Notes 15 : Wright-Fisher model

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 1.2].

1 Wright-Fisher Model

In the Wright-Fisher model, we have N diploid individuals, that is, each individual has two copies of each chromosome. Generations are non-overlapping. At each generation, each chromosome inherits its genetic material from a uniformly chosen chromosome from the previous generation, independently from all other chromosomes.

In its most basic form, the Wright-Fisher model overlooks many important details:

1. Mutation
2. Recombination
3. Sexes
4. Non-overlapping generations
5. Population size changes
6. Family size distribution
7. Population structure
8. Selection

And many others. We will include the first two later. The next four points can be dealt with using the robustness of the coalescent (with some caveats) which we will discuss briefly in the next lecture. The remaining two issues are somewhat trickier. The Wright-Fisher model is still useful in studying them as it serves as a null hypothesis which can be rejected based on data to provide evidence for the inadequacy of the model.

2 Fixation of a neutral mutation

Consider a particular locus which has two alleles A and a (for instance, a gene with two variants). Denote by X_t the number of A's in the population at time t. Under the Wright-Fisher model, X_t changes randomly from generation to generation—a phenomenon known as *genetic drift*.

Note that $X_t = 0$ and $2N$ are *absorbing states* (in the absence of further mutations). It is natural to ask:

- What is the probability that a particular allele is fixated?
- How fast does fixation occur?

The *fixation time* is defined as

$$\tau = \min\{t \,:\, X_t = 0 \text{ or } 2N\},$$

and is a stopping time. Recall that a *stopping time* is a $\{0, 1, \ldots, +\infty\}$-valued random variable such that the event $\{\tau \leq t\}$ depends only $\{X_0, \ldots, X_t\}$.

THM 15.1 *We have*

$$\mathbb{P}[X_\tau = 2N \,|\, X_0 = i] = \frac{i}{2N}.$$

Proof: Recall that a *martingale* is an (adapted) stochastic process $\{Z_t\}_{t\geq 0}$ satisfying

$$\mathbb{E}[Z_{t+1} \,|\, Z_0, \ldots, Z_t] = Z_t,$$

for all t. We claim that X_t is a martingale, indeed, the distribution of X_{t+1} given X_t is binomial with parameters $2N$ and $X_t/2N$, hence

$$\mathbb{E}[X_{t+1} \,|\, X_0, \ldots, X_t] = 2N\frac{X_t}{2N} = X_t.$$

By the martingale property, $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ for all t. This implies, along with the bounded convergence theorem (since $|X_t| \leq 2N$ and $\tau < +\infty$ a.s.),

$$i = \mathbb{E}[X_t \,|\, X_0 = i] = \mathbb{E}[X_\tau; \tau \leq t \,|\, X_0 = i] + \mathbb{E}[X_t; \tau > t \,|\, X_0 = i] \to \mathbb{E}[X_\tau \,|\, X_0 = i],$$

as $t \to +\infty$.

Finally,

$$i = \mathbb{E}[X_\tau \,|\, X_0 = i] = 2N\mathbb{P}[X_\tau = 2N \,|\, X_0 = i].$$

■

The previous theorem has an interesting consequence. When a new mutation arises in a population, its original frequency is 1 and it fixates with probability $1/2N$. If the rate at which mutations arise in each individual at a particular locus is μ, then the total rate of mutation in the population is $2N\mu$. Multiplying the two quantities, the rate at which mutations arise and fixate is μ. This explains why, when we discussed sequence evolution models in phylogenetics, we failed to distinguish between rates of mutation and rates of substitution.

3 Rate of convergence

We now look at how fast fixation occurs. Let

$$H_t^0 = \frac{2X_t(2N - X_t)}{2N(2N-1)},$$

be the *heterozygosity* (without replacement), that is, the probability that two randomly chosen chromosomes have a different allele. The proof of the following easy result introduces key ideas: *turning back time* and *coalescence*.

THM 15.2 (Rate of convergence) *We have*

$$\mathbb{E}[H_t^0] = \left(1 - \frac{1}{2N}\right)^t \mathbb{E}[H_0^0].$$

Proof: Pick two individuals at random at time t. Tracing their lineages backwards in time, at time 0 either:

- The two lineages have *coalesced*, that is, the same parent was chosen by both lineages at a particular generation. In that case, the two chromosomes necessarily inherit the same state. This happens with probability

  $$1 - \left(1 - \frac{1}{2N}\right)^t,$$

  by independence.

- The two lineages *have not coalesced*. In that case, the probability of inheriting a different allele is H_0^0 by symmetry. This event happens with probability

  $$\left(1 - \frac{1}{2N}\right)^t.$$

■

Taking a limit when $N \to \infty$ and rescaling time by $2N$, we see that the probability that two random lineages have not coalesced by time t is

$$\left(1 - \frac{1}{2N}\right)^{2Nt} \to e^{-t}.$$

In other words, in the limit of an infinte population, the coalescence time is exponential with mean 1. Similarly, for k samples, the limit distribution of the first coalescence (ignoring double coalescences which have probability $O(1/N^2)$) is exponential with mean $\binom{k}{2}^{-1}$.

Further reading

The material in this section was taken from Section 1.2 of the excellent monograph [Dur08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.