normalised Activity Statistics Classify Evolution in Geb As Unbounded. In Standish, R. K., Bedau, M. A., and Abbass, H. A., editors, *Proceedings of the Eighth International Conference on Artificial Life*, ICAL 2003, pages 173–181, Cambridge, MA, USA. MIT Press.

Chase, J. M. and Leibold, M. A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches*. University of Chicago Press. Google-Books-ID: Ssmcl_ubQUQC.

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual Review of Ecology and Systematics*, 31:343–366.

Chesson, P. and Kuang, J. J. (2008). The interaction between predation and competition. *Nature*, 456(7219):235–238.

Chow, S. S., Wilke, C. O., Ofria, C., Lenski, R. E., and Adami, C. (2004). Adaptive Radiation from Resource Competition in Digital Organisms. *Science*, 305(5680):84–86.

Clune, J., Goldsby, H. J., Ofria, C., and Pennock, R. T. (2010). Selective pressures for accurate altruism targeting: evidence from digital evolution for difficult-to-test aspects of inclusive fitness theory. *Proceedings of the Royal Society B: Biological Sciences*.

Cooper, T. F. and Ofria, C. (2003). Evolution of stable ecosystems in populations of digital organisms. In Standish, R. K., Bedau, M. A., and Abbass, H. A., editors, *Artificial Life VIII: Proceedings of the Eighth International Conference on Artificial life*, pages 227–232, Cambridge, MA. MIT Press.

Covert, A. W., Lenski, R. E., Wilke, C. O., and Ofria, C. (2013). Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences*, 110(34):E3171–E3178.

Craine, J. M., Ocheltree, T. W., Nippert, J. B., Towne, E. G., Skibbe, A. M., Kembel, S. W., and Fargione, J. E. (2013). Global diversity of drought tolerance and grassland climate-change resilience. *Nature Climate Change*, 3(1):63–67.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, New York, british first edition edition.

De Jong, K. A. (1975). *An Analysis of the Behavior of a Class of Genetic Adaptive Systems.* PhD Thesis, University of Michigan, Ann Arbor, MI, USA.

Deb, K. and Goldberg, D. E. (1989). An Investigation of Niche and Species Formation in Genetic Function Optimization. In *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 42–50, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Diggle, S. P., Griffin, A. S., Campbell, G. S., and West, S. A. (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, 450(7168):411–414.

Doebeli, M. and Dieckmann, U. (2003). Speciation along environmental gradients. *Nature*, 421(6920):259–264.

Dolson, E. (2018a). Code for Ecological Theory Provides Insights into Evolutionary Computation at the time of preprint/submission to journal.

Dolson, E. (2018b). Data, code, and supplemental figures for the MODES toolbox. Available at https://doi.org/10.5281/zenodo.2345140.

Dolson, E., Banzhaf, W., and Ofria, C. (2018a). Applying Ecological Principles to Genetic Programming. In Banzhaf, W., Olson, R. S., Tozier, W., and Riolo, R., editors, *Genetic Programming Theory and Practice XV*, pages 73–88, Cham. Springer International Publishing.

Dolson, E., Lalejini, A., Jorgensen, S., and Ofria, C. (2018b). Quantifying the tape of life: Ancestry-based metrics provide insights and intuition about evolutionary dynamics. In Ikegami, T., Virgo, N., Witkowski, O., Oka, M., Suzuki, R., and Iizuka, H., editors, *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial*

Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE), pages 75–82, Cambridge, MA. MIT Press.

Dolson, E., Lalejini, A., and Ofria, C. (2018c). Exploring genetic programming systems with MAP-Elites. *PeerJ Preprints*, 6:e27154v1.

Dolson, E. and Ofria, C. (2017). Spatial resource heterogeneity creates local hotspots of evolutionary potential. In *Proceedings of the European Conference on Artificial Life*, pages 122–129, Lyon, France. MIT Press.

Dolson, E. and Ofria, C. (2018). Visualizing the tape of life: Exploring evolutionary history with virtual reality. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '18, pages 1553–1559, New York, NY, USA. ACM.

Dolson, E., Perez, S., and Chaput, A. (2016a). emilydolson/resource-heterogeneity: Journal Submission.

Dolson, E., Perez, S., Olson, R., and Ofria, C. (2017). Spatial resource heterogeneity increases diversity and evolutionary potential. *bioRxiv*.

Dolson, E., Vostinar, A., and Ofria, C. (2015). What's holding artificial life back from open-ended evolution? *The Winnower*.

Dolson, E., Wiser, M. J., and Ofria, C. A. (2016b). The Effects of Evolution and Spatial Structure on Diversity in Biological Reserves. In *Artificial Life XV: Proceedings of the Fifteenth International Conference on Artificial Life*, pages 434–440, Cancun, Mexico. MIT Press.

Dolson, E. L., Banzhaf, W., and Ofria, C. (2018d). Ecological theory provides insights about evolutionary computation. *PeerJ Preprints*, 6:e27315v1.

Dolson, E. L., Lalejini, A., Jorgensen, S., and Ofria, C. A. (2019a). Interpreting the tape of life: Ancestry-based Analyses Provide Insights and Intuition about Evolutionary Dynamics. *Artificial Life*. status: In Press.

Dolson, E. L., Vostinar, A. E., Wiser, M. J., and Ofria, C. A. (2019b). The MODES toolbox: Measurements of Open-ended Dynamics in Evolving Systems. *Artificial Life*, 25(1).

Erwin, D. H. (1998). The end and the beginning: recoveries from mass extinctions. *Trends in Ecology & Evolution*, 13(9):344–349.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10.

Ferrière, R., Dieckmann, U., and Couvet, D. (2004). *Evolutionary conservation biology.* Cambridge University Press.

Fontaine, C., Guimarães, P. R., Kéfi, S., Loeuille, N., Memmott, J., van der Putten, W. H., van Veen, F. J. F., and Thébault, E. (2011). The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, 14(11):1170–1181.

Forest, F., Grenyer, R., Rouget, M., Davies, T. J., Cowling, R. M., Faith, D. P., Balmford, A., Manning, J. C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T. A. J., and Savolainen, V. (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, 445(7129):757–760.

Forman, R. T. (1995). *Land Mosaics: The Ecology of Landscapes and Regions*. Cambridge University Press.

Forman, R. T. T. and Godron, M. (1986). *Landscape ecology*. Wiley.

Fortuna, M. A., Zaman, L., Wagner, A. P., and Ofria, C. (2013). Evolving Digital Ecological Networks. *PLoS Comput Biol*, 9(3):e1002928.

Franklin, J. F. and Lindenmayer, D. B. (2009). Importance of matrix habitats in maintaining biological diversity. *Proceedings of the National Academy of Sciences*, 106(2):349–350.

Frederickson, M. E., Greene, M. J., and Gordon, D. M. (2005). Ecology: 'Devil's gardens' bedevilled by ants. *Nature*, 437(7058):495–496.

Fu, Y.-X. and Li, W.-H. (1999). Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology*, 56(1):1–10.

Galloway, J. N., Dentener, F. J., Capone, D. G., Boyer, E. W., Howarth, R. W., Seitzinger, S. P., Asner, G. P., Cleveland, C. C., Green, P. A., Holland, E. A., Karl, D. M., Michaels, A. F., Porter, J. H., Townsend, A. R., and Vorosmarty, C. J. (2004). Nitrogen cycles: past, present, and future. *Biogeochemistry*, 70:153–226.

Gavrilets, S. (2010). High-Dimensional Fitness Landscapes and Speciation. In Pigliucci, M. and Müller, G. B., editors, *Evolution—the Extended Synthesis*, pages 45–80. The MIT Press.

Gavrilets, S. and Vose, A. (2005). Dynamic patterns of adaptive radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(50):18040–18045.

Gillespie, R. (2004). Community Assembly Through Adaptive Radiation in Hawaiian Spiders. *Science*, 303(5656):356–359.

Goings, S. (2010). *Natural niching: Applying ecological principles to evolutionary computation*. Ph.D., Michigan State University, United States – Michigan.

Goings, S., Goldsby, H. J., Cheng, B. H., and Ofria, C. (2012). An ecology-based evolutionary algorithm to evolve solutions to complex problems. *Artificial Life*, 13:171–177.

Goings, S. and Ofria, C. (2009). Ecological approaches to diversity maintenance in evolutionary algorithms. In *IEEE Symposium on Artificial Life, 2009. ALife '09*, pages 124–130.

Goldberg, D. E., Richardson, J., and Grefenstette, J. J. (1987). Genetic algorithms with sharing for multimodal function optimization. In *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*, pages 41–49, Hillsdale, NJ. Lawrence Erlbaum.

Goldsby, H. J., Dornhaus, A., Kerr, B., and Ofria, C. (2012). Task-switching costs promote the evolution of division of labor and shifts in individuality. *Proceedings of the National Academy of Sciences*, 109(34):13686–13691.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., and Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50.

Goudard, A. and Loreau, M. (2008). Nontrophic Interactions, Biodiversity, and Ecosystem Functioning: An Interaction Web Model. *The American Naturalist*, 171(1):91–106.

Grabowski, L. and Magaña, J. (2014). Building on Simplicity: Multi-stage Evolution of Digital Organisms. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 113–120.

Grover, J. P. (1997). *Resource Competition.* Springer Science & Business Media.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

Haller, B. C., Mazzucco, R., and Dieckmann, U. (2013). Evolutionary Branching in Complex Landscapes. *The American Naturalist*, 182(4):E127–E141.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1):17–52.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Hara, A. and Nagao, T. (1999). Emergence of the cooperative behavior using ADG; Automatically Defined Groups. In Banzhaf, W., Daida, J. M., Eiben, A. E., Garzon, M. H., Honavar, V. G., Jakiela, M. J., and Smith, R. E., editors, *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation*, volume 1, pages 1039–1046, San Francisco, CA, USA. Morgan Kaufmann.

Harmon, L. J. and Harrison, S. (2015). Species diversity is dynamic and unbounded at local and continental scales. *The American Naturalist*, 185(5):584–593.

Harper, R. (2012). Spatial Co-evolution: Quicker, Fitter and Less Bloated. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO '12, pages 759–766, New York, NY, USA. ACM.

Harpole, W. S. and Tilman, D. (2007). Grassland species loss resulting from reduced niche dimension. *Nature*, 446(7137):791–793.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of Effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128.

Helmuth, T., McPhee, N. F., and Spector, L. (2016a). Effects of Lexicase and Tournament Selection on Diversity Recovery and Maintenance. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, GECCO '16 Companion, pages 983–990, New York, NY, USA. ACM. event-place: Denver, Colorado, USA.

Helmuth, T., McPhee, N. F., and Spector, L. (2016b). The Impact of Hyperselection on Lexicase Selection. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, pages 717–724, New York, NY, USA. ACM. event-place: Denver, Colorado, USA.

Helmuth, T. and Spector, L. (2015). General Program Synthesis Benchmark Suite. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, GECCO '15, pages 1039–1046, New York, NY, USA. ACM.

Helmuth, T., Spector, L., and Matheson, J. (2015). Solving Uncompromising Problems With Lexicase Selection. *IEEE Transactions on Evolutionary Computation*, 19(5):630–643.

Hendry, A. P. (2016). *Eco-evolutionary Dynamics*. Princeton University Press. Google-Books-ID: QMtIDAAAQBAJ.

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *The American Naturalist*, 163(2):192–211.

Holland, J. H. (1985). Properties of the Bucket Brigade. In *Proceedings of the 1st International Conference on Genetic Algorithms*, pages 1–7, Hillsdale, NJ, USA. L. Erlbaum Associates Inc.

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J. H., Lodge, D. M., Loreau, M., Naeem, S., Schmid, B., Setälä, H., Symstad, A. J., Vandermeer, J., and Wardle, D. A. (2005). Effects of Biodiversity on Ecosystem Functioning: A Consensus of Current Knowledge. *Ecological Monographs*, 75(1):3–35.

Hornby, G. S. (2006). ALPS: The Age-layered Population Structure for Reducing the Problem of Premature Convergence. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, GECCO '06, pages 815–822, New York, NY, USA. ACM.

Hu, J., Goodman, E., Seo, K., Fan, Z., and Rosenberg, R. (2005). The Hierarchical Fair Competition (HFC) Framework for Sustainable Evolutionary Algorithms. *Evolutionary Computation*, 13(2):241–277.

Hutchinson, G. E. (1961). The Paradox of the Plankton. *The American Naturalist*, 95(882):137–145.

Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C., and Baillie, J. E. M. (2007). Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. *PLOS ONE*, 2(3):e296.

Jankowski, K., Schindler, D. E., and Horner-Devine, M. C. (2014). Resource Availability and Spatial Heterogeneity Control Bacterial Community Response to Nutrient Enrichment in Lakes. *PLoS ONE*, 9(1):e86991.

Johnson, A., Strauss, E., Pickett, R., Adami, C., Dworkin, I., and Goldsby, H. (2014). More Bang For Your Buck: Quorum-Sensing Capabilities Improve the Efficacy of Suicidal Altruism. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 120–128.

Johnson, T. J. and Wilke, C. O. (2004). Evolution of Resource Competition between Mutually Dependent Digital Organisms. *Artificial Life*, 10(2):145–156.

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

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, 27(10):547–560.

Kokko, H., Chaturvedi, A., Croll, D., Fischer, M. C., Guillaume, F., Karrenberg, S., Kerr, B., Rolshausen, G., and Stapley, J. (2017). Can Evolution Supply What Ecology Demands? *Trends in Ecology & Evolution*, 32(3):187–197.

Korb, K. B. and Dorin, A. (2011). Evolution unbound: releasing the arrow of complexity. *Biology & Philosophy*, 26(3):317–338.

Kosmidis, I. (2017). brglm: Bias Reduction in Binary-Response Generalized Linear Models.

La Cava, W., Helmuth, T., Spector, L., and Moore, J. H. (2018). A Probabilistic and Multi-Objective Analysis of Lexicase Selection and ε-Lexicase Selection. *Evolutionary Computation*, pages 1–26.

Lafferty, K. D., Dobson, A. P., and Kuris, A. M. (2006). Parasites dominate food web links. *Proceedings of the National Academy of Sciences*, 103(30):11211–11216.

Lalejini, A., Dolson, E., and Jorgensen, S. (2018). stevenjson/ALife2018-Lineage: Final Release for Paper. https://doi.org/10.5281/zenodo.1254061.

Lalejini, A. and Ofria, C. (2016). The Evolutionary Origins of Phenotypic Plasticity. In *Proceedings of the Artificial Life Conference 2016*, pages 372–379. The MIT Press.

Lavergne, S. and Molofsky, J. (2007). Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences*, 104(10):3883–3888.

Le Gac, M., Plucain, J., Hindré, T., Lenski, R. E., and Schneider, D. (2012). Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 109(24):9487–9492.

Lehman, J. and Stanley, K. O. (2008). Exploiting Open-Endedness to Solve Problems Through the Search for Novelty. In *ALIFE*, pages 329–336.

Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evol. Comput.*, 19(2):189–223.

Lehmann, K. D. S., Goldman, B. W., Dworkin, I., Bryson, D. M., and Wagner, A. P. (2014). From Cues to Signals: Evolution of Interspecific Communication via Aposematism and Mimicry in a Predator-Prey System. *PLOS ONE*, 9(3):e91783.

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

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*, 138(6):1315–1341.

Lenski, R. E., Wiser, M. J., Ribeck, N., Blount, Z. D., Nahum, J. R., Morris, J. J., Zaman, L., Turner, C. B., Wade, B. D., Maddamsetti, R., Burmeister, A. R., Baird, E. J., Bundy, J., Grant, N. A., Card, K. J., Rowles, M., Weatherspoon, K., Papoulis, S. E., Sullivan, R., Clark, C., Mulka, J. S., and Hajela, N. (2015). Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1821).

Letten, A. D., Ke, P.-J., and Fukami, T. (2017). Linking modern coexistence theory and contemporary niche theory. *Ecological Monographs*, 87(2):161–177.

Li, X., Engelbrecht, A., and Epitropakis, M. G. (2013). Benchmark Functions for CEC'2013 Special Session and Competition on Niching Methods for Multimodal Function Optimization. Technical report, RMIT University, Australia.

Losos, J. (2010). Adaptive Radiation, Ecological Opportunity, and Evolutionary Determinism. *The American Naturalist*, 175(6):623–639.

MacArthur, R. H. and Wilson, E. O. (1967). *The Theory of Island Biogeography*. Princeton University Press.

MacPherson, B. and Gras, R. (2016). Individual-based ecological models: Adjunctive tools or experimental systems? *Ecological Modelling*, 323:106–114.

Maddamsetti, R., Lenski, R. E., and Barrick, J. E. (2015). Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics*, 200(2):619–631.

Mahfoud, S. W. (1992). Crowding and preselection revisited. *Urbana*, 51:61801.

Mahfoud, S. W. (1995). Niching methods for genetic algorithms. *Urbana*, 51(95001).

Maley, C. C. (1999). Four steps toward open-ended evolution. In Banzhaf, W., Daida, J. M., Eiben, A. E., Garzon, M. H., and Honavar, V. G., editors, *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation*, volume 2 of *GECCO'99*, pages 1336–1343, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Maynard Smith, J. (1992). Byte-sized evolution. *Nature*, 355(6363):772–773.

Maynard Smith, J. and Szathmary, E. (1997). *The major transitions in evolution*. Oxford University Press.

McCann, K. S. (2000). The diversity–stability debate. *Nature*, 405(6783):228–233.

McDonald, B. A. and Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annual Review of Phytopathology*, 40(1):349–379.

McGarigal, K. (2006). Landscape Pattern Metrics. In *Encyclopedia of Environmetrics*. John Wiley & Sons, Ltd.

McGarigal, K., Tagil, S., and Cushman, S. A. (2009). Surface metrics: an alternative to patch metrics for the quantification of landscape structure. *Landscape Ecology*, 24(3):433–450.

McKinney, M. L. (2006). Urbanization as a major cause of biotic homogenization. *Biological Conservation*, 127(3):247–260.

McPhee, N. F., Casale, M. M., Finzel, M., Helmuth, T., and Spector, L. (2016a). Visualizing genetic programming ancestries. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, GECCO '16 Companion, pages 1419–1426, New York, NY, USA. ACM.

McPhee, N. F., Donatucci, D., and Helmuth, T. (2016b). Using Graph Databases to Explore the Dynamics of Genetic Programming Runs. In *Genetic Programming Theory and Practice XIII*, Genetic and Evolutionary Computation, pages 185–201. Springer, Cham.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., Harrison, S. P., Hurlbert, A. H., Knowlton, N., Lessios, H. A., McCain, C. M., McCune, A. R., McDade, L. A., McPeek, M. A., Near, T. J., Price, T. D., Ricklefs, R. E., Roy, K., Sax, D. F., Schluter, D., Sobel, J. M., and Turelli, M. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology Letters*, 10(4):315–331.

Moritz, C. (2002). Strategies to protect biological diversity and the evolutionary processes that sustain it. *Systematic biology*, 51(2):238–254.

Moritz, C., Patton, J. L., Schneider, C. J., and Smith, T. B. (2000). Diversification of rainforest faunas: an integrated molecular approach. *Annual review of ecology and systematics*, 31(1):533–563.

Mouret, J.-B. and Clune, J. (2015). Illuminating search spaces by mapping elites. *arXiv:1504.04909 [cs, q-bio]*. arXiv: 1504.04909.

Mouret, J.-B. and Doncieux, S. (2009). Using Behavioral Exploration Objectives to Solve Deceptive Problems in Neuro-evolution. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, GECCO '09, pages 627–634, New York, NY, USA. ACM.

Muller, H. J. (1932). Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138.

Nahum, J. R., Godfrey-Smith, P., Harding, B. N., Marcus, J. H., Carlson-Stevermer, J., and Kerr, B. (2015). A tortoise–hare pattern seen in adapting structured and unstructured populations suggests a rugged fitness landscape in bacteria. *Proceedings of the National Academy of Sciences*, 112(24):7530–7535.

Noble, R. (2017). *ggmuller: Create Muller Plots of Evolutionary Dynamics*. R package version 0.5.2.

Ochoa, G., Tomassini, M., Vérel, S., and Darabos, C. (2008). A study of NK landscapes' basins and local optima networks. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 555–562. ACM.

O'Donnell, D. R., Parigi, A., Fish, J. A., Dworkin, I., and Wagner, A. P. (2014). The Roles of Standing Genetic Variation and Evolutionary History in Determining the Evolvability of Anti-Predator Strategies. *PLOS ONE*, 9(6):e100163.

Ofria, C., Dolson, E., Lalejini, A., Fenton, J., Jorgensen, S., Miller, R., Moreno, M. A., Stredwick, J., Zaman, L., Schossau, J., Gillespie, L., C G, N., and Vostinar, A. (2018). Empirical.

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

Oldford, W. (2016). qqtest: Self Calibrating Quantile-Quantile Plots for Visual Testing.

Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data*, 2(1):22.

Olson, R., Mirmomeni, M., Brom, T., Bruger, E., Hintze, A., Knoester, D., and Adami, C. (2013). Evolved digital ecosystems: Dynamic steady state, not optimal fixed point. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 126–133.

Ovaska, S. J., Sick, B., and Wright, A. H. (2009). Periodical switching between related goals for improving evolvability to a fixed goal in multi-objective problems. *Information Sciences*, 179(23):4046–4056.

Pacala, S. W. and Tilman, D. (1994). Limiting Similarity in Mechanistic and Spatial Models of Plant Competition in Heterogeneous Environments. *The American Naturalist*, 143(2):222–257.

Pianka, E. R. (1966a). Convexity, Desert Lizards, and Spatial Heterogeneity. *Ecology*, 47(6):1055–1059.

Pianka, E. R. (1966b). Latitudinal Gradients in Species Diversity: A Review of Concepts. *The American Naturalist*, 100(910):33–46.

Potter, M. A. and De Jong, K. A. (2000). Cooperative coevolution: An architecture for evolving coadapted subcomponents. *Evolutionary Computation*, 8(1):1–29.

Pugh, J. K., Soros, L. B., Szerlip, P. A., and Stanley, K. O. (2015). Confronting the Challenge of Quality Diversity. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, GECCO '15, pages 967–974, New York, NY, USA. ACM.

R Core Development Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ray, T. S. (1994). Evolution, complexity, entropy and artificial reality. *Physica D: Nonlinear Phenomena*, 75(1):239–263.

Ribeck, N. and Lenski, R. E. (2015). Modeling and quantifying frequency-dependent fitness in microbial populations with cross-feeding interactions. *Evolution*, 69(5):1313–1320.

Ripley, B. D. (1981). Mapped Point Patterns. In *Spatial Statistics*, pages 144–190. John Wiley & Sons, Inc.

Roopnarine, P. D., Angielczyk, K. D., Wang, S. C., and Hertog, R. (2007). Trophic network models explain instability of Early Triassic terrestrial communities. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1622):2077–2086.

Rouzic, A. L. and Carlborg, O. (2008). Evolutionary potential of hidden genetic variation. *Trends in Ecology & Evolution*, 23(1):33–37.

Rozen, D. E. and Lenski, R. E. (2000). Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *The American Naturalist*, 155(1):24–35.

Schluter, D. (2009). Evidence for Ecological Speciation and Its Alternative. *Science*, 323(5915):737–741.

Schluter, D., Price, T. D., and Grant, P. R. (1985). Ecological Character Displacement in Darwin's Finches. *Science*, 227(4690):1056–1059.

Schoener, T. W. (2011). The Newest Synthesis: Understanding the Interplay of Evolutionary and Ecological Dynamics. *Science*, 331(6016):426–429.

Scott, J. G., Fletcher, A. G., Anderson, A. R. A., and Maini, P. K. (2016). Spatial Metrics of Tumour Vascular Organisation Predict Radiation Efficacy in a Computational Model. *PLOS Computational Biology*, 12(1):e1004712.

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

Skusa, A. and Bedau, M. A. (2003). Towards a comparison of evolutionary creativity in biological and cultural evolution. *Artificial Life*, 8:233–242.

Soros, L. (2018). Necessary conditions for open-ended evolution. *Electronic Theses and Dissertations.*

Soros, L. and Stanley, K. (2014). Identifying necessary conditions for open-ended evolution through the artificial life world of Chromaria. In Sayama, H., Rieffel, J., Risi, S., Doursat, R., and Lipson, H., editors, *ALIFE 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 793–800, Cambridge, MA. MIT Press.

Spector, L. (2012). Assessment of problem modality by differential performance of lexicase selection in genetic programming: a preliminary report. In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, pages 401–408. ACM.

Standish, R. K. and Galloway, J. (2002). Visualising Tierra's tree of life using Netmap. In *ALife VIII Workshop proceedings*, page 171.

Stanley, K. and Miikkulainen, R. (2004). Competitive coevolution through evolutionary complexification. *J. Artif. Intell. Res. (JAIR)*, 21:63–100.

Stanley, K. O. and Soros, L. B. (2016). The role of subjectivity in the evaluation of open-endedness. Presentation delivered in OEE2: The Second Workshop on Open-Ended Evolution, at ALIFE 2016, in Cancún, Mexico, 4-8 July 2016.

Stein, A., Gerstner, K., and Kreft, H. (2014). Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecology Letters*, pages n/a–n/a.

Strona, G. and Lafferty, K. D. (2016). Environmental change makes robust ecological networks fragile. *Nature Communications*, 7:12462.

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., Dolson, E., Froese, T., Hickinbotham, S., Ikegami, T., McMullin, B., Packard, N., Rasmussen, S., Virgo, N., Agmon, E., Clark, E., McGregor, S., Ofria, C., Ropella, G., Spector, L., Stanley, K. O., Stanton, A., Timperley, C., Vostinar, A., and Wiser, M. (2016). Open-ended evolution: Perspectives from the OEE workshop in York. *Artificial Life*, 22(3):408–423.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D., and Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536:165.

Tilman, D. (1977). Resource Competition between Plankton Algae: An Experimental and Theoretical Approach. *Ecology*, 58(2):338.

Tilman, D., Isbell, F., and Cowles, J. M. (2014). Biodiversity and Ecosystem Functioning. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):471–493.

Tomassini, M. (2005). *Spatially Structured Evolutionary Algorithms: Artificial Evolution in Space and Time.* Natural Computing Series. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., and Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92(2):698–715.

Turner, C. B., Blount, Z. D., and Lenski, R. E. (2015a). Replaying evolution to test the cause of extinction of one ecotype in an experimentally evolved population. *PLoS ONE*, 10(11):e0142050.

Turner, C. B., Blount, Z. D., Mitchell, D. H., and Lenski, R. E. (2015b). Evolution and coexistence in response to a key innovation in a long-term evolution experiment with *Escherichia coli. bioRxiv*.

van Dam, A., Laidlaw, D. H., and Simpson, R. M. (2002). Experiments in Immersive Virtual Reality for Scientific Visualization. *Computers & Graphics*, 26(4):535–555.

Vandergast, A. G., Bohonak, A. J., Hathaway, S. A., Boys, J., and Fisher, R. N. (2008). Are hotspots of evolutionary potential adequately protected in southern California? *Biological Conservation*, 141(6):1648–1664.

Virgo, N., Agmon, E., and Fernando, C. (2017). Lineage selection leads to evolvability at large population sizes. *Proceedings of the European Conference on Artificial Life*, pages 420–427.

Vostinar, A. E., Fenton, J., Waters, C., and Ofria, C. (2018). Signals in the Dark: What factors select for the evolution of cooperation controlled by quorum sensing? *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 651–658.

Vostinar, A. E. and Ofria, C. (2019). Spatial Structure Can Decrease Symbiotic Cooperation. *Artificial Life*, pages 1–21.

Wagner, A. P., Zaman, L., Dworkin, I., and Ofria, C. (2013). Behavioral Strategy Chases Promote the Evolution of Prey Intelligence. *arXiv:1310.1369 [q-bio]*.

Walker, B. L. and Ofria, C. (2012). Evolutionary Potential is Maximized at Intermediate Diversity Levels. In *Proceedings of the 13th International Conference for Artificial Life*, pages 116–120, East Lansing, MI, USA. MIT Press.

Walker, J. C. (2011). The Evolution of Optimal Foraging Strategies in Populations of Digital Organisms. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11, pages 203–210, New York, NY, USA. ACM.

Webb, C. O. and Losos, A. E. J. B. (2000). Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. *The American Naturalist*, 156(2):145–155.

White, D. R., McDermott, J., Castelli, M., Manzoni, L., Goldman, B. W., Kronberger, G., Jaśkowski, W., O'Reilly, U.-M., and Luke, S. (2013). Better GP benchmarks: community survey results and proposals. *Genetic Programming and Evolvable Machines*, 14(1):3–29.

Whitley, D., Rana, S., and Heckendorn, R. B. (1998). The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. *Journal of Computing and Information Technology*, 7:33–47.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

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(6844):331–333.

Wilmers, C. C. and Getz, W. M. (2005). Gray Wolves as Climate Change Buffers in Yellowstone. *PLoS Biol*, 3(4):e92.

Winter, M., Devictor, V., and Schweiger, O. (2013). Phylogenetic diversity and nature conservation: where are we? *Trends in Ecology & Evolution*, 28(4):199–204.

Wiser, M. J. (2015). *An analysis of fitness in long-term asexual evolution experiments*. PhD thesis, Michigan State University.

Wiser, M. J., Dolson, E. L., Vostinar, A., Lenski, R. E., and Ofria, C. (2018). The boundedness illusion: Asymptotic projections from early evolution underestimate evolutionary potential. *PeerJ Preprints*, 6:e27246v2.

Wiser, M. J., Ribeck, N., and Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics*, pages 356–366, Brooklyn Botanic Garden, Brooklyn, NY.

Yedid, G., Ofria, C., and Lenski, R. (2009). Selective Press Extinctions, but Not Random Pulse Extinctions, Cause Delayed Ecological Recovery in Communities of Digital Organisms. *The American Naturalist*, 173(4):E139–E154.

Yedid, G., Ofria, C. A., and Lenski, R. E. (2008). Historical and contingent factors affect re-evolution of a complex feature lost during mass extinction in communities of digital organisms. *Journal of Evolutionary Biology*, 21(5):1335–1357.

Yedid, G., Stredwick, J., Ofria, C. A., and Agapow, P.-M. (2012). A Comparison of the Effects of Random and Selective Mass Extinctions on Erosion of Evolutionary History in Communities of Digital Organisms. *PLOS ONE*, 7(5):e37233.

Zaman, L. (2018). Investigating open-ended coevolution in digital organisms. In Ikegami, T., Virgo, N., Witkowski, O., Oka, M., Suzuki, R., and Iizuka, H., editors, *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 258–259, Cambridge, MA. MIT Press.

Zaman, L., Devangam, S., and Ofria, C. (2011). Rapid host-parasite coevolution drives the production and maintenance of diversity in digital organisms. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 219–226. ACM.

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E., and Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol*, 12(12):e1002023.