CHAPTER 13
Probabilistic Models

Chapter Goals:
- To give examples of dynamical models involving chance events
- To discuss how to incorporate chance events into simulations

Chapter Concepts:
- Demographic stochasticity
- Environmental stochasticity
- Birth-death process
- Wright-Fisher model
- Random genetic drift
- Individual-based model
- Moran model
- Coalescent theory

13.1 Introduction

All of the models considered so far have been deterministic; that is, the models predict that the system will be at one specific state at any given time. If a deterministic model forecasts that $n(t) = 50$, then the implication is that there will be exactly 50 individuals at time t. The real world is never so certain. Individuals may fail to reproduce or produce a bonanza crop of offspring simply by chance. Even if 50 is the most likely number of individuals, we might find 49 or 51 individuals, and there might even be some chance that the population is extinct or that it numbers in the millions. To account for such uncertainty, we must broaden our models. We need models that describe the realm of possible states; such models are known as "stochastic" or "probabilistic" models.

Definition 13.1: Stochastic Model
A model describing how the probability of a system being in different states changes over time.

Before embarking on the material in this chapter, first familiarize yourself with the principles of probability theory introduced in Primer 3. The core of this chapter focuses on developing stochastic models and simulating them, much as we did in Chapter 4 for deterministic models. Then, in Chapters 14 and 15, we introduce various methods that can be used to analyze stochastic models.

This chapter frequently relies on drawing random numbers from a probability distribution. Although computers cannot generate truly random numbers (everything they do is specified deterministically by computer code), there are many programs that generate "pseudo-random" numbers (see Press 2002). Pseudo-random numbers are determined by an algorithm in such a way that it is difficult to detect a pattern between successive numbers. For example, it is difficult to detect a pattern in the series: 1, 5, 9, 2, 6, 5, 3, 5, 8, 9, 7, 9, 3, 2, 3, . . . , but in fact these numbers are the digits in π (3.14159265358979323 . . .) that follow 3.14. Thus, we could use an algorithm that calculates π to get a series of pseudo-random integers between zero and nine. More sophisticated algorithms are described in Press (2002), which also discusses how random numbers can be drawn from different probability distributions (e.g., Poisson, binomial, normal,

etc.). In this chapter, we use the random number generators of *Mathematica* to simulate stochastic models, and we provide the code for generating each figure in the on-line supplementary material.

In the next four sections, we introduce the most fundamental stochastic models in ecology and evolution. Sections 13.2 and 13.3 describe stochastic models of population growth in discrete and continuous time, respectively. Similarly, sections 13.4 and 13.5 describe stochastic models of allele frequency change. To give a flavor for the breadth of stochastic models, we then explore three other models. Section 13.6 develops a stochastic model of cancer to illustrate how new models are explored. Section 13.7 introduces the concept of a spatially explicit stochastic model, which tracks the number and location of individuals within a population. Finally, section 13.8 is slightly more advanced and describes a relatively new and important branch of evolutionary theory, known as *coalescent theory*, which traces the ancestry of a sample.

13.2 Models of Population Growth

We begin by developing a stochastic model of population growth. A general deterministic model of population growth in discrete time is $n(t+1) = R\,n(t)$, where R might be a constant as in the exponential model or a function of the current density as in the logistic model. The equivalent stochastic model describes the *probability* of observing $n(t+1)$ individuals at time $t+1$, given that there are $n(t)$ individuals at time t. Again, a stochastic model might consider the passage of time to occur in discrete time steps or continuously.

Consider a species that reproduces once per season, at which point all of the parents die (i.e., nonoverlapping generations). To determine the number of individuals in the next generation, we must know the probability distribution describing the number of offspring per parent (Figure 13.1). That is, we must specify the probability that each reproducing parent is replaced in the next time unit by 0, 1, 2, etc. offspring. For a species with separate sexes, this exponential model counts only females and assumes that there are always enough males to fertilize these females. For a species that is hermaphroditic, each individual within the population is considered to be a reproducing parent. The distribution of offspring number will vary from species to species, but a simple (albeit arbitrary) choice is that the number of surviving offspring per parent follows a Poisson distribution (Figure P3.6). A Poisson distribution has only one parameter, μ, which gives both its mean and its variance. If the population size were initially $n(t) = 10$ and the mean number of offspring per parent, R, were 1.2, then there would be $n(t+1) = 1.2 \times 10 = 12$ offspring in a deterministic model. Even though 1.2 is the expected number of offspring per parent, however, any one parent will have a random number of offspring, which we draw from a Poisson distribution with mean $R = 1.2$. For example, the number of offspring per parent for each of the ten parents might be

4, 2, 0, 4, 1, 1, 1, 0, 0, 0

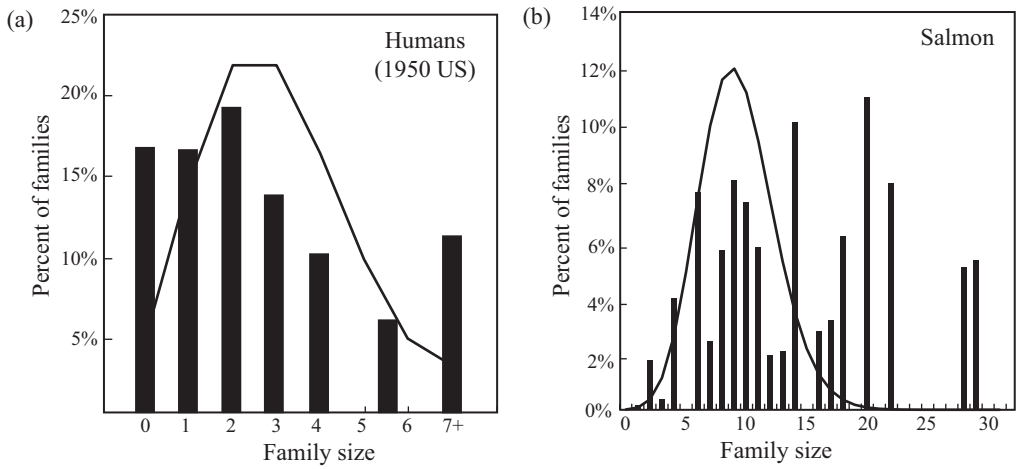

Figure 13.1: Family size distributions. (a) Distribution of family sizes for humans based on Kojima and Kelleher (1962). (b) Distribution of the number of offspring that survive and return to spawn per female in pink salmon, based on Figure 3b in Geiger *et al.* (1997). For these species, family size is more variable than predicted by the Poisson distribution (solid curves) with the same mean as the empirical distribution (histograms).

for a total of 13 surviving offspring. (We used *Mathematica* to draw these random numbers from a Poisson distribution.)

In this example, we expected the population to increase in size (from 10 to 12), but it actually increased even more (to 13). By chance, two of the parents left a surprisingly large number of offspring (four). Retracing our steps and drawing another random set of ten numbers from a Poisson distribution with mean $R = 1.2$ gives an entirely different outcome:

$$0, 1, 0, 1, 1, 1, 1, 3, 0, 1$$

for a total of 9 surviving offspring. In this case, the population size decreased.

To simulate population growth using a stochastic model, we could use random numbers to specify the number of offspring per parent in each subsequent generation. Given $n(t)$ parents at time t, the numbers of offspring per parent could be randomly drawn and the total set to $n(t + 1)$. Repeating the process to determine how many offspring are born to each of these parents would give us $n(t + 2)$. We could repeat this procedure for as many generations as desired. The simulation, however, would get slower and slower as the population size increased, because we must draw $n(t)$ random numbers, each one specifying the number of offspring per parent.

Fortunately, knowledge of probability theory can help us. We only care about the total number of offspring, and therefore we need only draw a single random number from a distribution that represents the sum of $n(t)$ draws from a Poisson distribution with mean R. The sum of $n(t)$ numbers drawn from a Poisson distribution with mean R is known to follow a Poisson distribution with mean $\mu = R\, n(t)$ (Supplementary Material P3.2). Thus, we can simulate a population in which $R = 1.2$ and $n(0) = 10$ by drawing a single random number from a Poisson with mean $\mu = 1.2 \times 10 = 12$. Using *Mathematica*, we obtained a random number of offspring equal to $n(1) = 21$. To get $n(2)$, we then

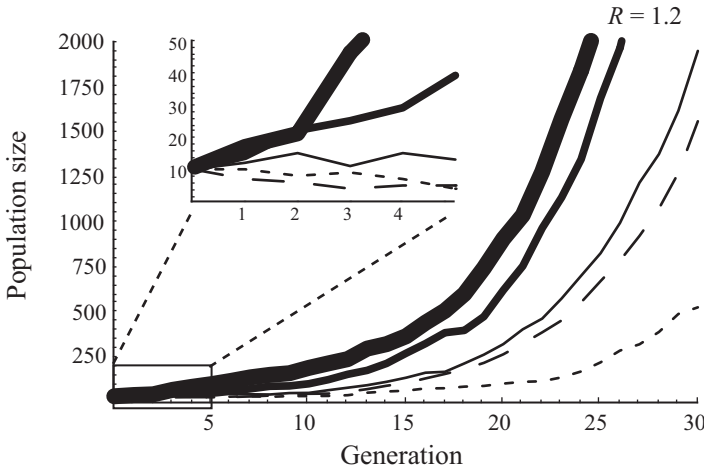

Figure 13.2: Stochastic model of exponential growth. Starting from a population of ten individuals, the number of individuals in generation $t + 1$ was drawn from a Poisson distribution with mean, $R\ n(t)$, where $R = 1.2$, until 30 generations had passed (first five generations are shown in the inset figure). This process was repeated five times (five curves).

drew a random number from a Poisson with mean $1.2 \times 21 = 25.2$, generating 26 offspring, by chance. In Figure 13.2, we show the resulting trajectory of population growth over 30 generations. Finally, we started the whole process over again from $n(0) = 10$ to generate the different curves (replicates).

The different curves in Figure 13.2 look as if they were drawn using different reproductive ratios R, but they weren't. In each case, $R = 1.2$. In the case of the top curve, the parents just happened to have more offspring early on in the simulation than in the case of the bottom curve. As in many stochastic models, there is a lot of variability in the outcome. Consequently, it is important to run several replicates of a stochastic simulation, starting with the same initial conditions and parameters, but drawing new random numbers each time step. We can then summarize the outcomes to draw conclusions. For example, we ran 100 replicate simulations with $n(0) = 10$ and $R = 1.2$. On average, 2470 offspring were alive after 30 generations. The standard deviation was 1739 offspring, indicating that the replicates varied substantially from one another. Indeed, the population had gone extinct in 3 of the 100 replicates. This variability in outcome is referred to as *demographic stochasticity*.

Variability in population size caused by chance differences in the number of surviving offspring per parent is known as *demographic stochasticity*.

The above simulations modeled exponential growth, where the mean number of offspring per parent, R, was the same regardless of population size. It is easy to incorporate density dependence by specifying how the mean of the Poisson distribution, $\mu = R(n)\ n(t)$, depends on the current population size. For example, we can run a stochastic simulation of the logistic model (3.5a) using $R(n) = 1 + r\ (1 - n(t)/K)$. If we let $r = 0.2$, R would again be 1.2 at low population sizes ($n(t) << K$). As the population size gets larger, however, the mean number of offspring per parent drops. With $n(0) = 10$, $r = 0.2$, and $K = 100$, the total number of offspring is Poisson distributed with mean $R(n(0))\ n(0) = (1 + 0.2\ (1 - 10/100))\ 10 = 11.8$. When we drew such a random number, we got $n(1) = 12$. In the next generation, the sum total number of offspring would follow a Poisson distribution with mean $\mu\ n(1) = (1 + 0.2\ (1 - 12/100))\ 12 = 14.1$, from which we drew a random number of $n(2) = 16$. Following this

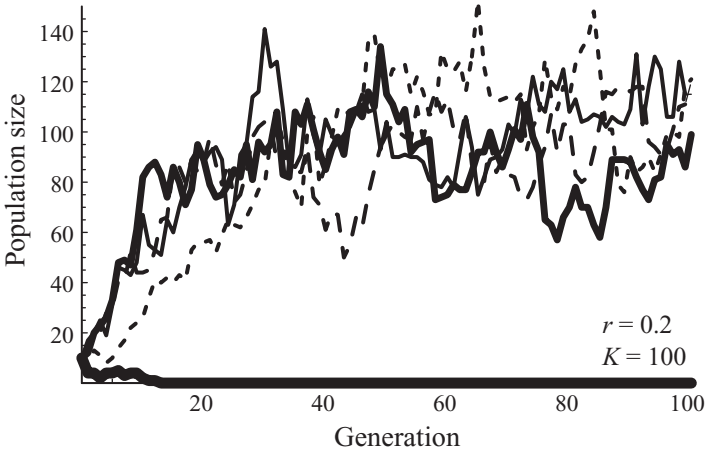

Figure 13.3: Stochastic model of logistic growth. Starting from a population of ten individuals, the number of individuals in generation $t + 1$ was drawn from a Poisson distribution with mean, $(1 + r\,(1 - n(t)/K))\,n(t)$, where $r = 0.2$ and $K = 100$, until 100 generations had passed. This process was repeated five times (five curves).

procedure for 100 generations and repeating the entire process five times gave us the data for Figure 13.3.

Although one replicate population out of five went extinct (again due to demographic stochasticity), the other four hovered around the carrying capacity of 100 and exhibited much less variability than Figure 13.2. Density dependence dampened the amount of demographic stochasticity by reducing the subsequent growth in those populations that happened to grow rapidly early on.

In Figures 13.2 and 13.3, we held the parameters R, r, and K constant, but environmental fluctuations can cause the parameters of a model to vary as well. This is referred to as *environmental stochasticity*. We can incorporate environmental stochasticity in the exponential growth model of Figure 13.2 by drawing the mean number of offspring per parent, R, from a probability distribution. For simplicity, assume that there are good years and bad years, with reproductive ratios R_g and R_b. If the chance that a year is good is p, the type of year will represent a Bernoulli random variable (Primer 3). We model environmental stochasticity by drawing a random number to determine the type of year. Specifically, each year, we draw a random number between 0 and 1 (uniformly); if the random number is less than p, the year is good; otherwise it is bad (Figure 13.4).

Variability in population size caused by chance fluctuations in the environment is known as *environmental stochasticity*.

The results in Figure 13.4 are dramatically different from Figure 13.2. The population size plummets during bad years, causing the trajectories to fluctuate wildly. Consequently, the risk of extinction is much higher. Indeed, out of 100 replicates with an average R of 1.2 and $n(0) = 10$, extinction occurred for 37 of the populations within 30 generations compared to only 3 with demographic stochasticity alone. Furthermore, the population size at generation 30 was smaller, on average (1775 versus 2470), with a much greater standard deviation (11,689 versus 1739).

These stochastic models of population growth exhibit fluctuations in population size regardless of the growth rate r. We also saw fluctuations in population size in the entirely deterministic model of logistic growth in discrete time when growth rates were high (Figure 4.2 and Box 4.1). Given data on changes over time in the size of a population, it can be difficult to determine the source of fluctuations (demographic stochasticity, environmental changes, or chaos).

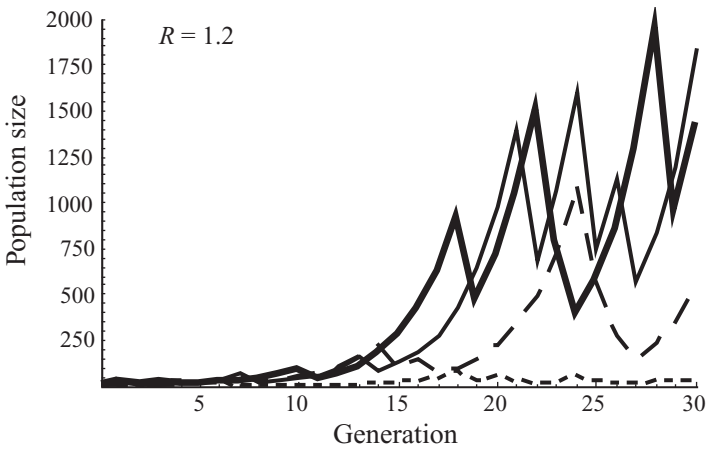

Figure 13.4: Exponential growth with demographic and environmental stochasticity. The probability of a good environment was $p = 0.7$. Each year, a random number, X, was drawn uniformly between 0 and 1 to determine if the current environment was good (if $X < 0.7$) or bad (if $X > 0.7$), where the reproductive factors in good and bad environments were $R_g = 1.5$ and $R_b = 0.5$, respectively. The average growth factor, $p\,R_g + (1 - p)\,R_b = 1.2$, is the same as Figure 13.2. One replicate went extinct after eight generations.

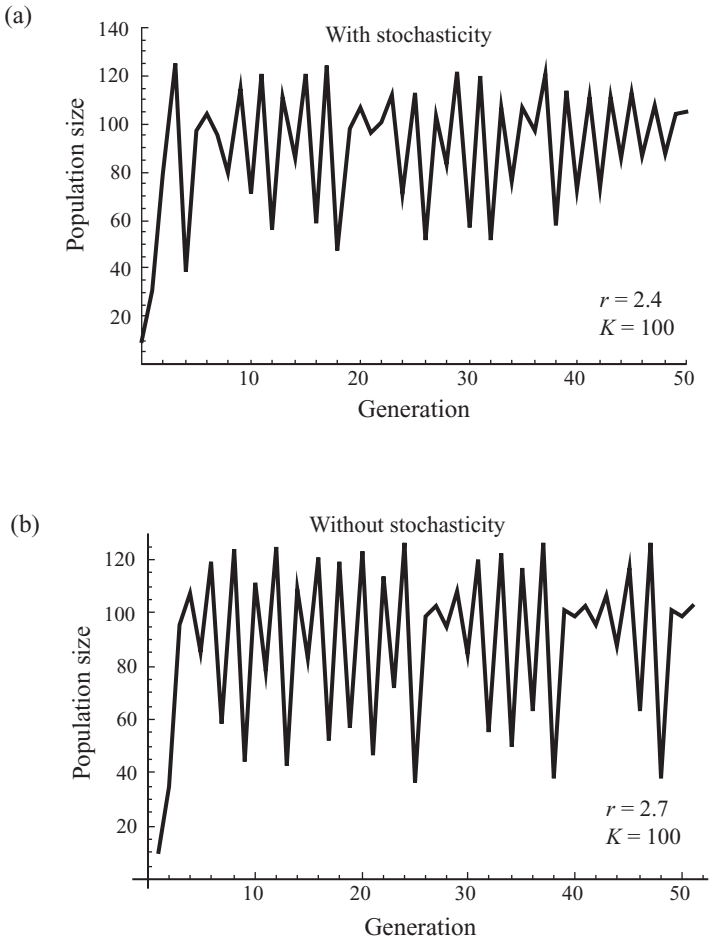

Figure 13.5: Random fluctuations or chaos? The logistic growth model was (a) simulated stochastically as in Figure 13.3 with $r = 2.4$ and (b) iterated deterministically as in Figure 4.2 with $r = 2.7$ and no stochasticity. The apparent randomness of the two trajectories have entirely different sources: (a) demographic stochasticity and (b) chaos. (Deterministically, a two-point cycle is expected with $r = 2.4$, and chaos is expected with $r = 2.7$.)

This point is illustrated in Figure 13.5, where panel (a) is a simulation of the stochastic logistic model with Poisson variation in the number of offspring per parent with $r = 2.4$ and panel (b) is a simulation of the deterministic logistic model (3.5a) with no variation in offspring number per parent and $r = 2.7$. These graphs look very similar, but they differ fundamentally in that the second graph is not random at all—each population size is exactly determined by the population size in the previous generation according to equation (3.5a).

More generally, several mechanisms can be acting simultaneously to affect population dynamics. The statistical field known as "time series analysis" was born to interpret data measured over time and to identify underlying dynamic forces. For example, spectral analysis determines whether there are cycles of particular frequencies within time series and can be used to ascribe these cycles to abiotic (e.g., climatic) and biotic (e.g., predator-prey) fluctuations (e.g., Loeuille and Ghil 2004). The interested reader is referred to Bjørnstad and Grenfell (2001), who review the literature on time series analysis applied to animal population dynamics, and to Kaplan and Glass (1995) for an introduction to time series analysis.

13.3 Birth-Death Models

In the previous section, generations were discrete and the entire population reproduced simultaneously. For populations in which reproduction is not synchronized, we need a different class of models. Imagine a vial of yeast. Yeast replicate by binary fission, but not every cell divides at the same time. If we were to track the population, we might see one cell divide and then another. Starting from only a few cells in the vial, we would initially observe few events per minute because there are so few cells replicating. As the population of cells expands, more and more new cells would be created each minute, causing cell "births" to occur in rapid succession.

How might we simulate this scenario? Let us start with a single cell. The chance that the cell replicates in any small unit of time Δt is $b\,\Delta t$, where b stands for the birth rate. As long as b is constant, the waiting time until the cell divides is exponentially distributed (see Definition P3.12) with mean $1/b$. For example, under nutrient-rich conditions, the mean time to cell division is approximately 90 minutes ($b = 0.011$ divisions per minute). In a simulation, we could draw a random number from the exponential distribution with parameter $\alpha = b$ to simulate the waiting time until cell division. Using *Mathematica*, we drew a waiting time of 83 minutes. Now we have two cells. As long as we don't care which cell divides, the total rate of cell division is twice what it was before, $\alpha = 2b$, and the distribution of waiting times is still exponential (Primer 3). Again using *Mathematica*, we drew a waiting time for the next cell division of 44 minutes from an exponential distribution with $\alpha = 0.022$. We could thus illustrate population growth as a series of steps rising from one cell to two cells at 83 minutes, to three cells after another 44 minutes, etc. To calculate the length of each step, we would draw a random number from an exponential distribution with mean $\alpha = b\,n(t)$ where $n(t)$ is the number of cells at time t (Figure 13.6).

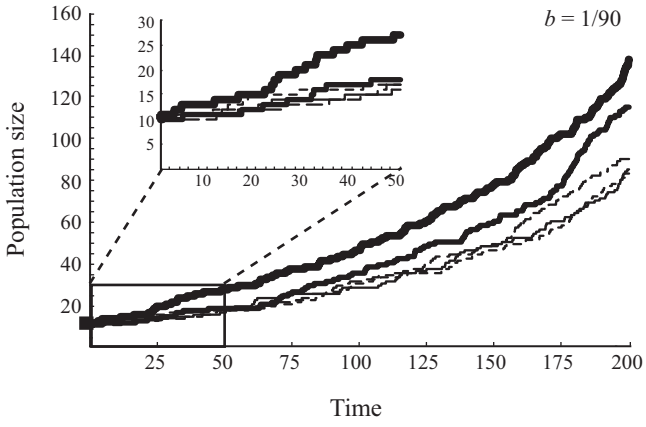

Figure 13.6: Birth process. A cell is chosen divide at a time randomly drawn from an exponential distribution with mean $b\,n(t)$, where $n(t)$ is the population size after the previous cell division. The division rate per cell was $b = 1/90$ per unit of time, and the initial population size was $n(0) = 10$. Five replicates are shown, and the inset shows the first 50 minutes.

This stochastic model is known as a *pure-birth process* or a *Yule process*, in honor of George Udny Yule (1924), who used this model to fit data on the number of species per genus assuming that speciation was akin to a birth. There is quite a bit of variation generated by a birth process, especially early on when few individuals are replicating (Figure 13.6 inset). The variation is not, however, as dramatic as in the stochastic model with discrete generations illustrated in Figure 13.2. In particular, the steps always rise upward because we allow only births within the population, but no deaths.

A *birth-death process* tracks changes to a population through births and deaths, assuming that only one event happens at a time.

We can extend the birth process to account for deaths by allowing individuals to die at rate d per individual per unit time as well as replicate at rate b. Such a model is known as a *birth-death process*. With $n(t)$ individuals, the waiting time until the next event happens, regardless of whether it is a birth or a death, depends on the total rate of events $\alpha = (b + d)\ n(t)$. When the event occurs, however, we must classify it as a birth or a death in order to track the resulting change in the population size.

In general, the chance that an event is a birth is given by $b/(b + d)$. For example, there is a 50% chance that the event is a birth when the birth and death rates are equal ($b = d$). This expression is fairly intuitive, but we can derive it formally using Rule P3.6. We wish to know the probability that a birth occurs in a time interval, Δt, given that either a birth or a death occurs in this interval. Using Rule P3.6, $P(\text{birth} \mid \text{birth or death}) = P(\text{birth} \cap \text{birth or death})/P(\text{birth or death})$. The event "birth $\cap$ birth or death" is read "birth and a birth or a death," and it can occur only if a birth occurs, which happens with probability $b\,\Delta t$; so $P(\text{birth} \cap \text{birth or death}) = b\,\Delta t$. Also, the probability of a birth or death is just $P(\text{birth or death}) = (b + d)\,\Delta t$. Therefore, we have $P(\text{birth} \mid \text{birth or death}) = b/(b + d)$.

We will analyze a birth-death process in Chapter 14, but to prepare for this analysis, let us summarize the behavior of the model in terms of the transitions possible in a small amount of time, Δt. Using an upper-case N to denote the random variable "population size", the probability that the population size at time $t + \Delta t$ is j, given that the population size at time t was i, is

$$p_{ji}(\Delta t) = P(N(t + \Delta t) = j \mid N(t) = i), \tag{13.1}$$

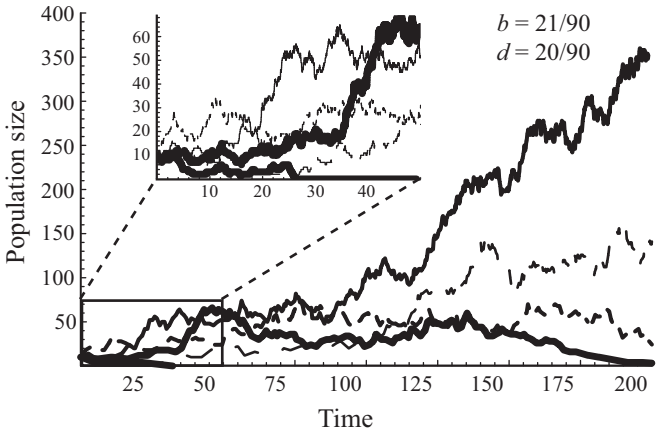

Figure 13.7: Birth-death process. The simulations of Figure 13.6 were repeated but with a birth rate of $b = 21/90$ and a death rate of $d = 20/90$, so that the net growth rate was $b - d = 1/90$ per unit time as in Figure 13.6. The inset figure shows the first 50 minutes. One population went extinct after 26 minutes.

where $p_{ji}(\Delta t)$ denotes the "transition probability" within a time period Δt. In a very short amount of time (so short that at most one event can occur) the transition probabilities $p_{ji}(\Delta t)$ are approximately

$$p_{ji}(\Delta t) = \begin{cases} b\,i\,\Delta t & \text{for } j = i + 1 \quad \text{(a birth)}, \\ d\,i\,\Delta t & \text{for } j = i - 1 \quad \text{(a death)}, \\ 1 - (b + d)\,i\,\Delta t & \text{for } j = i \quad \text{(no change)}, \\ 0 & \text{for } j \neq i - 1, i, i + 1 \quad \text{(other changes)}. \end{cases} \tag{13.2}$$

Figure 13.7 illustrates how adding deaths to the birth-process changes the dynamics (compare to Figure 13.6). Although the net growth rate ($b - d$) is the same (1/90), the inclusion of deaths causes the population to grow more erratically. In fact, one of the five replicates went extinct at $t = 26$.

So far, we have assumed that the per capita birth and death rates are constant, regardless of population size. It is easy to generalize this birth-death model to incorporate density dependence, by making either the birth or death rate a function of the number of individuals. Although it is possible to incorporate density dependence in a number of different ways, it is often assumed that competition among individuals acts to reduce the replication rate, and that the death rate remains constant (Renshaw 1991). For example, the per capita birth rate might decrease linearly with population size, as in the logistic model, giving the transition probabilities

$$p_{ji}(\Delta t) = \begin{cases} b\,i\left(1 - \dfrac{i}{K}\right)\Delta t & \text{for } j = i + 1 \quad \text{(a birth)}, \\ d\,i\,\Delta t & \text{for } j = i - 1 \quad \text{(a death)}, \\ 1 - \left(b\left(1 - \dfrac{i}{K}\right) + d\right)i\,\Delta t & \text{for } j = i \quad \text{(no change)}, \\ 0 & \text{for } j \neq i - 1, i, i + 1 \quad \text{(other changes)}. \end{cases} \tag{13.3}$$

Here, the probability of a birth is zero at K, which represents a limit to the population size. To revise the simulations, all we have to do is update the birth and death rates each time the population size changes. It is also possible to

incorporate temporal variation in the birth and death rates due to environmental fluctuations; such models are known as "nonhomogeneous birth-death processes."

Birth-death models have been applied to many other biological problems. For example, birth-death models have been used to describe changes in the number of repeats at microsatellites, which are stretches of DNA containing several copies in a row of a short motif (e.g., GAGAGAGA . . .) (Edwards et al. 1992; Ohta and Kimura 1973; Valdes et al. 1993). The birth-death model has also been used to describe the process of speciation (akin to birth) and extinction (akin to death), providing an interesting null model to describe the generation of biodiversity (Harvey et al. 1994; Nee et al. 1994a; Nee et al. 1995; Purvis et al. 1995). We will return to birth-death models in Chapter 14, where we describe analytical techniques that can be used to determine such things as the probability that the system is at any particular size, the probability of extinction, and the expected time until extinction.

13.4 Wright-Fisher Model of Allele Frequency Change

Next, we turn to a class of stochastic models that have played an important role in evolutionary biology. In the previous sections, the stochastic models focused on the total number of individuals within a population. Stochastic models can also be used to track the frequency of various types. We will again consider two different types of models. In this section, as in section 13.2, we assume that the entire population reproduces simultaneously, so that the generations are discrete and nonoverlapping. In the next section, we assume that generations are overlapping and that individuals are born and die at random points in time, as in the birth-death model of section 13.3. Again, the focus here will be on the development of these models and their simulation, laying the groundwork for the analytical techniques presented in subsequent chapters.

Consider a population that has a constant size, N, and only two types of individuals (A and a), as in the one-locus, two-allele haploid model (see extension to diploids in Problem 13.4). The deterministic model of this process, equation (3.8c), predicts that the frequency of type A at time $t + 1$ will be exactly $p(t + 1) = W_A\, p(t)/(W_A\, p(t) + W_a\, (1 - p(t)))$, where W_i represents the relative fitness of each type. By chance, however, individuals of type A might happen to leave more or fewer offspring in any given generation, so that $p(t + 1)$ will have a probability distribution centered around this deterministic prediction.

In the *Wright-Fisher model*, N offspring are sampled with replacement from the parental generation, which then dies. This sampling process causes random fluctuations in allele frequencies.

We first tackle the so-called "neutral" case where individuals are equally fit ($W_A = W_a = 1$). If the population size remains constant at N, and if the initial frequency of type A is $p(0)$, we can imagine individuals producing an infinite number of propagules (seeds, spores, etc.) from which a total of N surviving offspring are sampled. This thought experiment implies that the number of copies of allele A among the offspring should be binomially distributed with a mean of $N\, p(0)$ (see Primer 3). Thus, to simulate the Wright-Fisher model, we draw a random number from the binomial distribution with parameters N and $p(0)$.

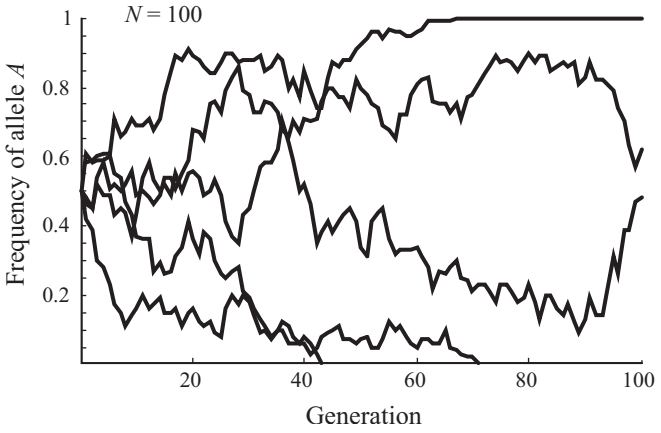

Figure 13.8: The Wright-Fisher model without selection. Each generation, offspring were chosen by randomly drawing from the alleles (*A* and *a*) carried by the parents with replacement (equivalent to binomial sampling). The population was assumed to be haploid and of constant size, $N = 100$. The frequency of allele *A* is plotted over time, starting with $p(0) = 0.5$.

The result j is the number of copies of the allele in the next generation, and $p(1) = j/N$. When we did this for an initial allele frequency of $p(0) = 1/2$ in a population of size $N = 100$, we drew 42 copies of allele A, so $p(1) = 0.42$. We can then find $p(2)$ by drawing a random number from a binomial distribution with parameters N and $p(1)$. Extending this process over 100 generations and repeating it five times, gave us the data for Figure 13.8.

The results are completely different from what we would expect based on the deterministic model of Chapter 3. According to equation (3.8c), when relative fitnesses are equal ($W_A = W_a = 1$), the allele frequency should stay constant ($p(t+1) = p(t)$). In Figure 13.8, however, the allele frequencies rise and fall by chance over time. This process, whereby random sampling of offspring causes allele frequencies to vary from their deterministic expectation, is known as *random genetic drift*. These chance events led to the loss of the *A* allele at generation 43 in one replicate and at generation 71 in another. Conversely, the *A* allele became fixed within the population at generation 67 in a third replicate. A polymorphism remained in two of the replicates at generation 100, but eventually the *A* allele would have been lost or fixed had we continued to run the simulations.

The process whereby random sampling of offspring causes allele frequencies to vary from their deterministic expectation is known as *genetic drift*.

Here, we have been using simulations to determine the probability that, at some future point in time, the population will be composed of a certain proportion $p(t)$ of type A. There is a faster way to calculate this probability distribution in small populations, which will provide us with a good background for the analysis in Chapter 14. First, because sampling N surviving offspring randomly and independently from all possible offspring is described by a binomial distribution, we can use Definition P3.4 to write down the probability that there are j individuals of type A at time $t+1$ given that there were i individuals of type A at time t. Using an upper-case X to denote the random variable "number of type A individuals", the transition probabilities for the Wright-Fisher model are

$$\begin{aligned} p_{ji} &= P(X(t+1) = j \mid X(t) = i) \\ &= \binom{N}{j}\left(\frac{i}{N}\right)^j\left(1 - \frac{i}{N}\right)^{N-j}, \end{aligned} \tag{13.4}$$

where p_{ji} denotes the transition probability within one generation. Equation (13.4) is the formula for the binomial distribution (Definition P3.4), but with p written as i/N.

We can use (13.4) to describe a "transition probability matrix" for the Wright-Fisher model, which gives the probability of going from any state i to any state j in one generation. Because we could have anywhere from 0, 1, 2, to N copies of type A, this matrix has $N + 1$ rows and columns. For example, in a population of size four, the transition probability matrix is

$$\mathbf{M} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{81}{256} & \frac{1}{16} & \frac{1}{256} & 0 \\ 0 & \frac{108}{256} & \frac{4}{16} & \frac{12}{256} & 0 \\ 0 & \frac{54}{256} & \frac{6}{16} & \frac{54}{256} & 0 \\ 0 & \frac{12}{256} & \frac{4}{16} & \frac{108}{256} & 0 \\ 0 & \frac{1}{256} & \frac{1}{16} & \frac{81}{256} & 1 \end{pmatrix}. \tag{13.5}$$

Each column sums to one because a population that starts with i copies of the allele must have some number between 0 and N copies in the next generation: $\sum_{j=0}^{N} p_{ji} = 1$. The first and last columns are particularly simple because there is no mutation; if nobody is type A ($i = 0$; first column) or if everybody is type A ($i = N$; last column), then no further changes are possible.

The helpful part about writing (13.5) in matrix form is that it can be iterated using the rules of matrix multiplication (Primer 2). $\mathbf{M}^2$ tells us the probability that there are j copies at time $t + 2$ given that there were i copies at time t. In general, $\mathbf{M}^t$ tells us the probability that there are j copies at time t given that there were i copies at time 0. For example, calculating $\mathbf{M}^{1000}$ using equation (13.5) (using a mathematical software package) gives

$$\mathbf{M}^{1000} = \begin{pmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.75 & 1 \end{pmatrix}. \tag{13.6}$$

(The zeros in the middle of this matrix aren't exactly zero, but they are less than 10^{-126}.)

We can also represent the initial state of the system using a vector

$$\begin{pmatrix} P(X(0) = 0) \\ P(X(0) = 1) \\ P(X(0) = 2) \\ P(X(0) = 3) \\ P(X(0) = 4) \end{pmatrix}. \tag{13.7}$$

For example, if the population initially had two copies of the allele, then $P(X(0) = 2) = 1$ and all other entries in this vector are zero. Multiplying $\mathbf{M}^{1000}$ on the right by this initial vector, we find

$$\begin{pmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.75 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0 \\ 0 \\ 0 \\ 0.5 \end{pmatrix}.$$

The vector on the right indicates that there is a 50% chance that type A will be lost ($j = 0$) after 1000 generations and a 50% chance that type A will be fixed ($j = 4$). If instead, the system initially had one copy of the allele, then $P(X(0) = 1) = 1$ and the remaining terms in vector (13.7) are zero. Now when we multiply $\mathbf{M}^{1000}$ on the right by this initial vector, we find that there is a 75% chance that type A will be lost and a 25% chance that it will be fixed after 1000 generations. These results suggest that if we start with i copies of type A, then type A will eventually be lost with probability $1 - i/N$ and fixed with probability i/N.

Writing this stochastic model in terms of a transition probability matrix suggests that we could apply the matrix techniques used in Primer 2 and Chapters 7–9 to understand stochastic models. This is exactly right, and we shall do so in the next chapter. Once again, eigenvalues and eigenvectors play a key role in analyzing stochastic models. At least for small population sizes, however, we can get an exact numerical solution just by calculating $\mathbf{M}^t$, against which we can check any theoretical prediction.

The Wright-Fisher model can be extended to incorporate fitness differences, mutation, multiple loci, etc. In reality, many of these processes are themselves stochastic, but a shortcut is often taken by assuming that these processes affect the number of propagules and their allele frequencies. If the number of propagules is very large, then these processes can be described by a deterministic recursion (e.g., using equation (3.8c) for selection or $(1 - \mu)\, p(t) + \nu\,(1 - p(t))$ for the allele frequency after mutation). As a consequence, sampling occurs only once, when the N adult individuals are chosen from the propagules.

As an example, Figure 13.9 illustrates simulations of the Wright-Fisher model with selection. In this figure, the A allele is 10% more fit than the a type and begins at a frequency of $p(0) = 0.05$. The simulations are run for populations of size (a) $N = 100$ and (b) $N = 10{,}000$. In both cases, the alleles rise in

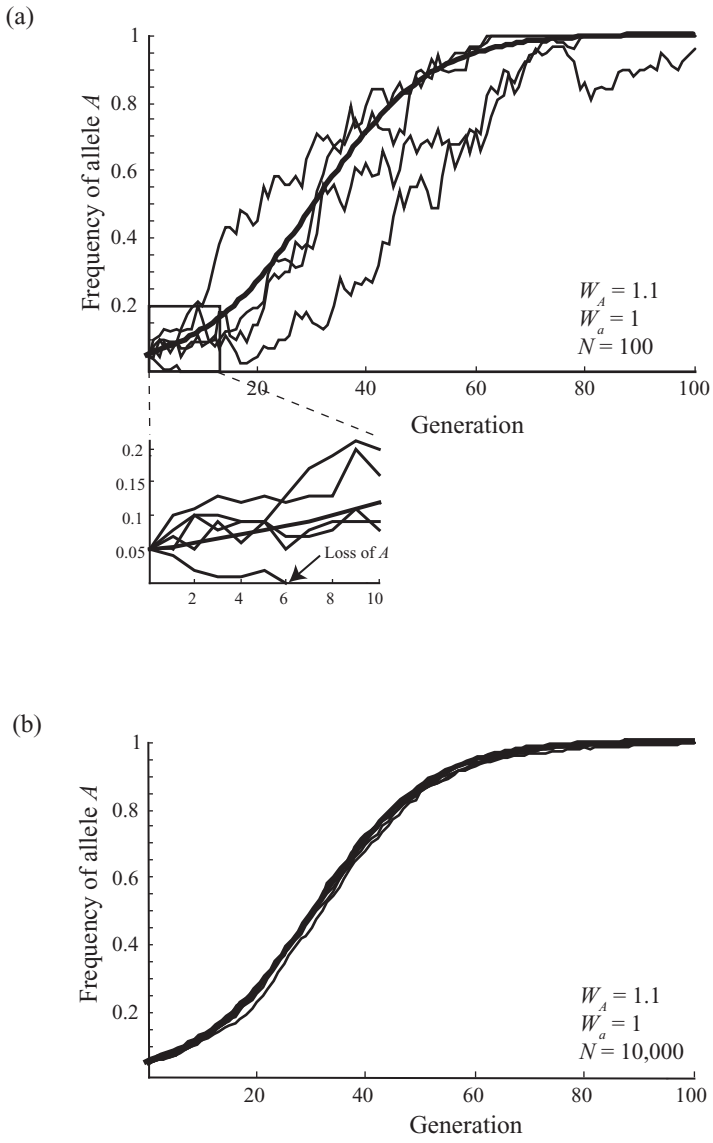

Figure 13.9: The Wright-Fisher model with selection. Simulations were carried out as in Figure 13.8 except that offspring alleles were more likely to be chosen from parents carrying the more fit *A* allele ($W_A = 1.1$, $W_a = 1$). For the thin curves, population size was set to (a) $N = 100$ and (b) $N = 10{,}000$ (five replicates each). For comparison, the thick curves illustrate the deterministic trajectory ($N = \infty$), obtained by iterating equation (3.8c). The frequency of allele *A* is plotted over time, starting with $p(0) = 0.05$. The favorable allele *A* was lost at generation 6 in one replicate with $N = 100$ (inset).

frequency towards fixation within 100 generations, roughly following the S-shaped trajectory seen in deterministic models (bold curve). When the population size is small, the Wright-Fisher model exhibits more variability around the deterministic trajectory than when the population size is large. This is consistent with the fact that the variance in the frequency of allele *A* due to sampling should be $p\ (1 - p)/N$ under the binomial distribution (see section P3.3.1). In fact, when *N* is only 100, we observe extinction of the beneficial allele in one of the five replicates (see inset figure). When *N* is 10,000, however, none of the replicates go extinct, and there is little variability in the trajectory.

These figures illustrate an important point: adding stochasticity to a model need not cause major changes to the results. In populations of small size ($N = 100$), we have seen allele frequency change when there should have been none (the neutral case, Figure 13.8), and we have witnessed the loss of a

beneficial allele, which we would expect to fix (Figure 13.9a). In populations of large size, however, there is less random genetic drift. Consequently, when the amount of chance (here represented by variation in samples from the binomial distribution) is small relative to other forces like selection, stochastic models can behave very much like deterministic models.

To simulate more than two types within a population (e.g., more than two alleles at a locus, or multiple genotypes at two loci), the above method must be modified by drawing from a multinomial distribution, with parameters N and p_i where the p_i are the frequencies of the various types, so that $\sum_{i=1}^{c} p_i = 1$ when there are c types (see Definition P3.5). This works well unless the number of types becomes very large. For example, with two alleles at each of 100 loci, there are 2^{100} possible haploid genotypes (Box P3.1). This number is greater than 10^{30}, which is much larger than any population. When there are too many types, drawing random numbers from a multinomial distribution grinds to a halt. What else can you do? The alternative is to develop an *individual-based model* where you mimic the production of each offspring within the population, one at a time (Deangelis and Gross 1992).

An *individual-based model* is a simulation where each individual is tracked explicitly, along with its properties (e.g., genotype, location, age, etc.).

Typically, in an individual-based model for the above process, you randomly draw a gamete from a mother and a gamete from a father within the parental population, and unite these to form a diploid offspring (followed by meiosis if you wish to produce a haploid offspring). If the fitness of the offspring is W and the maximum fitness is $W_{\max}$, you can then test to see if your offspring survives selection by drawing a random number uniformly between 0 and 1. If that random number is less than $W/W_{\max}$ the offspring becomes one of the N surviving individuals in the next generation, otherwise you start again by choosing new parents at random. This procedure works well unless the population size is very large.

13.5 Moran Model of Allele Frequency Change

In the Wright-Fisher model, we assumed that the entire population reproduced simultaneously. Intuitively, one might think that random genetic drift would be exaggerated by having the entire population replicate at once. To check this intuition, we explore a model where only one individual reproduces at a time. One way to do this would be to expand the birth-death process to allow multiple types of individuals (e.g., types of alleles) and to track the numbers of each type. In this case, you would observe both changes to the population size and to the frequencies of each type. But what if you wanted to hold the population size constant, to compare the results to those of the Wright-Fisher model?

The easiest way to adapt the birth-death process, holding the population size constant, is to couple each birth event with a death event. Whenever an individual is chosen to give birth, another individual is randomly chosen to die. Typically, the individual chosen to die can be any individual in the population, including the parent of the new offspring, but not the new offspring itself. It is also typical to track the population only at those discrete points in time where a birth-death event occurs, measuring time in terms of the number

In the *Moran model*, a randomly chosen individual reproduces, followed by the death of a randomly chosen individual. This sampling process also causes genetic drift.

of events that have happened rather than in chronological time. This evolutionary model is known as the *Moran model* (Moran 1962).

We focus on a population of size N with only two types A and a, where the number of copies of A is i and the frequency of A is $p = i/N$. If all individuals are equally fit, then the chance that an A-type parent is chosen to replicate is p. Thus, after one birth-death event, the number of copies of A goes up by one if the individual chosen to replicate carries the A allele (with probability p) and the individual chosen to die carries the a allele (with probability $1 - p$), giving an overall probability of $p\,(1 - p)$. Similarly, the number of copies of A goes down by one (if a replicates and A dies), with probability $(1 - p)\,p$. Finally, the number of copies stays the same if the individual chosen to replicate is the same type as the individual chosen to die, which happens with probability $p^2 + (1 - p)^2$. These calculations allow us to write down the probability of going from i copies of type A to j copies:

$$p_{ji} = P(X(t + 1) = j \mid X(t) = i), \tag{13.8}$$

where p_{ji} denotes the transition probability after one birth-death event, and $X(t)$ is a random variable representing the number of copies of type A at time t. For the Moran model, the transition probabilities p_{ji} are

$$p_{ji} = \begin{cases} p\,(1 - p) & \text{for } j = i + 1 \quad \text{(increase by one)}, \\ (1 - p)\,p & \text{for } j = i - 1 \quad \text{(decrease by one)}, \\ p^2 + (1 - p)^2 & \text{for } j = i \quad \text{(no change)}, \\ 0 & \text{for } j \neq i - 1, i, i + 1 \quad \text{(other changes)}, \end{cases} \tag{13.9}$$

where $p = i/N$. The key assumption of the Moran model is that the transition probability is zero for transitions that differ from the current state by more than one A allele.

Figure 13.10 illustrates the outcome of five replicate simulations of the Moran model starting with $i = 50$ copies of type A in a population of size $N = 100$. The simulations look similar to those from the Wright-Fisher model without selection (Figure 13.8). There are differences, however, as the inset figure shows. The allele frequency only jumps by $+/-\ 1/N$ in the Moran model, whereas much larger jumps can occur in the Wright-Fisher model. The main qualitative difference, however, is the scale along the x axis. There are only 100 generations represented in Figure 13.8 of the Wright-Fisher model, but 10,000 birth-death events represented in Figure 13.10 of the Moran model. You might be tempted to conclude that the Moran model exhibits less drift, but this is not a fair comparison. One time step in the Wright-Fisher model involves N births followed by the death of all N parents and so is more equivalent to N birth-death events in the Moran model. Thus, Figures 13.8 and 13.10 both represent the same total number of generations (100) with $N = 100$. Over this time period, and with only five replicates each, it is unclear which model exhibits more drift.

Given that no clear conclusions emerge from a few replicate simulations, we must run many more replicate simulations to compare the Wright-Fisher and Moran models. Starting with $p(0) = 0.5$ in a population of size 100, we ran 500

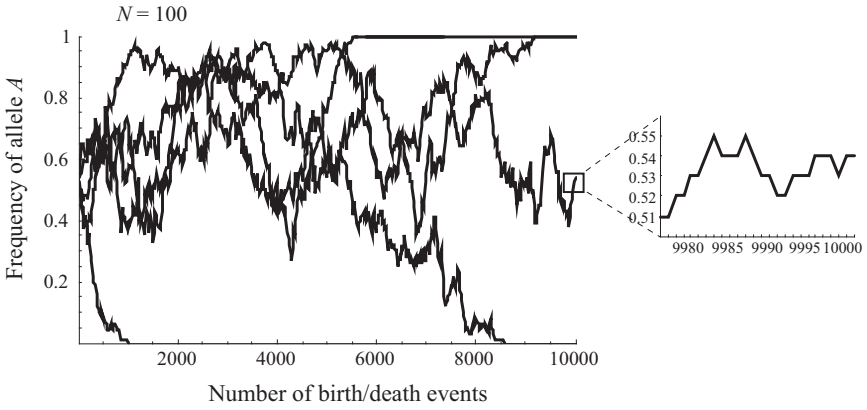

Figure 13.10: The Moran model without selection. At each time step, an individual was randomly chosen to give birth, after which an individual other than the new offspring was randomly chosen to die. The population was assumed to be haploid and of constant size, $N = 100$. The frequency of allele A is plotted over time, starting with $p(0) = 0.5$.

replicate simulations until fixation or loss of type A. By coincidence, the A type was lost 48.4% of the time using both the Moran and the Wright-Fisher model. In the Moran model, however, it took only 66.3 generations ($SE = 2.2$), on average, which was approximately half the time until loss or fixation in the Wright-Fisher model (133.3 generations with $SE = 4.6$).

The above results show that polymorphism is lost significantly faster in the Moran model than in the Wright-Fisher model. This result seems counterintuitive, because the Moran model makes only little jumps in frequency, whereas the Wright-Fisher model can make large jumps. A clue that can help us to understand this result is provided by the variance in reproductive success in the two models. When reproductive success is more variable, stochasticity (here, random genetic drift) plays a stronger role, and polymorphism will be lost by chance more rapidly.

In the Wright-Fisher model, the variance in reproductive success of single individuals, σ_r^2, is given by the binomial variance $N\,p\,(1-p)$ from equation (P3.4), when there is a single individual (i.e., with $p = 1/N$). Thus, $\sigma_r^2 = 1 - 1/N$. To calculate the variance in reproductive success over a single birth-death event in the Moran model, we use the formula for calculating variance (Definition P3.3), summing the squared change in number of copies over all possible transitions using (13.9):

$$\begin{aligned} &p(1-p)(+1)^2 + (1-p)p(-1)^2 + (1-2p(1-p))(0)^2 \\ &\quad = 2p(1-p). \end{aligned}$$

Because the variance of a sum of independent random variables is the sum of the variances (Table P3.1), the total variance in reproductive success over N such birth-death events is $2Np(1-p)$ per generation. Again, because we are interested in the variance in reproductive success of a focal individual, we set $p = 1/N$, demonstrating that $\sigma_r^2 = 2 - 2/N$. Thus, the Moran model exhibits

twice the variance in reproductive success, and consequently more random genetic drift, than the Wright-Fisher model (Ewens 1979). At an intuitive level, the Moran model is more variable because sampling occurs twice, when choosing which individual replicates and when choosing which individual dies.

The Moran model can be extended to incorporate processes such as selection and mutation by modifying the transition probabilities (Problem 13.6). For example, selection can be incorporated by altering the chance that an individual is chosen to reproduce. With selection, type A is chosen to give birth with a probability, p', equal to the frequency of type A weighted by its fitness (W_A) divided by the mean fitness: $p' = W_A\, p/\overline{W}$, where $\overline{W} = W_A\, p + W_a\,(1 - p)$. That is, p' is the same as the frequency change due to one generation of selection in the standard deterministic model of haploid selection (see equation (3.8c)). Making the key assumption that only one birth-death event occurs per time step, the transition probabilities are

$$p_{ji} = \begin{cases} p'(1-p) & \text{for } j = i+1 & \text{(increase by one)}, \\ (1-p')\,p & \text{for } j = i-1 & \text{(decrease by one)}, \\ p'p + (1-p')(1-p) & \text{for } j = i & \text{(no change)}, \\ 0 & \text{for } j \neq i-1, i, i+1 & \text{(other changes)}. \end{cases} \tag{13.10}$$

Here we have assumed that individuals are chosen at random to die, because we did not want to impose two bouts of selection on the population per generation. Other choices are equally plausible, however. You could impose viability selection on the death probabilities instead of (or in addition to) fertility selection on the birth probabilities.

In Chapter 14, we shall derive several important results using the Moran model, including the probability of fixation (or loss) and the time until fixation (or loss). These analytical results assume that there are only two types of individuals, so that we can count the number of one type and infer the number of the other. You can explore the Moran with more than two types by running simulations akin to Figure 13.10 by developing appropriate rules for who gives birth and who dies.

13.6 Cancer Development

The above examples are well-known and provide good background for how stochastic models can be constructed. In this section, we develop another example and model the occurrence of retinoblastoma, a cancer of the eye. This example will help illustrate how stochasticity can be incorporated into models investigating a wide variety of problems in biology.

Retinoblastoma is the most common eye cancer among children, with a worldwide incidence of about 5 in 100,000 children (Knudson 1971, 1993). The genetics of retinoblastoma are highly unusual. The mutation responsible for heritable cases of retinoblastoma occurs at the RB-1 gene on the long arm of chromosome 13 (Lohmann 1999). RB-1 is a tumour suppressor gene, and mutations in this gene disrupt control of the cell cycle. At a cellular level, the RB-1 mutation is recessive; the cell cycle is normal as long as there is one wild-type

allele in the cell. At an individual level, however, the mutation is dominant with a *penetrance* ρ of about 95%, meaning that about 95% individuals born with one mutant and one wild-type allele develop eye cancer (Knudson 1971, 1993).

How can a heterozygous individual get cancer when heterozygous cells are normal? The resolution of this paradox lies in the fact that somatic mutations occur sporadically during development, causing some cells in the eye to lose heterozygosity. It is those few mutant cells that lose their one copy of the wild-type allele that are responsible for retinoblastoma. Loss of heterozygosity (LOH) can occur by several mechanisms during mitosis (Lohmann 1999), including gene deletion, chromosome loss, mitotic recombination, and point mutations. Understanding the development of retinoblastoma requires a stochastic model, because the chance timing of mutational events determines whether cancer develops, as well as its severity.

Figure 13.11 illustrates the development of the vertebrate retina. The single-celled zygote undergoes five binary cell divisions to reach the 32-celled blastula stage. Experiments performed at this stage in *Xenopus* indicate that only nine of these blastomere cells (*a* through *i*) contribute to the retina of each eye (Huang and Moody 1993). These cells then undergo a series of n cell divisions. Averaged over the 32 blastomere cells, n must be $\sim$41 to account for the approximately 10^{14} cells in the human body (Moffett et al. 1993). The retina is composed of $\sim 1.5 \times 10^8$ cells (Bron et al. 1997; Dreher et al. 1992), but only three of the seven major retinal cell types (horizontal, amacrine, and Müller cells) appear to have the potential to proliferate into retinoblastoma in RB-1 homozygous mutant cells (Chen et al. 2004). Based on counts of these three cell types (Dreher et al. 1992; Van Driel et al. 1990), the total number of retinal cells that have the potential to cause retinoblastoma in one fully formed eye, C, is $\sim 2 \times 10^7$.

The experiments of Huang and Moody (1993) also indicate that different fractions of retinal cells descend from each blastomere cell (see inset table in Figure 13.11). We will call these fractions f_a through f_i. For example, the cell D1.1.1 (marked as "*a*") contributes 49.7% of the cells in the left retina. The exact cell fate is determined later in development, so each blastomere contributes to the different cell types in the retina (Huang and Moody 1993). We incorporate these observations by letting $f_y\, C$ equal the number of susceptible cells contributed by the blastomere cell y to the left retina.

To model stochastic mutation, we assume that mutations occur during DNA replication (i.e., at discrete points in time). Whether a daughter cell produced by a heterozygous parent cell is mutant represents a random variable with two possible outcomes (a Bernoulli trial): with probability μ it is mutant, and with probability $1 - \mu$ it remains heterozygous. Unfortunately, we do not know exactly when each progenitor cell divides in the development of the retina. As a preliminary map of development, we considered Figure 13.12. Phase 1 consists of the five cell divisions leading to the blastula. In phase 2, cell divisions produce all of the cell types in the body, and we assume that only one daughter cell per division remains in the lineage leading to the retina. In phase 3, the stem cells of the retina proliferate, with all daughter cells contributing to the retina. The number of divisions in phase 3, m_y, is chosen to ensure that

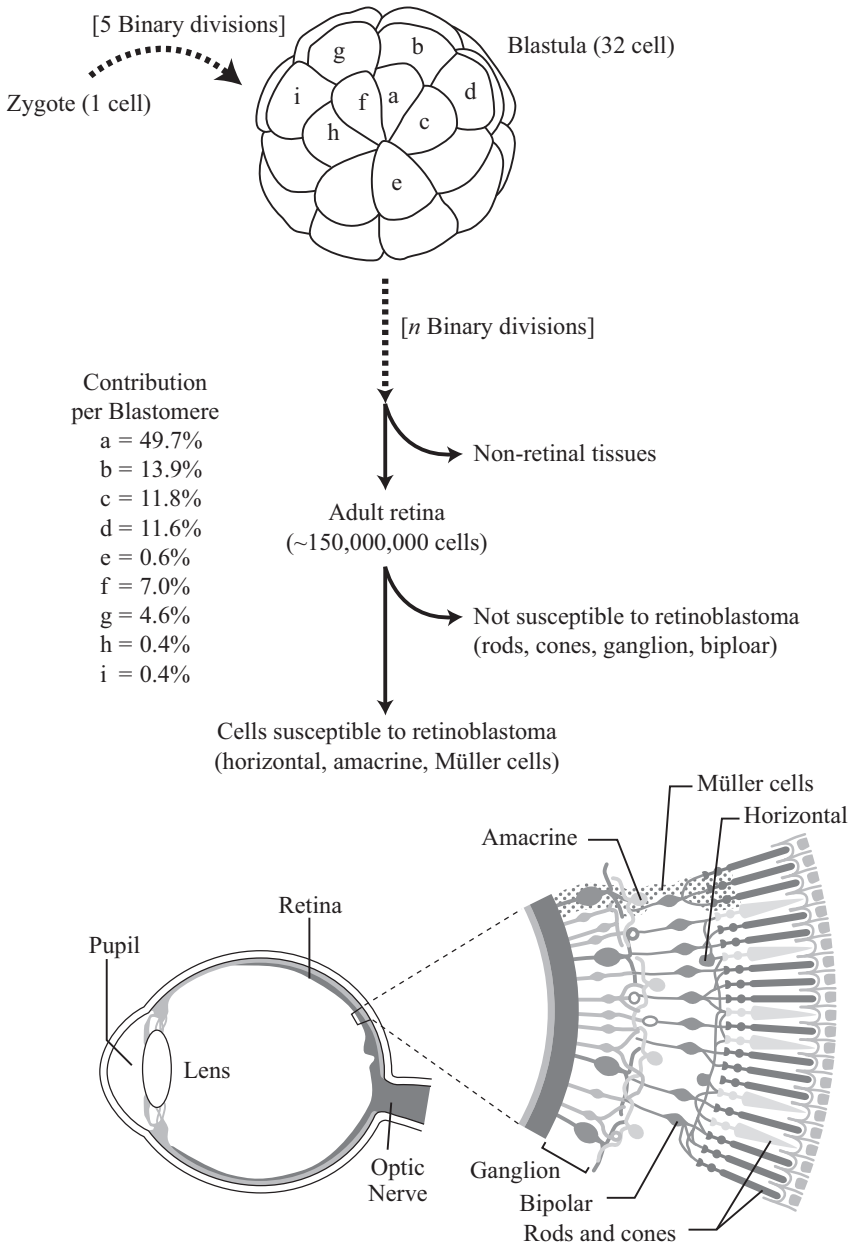

Figure 13.11: Development of the retina. Development from the zygote (top left), through the blastula stage (top center), to the eye (bottom center) is illustrated (http://webvision.med.utah.edu). Percentages indicate the fraction of retinal cells of the left eye derived from each of the blastomeres marked *a* through *i* (Huang and Moody 1993).

blastomere y contributes the appropriate number of susceptible cells to the fully developed retina, $f_y C$. (For a more precise calculation, we allow a fraction p_y of the cells to undergo an additional cell division to get exactly $f_y C$ cells.)

The bulk of mutations causing a loss of heterozygosity are likely to happen when there are many cells (i.e., many Bernoulli trials), which occurs when the

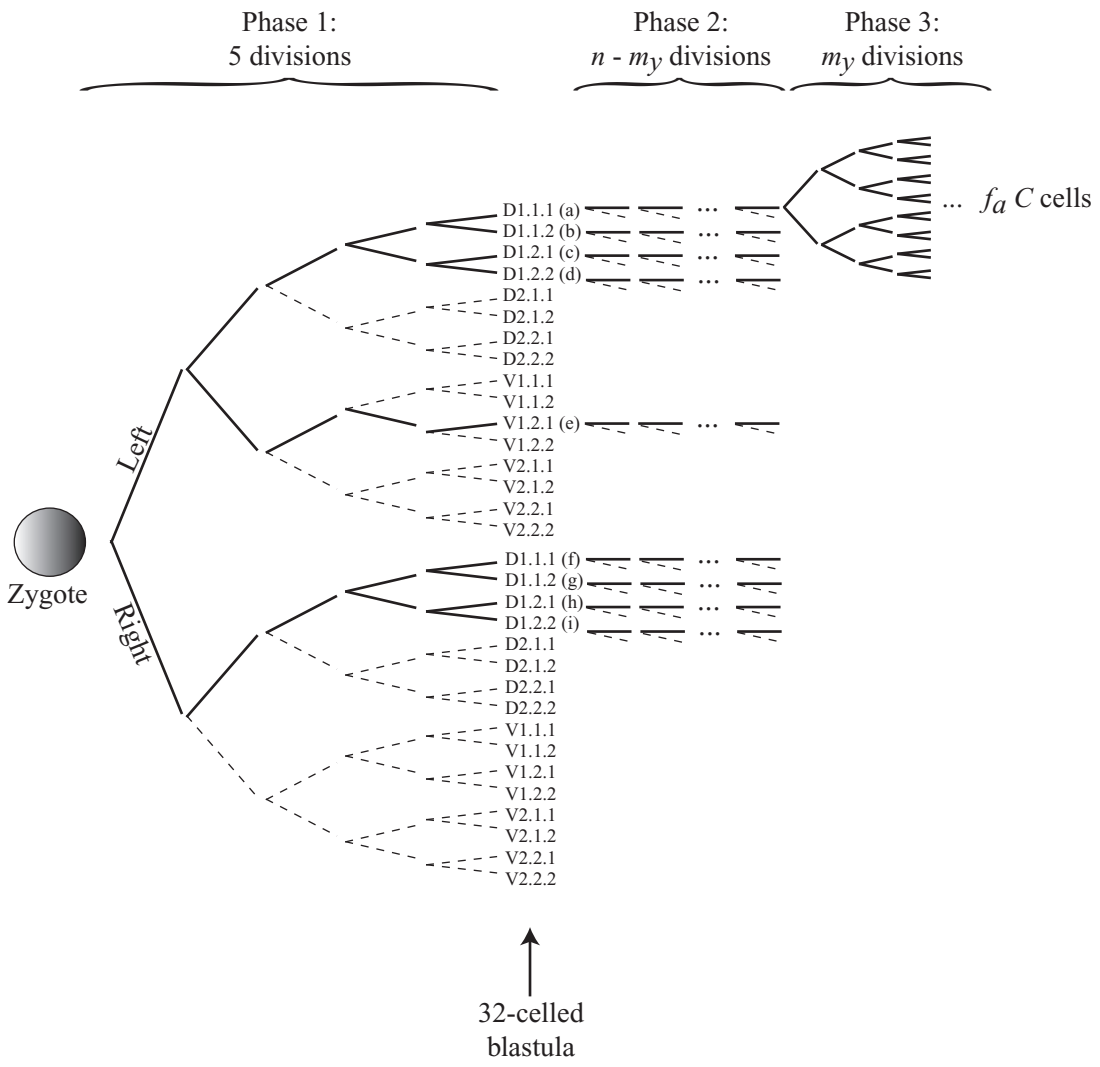

Figure 13.12: A cell-lineage map leading to the eye. With time proceeding from left to right, lines connect parent cells to daughter cells; solid lines indicate lineages that contribute to the pool of susceptible retina cells, while dashed lines indicate lineages that do not contribute to the retina. Phase 1 consists of the five cell divisions from the zygote to the blastula. Phase 2 consists of the cell divisions between the blastula and the stem cells that generate the retina. Phase 3 consists of the proliferation stage during which the retina develops from a series of binary divisions. The exact details in phases 2 and 3 are not known.

eye is nearly fully developed (phase 3 of Figure 13.12). Thus, we might expect that the exact number of cell divisions during phase 3, m_y, would be much more critical than the number in phase 2, $n - m_y$.

Our goal is to characterize the probability that retinoblastoma occurs and in what form: in one eye or both, and with multiple tumors per eye or only one. If we carried out a Bernoulli trial for every daughter cell illustrated in Figure 13.12, however, simulating development would be quite slow. We can speed up the process by simulating mutations among the $x(t)$ daughter cells produced at cell

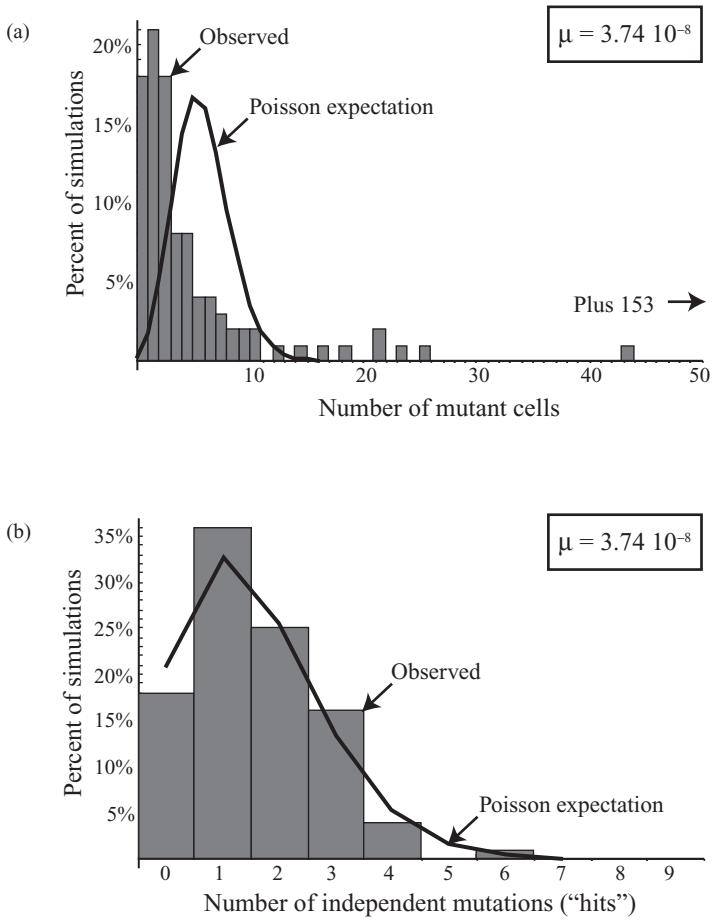

Figure 13.13: A stochastic model of mutation leading to retinoblastoma. Simulations were based on the exact sequence of cell replication described in Figure 13.12 and replicated 100 times. Starting with a heterozygous zygote, the number of cells that lose the wild-type allele in the *t*th round of cell division was drawn randomly from a binomial distribution with parameters $x(t)$ (the number of daughter cells) and μ (the mutation rate). (a) A histogram of the total number of mutant cells per eye. (b) A histogram of the number of distinct mutational events leading to the cancerous cells in an eye. Curves illustrate a Poisson distribution with the same mean as the observed distribution. $\mu = 3.74 \times 10^{-8}$, $C = 2 \times 10^7$, $n = 41$.

division t by the parent cells that remain heterozygous. The number of these daughter cells that lose the wild-type allele is a random variable drawn from a Binomial distribution with probability μ and a number of trials equal to $x(t)$. The daughter cells that remain heterozygous then produce $x(t + 1)$ daughter cells, and the process continues. All mutant cells and their descendents are kept track of separately, as these are assumed to remain mutant. We used this method to generate the histograms in Figure 13.13, replicating the process of development 100 times and using a mutation rate of $\mu = 3.74 \times 10^{-8}$ per daughter cell (as estimated below). Figure 13.13a illustrates the total number of homozygous mutant cells that developed within the left eye of each simulated "individual." Many of these mutant cells descended from the same mutation. Figure 13.13b illustrates the number of independent mutations that led to the observed number of mutant cells in the left eye.

These two histograms tell an interesting story. In the second histogram, the number of mutational events closely follows a Poisson distribution, as expected if mutations occur independently at a small rate in a large number of Bernoulli trials (recall that the Poisson distribution is an excellent approximation to

the binomial distribution in this case). Technically, the LOH (loss of heterozygosity) mutations do not occur independently, because the descendants of a LOH mutation cannot have a further LOH mutation. Nevertheless, because most mutations happen late in development when there are many cells, there is little opportunity for further mutation. The first histogram is decidedly not Poisson and has "fat tails" (leptokurtosis). That is, there is a much higher probability of observing many LOH cells, or none, than expected based on the mean number of LOH cells.

The great variability in outcomes observed in our model is typical of a *jackpot* distribution. A jackpot distribution is one where there is a small chance of getting a very large outcome (akin to the small chance of winning a lottery). Such a distribution arises naturally when modeling mutation in a growing population of cells, because there is a small chance that a mutation happens early and is carried by many descendent cells. Although our model incorporates more developmental details, the results are fundamentally similar to a model developed by Luria and Delbruck (1943). These authors carried out a series of experiments growing bacteria in liquid culture and afterwards exposing the cells to a novel environment (a bacteriophage). They then counted up the number of resistant cells and observed a jackpot distribution—some cultures contained many resistant cells while most had few. Luria and Delbruck then used a mathematical model of mutation to demonstrate that mutations must have occurred during the growth of the population, before exposure to the novel environment, and not in response to the novel environment—only then is a jackpot distribution expected. This result became a cornerstone of modern genetics. The Luria-Delbruck model also forms the basis for an important method used to calculate mutation rates, known as the *fluctuation test*.

The results of Figure 13.13b can be used to predict the form of retinoblastoma. The probability that an eye is not affected is estimated by the height of the bar at 0: $p_0 = 0.18$. Using this estimate, we can calculate the probability of observing no retinoblastoma, retinoblastoma in one eye (unilateral), and retinoblastoma in both eyes (bilateral) among individuals that inherit the RB-1 mutant allele. Assuming that the two eyes represent independent sampling events, each with a probability p_0 of being unaffected, these probabilities are given by the binomial distribution:

$$P(\text{no retinoblastoma}) = p_0^2 = 0.032,$$
$$P(\text{unilateral retinoblastoma}) = 2p_0(1 - p_0) = 0.295,$$
$$P(\text{bilateral retinoblastoma}) = (1 - p_0)^2 = 0.672.$$

Furthermore, there is a pretty high probability, 46%, that an eye contains multiple tumors (summing the bars from 2 onward in Figure 13.13b).

Even with the fairly complicated model of development illustrated in Figure 13.12, we can make some general predictions using the probability theory introduced in Primer 3. To do so, we need to derive formulas for the values of p_0, p_1, and p_{2+}, rather than estimating them from simulations. Calculating p_0 is

the most straightforward, so we focus only on p_0 and on the questions that can be answered with this quantity.

If S is the total number of daughter cells produced throughout the development of one eye, then the probability that none of these are mutant is

$$p_0 = (1 - \mu)^S \tag{13.11}$$

(Definition P3.4 with $k = 0$), where μ is the mutation rate per daughter cell. Equation (13.11) provides us with a way to relate the mutation rate to the penetrance of the mutation (i.e., to the probability that an individual is not affected). To calculate S, we count the number of daughter cells ever produced that contribute to the susceptible population of retinal cells in one eye. Using Figure 13.12, S is very nearly 4×10^7, almost all of which arise in Phase 3.

Using equation (13.11), the probability of being free of symptoms in both eyes is given by $p_0^2 = (1 - \mu)^{2S}$. One minus this quantity gives the probability of getting a tumor in at least one eye (the penetrance): $\rho = 1 - (1 - \mu)^{2S}$. We can rearrange this equation to solve for the mutation rate: $\mu = 1 - (1 - \rho)^{1/(2S)}$. Given the observed penetrance, $\rho = 0.95$, and $S = 4 \times 10^7$, the estimated mutation rate is $\mu = 3.74 \times 10^{-8}$ per daughter cell produced, as used above.

Is this estimated mutation rate per daughter cell reasonable? The observed mutation rate at RB-1 is 8×10^{-6} per individual generation (Knudson 1993). The number of cell divisions within humans has been estimated as ~179 divisions from zygote to zygote (averaged across sexes and assuming a generation time of 25 years; Vogel and Rathenberg 1975). Thus, the observed mutation rate corresponds to a mutation rate of 4.47×10^{-8} per cell division, which is reasonably close to our estimated mutation rate of 3.74×10^{-8}.

We can also use equation (13.11) to predict the form of retinoblastoma:

$$P(\text{no retinoblastoma}) = p_0^2 = (1 - \mu)^{2S},$$
$$P(\text{unilateral retinoblastoma}) = 2p_0(1 - p_0) = 2(1 - \mu)^S\,(1 - (1 - \mu)^S),$$
$$P(\text{bilateral retinoblastoma}) = (1 - p_0)^2 = (1 - (1 - \mu)^S)^2.$$

Using $\mu = 3.74 \times 10^{-8}$, these calculations predict that, of individuals initially carrying the RB-1 mutation, 5% should be symptom free, 35% should develop unilateral retinoblastoma, and 60% should develop bilateral retinoblastoma. These predictions are consistent with observations (Knudson 1971) and with the simulation results presented above. Interestingly, these results depend only on μ and S.

Our model of retinoblastoma could be improved by taking into account a more sophisticated version of development than illustrated in Figure 13.12. Yet our model provides insight into which details matter most. As mentioned earlier, the exact number of cell divisions during phases 1 and 2 has a negligible influence on the number of mutations that arise. In fact, our results were nearly unchanged when we replaced Figure 13.12 with a simple series of binary cell divisions. While our results are not sensitive to events during phases 1 and 2,

they would be sensitive to events late in development, including the exact number of cells in the retina (C) and the extent of cell births and deaths in phase 3.

13.7 Cellular Automata—A Model of Extinction and Recolonization

In previous models, we ignored the spatial location of individuals. Space often matters, however, because individuals tend to interact and breed locally and might not migrate over long distances relative to the range of the species. For example, HIV is highly spatially structured in different tissues within an infected individual (Frost et al. 2001). Only by accounting for this structure do models generate reasonable predictions for the level of genetic variability observed in HIV and the ability of HIV to respond to antiretroviral drugs. Although some models of spatially structured populations are analytically tractable (see Chapter 15 as well as examples in Nisbet and Gurney (1982) and Renshaw (1991)), many are not. Numerical analysis of spatial models has thus played an important role in biology.

A commonly used type of spatial model is a *cellular automaton*. An *automaton* is a machine or robot that carries out a series of instructions. A cellular automaton is an array of automata arranged in a lattice or grid, where each automaton is assigned its own position or *cell*. Typically, the grid lies in one or two dimensions, and cell shapes are uniform (as in the square grid illustrated in Figure 13.14). But the exact size and shape of a cellular automaton is flexible. One of the more famous cellular automata is the game of life, invented by John H. Conway to mimic births and deaths in a spatially arranged population (Gardner 1983).

To simulate a biological process on a cellular automaton, you must first specify the initial states of each cell and the instructions that each automaton

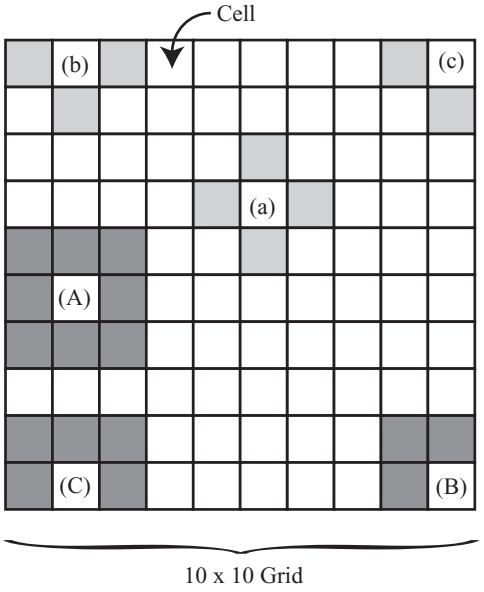

Figure 13.14: A cellular automaton. Each cell in this 10 × 10 grid is either inhabited or empty and can receive migrants from n nearest neighbor cells. Allowing migration from only the vertical and horizontal nearest neighbors, the light grey cells are potential sources of migrants to the focal cells: (a) a center cell ($n = 4$), (b) an edge cell ($n = 3$), (c) a corner cell ($n = 2$). Allowing migration from the vertical, horizontal, and diagonal nearest neighbors, the dark gray cells are potential sources of migrants to the focal cells: (A) a center cell ($n = 8$), (B) an edge cell ($n = 5$), (C) a corner cell ($n = 3$).

must use to determine their state in the next time step. The instructions to be carried out typically depend on the states of the surrounding cells. For example, the original game of life was played on a square grid, with each cell being dead (0) or alive (1). The number of live cells in the eight cells surrounding a given focal cell was then counted (n). If n was two, the state of the focal cell (alive or dead) remained unchanged. If n was three, the focal cell was set to 1 (alive) regardless of what it was before. In all other cases, the focal cell was set to 0. These cases roughly describe survival, birth, and death in the presence of local reproduction and competition. The exact rules were not chosen to portray growth in any particular species, per se, but to generate interesting spatial patterns, without exploding or imploding too rapidly.

In the game of life, the rules for updating the cells are deterministic, but stochastic rules are commonly used in cellular automaton models. As an example, we develop a cellular automaton model of extinction and recolonization (see Problems 5.12 and 5.13). In the nonspatial model, a fraction of patches, $p(t)$, is occupied at time step t. Of the $1 - p(t)$ unoccupied sites, a fraction $m\,p(t)$ are recolonized from occupied patches. Subsequently, each occupied site suffers a risk of extinction e through catastrophic events such as fire or disease. The resulting recursion equation for the deterministic model is (see Problem 5.12)

$$p(t + 1) = (1 - e)(p(t) + m\,p(t)\,(1 - p(t))). \tag{13.12}$$

To be more realistic, we model a spatial version of this model, where each cell in a 10×10 square grid represents a patch (empty or occupied) and where extinction and recolonization are stochastic events. Recolonization of an empty patch at position $\{i,j\}$ occurs with probability $m\,f_{i,j}$, where m is the recolonization rate per patch and $f_{i,j}$ is the number of neighboring patches that are occupied. This process is repeated for each unoccupied cell in the grid. In our simulations, we considered the neighborhood size to consist of the eight nearest cells (Figure 13.14A). We run into a problem, however, when we consider cells on the edge of the grid, which don't have eight neighbors. There are two approaches for handling the edges of a grid. First, the grid can be "wrapped around" to make a torus (a donut shape), so that, for example, a cell on the left edge can receive migrants from cells on the right edge. This procedure ensures that the edge cells and the central cells follow the same rules and is thought to represent populations larger than the grid size more accurately. The second approach assumes that the environment outside of the habitat is inhospitable, so that edge and corner cells really have fewer neighbors (Figure 13.14). We used this second approach in our simulations.

Next, we consider extinction. For each occupied site on the grid, we choose a random number uniformly between 0 and 1. If the random number is less than e, the population goes extinct, otherwise the site remains occupied.

Simulations of this extinction-recolonization model are illustrated in Figure 13.15 (the *Mathematica* code used to generate the figure is available on the book website). Colonization causes the spread of populations to adjacent cells and generates clusters of occupied cells. Extinction, however, causes sites

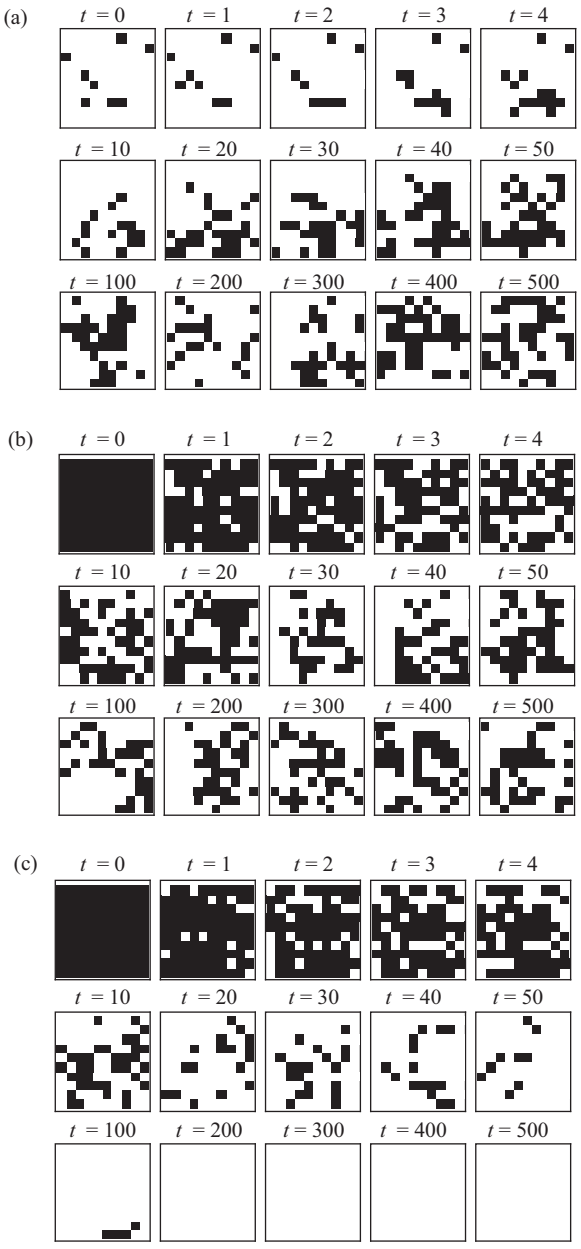

Figure 13.15: Simulations of the extinction-recolonization model. Occupied (black) and empty (clear) patches are shown on a 10×10 grid of sites. The simulations are run for 500 generations, and a snapshot of the metapopulation is shown at several intermediate time points. (a) $m = 0.05$, $e = 0.16$, initial fraction of filled sites = 10%, final fraction of filled sites = 41%, (b) $m = 0.05$, $e = 0.16$, initial fraction of filled sites = 100%, final fraction of filled sites = 36%, (c) $m = 0.03$, $e = 0.16$, initial fraction of filled sites = 100%, final fraction of filled sites = 0% (extinction).

that were previously occupied (black) to become empty (clear). Over time, the grid approaches a balance between filled and empty sites that is roughly the same whether 10% of sites were initially filled (Figure 13.15a) or 100% (Figure

13.15b). Eventually, however, the ensemble of populations goes extinct, but this takes many generations (t = 23,343 and 31,282 generations in the simulations of Figures 13.15a and 13.15b, respectively). In contrast, if we reduce the migration rate or increase the extinction rate, extinction happens much more rapidly, on average (e.g., t = 144 in the simulations of Figure 13.15c).

The main advantage of cellular automaton models is that they allow us to explore the dynamics of a population arranged over space, so that we can determine how summary statistics such as the mean extinction time of a species depend on the spatial arrangement and connectedness of populations.

13.8 Looking Backward in Time—Coalescent Theory

In all of the stochastic models considered so far, we imagine time running forward. While this is natural, there are some problems for which it is faster and easier to imagine time running backwards. This isn't as crazy as it seems. For example, when you draw your family tree, you start with yourself in the present and go backward in time through your parents, grandparents, etc. This example provides a good explanation for why you might want to run time backwards. To draw your family tree forward in time, you would have to start with every individual alive in, say, 1700 A.D. Then you would figure out who gave birth to whom, and draw every generation to the present day. Having traced every family lineage to the present, you would then throw out almost all of this information, keeping only those lineages that led to you. Ridiculous. Working backward in time thus makes sense when you are interested in a focal individual living in the present and in the historical processes leading up to that individual.

Here we explore a model based on the assumptions of the Wright-Fisher model, but that is run backward in time. The analysis of this model has led to an important new branch of mathematical biology known as *coalescent theory*. We begin in the present (t = 0), focusing on a certain number of alleles (n) sampled from a population of N individuals. To simplify the situation, we assume that the population is haploid, but all of the following results apply to a randomly mating diploid population as long as we replace N with $2N$, the number of alleles in a diploid population of N individuals.

In the current generation (t = 0), there is a different individual alive for each of the n alleles. For now, we do not keep track of whether the alleles encode the same DNA sequences, but rather only whether the alleles are carried by different individuals. In the previous generation (t = 1), there is some chance that two of the alleles descended from the exact same parent allele, meaning that there were only $n - 1$ different parent alleles that gave birth to the n alleles sampled today (Figure 13.16). This event, whereby n offspring alleles descended from only $n - 1$ parent alleles, is known as a *coalescent event*, and represents two lineages coming together ("coalescing") into one. If we predict that the alleles in our sample are likely to have coalesced in the recent past, then these alleles should be closely related and similar to one another. Conversely, alleles that are predicted to coalesce in the distant past should be less similar. Coalescent theory has had such a great impact because it predicts the relatedness among

Coalescent theory describes the probability that alleles in a sample descend from the same ancestral allele at time t in the past.

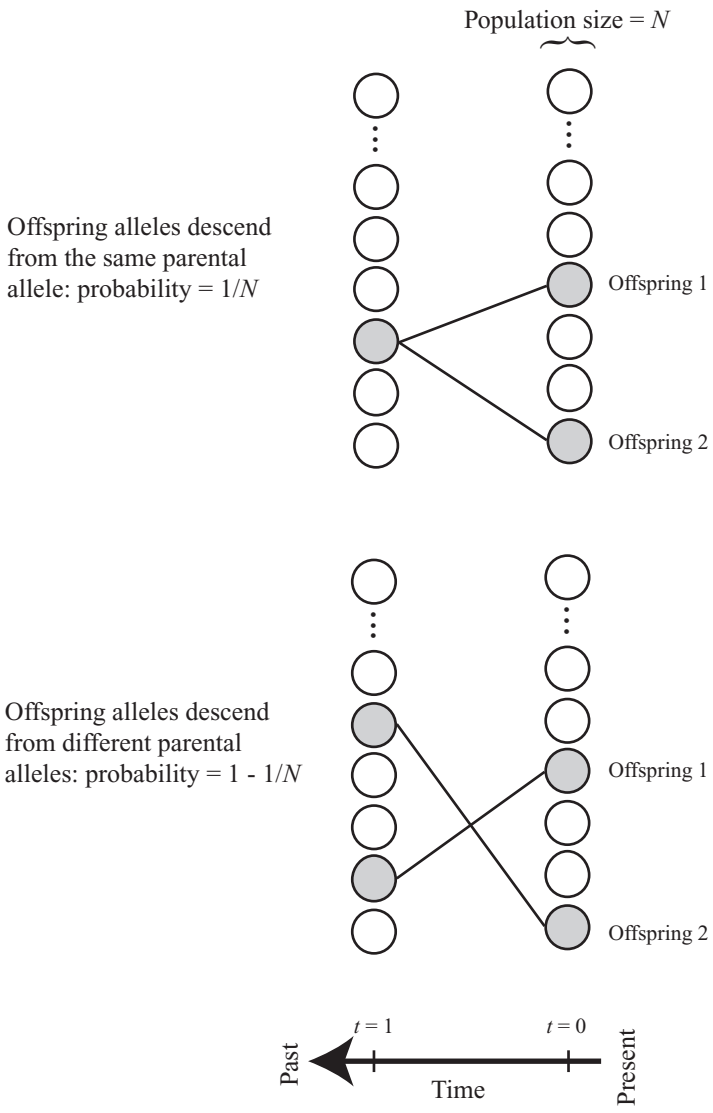

Figure 13.16: Descent of alleles from parents to offspring. The probability that two focal alleles (shaded circles) were born from the same parent allele is $1/N$ (top); this is called a coalescent event. Otherwise, the two focal alleles were born from different parent alleles (bottom). In Figures 13.16–13.18, time runs from the present on the right to the past on the left.

samples of individuals, predictions that can be tested using the DNA sequences carried by these individuals (Felsenstein 2004; Hudson and Kaplan 1995; Rosenberg and Nordborg 2002).

Our first aim is to describe the chance that a coalescent event happens t generations in the past, starting with a sample of only two alleles. The time in the past at which these two alleles coalesce, T_2, is the random variable of interest, and we seek the probability distribution for T_2. Given that all individuals in the population reproduce simultaneously (as in the Wright-Fisher model), what is the probability that two alleles descend from the same parent allele? The first sampled allele must have had some parent (with probability equal to one), but the second sampled allele could have had the same parent (with probability $1/N$) or a different parent (with probability $1 - 1/N$). Thus, the probability that there was a coalescent event in the previous generation ($t = 1$) is $1/N$. If the

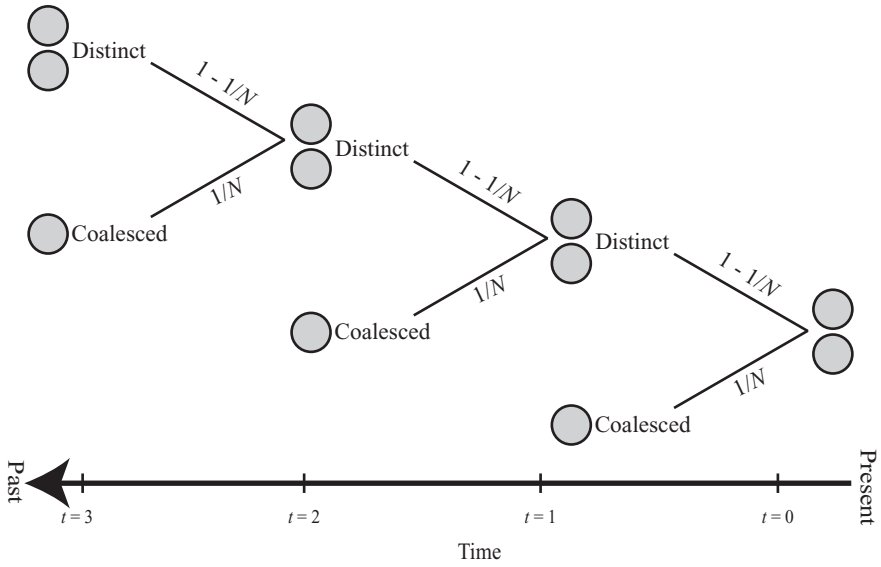

Figure 13.17: A decision tree for the coalescent model. Given two sampled alleles (shaded circles), there are two possibilities: the alleles share the same parent in the previous generation (coalesce) or they remain distinct. Once the two alleles coalesce, the coalescent process is over, and we no longer trace their history. The probability of one particular outcome is calculated as the product of the probabilities along the path to that outcome. For example, the probability of a coalescent at time $t = 3$ is given by $(1 - 1/N)$ $(1 - 1/N)$ $(1/N)$.

alleles coalesced, we know how related the sampled alleles are: they are siblings. If the alleles did not coalesce, then we are right back to where we started: with two alleles (now at $t = 1$), which may or may not have descended from the same ancestral allele at $t = 2$. And, again, the probability that they are descended from the same ancestral allele at $t = 2$ is $1/N$, in which case the alleles represent cousins. We can write all of these possibilities in the form of a decision tree (Figure 13.17).

From this decision tree, we can calculate the probability that the sampled alleles coalesce at generation t, counted backward in time. The probability of coalescence in the current generation is zero, $P(T_2 = 0) = 0$, because we know we sampled two different alleles. We have already figured out that the probability of coalescence at time $t = 1$ is $P(T_2 = 1) = 1/N$. The probability that the coalescent event occurs at time $t = 2$ is $P(T_2 = 2) = (1 - 1/N)\ 1/N$, which equals the probability that the alleles had not coalesced at time $t = 1$ (contributing the $1 - 1/N$ term) but did coalesce at time $t = 2$. In general, the probability that the two alleles coalesce t generations in the past is given by

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}. \tag{13.13}$$

Expression (13.13) is what we are after—it is the probability distribution for the time to coalescence for two alleles, T_2. A comparison of equation (13.13) with Definition (P3.7) reveals that the waiting time for the coalescence of two

alleles has a geometric probability distribution with parameter $p = 1/N$. As a result, we can immediately use the properties of the geometric distribution to infer that the mean time until coalescence of two alleles in a haploid population of size N is $E[T_2] = 1/p$ or N generations. Thus, the larger the population, the less likely it is that our two sampled individuals are close relatives. We also know that a geometric random variable has a variance $(1 - p)/p^2$ (see Table P3.2), from which we can calculate the variance in coalescent times as $N(N - 1)$. Thus, the exact time of coalescence is extremely variable.

If it takes N generations, on average, for a sample of two alleles to coalesce, you might think that it would take an incredibly long time for a sample of n alleles to coalesce into one allele. Our next aim is to find the probability distribution for the time, M_n, until the most recent common ancestor (MRCA) of n alleles sampled from a population of size N. The random variable M_n can be viewed as the sum of several independent random variables representing the time until each coalescent event: the time until the n sampled alleles coalesce into $n - 1$ alleles, plus the time that the $n - 1$ alleles coalesce into $n - 2$ alleles, etc., until only one allele remains. Defining T_i as the time until i alleles coalesce into $i - 1$ alleles, $M_n = T_n + T_{n-1} + \cdots + T_2$. We have already calculated the probability distribution for T_2; now we need to find the probability distribution for T_i when there are more than two alleles.

When there are i sampled alleles, the probability that none of them coalesce in the previous generation is given by the probability that each allele descends from different parents. The first allele must descend from some parent allele (probability = 1), the second allele must descend from a different parent allele from the first (probability = $1 - 1/N$), the third allele must descend from a different parent allele than either the first or the second (probability = $1 - 2/N$), etc. Writing $p(i)$ as the probability that there is at least one coalescent event in the preceding generation, the probability that there is *not* a coalescent event, $1 - p(i)$, is the probability that all i alleles descended from different parent alleles:

$$1 - p(i) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{i-1}{N}\right). \tag{13.14}$$

Standard coalescent theory makes the assumption that the population size N is large relative to the sample size i, and then approximates (13.14) using a Taylor series (Recipe P1.2, Primer 1). Assuming $1/N$ to be small and ignoring terms that are $O(1/N^2)$, equation (13.14) becomes

$$\begin{aligned} 1 - p(i) &\approx 1 - \frac{1}{N} - \frac{2}{N} \cdots - \frac{i-1}{N} \\ &= 1 - \frac{1}{N}\sum_{j=1}^{i-1} j \\ &= 1 - \frac{i(i-1)}{2N}, \end{aligned} \tag{13.15}$$

where Rule A1.18 is used to evaluate the sum. Therefore, the probability of at least one coalescent event is $p(i) = i(i - 1)/(2N)$, which is sometimes written

using the binomial coefficient as $p(i) \approx \binom{i}{2}/N$ (see Box P3.1). Technically, $p(i)$ describes the probability of *one or more* coalescent events in the preceding generation, but it is very unlikely that more than one coalescent event occurs when the population size is much larger than the sample size ($N >> i$). In this case, the probability that i alleles are descended from $i - 1$ alleles is very nearly equal to $p(i)$.

Following the same logic leading up to equation (13.13), the probability that it takes t generations for i alleles to coalesce into $i - 1$ alleles (represented by the random variable T_i) is

$$P(T_i = t) = (1 - p(i))^{t-1}\, p(i). \tag{13.16}$$

Again, this is a geometric distribution, now with parameter $p(i)$. As a result, the mean time until i alleles coalesce to $i - 1$ alleles is $1/p(i)$, or $2N/(i\ (i - 1))$ generations.

Given the above results, the time until the most recent common ancestor of n alleles, M_n, is given by the sum of several geometrically distributed random variables, T_i, each with their own parameter $p(i)$. Unfortunately, the probability distribution for a random variable given by the sum of different geometric random variables is not known in any simple form. Nevertheless, we can derive the expected (or mean) time until the MRCA for n alleles: $E[M_n] = E[T_n + T_{n-1} + \cdots + T_2]$ as $E[M_n] = E[T_n] + E[T_{n-1}] + \cdots + E[T_2]$, because the expectation of a sum equals the sum of the expectations (Table P3.1). Therefore, the expected time until the MRCA of n alleles is

$$\begin{aligned} E[M_n] &= \frac{2N}{n(n-1)} + \frac{2N}{(n-1)(n-2)} + \cdots + N \\ &= 2N \sum_{i=2}^{n} \frac{1}{i\,(i-1)} = 2N\,\frac{n-1}{n}. \end{aligned} \tag{13.17}$$

The last sum in (13.17) can be evaluated by induction (see Problem 13.8).

Result (13.17) is pretty amazing. The average time until the MRCA of all n alleles in a sample is less than twice the average time until the ancestor of only two alleles, no matter how large the sample. When there are lots of alleles, not much time passes before a coalescent event takes places because there are many pairs of alleles that could potentially coalesce.

Now that we have described the probability distribution for coalescent times, we can simulate these coalescent events to give us a better feeling for the ways in which a sample of alleles are likely to be related. To carry out these simulations starting with n alleles, we draw a random number from a geometric distribution with parameter $p(n)$ to get the time frame over which there remain n distinct alleles within the population. At this randomly drawn time, a coalescent event occurs, and we join together the branches for two alleles. We then repeat the process for the remaining $n - 1$ alleles, until we reach the MRCA. We carried out this coalescent simulation starting with a sample of ten alleles in Figure 13.18, repeating the process four times to illustrate the variability in tree length and tree shape. Notice that coalescent events occur more rapidly near the present ($t = 0$) because there are more pairs of alleles that can potentially

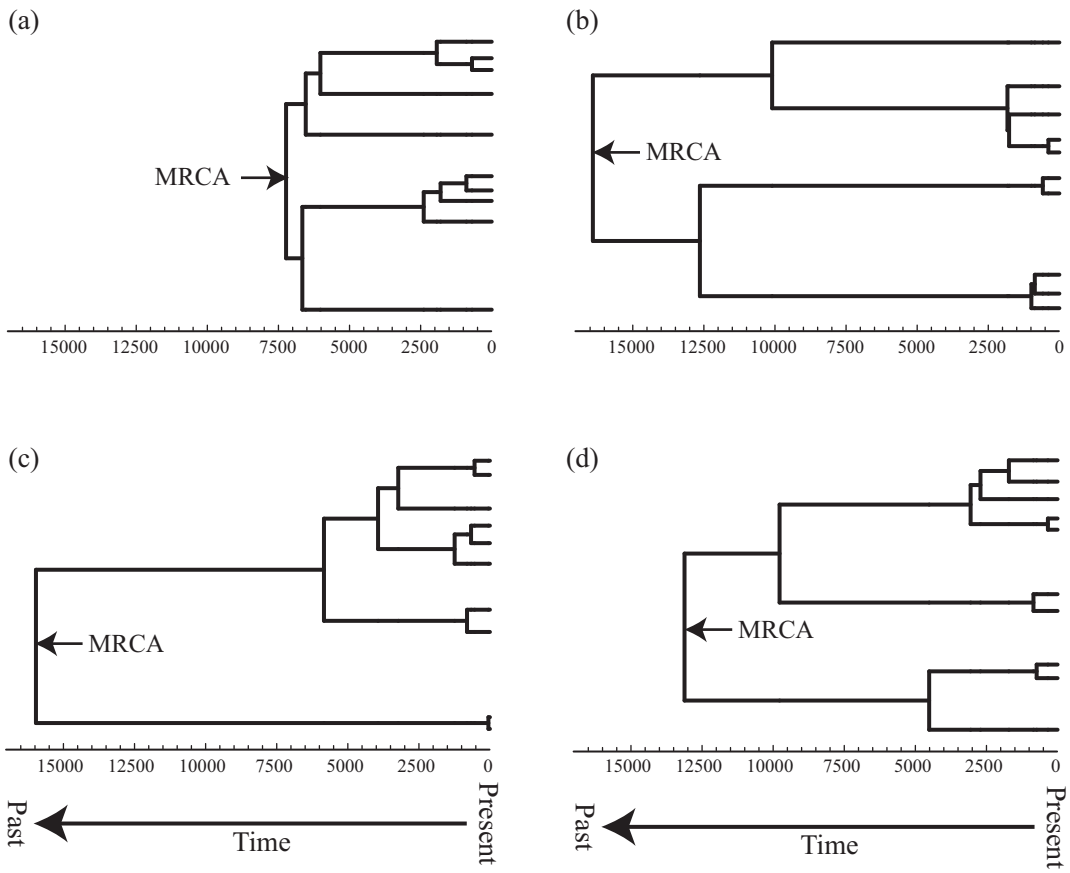

Figure 13.18: Coalescent simulations. Independent simulations of the coalescent process are illustrated in each panel. Starting at the present ($t = 0$, far right) with a sample of $n = 10$ alleles in a population of size $N = 10{,}000$, the time until a coalescent event was determined by randomly drawing times from the geometric distribution given by equation (13.16). At that point, two branches were randomly merged. The process was repeated until only one lineage remained (the most recent common ancestor, MRCA).

coalesce and that the final coalescent event between two alleles is, on average, the longest one.

The coalescent thus provides us with a description of the types of phylogenetic trees expected under the null model of a population of constant size in the absence of selection. While this description is interesting in and of itself, the real value of coalescent theory is that mutation can be overlaid on top of the phylogenies to describe patterns expected within a sample of DNA sequences. Each time a mutation occurs, it alters the DNA sequence in that individual and all of its descendants.

If mutations occur continuously over time at a constant rate, mutations can be imagined as raining down on the phylogenies drawn in Figure 13.18. The total number of mutations would then follow a Poisson distribution with a mean equal to the mutation rate times the total amount of time represented by all of the branches on the tree. For shorter trees (e.g., Figure 13.18a), we would thus expect fewer mutations and less genetic variability among the sampled sequences than in longer trees (e.g., Figure 13.18b).

Alternatively, if mutations occur only at discrete points in time (e.g., at meiosis), then the total number of mutations that occur in the history of a sample would follow a binomial distribution with parameters equal to the mutation rate and the total number of events (meioses) throughout all of the branches in the tree. Fortunately, because the binomial distribution converges upon a Poisson distribution when events are rare and when there are a large numbers of trials (see section P3.3.6 and Appendix 5), it makes little difference whether we model mutations as arising continually over time or at specific points in the life cycle (e.g., meiosis). Here, we assume that mutation is a continuous process, as is more common in coalescent theory.

To give you a flavor of the sorts of results that can be generated using coalescent theory, we will illustrate how to calculate two important quantities: (i) the probability that two sampled alleles are genetically identical and (ii) the number of segregating sites in a sample of n alleles. Throughout, we assume that mutations occur at rate μ per generation per sequence and that each mutation is unique (i.e., changes a different base pair within the sequence).

Let us consider the first question—what is the probability that two sampled alleles are identical? This question is impossible to answer without knowing how related the alleles are. Fortunately, the probability distribution (13.13) tells us the probability that the two alleles last shared a common ancestor at time t in the past, P(coalescence at time t). But how do we rewrite the probability that the alleles are identical using this information? The answer is to use the law of total probability (Rule P3.8). Because the coalescent times represent a set of mutually exclusive events, we can rewrite the probability that the two alleles are identical by summing the conditional probability that the alleles are identical given their coalescent time, over all coalescent times:

$$P(\text{alleles identical}) = \sum_{t=1}^{\infty} P(\text{alleles identical}|\text{coalescence at time } t)P(\text{coalescence at time } t).$$

The benefit of this expression is that we have now broken down the probability into pieces that are easier to calculate. The probability of coalescence at time t is given by (13.13): P(coalescence at time t) = $(1 - 1/N)^{t-1}$ $(1/N)$. And P(alleles identical | coalescence at time t) equals the probability that no mutations occur along either of the two branches leading from the common ancestor to the two sequences, amounting to a total branch length of $2t$ if they coalesced at time t. The probability of no mutations in this time period is given by the probability of drawing no events ($k = 0$) from a Poisson distribution with an expected number of mutations of $2\mu t$ (see Definition P3. 6). This probability is P(alleles identical | coalescence at time t) = $e^{-2\mu t}$. Summing over all coalescent times, the probability that the two sequences are identical is

$$\begin{aligned} P(\text{alleles identical}) &= \sum_{t=1}^{\infty} e^{-2\mu t}\left(1 - \frac{1}{N}\right)^{t-1}\frac{1}{N} \\ &= \frac{1}{1 + (e^{2\mu} - 1)N}. \end{aligned} \tag{13.18a}$$

The sum in the first line of (13.18a) can be interpreted as a constant (here $e^{-2\mu}$) raised to the power of the random variable (here t), and averaged over its probability distribution (here the geometric distribution). This sum defines the moment generating function of a distribution (see Appendix 5). Thus, rather than having to simplify this summation from scratch, we can use the moment generating function of the geometric distribution (see Table P3.2 with $z = -2\mu$) to obtain the second line in (13.18a).

Equation (13.18a) can be simplified further by assuming that the mutation rate is small but that $N\mu$ is not small. First, (13.18a) is rewritten in terms of $\theta = 2N\mu$, the expected number of differences between two sequences in a haploid population (see Problem 13.9), by replacing N with $\theta/(2\mu)$. Next, the limit of (13.18a) is taken as the mutation rate goes to zero but θ is held constant (see Appendix 2). The probability of identity given by (13.18a) is then very nearly equal to

$$P(\text{alleles identical}) = \frac{1}{1 + 2N\mu}. \tag{13.18b}$$

This result makes qualitative sense. Two alleles will be more similar when P(alleles identical) is near one, which requires a low mutation rate and/or a small population size, so that the average coalescent time is short.

Next let us consider the second question—how many nucleotide sites in the DNA are likely to vary within a sample of n alleles? Any mutation that occurs in the history of the sample since the most recent common ancestor will cause a nucleotide difference between some individuals in the sample. We refer to this polymorphic nucleotide as a *segregating site*. Thus the number of segregating sites in the sample is equal to the number of mutations that occur over all of the branches of the tree. The number of segregating sites, S, is a random variable because mutations occur randomly over the tree and the length of the tree is determined by the random coalescence times. In general we might attempt to derive the probability distribution for S, but here we focus only on its expected value (i.e., the mean number of segregating sites, $E[S]$). Again, it would be impossible to calculate the number of segregating sites without knowing how much time has passed along the lineages leading to the present day sample from their most recent common ancestor. To proceed, we condition on the total length of a tree, L, summed over all branches and use the law of total expectation (Rule P3.9) to rewrite $E[S]$ as

$$E[S] = \sum_l E[S|L = l]\, P(L = l). \tag{13.19}$$

Equation (13.19) decomposes our problem into smaller pieces that can be evaluated.

The term $E[S \mid L = l]$ is the expected number of segregating sites given that the total tree length is l generations, which will be a Poisson random variable with mean $\mu\, l$. Plugging this result into equation (13.19), the expected number of segregating sites becomes $E[S] = \mu \sum_l l\, P(L = l)$, or just $E[S] = \mu\, E[L]$ (see Definition P3.2).

Finally, we need to calculate $E[L]$. We already know that the coalescent time when there are i distinct alleles is geometrically distributed with mean $2N/(i\,(i-1))$ generations. The sum of the branch lengths within a phylogenetic tree while there are i alleles is thus, on average, $2N/(i\,(i-1))$ multiplied by i, the number of branches in the tree during this period. Summing over all possible numbers of branches, i, the mean total length of a tree, $E[L]$, is

$$E[L] = \sum_{i=2}^{n} i \frac{2N}{i\,(i-1)} = 2N \sum_{i=2}^{n} \frac{1}{(i-1)} = 2N \sum_{i=1}^{n-1} \frac{1}{i}. \tag{13.20}$$

The expected number of segregating sites within a sample is therefore given by

$$E[S] = \mu\, E[L] = 2N\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}.$$

Interestingly, both the probability of identity and the expected number of segregating sites depend on the same quantity $\theta = 2N\mu$, which measures the expected difference between two sequences under the Wright-Fisher model in a haploid population (see Problem 13.9). (In a diploid population, $\theta = 4N\mu$.) The fact that these quantities should be related to one another was used by Tajima (1983) to test the null hypothesis that genes have been evolving neutrally, without selection (Tajima's D statistic).

The power of coalescent theory is that one can obtain the expected values of various properties (like the number of segregating sites) and compare them against data. You might wonder, however, whether it is reasonable to assume that the Wright-Fisher model is correct. Indeed, one way to think about coalescent theory is that it provides us with a null hypothesis that should hold if the history of the sample involved nothing other than neutral sampling in a population of constant size. If the patterns within the sampled sequences do not match these expectations, then we infer that some other process is happening. This other process might have been selection, but it might also have been changes to the population size and/or migration. Although it is difficult to extend coalescent theory to describe selection (see Krone and Neuhauser 1997; Neuhauser and Krone 1997), the theory has been extended in various ways to account for changing population size, migration, founder events, etc. (Felsenstein 2004; Hudson and Kaplan 1995; Rosenberg and Nordborg 2002).

13.9 Concluding Message

In this chapter we have described how many of the classic deterministic models introduced in Chapter 3 can be extended to incorporate stochasticity. Discrete-time models of exponential and logistic growth were extended in section 13.2 to allow for variation in family size. Continuous-time models of population growth were extended in section 13.3, using a birth-death model that allows for replication and death at random points in time. Population-genetic models of allele frequency change were extended in section 13.4 using the classic Wright-Fisher model (where reproduction is simultaneous) and in section 13.5 using the Moran model (where only one individual reproduces at a time).

In addition to these classic models, we developed three other stochastic models, chosen to illustrate different ways in which chance events can be modeled. The model of retinoblastoma, a cancer of the eye, describes the stochastic way in which mutations arise during development. The cellular automaton model of extinction-recolonization describes the stochastic way in which populations spread over space, as well as the stochastic nature of extinction. Finally, the coalescent model describes the stochastic way in which individuals are related to one another. In each of these models, chance plays a key role in determining the outcome, as we explored through the use of probability theory (Primer 3) and simulation. In the next two chapters, we describe methods of analyzing such stochastic models that allow us to draw more general conclusions than are possible from simulations alone.

Problems

Problem 13.1: In the text, we considered an environment that varies over time between a good state and a bad state, where the probability of being in either state at time $t + 1$ is given by

$$\begin{pmatrix} P_g(t+1) \\ P_b(t+1) \end{pmatrix} = \begin{pmatrix} p_{gg} & p_{gb} \\ p_{bg} & p_{bb} \end{pmatrix} \begin{pmatrix} P_g(t) \\ P_b(t) \end{pmatrix}$$

Generalize this model by allowing three states: good (g), bad (b), and recovering (r). Assume that if the environment is bad, it either remains bad or enters the recovering state, that if the environment is recovering, it either continues to recover or enters the good state, and that if the environment is good, it either remains good or turns bad. No other transitions are possible. Write down the matrix equation describing these transition probabilities.

Problem 13.2: Here you will simulate the model of exponential growth with environmental and demographic stochasticity when the environment is correlated from year to year. As in Figure 13.4, assume that there are good ($R_g = 1.5$) and bad ($R_b = 0.5$) environments. Now, if the current environment is good, assume that it remains good with probability 0.85, while if the current environment is bad it remains bad with probability 0.65:

$$\begin{pmatrix} P_g(t+1) \\ P_b(t+1) \end{pmatrix} = \begin{pmatrix} 0.85 & 0.35 \\ 0.15 & 0.65 \end{pmatrix} \begin{pmatrix} P_g(t) \\ P_b(t) \end{pmatrix}$$

These numbers were chosen so that, over the long run, the environment is good 70% of the time as in Figure 13.4, but now the environment has a higher probability of remaining in the same state for several years in a row (given by the diagonal elements). (a) Generate figures for these simulations as in Figure 13.4. (b) Simulate the model 100 times and count how often the population goes extinct within 30 generations. [Hint: Use the state of the environment in the previous generation and the transition matrix to obtain the probability that the environment will be good in the next time step, p. Use this p value as we did in Figure 13.4.]

Problem 13.3: Alter the birth-death model of population growth to allow death rates to depend on the current population size, *i,* rather than birth rates. (a) Determine the transition probabilities (see equation (13.3) for density-dependent birth rates). (b) According to your answer to (a), at what population size is there the lowest probability of dying per time? Does this make sense?

Problem 13.4: The Wright-Fisher model can be extended to describe a population of N diploid individuals rather than N haploid individuals if it is assumed that alleles can come from any parent and are united at random in the offspring. (a) Write down the equivalent to equation (13.4) for a diploid population with N individuals. (b) Write down the equivalent to equation (13.13) for a diploid population with N individuals. (c) Infer the average time to coalescence for a pair of alleles drawn from a diploid population. (d) Explain why your answer for the diploid model differs from the expected coalescence time of N generations for a haploid population.

Problem 13.5: (a) For the Moran model (13.8), write p_{ji} in the form of a transition probability matrix for a population of size $N = 4$. Confirm that the columns sum to one. (b) Repeat for a population of any size N by filling in the "—" entries in the matrix

$$\begin{pmatrix} — & — & — & \cdots & — \\ — & — & — & \cdots & — \\ — & — & — & \cdots & — \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ — & — & — & \cdots & — \end{pmatrix}.$$

The transition probability matrix for the Moran model is tridiagonal, with zeros everywhere except the entries on, immediately above, or immediately below the diagonal.

Problem 13.6: Modify equations (13.9) for the Moran model of allele frequency change to take into account mutation. Assume that mutations occur only during reproduction (births) and that there is a probability μ that allele A mutates to a and a probability ν that allele a mutates to A.

Problem 13.7: In Figure 13.19, we show an empty cell (A) surrounded by cells in a cellular automaton (dark cells are occupied, white cells are empty). A monsoon causes the top left cell to go extinct in the current generation. What is the probability that the central cell is recolonized at the next time step if (a) the eight nearest neighbors serve as a migrant source and extinction occurs before migration, (b) the eight nearest neighbors serve as a migrant source and migration occurs before extinction, and (c) the four nearest neighbors serve as a migrant source and extinction occurs before migration. [Define m and e as in the derivation of (13.12) in the text.]

Problem 13.8: Prove that $\sum_{i=2}^{n} 1/(i\,(i-1)) = (n-1)/n$, which was needed to derive the average time until the most recent common ancestor of a sample of n alleles,

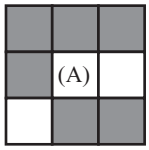

Figure 13.19: Recolonization from neighboring sites

equation (13.17). (a) Calculate and simplify the sum $\sum_{i=2}^{n} 1/(i\,(i-1))$ for $n = 2$, 3, and 4. The resulting pattern suggests that $\sum_{i=2}^{n} 1/(i\,(i-1)) = (n-1)/n$. (b) Assume that this equation holds true for $n - 1$, so that $\sum_{i=2}^{n-1} 1/(i(i-1)) = (n-2)/(n-1)$. Show that you can add $1/(n\,(n-1))$ to both sides of the equation to prove that the equation also holds true for n. Because the equation is correct for $n = 2$ and remains true every time n is increased by one, you have proven by induction that $\sum_{i=2}^{n} 1/(i\,(i-1)) = (n-1)/n$.

Problem 13.9: Calculate the average number of differences between two sequences, $E[D]$, using coalescent theory when mutations occur continuously over time at a rate μ per generation per sequence length. (a) Rewrite $E[D]$ in terms of $E[\mathrm{D} \mid$ coalescence at time $t]$ using the law of total expectation (Rule P3.9). (b) Calculate $E[\mathrm{D} \mid$ coalescence at time $t]$. (c) Use the probability distribution for coalescent times, P(coalescence at time t) $= (1 - 1/N)^{t-1}\, 1/N$, as well as your answers to parts (a) and (b) to determine the average number of differences between two sequences, $E[D]$.

Further Reading

For more information on the mathematical underpinnings of stochastic models, see

- Taylor, H. M., and S. Karlin. 1998. *An Introduction to Stochastic Modeling*. Academic Press, San Diego.
- Allen, L.J.S. 2003. *An Introduction to Stochastic Processes with Applications to Biology*. Pearson/prentice Hall, Upper Saddle River, N.J.

For more examples of stochastic models in ecology, see

- Renshaw, E. 1991. *Modelling Biological Populations in Space and Time*. Cambridge University Press, Cambridge.
- Nisbet, R. M., and W.S.C. Gurney. 1982. *Modelling Fluctuating Populations*. Wiley, Chichester.
- Hubbell, S. P. 2001. *The Unified Neutral theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, N.J.

For more examples of stochastic models in evolution, see

- Ewens, W. J. 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin.
- Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row, New York.

References

Bjørnstad, O. N., and B. T. Grenfell. 2001. Noisy clockwork: Time series analysis of population fluctuations in animals. *Science* 293:638–43.

Bron, A. J., R. C. Tripathi, B. J. Tripathi, and E. Wolff. 1997. *Wolff's Anatomy of the Eye and Orbit*. Chapman & Hall Medical, London.

Chen, D., I. Livne-bar, J. Vanderluit, R. Slack, M. Agochiya, and R. Bremner. 2004. Cell-specific effects of RB or RB/p107 loss on retinal development implicate an intrinsically death resistant cell-of-origin in retinoblastoma. *Cancer Cell* 5:539–551.

Deangelis, D. L., and L. J. Gross (eds.). 1992. *Individual-Based Models and Approaches in Ecology: Populations, Communities, and Ecosystems*. Chapman and Hall, New York.

Dreher, Z., S. R. Robinson, and C. Distler. 1992. Müller cells in vascular and avascular retinae: A survey of seven mammals. *J. Comp. Neurol.* 323:59–80.

Edwards, A., H. A. Hammond, L. Jin, C. T. Caskey, and R. Chakraborty. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12:241–253.

Ewens, W. J. 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Frost, S.D.W., M.-J. Dumaurier, S. Wain-Hobson, and A.J.L. Brown. 2001. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. *Proc. Natl. Acad. Sci. U.S.A.* 98:6975–6980.

Gardner, M. 1983. *Wheels, Life, and other Mathematical Amusements*. W. H. Freeman, New York.

Geiger, H. J., W. W. Smoker, L. A. Zhivotovsky, and A. J. Gharrett. 1997. Variability of family size and marine survival in pink salmon (Oncorhynchus gorbuscha) has implications for conservation biology and human use. *Can. J. Fish. Aquat. Sci.* 54:2684–2690.

Harvey, P. H., E. C. Holmes, A. Ø. Mooers, and S. Nee. 1994. Inferring evolutionary processes from molecular phylogenies. Pp. 313–333 in R. W. Scotland, D. J. Siebert and D. M. Williams, eds. *Models in Phylogeny Reconstruction*. Clarendon Press, Oxford.

Huang, S., and S. A. Moody. 1993. The retinal fate of Xenopus cleavage stage progenitors is dependent upon blastomere position and competence: Studies of normal and regulated clones. *J. Neurosci.* 13:3193–3210.

Hudson, R. R., and N. L. Kaplan. 1995. The coalescent process and background selection. *Philos Trans. R. Soc. London, Ser. B* 349:19–23.

Kaplan, D., and D. Glass. 1995. *Understanding Nonlinear Dynamics*. Springer-Verlag, New York.

Knudson, A. G. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U.S.A.* 68:820–823.

Knudson, A. G. 1993. Antioncogenes and human cancer. *Proc. Natl. Acad. Sci. U.S.A.* 90:10914–10921.

Kojima, K., and T. M. Kelleher. 1962. The survival of mutant genes. *Am. Nat.* 96:329–343.

Krone, S. M., and C. Neuhauser. 1997. Ancestral Processes with Selection. *Theor. Popul. Biol.* 51:210–37.

Loeuille, N., and M. Ghil. 2004. Intrinsic and climatic factors in North-American animal population dynamics. *BMC Ecol.* 4:6.

Lohmann, D. R. 1999. RB1 gene mutations in retinoblastoma. *Hum. Mutat.* 14:283–8.

Luria, S. E., and M. Delbruck. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.

Moffett, D. F., S. B. Moffett, and C. L. Schauf. 1993. *Human Physiology—Foundations and Frontiers*. Mosby, St. Louis.

Moran, P. 1962. *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. London Ser. B* 344:77–82.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1995. Estimating extinction from molecular phylogenies. Pp. 164–182 in J. L. Lawton and R. M. May, eds. *Extinction Rates*. Oxford University Press, Oxford.

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. *Philos. Trans. R. Soc. London, Ser. B* 344:305–311.

Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples in models with selection. *Genetics* 145:519–534.

Nisbet, R. M., and W. S. C. Gurney. 1982. *Modelling Fluctuating Populations*. Wiley, New York.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22:201–204.

Press, W. H. 2002. *Numerical Recipes in C++ : The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1992. *Numerical recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Purvis, A., S. Nee, and P. H. Harvey. 1995. Macroevolutionary inferences from primate phylogeny. *Proc. R. Soc. London, Ser. B* 260:329–333.

Renshaw, E. 1991. *Modelling Biological Populations in Space and Time*. Cambridge University Press, Cambridge.

Rosenberg, N. A., and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3:380–390.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133:737–749.

Van Driel, D., J. M. Provis, and F. A. Billson. 1990. Early differentiation of ganglion, amacrine, bipolar, and Muller cells in the developing fovea of human retina. *J. Comp. Neurol.* 291:203–219.

Vogel, F., and R. Rathenberg. 1975. Spontaneous mutation in man. *Adv. Hum. Genet.* 5:223–318.

Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philos. Trans. R. Soc. London, Ser. B* 213:21–87.