Chapter Four

Evolutionary Dynamics: Mutations, Selection, and Diversiy

4.1 INTRODUCTION

The modern study of evolution via natural selection is now more than 150 years old. Yet the process of evolution is many billions of years old. Evolution has shaped the emergence of life, the diversification of a multitude of forms, and is essential to understanding global-scale challenges ranging from adaptation to climate change to the spread of antibiotic resistant pathogens. Yet the core mechanisms of how evolution works remains controversial to some, and poorly understood by others – both within and without the scientific community. Hence, this chapter aims to explain how evolution works, both as a 'historical' as well as an ongoing process.

To begin, it is worth recalling a few key definitions:

Evolution : Any change in the genetic makeup of individuals in a population over subsequent generations.

Evolution via natural selection : Any *nonrandom* changes in the genetic makeup of a population due to the *differential reproduction/survival* of the individuals.

The differences are important. Organisms have genomes that contain the code of life. These genomes are usually DNA-based (for cells and viruses) and sometimes RNA-based (for viruses). The sequence of chemical letters that make up genomes are copied during reproduction. The copying process is done with high fidelity, but it is not perfect. The error rate of genome replication varies with organism and, indeed, with genome length of organisms. Hence, there is a chance that the genome of an offspring will have a different genome than that of their parent. For haploid organisms (i.e., with with a single copy of their genome), then mutations alone are sufficient to generate differences between mother and daughters. There are many other factors that influence the genetic make-up of organisms at the molecular level. For example, large sections of genomes can be deleted, inserted, swapped with other genomes. Moreover, for organisms with more than one copy of each genome – like diploid sexual organisms – then the offspring genomes reflect site-by-site differesnce in the combinations of copies passed on from the two parents.

The list of processes above is partial. But this chapter is not meant to review all the ways in which genomes can change upon reproduction. Irrespective of the mechanism, the question that is central to this chapter is simply: how do changes in a genome translate into changes in the *frequencies* or *numbers* of distinct "genotypes" in a population? This question can be addressed in many ways. This chapter will introduce mathematical models in concert with *experimental* approaches to address the problem of assessing evolutionary dynamics. The rationale for doing so is twofold. First, direct tests of evolution in action provide the context to probe both changes and consequences over time. How fast does evolution operate, what are the shape of fitness landscapes, and are populations limited by mutations or, instead, by the competition between many paths to adaptation? Second, by archiving samples in evolution experiments, it is also possible to probe the udnerlying genomic mechanisms underlying observed changes, a process that has become both increasingly accessible and indeed common given the ongoing revolution of sequencing in the past 30 years.

The popularization and dissemination of the concept that studies of the evolution at the population scale could be an experimental science is due, in no small part, to the seminal work of Richard Lenski (Lenski et al. 1991; Elena and Lenski 2003; Sniegowski et al. 1997; Lenski and Travisano 1994; Wiser et al. 2013; Good et al. 2017). Lenski and his team have now studied the evolutionary dynamics of bacteria for over 60,000 generations. At the outset, Lenki's key insight was to recognize that many of the principles underlying evolution via natural selection could be probed, in a forward sense, by growing an organism – *Escherichia coli* – in relatively, simple conditions: a shaken flask. The bacteria were inoculated in DM25 growth medium – a media in which the principle glucose carbon source is available at relatively low concentrations. The media also contained small amounts of citrate – a point that will have relevance in a moment. The bacteria grow for approximately 6-7 generations per day, and then 1% of the population is transferred to fresh media the subsequent day. In addition, the population is archived on a regular basis. This process of growth and diluation has continued day after day, month after month, and now year after year for more than 60,000 generations – a span of over 1 million human years (if we are to translate, somewhat dubiously, this time-scale into anthropogenic terms). The "Long-term Evolution Experiment" (or LTEE) has enabled many levels of insights into how evolution works in practice as well as stimulated new methodological developments to probe genomic variation amongst individual cells in a population.

The result has been remarkable in many ways. First, as was expected, the bacteria adapted to their new environmental conditions. By "adapted", the bacteria grow faster and evolved to have higher fitness than do their ancestors (see Figure 4.1). In essence, the bacteria have changed over time, often in rapid transitions, Second, the advantage of performing experimental evolution is that the genomic basis for such evolution can be probed systematically, by sequencing and comparing changes in genotypes across thousands of generations. In doing so, it is also possible to ask: what is the shape of fitness landscapes, locally at

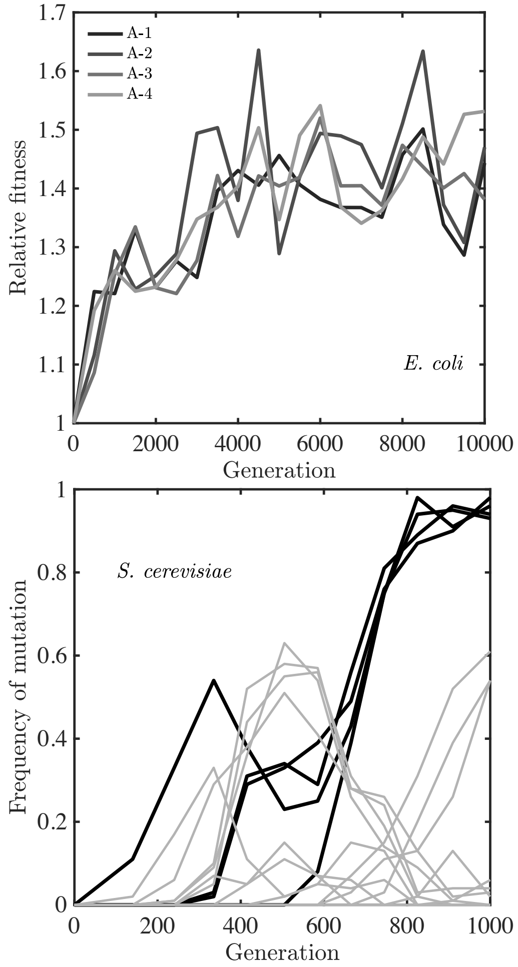


Figure 4.1: Reproducibility and chance in evolutionary dynamics. (Top) Relative fitness of evolved *E. coli* strains compared to ancestral strain in the first 10,000 generations of the long-term evolution experiment (Lenski and Travisano 1994); (Bottom) Relative frequency of different mutations in the first 1200 generations of an evolving *S. cerevisiae* population (Lang et al. 2013).

least. That is to say: given a particular genotype, how many nearby mutations lead to organismal death, to increased fitness, or to organisms with slightly worse fitness than the 'wildtype'. A long line of papers nad reviews of the LTEE are available elsewhere, and hopefully this brief introduction whets your appetite to learn more (e.g., (Kawecki et al. 2012; Good et al. 2017; Blount et al. 2018)).

Over time, the emergent understanding from the LTEE was that a bacterial genotype would adapt increasingly slowly to present conditions. The adaptations are enabled by mutations that enhance fitness, e.g., improving uptake mechanisms, shortening the lag period, and other means of growing in the same, fixed environment. In the same vein, the work of many others have branched out from this initial premise. That is: take a microbe, choose a condition, let the microbe grow again and again in this condition, and then use the changes in the genotype and perhaps even the diversifying community to understand how evolution works now and perhaps even some of the conditions of how evolution has worked in the past given the appropriate choice of conditions. This gets particularly interesting insofar as the selective condition enables the experiment to recapitulate a major evolutionary transition, e.g., the evolution of multicellularity in a test tube (Ratcliff et al. 2012).

At this point, you have may have the impression that the space of genotypes is replete with beneficial adaptations despite the fact that there are many ways for mutations to have a deleterious effect on fitness. How easy it is to adapt lies in the eyes of the beholder. But, the visualiation of increases in fitness over 10,000 generations belies something else: punctuated evolution. That is, when viewed at higher resolution, it is evident that there are *discrete* jumps in fitness related directly to cell size. The bacteria cells evolved larger cell sizes concurrent with increases in fitness. It could be that cell size was incidental to fitness. Yet, for various reasons, Santiago Elena, Rich Lenski and colleagues favored the hypothesis that cell size was of direct fitness benefit (probably related to increased nutrient uptake) or, at least, pleitropically related to those traits that conferred a fitness benefit (Elena et al. 1996). In this interpretation, individual bacteria would, on occasion, mutate their genotype, modify their cell size, thereby increasing their growth rate (see Figure 4.2). This is, in essence, evolution via natural selection – a concept that applies to Darwin's finches and their evolution of beak size in relationship to environmental disturbances and changing food availability (Grant and Grant 2014) just as well as to bacteria growing in shaken flasks given a stable food source. But, is that all there is to the process – stasis interrupted by occasional selective sweeps?

This chapter will introduce the key features of models that bridge the gap between mechanisms of evolutionary change tne the resulting changes in populations, particularly with a focus on asexual organisms. In doing so, the text is motivated by a number of questions. How fast should such a mutation "fix", i.e., go from a single individual to (nearly) the entire population? How does the rate of adaptation depend on mutation rate and other features, e.g., background levels of mutants? And, are populations "beneficial mutation limited", i..e, only occasionally experiencing beneficial mutations? Or, instead, are such beneficial

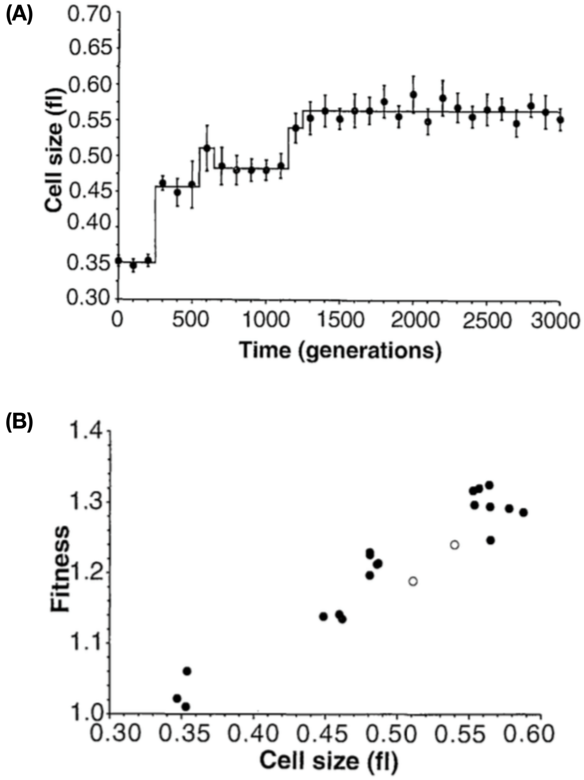


Figure 4.2: The punctuated growth of cell size in the LTEE experiment over 3000 generations. (A) Evidence of punctuated changes in cell size over 3000 generations, where 1 fl = 10^{-15} L. (B) Evidence that increases in cell size are related to increases in fitness. Reproduced from (Elena et al. 1996)

mutations in amply supply such that genotypes with different mutations compete with each other, perhaps even with other mutations that have no benefit one way or another but nonetheless hitchike along for the ride.

This last question also raises a puzzle framed by the second panel in Figure 4.1. This panel shows changes in the frequency of mutations measured in evolving populations of yeast, albeit those that have been sequenced with sufficient time resolution to be able to witness the spread of mutations in a population. Notably, the timing of these increases suggests a problem: is it really possible that multiple chance mutations swept through the population at nearly precisely the same time (and speed) – see the darker lines representing mutations that reach nearly 100% frequency in Figure 4.1. Or, instead, is it possible that the sweep of multiple concurrent mutations represents a balance between selection and chance? The mechanistic basis for evolution via natural selection suggests that these reproducibility and chance both play a role in evolution. But, how big a role do these two factors play and how would we recognize the balance from experiments? In other words, if you were to measure the changes in frequencies in Figure 4.1 in the case of evolving yeast, how would you interpret it? Does the rise of many mutations at or near the same moment in time reflect a hallmark of reproducibile evolution or, instead, a hybrid, in which certain reproducible features are mixed with chance? This chapter introduces models and methods to approach and perhaps even answer this question.

4.2 SELECTION AND THE DISAPPEARANCE OF DIVERSITY

The term "survival of the fittest" has long been invoked to describe dynamics in which a subset of individuals outcompete others. The 'more fit' subset are those with traits that increase their competitiveness, fecundity and/or ability to survive. These traits depend strongly on the organism of interest. In the classic example of Darwin's finches, the traits are linked to body and beak size, given that variation in the size of available seed resources imposes strong selective pressures on the size and shape of finch beaks (Grant and Grant 2014). For bacteria, the traits may include changes to surface moeities that confer resistance to viral infection or to the effectiveness of efflux pumps that pump out otherwise lethal antimicrobials out of the cell and back into the environment (Madigan et al. 2009). For fish like the ancient Tiktaalik, it may be the emergence of proto-appendages that enable them to walk on land; features that could, over time, become the genesis of an entirely new form of life (Shubin 2008) These examples are so disparate that it would seem hard to invoke a single or even limited set of modeling frameworks that could accomodate each. This diversity often poses what seems like an innsurmountable obstacle to biologists when faced with the challenge to build a model. It seems like every detail must be included. Indeed, these examples include highly disparate phenotypes and associated genotypes. Yet, what is common is that irrespective of size, scale, or evolutionary epoch, all of these organisms must reproduce. If abstracted in a certain way, the change in the number of individuals over time may be predictable, and comparable, across systems, even if highly detailed models may be necessary to explain both the changes in phenotypes and of growth rates. In that sense, the path of model building of evolutionary dynamics resembles an argument made by Oliver Wendell Holmes Jr. on the (ir)relevance of certain details when trying to move from legal cases to a 'general form':

> The reason why a lawyer does not mention that his client wore a white hat when he made a contract ... is that he foresees that the public force will act in the same way whatever his client had upon his head. It is to make the prophecies easier to be remembered and to be understood that the teachings of the decisions of the past are put into general propositions and gathered into textbooks, or that

> statutes are passed in a general form.
> -Oliver Wendell Holmes, The Path of the Law, 1897

Let us then move to these general propositions and their prophecies, keeping in mind that they make certain assumptions, often strong assumptions, about the drivers of evolution. And we will soon explore the many ways that conditions may well violate these assumptions leading to altogether new conclusions. In doing so, let's consider perhaps the simplest model of evolution, in which the size of the population remains fixed.

4.2.1 Replicator Dynamics

Consider two populations, A and B, growing at rates r_A and r_B. For a continous process, one would expect that the population of A would grow like $n_A(0)e^{r_At}$ and that of B would grow like $n_B(0)e^{r_Bt}$. Although both are seemingly better off from an abundance standpoint, the relative amount of an exponential population may nonetheless decline exponentially fast. Note that this paradoxical idea is of practical relevance, it explains why variants of SARS-CoV-2 can take-over a population, in a logistic sense, even as the variant and wild-type strains increase in abundance. If we term x_A and x_B as the relative proportion of the two types then after a long period of time one expects $x_B \sim e^{(r_B-r_A)t}$, which means that the relative population of B would decreases exponentially whenever $r_B < r_A$. Moreover, such exponential growth cannot continue unabated. Hence, it is worthwhile to develop a simple model of replicators that continuously reproduce in a system in which the total population remains constant. Practically speaking this means that whatever exponential growth of individual populations must be compensated by a time-dependent loss rate $\omega(t)$, i.e.,:

$$\frac{\mathrm{d}n_A}{\mathrm{d}t} = r_An_A - \omega(t)n_A, \tag{4.1}$$

$$\frac{\mathrm{d}n_B}{\mathrm{d}t} = r_An_B - \omega(t)n_B. \tag{4.2}$$

Figure 4.3 illustrates this concept via the concept of a turbidostat. A turbidostat is a device used to measure the turbidity of a solution. If bacteria are grown in a flask, then the turbidity is a proxy for density, i.e., more bacteria correspond to a more turbid solution and less bacteria correspond to a less turbid (more transparent) solution. Although such feedback-control may yield a nearly constant population size, it is worthwhile to impose stricter controls for the purposes of theory. Hence, we need to identify which choice of $\omega(t)$ would guarantee that $n_A(t) + n_B(t) = N$, a fixed target population size.

If the total population is fixed, then we should expect that $\frac{\mathrm{d}N}{\mathrm{d}t} = 0$, or by adding together the changes $\mathrm{d}n_A/\mathrm{d}t$ and $\mathrm{d}n_B/\mathrm{d}t$ leads to

$$r_An_A + r_Bn_B - \omega(t)\left(n_A + n_B(t)\right) = 0. \tag{4.3}$$

Identifying $N(t) = n_A + n_B$ this equation becomes

$$\omega(t)N = r_An_A + r_Bn_B \tag{4.4}$$

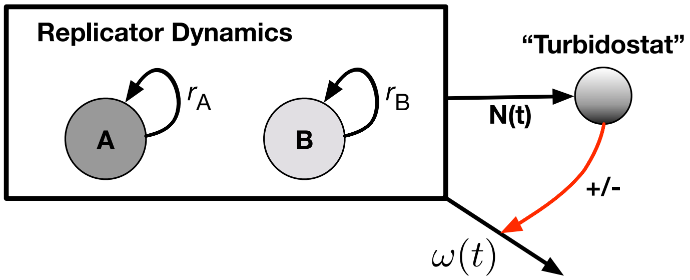


Figure 4.3: Schematic of a turbidostat in which local changes in total population size modulate the rate of population death so as to maintain a fixed population size. In this schematic, the negative or positive regulation acts to keep the population set to a constanat. Although conceptual in design, measurements of optical density could be used to provide feedback control on the rate of turnover. Moreover, this notion also illustrates the conceptual basis of replicator equations in which the total population is assumed to be strictly fixed.

or

$$\omega(t) = r_A x_A + r_B x_B \tag{4.5}$$

where $x_A = n_A/N$ and $x_B = n/B/N$ represent the fraction of types A and B in the population. This idealized turbidostat has a loss rate equal to the average growth rate, i.e., what we term the average fitness, i.e., $\langle r(t)\rangle = \omega(t)$. Altogether this yields a model that is widely known as replicator dynamics:

$$\frac{\mathrm{d}x_A}{\mathrm{d}t} = r_A x_A - \langle r\rangle x_A, \tag{4.6}$$

$$\frac{\mathrm{d}x_B}{\mathrm{d}t} = r_A x_B - \langle r\rangle x_B, \tag{4.7}$$

and as should be evident, $\dot{x}_A + \dot{x}_B = 0$. These equations imply that populations will increase in proportion insofar as their growth rates at a given moment in time exceed that of the average fitness. Such ideas extend naturally to 3, 4 or arbitrary numnber of populations. However, for two populations it is possible to describe the dynamics in terms of $x \equiv x_A$, from which x_B can be deduced, i.e., $x_B(t) = 1 - x_A(t)$. In that case, we now have

$$\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}t} &= r_A x - (r_A x + r_B(1-x))\,x && (4.8)\\
&= x\,(r_A - r_A x - r_B(1-x)) && (4.9)\\
&= x\,(r_A(1-x) - r_B(1-x)) && (4.10)\\
&= x(1-x)\quad \overbrace{[r_A - r_B]}^{\text{selection coefficient}} && (4.11)
\end{aligned}$$

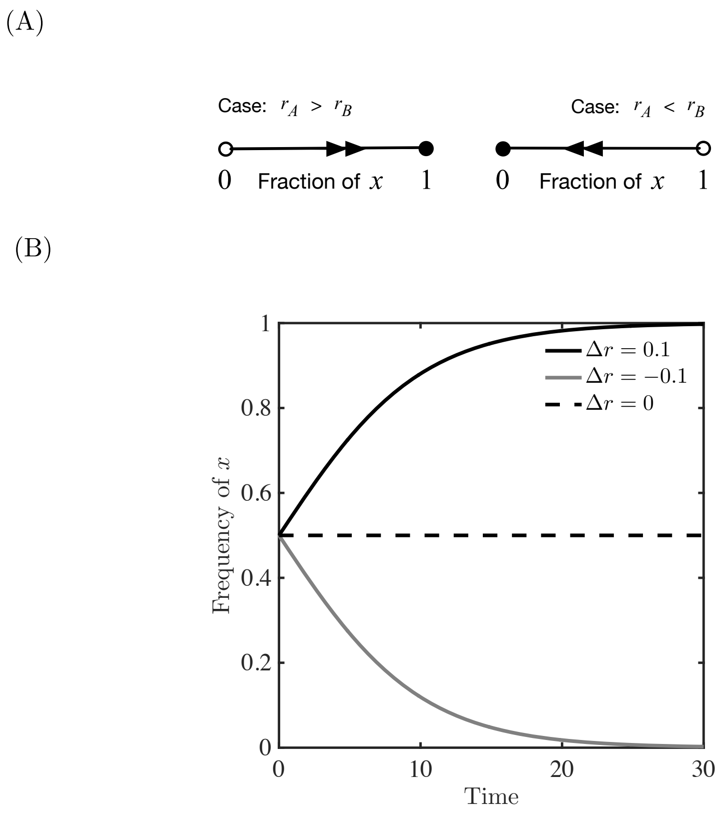


Figure 4.4: Long-term dynamics in replicator dynamics. (A) One-dimensional state space representation with two cases. When $r_a > r_b$ then the system converges to $x = 1$ (associated with the type-A replicator) and the converse holds when $r_a < r_b$. (B) Simulation of replicator dynamics with $\Delta r = 0.1, -0.1$ and 0 using the standard replicator dynamics and initializing the system with $x_0 = 0.5$ (an equal balance of type-A and type-B replicators).

This last equation is the logistic growth equation where the selection coefficient, i.e., $\Delta r = r_A - r_B$, determines which of the two types grow to dominate (see Figure 4.4).

The solution to the logistic growth equation is

$$x(t) = x_0 \left[\frac{e^{\Delta r t}}{1 + x_0 \left(e^{\Delta r t} - 1 \right)} \right] \tag{4.12}$$

which generalizes to the interactions of S different interacting types, $x_1, x_2, \ldots, x_S$. Of note, the Technical Appendides describe how to use changes in frequency amongst competing strains to estimate the selection coefficient from experiments (a process used in (Lang et al. 2013)). Given dynamics that are strictly determined by Malthusian growh rates, then it seems the fittest should indeed survive, i.e., $x_i \to 1$ if r_i is the maximum of all the growth rates $\{r_1, r_2, \ldots, r_S\}$. As a consequence, the other populations should go to zero given pure Malthusian growth competition amongst diverse replicates. In essence, selection eliminates diversity. The model also suggests that if $\Delta r = 0$, then the proportions will remain constant. Although this may be true in a theoretical sense for an infinitely large population, later in these notes I will show how the process of genetic drift leads to changes in the frequency of types even when there are no intrinsic fitness differences. Yet, for cases where fitness differences are present, the finding that a subset will dominate raises the question: are there any general conclusions we can make about what drives the rate at which selection can, in theory, eliminate diversity? This question is at the heart of the premise underlying Fisher's fundamental theorem of natural section.

4.2.2 Fisher's Fundamental Theorem

There are not many theorems in the life sciences – at least not many considered relevant enough to be discussed by practicing life scientists. It may be worthwhile to ask why. First, theorems require strict assumptions. These 'given-s' may be mathematically reasonable, but new discoveries often give way to exceptions. Second, to the extent that one can find theorems in mathematical biology papers, these theorems rarely percolate up to directly influence the practice of life scientists. Some theorems may prove things that are already intuitive, and although some may be satisfied with such additional rigour, it is critical to ask: what is the added value gained? Fisher's Fundamental Theorem is the exception that proves the rule.

Ronald Fisher, a statistician and evolutionary biologist, proposed the following theorem: the average rate of increase in fitness is equal to the variance of fitness in a population. In essence, more variation can somehow drive faster rates of change, presumably accelerating the rise of the most fit. One way to demonstrate this theorem is to consider the case of S replicators each with a population fraction x_i whose dynamics can be described as

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = r_i x_i - \langle r \rangle x_i \tag{4.13}$$

where $\langle r \rangle = \sum_{i=1}^{S} r_i x_i$ is the average fitness. The change in the average fitness is then

$$\begin{aligned}
\frac{\mathrm{d}\langle r\rangle}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_i^S r_i x_i\right) && (4.14)\\
&= \sum_i^S r_i \frac{\mathrm{d}x_i}{\mathrm{d}t} && (4.15)\\
&= \sum_i^S r_i \left(r_i - \langle r\rangle\right) x_i && (4.16)\\
&= \left(\sum_i^S r_i^2\right) - \left(\langle r\rangle \sum_i^s r_i x_i\right) && (4.17)\\
&= \langle r^2\rangle - \langle r\rangle^2 && (4.18)
\end{aligned}$$

where the last line shows that the change in mean fitness is equal to the difference between the average squared fitness and the average fitness squared. This is precisely the definition of the variance, i.e.,

$$\frac{\mathrm{d}\langle r\rangle}{\mathrm{d}t} = \mathrm{Var}(r). \qquad (4.19)$$

The variance is positive, by definition so that the mean fitness seems to always increase. But this concept seems incompatible with our earlier finding that selection purges diversity. The resolution is found by realizing that the very process of selection operating to increase mean fitness also purges diversity, thereby slowing down the increase of mean fitness. Asymptotically, one expects that $\mathrm{Var}(r) \rightarrow 0$, i.e., the system increases in fitness, while purging itself of variation – the very grist for the acceleration in fitness. Notably, the collapse of diversity happens at the fastest rate when the population is, itself, most diverse.

These results can be seen in a model of 10 competing populations, shown in Figure 4.5. As is apparent, the replicator with the highest growth rate eventually dominates the system and the replicator with the lower growth rate (despite having the highest initial frequncy) is rapidly purged from the system. This is the essence of frequency independent selection, insofar as a faster replicator always outgrows the competitors (in relative terms). In doing so, the speed of adaptation increases, at least at the beginning, until it begins to approach 0 (Figure 4.5-right. The trajectory of adaptation precisely matches that of the measured variance during the simulation. Hence, Fisher's Theorem is true, in a mathematical sense, but is also of limited value, in a biological sense, as it has many requirements that continue to limit its utility in understanding all but the most constrained of evolutionary dynamics. The key insight is that variation is the grist upon which the evolutionary wheel operates. The other insight is that selection is relative; and that fitness need not always go up in an absolute sense.

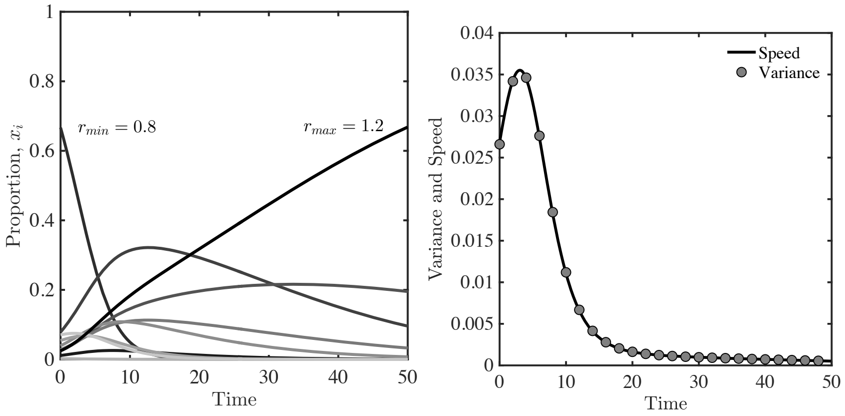


Figure 4.5: Dynamics of $S = 10$ populations competing via replicator dynamics. (A) The faster replicator dominates the system, all other populations will eventually go to zero. (B) Variance-speed relationship with time – the values coincide at all times. Note that the system increases in mean fitness at its maximal value precisely when there is the greatest variance in fitness, and that both decay to zero as the system is dominated by the fastest replicator.

4.3 MECHANISMS THAT RESTORE DIVERSITY

There are many mechanisms to restore diversity. But, rather than list/examine each, this chapter will examine two: mutation-selection balance and negative frequency dependent selection. This first example is important because it shows the dueling tension between selection that can act to purge diversity and the process of mutation that increases dviersity. The second example is often invoked when context matters, i.e., when ecology or feedback between organisms (or the environment) means that there is not a unique, global optimal. Instead, the very same reasons that might explain the rise of a type from rare may also explain its failure to eliminate competitors when abundant.

4.3.1 Mutation-selection balance

Consider two variants, an ancestral population which we term the wild-type and a mutant population that has a deleterious mutation. The wild-type replicates at a rate r and the mutant replicates at a rate $r(1-c)$ where c is the cost the mutation. This model of wild-types and mutants is inspired by the dynamics of RNA viruses for which mutations occur frequently. Some of these mutations may be lethal such that the RNA virus is not able to infect or replicate a target host cell. Such offspring cannot survive. Yet what about those daughter viruses that bear a partial cost, how can they continue to persist? To answer this question, consider the following replicator dynamics describing changes in the

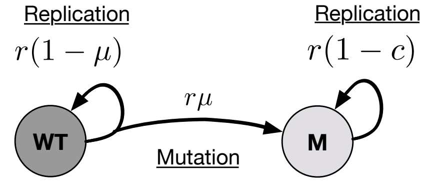


Figure 4.6: Mutation-selection balance in which a WT replicates at a rate r, with mutation rate μ, and the mutant replicates at a rate $r(1-c)$ where c denotes the relative cost of mutations.

frequences of wild-type individuals, x, and mutants, y:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = r(1-\mu)x - \langle r \rangle x \tag{4.20}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = r\mu x + r(1-c)y - \langle r \rangle y \tag{4.21}$$

where μ is the mutation probability per replication and $\langle r \rangle = rx + r(1-c)y$ is the average growth rate (see Figure 4.6). Because replicator dynamics describe changes in the relative growth rate, we can rescale these equations in terms of a dimensionless time, $\tau \equiv rt$, where a change of 1 in τ represents a typical generation time. The re-scaled dynamics become:

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = (1-\mu)x - \tilde{r}x \tag{4.22}$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = \mu x + (1-c)y - \tilde{r}y \tag{4.23}$$

where $\tilde{r} = x + (1-c)(1-x)$ given that frequencies must add to 1, i.e., $y = 1-x$. Given that the dynamics of the wild-type completely specify the mutant fraction, this system of equations has an equilibrium when $\frac{\mathrm{d}x}{\mathrm{d}t} = 0$ or when $1-\mu = x + (1-c)(1-x)$. At equilibrium, this means that

$$x^* = 1 - \frac{\mu}{c}. \tag{4.24}$$

Hence, the wild-type should persist insofar as the relative costs of mutations exceed the mutation probability. When this occurs, then $y^* = \frac{\mu}{c}$, and the system coexists due to mutation-selection balance (see Figure 4.7-left). In a mutation-selection balance scenario, the more fit wild-type generates a constant flux of new mutants, which themselves replicate, albeit at a lower rate. For RNA viruses, mutation probability per replication can approx 1/1000 (see Figure 4.7-right). Hence mutants will not be able to outcompete wild-type individuals so long as costs exceed more than a 0.1% reduction in fitness. Yet, the wild-type individuals will not be able to completely displace the mutants, given that their

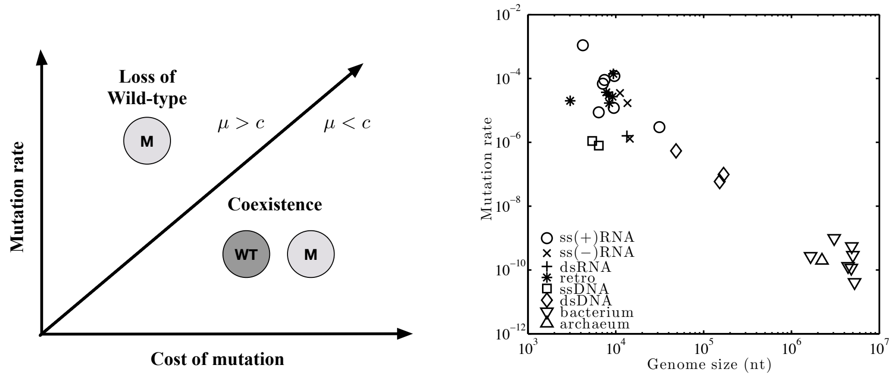


Figure 4.7: Mutation-selection balance. (Left) Expected dynamics in the $c - \mu$ phase plane. Wild type and mutants coexist when $\mu < c$, i.e., the levels of mutation are lower than the relative fitness costs of the mutant compared to the wild-type. However when $\mu > c$ then the mutant populations grows and eventually eliminates the wild-type. (Right) Relationship between mutation rate and genome length in different viral and microbial organisms. Adapted from Sanjuan (2004).

success continues to represent the source of new variants. Note that in effect, here the system is driven to a lower-fitness in terms of replicator speed; albeit to a higher fitness when accounting for the mutational costs on fitness.

4.3.2 Frequency-dependent selection

Mutation selection balance can give rise to coexistence of types. However, such coexistence seems fragile; as the continued persistence of the mutant type requries the continual renewal of its population via new mutations. Instead, consider two replicators each of whose frequency changes according to the following:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = r_i x_i - \langle r \rangle x_i \tag{4.25}$$

where $\langle r \rangle = \sum_{i=1}^{S} r_i x_i$ is the average fitness. When r_i is fixed, then this leads to the elimination of diversity, in which evolution rapidly purges diversity (indeed, at a speed proportional to the variance). However, what happens if the fitness of each strain is context dependent. A strain may produce some form of public good, like the bacteria *Pseudomonas aeruginosa* which secretes an extracellular enyzmes called a siderophore that binds to non-soluble iron, thereby enabling it to be taken up – at least potentially – by cells. Yet, when abundant a producer strain might be less fit than a cheater strain that can take advantage of the iron without incurring the costs of producing and secreting costly siderophores. In this case then, the frequency of the producer strain $r_1(x_1)$ might decrease with

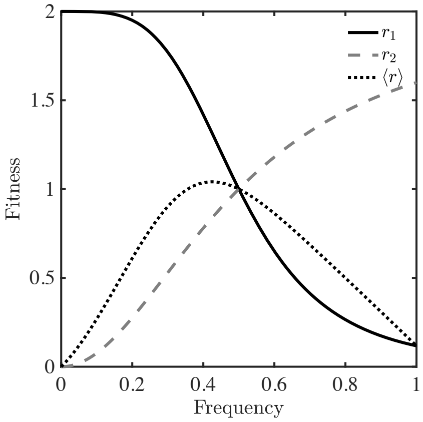


Figure 4.8: Illustration of negative frequency dependence for two frequency-dependent fitness functions, here $r_1(x)$ decreases with x and likewise $r_2(y)$ decreases with $y \equiv 1-x$. The average fitness is equal to $\langle r \rangle = r_1 x + r_2(1-x)$. The fixed point of this system occurs when $r_1 = r_2$ (denoted by the intersection of curves). However, note that this equilibrium does not correspond to the largest average fitness which is at a smaller value of x. In this example, $r_1(x) = \frac{2}{1+x^4/0.5^4}$ and $r_2(x) = \frac{2x^2}{x^2+0.5^2}$.

its own frequency x_1. If that is the case then the strain with the larger relative fitness will shift with frequency. As one conceptual example, the curves in Figure 4.8 depict a case where type 1 has a higher fitness than does type 2 when rare but has a lower fitness when abundant. Recall that for two populations where $x \equiv x_1$, $y \equiv x_2$ and $x + y = 1$, the replicator dynamics reduces to

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(1-x)\left(r_1(x) - r_2(x)\right). \tag{4.26}$$

Hence, it is apparent that the critical value x_c corresponding to equivalent fitness values are equivalent $(r_1(x_c) = r_2(x_c))$ will represent a stable equilibrium irrespective of the particular shape of r_1, insofar as $\partial r_1(x)/\partial x < 0$ and $\partial r_2(x)/\partial(x) > 0$. This implies that frequency should increase for $x < x_c$ and decrease for $x > x_c$, leading to a stable fixed point of the system at $x^* = x_c$, corresponding to a mixed system with two strains. Yet, notably, this simple example also illustrates that this fixed point $x^* = 0.5$ need not be the maximum average fitness, which occurs for a value $x < x^*$. This simple example reinforces the point that the selection process need not act to maximize the fitness for the entire population.

4.4 STOCHASTICITY IN THE EVOLUTION OF POPULATIONS – BASELINE EXPECTATIONS

4.4.1 Reproduction and survival - a recipe for evolution without natural selection

Evolution denotes the heritable change in the frequency of genotypes from one generation to the text. Hence, any description of evolution must include at least two genotypes and some way of transmitting, with fidelity, the genotype from parent to offspring (i.e., from one generation to the next). Put concretely, consider a population with N individuals in which there can be one or more types, what we refer to as genotypes. For asexual organisms, we can envision a scenario in which they produce n_i offspring each, such that there is a pool of new individuals of size $N_{off} = \sum_{i=1}^{S} n_i$ where S denotes the number of types in a population. Further, there need not be any difference in the expected number of offspring between types, such that the average number of offspring $b \equiv \langle n_i \rangle$ is the same for all genotypes. A key approximation underlying a large class of evolutionary dynamics models is that the nuber of individuals remains constant over time, i.e., from one generation to the next. Hence, in this model, all N of the "mothers" die while N of the N_{off} offspring survive to become the next generation of mothers. In such a case, differences in genotype frequencies will emerge even if there is no inherent differences in the expected number of offspring or the survivorship of offspring, The notion that the stoachsticty of reproduction and survival can lead to 'neutral' evolutionary dynamics is, in essence, the central assumption of the Wright-Fisher model of population genetics (see Figure 4.9).

There are different ways to think about these dyanmics. For example, one could envision the state of the system as being a set of N id-s, or unique numbers, each referring to the genotype identity of the individuals:

$$t : \overbrace{1\ 10\ 2\ 3\ 1\ 1\ 1\ 2\ 2\ 4\ ....\ 8\ 10\ 1}^{N \text{ in total}} \tag{4.27}$$

However, if there's no difference between individuals of the same genotype, then we could also think of the composition of a population in terms of a set of pairs (id, number). For example, with three types in a population of 100, this might be:

id	1	2	3
number	60	10	30

Hence, the "state space" is a set of S numbers, referring to the S genotypes, such that $n_i = N$. In the case that there are only two types, then we can refer to the types as A and B, such that $n_A + n_B = N$. Because the sum is fixed, then we only need one number, $n \equiv n_A$. This implies that the Wright-Fisher process has the effect of changing the number n from one generation to the next. This also means that the Wright-Fisher models is a form of Markov Chain. The word

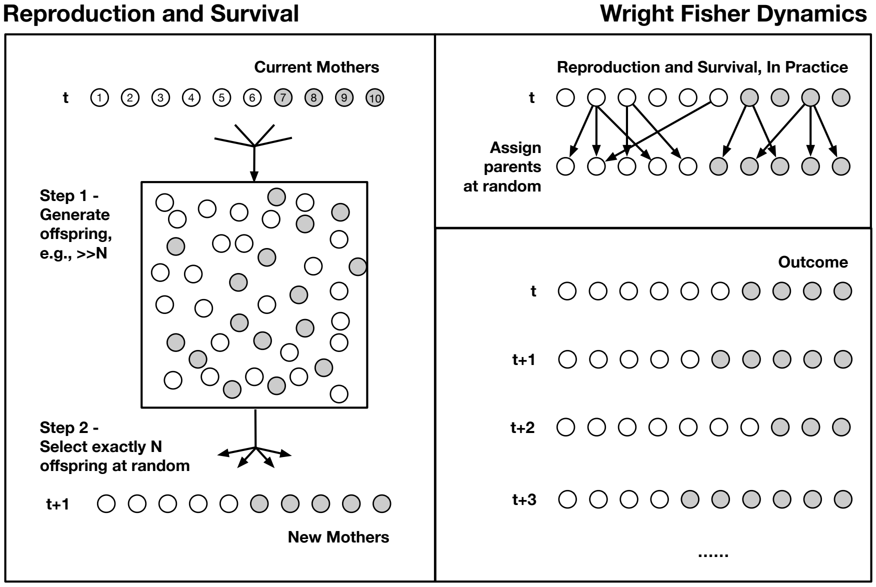
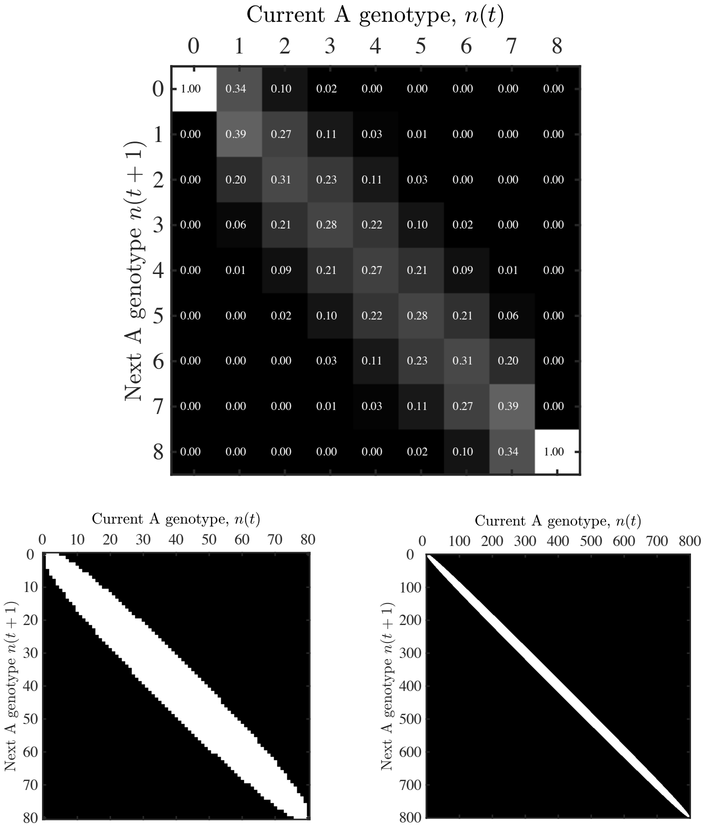

Reproduction and Survival
Wright Fisher Dynamics
Current Mothers
t
1
2
3
4
5
6
7
8
9
10
Step 1 - Generate offspring, e.g., >>N
Step 2 - Select exactly N offspring at random
t+1
New Mothers
Reproduction and Survival, In Practice
t
Assign parents at random
Outcome
t
t+1
t+2
t+3
......

Figure 4.9: Neutral population dynamics via Wright Fisheer (Left) Schematic of the two steps of the WF model given non-overlapping generations. Step 1: Random production of offspring. Step 2: selection of N offspring for the next generation. (Right) Expected outcome – parent-offspring relationships define differences in genotype frequency that change stochastically over time, even if the average frequency is expected to remain unchanged.

'Markov' denoes the memoryless nature of the stochastic process (depending only on the current state) and the word 'chain' denotes the fact that there are a proscribed sequences of events that are possible. What then are the expected dynamics of n_t and how much difference do we expect between trajectories, i.e., what is the expected variance?

4.4.2 Population genetic models of non-overlapping generations

A system in state n at generation t can become a system in state n' in the next generation $t+1$. For example, if $n = 50$ now, then it is possible that there may be 50, or perhaps more or less, individuals of type A at the next generation. Formalizing this requires considering the WF process in terms of *conditional probabilitites*. That is, conditional upon the fact that the system has precisely n individuals of type A then $T(n'|n)$ denotes the probability that the system will have precisely n' individuals of type A at the next generation. One restriction



	0	1	2	3	4	5	6	7	8
0	1.00	0.34	0.10	0.02	0.00	0.00	0.00	0.00	0.00
1	0.00	0.39	0.27	0.11	0.03	0.01	0.00	0.00	0.00
2	0.00	0.20	0.31	0.23	0.11	0.03	0.00	0.00	0.00
3	0.00	0.06	0.21	0.28	0.22	0.10	0.02	0.00	0.00
4	0.00	0.01	0.09	0.21	0.27	0.21	0.09	0.01	0.00
5	0.00	0.00	0.02	0.10	0.22	0.28	0.21	0.06	0.00
6	0.00	0.00	0.00	0.03	0.11	0.23	0.31	0.20	0.00
7	0.00	0.00	0.00	0.01	0.03	0.11	0.27	0.39	0.00
8	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.34	1.00

Figure 4.10: Transmission matrix in WF given the current state n in columns and the next state $n(t+1)$ in the rows. (Top) Transition values given a population of size $N = 8$; (Left-bottom) Transition values $T > 0.01$ for $N = 80$; (Right-Bottom) Transition values $T > 0.01$ for $N = 800$;

is that the system must go somewhere, i.e., $n' \in 0, 1, 2, \ldots, N$. In other words,

$$\sum_{n'=0}^{n} T(n'|n) = 1. \tag{4.28}$$

The structure of the transition matrix can be seen in the following transmition matrix for $N = 8$ (Figure 4.10) and then for $N = 80$ and then for $N = 800$. As is apparent, as N increases, the probability that the next state will be close to the current state becomes more tightly centered (in relative terms) on the current value – a point that is more readily seen when considering those states that might occur next with a probability exceeeding 0.01. In essence, $T(n'|n)$ is the probability that a type A individual is selected n' out of the N trials, one for each individual in the next generation. This should remind you of something:

flipping coins, albeit biased coins. The bias comes in because the probability of choosing a type A individual is n/N and not necessarily 1/2. Therefore given the probability of "success" of n/N (type A) and the probability of "failure" of $1 - n/N$ (type B), this probability is equivalent to a binomial distribution, or:

$$T(n'|n) = \binom{N}{n'} p_A^{n'} p_B^{N-n'} \tag{4.29}$$

or

$$T(n'|n) = \binom{N}{n'} \left(\frac{n}{N}\right)^{n'} \left(1 - \frac{n}{N}\right)^{N-n'}. \tag{4.30}$$

The expected mean of a binomial process is equal to the number of trials, N, multiplied by the probabilty of 'success', i.e., n/N - the fraction of type A individuals, such that the expected number of type A individuals remains the same from one generation to the next, i.e., $\langle n \rangle = n_0$. Formally speaking, this means that a WF process is a martingale, such that expected value of the Markov chain does not change with time. This martingale property becomes essential to pose a cruciall question concerning WF dynamics. Given that the $n(t = 0) = n_0$, what is the probability of fixation, i.e., that the system will converge to entirely dominated by type A individuals, $n \to N$? The potential paradox is that the expecte number of type A individuals should not change, even though it seems that variation may drive the system to one of two extremal outcomes.

4.4.3 Variation and fixation

One way to think about fixation is to note that the WF model represents a Markov Chain with two 'absorbing states'. An absorbing state denotes one in which $T(n'|n_a) = \delta_{n',n_a}$ such that the only transition is to remain in the same place. For an evolutionary dynamic model without mutation or immigration, there are two absorbing states, $n = 0$ and $n = N$, corresponding to the elimination and fixation of type A respectivelly. Although the mean value of n is expeted to remain constant, the variance in the next generation is non-zero, i.e., $\mathrm{Var}(n'|n) = n(1 - n/N)$ for a binomial process. This means that a typical realization is often different than the prior state, and maximally different at the point $n = N/2$, where the types are in equal proportion. How much variation should be expected over time across evolutionary trajectories given that the initial conditions were shared? In other words: how repeatable is evolution given drift?

One answer can be approached by calculating the probability that two randomly chosen individuals in the population are of different genotypes. Formally, this is termed 'heterozygosity' or:

$$H = 2\frac{n}{N}\left(1 - \frac{n}{N}\right) \tag{4.31}$$

where n is the number of A genotypes, $N - n$ is the number of B genotypes, and therefore n/N is the fraction of A genotypes and $1 - n/N$ is the fraction of B genotypes. The expected dynamics of heterozygosity, $H(t)$ can be computed by calculating the value one generation ahead, where $E(H)' \equiv \langle H(t+1) \rangle$:

$$E(H') = E\left(2\frac{n'}{N}\left(1 - \frac{n'}{N}\right)\right). \tag{4.32}$$

By expanding out these expectations it is possible to show (see Technical Appendices) that for large populations, $N \gg 1$,

$$H(t) \approx H_0 e^{-t/N}. \tag{4.33}$$

This equation implies that the WF model of neutral drift will eventually lead to a state in which there is no variation between individuals, i.e., to one of the two absorbing states, either local extinction or fixation. The question now is: which one and with which probability?

To answer this question, we must turn back to the martingale property of the Markov chain. At equilibrium, we expect there to be a steady state probability distribution

$$P^*(n) = \pi\delta_{n,N} + (1-\pi)\delta_{n,0} \tag{4.34}$$

where π is the fixation probability (and not the number 3.14159...) and the δ denotes the Kronecker delta equal to 1 when the two arguments in its subscripts are equal and zero otherwise. At equilibrium, it should be possible to calculate the expected value of n, or

$$E^*(n) = \Sigma_n P^*(n) = \pi N + 0 \times (1-\pi) = \pi N. \tag{4.35}$$

However, by the martingale property, this value πN must equal the initial value n_0, such that $\pi = n_0/N$. In other words, the probability of fixation is equal to the initial fraction of type A genotypes in the population. Likewise, the probability of extinction is equal to the initial fraction of type B genotypes in the population. Another way to think about this is that eventually one of the current individuals will be the ancestor of all extant individuals. And, since each individual has an equal chance of being that progenitor, then the probability that the progenitor is type A is n_0/N or the probability of randomly choosing a type A individual. Hence, in the absence of selection, even the case of replication with equivalent fitness values will eventually lead to the extinction of diversity, albeit over time scales set by the size of the population.

4.5 EVOLUTIONARY DYNAMICS WITH STOCHASTICITY AND SELECTION

This next section combines principles develope thus far that suggest that drift can lead to the local fixation or extinction of a population and that selection is a powerful force to drive one population to fixation at the expense of others. Both drift and selection seem likely to make it harder for coexistence to emerge,

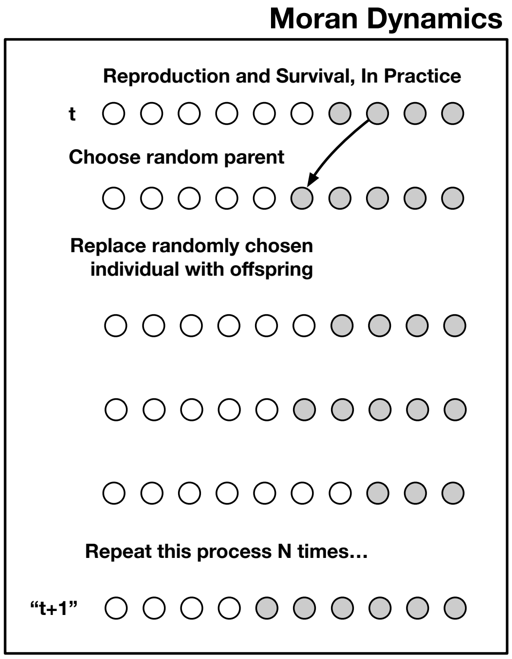

Moran Dynamics
Reproduction and Survival, In Practice
t
Choose random parent
Replace randomly chosen individual with offspring
Repeat this process N times...
"t+1"

Figure 4.11: Schematic of the Moran model, in which a randomly selected parent reproduces and replace any one of the N current individuals with one of their offspring. A total of ~ N such events are equivalent to a single generation of the WF model.

yet unlike drift, the process of selection provides the basis for understanding the extent to which the outcomes of evolution are repeatable and predictable. Building simple evolutionary models that incorporate both selection and drift is a first step to parsing patterns emerging from experimental evolutionary data whether in bacteria and yeast.

4.5.1 The Moran model of population genetics

Here, we consider a variant of evolutionary dynamics in which there are exactly N individuals, but the dynamical steps involve single replacement events only. The Moran model includes two steps. In the first step, a random individual is chosen to give birth, i.e., to replicate. In the second step, a randomly chosen indiviual is selected to die, and is replaced by the offspring of the first individual, as is show in Figure 4.11. In this case, the transition matrix is different, in that given a system in state n at time t, there can be either $n-1$, n or $n+1$ individuals of type A at time $t+1$. Here, it appears evolution moves more slowly than it does in the WF case. Another way to think about this is that it takes more steps of the Moran model to be equivalent to a single generation in the Wright Fisher model – formally speaking this is on the order of N steps. In the event that all individuals have the same fitness then,

$$T(n'|n) = \begin{cases} \frac{n}{N}\left(1-\frac{n}{N}\right) & n' = n+1 \\ \frac{n}{N}\left(1-\frac{n}{N}\right) & n' = n-1 \\ 1 - T(n-1|n) - T(n+1|n) & n' = n \\ 0 & otherwise \end{cases} \tag{4.36}$$

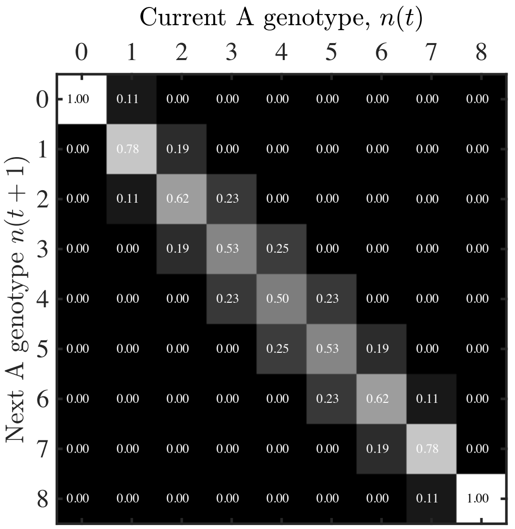
The transition matrix for the Moran model is shown in Figure 4.12 noting that the only non-zeros values lie on or adjacent to the diagonal.

The properties of the Moran model are quite similar to that of Wright-Fisher, namely that the expected number remains constant (i.e., the process is a Martingale) and that the heterozygosity declines exponentially (see Technical Appendices). The major difference is that the timescales are distinct and it takes on the order of N Moran steps (again, see Technical appendices) to be equivalent to a single generation step in the Wright Fisher model. The other key point to recall is that the transition matrix for the Moran model will have positive values either on or just adjacent to the diagonal. This is unlike the Wright Fisher model whose probabilities will be centered on the diagonal, but have nonzero weight (nearly) everywhere. The Moran model is also useful as a means to consider the combination of selection and drift – and how both converge to yield realized, evolutionary dynamics.

4.5.2 Selection in light of stochasticity

Finally, selection. There are multiple mechanisms by which selection can change the state of populations. Advantageous changes in genotype can increase fitness relative to the background “wild-type”. In a deterministic model, the more fit the organism is, the faster it takes over the population. But, that is merely a quantitative issue. The deeper implication of deterministic models is that a better fit mutant *always* outperforms the wild-type. That is not the case. Consider one better-fit individual that has a 10% growth rate advantage compared to all other individuals in a population. This mutant may, by chance, die before reproducing. Hence, it is obvious that selective advantages, in and of themselves, do not guarantee success. Moreover, it is also possible for the converse to occur. That is, a mutant with a fitness disadvantage may, given small populations, grow in number and reach fixation. A model will help us conceptualize the relative importance of selective advantages, initial sizes of mutant populations, and the total size of the population, while estimating the chance of fixation as well as the time-scale over which fixation occurs.

To explore the impact of selection requires modifying the Moran model. Instead of assuming strict neutrality, we will assume that type A individuals have a slightly higher chance of being chosen for replication, i.e., by a factor of $(1+s)$ relative to that of type B individuals. The value s denotes the fitness



	0	1	2	3	4	5	6	7	8
0	1.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.78	0.19	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.11	0.62	0.23	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.19	0.53	0.25	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.23	0.50	0.23	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.25	0.53	0.19	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.23	0.62	0.11	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.78	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	1.00

Figure 4.12: Transmission matrix in the Moran model given the current state n in columns and the next state $n(t+1)$ in the rows. Note that the only non-zero terms are those that are on the diagonal or adjacent to the diagonal. Note also that in this population the total size is 8, such that when $n = 0$ or $n = 8$, then the only non-zero term is $T(0|0)$ and $T(8|8)$ respectively, given that such systems have reached local extinction or fixation.

difference, such that $s > 0$ represents a selective advantage and $s < 0$ represents a selective disadvantage. The rescaled transmission matrix can be written as

$$T(n'|n) = \begin{cases} (1+s)\frac{n}{N}\left(1-\frac{n}{N}\right) & n' = n+1 \\ \frac{n}{N}\left(1-\frac{n}{N}\right) & n' = n-1 \\ 1 - T(n-1|n) - T(n+1|n) & n' = n \\ 0 & otherwise \end{cases} \tag{4.37}$$

Given this transmission matrix, what is the probability that type A fixes given that there are initially n individuals of type A in the population?

To answer this question, denote π_n as the probability of fixation given that there are n individuals of type A. Because the Moran process is a Markov process, then π_n is independent of time, e.g., the future chance of fixation will be the same whether or not the system has $n = 20$ individuals at time 0, 5 or 500. We can leverage this Markovian process to write down a recurrence relationship amongst fixation probabilities

$$\pi_n = T(n+1|n)\pi_{n+1} + T(n-1|n)\pi_{n-1} + T(n|n)\pi_n \tag{4.38}$$

In essence, this says that the probability of fixation given n individuals can be decomposed into the probability of fixation given the three possible outcomes at the next moment ($n-1$, n, and $n+1$) multiplied by their corresponding fixation probabilities.

After some tedious algebra (see the Technical Appendices), it is possible to conclude that

$$(\pi_{n+1} - \pi_n) = \frac{\pi_n - \pi_{n-1}}{1+s}. \tag{4.39}$$

This recurrence relationship provides a route to find π_n. It helps that we already know 2 values, irrespective of s. That is, $\pi_0 = 0$ and $\pi_N = 1$. Hence, the recurrence relationship can help once we realize that if $\pi_1 = C$ then we can use the "boundary conditions" to write

$$\pi_2 - \pi_1 = \frac{C}{1+s} \tag{4.40}$$

and by extension

$$\pi_{n+1} - \pi_n = \frac{C}{(1+s)^n}. \tag{4.41}$$

By summing over all values from $k = 0$ to $n-1$ one can arrive at the final result:

$$\pi_n = \frac{1-(1+s)^{-n}}{1-(1+s)^{-N}} \tag{4.42}$$

which in the limit of very large populations becomes

$$\pi_n = 1-(1+s)^{-n}. \tag{4.43}$$

This allows us to ask: what is the chance that a new mutant in a very large population have of reaching fixation? The answer should be π_1, or

$$\pi_1 = \frac{s}{1+s} \tag{4.44}$$

Moreover, if the fitness is much less than 1, then the relative fitness advantage s denotes the approximate probability of fixation. That is a 10% fitness advantage yields an approximately 10% chance of fixation in an infinite population. Seems pretty big! Likewise, for deleterious mutations, then

$$\pi_1 = (1+s)^{N-1} \tag{4.45}$$

so that when $N|s| \gg 1$ it becomes virtually impossible for a deleterious mutation to fix. Hence, with a 10% fitness advantage, then populations beyond 10 or so individuals would rarely have such a mutant fix, whereas very weakly deleterious mutation, e.g., with a 0.01% fitness disadvantage might reach fixation in a population of 1000. That is to say: fitness need not always increase in real populations, no matter what Fisher's fundamental theorem might tell you in the limit of infinite populations and deterministic dynamics!

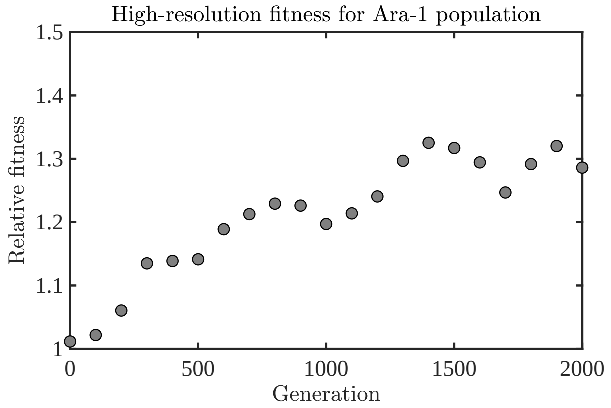


Figure 4.13: High-resolution sampling of fitness in 2000 generations reveals stepwise changes in fitness.

4.6 SWEEPS OR HITCHIKING OR BOTH?

4.6.1 Apparent punctuated equilibrium, sweepts, and the LTEE

The start of this chapter began with a puzzle: are evolutionary dynamics repeatable and, if so, are they repeatable at the scale of genotypes or phenotypes (or both)? The work of Richard Lenski and colleagues has shown that evolutionary dynamics is repeatable, and that organisms (like *E. coli*) that face the same selection pressure can evolve, independently, leading to systematic changes in fitness. In the case of the LTEE, these changes also coincided with changes in cellular physiology: cells got bigger – though surprisingly, Lenski and Travisano were unable to prove the direct mechanistic between cell size and fitness (Lenski and Travisano 1994). But what they did find was the fitness increased in each lineage in a seemingly continuous fashion. That continuos increase belied a mechanism of rapid sweeps. Figure 4.13 shows that the fitness of the Ara-1 population was characterized by plateaus and then rapid changes in fitness. These rapid changes appear to support Stephen Jay Gould and Niles Eldredge's proposal of punctuated equilibrium (Gould and Eldredge 1993). The population remains in an evolutionary stasis (from the perspective of fitness) until a mutation that confers a benefit appears and then rapidly spreads in the population, leading to a relatively rapid shift in fitness. Indeed, this work suggests that *E. coli* are limited – in a mutational sense – such that only when beneficial mutations appear do they rapidly spread in the system. Yet, as limited, it would also suggest that the timing of these jumps will vary across experiments and associated populations.

One of the reasons why there can be variation in evolutionary dynamics is that, as we have shown, not all beneficial mutants fix. A beneficial mutant with fitness benefit s will invade a fraction $s/(s+1)$ of the time in the infinite population limit. Moreover, there can be extended periods where the 'more-fit' mutant persists at relatively low frequencies. It is only once the number of mutants with a fitness advantage exceed some critical level, usually on the order of $1/s$, that they increase exponentially. That is, if there is a fitness advantage of 10% then it requires 10 individuals (irrespective of the final population size) to likely lead to fixation, whereas it would take 100 individuals if the fitness advantage is only 1%. Hence, even a selectively beneficial gene will experience predominantly the stochastic effects of drift until it reaches a critical subpopulation size. These insights leads to the open question to many is whether or not mutants with beneficial alleles (or genotypes) appear frequently or rarely. The use of such adjectives implies a time scale of reference. Hence, for reference, consider the time scale to be the fixation or "sweep" time – this should be on the order of $\tau_s = \log N/s$. In that event, the following operational definitions will be helpful:

Limiting beneficial mutation regime: new mutations that appear that benefit the orgnanism happen very slowly (e.g., more slowly than the time over which such mutations can sweep).

Overlapping beneficial mutation regime: new mutations that appear that benefit the orgnanism happen relatively quickly (e.g., faster than the time over which such mutations can sweep).

Returning to the first dataset from Lang et al. (Lang et al. 2013), the question becomes what evidence, if any, is there to suggest that real populations may be in the overlapping benficial mutation regime – a regime that may not necessarily have been self-evident when looking at evolutionary dynamics after long intervals in measurement. The early data from the LTEE suggests that for this particular experiment: *E. coli* growing in a limited carbon resource environment may be mutationally limited. But, thankfully for the study of evolutionary biology, sequencing capabilities have increased so it is possible to delve in greater detail, not just in terms of the temporal resolution of fitness but also of complete genomes. It turns out there were many mutations, some of small effect, some of large effects, and some that surprised even the experimental team (see the Cit+ story for more (Blount et al. 2008)). But to see evidence tha sheds light on the question of the operative mutational regime for microbes, we will turn to another lab workhorse - Baker's yeast, i.e., *Saccharomyces cerevisiae.*

4.6.2 Evidence for clonal interfence and hitchiking in microbial populations

Lang et al. (Lang et al. 2013) used an experimental evolution approach to analyze competition of yeast in a serial dilution framework. That is, each day a fraction of the flask was resuspended in fresh media, allowed to grow, resuspended in fresh media, and so on. Then, samples were preserved approximately every eighty generations over a one-thousand generation experiment, i.e., at generations 0, 140, 240, 335, 415, 505, 585, 665, 745, 825, 910, and 1000. At each

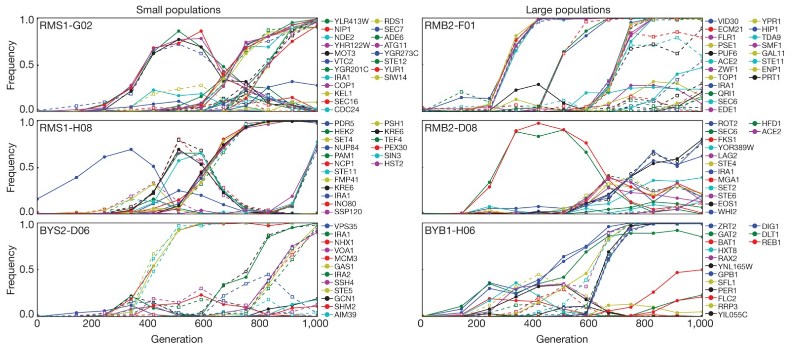


Figure 4.14: Trajectories of spontaneous mutations in evolving yeast populations. Solid lines denote synonymous mutations and dashed lines denote nonsynonymous or intergenic mutations. Reproduced from (Lang et al. 2013).

timepoint, the entire population was sequenced using a whole genome based approach that could use reference genomes to identify mutations across 40 distinct populations. The team identified 1,020 mutations. As such, each particular experiment yielded a unique set of mutational trajectories (see Figure 4.14). Via a high-resolution approach it is apparent that the system is not mutationally limited. That is, multiple examples reveal the emergence of mutations that appear to be on their way to sweeping when another set of mutations appear and then start to incrase, and then another, and so on. But the challenge is that because the experiment could identify mutations and their frequency, it is not immediately apparent how (or whether) these mutations are linked. Yet there are clues.

As is apparent, there are coincident increases in the frequency of multiple mutations in a population (see the overlapping lines in both the small and large population figures in Figure 4.14). This could mean that the yeast genome luckily experienced multiple concurrent beneficial mutations, or perhaps many mutations all of a similar effect size. But, if so, how come so many mutations appeared all at once? Instead, it may be that some of these mutations were not beneficial, but were actually neutral, and therefore that real evolutionary dynamics must be expanded from the simplified concept of sweeps and stasis to one in which there is a bubbling up of mutation after mutation that sometimes sweep in cohorts to partial fixation. This is the subject of one of the problem sets, but at least some additional narrative evidence is worthwhile to consider. At first, a beneficial mutation appears and begins to sweep through the population. Let's call this genome A. But then another mutation appears on a different genome (genome u), potentially with neutral mutations in the background, and genome u begins to outcompete the original sweeping genome. However, then, a new mutation appears in the A background, let's call this genome AB and it (and its neutral mutations) reover and outcompete u and its mutations. This

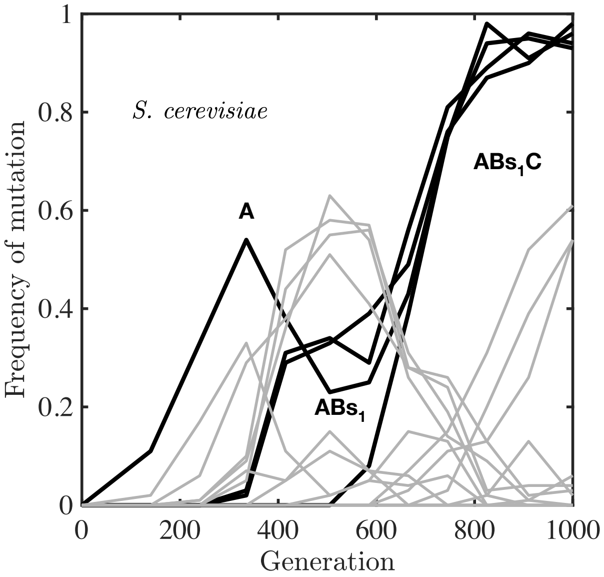


Figure 4.15: A sequence of beneficial mutations and neutral mutations that hitchike offer a parsimonious explanation of the concurrent rise of multiple mutations in the population. Reproduced from (Lang et al. 2013).

process repeated many times over, with variation, i.e., with non-neutral mutations recurring time and again, with neutral mutations also sweeping (although not repeating). Figure 4.15 shows precisely this point. One way to verify this is to ask if the genomic sites of non-neutral mutations repeat (hint: they do) and whther or not the sites of synonymous mutations repeat (hint: they do not). In the end, this experiment revealed novel evolutionary dynamics by looking with higher resolution – finding that sweeps were really comprised of near-sweeps, hitchiking, and reproducibility despite the apparent randomness of many of the sweeping mutations. Like the old adage a watched pot never boils, perhaps a carefully observed evolution experiment never really sweeps!

4.7 TAKE-HOME MESSAGES

- Evolutionary dynamics arise due to a combination of deterministic and stochastic processes.
- There are multiple mechanisms that can lead to the purging of diversity, insofar as pure replication implies that the type that grows the fastest should outcompete competing replicators.
- Although Fisher's Fundamental Theorem suggests that fitness must always

increases, this theorem has severe restrictions and is of limited value in practice.

- There are multiple evolutionary mechanisms that can restore diversity including mutation-selection balance and negative frequency-dependent selection. Other processes, including sexual reproduction and heterozygote fitness advantages are outside the scope of this chapter.
- The study of how stochasticity impacts evolutionary dynamics is usually framed in terms of models of neutral drift, e.g., the Wright-Fisher or Moran model.
- In both cases, stochastic drift leads to either local extinction or fixation in finite populations over time scales set by the size of the population.
- In the event of selective benefits, then even rare mutants can lead to fixation, but such fixation is not inevitable and depends on the size of the population and the selective advantage.
- Recent observations of contemporary evolutionary dynamics suggests that unlike assumptions of slow evolution and limiting beneficial mutations, that natural systems often operate with many mutations competing at once (i.e., a term called clonal interference) even as many neutral mutations are linked to beneficial mutations as part of multi-mutation sweeps (i.e., via a process termed genetic hitchhiking).

4.8 HOMEWORK PROBLEMS

These problems builds upon a growin set of computational tools, including methods elaborated on in the associated lab guide. In order to work on these problems you should be prepared to

- Translate a Markov Chain into a numerical representation
- Simulate an ensemble of trajectories
- Compare individual trajectories to that expected given the transition probabilities
- Apply basic principles of Markov Chain theory to the problem of population genetics
- Observe the fixation of neutral alleles due to genetic drift

These problems leverage these skills and the theory in this chapter to deepen your understanding of population genetics and evolutionary dynamics, including a chance to analyze recent data. The overall objective is to explore the principles by which key factors, including population size and selection coefficients, play in shaping time scales over which population structure changes.

Hint: for the Moran model you will want to implement stochastic birth-death events, one at a time, that is one random individual is chosen to die and one random individual replaces it. Hence one needs about N Moran events to be equivalent to one WF step. The key here is to think about how to incorporate differences in the probability of reproduction (e.g., birth) of some individuals relative to others.

Problem 1. How Fast is Drift? Part 1.

Here you will build on the computational laboratory analysis of the Wright Fisher (WF) model to examine variability in the speed of fixation/extinction given neutral genetic drift. First, using the WF model with N=200 individuals, begin with 100 type A individuals and show three representative samples that reach either fixation or extinction. What is the approximate number of time steps you expect the process to take? Next, run an ensemble of stochastic trajectories (hint: use 100 to have sufficient statistics). Do the same starting with 25, 50, 100, 150 and 175 individuals of type A. What is the relationship between the fraction of A individuals and the probability of fixation - and how does it compare to theoretical expectations?

Problem 2. How Fast is Drift? Part 2.

Using N=100 individuals of which 50 are type A, run a WF model using an ensemble of at least 200 trials. In each case track how long it takes for A either to go extinct or to fix? What do you find? If you were to decrease/increase the initial fraction of individuals type A to either 20 or 80, how long would the process take? Finally, using the Markov Chain transition matrix, find the theoretical expectation for the time-varying probability of extinction and fixation. Explain your rationale and comparisons to the results of stochastic realizations amongst the ensemble.

Problem 3. A Sense of Scale

Develop a Moran model where Type A individuals have an enhanced fitness of $(1+s)$ relative to type B individuals. That is their reproductive chances are enhanced by a factor of $(1+s)$ though their death rates are the same. Starting with a single individual in a population of N=500, simulate the Moran process until the mutant population goes extinct or fixes (Hint: don't simulate every individual, but rather the cohort of A-s and B-s all at once). Use $s = 0.05$ and find three examples in which the population fixes. How many simulations did it take to get these 3 examples? Explain your findings and what it says about the invasibility of beneficial mutations of (relatively) small effect.

Problem 4. A Sense of Scale Continued

Using your Moral model, modulate the mangitude of the beneficial mutaation s from 0.05 to 0.5 in increments of 0.05 use ≈ 100 ensembles and answer the question: how often does a beneficial mutation fix? How does this compare to theory? Finally, recall that we expected a critical number of individuals with a beneficial mutation so that the mutation would almost certainly fix. We identified this scale as 1/s. Focus on the case $s = 0.05$ and re-do your ensemble analysis by varying the initial number of mutants and its effect on the probability of fixation. What do you conclude about the critical establishment size in this model?

Problem 5. Overlapping or limiting beneficial mutations

The paper of Lang et al. (Lang et al. 2013) measures the change in mutations with time. Here, you will explore these patterns focusing on one population G7. First, load the data and plot the frequency of mutations over time. Using two lines of quantitative evidence, make an argument as to why this population experiment is in the limiting beneficial mutation regime or the overlapping beneficial mutation regime. This problem is intentionally open-ended.

Additional Challenge Problems.

- How does the time to fixation/extinction scale both with N the size of the population, s the fitness benefit of the mutation, and x the initial fraction of the population that has a beneficial mutation? Can you identify any potential cross-over regimes based on the benefit of the mutation and the population size?
- The paths in the Moran model look diffusive. Can you develop a scaling formalism for the change in the relative frequency of a neutral allele in a population of size N given there is x fraction of the neutral allele present?
- Develop a clustering algorithm to automatically group mutants together that are likely on the same genome using the Lang et al. data.

4.9 TECHNICAL APPENDICES

Dynamics of heterozygosity: The heterozygosity of a population is equal to the probability that two randomly chosen individuals in a population of size N have different genotypes (or alleles). As noted in the main text, this is defined as:

$$H = 2\frac{n}{N}\left(1 - \frac{n}{N}\right) \tag{4.46}$$

where n is the number of A genotypes, $N-n$ is the number of B genotypes, and therefore n/N is the fraction of A genotypes and $1-n/N$ is the fraction of B genotypes. The expected dynamics of heterozygosity, $H(t)$ can be computed by calculating the value one generation ahead, where $E(H)' \equiv \langle H(t+1)\rangle$:

$$E(H') = E\left(2\frac{n'}{N}\left(1 - \frac{n'}{N}\right)\right). \tag{4.47}$$

This is formally equal to selecting A then B or B then A when randomly selecting two individuals in the population; either way the genotypes are different. The expected value of H' depends on the frequency and variance of genotypes at the next generation.

Heterozygosity in the Wright-Fisher model: In the WF model, the dynamics of heterozygosity are:

$$\begin{aligned}
E(H') &= \frac{2}{N}E(n') - \frac{2}{N^2}E\left((n')^2)\right) && (4.48)\\
&= \frac{2n}{N} - \frac{2}{N^2}\left[Var(n) + (E(n'))^2\right] && (4.49)\\
&= \frac{2n}{N} - \frac{2}{N^2}\left[n(1-n) + n^2\right] && (4.50)\\
&= \frac{2n}{N} - \frac{2n}{N^2} + \frac{2n^2}{N^3} - \frac{2n^2}{N^2} && (4.51)\\
&= \frac{2n}{N}\left[1 - \frac{1}{N} + \frac{n}{N^2} - \frac{n}{N}\right] && (4.52)\\
&= \frac{2n}{N}\left(1 - \frac{n}{N}\right)\left(1 - \frac{1}{N}\right) && (4.53)\\
&= H\left(1 - \frac{1}{N}\right). && (4.54)
\end{aligned}$$

The final result implies that $E(H(t+1)) = H(t)\left(1 - \frac{1}{N}\right)$ or more generally

$$E(H(t)) = H_0\left(1 - \frac{1}{N}\right)^t, \tag{4.55}$$

i.e., heterozygosity is expected to decay exponentially such that when $N \gg 1$ then

$$E(H(t)) = H_0 e^{-t/N} \tag{4.56}$$

with a time-scale of decay on the order of N generations. This result and its implications are discussed in the main text.

Heterozygosity in the Moran model: In the Moran model, the dynamics of heterozygosity are:

$$E(H') = E\left(2\frac{n'}{N}\left(1-\frac{n'}{N}\right)\right). \tag{4.57}$$

where $E(n') = n$ by the Martingale property and

$$\begin{align} E(n'^2) &= \Sigma\,(n')^2\,P(n') \tag{4.58}\\ &= (n-1)^2T(n-1|n) + n^2T(n|n) + (n+1)^2T(n+1|n) \tag{4.59}\\ &= (n^2-2n+1)T(n-1|n) + n^2(1-T(n-1|n)-T(n+1|n)) + \notag\\ &\quad (n^2+2n+1)T(n+1|n) \tag{4.60}\\ &= n^2 + 2\frac{n}{N}\left(1-\frac{n}{N}\right), \tag{4.61} \end{align}$$

given that $T(n-1|n) = T(n+1|n) = \frac{n}{N}\left(1-\frac{n}{N}\right)$ for all $n > 0$ and $n < N$. Combining terms yields:

$$\begin{align} E(H') &= \frac{2n}{N} - \frac{2}{N^2}\left[n^2 + \frac{2n}{N}\left(1-\frac{n}{N}\right)\right], \tag{4.62}\\ &= \frac{2n}{N} - \frac{2n^2}{N^2} - \frac{4n}{N^3} + \frac{4n^2}{N^4}, \tag{4.63}\\ &= \frac{2n}{N}\left(1-\frac{2}{N^2}\right) - \frac{2n^2}{N^2}\left(1-\frac{2}{N^2}\right), \tag{4.64}\\ &= \frac{2n}{N}\left(1-\frac{n}{N}\right)\left(1-\frac{2}{N^2}\right) \tag{4.65}\\ & \tag{4.66} \end{align}$$

which is equivalent to

$$E(H') = H(n)\left(1-\frac{2}{N^2}\right). \tag{4.67}$$

such that for $N \gg 1$

$$E(H(t)) = H_0e^{-2t/N^2} \tag{4.68}$$

implying that it takes on the order of N Moran steps to yield one generation in the WF model.

Fixation in the Moran model with selection: Consider the following recurrence relationship amongst fixation probabilities

$$\pi_n = T(n+1|n)\pi_{n+1} + T(n-1|n)\pi_{n-1} + T(n|n)\pi_n \tag{4.69}$$

for the Moran model with selection. This can be written as

$$\begin{aligned} \pi_n &= \frac{T(n+1|n)\pi_{n+1} + T(n-1|n)\pi_{n-1}}{1-T(n|n)}\\ &= \frac{T(n+1|n)\pi_{n+1} + T(n-1|n)\pi_{n-1}}{T(n-1|n) + T(n+1|n)} \end{aligned}$$

Recall that the difference between $T(n-1|n)$ and $T(n+1|n)$ is a factor of $(1+s)$. Hence, We can rewrite this as

$$\pi_n = \frac{1}{2+s}\pi_{n-1} + \frac{1+s}{2+s}\pi_{n+1} \tag{4.70}$$

or by re-arranging terms

$$(2+s)\pi_n = \pi_{n-1} + \pi_{n+1}(1+s) \tag{4.71}$$

and again by re-arranging

$$\pi_n - \pi_{n-1} = (1+s)\left(\pi_{n+1} - \pi_n\right) \tag{4.72}$$

and finally

$$\left(\pi_{n+1} - \pi_n\right) = \frac{\pi_n - \pi_{n-1}}{1+s} \tag{4.73}$$

as examined in the main text.

Measuring the strength of selection given change in the frequency of variants: Consider two variants growing in a population, such that the frequency of type A grows from an initial value of x_0 over time as $x(t)$ given a selective growth advantage of s

$$x(t) = x_0\left[\frac{e^{st}}{1+x_0\left(e^{st}-1\right)}\right]. \tag{4.74}$$

Given a measurement time t_1 and an observation of $x_1 = x(t_1)$, what is the estimated selective growth advantage? This framework is the standard approach to estimate fitness advantages in direct competition between two types. In this case, Eq. (4.74) can be rewritten as

$$\begin{aligned}
x_1\left(1+x_0\left(e^{st}-1\right)\right) &= x_0 e^{st} && (4.75)\\
x_1(1-x_0) + x_1 x_0 e^{st} &= x_0 e^{st} && (4.76)\\
x_1(1-x_0) &= x_0(1-x_1)e^{st} && (4.77)\\
e^{st} &= \frac{x_1}{1-x_1}\frac{1-x_0}{x_0} && (4.78)
\end{aligned}$$

such that finally, the estimated selection coefficient is:

$$\hat{s} = \frac{1}{t_1 - t_0}\log\left[\frac{x_1}{1-x_1}\frac{1-x_0}{x_0}\right]. \tag{4.79}$$

This estimate is used in practice to convert frequency changes to estimates of the selection coefficient.