Quantifying the Tape of Life: Ancestry-based Metrics Provide Insights and Intuition about Evolutionary Dynamics

Emily Dolson[1,2,3], Alexander Lalejini[1,2,3], Steven Jorgensen[1,2] and Charles Ofria[1,2,3]

[1]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI, 48824
[2]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824
[3]Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI, 48824
dolsonem@msu.edu

Abstract

Fine-scale evolutionary dynamics can be challenging to tease out when focused on broad brush strokes of whole populations over long time spans. We propose a suite of diagnostic metrics that operate on lineages and phylogenies in digital evolution experiments with the aim of improving our capacity to quantitatively explore the nuances of evolutionary histories in digital evolution experiments. We present three types of lineage measurements: lineage length, mutation accumulation, and phenotypic volatility. Additionally, we suggest the adoption of four phylogeny measurements from biology: depth of the most-recent common ancestor, phylogenetic richness, phylogenetic divergence, and phylogenetic regularity. We demonstrate the use of each metric on a set of two-dimensional, real-valued optimization problems under a range of mutation rates and selection strengths, confirming our intuitions about what they can tell us about evolutionary dynamics.

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, we have developed a standardized suite of diagnostic metrics to summarize short-term evolutionary dynamics within a population by measuring lineages and phylogenies. Here, we 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 we are interested in (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 use phenotypes and genotypes as the states in the sequence, but we could just as easily use some other descriptor of the lineage's position in the fitness landscape 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, we explore a subset of these metrics that we 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, we 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, we discuss its application and our intuition for what it can tell us about evolution. We evaluate our intuition on a set of two-dimensional, real-valued optimization problems under a range of mutation rates and selection strengths. For this work, we restrict our attention to asexually reproducing populations; however, we suggest how these metrics can extend to sexual populations.

In addition to demonstrating a range of metrics that are useful to digital evolution research, we 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, we are unaware of efforts to formalize a suite of quantitative lineage and phylogeny-based metrics for computational evolution.

Metrics

Code for all of our metrics is open source and available in the Empirical library (`https://github.com/devosoft/Empirical`). Empirical is a C++ library built to facilitate writing efficient and easily sharable scientific software. Empirical is a header-only library, so adding these metrics to an existing project has minimal overhead.

Lineage Metrics

Each of the three lineage metrics that we 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 1 is given as a toy example to help guide our discussion of these metrics. While we limit our focus to three lineage metrics, this abstraction 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. One possibility is to compress sexual lineages into linear sequences of states by modeling sexual reproduction events as asexual reproduction events, designating one parent to be a part of the lineage and considering the genetic contributions of other parents as 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). Alternatively, we can extend our 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, calculating a given metric for each of them and then averaging the resulting values together. Another approach would be to divide an organism into its constituent parts that are inherited atomically (such as genes or instructions, depending on the representation); an organism would then be viewed as a collection of lineages rather than a single one. Assessing the efficacy of these and potentially other approaches would be a useful line of research to pursue in the future.

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 tasks to be successful, we might define state as the set of tasks performed by an individual. In this scenario, lineage length would only increase when the set of tasks performed by an ancestor changes; sequential ancestors that perform the identical sets of tasks would be compressed into a single state in the sequence, even if other traits differ.

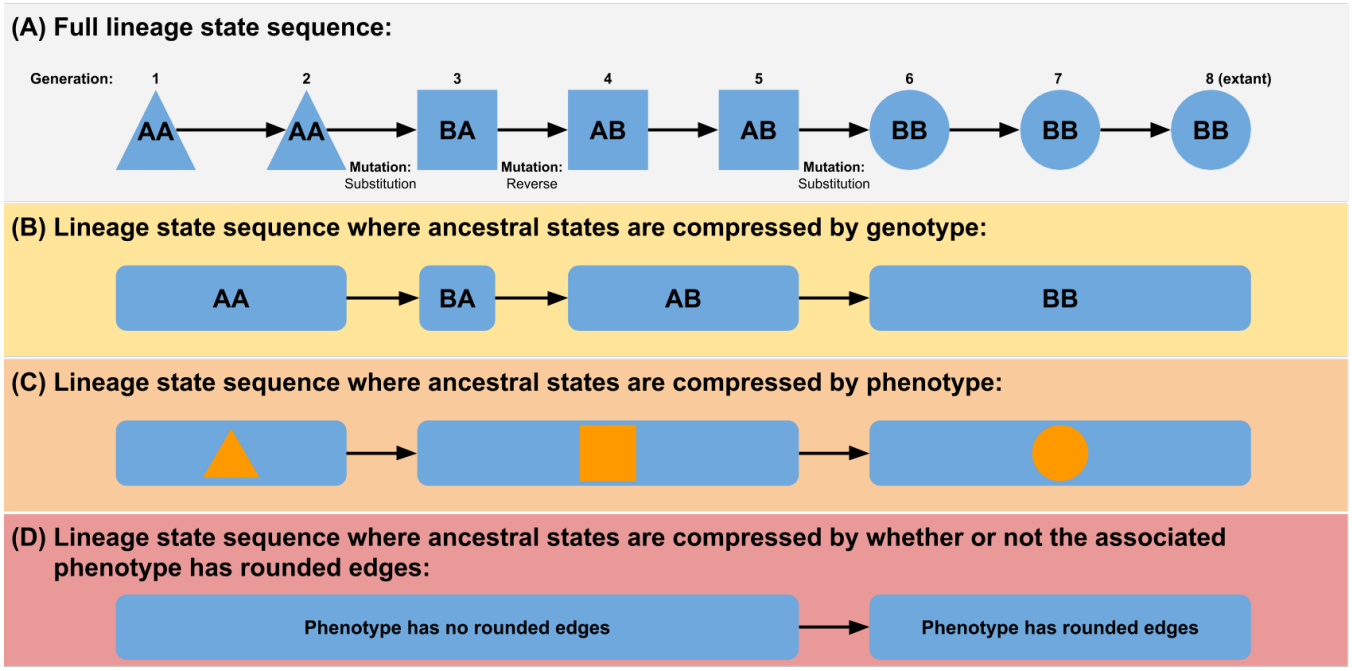

Figure 1: Four methods of representing a lineage. This example lineage has accumulated three mutations (one reverse mutation and two substitutions) and gone through three distinct phenotypes. In (A), each state along the lineage represents a single individual; lineage length is the number of generations 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.

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.

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.

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; here we use the most fit). 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.

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 we refer generally to items in a phylogeny, we 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 we 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 paper, we 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. 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, we exclude pre-MRCA taxa from our analyses. However, we will not remove non-branching internal nodes, as these only serve to make our phylogenies more informative.

Here we provide a high-level summary of phylogeny metrics that we 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).

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 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).

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).

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

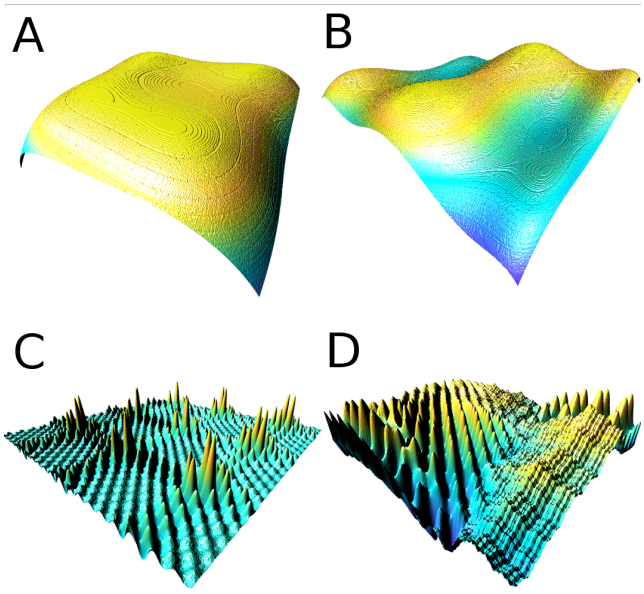

Figure 2: The fitness landscapes used in this experiment: A) Himmelblau, B) Six-humped Camel Back, C) Shubert, and D) Composition Function 2. Interactive versions available at `https://emilydolson.github.io/fitness_landscape_visualizations`.

each taxon in the population.

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.

Test Problems

To understand the metrics defined above, the test problems used need to be well understood and studied. The benchmark functions from the GECCO Competition on Niching Methods meet both of these requirements and allow us to visualize the actual fitness landscape, due to the low dimensionality of the problems (Li et al., 2013). For each problem, the X and Y coordinates offered by a given organism are translated by the function into a fitness value. We chose a diverse subset of these functions (Himmelblau, Shubert, Composition Function 2, and Six-Humped Camel Back) as our test problems in order to gain a broad understanding of our metrics. We used the implementations of these problems at `https://github.com/mikeagn/CEC2013` (C++ for fitness calculations during evolution, Python for post-hoc analysis). Figure 2 shows the fitness landscapes defined by each of our four chosen test problems.

For each test problem, we evolved populations of 1000 organisms under a range of mutation rates and selection strengths for 5000 generations. Each organism's genome consisted of two floating point numbers that defined its position in the fitness landscape. We initialized populations by randomly generating a number of organisms equal to the population size. To determine which organisms reproduced each generation, we used tournament selection. We evolved populations under five different tournament sizes: one, two, four, eight, and sixteen. Tournament size represents 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). 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. We prevented mutations from causing a value to exceed the valid domain of the given problem. For each problem and tournament size, we evolved populations at eight mutation rates: 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02, and 1e-01.

We also ran a second set of experiments to explore the impact of ecological dynamics on these metrics. For these experiments, we generated a stable ecology using the Eco-EA algorithm as a selection technique (Goings et al., 2012). Eco-EA is a technique for creating niches that promote stable diversification in the context of an evolutionary algorithm. In our test problems, we created niches associated with spatial locations across the fitness landscape. For all experimental conditions, we ran ten replicates, each with a unique random number seed. Our experiment is implemented using the Empirical library; our implementation is included in the supplemental material for this paper (Lalejini et al., 2018).

Data Analysis

3D visualizations

In order to make these metrics useful, we must have an accurate understanding of how various measurements correspond to the actual behavior of lineages. The most direct way to confirm our expectations is to 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). Creating such a visualization entails condensing a large quantity of information into a limited space. When projected onto two dimensions, lineages can obscure parts of the fitness landscape (and each other). To mitigate this problem, we used the A-Frame framework (A-Frame authors, 2018) to build a three-dimensional data visualization (see

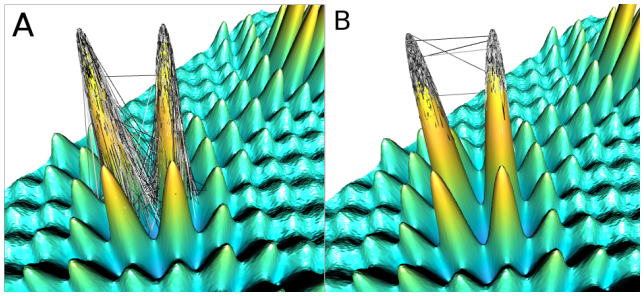

Figure 3: A close-up on two adjacent peaks in the Shubert function fitness landscape. Lineages are depicted as paths fading from white to black over evolutionary time. The lineages shown here evolved under a mutation rate of 0.01. A) Was evolved using a tournament size of 2, whereas B) was evolved using a tournament size of 16. These figures neatly illustrate how increased tournament size keeps the lineage near the tops of the peaks.

Figure 3) described in detail in our companion paper (Dolson and Ofria, 2018). For the data interpretation in this paper, we used an Oculus Rift to provide us with fine-grained control of which part of the visualization we were looking at. 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`.

Metric analysis

We analyzed trends in our metrics using the R Statistical Computing Language (R Core Team, 2017). Specifically, we used the ggplot2 library for all graphs included in this paper (Wickham, 2009). All analysis scripts are available in the supplemental material for this paper (Lalejini et al., 2018)..

Results and Discussion

Overall, our 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 4). 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, 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

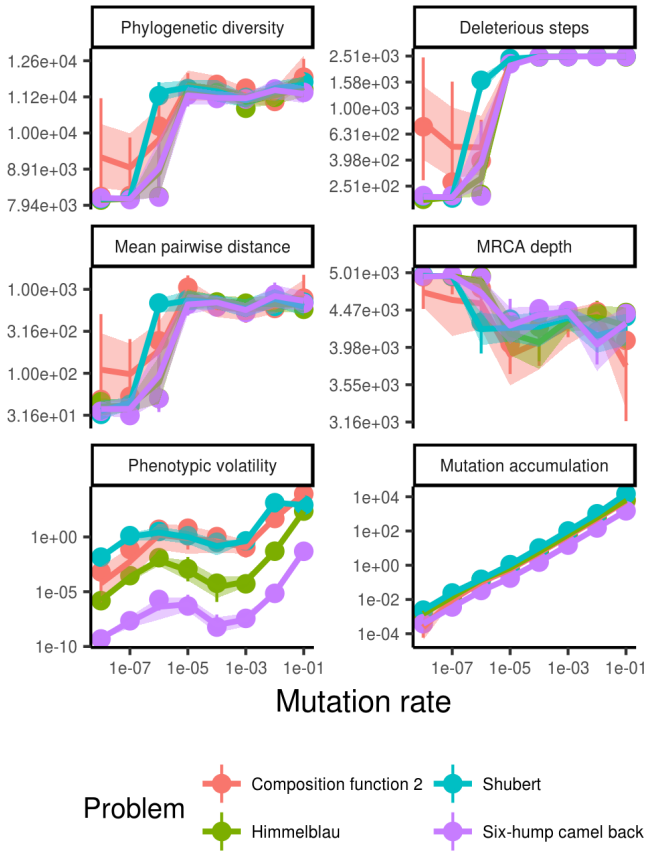

Figure 4: Values of example metrics across different mutation rates for each of the four problems. All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All experiments shown here used a tournament size of 4. 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.

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 5). 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 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

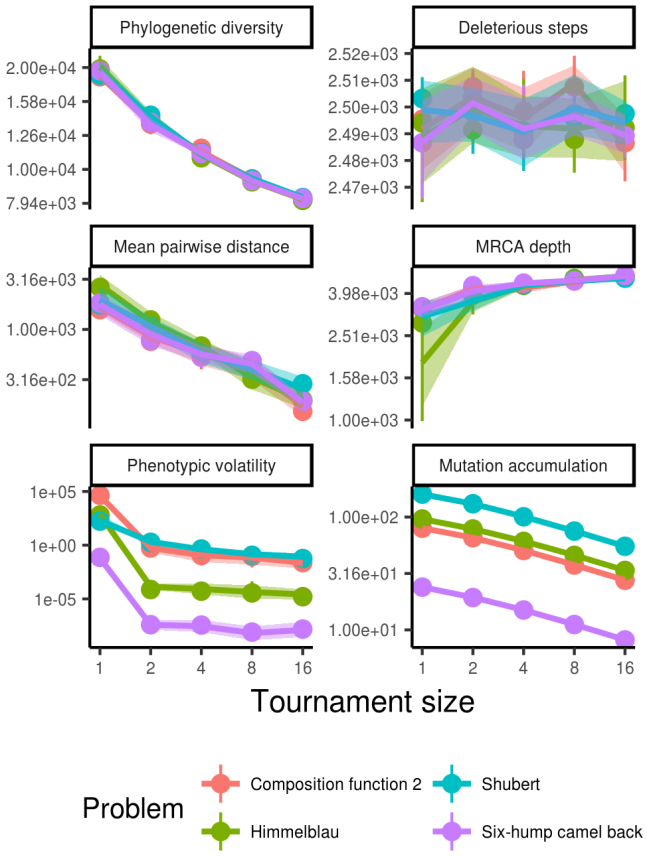

Figure 5: Values of example metrics across different tournament sizes for each of the four problems. All experiments shown here used a mutation rate of 0.001. All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. 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.

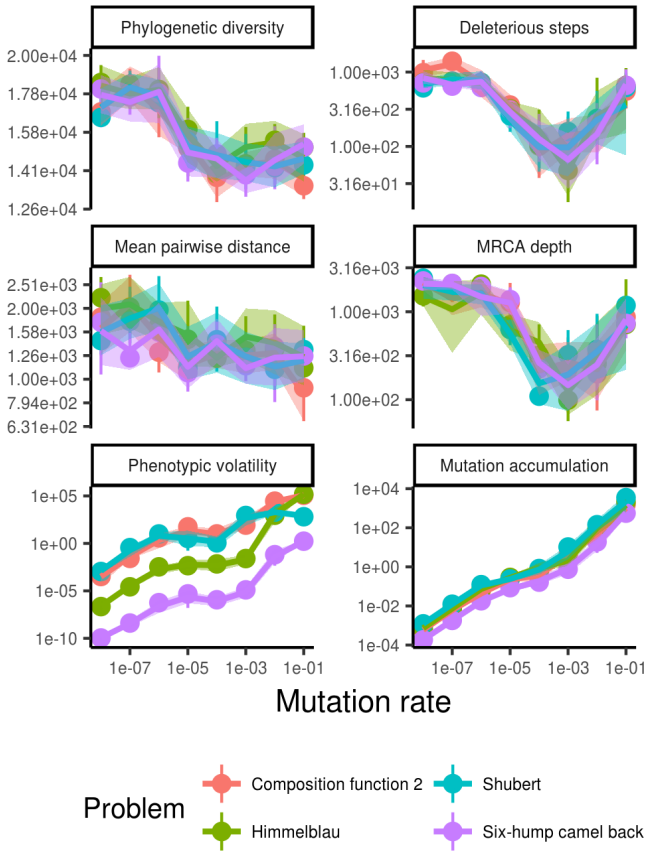

Figure 6: Values of example metrics across different mutation rates for each of the four problems under a diversity-preserving selection regime, Eco-EA. All lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All 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.

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 4, 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. 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 our metrics respond to ecological coexistence.

Indeed, the presence of stabilizing ecological dynamics substantially changes the values we observe for most metrics (see Figure 6). 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 the underlying mechanisms behind

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

Conclusions

Our goals for this work are two-fold: 1) to suggest a set of metrics that will improve 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, we will expand our suite of lineage and phylogeny metrics, compiling accessible descriptions and examples of each metric.

We have demonstrated that these metrics behave reasonably on a set of toy problems with simple organisms. Having established baseline expectations for their responses to common conditions, our next step is to apply these metrics to more complex scenarios: populations of digital organisms that we evolve in a variety of qualitatively different environments where we would expect to observe a wide range of evolutionary dynamics. It is under these conditions that we expect the true value of these metrics to become clear.

Acknowledgements

We thank members of the MSU Digital Evolution Lab for helpful comments and suggestions on this manuscript. This research was supported by the National Science Foundation (NSF) through the BEACON Center (Cooperative Agreement DBI-0939454), Graduate Research Fellowships to ED and AL (Grant No. DGE-1424871), and NSF Grant No. DEB-1655715 to CO. Michigan State University provided computational resources through the Institute for Cyber-Enabled Research and the Digital Scholarship Lab. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or MSU.

References

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

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

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

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

Burlacu, B., Affenzeller, M., Kommenda, M., Winkler, S., and Kronberger, G. (2013). Visualization of genetic lineages and inheritance information in genetic programming. page 1351. ACM 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.

Dolson, E. and Ofria, C. (2018). Visualizing the tape of life: exploring evolutionary history with virtual reality. In *GECCO 18 Companion: Genetic and Evolutionary Computation Conference Companion*, page 7, Kyoto, Japan. ACM.

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

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.

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

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.

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.

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

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. page 10.

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.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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.

Virgo, N., Agmon, E., and Fernando, C. (2017). Lineage selection leads to evolvability at large population sizes. pages 420–427. MIT Press.

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.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

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