Building a heuristic understanding of natural selection

William R. Shoemaker

December 3rd, 2025

Outline

In this lecture we will provide an overview of the basic timescales and dynamics of a population evolving under selection. We will emphasize asymptotic and heuristic approaches to build an intuition on what is and is not to be expected of an evolving ideal population (e.g., no ecology, spatial structure, etc.). We will reference studies that directly use our results from theory to investigate both natural and experimental microbial populations. Specifically, how 1) over short timescales one can use heuristics to quantify the evolutionary dynamics of individual *de novo* mutations in laboratory yeast experiments and 2) over long timescales interpret patterns of molecular evolution in the human gut microbiome. A final (optional) section derives one of our core items of interest, the probability of fixation p_{fix}, via exact solutions.

Asymptotic and heuristic analysis

We will be working with a Langevin equation that governs the dynamics of the frequency of a mutant allele with selective advantage s over the wild type in a population of size N. There are two ingredients to consider: 1) selection as deterministic logistic growth and 2) random genetic "drift" (i.e., multiplicative noise)

$$\frac{df}{dt} = \underbrace{s \cdot f(1-f)}_{\text{selection}} + \underbrace{\sqrt{\frac{f(1-f)}{N}}\eta(t)}_{\text{genetic drift}} \tag{1}$$

where $\eta(t)$ is a standard Brownian noise term. While Eq. 1 appears straightforward, it does not admit a closed form solution. Therefore, we will work to build a heuristic understanding of the molecular evolutionary dynamics of a mutation at a single site. Such an approach is advantageous in that it allows us to relax some of our assumptions (e.g., that N or s is constant with time). It also takes advantage of how the dynamics of real populations are often the outcome of either large (e.g., N) or small numbers (e.g., s or the mutation rate),

the product of which is either very large or very small. We can then obtain a reasonable picture solely considering the parameter regimes $N|s| \ll 1$ and $N|s| \gg 1$. This manuscript will primarily present previously derived results [1] (Ben's notes), with similar physical heuristic arguments having been applied elsewhere to population genetic questions [2, 3].

Dynamics under neutrality

When $N|s| \to 0$ our Langevin reduces to

$$\frac{df}{dt} = \sqrt{\frac{f(1-f)}{N}}\eta(t) \tag{2}$$

Because selection if unbiased $d_t \langle f(t) \rangle = 0$ and $\langle f(t) \rangle = f(0)$. The mutation must ultimately become fixed or go extinct, meaning

$$\lim_{t\to\infty} \langle f(t) \rangle = 1 \cdot p_{\text{fix}} + 0 \cdot (1 - p_{\text{fix}}) \tag{3}$$

Together these relations tell us that the probability of fixation of a neutral mutation is its frequency. It is also worth noticing how the ensemble average of a neutral mutation can be a misleading measure of the *typical* value of $f(t)$ over extended timescales. So what does the trajectory of a neutral mutation look like and how long does it take?

First, consider the regime $f(t) \ll 1$, where the mutant changes frequency with infinitesimal increments δf as

$$\delta f \equiv f(t + \delta t) - f(t) = \sqrt{\frac{f(t)\delta t}{N}} \cdot \mathcal{N}(0, 1) \tag{4}$$

But in real populations we measure allele frequencies over finite increments of time, Δt, not infinitesimal, meaning that finite changes in f alter the diffusion constant. We can investigate the consequences of the finite timescale using a crude form of integration. First, Eq. 4 is a reasonable approximation if $\Delta f \ll f$. Over this same timescale $\langle f(t) \rangle$ is a reasonable approximation of the typical value of $f(t)$. However, this approximation breaks down for times of order $\Delta t \sim Nf$ when $\Delta f \sim \pm f$. The perturbation will be negative with probability $\frac{1}{2}$ (genetic drift is unbiased) and likely go extinct. With probability $\frac{1}{2}$ the perturbation is positive, meaning that the mutation survives and the process repeats with an initial frequency of $2f$. After $2Nf$ generations the mutation will be near extinct or have grown to frequency $4f$. Iterating this process k times gives us the frequency $f(t + \Delta t) = f(t)2^k$ with probability $(\frac{1}{2})^k$ and it occurs after time $\Delta t \sim Nf(t) \sum_{k'=0}^{k-1} 2^{k'}$. There are two equivalent ways to summarize the behavior of the trajectory, depending on whether one manipulates Δt or Δf (as one theoretically do in experiments).

1. Survive for Δt generations, growing to size $f + \frac{\Delta t}{N}$ with probability $\frac{Nf}{Nf+\Delta t}$
2. Reach size $f + \Delta f$ over time $\Delta t = N\Delta f$ with probability $\frac{f}{f+\Delta f}$

Note how these results scale linearly with Δt, as opposed to the square root dependence over short timescales.

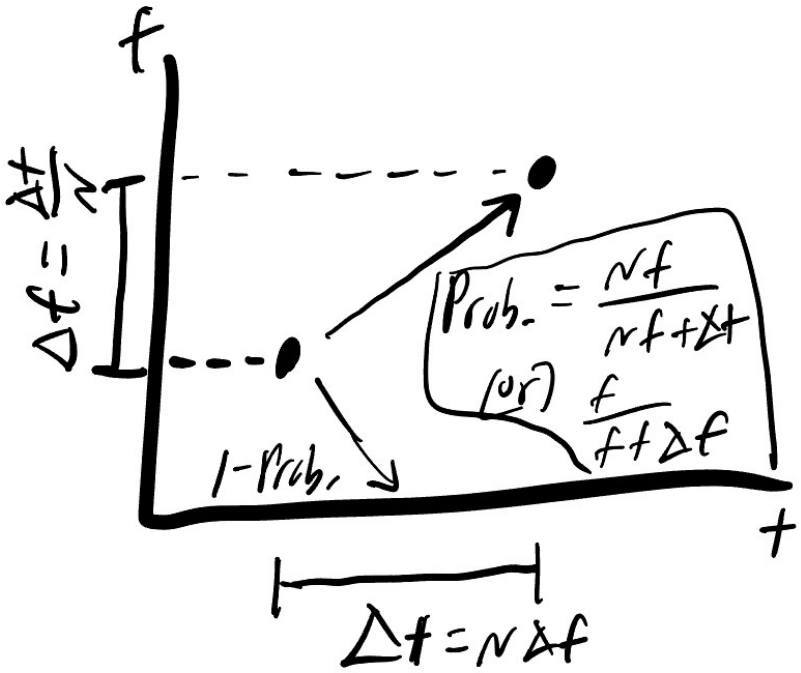

Figure 1: **Neutral dynamics in our heuristic analysis.**

Fixation occurs when $f \sim \mathcal{O}(1)$. Using the above results we obtain the probability and timescale of fixation

$$p_{\text{fix}} \sim \frac{f}{f+1} \sim f \tag{5}$$

$$p_{\text{fix}} \sim \frac{f}{f+1} \sim f \tag{6a}$$

$$T_{\text{fix}} \sim N \cdot 1 \sim N \tag{6b}$$

Dynamics under selection

In the opposite limit $N|s| \to \infty$ one would naively expect that genetic drift can be ignored, providing us with a deterministic ODE of logistic growth

$$\frac{df}{dt} = s \cdot f(1-f) \tag{7}$$

the solution of which is

$$f(t+\Delta t) = \frac{f(t)e^{s\Delta t}}{1-f(t)+f(t)e^{s\Delta t}} \tag{8}$$

This equation seems reasonable, but upon closer inspection we find that it predicts $p_{\text{fix}} \approx 1$ and $T_{\text{fix}} = \frac{1}{s}\ln\left(\frac{Nf}{1-f}\right) \to \infty$ generations for fixation to occur!

These counterintuitive results arose because we ignored genetic drift, something that cannot be done even in the limit $N|s| \to \infty$. Let's again examine frequency increments in the limit $f \ll 1$

$$\delta f \sim sf(t)\delta t + \sqrt{\frac{f(t)\delta t}{N}} \cdot \mathcal{N}(0,1) \tag{9}$$

As before, this is a reasonable approximation for Δt where $\Delta f \ll f$, but breaks down when $\Delta f \sim f$, or

$$f \sim |s| f \Delta t + \sqrt{\frac{f\Delta t}{N}} \cdot \mathcal{N}(0,1) \tag{10}$$

Selection is then dominant when $N|s|f \gg 1$ but drift dominates when $N|s|f \ll 1$. Such qualitatively different behavior suggests that a complete approximation requires a piecewise model

$$\frac{df}{dt} \approx \begin{cases} \sqrt{\frac{f}{N}}\,\eta(t) & \text{if } f \lesssim \frac{1}{N|s|} \\ sf(1-f) & \text{if } f(1-f) \gg \frac{1}{N|s|} \\ \sqrt{\frac{1-f}{N}}\,\eta(t) & \text{if } 1-f \lesssim \frac{1}{N|s|} \end{cases}$$

In terms of experiments one can actually perform, mutant lineages are usually mixed with their ancestor at frequencies that exceed $\frac{1}{N|s|}$, meaning that once can estimate s using Eq. 8

$$s = \frac{1}{\Delta t}\ln\left(\frac{1-f(t)}{f(t)} \cdot \frac{f(t+\Delta t)}{1-f(t+\Delta t)}\right) \tag{11}$$

This is the standard formula used to infer the strength of selection in experiments (e.g., the fitness trajectories from the Long-Term Evolution Experiment (ref)). However, this formula cannot be directly applied to *de novo* mutations as $\frac{1}{N} \ll \frac{1}{N|s|}$, meaning that the probability of fixation is shaped by both drift and selection.

Mutations that start at $f(t) \ll \frac{1}{N|s|}$ initially evolve neutrally, reaching size $\frac{1}{N|s|}$ with probability $N|s|f(t)$ over time $\Delta t \sim \frac{1}{s}$. At this point selection dominates if the mutation is beneficial, meaning that the dynamics can be approximated as deterministic, growing logistically until they reach size $1 - \frac{1}{Ns}$ after

$\Delta t \approx \frac{2}{s}\ln(Ns-1) \sim \frac{1}{s}\ln(Ns)$ generations. At this point the beneficial mutation will drift to fixation within $\frac{1}{s}$ generations. In contrast, deleterious mutations deleterious mutations cannot exceed a frequency of $\frac{1}{N|s|}$ and will typically drift to fixation over time $\Delta t \sim \frac{1}{|s|}$. If a deleterious mutation finds itself above frequency $1 - \frac{1}{N|s|}$ it will fix with probability $N|s|(1-f)$. Combining the above results, for beneficial mutations we obtain

$$p_{\text{fix}}(s > 0, f) \sim \begin{cases} 1 & \text{if } f \gtrsim \frac{1}{Ns} \\ Ns & \text{if } f \lesssim \frac{1}{Ns} \end{cases}$$

whereas for deleterious mutations

$$p_{\text{fix}}(s < 0, f) \sim \begin{cases} 0 & \text{if } 1 - f \gtrsim \frac{1}{N|s|} \\ N|s|(1-f) & \text{if } 1 - f \lesssim \frac{1}{N|s|} \end{cases}$$

For a *de novo* mutation everything simplifies to

$$p_{\text{fix}}(s, 1/N) \sim \begin{cases} s & \text{if } s > 0 \\ 0 & \text{if } s < 0 \end{cases}$$

Given that a mutation will become fixed, the time it takes to do so is then

$$T_{\text{fix}} \approx \frac{1}{s} + \frac{2}{s}\ln(Ns) + \frac{1}{s} \sim \frac{1}{s}\ln(Ns) \tag{12}$$

Unlike neutral mutations which spend equal amounts of time at intermediate frequencies, much of the fixation time of beneficial mutations is spent at frequencies where selection does not occur. The dynamics of a mutation are then *effectively* neutral if $|s| \lesssim \frac{1}{N}$, known as the *drift barrier* as it represents limitations on the fidelity of selection [4, 5]. One can calculate the strength of selection on cellular features (e.g., # ATP required to build an excess nucleotide vs. # ATP required to build a cell) and (with several grains of salt) estimate N for a given species, providing the means to assess whether natural selection is feasible [6, 7]

The above derivations have played a crucial role in the design and interpretation of microbiological evolution experiments, specifically the dynamics of *barcoded lineages*. While whole genome sequencing has substantially declined in cost, it is still prohibitively expensive to sequence thousands of individual isolate genomes in order to track the dynamics of *de novo* mutations. Instead, researchers can genetically design strains of asexual (i.e., no recombination) microorganisms where random short nucleotide sequences are inserted in a specific location for individual cells. The cells are then allowed to grow in a batch culture setting, where they acquire *de novo* beneficial mutations. These mutations are physically linked to a specific barcode as they are both on the same chromosome. Therefore, sequencing only the barcodes provides information about

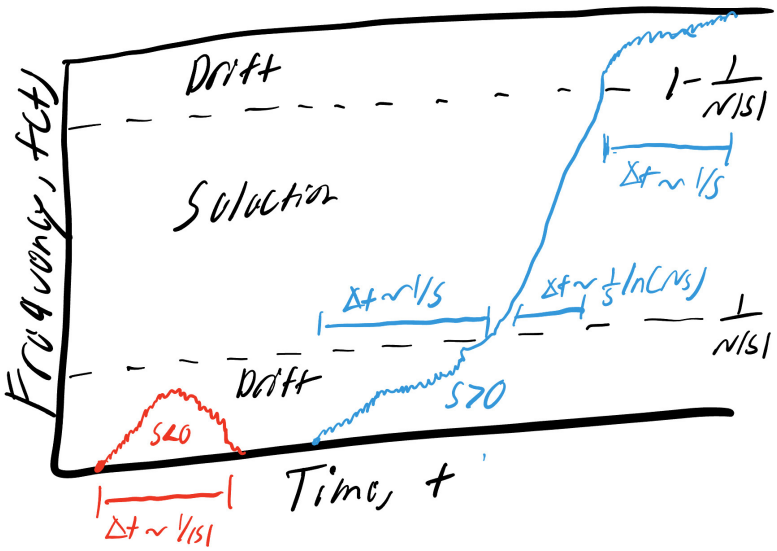

Figure 2: **Piecewise picture of evolutionary dynamics under selection.**

the beneficial mutation. The results presented above provide the means to accurately infer s while accounting for both selection and drift, providing the means to empirically investigate the *distribution of fitness effects* of *de novo* mutations [8].

Exact results

Under the Itô $\leftrightarrow$ Fokker–Planck correspondence we can write out the above Langevin as the partial differential equation of the probability density $\phi(f,t)$.

$$\frac{\partial\phi(f,t)}{\partial t} = \frac{\partial^2}{\partial^2 f}\left[\frac{f(1-f)\phi(f,t)}{2N}\right] - \frac{\partial}{\partial f}\left[sf(1-f)\phi(f,t)\right] \tag{13}$$

This equation *is linear in* $\phi(f,t)$, so formal solutions can be derived [9]. However, this solution can be unwieldy. Instead, it is easier instead to work with the *Moment Generating Function* (MGF) which can be obtained by performing a Laplace transformation over the domain f.

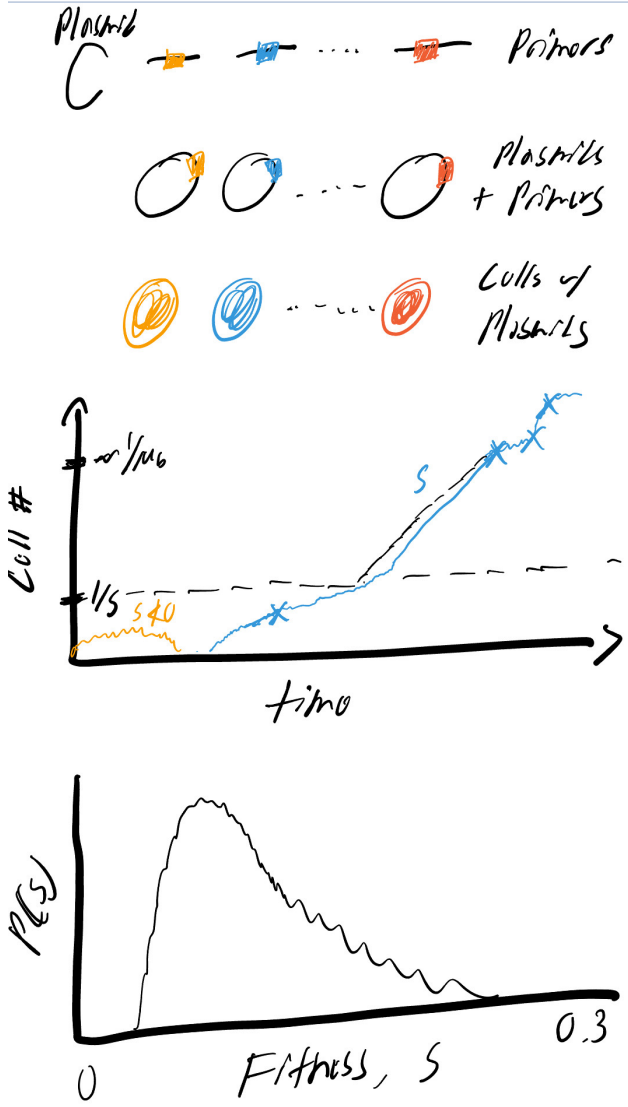

Figure 3: **Barcoding as a means to infer fitness effects of individual beneficial mutations.**

$$\frac{\partial H(z,t)}{\partial t} = \frac{\partial}{\partial t}\int_0^1 e^{-zf}\phi(f,t)df \tag{14a}$$

$$= \int_0^1 e^{-zf}\left(\frac{\partial^2}{\partial^2 f}\left[\frac{f(1-f)\phi(f,t)}{2N}\right] - \frac{\partial}{\partial f}\left[sf(1-f)\phi(f,t)\right]\right)df \tag{14b}$$

$$= z^2\mathcal{L}\left\{\frac{f(1-f)\phi(f,t)}{2N}\right\} - z\mathcal{L}\left\{sf(1-f)\phi(f,t)\right\} \tag{14c}$$

$$= \left[sz - \frac{z^2}{2N}\right]\left[\frac{\partial H(z,t)}{\partial z} - \frac{\partial H^2(z,t)}{\partial z^2}\right] \tag{14d}$$

where, by definition, $H(z,0) = e^{-zf}$. We have ignored terms containing boundaries $f = 0, 1$ as they represent absorbing boundaries. As $t \to \infty$ the mutation will either fix or go extinct depending on p_{fix}.

$$\lim_{t\to\infty} H(z,t) = 1 \cdot (1 - p_{\text{fix}}) + e^{-z} \cdot p_{\text{fix}} \tag{15}$$

We can derive p_{fix} by connecting the final state to the initial condition at $t = 0$. This would ordinarily require the time-dependent solution, but we can use a previously recognized trick where at $z = 2Ns$ the time derivative of $H(z,t)$ vanishes, allowing us to connect the initial and final values [10].

$$\lim_{t\to\infty} H(2Ns,t) = \lim_{t\to\infty} H(2Ns,0) \tag{16}$$

from which we solve

$$p_{\text{fix}}(s,f) = \frac{1 - e^{-2Nsf}}{1 - e^{-2Ns}} \tag{17}$$

Notice how genetic drift and selection occur together as the compound parameter $N \cdot s$, where the limiting behavior of Eq. 17 is consistent with the heuristic results derived earlier.

Because we have the full solution we can go one step further and use the expression to model the rate of *substitutions* in a population (i.e., all individuals have the same mutation). Assume that mutations enter the population at rate $\mu N \ll 1$ meaning that there is at most one mutation segregating in the population at any point in time. So the population evolves as a sequence of non-overlapping mutation segregation events. By definition mutations enter the population at frequency $f = 1/N$, giving us

$$p_{\text{fix}}(s,1/N) = \frac{2s}{1 - e^{-2Ns}} \tag{18}$$

so that successful mutations (i.e., mutations that become substitutions) occur at rate

$$\lambda = N\mu p_{\text{fix}}(s, 1/N) = \frac{2Ns\mu}{1 - e^{-2Ns}} \tag{19}$$

This result allows us to predict how genetic changes accumulate over time within a population, i.e., how a population *diverges* from its ancestral state over time due to selection. Often in population genetics we are interested in the probability of observing a mutation at a given site, d_N. We can actually calculate this quantity in real data (after considering issues such as sampling) and understand its value in relation to the above model. A single individual from a population will contain the mutation with probability $f(t)$. Imagine we have a large number of *replicate* populations, then $f(t)$ will vary from population to population. We can then obtain the divergence by averaging over this *statistical ensemble*

$$d_N = \langle f(t) \rangle \tag{20}$$

We cannot evaluate this quantity for the general case, but over extended timescales when $N\mu \ll 1$ the mutation will either be fixed or has yet to arise in the bulk of the populations. We can then approximate divergence as

$$d_N \approx \lambda t = \frac{2Ns\mu t}{1 - e^{-2Ns}} \tag{21}$$

Because we rarely know the divergence time or μ it is useful to remove the dependence. We can do this by observing that *neutral* sites diverge at $d_S = N\mu \cdot \frac{1}{N} \cdot t = \mu t$. Where S specifies synonymous sites (as opposed to N, which represents nonsynonymous). It is worth taking a moment to appreciate the implications of this result, as the independence of the rate of neutral evolution from N is crucial prediction that spurred the development of the neutral theory of molecular evolution [11]. Taking the ratio, we obtain

$$\frac{d_N}{d_S} = \frac{2Ns}{1 - e^{-2Ns}} \tag{22}$$

where $d_N/d_S > 1$ and $d_N/d_S < 1$ indicate positive and negative selection, respectively. This quantity is both of historical interest [12, 13] and is actively being investigated in microbial populations in the human gut [14, 15].

References

[1] Good, B. Molecular evolution in rapidly evolving populations (2016).

[2] Desai, M. M. & Fisher, D. S. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798 (2007).

[3] Fisher, D. S. Evolutionary dynamics lectures les-houches summer school, 2006 .

[4] Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).

[5] Lynch, M. Evolutionary layering and the limits to cellular perfection. *Proceedings of the National Academy of Sciences* **109**, 18851–18856 (2012).

[6] Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences* **112**, 15690–15695 (2015).

[7] Karakoç, C., Shoemaker, W. R. & Lennon, J. T. Evolutionary bioenergetics of sporulation. *bioRxiv* 2025–08 (2025).

[8] Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).

[9] Kimura, M. Stochastic processes and distribution of gene frequencies under natural selection (1955). *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers. The University of Chicago Press, Chicago* **144** (1994).

[10] Houchmandzadeh, B. & Vallade, M. Alternative to the diffusion equation in population genetics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **82**, 051913 (2010).

[11] Kimura, M. *The neutral theory of molecular evolution* (Cambridge university press, 1985).

[12] Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution* **3**, 418–426 (1986).

[13] McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the adh locus in drosophila. *Nature* **351**, 652–654 (1991).

[14] Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS biology* **17**, e3000102 (2019).

[15] Liu, Z. & Good, B. H. Dynamics of dn/ds within recombining bacterial populations. *bioRxiv* 2025–09 (2025).