Wright-Fisher Models, Approximations, and Minimum Increments of Evolution

William H. Press
The University of Texas at Austin

January 10, 2011

1 Introduction

Wright-Fisher models [1] are idealized models for *genetic drift*, the process by which the the population frequency of an allele varies with time stochastically, and, in particular, may disappear from the population entirely, or may *fix* in 100% of the population. Wright-Fisher models can be applied to the dynamics of neutral (or nearly neutral) mutations – the vastly dominating case, as emphasized by Kimura [2] – or to the case of alleles that have a fitness advantage or disadvantage, as parameterized by a selection coefficient s.

Wright-Fisher models make three idealized assumptions: [3] (1) Generations are taken to be discrete, so that the population evolves by a discrete-step Markov process. (2) The population size is taken to be fixed, so that alleles compete only against other alleles and not against an external environment. (3) Random mating is assumed. None of these assumptions hold in any real population. Nevertheless, Wright-Fisher has proved to be a useful intuitive guide in real cases, and also a foundation on which more complicated population models can build. [4]

2 A Specific Wright-Fisher Model

There are many closely-related formulations of Wright-Fisher models. The one used here is as follows.

We consider a population of N diploid individuals, so that there are at most $2N$ copies of an allele in the population. N of course actually means N_e the *effective population size* [5, 3], a distinction that is relevant only when considering what happens when the idealizations fail (beyond our discussion here). One generation in a stable total population size produces N offspring.

Let us focus on some allele x that occurs n times (out of $2N$). We can define the allele's probability of occurrence by

$$p \equiv \frac{n}{2N} \tag{1}$$

That is, the number of x alleles is parameterized as $2Np$.

The offspring of a random mating carries 0, 1, or 2 copies of the allele x with probabilities

$$\begin{aligned} P(0) &= (1-p)^2 \\ P(1) &= 2p(1-p) \\ P(2) &= p^2 \end{aligned} \tag{2}$$

which of course sum to 1 for any p.

But let us introduce the possibility that an x allele carries a selective advantage or disadvantage. The definition of an allele's selection coefficient s used here is that an individual with one copy of the allele (heterozygous) should produce a factor $(1+s)$ more offspring per mating than an individual lacking the allele. Here s may be positive or negative. The value $s = -1$ represents an immediately lethal mutation.

The model must also parameterize in some way the case of an individual carrying two copies of the x allele (homozygous). In principle this is a second parameter, independent of s. The homozygous individual could in principal produce many more, or many fewer, than a factor $(1+s)$ offspring. However, for mathematical simplicity, most Wright-Fisher models collapse the two parameters by assuming that the homozygous x individual produces $(1+s)^2$ more offspring. For small s this is approximately twice the advantage (in excess offspring) as the heterozygous x individual, since $(1+s)^2 \approx 1+2s$.

When a favorable mutation is destined to fix in the population, the heterozygous case will be typical early on, while p is small; but the homozygous case will be typical later, when p approaches 1. So, we can view the error introduced by our collapsing to a single s as being roughly equivalent to ignoring a (presumed weak) functional dependence of s on p.

Now, equation (2) is replaced by an equation with the fitness bias,

$$\begin{aligned} P(0) &\propto (1-p)^2 \\ P(1) &\propto 2p(1-p)(1+s) \\ P(2) &\propto p^2(1+s)^2 \end{aligned} \tag{3}$$

The right hand sides sum to $(1+ps)^2$, so to convert the proportion signs to equalities, we must normalize and divide the above equation by this factor.

It is a straightforward exercise to compute the mean and variance of the normalized distribution (3),

$$\begin{aligned} \text{mean} &= 2p\frac{1+s}{1+ps} \\ \text{var} &= 2p(1-p)\frac{1+s}{(1+ps)^2} \end{aligned} \tag{4}$$

So, when s is positive, an individual's mean number of x alleles goes, in one generation of random mating, from $2p$ to something larger than this by a factor $(1+s)/(1+ps)$. (And correspondingly, the number of alleles decreases when s is negative.)

3 Formulation as a Binomial Markov Process

For the population as a whole, one generation of mating (N offspring) is a random process that alters n, the total number of x-alleles. From equation (4), and the additivity of mean and variance for (independent) random matings,

$$\begin{aligned} \langle n \rangle &= N \times \text{mean} = 2Np\frac{1+s}{1+ps} \\ \text{Var}(n) &= N \times \text{var} = 2Np(1-p)\frac{1+s}{(1+ps)^2} \end{aligned} \tag{5}$$

If we momentarily set $s=0$ in equation (5), the case for a neutral mutation, we immediately see that n_{t+1} in generation number $t+1$ is generated by a binomial process whose parameters depend on N and on n_t of the previous generation. That is,

$$n_{t+1} \sim \text{Binomial}(2N, p) = \text{Binomial}\left(2N, \frac{n_t}{2N}\right) \tag{6}$$

where $\sim$ here means "is drawn from". In other words, we have a Markov process, where (for constant N), each n_{t+1} depends only on the immediately preceding n_t.

To understand the case $s \neq 0$, define

$$q \equiv \frac{p(1+s)}{1+ps}, \qquad \text{so that} \qquad 1-q = \frac{1-p}{1+ps} \tag{7}$$

Note that as p varies from 0 to 1, q also varies from 0 to 1, for any value of s. Now we can recognize that equation (5) is a binomial process even in the general case of nonzero s, namely

$$n_{t+1} \sim \text{Binomial}(2N, q_t) \tag{8}$$

where q_t relates back to n_t through equations (7) and (1).

There are two different ways that one can use equation (8) in numerical modeling of the Wright-Fisher model here discussed. One can use it in Monte Carlo fashion, advancing a single n through successive generations by drawing from a random binomial generator. Or, one can advance the entire probability distribution of n one generation at a time, by calculating the matrix elements of the Markov process implied by (8) and doing the matrix multiplication [6]. In this latter case, there is no random number generator, and the results are "exact" up to numerical roundoff and truncation errors. Various approximation schemes, not described here, for dealing with large values of N and large numbers of generations can also be implemented within a Markov matrix formulation. Diffusion theory [4] can be captured within such approximations; but a Markov matrix formulation can be arbitrarily more exact than diffusion theory, depending on the computational workload that one can afford.

4 Numerical Approximations to Exact Results

From numerical models like those described above, one can obtain detailed results for models across a wide range of N and s. It is useful to summarize some of these more-or-less exact results as "fitting formulas" that capture the asymptotic behavior in various regimes, while maintaining tolerable accuracy in complicated overlap regions. Note the distinction between obtaining exact analytic results for approximations of the original model, as in the case of the diffusion approximation [4], and obtaining approximate fitting formulas to the numerically implemented exact model, as here. By "tolerable accuracy" we usually mean on the order of 10–20%, although a factor of 2 might be tolerable in some cases. Some fitting formulas thus obtained here follow.

Probability of Fixation

A mutation is said to fix when it becomes universal in the population, that is, $p = 1$ and $n = 2N$. Our methods don't improve on the standard approximation,

$$P_{\text{fix}} \approx \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \tag{9}$$

This formula has several asymptotic cases: When the selection coefficient is positive and $O(1)$, fixation is highly likely,

$$P_{\text{fix}} \approx 1 - e^{-2s} \tag{10}$$

When s is of either sign but small in relation to the inverse of the population size, $|s| \ll 1/(4N)$, then Kimura's neutral result obtains,

$$P_{\text{fix}} \approx \frac{1}{2N} \tag{11}$$

which can be understood as the uniform probability that any one of the $2N$ alleles in the population will ultimately prevail in a neutral coalescent process.

When s is positive, $Ns \gg 1$ but $s \ll 1$, the fixation probability increases linearly with s,

$$P_{\text{fix}} \approx 2s = \frac{1}{2N}(4Ns) \tag{12}$$

So the fixation probability is larger, relative to the neutral case, by a factor $4Ns$.

When s is negative and $|s| \gtrsim 1/(4N)$, the fixation probability goes exponentially to zero as

$$P_{\text{fix}} \approx e^{-4N|s|} \tag{13}$$

Mean Time to Fixation

For alleles that do fix, we can ask what is their distribution of times to fixation, measured in generations. A simpler question is what is the mean of that distribution. A fitting formula that is tolerably correct in all regimes is

$$\langle t_{\rm fix}\rangle \approx \frac{4N}{1+\frac{3}{8}N|s|}\left(\frac{1+\frac{1}{2}(\ln N)|s|}{1+|s|}\right) \tag{14}$$

When $N|s| \ll 1$ (approximately neutral case), the mean fixation time is about $4N$. When $Ns \gg 1$ but $s \ll 1$ (regime of fixation probability linear in s), equation (14) becomes

$$\langle t_{\rm fix}\rangle \approx \frac{32}{3|s|} + \frac{16}{3}\ln N \tag{15}$$

so fixation is not only more likely, but also occurs more rapidly in proportion to s getting larger. When selection per generation is strong, $|s| \gtrsim 1$, we have

$$\langle t_{\rm fix}\rangle \approx \frac{16}{3}\frac{\ln N}{|s|} \tag{16}$$

That is, highly positive mutations take longer to fix in larger populations, albeit only logarithmically.

Mean Time to Loss

If an allele doesn't eventually fix, then it will eventually be lost from the population. The mean time to loss, measured in generations, is tolerably fit by

$$\langle t_{\rm lose}\rangle \approx 2\ln\left(\frac{N}{1+N|s|}\right)+2 \tag{17}$$

which, because of the logarithm, can never be very large.

Relation Between Mutation and Fixation Rates

Let μ be the mutation rate per individual per generation for the particular nucleotide and A,C,G,T outcome associated with allele x. For the neutral case, a population of N will generate a mean of $2N\mu$ such mutations per generation, each of which will fix with probability $1/(2N)$. Thus the mean number of fixations of x per generation $\mu_{\rm fix}$ is just μ, identical to the mutation rate at which they are generated, as Kimura famously pointed out. When this number is small (the usual case) it can be interpreted as the probability per generation of fixing the allele x.

For the case of nonzero s, the interesting case is when $N|s| \gg 1$, so that selection matters. When s is positive but small, $s \ll 1$, we multiply the population rate of mutation $2N\mu$ by the fixation probability $2s$ to get

$$\mu_{\rm fix} = 4Ns\,\mu \tag{18}$$

So the fixation rate in the population is larger than the mutation rate by a factor $4Ns$.

5 Smallest Observable Selections on Evolutionary Time Scales

Suppose that, because of an environmental change or an opportunity created by a previous mutation, a mutation becomes significantly favorable, $4Ns > 1$, at generation $t = 0$. How long does it take, on average, before it fixes in the population?

The answer is the sum of two terms. First there is the time until a mutation that is destined to fix appears in the population. Second, there is the time that it spends in the population fixing. Actually these are not statistically independent, but for illustrative purposes we pretend that they are.

The rate at which mutations fix is $(2N\mu) \times (2s) = 4Ns\mu$, so the mean waiting time for the first one after $t = 0$ is the reciprocal of this. By equation (15) the fixation time is, up to a small additive logarithm, 32/(3s). So the total is

$$T = \frac{1}{s}\left(\frac{1}{4N\mu} + \frac{32}{3}\right) \tag{19}$$

Thus, the number of generations required is at least $\sim 10/s$, and can be much larger if we are "mutation starved" with $\mu N \ll 1$.

An interesting way to look at equation (19) is as putting limits on the smallest mutations that can be observed on a known evolutionary time scale. In other words, solve equation (19) for s in terms of μ, N, and T (a number of generations over which we observe significant adaptation by accumulated mutations). Then smaller values of s don't have time to happen. That is,

$$s \gtrsim \frac{1}{T}\left(\frac{1}{4N\mu} + \frac{32}{3}\right) \tag{20}$$

Let's try putting in numbers for vertebrate evolution. Little is actually known about the value μ, but it is often estimated as 10^{-8} or 10^{-9}. [7, 8]

As a first example, consider Homo sapiens during the 10^6 or so years of his evolution, comprising 10^5 or so generations. Population studies lead to estimates of the effective population N on the order of 10^4. [9] So $4N\mu \ll 1$, and we get

$$s > \frac{1}{4NT\mu} \approx 10^{-1} \tag{21}$$

So, with these values, we see that humans must have evolved by mutations with huge ($> 10^{-1}$) individual selection coefficients. Indeed, the uncertainty in even the order of magnitude of the input quantities allows the possibility of $s \gtrsim 1$, which would effectively stop all incremental adaptation by natural selection, something that we know to be contrary to fact. This is suggestive of a mutation rate μ that is itself under mutational selection and kept large enough to allow gradual adaptation via large numbers of small mutations to proceed. If we imagine some necessary maximum value of s, say $s_{\min} = 10^{-2}$, then we can write equation (19) yet again (for human) as

$$\mu \sim \frac{1}{4NTs_{\min}} \sim 10^{-8} \tag{22}$$

By contrast, now consider some notional species of fish with $N = 10^{10}$ and 10^6 generations of available time to evolve. Now $4N\mu \gg 1$, so the bound on s is

$$s > \frac{32}{3T} \approx 10^{-5} \tag{23}$$

In this case, mutations of exquisitely small effect in terms of s are evolutionarily effective, as Darwin understood. Thus, humans and (notional) fish appear to evolve in very different regimes of selection coefficient s, with possibly important effects on the nature of their respective evolutions.

References

[1] Wright S (1931) Evolution in Mendelian populations, *Genetics* **16**, 97–159.

[2] Kimura M (1968) Evolutionary rate at the molecular level, *Nature* **217**, 624–626.

[3] Jobling MA, Hurles ME, Tyler-Smith C (2004) *Human Evolutionary Genetics* (New York: Garland) §5.3.

[4] Ewens WJ (2004) *Mathematical Population Genetics, I. Theoretical Introduction*, 2nd ed. (New York: Springer).

[5] Gillespie JH (2004) *Population Genetics: A Concise Guide*, 2nd ed. (Baltimore: Johns Hopkins), §2.7.

[6] Press WH, Teukolsky SA, Vetterling WT, and Flannery BP (2007) *Numerical Recipes: The Art of Scientific Computing*, 3rd Ed. (Cambridge, UK: Cambridge University Press), §16.3.

[7] Nachman MW, Crowell SL (2000) Estimate of the Mutation Rate per Nucleotide in Humans, *Genetics* **156**, 297–304.

[8] Haag-Liautard C, Dorris M, Maside X Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila, *Nature* **445**, 82–85.

[9] Jobling MA et al., op. cit., §6.2.