Contingency vs. Convergence: Evolutionary Dynamics in an Epistatic Landscape
Maths 242 Project Proposal

[redacted]

December 12, 2021

Abstract

We propose a tentative framework of studying historical contingency and convergence depending on the fitness landscape.

1 Introduction

Does evolutionary dynamics repeat itself if we "replay the tape of life"? [Gou90]. The full answer to this question is unsurprisingly mixed and nuanced. On one hand, paleontologists and microbiologists have observed that different species or populations can develop very similar complex traits independently, eg. the biological eye [Mor06], or different species of yeast developing the same kind of behavior in the Long-Term Evolution Experiment (LTEE) [GMB+17]. Such phenomenon of path-independent evolutionary trait in the long run is coined as convergence, or parallelism in microbial evolutions (there are subtle differences between the two concepts, which go beyond the scope and purpose of this project).

On the other hand, people have observed historical contingency in evolutionary trajectories, (eg. rare mutations open up new evolutionary pathways and thus certain complex trait is path-dependent). This observation was most famously (or rather infamously) theorized by Gould as "punctuated equilibrium". Historical contingency has also been observed in microbial laboratories: eg. Lenski's LTEE revealed a background-dependent rare mutant of *E. coli* can utilize citrate as carbon source, which wild-type *E. coli* cannot do [BBL08].

There is thus a long history of debate between contingency and convergence as the dominant driving force on evolutionary dynamics. It is known that evolutionary dynamics is essentially governed by parameters including the population size, mutation rate per individual U and the fitness landscape $X(\vec{g})$, where $X(\vec{g})$ denotes the fitness of a specific genotype $\vec{g}$. Perhaps surprisingly, rich and complex dynamics can happen even with a simple fitness landscape without epistasis, for example in a small population where drift is a dominant force.

For our purpose, we want to single out the effect of the fitness landscape and ask how contingency or convergence phenomenon would change based on the "ruggedness" of the fitness landscape. Therefore, we would choose the parameter regimes that ignore drift ($N \to \infty$) and maybe with strong selection and weak mutations (so that the selective adaptive walk is most pronounced), and still maintain somewhat realistic and relevant for real organisms.

Even with simplifications of the model, the task is still difficult due to the extremely high dimension of genotype space. In the following chapters, we introduce a framework of exploring the question of historical contingency depending on the fitness landscape. The main steps involve identifying a rugged fitness landscape simple but good enough to generate contingency behavior, choosing workable models to describe the evolutionary dynamics and finally quantifying the contingency phenomenon with some metric.

2 Choice of Evolutionary Dynamical Models

There exists a large set of viable models to describe evolutionary dynamics, depending on the level of details in the problem. Of these, two specific models stood out and we will focus on them in this chapter.

2.1 Quasi-Species Equation

The quasi-species model was first developed to describe the evolution of an infinite population of DNA sequences [EMS89]. If we denote f_i the frequency of the genotype i, Q_{ij} the mutation matrix and X_i the fitness of genotype i (which is taken to be fixed constant), then the time dynamics of f_i is

$$\frac{df_i}{dt} = \sum_{j=1} f_j(t) X_j Q_{ji} - f_i \bar{X} \tag{1}$$

where $\bar{X}(t) = \sum_i f_i(t) X_i$ is the average fitness of the population. Notice that we depart from the traditional notation for the fitness and frequency in the literature, to be consistent with the fitness class model we will introduce later.

The mutation matrix is easy to be calculated from per-site mutation rate μ as

$$Q_{ij} = \frac{\mu}{|\mathcal{A}| - 1}^{d(i,j)} (1-\mu)^{L-d(i,j)} \tag{2}$$

where $d(i,j)$ is the hamming distance between the two genotypes and L is the total sites of the genome, and $|\mathcal{A}|$ is the genetic alphabet cardinality (eg. $|\mathcal{A}| = 4$ for the real genetic code "A,T,C,G").

Even without drift, the full dynamics of the quasi-species equation is quite complicated as the average fitness depends on f_i. However, the equilibrium distribution is easy to obtain since it is equivalent to solving an eigenvalue problem. This will aid us to characterize the fitness landscape given the equilibrium distribution f_i^*, as we will see later.

We can perhaps only keep track of the single or double point mutations (i.e. $d(i,j) \leq 2$ as the multi-point mutation becomes exponentially improbable. This will simplify the mutation matrix and thus might aid us to write down the full dynamics of the quasi-species equation.

2.2 Distribution of Fitness Classes

The quasi-sepcies equation has the benefit of completely characterizing the evolutionary dynamics of each genotype but it may not be the most feasible model due to the large dimesion of the genotype space and degeneracy in fitness. Here we consider instead the evolution of fitness classes, where the genotypes are grouped together according to their fitness strengths.

The dynamical equation for fitness distribution $f(X,t)$ can be written as

$$\frac{\partial f(X,t)}{\partial t} = (X - \bar{X}) f(X,t) + U \int \rho(s)(f(X-s,t) - f(X,t))ds + \sqrt{\frac{f}{N}}\eta(X,t) \tag{3}$$

where again $\bar{X} \equiv \int X f(X,t) dx$ is the average fitness, and the selection term is the same as before, except grouped with the same fitness classes. Notice however the mutation term is easier than the quasi-species equation, and the total effect of mutation matrix and fitness landscape is combite into the distribution of fitness (DFE) $U\rho(s)$. Lastly, the drift term is a complicated object that depends on the population size and is constrained to fix the population size to be constant. In the large population (deterministic) limit, we can neglect the drift term and focus on the quantity $\langle f(X,t) \rangle$ averaged over the statistical ensemble.

Using this fitness class model, we can easily compute the average fitness dynamics

$$\bar{X}(t) = \int s U \rho(s) p_{fix}(s) \tag{4}$$

where $p_{fix}(s)$ is the fixation probability of a mutation of effect size s that can be calculated from the evolutionary equation. Thus given a hypothesized DFE $U\rho(s)$, we can compare the theoretical

prediction with the experimental fitness trajectory data (similarly with mutation accumulation), which was described in [GD14].

The main challenge, though is to relate the DFE to the epistatic structure of the fitness landscape in a useful way, which we will explore next.

In a nutshell, the quasi-species equation has main advantage of characterizing the deterministic dynamics completely for genotypes given a fitness landscape but can be very inefficient due to degeneracy; whereas the fitness class model presents exactly the opposite challenge. For our purpose of the project, choosing models depending on the fitness landscape perhaps is the right path forward.

3 Characterizing the Fitness Landscape

The first step is to construct a realistic fitness landscape that could generate the behavior we desire.

3.1 Fitness Landscape Inference

To get a sense of what a real fitness landscape could look like, we first turn to a case study of inference from the viral population data.

[SDGM+15]Seifert et. al used the Swiss HIV Cohaort Study data from two patients and tried to infer from the sequencing genotypic data what their corresponding fitness is. They assumed that by the time of sequencing, the viral evolution has reached equilibrium distribution governed by the quasi-speicies equation. Therefore, the equilibrium distribution is given by Equation (1) as

$$\bar{X}(f, X)\vec{f^*} = Q^T \text{diag}(X)\vec{f^*} \tag{5}$$

Thus the fitness landscape vector $\vec{X}$ can be inferred by solving the linear algebra equation. Since the absolute magnitude of fitness does not matter in the equilibrium, we can assume $\bar{X} = 1$ in equilibrium, thus removing one degree of freedom. Then the fitness is given by

$$\vec{X} = \text{diag}(f)^{-1} Q^{-1} \vec{f^*} \tag{6}$$

With Bayesian MCMC sampling technique (omitted here), the fitness landscape thus can be obtained and it is visualized in Figure 1.Thus we see that there is a vast majority of nearly neutral haplotype network, connected with a few really fit mutants in the fitness landscape.

The biggest challenge of this method of inference is exactly the assumption of reaching evolutionary equilibrium for the viral population, which might not be so bad after-all if we assume this is some sort of quasi-equilibrium.

Another step towards analyzing the epistatic structure of the inferred fitness landscape is to identify the quantitative trait locus or measure the hamming distance between different mutants to get a better sense of how the really fit mutant emerges from epistatic interactions. This hopefully would aid us to construct a realistic model of fitness landscape.

3.2 Microscopic vs. Macroscopic Epistasis

In the two preceding models of evolutionary dynamics, we have identified $\vec{X_i}$ (fitness vector) and $\rho(s)$ as describing the action of fitness background. Thus the crux of the challenge, if using the fitness class model, is to relate $\rho(s)$ to $\vec{X_i}$ that describes the genotype fitness. Traditionally the epistatic interactions refer to the fitness effect of certain mutation being dependent on its genetic background. However, this does not directly translate to a non-constant DFE in the fitness class model. Good et. al [GD14] especially distinguished the two as microscopic and macroscopic epistasis. They showed that even without microscopic epistasis, there can still be seemingly epistatic phenomon such as the diminishing return of fitness improvement due to "running out of beneficial mutation" effect. Similarly, we can equally imagine a world where $\rho(s)$ does not change with microscopic epistasis.

In full characterization, the DFE $\rho(s|\vec{f})$ depends on the distribution in the genotype space $\vec{f}$. To simplify, in the literature of uncorrelated landscape people assume that $\rho(s|\vec{f}) = \rho(s|\vec{X})$ where $\vec{X}$ is the distribution in the fitness space rather than the genotype space. Much progress then can be made by assuming certain structure of $\rho(s|X)$, (eg. exponential due to extreme value statistics assumption [Orr03] or diminishing return dependent on X [WRL13].However, to model the contingency

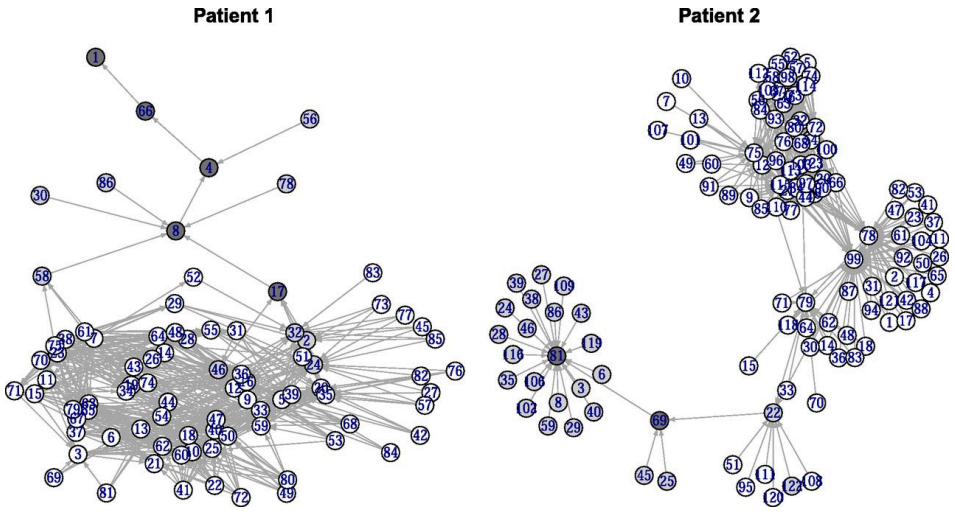

Figure 1: **Fitness landscape inference from [Orr03]**.Rank fitness landscapes of the haplotypes in each of two patients. A directed edge $i \to j$ exists between haplotypes i and j if the posterior fitness difference $X_j - X_i$ can credibly be inferred to be > 0, i.e., if, given the model, there is evidence for haplotype j being fitter than haplotype i. Both graphs possess the transitive property; i.e., if j is fitter than i (indicated by an edge $i \to j$) and k is fitter than j, then k is also fitter than i and a directed path exists from i to k. Dark gray vertices possess credibly larger than average, light gray vertices possess average, and colorless vertices possess lower than average fitness $\bar{X} = 1$.

phenomenon, we really need the details of microscopic epistatisis probably localized in some region of the genotype fitness landscape, but not necessarily localized in $\rho(s|\vec{X})$, as we can imagine there is some lower fitness plateau followed by a sudden jump to a really fit mutant in the nearby genotype space due to strong microscopic epistasis (Figure 2).

Therefore, modeling contingency phenomenon needs a smart combination of thinking about both the microscopic and macroscopic epistasis.

3.3 Criteria of a Contingency Fitness Landscape

We now turn to a first attempt at characterizing the fitness landscape for contingent adaptations. As argued by [BBL08] contingent adaptations should tend to be complex and require multiple steps, some of which might not even be beneficial given other more fit evolutionary pathways.

Furthermore, the prevalent microscopic epistasis besides such rare and pronounced contingency epistasis can be tuned antagonistically, as described by [DWF07] as a more realistic model for long-term evolution.

Intrigued by the visualization of haplotype network, we attempt to draw here a mutation network in the fitness landscape to describe in a very conceptual level how contingency could happen (Figure 2).

3.4 Toy Model for Contingency Fitness Landscape

Inspired by the conceptual visualization, we try to come up with concrete construction of epistatic structure. Stuart Kauffman's NK model is a simple way of generating random fitness landscapes of tunable ruggedness, which can be written as

$$X(\vec{a}) = \sum_{i=1}^{L} b_i((a_j)_{j \in \vec{e_i}}) \tag{7}$$

where $\vec{a}$ denotes the genotype of L sites and b_i denotes the fitness contribution of allele a_i, and e_i is the interaction structure of locus i.

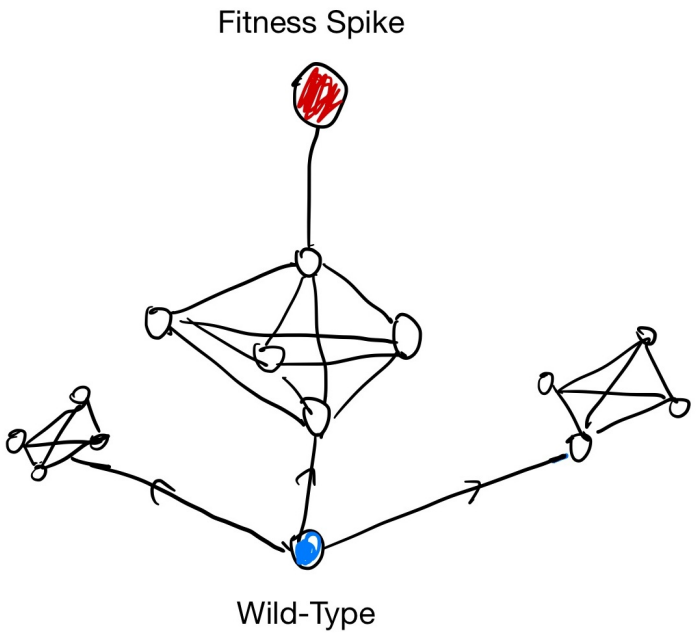

Figure 2: **Conceptual fitness landscape for a contingent mutation.**The rare yet really fit mutation (colored red) is dependent on which nearly-neutral haplotype networks evolution chooses among its potential pathways.

In addition to this, we can add some really large fitness peaks adjacent to a large nearly-neutral haplotype network in an ad-hoc way, just as depicted in Figure 2. The exact details have not been thought through, but it might simply involve tuning the b_i's in some region.

Once the fitness landscape is complete, we need to translate it into the macroscopic DFE $U\rho(s|\vec{f})$, which again is a formidable task. This can be done in a perturbative way perhaps.

Finally, we need to characterize the ruggedness of the landscape using a viable metric.

4 Quantifying the Contingency Phenomenon

After generating a realistic yet simple fitness landscape for the contingency dynamics, we now should quantify exactly what we mean by historical contingency. In the deterministic model, there is no such sense of "replaying the tape of life" since by defination it is deterministic. So we can turn to the leakage rate to the fitness spikes as a way of describing the contingency phenomenon. For example, we can calculate the leakage rate given initial population centered at some random point in the fitness landscape, versus the leakage rate given the initial population centered exactly nearby the fitness spikes (Figure 2).

The exact metric should be carefully designed and requires more reading into the literature.

5 Summary

In conclusion, interesting phenomenon of contingency and parallelism have been observed in evolutionary dynamics, but it remains a challenge to quantitatively characterize them depending on the fitness landscape. In this project we aim to generate simple and realistic fitness landscape in describing the contingency and hope to discover a law that relates the contingency to the ruggedness of the fitness landscape. To do this, we pick the fitness class model to describe the evolutionary dynamics, but have to work hard to get the distribution of fitness effect (DFE) right that describes the microscopic epistatic structure. Simplifying our model is key to the success of this project and we welcome any suggestion on how to streamline the assumptions and neglect irrelevant details.

References

[BBL08] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. 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, 2008.

[DWF07] Michael M Desai, Daniel Weissman, and Marcus W Feldman. Evolution Can Favor Antagonistic Epistasis. *Genetics*, 177(2):1001–1010, 10 2007.

[EMS89] Manfred Eigen, John McCaskill, and Peter Schuster. The molecular quasi-species. *Adv. Chem. Phys*, 75(149-263):6, 1989.

[GD14] Benjamin H Good and Michael M Desai. The Impact of Macroscopic Epistasis on Long-Term Evolutionary Dynamics. *Genetics*, 199(1):177–190, 11 2014.

[GMB+17] Benjamin H. Good, Michael J. McDonald, Jeffrey E. Barrick, Richard E. Lenski, and Michael M. Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, Nov 2017.

[Gou90] Stephen Jay Gould. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company, 1990.

[Mor06] Simon Conway Morris. Evolutionary convergence. *Current Biology*, 16(19):R826–R827, 2006.

[Orr03] H Allen Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526, 2003.

[SDGM+15] David Seifert, Francesca Di Giallonardo, Karin J. Metzner, Huldrych F. Günthard, and Niko Beerenwinkel. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, 199(1):191–203, 2015.

[WRL13] Michael J Wiser, Noah Ribeck, and Richard E Lenski. Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367, 2013.