Coalescents and Neutral Sampling

S. Sawyer — Vs. January 11, 2005 — ©

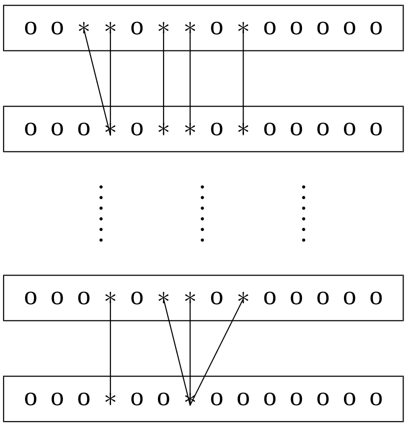
2.1. Introduction

Consider a population that is maintained at $2N$ genes for a large number of generations. Assume that each gene in each generation has a unique parental gene, the parental gene is equally likely to be any of the genes in the preceding generation, and the choices of parents for different genes are independent. Then two genes in the same generation will have the same parent with probability $1/2N$ and will have different parents with probability $1 - 1/2N$. (This is called the standard Wright-Fisher model.)

In general, a sample of r genes from the same generation has $r_1 \le r$ distinct parents. These r_1 parental genes have $r_2 \le r_1$ distinct parents in the generation that precedes them. In general the original sample has $r_n \le r$ distinct ancestors in the n^{th} generation before the starting generation, where $r_n \le r_{n-1} \le \ldots \le r_2 \le r_1 \le r$.

Figure 1.1 illustrates a sample of $r = 5$ genes along with its ancestors for $2N = 14$. The sample has $r_1 = 4$ distinct parents, and in an earlier generation has $r_m = 4$ and $r_{m+1} = 2$. If N is large, then $r_{n+1} = r_n$ most of the time, since the probability that any pair of genes has the same parent is $1/2N$, and $r_{n+1} = r_n = r$ with probability bounded by $(1 - 1/2N)^r$. However, as long as N is finite, $r_n < r$ for some n will eventually occur with $r_{n_1} < r_{n_1+1} = r$ for some finite n_1 ($n_1 = 1$ in Figure 1.1). Similarly, if $r_{n_1} > 1$, there exists $n_2 > n_1$ such that $r_{n_2} < r_{n_2+1} = r_{n_1}$. This process continues until $r_{n_k} = 1$ for some k. That is, until the sample of r genes has a unique most-recent common ancestor (or MRCA).

Figure 1.1.
Ancestors of a sample of $r = 5$ genes


o o * * o * * o * o o o o o
o o o * o * * o * o o o o o
o o o * o * * o * o o o o o
o o o * o o * o o o o o o o

The pattern of ancestors of a sample of r genes since their MRCA is called the *coalescent process* of the r genes. Note that this process coalesces going backwards in time, rather than as time progresses into the future.

2.2. Properties of the Coalescent

We will show below that, in the limit as $N \to \infty$, the coalescent process can be described statistically in the following way:

(i) Any individual pair of genes had its first common ancestor $n \approx 2Nt$ generations ago, where $t \approx n/2N$ is an exponentially distributed random variable with mean one.

(ii) For a sample if r genes, let t_r^{2N} be the number of generations until at least one pair from the sample has a common parent. Then, as $N \to \infty$, $t_r^{2N} \approx t_r$, where t_r is exponentially distributed with mean $E(t_r) = 2/r(r-1)$.

(iii) In the ancestral pedigree of the r genes, in the limit as $N \to \infty$, at most one pair of genes has a common parent in any one generation. Thus each coalescent step in the coalescent process has exactly one pair of genes with a common parental gene. The mean times between coalescent steps are independent exponentially distributed random variables t_k with means $E(t_k) = 2/(k(k-1))$, where k is the number of ancestors just before the branching event (or just after the branching event, if we looks forwards to the present instead of backwards into the past).

These properties imply that, in the limit as $2N \to \infty$, the time since the MRCA of r genes (in units of $2N$ generations) has expected value

$$E(T_r) = \sum_{j=2}^{r} E(t_j) = \sum_{j=2}^{r} \frac{2}{j(j-1)} = 2\sum_{j=2}^{r}\left(\frac{1}{j-1} - \frac{1}{j}\right) = 2\left(1 - \frac{1}{r}\right) \;\le\; 2 \qquad (2.1)$$

Thus the expected time since the common ancestor of any set of r genes is less than $4N$ generations, no matter how large r is, as long as r is fixed as $N \to \infty$.

Note that mutation was not mentioned in this description. Genes have the same allelic type as the parental gene unless there has been an intervening mutation. If there is no mutation, each gene in a sample of r genes has the same allelic type as the MRCA. Note that mutations that are selectively neutral do not affect the coalescent process. Thus selectively neutral mutation can be analyzed by assuming that it acts on the coalescent after it is formed. See below for an example of how this can be used to estimate an inbreeding coefficient in the present.

Before proving (i,ii,iii) above, let's first discuss some of the consequences of this particular probability model.

2.3. The Fate of a Parental Gene

Consider the Wright-Fisher model from the point of view of a parental gene. Each offspring gene in the next generation will have that gene as its parent with probability $1/2N$. It will choose another gene to be its parent with probability $1 - 1/2N$. Since these choices are independent, the number of offspring Y of a particular parental gene has a Bernoulli distribution based on $2N$ independent trials with probability of success $1/2N$ for each chose. That is,

$$P(Y = n) \;=\; \binom{2N}{n}\left(\frac{1}{2N}\right)^{n}\left(1 - \frac{1}{2N}\right)^{2N-n} \qquad (0 \le n \le 2N) \qquad (3.1)$$

where $\binom{2N}{n} = (2N)!/(n!(2N-n)!)$.

In general, a random variable B has a *Bernoulli distribution* with parameters M and p (abbreviated $B \approx B(M,p)$) if

$$P\big(B = n\big) \;=\; \binom{M}{n} p^n (1-p)^{M-n} \qquad (0 \le n \le M) \qquad (3.2)$$

Thus $Y \approx B(2N, 1/2N)$ by (3.1). The *Poisson Limit Theorem* for Bernoulli random variables $B \approx B(M,p)$ says that

$$\lim_{\substack{M\to\infty,\ p\to 0,\\ Mp\to c}} P\big(B = n\big) \;=\; e^{-c}\frac{c^n}{n!}, \qquad (c \ge 0,\ n \ge 0) \qquad (3.3)$$

In general, a random variable V whose distribution is given by the right-hand side of (3.3) is said to have a *Poisson distribution with mean c*, or $V \approx \mathrm{Poi}(c)$ for short.

The limit (3.3) is not difficult to obtain: We can rearrange terms in (3.2) to obtain

$$P(Y = n) \;=\; \frac{(Mp)^n}{n!}\left(\frac{M(M-1)\dots(M-n+1)}{M^n}\right)\left(1 - \frac{Mp}{M}\right)^{M}\left(1 - \frac{Mp}{M}\right)^{-n} \qquad (3.4)$$

Assume $Mp = c$ in (3.3) and let $M \to \infty$ for a fixed value of n. Since $\lim_{M\to\infty}\big(1 - (c/M)\big)^M = e^{-c}$, the limit as $M \to \infty$ of the expression in (3.4) is $\frac{c^n}{n!}e^{-c}$. The general case of (3.3) is similar. (***Exercise.*** Prove (3.3) from (3.2) for fixed n. Give all of the details.)

Since $Y \approx B(2N, 1/2N)$ in (3.1), the number of offspring genes of any parental gene for large N is approximately Poisson with mean one (that is, $\mathrm{Poi}(1)$). In particular $P(Y = 0) \approx e^{-1} = 0.36788$. Thus if N is large, about 37% of genes in any particular generation will have no survivng offspring in the next generation.

The distribution of the offspring of *r* parents: Let $Y_1, Y_2, \dots, Y_r$ be the numbers of descendents of r genes in the following generation, where Y_i is the number of descendents of the i^{th} parental gene. Each individual Y_i has the binomial distribution $B(2N, 1/(2N))$. However, the random variables Y_i are not independent: If the i^{th} parent has a huge number of offspring, then the numbers of offspring of the other parents must be less, since the total number of offspring is constratined to be $2N$.

Under the assumptions of Section 1, the random variables $(Y_1, \dots, Y_r)$ together have a *multinomial distribution*:

$$\begin{aligned} &P(Y_1 = n_1,\, Y_2 = n_2,\, \dots\,, Y_r = n_r) \\ &\quad = \binom{M}{n_1\; n_2\; \dots\; n_r\; M_r} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r} (1 - p_1 - \dots - p_r)^{M_r} \end{aligned} \tag{3.5}$$

where $M = 2N$, $p_i = 1/2N$ for $1 \le i \le r$, $M_r = M - n_1 - n_2 - \dots - n_r$, and

$$\binom{M}{n_1\; n_2\; \dots\; n_r\; M_r} = \frac{M!}{n_1!\; n_2!\; \dots\; n_r!\; M_r!} \tag{3.6}$$

Note the slight difference in notation between the *binomial coefficients* in (3.1) and the *multinomial coefficients* in (3.6): For example

$$\binom{7}{3} = \frac{7!}{3!\; 4!} = \frac{7(6)5}{1(2)3} = 35$$

and

$$\binom{7}{2\; 3\; 2} = \frac{7!}{2!\;\; 3!\;\; 2!} = \frac{7(6)(5)4}{1(2)1(2)} = 210$$

Thus the binomial coefficient $\binom{7}{3}$ is $\binom{7}{3\; 4}$ if we write it as in (3.6). The illustration above also suggests that multinomial coefficients tend to be larger than binomial coefficients.

Now suppose $M \to \infty$ and $p_i \to 0$ in (3.5) in such a way that $Mp_i \to c_i$ for $1 \le i \le r$. Then the limiting distribution of each individual Y_i has a Poisson distribution (3.3) with $c = c_i$. It also turns out that, in the limit, the Poisson random variables Y_i are *independent* Poisson random variables with parameters c_i. This is the content of the second exercise below.

The distribution of offspring with selection: Darwinian selection is modeled in the Wright-Fisher model by making the parental genes appear slightly larger (more probable) or smaller (less probable) in the sampling process that determines the next generation.

Specifically, assign *fitnesses* $w_i \ge 0$ $(1 \le i \le 2N)$ to the parental genes and assume (without loss of generality) that the r parental genes are genes corresponding to $w_1, w_2, \dots, w_r$. The offspring choose their parents independently as before, but now with probability $p_i = w_i/w$ of selecting the i^{th} parent for $w = \sum_{i=1}^{2N} w_i > 0$ instead of $p_i = 1/2N$. The joint distribution of the numbers of descendents of the first r parents is now the multinomial distribution

$$\begin{aligned} &P(Y_1 = n_1,\, Y_2 = n_2,\, \dots\,, Y_r = n_r) \qquad (3.7) \\ &= \binom{M}{n_1\; n_2\; \dots\; n_r\; M_r} \left(\frac{w_1}{w}\right)^{n_1} \left(\frac{w_2}{w}\right)^{n_2} \dots \left(\frac{w_r}{w}\right)^{n_r} \left(\frac{w_{r+1} + \dots + w_{2N}}{w}\right)^{M_r} \end{aligned}$$

If $w_1 = w_2 = \dots = w_{2N}$, then $w_i/w = p_i = 1/2N$, and (3.7) reduces to the “selectively neutral” case (3.5) with $p_i = 1/2N$.

Exercise 3.1. Prove that $E(X) = \sum_0^\infty nP(X = n) = c$ if $X \approx \mathrm{Poi}(c)$, and thus that $X \approx \mathrm{Poi}(c)$ really does have mean c.

Exercise 3.2. Show that, for large N, the numbers of offspring of r different parental genes are r *independent* Poisson-distributed random variables with mean one, in the absence of selection. (*Hint*: Use (3.5) with $M = 2N$ and simplify the factorials.)

Exercise 3.3. Show that the results of the last Exercise are true even with selection. Specifically, assume that the first r parental alleles have fitnesses $w_1, w_2, \ldots, w_r$ as in (3.7), and that the remaining genes in the parental generation have fitnesses $w_j = 1$ for $r + 1 \le j \le 2N$. Prove that, as $N \to \infty$, the limiting variables Y_i in (3.7) are independent Poisson random variables with means μ_i for some choice of constants $\mu_1, \ldots, \mu_r$.

What are the μ_i? Are they same as you expected from (3.7)? Why?

Remarks. The last two exercises shows that the number of descendents over time of an initial parental gene forms a *branching process* with offspring distribution $\mathrm{Poi}(\mu_1)$.

Recall that two random variables X and Y are independent if

$$\Pr(X = a, Y = b) \;=\; \Pr(X = a)\Pr(Y = b)$$

for all values of a and b, and r random variables $X_1, X_2, \ldots, X_r$ are independent if

$$\begin{aligned} &\Pr(X_1 = i_1, X_2 = i_2, \ldots, X_r = i_r) \\ &= \Pr(X_1 = i_1)\,\Pr(X_2 = i_2)\,\ldots\,\Pr(X_r = i_r) \end{aligned}$$

for all choices of integers $i_j \ge 0$.

2.4. Proofs of Coalescent Results

A key mathematical step in the derivation of the properties of the coalescent will be

Lemma 4.1. *For any constant c*

$$\lim_{\substack{N\to\infty \\ n/2N\to t}} \left(1 - \frac{c}{2N} + O\left(\frac{1}{N^2}\right)\right)^n \;=\; e^{-ct}, \qquad 0 \le t < \infty \tag{4.1}$$

The expression $O(1/N^2)$ in (4.1) is a example of a useful notation due to a mathematician named Landau. In Landau's notation, $O\big(f(N)\big)$ means any mathematical expression which can be written

$$A(N)f(N) \tag{4.2}$$

where $A(N)$ is bounded; i.e., such that there exists a constant Ω such that $|A(N)| \le \Omega$ for all $N \ge 1$. (If we were concerned about limits as $N \to 0$, this would be a bounded with $|A(n) \le \Omega$ for $N \le 1$.)

The implicit function $A(N)$ in $O\big(f(N)\big)$ need not be the same in successive uses of $O\big(f(N)\big)$. Thus

$$\begin{aligned} O(1/N) \;+\; O(1/N) \;&=\; O(1/N) \\ O(1/N) \;+\; O(1/N^2) \;&=\; O(1/N) \\ 3\,O(1/N) \;&=\; O(1/N) \\ (1 \;-\; 2/N)(1 \;+\; 5/N) \;&=\; 1 \;+\; 3/N \;+\; O(1/N^2) \end{aligned} \tag{4.3}$$

are all correct. In the first three relations a new implicit bounded function $A(N)$ for the "O" on the right-hand side of (4.3) can be found in terms of the implicit functions $A(N)$ on the left-hand side.

The last equation follows from the identity $(1-2x)(1+5x) = 1+3x-10x^2$ for $x = 1/N$, so that $A(N) = 10$ in this case.

When Landau's notation is used, the range of N for which $A(N)$ is bounded is understood ($N \ge 1$ in (4.1)). For example, by Taylor's theorem

$$\log(1-x) \;=\; -x + O(x^2) \tag{4.4}$$

for small enough x. The function $A(x) = O(x^2)/x^2$ in (4.4) is

$$A(x) = (\log(1-x)+x)/x^2$$

Then $\lim_{x\to 0} A(x) = 1/2$, so that $A(x)$ is bounded for $x \le 0.50$, but $A(x)$ is unbounded as $x \uparrow 1$. Thus the implicit $A(x)$ in (4.4) is bounded only for $|x| \le c < 1$, where the implicit bound depends on c.

Proof of Lemma 4.1. Take the logarithm of the left-hand side of (4.1) and apply (4.4). If $x = c/2N + O(1/N^2)$ in (4.1), then $x^2 = (c/2N+O(1/N^2))^2 = (c/2N)^2(1+O(1/N))^2 = O(1/N^2)$ as in (4.2). Hence in (4.4) $-x+O(x^2) = -c/2N + O(1/N^2) + O(1/N^2) = -c/2N + O(1/N^2)$, where N is assumed to be sufficiently large so that $|x| \le c < 1$. The logarithm of the left-hand side of (4.1) then equals

$$\begin{aligned} & n\ \log\left(1 - \frac{c}{2N} + O\left(\frac{1}{N^2}\right)\right) \qquad (4.5)\\ &= n\left(-\frac{c}{2N} + O\left(\frac{1}{N^2}\right)\right) \;=\; -c\frac{n}{2N} + O\left(\frac{n}{N}\frac{1}{N}\right)\\ &= \frac{n}{2N}\left(-c + O\left(\frac{1}{N}\right)\right) \;\to\; -ct \qquad \text{as } n/2N \to t \end{aligned}$$

by (4.4). Taking exponentials in (4.5) implies (4.1).

Our first asymptotic result for coalescents is as follows. As before, given a sample of r genes, let t_r^{2N} be the number of generations into the past until at least one pair of genes has a common parent, or equivalently until the ancestors of the r genes have fewer than r parents. Thus $t_r^{2N} > 1$ if and only if the initial r genes have r distinct parents. The asymptotic distribution of t_r^{2N} for large N is given by

Lemma 4.2. *If t_r^{2N} is as above, then for all $t \ge 0$*

$$\lim_{N\to\infty} \Pr\left(\frac{t_r^{2N}}{2N} > t\right) \;=\; e^{-\frac{r(r-1)}{2}t} \tag{4.6}$$

Proof. By definition

$$\begin{aligned} \Pr(t_r^{2N} > n) \;&=\; P(r \text{ genes have } r \text{ distinct parents for}\\ &\qquad\qquad \text{at least the past } n \text{ generations}) \qquad (4.7)\\ &=\; \Pr(t_r^{2N} > 1)^n \end{aligned}$$

since the consecutive sampling events that determine the parents are independent over n generations. Let the r genes in the present generation be labeled as $\#1, \#2, \ldots, \#r$. Then

$$\begin{aligned} \Pr(t_r^{2N} > 1) \;=\; \Pr(\ &\#1 \text{ chooses parent } R_1\\ &\#2 \quad \text{''} \quad \text{''} \quad R_2 \ne R_1\\ &\#3 \quad \text{''} \quad \text{''} \quad R_3 \ne R_1, R_2\\ &\cdots \qquad\qquad \cdots \quad \cdots\\ &\#r \quad \text{''} \quad \text{''} \quad R_r \ne R_1, \ldots, R_{r-1})\\ \;=\; &\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{r-1}{2N}\right) \end{aligned}$$

since the choices of parents are independent and (for example) gene #r's choice of parent must avoid the parents of genes #1 through #$r-1$. The product of the $r-1$ factors above can be expanded into a sum of 2^{r-1} terms, which can be written

$$\begin{aligned}\Pr(t_r^{2N} > 1) \;&=\; 1 - \left(\sum_{i=1}^{r-1} i\right)\frac{1}{2N} \;+\; \text{terms with factors } \frac{1}{N^2},\ \frac{1}{N^3},\dots \\ &=\; 1 - \frac{r(r-1)}{2}\frac{1}{2N} \;+\; O\left(\frac{1}{N^2}\right) \qquad (4.8)\end{aligned}$$

where the implicit function $A(N)$ in $O(1/N^2)$ depends on r. This is because the sum of the last $2^{r-1}-r$ terms in (4.8) is $1/N^2$ times a sum which is bounded for $N \ge 1$ by a constant that depends on r. Hence by (4.7), if $t = t(n) \sim n/(2N)$,

$$\Pr\left(\frac{t_r^{2N}}{2N} > t\right) \;\approx\; \Pr(t_r^{2N} > n) \;=\; \left(1 - \frac{r(r-1)}{2}\frac{1}{2N} \;+\; O\left(\frac{1}{N^2}\right)\right)^n$$

and by Lemma 4.1

$$\lim_{N\to\infty} \Pr\left(\frac{t_r^{2N}}{2N} > t\right) \;=\; e^{-\frac{r(r-1)}{2}t} \qquad (4.9)$$

which completes the proof of Lemma 4.2.

Remarks. (1) A random variable X has an *exponential distribution with mean* λ (abbreviated $X \approx \mathrm{Exp}(\lambda)$) if $\lambda > 0$ and

$$P(X \ge t) \;=\; e^{-t/\lambda} \qquad \text{for} \quad 0 \le t < \infty \qquad (4.10)$$

If $X \approx \mathrm{Exp}(\lambda)$, then $P(X \ge t) = e^{-t/\lambda} = P(X_1 \ge t/\lambda)$ where $X_1 \approx \mathrm{Exp}(1)$, since $\Pr(X_1 \ge t) = e^{-t}$. Thus $X \approx \lambda X_1$ where $X_1 \approx \mathrm{Exp}(1)$ since $P(X \ge t) \equiv P(\lambda X_1 \ge t) = \exp(-t/\lambda)$ for $t \ge 0$. Equation (4.9) says that $t_r^{2N}/2N$ has a limiting exponential distribution with mean $\lambda = 2/r(r-1)$, or, equivalently,

$$\frac{t_r^{2N}}{2N} \;\approx\; \frac{2}{r(r-1)}X_1 \qquad \text{for large } N,\ \text{where } P(X_1 \ge t) \equiv e^{-t} \qquad (4.11)$$

In (4.11), "$\approx$" is used in both the senses of "approximately" (for large N) and "has the same probability distribution as".

Exercise 4.1. If $P(X > t) = \int_t^\infty f(x)\,dx$ is differentiable, then the *expected value* of X is

$$E(X) = \int_0^\infty xf(x)\,dx = \int_0^\infty x\left(-\frac{d}{dx}P(X > x)\right)\,dx$$

Prove that $E(Y) = \lambda$ if $Y \approx \mathrm{Exp}(\lambda)$, so that the random variable $Y \approx \mathrm{Exp}(\lambda)$ really does have mean λ.

Remarks. (2) The limit (4.9) can also be obtained from the Poisson Limit Theorem (3.3). Consider the process of Lemma 4.2 going backwards in time, where a coalescent event (i.e., some pair of ancestral genes has a common parent) is considered a "success". Until the first "success" occurs, the $r' = r$ parental genes in the previous generation are distinct genes chosen at random from that generation, so that the probability of success remains the same. Then

$$\Pr(t_r^{2N} > n) \;=\; P(0 \text{ “successes” in } n \text{ generations}) \tag{4.12}$$

for a Bernoulli trial with $M = n$ trials and probability of failure q given by (4.8) (so that the probability of success is $p = 1 - q$). As $N \to \infty$ and $n/2N \to t$

$$\begin{aligned} Mp \;=\; np \;&=\; \frac{r(r-1)}{2}\frac{n}{2N} \;+\; O\Big(\frac{n}{N^2}\Big) \\ &=\; \frac{n}{2N}\left(\frac{r(r-1)}{2} + O\left(\frac{1}{N}\right)\right) \;\to\; c = \frac{r(r-1)}{2}t \end{aligned}$$

and by (3.3) and (4.12)

$$\lim_{\substack{N\to\infty \\ n/2N\to t}} \Pr(t_r^{2N} > n) \;=\; P(\mathrm{Poi}(c) = 0) \;=\; e^{-c}\frac{c^0}{0!} = e^{-c} = e^{-\frac{r(r-1)}{2}t}$$

which is (4.9). The connection between (3.3) and Lemma 4.1 should not be too surprising, since the most direct proofs of the Poisson Limit Theorem (3.3) depend on something like Lemma 4.1.

Corollary 4.2. *Label a sample of r genes as $\#1, \#2, \ldots, \#r$, and let m_{ij} be the number of generations into the past until the pair $\#i$ and $\#j$ have a common ancestor. Then for each i, j with $1 \le i < j \le r$,*

$$\lim_{N\to\infty} \Pr\left(\frac{m_{ij}}{2N} > t\right) \;=\; e^{-t}, \qquad 0 \le t < \infty \tag{4.13}$$

Proof. Apply Lemma 4.2 with $r = 2$.

2.5. “Special Events” and the Regularity of the Coalescent

Assume that a sample of r genes in the present has r_n distinct ancestors n generations in the past. A “special (coalescent) event” occurs in generation n if $r_{n+1} \le r_n - 2$; i.e., if the r_n genes in that generation have $r_n - 2$ distinct parents or fewer. This can only happen if two or more distinct pairs of genes each have a single parent, or else if three or more genes have the same parent.

The next result says that these “special events” do not occur at all in the time scale $t \approx n/2N$ for large N. That is, in this time scale, all coalescent events are due to a unique pair in that that generation having the same parent, with the remaining individuals in that generation having other distinct parents. Thus, except for an event of small probability, either $r_{n+1} = r_n$ or $r_{n+1} = r_n - 1$ for all $n \le C(2N)$.

To show this, let $\widetilde{n}_{\mathrm{spec}}$ be the number of generations into the past until the first special event. Then

Lemma 5.1. *For $\widetilde{n}_{\mathrm{spec}}$ as above and all $t \ge 0$,*

$$\lim_{N\to\infty} \Pr\left(\frac{\widetilde{n}_{\mathrm{spec}}}{2N} > t\right) \;=\; 1$$

Proof. Suppose that a sample of s genes are labeled $\#1, \#2, \ldots, \#s$. Then

$$P_s(\widetilde{n}_{\text{spec}} = 1) \;=\; \sum_{i=2}^{s} \sum_{j=i+1}^{s} P_s\Big(\begin{array}{l}\text{Genes } \#1,\ldots,\#i{-}1 \text{ have distinct parents, } \#i \\ \text{has the same parent as one of } \#1,\ldots,\#i{-}1, \\ \#i{+}1,\ldots,\#j-1 \text{ have new distinct parents,} \\ \#j \text{ has the same parent as one of } \#1,\ldots,\#j-1, \\ \text{no conditions on the parents of } \#j{+}1,\ldots,\#s\text{)}\end{array}$$

$$= \sum_{i=2}^{s} \sum_{j=i+1}^{s} \left(1-\frac{1}{2N}\right)\cdots\left(1-\frac{i-1}{2N}\right)\frac{i-1}{2N}\left(1-\frac{i}{2N}\right)\cdots\left(1-\frac{j-2}{2N}\right)\frac{j-2}{2N}$$

Hence

$$P_s(\widetilde{n}_{\text{spec}} = 1) \;\le\; \frac{s^4}{(2N)^2} \tag{5.1}$$

since the first sum above has $\le s^2$ terms, each of which is bounded by $s^2/(2N)^2$. In general $\Pr(\widetilde{n}_{\text{spec}} > n) \le \Pr(\widetilde{n}_{\text{spec}} > 1)^n$, since the number of distinct ancestors of the r present genes decreases in prior generations. Similarly

$$\begin{aligned}\Pr(\widetilde{n}_{\text{spec}} > n) \;&\ge\; \left(\min_{1\le s\le r} P_s(\widetilde{n}_{\text{spec}} > 1)\right)^n \\ &=\; \left(1-\max_{1\le s\le r} P_s(\widetilde{n}_{\text{spec}} = 1)\right)^n \\ &\ge\; \left(1-\frac{r^4}{(2N)^2}\right)^n\end{aligned}$$

by (5.1). Hence

$$\Pr\left(\frac{\widetilde{n}_{\text{spec}}}{2N} > t\right) \;=\; \Pr(\widetilde{n}_{\text{spec}} > n) \;\ge\; \left(1-\frac{r^4}{(2N)^2}\right)^n \tag{5.2}$$

where $n = [2Nt]$ is the greatest integer $n \le [2Nt]$ as before. Thus

$$\lim_{N\to\infty} \Pr\left(\frac{\widetilde{n}_{\text{spec}}}{2N} > t\right) \;=\; 1, \qquad 0 \le t < \infty$$

by Lemma 4.1 with $c = 0$. This completes the proof of Lemma 5.1.

2.6. The Coalescent

The results in the previous sections have a number of interesting consequences.

Assume we have r genes in the present generation. Let n_j be the total number of past generations in which the r genes have exactly j distinct ancestors ($2 \le j \le r$). Note that there is only one ancestor in all generations before the most recent common ancestor (MRCA) of the sample of r genes, and always $j \ge 2$ ancestors in any generation strictly between the MRCA and the present. Thus, if the initial generation is counted in n_r, the number of generations G_r since the most recent common ancestor can be written

$$G_r = n_r + n_{r-1} + n_{r-2} + \cdots + n_2 \tag{6.1}$$

Ignoring "special events", the random variables n_j are independent and have geometric distributions by (4.7). Similarly, by Lemma 4.2, $n_j/2N \approx t_j$ for large N, where $t_j \approx \mathrm{Exp}(2/(j(j-1)))$ is exponentially distributed with mean $2/j(j-1)$. Thus as $N \to \infty$ the limiting distribution

$$\begin{aligned} \frac{G_r}{2N} &\approx T_r = t_r + t_{r-1} + \cdots + t_2 \qquad \text{where} \\ t_j &\approx \mathrm{Exp}\left(\frac{2}{j(j-1)}\right) \approx \frac{2}{j(j-1)} X_j, \qquad X_j \approx \mathrm{Exp}(1) \end{aligned} \tag{6.2}$$

Lemma 4.2 implies $n_j/2N \approx t_j$ in the sense that $P(n_j/2N \ge t) \to P(t_j \ge t)$ for all $t \ge 0$. It follows from the same proof that $E(n_j/2N) \to E(t_j)$ for $2 \le j \le r$ as well, and $E(G_r/2N) \to E(T_r)$ where

$$\begin{aligned} E(T_r) &= \sum_{j=2}^{r} E(t_j) = \sum_{j=2}^{r} \frac{2}{j(j-1)} E(X_j), \qquad X_j \approx \mathrm{Exp}(1) \\ &= 2\sum_{j=2}^{r}\left(\frac{1}{j-1} - \frac{1}{j}\right) = 2\left(1 - \frac{1}{r}\right) \le 2 \end{aligned}$$

since the sum telescopes and $E(X_j) = 1$ if $X_j \approx \mathrm{Exp}(1)$. Thus $E(T_r) \le 2$, and the expected number of generations since the most recent common ancestor of a sample of size r is $E(G_r) \approx E(2NT_r) \le 4N$ for large N, no matter how large r is.

Exercise 6.1. Assume in the Wright-Fisher model that each gene mutates in each generation with some probability u. Each mutant gene is entirely new to the population and is selectively equivalent to all preexisting genes, so that mutation does not affect the coalescent.

Let $H(2N, u)$ be the probability that the two genes in a randomly chosen individual are different. (This is called the *probability of heterozygosity*, and is the same as the probability that any two genes in the population are different. Similarly, $I(2N, u) = 1 - H(2N, u)$ is the *probability of homozygosity*.)

Prove that

$$I(2N, u) = 1 - H(2N, u) = \sum_{n=1}^{\infty}\left(1 - \frac{1}{2N}\right)^{n-1} \frac{1}{2N}\,(1-u)^{2n} \tag{6.3}$$

Use this identity to prove

$$\lim_{\substack{N\to\infty,\ u\to 0,\\ 4Nu\to\theta}} H(2N, u) = \frac{\theta}{1+\theta} \tag{6.4}$$

As a check, note that this is equivalent to a special case ($r = 2$) of Theorem 7.1 in Section 7 below, which will be derived by a different argument.

(*Hint*: Consider the coalescent pedigree that connects the two genes to their MRCA. Then $I(2N, u)$ is the probability that no mutations have occurred in any of the links of this pedigree. Sum over the number of generations since the MRCA.)

2.7. An Inbreeding Coefficient for r Genes

Asume that we are following a genetic locus in which each gene undergoes mutation with probabily u per gene per generation. Each mutant gene is of a type that is new to the population. Mutant genes are selectively equivalent to the original genes.

After the population has reached equilibrium in time, let $I(r, u, 2N)$ be the probability that a sample of r genes is homogeneous in the sense that there is only one allelic type represented in the sample. We can use the properties of the coalescent to calculate the limit of $I(r, u, 2N)$ as $N \to \infty$, $u \to 0$, and $4Nu \to \theta$.

Theorem 7.1. *The probability $I(r,u,2N)$ that r randomly-chosen genes are identical satisfies*

$$\lim_{\substack{N\to\infty,\ u\to 0,\\ 4Nu\to\theta}} I(r,u,2N) \;=\; I(r,\theta) \;=\; \frac{(r-1)!}{(\theta+1)(\theta+2)\cdots(\theta+r-1)} \tag{7.1}$$

Proof. In any generation n with $r_n = j$ distinct ancestors of the r genes, the coalescent pedigree has j pedigree links from the preceding generation (that is, the generation closer to the MRCA). Let n_j be the total number of generations after the MRCA in which $r_n = j$ as in (6.1). The random coalescent pedigree then has a total of $L = \sum_{j=2}^{r} jn_j$ links since the common ancestor.

Mutation occurs independently on these L links with probability u per generation per pedigree link, with each new allelic type radiating upwards through the pedigree from the link to the present or until encountering a later mutation. Since every mutant allele is new, the initial sample of size r will all be of the same type if and only if there has been no mutations in any of the $\sum_{j=2}^{r} jn_j$ links. For fixed u and $2N$, this has probability

$$I(r,u,2N) \;=\; E\left((1-u)^{\sum_2^r jn_j}\right) \;=\; \prod_{j=2}^{n} E\left((1-u)^{jn_j}\right) \tag{7.2}$$

since the n_j are independent random variables. Suppose $4Nu \to \theta \ge 0$ as $N \to \infty$. Then by (4.4)

$$\begin{aligned}\log(1-u)^{jn_j} \;&=\; jn_j(-u + O(u^2))\\ &=\; -2Nu\frac{jn_j}{2N}\left(1 + O\left(\frac{4Nu}{N}\right)\right) \;\approx\; -\frac{1}{2}\,\theta j\,t_j\end{aligned}$$

where $n_j/2N \approx t_j \approx \mathrm{Exp}\big(2/(j(j-1))\big)$ as in (6.2).

Thus $E\left((1-u)^{jn_j}\right) \to E\left(e^{-(1/2)\theta j t_j}\right)$ and

$$\lim_{\substack{N\to\infty,\ u\to 0,\\ 4Nu\to\theta}} I(r,u,2N) \;=\; \prod_{j=2}^{r} E\left(e^{-\frac{1}{2}\theta j t_j}\right), \qquad t_j \approx \frac{2}{j(j-1)}X_j \tag{7.3}$$

for $X_j \approx \mathrm{Exp}(1)$. In particular, the limit $I(r,\theta) = \lim I(r,u,2N)$ exists. Since $\frac{1}{2}\theta j t_j \approx \theta X_j/(j-1)$ where X_j are independent $\mathrm{Exp}(1)$,

$$\begin{aligned}I(r,\theta) \;&=\; \prod_{j=2}^{r} E\left(e^{-\theta X_j/(j-1)}\right) \;=\; \prod_{j=1}^{r-1} E\left(e^{-\theta X_j/j}\right), \qquad X_j \approx \mathrm{Exp}(1)\\ &=\; \prod_{j=1}^{r-1}\int_0^\infty e^{-\theta t/j}e^{-t}\,dt \;=\; \prod_{j=1}^{r-1}\frac{1}{1+\theta/j} \;=\; \prod_{j=1}^{r-1}\frac{j}{j+\theta}\\ &=\; \frac{(r-1)!}{(\theta+1)(\theta+2)\cdots(\theta+r-1)}\end{aligned} \tag{7.4}$$

2.8. The Ewens Sampling Formula

In Chapter 2, we derive the probability of *any* configuration of r genes into allelic types, specifically the probability that the r genes are composed of k different mutant types with n_i genes of the i^{th}

type ($1 \leq i \leq k,\ \sum_{i=1}^{k} n_j = r$). Specifically, under the same conditions as in Section 4 (i.e., (7.3)), this probability is

$$\frac{\theta^k\, r!}{L_r(\theta)\, n_1 \ldots n_k\, \beta(1)! \ldots \beta(r)!} \tag{8.1}$$

Here $L_r(\theta) = \theta(\theta+1)\ldots(\theta+r-1)$ is the r^{th} "ascending factorial power" of θ and (for $1 \leq j \leq r$) $\beta(j)$ is the number of alleles i such that $n_i = j$. Thus

$$\sum_{j=1}^{r} \beta(j) = k \qquad \text{and} \qquad \sum_{j=1}^{r} j\beta(j) = \sum_{i=1}^{k} n_k = r$$

This formula is called the *Ewens Sampling Formula (ESF)*. Note that (7.4) is a special case. It was conjectured by Ewens (1972) and proven rigorously by Karlin and MacGregor in the same year (1972).

The ESF has been used to prove that observed variants in particular genes must have selective significance, since the observed configuration of genes is inconsistent with the ESF. As any example, Singh, Lewontin, and Felton (1976) used a method called electrophoresis to examine 146 genes at a locus controlling a particular enzyme in the fruit fly *Drosophila pseudoobscura*. They found the configuration

$$68^1\ 11^1\ 8^1\ 6^2\ 5^2\ 3^7\ 2^3\ 1^{10} \tag{8.2}$$

The notation in (8.2) means 68 genes of one allelic type, 11 genes of another allelic type, 8 genes of a third type, two alleles with 6 genes each, . . . , and ten alleles each represented by exactly one gene in the sample.

Note that the allelic class with 68 genes seems unusually large in comparison with the diversity of the other 26 alleles. The problem is to decide whether the configuration (8.2) is consistent with the coalescent with selectively neutral muation, or equivalently with the ESF. This question was answered in the negative by Watterson (1978a), who derived a statistical test to conclude that (8.2) was highly inconsistent with neutrality at equilibrium. See Chapter 2 for more discussion of this example. There has also been some criticism of the use of electrophoresis for this problem.

For a long time, it has been an open problem whether most observed genetic polymorphisms are due to selective effects or, instead, are the result of a balance between genetic drift and selectively neutral mutation. One piece of evidence is the fact that many different polymorphic loci in the fruit fly, *Drosophila pseudoobscura*, appear to have the same alleles in roughly the same proportions in populations ranging from central California to Colombia (Dobzhansky and Powell 1975, p551; Prakash, Lewontin, and Hubby 1969). Under the neutral theory, this could happen only if mutant alleles for each locus were carried over this entire range before becoming extinct by mutation and genetic drift. The probability of this happening at several loci appears to be vanishingly small, but depends critically on the variance of the migration per generation for *Drosophila* (Sawyer 1976, 1977; Sawyer and Fleischman 1979).

On the other hand, a genetic study has been made of the Alaskan pink salmon (*Oncorhynchus gorbuscha*), which spawns in North American rivers and return from the ocean in exactly two years. Thus salmon populations in even and odd years should be genetically isolated, and in fact show genetic divergence at three enzyme loci (Aspinwall 1974). It is hard to see how this could be caused by selection. This could be viewed as evidence of selective neutrality and genetic drift, but might conceivably be the result of some extreme environmental effect in a particular even or odd year.