Rank epistasis: A new model for analyzing epistatic interactions in the absence of quantifiable fitness scores

Acacia Ackles[1,2,3,†,*], Austin J. Ferguson[1,2,4,†], Connor Grady[1,5,6,†]

1 BEACON Center for the Study of Evolution in Action
2 Ecology, Evolutionary Biology, and Behavior, Michigan State University
3 Integrative Biology, Michigan State University
4 Computer Science and Engineering, Michigan State University
5 Biomedical Engineering, Michigan State University
6 Institute for Quantitative Health Science and Engineering, Michigan State University

†These authors contributed equally to this work.
* Email: alackes@msu.edu

Abstract

Understanding how genes interact with one another is crucial for understanding how genotypes become phenotypes. Genotypes with high levels of epistasis – that is, where the one gene's expression affects the expression of one or more other genes – are often found when organisms possess complex structures performing complex functions. Therefore, quantifying the degree of epistasis in a genome is crucial for our understanding of genomic complexity. However, in biological systems, it is often difficult to directly measure epistasis, since existing metrics rely on precise fitness metrics that are difficult or impossible to obtain. Here we propose a rank-based model of epistasis and test this model on several kinds of static and changing fitness landscapes. We find evidence that this rank-based model is effective at measuring epistasis in changing environments or newly adapting populations, but falls short when measuring populations at convergence. Nevertheless, this rank-based metric could be applied to biological systems in order to improve our understanding of epistasis on a locus and population level.

Introduction

Epistasis is where the expression of one gene affects the expression of one or more genes. Because of this, epistasis is a key component of the complex genetic architecture that affects fitness in biological and digital organisms. In evolving systems, epistasis can have a major effect on the fitness of the organisms [1]. As the evolutionary process progresses, accumulated mutations shape the fitness landscape [2]. Determining the loci which have the greatest effect on epistasis and the fitness landscape is a major goal for many researchers.

Epistasis can be measured in both digital and biological organisms. It has been shown in digital organisms, that while studying the mutational effects on epistasis in 'fit' and 'flat' populations, flat genomes showed no epistasis and fit populations showed positive epistasis [3]. The metrics used to calculate epistasis in digital evolution often rely on absolute fitness of the organisms which is routinely calculated in a computational evolutionary environment. In biological systems, absolute fitness is not always known and fitness is measured in relation to the other organisms in the population. To study epistasis in biological systems, it is common to generate known mutants and compare their fitness to the wildtype or other known mutants [4, 5]. Another approach is to use genome wide association studies (GWAS) to infer genetic interactions of two known fitness states [6, 7].

While the metrics developed in digital organisms are very effective in computational evolution, they may not be well suited for use in biology. In order to address this, we propose a rank-based epistasis metric that avoids the requirement of absolute fitness measurements of organisms. To test and further this metric, here we utilize the speed and flexibility of computational evolution. In this work we evolve bit-string genomes in different landscapes to access the applications and limits of this rank based metric. Effective implementation of this rank based epistasis metric will save researchers time and aid researchers studying epistasis in a more biologically relevant way.

Materials and methods

Rank Epistasis Metric

Evaluating organisms in a population on the rank epistasis model consists of four main steps. The current model assumes a binary genome (each site can only take on two values).

1. Rank organisms in the population. Each organism is compared to other organisms to produce a complete ranking. In this digital system we utilize each organism's calculated fitness score, but note that the absolute score is not necessary provided a complete ranking can be decided.

2. Acquire one-step mutants at one locus. Select a single locus in the genome. For each individual in the population, mutate the genome at that locus to produce a single-step mutant.

3. Rank mutants. Using the same procedure as in *1*, rank the new single-step mutants to construct another complete ranking.

4. Calculate the edit distance between the two rankings. Using the organism rankings from steps *1* and *3* as lists of individuals sorted by rank, calculate the edit distance between the two lists.

This allows us to calculate the epistasis at a single locus, compared to the per-organism calculations in other epistatic measurements [3, 4]. However, for the purposes of this study the per-locus epistasis is aggregated to get a measure of epistasis in the population at one point in time. Unless otherwise noted, here we use the Levenshtein edit distance. Preliminary testing shows promise in weighting the edit distances by the number of unique genomes in the population. These data suggest this additional factor helps observe effects of convergence, though it is not included in this work.

NK Fitness Landscapes

To evaluate the efficacy of this model in different scenarios, we used three distinct classes of fitness landscapes all based on the NK landscape model [9]. This model was chosen because the degree of interaction between sites in the genome can be adjusted at-will, allowing for an easy comparison between populations with different levels of epistatic interaction.

Classical NK

In a classical NK landscape, we generate a fitness table of N columns and 2^K rows with random values between 0 and 1 in each cell. Each locus in an organism's genome corresponds to one of the N columns. We combine that locus and the following $K-1$ loci to create a K-bit string. The fitness at that locus is then equal to the value in the column's cell corresponding to the bitstring. The fitness of an organism is the sum of the fitness at each locus. A complete explanation of NK fitness scoring can be found at [9].

NK Treadmill

One limitation of classic NK landscapes is that populations tend to converge rapidly. To examine constantly adapting populations, we use a modification of the NK landscape we call the *NK Treadmill.*

The NK Treadmill landscape is an oscillatory landscape with fitness values generated by a modified sine function we call the *triangle sine.* This modified function has consistent slope in contrast with the normal sine function, which has itself a constantly changing slope. The triangle sine is defined as:

$$tSin(x) = 0.25\pi \sum_{i}^{N} -1^{\frac{i(2i+1)}{2}} \sin((2i+1)x)$$

Each entry in the NK table is a tuple, (α, β), with both values between 0 and 1. For each group of K sites, fitness for the group is determined by the function:

$$W = \frac{(1 + tSin((\beta + 0.5)t + 2\alpha\pi)}{2}$$

where t is the current generation.

Fit vs. Flat

To investigate whether this metric can distinguish different levels of epistasis in differently adpated populations, we examined populations on an NK landscape with two fitness peaks: a high, ”fit” peak, and a lower, ”flat” peak, as in [2].

In this landscape, fitness score was based solely on the composition of groups of K sites and not their position in the genome. The score for each group of for $K = 3$ is shown in Table 1.

Bitstring	Score
0 0 0	1
0 0 1	0
0 1 0	2
1 0 0	0
1 0 1	2
1 1 0	0
1 1 1	1

Table 1. NK fitness scores for Fit and Flat landscapes.

Thus a genome of alternating bits has the highest score, but a single mutation creates a much larger impact to score, while a genome of uniform bits has a mediocre score but a single mutation creates only a small fitness impact.

We expect, then, that under high mutation rates, populations will converge at the lower peak, while under low mutation rates they will converge at the higher peak.

Comparison Metrics

To compare out metric to existing measures of epistasis, we also calculate epistasis in our populations using the metric discussed in Elena and Lenski (1997). A strong correlation between our rank-based metric and the epistasis measure in that paper, β, would provide evidence that we are directly measuring epistatic activity and not a related signal.

Evolutionary System and Analysis Pipeline

Experiments in this paper were conducted using the Modular Agent Based Evolver framework (MABE) [8]. All experiments used a $N = 200$-bit genome, and insertion-deletion (indel) mutations ranging in from 1 to 3 bits were used at rate 0.001, 0.01, 0.1 per locus. Populations consisted of 100 organisms and evolved for 2,500 generations. Tournament selection with a size of 7 was employed to select organisms for the next generation. Epistatic metrics were taken every 100 generations, and each experimental condition was conducted with 50 replicates.

All analysis was done using the R statistical computing language [10]. Plots were created using the ggplot2 R package [11]. Source code, data analyses, and additional figures are availble as supplemental material at `https://github.com/alackles/cse845`.

Results and discussion

Convergence muddles rank-based metric

The fit and flat fixed NK landscape provides little challenge to evolving organisms. As we expected, populations ascended to the highest, alternating-bit peak under low mutation rates (0.001) and settled in the lower, flat peaks under the highest mutation rate (0.1). Populations in many replicates converged, but more so in the low mutation rate treatment with an average of less than 16 unique genotypes at finish, compared to the two higher mutation rates that together average just over 90 unique genotypes.

The extreme case sees a population converging to a single genotype, and indeed this is often observed with the lowest mutation rate. In this case, mutating a single locus may cause a shift in fitness, but that shift is seen in all organisms and thus the ranking does not change. This yields an edit distance of zero. This should all be considered when looking at the fixed NK landscape results (see Figure 1).

At the 0.05 level, the distributions of edit distance at the end of each replicate was statistically significant in all pairwise combinations of mutation rates (using a paired Wilcoxon test for shared random number seeds and using Bonferroni correction for multiple comparisons). Between the 0.1 and 0.01 mutations rates $p = 3.151606e - 04$, between 0.1 and 0.001 $p = 5.329071e - 14$ and between 0.01 and 0.001 $p = 1.065814e - 14$.

Therefore these results do not match our original hypothesis; we expected to measure higher epistasis at the fit peak (lower mutation rate) than the flat peak. However, this is potentially due to the increased convergence in populations at lower mutation rates biasing the edit distance to be lower. Future work must be conducted to disentangle these effects.

Edit distance varies expectedly with K

Unlike the fixed NK landscape, the classical, randomly-generated landscapes provide considerable levels of noise and amount to more difficult landscapes for the evolving

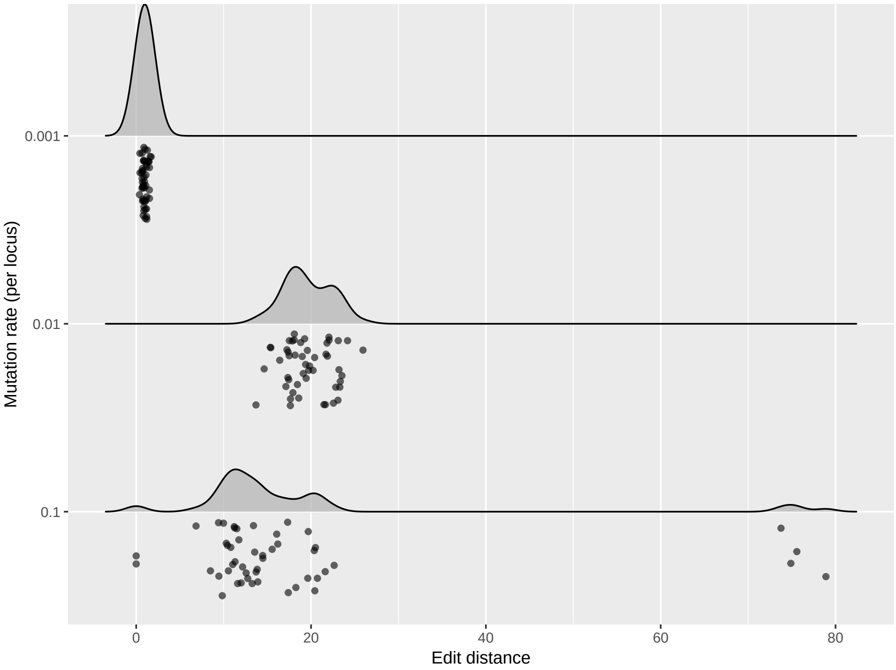

Fig 1. Raincloud plot showing the mean edit distance (averaged over all loci) for all replicates for the fixed NK "fit vs. flat" landscape at the end of the allotted 2,500 generations. For each mutation rate, the density of points is shown with the actual points plotted underneath for clarity.

organisms. These attributes make examining the rank-based metric more feasible in this domain, as populations are less likely to converge en masse.

Here we observe the same phenomenon at each mutation rate: increasing the value of K increases the edit distance where the population stabilizes. While this is seen in all mutation rates (though 0.1 is high enough to "boil" most genomes), we narrow in on the 0.01 per bit mutation rate, as shown in Figure 2 (all figures available in the supplemental material).

Edit distance is stratified according the value of K, which is what we expect due to the nature of NK landscapes. K is the number of consecutive bits used to encode a gene. Therefore, K directly affects the degree of epistasis; increasing K means that each bit will interact with more bits. This is indeed what we see, which is presented in Figure 3. The difference between K values at a given generation is statistically significant at the 0.05 level for all generations checked (every 500^{th} generation in the range [500, 2500], all $p \leq 2.872103e-09$, paired Wilcoxon test using Bonferonni correction for multiple comparisons). While further testing is required to see exactly how closely the rank-based metric tracks the K value of the landscape, this result provides early evidence that the metric measures some facet of epistasis.

Rapidly changing environments increase epistasis measurement

Running the treadmill NK landscape requires a parameter, velocity, and for this work we used three values of velocity: 0.1, 0.05, and 0.01. This parameter determines the speed at which the landscape changes, with higher velocity naturally causing more rapid changes.

Results from the 0.01 mutation rate and $K = 3$, selected because they show the *least*

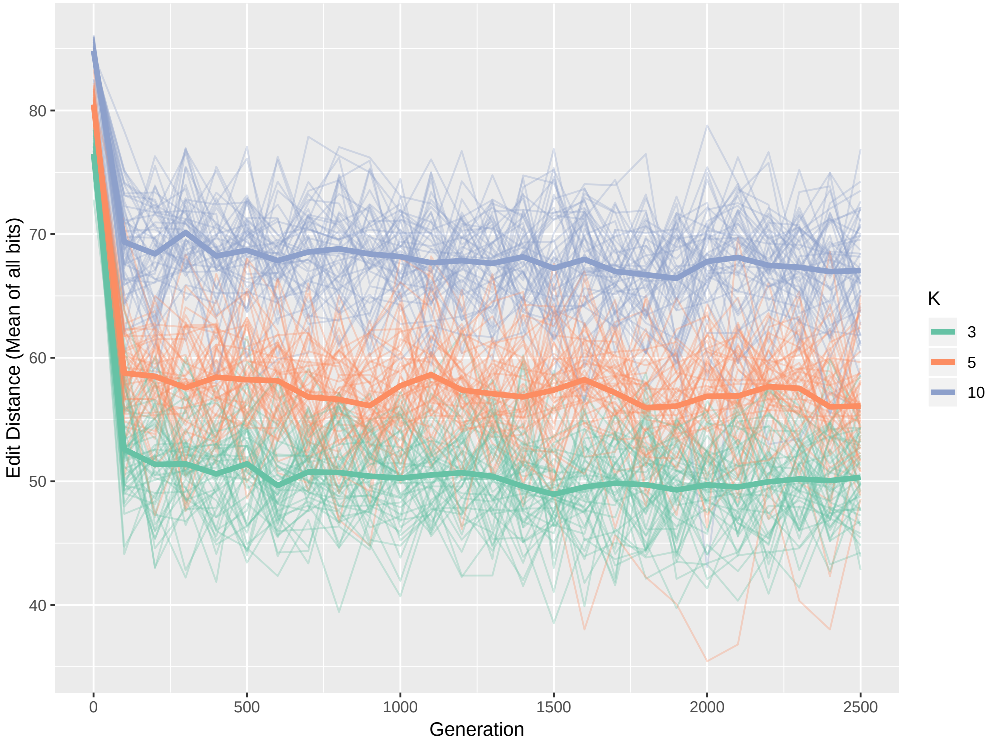

Fig 2. Edit distance of rank-based metric over time for random NK landscapes at a mutation rate of 0.01. Each replicate is shown as a semi-transparent line, colors show the different values of K, the number of bits in each gene. The grand mean of each K value is overlain as a thicker, opaque line.

effect compared to the moderate mutation rate and other K values, are shown in Figure 4. It should be noted that these results use the Damerau-Levenshtein edit distance, but preliminary tests show the difference is negligible. Here, all pairwise differences are significant at the 0.05 level (all $p \leq 0.007341238$) except for 0.1 and 0.05 velocity at 2000 ($p = 0.4511180$) and 2500 generations ($p = 0.2640401$, paired Wilcoxon test with Bonferonni corrections for multiple comparisons).

While an increase in velocity results in an increase in the rank epistasis metric, we hypothesize this is *not* a results of increased epistasis with higher velocity, but rather a decrease in convergence. As the landscape quickly shifts, the population must continually adapt. This continuous adaptation is likely to maintain more diversity than a single peak. However, further work is required to truly tell if the observed effect is due to increased epistasis, reduced convergence, or some combination of both.

Rank-based epistasis correlates with traditional measure when not converging

We also compared the rank-based epistasis metric to a known metric to look for correlation between the two. Specifically, we used the epistasis calculation described in [4]. In that work, the logarithm of fitness for a mutant k mutations away from the reference organism is described by the quadratic $-\alpha k - \beta^2 k$. β describes the non-linearity as we increase the number of mutations, making it a measure of epistasis. Therefore here we focus on the values of β.

Figure 5 shows the rank-based metric plotted against the known epistasis formula at

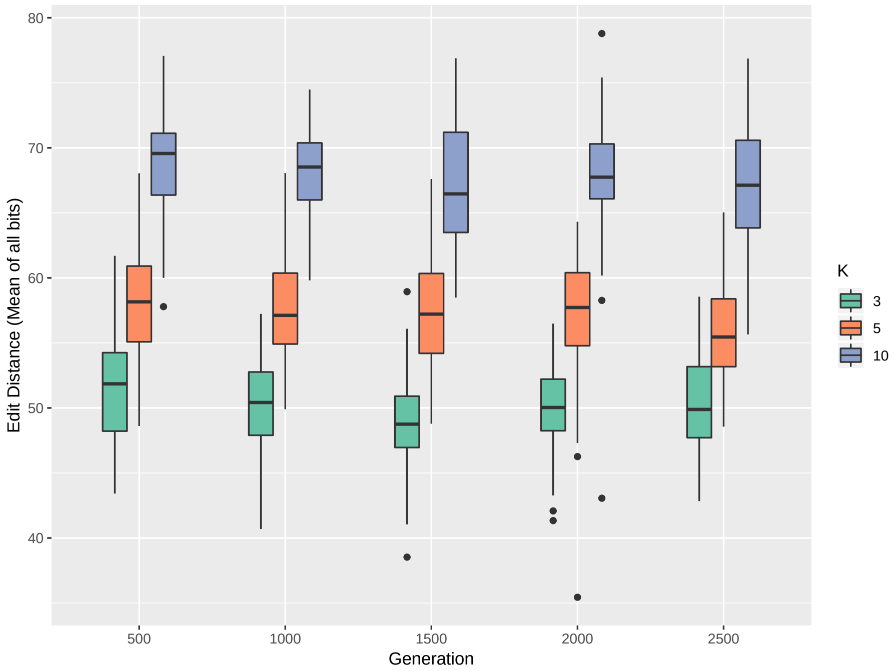

Fig 3. Boxplots showing the edit distance of the rank-based metric at several points in the evolutionary run and at three levels of K (3, 5, 10). Each boxplot represents the 50 replicates of that treatment at each time point.

generation 2,500 using the classical NK landscape as a representative sample. The 0.001 mutation rate shows again that the metric suffers when populations converge; with almost all points being near either no (0) or complete (200) epistasis. At the lowest mutation rate there is not significant correlation at the 0.05 level. The two higher mutation rates, 0.01 and 0.1, do not see genotypic convergence as strong as the 0.001 mutation rate. At these higher mutation rates we observe a positive correlation between the rank-based metric and the more traditional epistasis measurement. These correlations are weakly positive (Pearson's r; $0.34 < r < 0.4$) for the 0.01 mutation rate and moderately to strongly positive ($0.65 < r < 0.84$) for the 0.1 mutation rate (all $p \leq 1.527782e-02$).

This situation-depend correlation provides further evidence that the rank-based metric is in some ways measuring epistasis. More work is required, however, to understand under what scenarios this metric is appropriate and how it can be modified to handle more situations.

Conclusion

We assessed our newly developed rank epistasis model on three versions of an NK fitness landscape, each with varying results. This model shows promise for analyzing epistatic interactions at single loci for populations in active adaptation, but falls short when confronted with a population which is approaching convergence.

To extend this model for future use, we would want to develop a protocol for creating single-step mutants of genomes with more than two alleles. One could in theory create all single-step mutants and average the edit distance across all combinations of single-step mutants for every individual in the population, but this quickly becomes combinatorially infeasible. Other options include mutating the same site for all

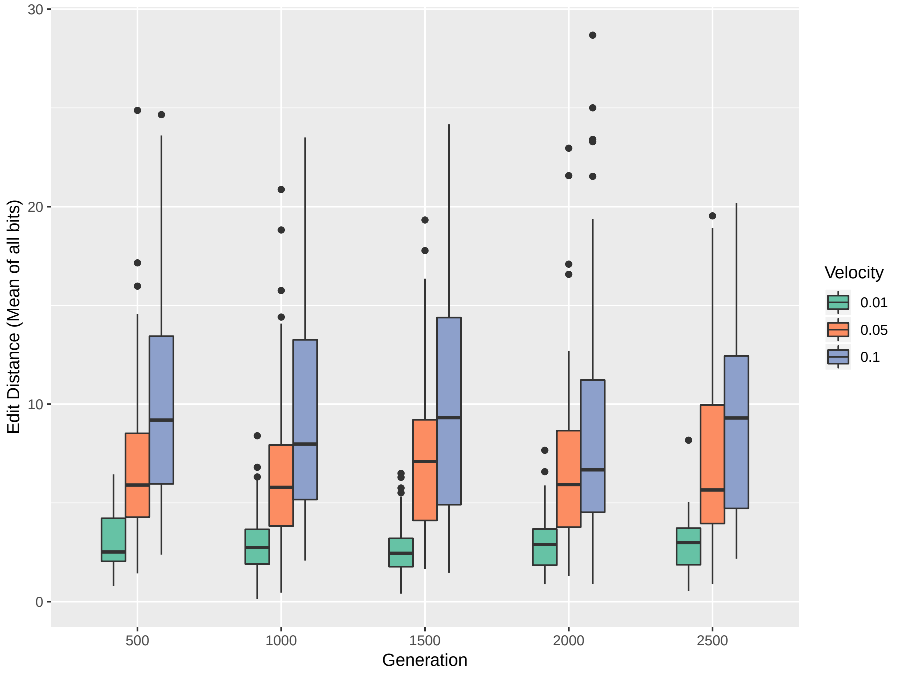

Fig 4. Boxplots showing the average edit distance per locus at several points in time. Three different values of velocity for the NK treadmill experiment are shown. Each box represent 50 replicates. Boxes shown are at the lowest mutation rate (0.001 per locus) and at $K = 3$.

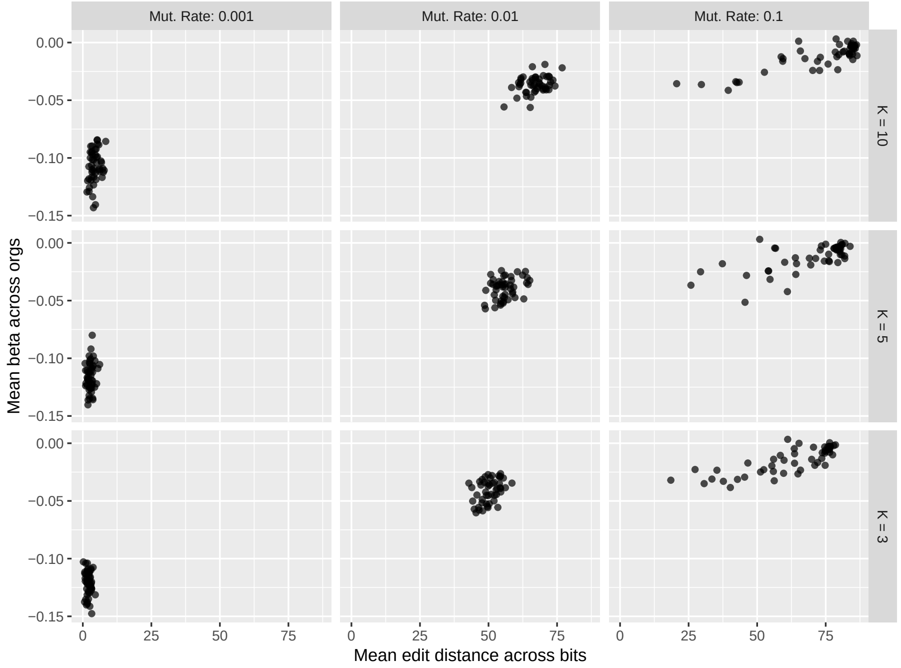

Fig 5. Each point represents one particular replicate at generation 2,500 from the classical NK landscape treatment. The x-axis was calculated using the new rank-based metric, while the y-axis was calculated using the epistasis formula from [4].

individuals to the same allele and taking some average of all possible alleles. Further work will need to be done to determine what type of mutation schema is both informative and computationally reasonable.

A model which allows for multiple kinds of single-step mutations would be applicable to many questions in biological systems. Generating a full landscape of single-step mutations is rapidly coming within reach of wet lab systems. Some work has been done, for example, on mapping the full fitness landscape of cells in cancerous tumors [13]. Applying rank epistasis metrics to such rapidly adapting biological populations can help us understand which sites are biologically important, and how these sites interconnect with the rest of the genome.

Understanding epistasis at the site-by-site level without need for a precise calculation of an individual's overall fitness can help us understand genomic interactions at greater depth and greater scale. By examining epistatic interactions at the individual locus level as it relates to the population level, we can expand our understanding of genome dynamics and learn more about the complex systems and structures of biological genetics.

Acknowledgments

The authors would like to thank Clifford Bohm for assistance in designing this study. This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. 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 National Science Foundation. Michigan State University provided computational resources through the Institute for Cyber-Enabled Research.

References

1. Ferretti L, Schmiegelt B, Weinreich D, Yamauchi A, Kobayashi Y, Tajima F, Achaz G. Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. J Theor Biol. 2016;396:132–143.

2. Wilke C, Wang J, Ofria C, Lenski R, Adami C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature. 2001;412(6844):331–333.

3. Franklin J, Labar T, Adami C. Mapping the Peaks: Fitness Landscapes of the Fittest and the Flattest. Artif Life. 2019;(25)3:250–262.

4. Elena S and Lenski R Test of synergistic interactions among deleterious mutations in bacteria. Nature 1997;390(6658):395-8.

5. Trindade S, Sousa A, Xavier K, Dionisio F, Ferreira M, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. PLoS Genet. 2009;5(7):e1000578.

6. Ritchie M. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. Ann Hum Genet. 2011;75(1):172–182.

7. Tian F, Bradbury P, Brown P, Hung H, Sun Q, Flint-Garcia S, Rocheford T, McMullen M, Holland J, Buckler E. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genet. 2011;43(2):159–162.

8. Bohm C, C G N, Hintze A MABE (Modular Agent Based Evolver): A framework for digital evolution research Proc 14th Artif Life Conf. 2017;76–83.

9. Kauffman S, Levin S. Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol. 1987 128(1):11–45.

10. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2017.

11. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag.; 2009.

12. R Core Team R: A language and environment for statistical computing. (2013).

13. Rogers Z, McFarland C, WInters I, Seoane J, Brady J, Yoon S, Curtis C, Petrov D, Winslow M. Mapping the in vivo fitness landscape of lung adenocarcinoma tumor suppression in mice. Nature Genet. 2018;50:483–486.