A COMPUTATIONALLY TRACTABLE MODEL FOR THE EVOLUTION OF A GENOME'S SIZE AND ITS GROWING FITNESS LANDSCAPE

August 26, 2025

Master's Thesis
Faculty of Science, University of Bern

Handed in by
Jacob Hugh Riina

Supervisors
Prof. Dr. Claudia Bank
Dr. Stephan Peischl

Introduction

Genome Evolution

Since life's origins approximately 3.5 billion years ago, it has generally increased in complexity, and understanding the processes that generate this complexity is one of the main goals of evolutionary biology. One such mechanism of generating complexity is through genome evolution: by changing the size of their genomes, the earliest living cells - those with a cell membrane, which encode and pass on information through their DNA - were able to create *evolutionary innovations*: new metabolism, physiology or regulation, which augumented their ability to thrive in a variety of environments. At the increased cost of replicating and proofreading more DNA, organisms have since been able to evolve new genes allowing them to metabolize different carbon, nitrogen, or complex nutrient sources, various types of receptors and secretion systems to interact, competitively or mutualistically, with others, and the internal regulatory networks required for them to decide which behaviors are optimal under which circumstances, or even to coordinate multicellularity. Several billion years later, we see the consequences of this incessant evolutionary process: the stunning diversity amongst organisms in the tree of life, each a product of the series of environments they have adapted to. We can say, then, that the content of an organism's genome is not solely the result of how it has evolved to fit its current environment, but also the entire history of environments that its ancestors have evolved in; as always, "Nothing in biology makes sense except in the light of evolution".

Understanding this facet of evolution is [difficult]. Unlike with [allelic mutation], genome evolution occurs up to several orders of magnitude less freqently, and is more heterogenous in its manifestation. We define here genome evolution as a mutational process resulting in either the expansion or reduction in the size of the functional genome, proceeding from one of several causes: among genome-increasing mechanisms, there exist duplications from repeat region replication errors, transposon activity, or whole genome duplications, as well as genes arising from noncoding sequence through de novo gene birth, while among genome-reducing mechanisms are pseudogenization and deletion via (some recombination method). Owing to a large protein fold space, evolution has generally favored mechanisms of duplication and neofunctionalization amongst new genes through horizontal gene transfer, transposon activity, or whole-genome duplications. (by doing so the cell acheives two things: 1. relieve pleiotropic effects, 2. lessen time to navigate sequence space). Alternatively, unused regions of the geome may also see genes spontaneously arise. (this process is much slower and harder to observe). In either case, these new genes are expected to accumulate neutrally, and over time have the opportunity to gain beneficial functionality. Although organisms are effective at removing unnecessary genes through pseudogenization, (this is a slow process, and still does not account for historical contingencies) Different organisms may happen to adapt to the same environment in different way

We do however see in even shorter evolution experiments that not only mutation of alleles, but also genomic abberations appear that explain increases in fitness. (noted rate of nucleotide gain per generation stats here)

Fitness Landscape Theory

As has been noted by studies of the *de novo* evolution of genes, their emergence out of unused genetic substrate remains difficult to observe experimentally. (why? duplication vs de novo navigation of landscapes). There is thus a usefulness to generation of models that allow us to form hypotheses about these difficult to test systems.

Although fitness landscape theory has been used extensively to model adaptation on static genomes, it is less well developed on the scale of long term evolution, where changes in the genome size also arise. With this work we extend the Rough Mount Fuji model to incorporate genome evolution alongside allelic mutation

Methods

The Rough Mount Fuji Model

The Rough Mount Fuji (RMF) model allows for tunable epistasis [1] and so enables the investigation its effects as landscapes scale. In this model we represent genotypes as binary vectors, and their permutations in genotype space construct a network where each has a hamming distance of 1 from its neighbors, and the degree of each node is equivalent to the number of loci L in the model (see Figure **??**).

The assignment of fitness values to genotypes entails sampling two probabilistically determined components: Locus-wise additive effects $a = (a_1, \cdots, a_i) \sim$ iid. $\mathcal{N}(0, 1)$ are drawn from a normal distribution, and each genotype g recieves a unique epistatic component $e_g \sim \mathcal{N}(0, \sigma_e)$, also drawn from a normal distribution. Fitness values for some arbitrary genotype can then be assigned as the sum of the product of the additive effects and the genotype, plus the epistatic component, as an exponential:

$$f_g = exp(e_g + \sum_{i=0}^{L} a_i * g_i)$$

Epistasis is represented in the ratio between the variance of additive components and epistatic components, and we control epistasis on the landscape by varying the epistatic σ_e - in cases where the epistatic component is small the additive component dominates and the landscape becomes easily predictable, whereas when epistatic effects are large and outweigh the additive component, there is virtually no correlation between different genotypes and ruggedness is high. Now able to reproducibly retrieve some epistatic component for any genotype by providing the CBRNG with the base-10 representation of that binary genotype, we are no longer obligated to generate fitnesses for all genotypes on the landscape in advance and instead need only do so upon encountering that genotype during simulation.

A discrete-generation RMF model with Wright-Fisher dynamics was implemented based on [2] and previous software [3]. We model a static N total individuals mutating and reproducing simultaneously in each generation, with genotype frequencies governed by binomial and multinomial sampling, respectively. To initialize the simulation, we specify the number of loci in the model and consequently the length of the vector of additive effects to generate, the number of randomly distributed initial genotypes on the landscape and the number of individuals per genotype, the simulation length, the variance σ of the epistatic component, and the mutation rate μ - the probability for an individual to generate a mutant of hamming distance 1 from its parent. An initial report table is then generated with a description of initial genotypes on the landscape, their population sizes, and fitnesses. We proceed to instantiate subsequent generations by first generating mutants from each extant genotype with population $\mathbb{P}$ through binomial sampling: Binomial$(n = \mathbb{P}, p = \mu)$. We adjust the corresponding population counts and calculate fitnesses for new genotypes if necessary, and then assign from i extant genotypes for all individuals in the new generation of size N new genotypes according to relative fitness: Multinomial$(N, p' = (p'_1, \cdots, p'_i))$ given $p'_i = p_i * \frac{\omega_i}{\bar{\omega}}$, where p_i is some genotype's current frequency, ω_i is its fitness, and $\bar{\omega}$ is the average fitness across all extant genotypes. Therefore, highly fit mutants gain an advantage during sampling and are less likely to be eliminated through drift. The simulation is run for the specified number of generations and upon terminating a final report is also provided (Figure **??**).

Counter-Based Random Number Generation

The facile generation of random epistatic components was performed using a Julia implementation [4] of the Random123 package from D.E. Shaw Research [5]. Particularly, the ARS1x generator was used, with a period of 2^{128}, this being the number of loci that can be modeled in a genotype network without reseeding.

Scripting

Results

Additive Limit

Uncorrelated Limit

Time-Dynamic Outputs

Endpoint Outputs

Does endpoint exist under this model?

Discussion

Code and Data Availability

References

[1] Johannes Neidhart, Ivan G Szendro, and Joachim Krug. "Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model". In: *Genetics* 198.2 (Oct. 2014), pp. 699–721. ISSN: 1943-2631. DOI: 10.1534/genetics.114.167668. (Visited on 09/16/2024) (cit. on p. 2).

[2] Juan Li, André Amado, and Claudia Bank. "Rapid Adaptation of Recombining Populations on Tunable Fitness Landscapes". In: *Molecular Ecology* 33.10 (2024), e16900. ISSN: 1365-294X. DOI: 10.1111/mec.16900. (Visited on 01/22/2025) (cit. on p. 2).

[3] André Amado, Juan Li, and Claudia Bank. "STUN: Forward-Time Simulation on TUnable fitNess Landscapes in Recombining Populations". In: *Bioinformatics Advances* 3.1 (Jan. 2023), vbad164. ISSN: 2635-0041. DOI: 10.1093/bioadv/vbad164. (Visited on 09/16/2024) (cit. on p. 2).

[4] *JuliaRandom/Random123.Jl*. JuliaRandom. Apr. 2024. (Visited on 01/21/2025) (cit. on p. 3).

[5] John K. Salmon et al. "Parallel Random Numbers: As Easy as 1, 2, 3". In: *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '11. New York, NY, USA: Association for Computing Machinery, Nov. 2011, pp. 1–12. ISBN: 978-1-4503-0771-0. DOI: 10.1145/2063384.2063405. (Visited on 01/21/2025) (cit. on p. 3).