CHAPTER 7
Population Genetics Models

One theory of evolution holds that favourable mutations are relatively rare while in contrast selectively neutral mutations are common and account for much of the diversity between individuals observed at the molecular level. In this chapter a stochastic model is discussed which provides some insight into the behaviour of a population subject to recurrent neutral mutation. The model, introduced below, is closely related to the migration processes of Chapters 2 and 6 and the invasion model of Chapter 5, but the major motivation for its inclusion is the use made of reversibility in Section 7.2 to elucidate some of its properties.

7.1 NEUTRAL ALLELE MODELS

Consider a population of M individuals in which the individuals are of various genetic (or allelic) types. Suppose there are J types (or alleles) altogether, and let n_j be the number of individuals of allelic type j. The mechanism by which the population reproduces is as follows. Individuals die at rate μ. When a death occurs an individual, chosen at random from amongst the remaining $M-1$ individuals, gives birth. The offspring is of the same allelic type as the parent with probability $1-u$ and is a mutation of a different allelic type with probability u. When a mutation occurs the mutant individual is equally likely to have any of the other $J-1$ allelic types, excluding his parent's type. In this model the population size remains constant at M and the alleles are neutral, in the sense that an individual's type does not affect his ability to survive or to produce offspring.

It follows from the above description that the Markov process $\mathbf{n} = (n_1, n_2, \ldots, n_J)$ has transition rates

$$q(\mathbf{n}, T_{jk}\mathbf{n}) = \mu \frac{n_j}{M} \frac{n_k(1-u)+(M-n_k-1)u/(J-1)}{M-1}$$

These are of the form (6.2), and hence the equilibrium distribution is given by Theorem 6.1. Indeed these transition rates are a special case of the form (6.9), and so from expression (6.10) it follows that the equilibrium distribution is

$$\pi(\mathbf{n}) = \binom{-Jf}{M}^{-1} \prod_{j=1}^{J} \binom{-f}{n_j} \tag{7.1}$$

where

$$f=\frac{(M-1)u}{J(1-u)-1} \tag{7.2}$$

The above model is of more interest when there are an infinite number of alleles, so that when a mutation occurs the mutant individual has a completely new allelic type, never before represented in the population. Letting $J\to\infty$ in expression (7.1) causes problems: the probability that any *given* allele is present in the population will tend to zero. It is more helpful to describe the population by the process $\mathbf{M}=(M_1, M_2, \ldots, M_M)$ where M_i is the number of alleles represented in the population by i individuals. Thus

$$\sum_{i=1}^{M} iM_i = M \tag{7.3}$$

It follows from the distribution (7.1) that the equilibrium distribution for the process $\mathbf{M}$ is

$$\pi_M(\mathbf{M})=\frac{J!}{M_1!\,M_2!\cdots M_M!\,(J-\sum M_i)!}\binom{-Jf}{M}^{-1}\binom{-f}{1}^{M_1}\binom{-f}{2}^{M_2}\cdots\binom{-f}{M}^{M_M} \tag{7.4}$$

Now let $J\to\infty$ with u held constant; then from (7.2)

$$Jf\to\frac{(M-1)u}{1-u}=\nu \tag{7.5}$$

say. In the limit the form (7.4) becomes (Exercise 7.1.3)

$$\pi_M(\mathbf{M})=\binom{\nu+M-1}{M}^{-1}\prod_{i=1}^{M}\left(\frac{\nu}{i}\right)^{M_i}\frac{1}{M_i!} \tag{7.6}$$

for $\mathbf{M}$ satisfying (7.3). We shall call the resulting process $\mathbf{M}$ the *infinite alleles model.*

The distribution (7.6) is strikingly similar to the equilibrium distribution for the family size process discussed in Section 2.4, and we shall now show that the relationship is not coincidental. Consider a family size process in which individuals with a new allelic type join the population at rate $\lambda\nu$, individuals give birth to new individuals of the same allelic type at rate λ, and individuals die at rate μ (this is a slight change from the process considered in Section 2.4: we have replaced ν by $\lambda\nu$). It follows from the discussion contained in Section 2.4 that if M_i is the number of alleles represented in the population by i individuals then $\mathbf{M}=(M_1, M_2, \ldots)$ is reversible with equilibrium distribution

$$\pi(\mathbf{M})=\prod_{i=1}^{\infty}e^{-\alpha_i}\frac{\alpha_i^{M_i}}{M_i!}$$

where

$$\alpha_i = \frac{\nu}{i}\left(\frac{\lambda}{\mu}\right)^i$$

provided $\lambda < \mu$. Thus $M_1, M_2, \ldots$ are independent, each with a Poisson distribution. Now suppose that we truncate the process $\mathbf{M}$ by forbidding transitions which would cause the total number of individuals alive to drop below $M-1$ or rise above M. Then Corollary 1.10 shows that the equilibrium distribution will have the form

$$\pi(\mathbf{M}) = B\prod_{i=1}^{M}\frac{\alpha_i^{M_i}}{M_i!} \tag{7.7}$$

for $\mathbf{M}$ such that $\sum iM_i = M-1$ or M. How does the process $\mathbf{M}$ behave when its state space is truncated in this way? Well, when the population size is M any particular existing individual dies with probability intensity μ. When the population size is $M-1$ any particular existing individual gives birth to another individual of the same type with probability intensity λ, and with probability intensity $\lambda\nu$ a mutant individual of a new allelic type is born. The proportion of individuals born which are mutations is $\nu/(\nu+M-1)$, which equals u, by relation (7.5). When the population size is $M-1$ the process thus behaves as if each existing individual gives birth at rate $(\nu+M-1)\lambda$, and with probability u the individual born is a mutation. If we now let $\lambda \to \infty$ the births occur immediately after deaths and we obtain the infinite alleles model which led to the distribution (7.6). Distribution (7.7) can be rewritten

$$\pi(\mathbf{M}) = B\left(\frac{\lambda}{\mu}\right)^{\sum iM_i}\prod_{i=1}^{M}\left(\frac{\nu}{i}\right)^{M_i}\frac{1}{M_i!}$$

for $\mathbf{M}$ such that $\sum iM_i = M-1$ or M, and as $\lambda \to \infty$ this approaches the distribution (7.6), as of course it must do.

Often it is not possible to observe the entire population, but only a sample from it. We shall conclude this section by obtaining the sampling distribution for a sample from the infinite alleles model. Say that a set of M individuals has the *description* $\mathbf{M} = (M_1, M_2, \ldots)$ if there are M_i alleles represented by i individuals in the set. The following theorem shows that the equilibrium distribution (7.6) of the infinite alleles model has a rather interesting property.

Theorem 7.1. *Suppose that a random sample of size m is chosen without replacement from a population of size M, $m \le M$. If the population has the description $\mathbf{M}$ with probability $\pi_M(\mathbf{M})$ then the sample has the description $\mathbf{m}$ with probability $\pi_m(\mathbf{m})$.*

Proof. Consider the process **M** described above in which the population size fluctuates between $M-1$ and M. The equilibrium distribution is given by (7.7) and hence the description of the population conditional on the population size being $M-1$ is $\pi_{M-1}(\mathbf{M})$, and conditional on it being M is $\pi_M(\mathbf{M})$. Now when the population size drops from M to $M-1$ the effect is the same as that of choosing a random sample of size $M-1$. Thus when a random sample of size $M-1$ is chosen from a population whose description is **M** with probability $\pi_M(\mathbf{M})$, the description of the sample must be **m** with probability $\pi_{M-1}(\mathbf{m})$. This establishes the theorem for the case $m=M-1$. For general m the theorem follows from the observation that one way to choose a sample of size m is first to choose a sample of size $M-1$, then from this to choose a sample of size $M-2$, and so on until only m individuals are included.

Theorem 7.1 shows that the sampling distribution for a sample of size m from a population of size M is the same as the equilibrium distribution for a population of size m, provided both populations have the same value of ν.

Exercises 7.1

1. Show that the model described at the beginning of this section remains a reversible migration process if the mean lifetime an individual of allelic type j is μ_j^{-1} and if the offspring of an individual of type j is of type k with probability $up_k(1-p_j)$, where $\sum p_j = 1$.
2. A population of size M is divided into J types in accordance with expression (7.1), and a random sample of size m ($\leq M$) is chosen from it. Write down the conditional distribution for the composition of the sample given the composition of the population. Deduce that the sample will divide into J types in accordance with expression (7.1), but with M replaced by m.
3. By considering the coefficients of x^M in the power series expansions of the identity

$$(1-x)^{-\nu} = \prod_{j=1}^{\infty} e^{\nu x^j/j}$$

show that the distribution (7.6) sums to unity. Deduce (7.6) from (7.4).

4. Suppose that **M** is distributed according to expression (7.6). Show that if a random sample of size m has the description **m**, with a particular allele represented by i individuals in the sample, then an individual randomly selected from the remaining $M-m$ members of the population is of that particular allelic type with probability $i/(m+\nu)$.
5. The heterozygosity of a population is defined to be the probability that two distinct individuals, chosen at random from the population, are of different allelic types. Deduce from Theorem 7.1 that the heterozygosity

of a population is $\nu/(\nu+1)$ and that the whole population is of the same allelic type with probability

$$\binom{\nu+M-1}{M-1}^{-1}$$

6. If $\pi_M(\mathbf{M})$ is given by expression (7.6) then Theorem 7.1 has established the following.
   (a) If a random sample of size m is chosen from a set of size M whose description is $\mathbf{M}$ with probability $\pi_M(\mathbf{M})$, then the description of the random sample is $\mathbf{m}$ with probability $\pi_m(\mathbf{m})$.

   Show also that

   (b) If an individual is chosen at random from a set of size M whose description is $\mathbf{M}$ with probability $\pi_M(\mathbf{M})$ and if the individual is found to be of the same allelic type as exactly $M-m-1$ others in the set then the remaining m individuals form a set whose description is $\mathbf{m}$ with probability $\pi_m(\mathbf{m})$.
7. (*Hard*) The preceding exercise has a converse. Suppose that $\pi_M(\mathbf{M})$ is a probability distribution over descriptions of a set of size M and that $\pi_M(\mathbf{M})>0$ for all $\mathbf{M}$ satisfying equation (7.3). Show that if statements (a) and (b) hold for all m and M satisfying $m<M$ then $\pi_M(\mathbf{M})$ must be of the form (7.6).
8. If $\mathbf{M}$ is distributed according to (7.6) with ν an unknown parameter show that the number of alleles in the population, $\sum M_i$, is a sufficient statistic for ν. If only a random sample from the population is observed deduce that the number of alleles present in the sample is sufficient for ν.
9. In the models described in this section it has been assumed that individuals' lifetimes are exponentially distributed with mean μ^{-1}. In fact the equilibrium distributions (7.1) and (7.6) remain unaltered if individuals' lifetimes are arbitrarily distributed. Establish this by considering an infinite server queue at which arrival rates are of the form (3.27) with

$$\Psi(\mathbf{n})=\psi\left(\sum_{j=1}^{J} n_j\right)\prod_{j=1}^{J}(f+n_j-1)!$$

   If lifetimes are exponentially distributed then the number of offspring an individual has is geometrically distributed. If lifetimes are constant show that the number of offspring an individual has is binomially distributed.
10. The infinite alleles model and the family size process can be regarded as special cases of a more general stochastic population model. Let M_i be

the number of families of size i and let $M = \sum iM_i$ be the total population size. Suppose that $\mathbf{M} = (M_1, M_2, \ldots)$ is a Markov process with transition rates

$$\begin{aligned} q(\mathbf{M}, T_{i,i+1}\mathbf{M}) &= iM_i\lambda(M) \qquad i = 1, 2, \ldots \\ q(\mathbf{M}, T_{i,i-1}\mathbf{M}) &= iM_i\mu(M) \qquad i = 2, 3, \ldots \\ q(\mathbf{M}, T_{\cdot 1}\mathbf{M}) &= \nu\lambda(M) \\ q(\mathbf{M}, T_{1\cdot}\mathbf{M}) &= M_1\mu(M) \end{aligned}$$

Thus birth, death, and immigration rates are affected by the total population size. Verify that in equilibrium the process $\mathbf{M}$ is reversible with

$$\pi(\mathbf{M}) = B\left\{\prod_{r=1}^{M} \frac{\lambda(r-1)}{\mu(r)}\right\} \prod_{i=1}^{\infty} \left(\frac{\nu}{i}\right)^{M_i} \frac{1}{M_i!}$$

Show that if a random sample of size m is taken from the population then its description $\mathbf{m}$ has distribution (7.6) with M replaced by m. Show that this property of a random sample remains true if the population size at time $t = 0$ is zero and the functions $\lambda(M)$ and $\mu(M)$ are replaced by functions of time $\lambda(M, t)$ and $\mu(M, t)$. Deduce that the property remains true if $\lambda(M, t)$ and $\mu(M, t)$ are themselves stochastic processes.

11. Suppose the family size process (or more generally the process described in the previous exercise) is used to model the numbers of moths present in a particular location, with each family representing a distinct species of moth. Deduce from Exercise 7.1.4 that if a trap catches a random sample of m moths then the expected number of species caught is

$$\sum_{i=1}^{m} \frac{\nu}{\nu + i - 1}$$

Observe that this depends upon the parameter ν alone, unlike the corresponding relationship based upon the expected value of m, found in Exercise 2.4.7.

12. Consider a population process in which the population size fluctuates between M and M' as follows. After the population size has been M for a certain time a reproduction period is entered, when only births and immigrations are allowed. During this reproduction period each individual gives birth at rate λ to a new member of its family, and immigrants arrive at rate $\nu\lambda$ to found new families. Individuals appearing during the reproduction period are allowed to give birth themselves during that same period, and the period ends when the population size reaches M'. Then a random sample of M individuals is chosen from the

population to form the next generation and the procedure is repeated. Show that if the process **M** is observed in discrete time just before each reproduction period begins then it is reversible with equilibrium distribution (7.6). Show that the same is true if the process is observed immediately *after* each reproduction period, with M replaced by M' in expression (7.6).

13. Amend the model described in the previous exercise so that the reproduction period ends when the population size reaches $2M$, and then suppose the original M individuals die, leaving a new generation of size M. Show that the process **M** observed just before each reproduction period begins is identical to that for the unamended model with M' infinite.
14. Suppose that in the model of the previous exercise individuals appearing during a reproduction period are not allowed to give birth. The resulting model can be viewed as follows. The parent of each individual in the new generation is chosen independently and at random from among the M individuals alive in the previous generation, and each birth may result in a new mutation, mutations arising independently over the M births. Observe that the number of offspring an individual has is binomially distributed and that when an allele appears for the first time it is represented by just one individual in the population. Deduce that the process **M** obtained from this model is not reversible.

7.2 THE AGE OF AN ALLELE

Suppose that a population has been evolving according to the infinite alleles model for a long period. It is observed at time t, and it is found that K alleles are present in the population, M_i of them represented by i individuals, for $i = 1, 2, \ldots$. What does this information tell us about the ages of the K alleles?

We would like to be able to deduce from the reversibility of the process $\mathbf{M}(t)$ that the future of an allele represented in the population is stochastically similar to its past. Unfortunately such a conclusion does not immediately follow, since from a realization of the process $\mathbf{M}(t)$, $-\infty < t < +\infty$, it is not possible to discern the progress between first occurrence and eventual extinction of a particular allele. The problem could be overcome by using the finite allele model of the previous section and a limiting argument, but we shall use an alternative labelling method.

Suppose that the allelic type of each individual in the population is associated with an integer in the range from zero to M. When a non-mutant individual is born its allelic type is that of its parent. When a mutant individual is born its allelic type is not that of its parent, nor that of any of the other alleles present in the population. Since the population size is M

there will be at least one integer in the range from zero to M which can be assigned to the new allele; if there is more than one possible integer the choice is made at random. The integer associated with an allele is not intended to represent any physical characteristic of the allele—it simply labels it. As time progresses the same label will be used repeatedly for different alleles, but note that after an allele becomes extinct an interval will elapse before its label is used again. Describe the state of the population by

$$(\mathbf{M}, \mathbf{l}) = (M_1, M_2, \ldots, M_M;\ l(1,1), l(1,2), \ldots, l(1, M_1);$$
$$l(2,1), l(2,2), \ldots, l(2, M_2);$$
$$\ldots, l(M, M_M))$$

with $l(i,k)$ the label of the kth allele among those represented in the population by i individuals. The effect of the death of an individual whose allelic type had j representatives is to be equivalent to randomly choosing one of the M_j labels from among $l(j,1), l(j,2), \ldots, l(j, M_j)$ and inserting it at random into one of the $M_{j-1}+1$ positions among $l(j-1,1)$, $l(j-1,2), \ldots, l(j-1, M_{j-1})$. A birth is to be dealt with similarly.

Theorem 7.2. *The labelled process* $(\mathbf{M}, \mathbf{l})$ *is reversible and has equilibrium distribution*

$$\pi(\mathbf{M}, \mathbf{l}) = \frac{(M - \sum M_i + 1)!}{(M+1)!} \binom{\nu + M - 1}{M}^{-1} \prod_{i=1}^{M} \left(\frac{\nu}{i}\right)^{M_i} \frac{1}{M_i!}$$

Proof. The detailed balance conditions are easily checked and the result follows from these. Observe that the first term of the equilibrium distribution is the reciprocal of the number of distinct orderings of $\sum M_i$ different labels. In equilibrium $\mathbf{M}$ has the distribution (7.6) and given $\mathbf{M}$ every possible arrangement of labels, $\mathbf{l}$, is equally likely.

A realization of the labelled process $(\mathbf{M}, \mathbf{l})$ for $-\infty < t < \infty$ allows us to trace the history of any particular allele from the point in time when it first appears until it becomes extinct. Consider the problem of estimating, from an observation $\mathbf{M}$ on the process at a particular time t, the age of an allele present in the population. From the reversibility of the process $(\mathbf{M}, \mathbf{l})$ we see that the age of the allele has the same distribution as the time to extinction of that allele. But to calculate the distribution of the time to extinction it is not necessary to have the frequencies of the other alleles in the population: the future frequency of an allele in the population at time t follows a

random walk with transition intensities

$$q(j, j-1) = \mu \frac{j}{M}\left(\frac{M-j}{M-1} + \frac{j-1}{M-1}\, u\right)$$
$$q(j, j+1) = \mu \frac{M-j}{M}\frac{j}{M-1}\,(1-u) \tag{7.8}$$

Thus the age distribution of an allele represented in the population by j individuals can be calculated (Exercise 7.2.1); it depends upon j but not upon the entire state $\mathbf{M}$.

The rest of this section will be devoted to some other consequences of Theorem 7.2, but before we embark on these it is worth clarifying the contribution reversibility makes to the solution of problems concerning the age of an allele. The labelled process $(\mathbf{M}, \mathbf{I})$ arising from some genetic models (e.g. that described in Exercise 7.1.14) is not reversible. Nevertheless, if it is a stationary Markov process then the reversed process obtained from it will also be a stationary Markov process, and given the state $(\mathbf{M}, \mathbf{I})$ at a fixed point in time the age of an allele in the original process will have the same distribution as the time to extinction of that allele in the reversed process. The difficulty is that the reversed process is likely to have extremely complex transition rates. The results obtained in this section follow from the tractability of the reversed process in the case where $(\mathbf{M}, \mathbf{I})$ is reversible.

Corollary 7.3. *If at a given time an allele is represented by j individuals in the population the probability that this allele is the oldest of the alleles then existing is j/M.*

Proof. Suppose that at a given time M different alleles are represented in the population; thus no two individuals have the same allelic type. The probability that a particular allele (*not* individual) will outlive the other $M-1$ alleles is M^{-1}, by symmetry. It follows from this that if at a given time an allele is represented by j individuals in the population the probability that this allele will outlive the other alleles then existing is j/M. The reversibility established in Theorem 7.2 implies that the probability this allele is the oldest is also j/M.

Corollary 7.4. *At a given time the age of the oldest allele is independent of its frequency in the population.*

Proof. The time to extinction of all the alleles currently existing in the population does not depend upon $\mathbf{M}$; it is simply the time a random walk

with transition rates (7.8) takes to reach zero starting from M. The distribution of this time will remain unchanged if we are given which of the currently existing individuals is the ancestor of the last surviving individual with a currently existing allelic type, and so will remain unchanged if we are given which allele will survive the longest. Interchanging times to extinction with ages, as Theorem 7.2 allows us to do, gives the result.

An extension of Corollary 7.4 is given in Exercise 7.2.2.

Corollary 7.5. *In equilibrium the probability that the oldest allele existing is represented by i individuals in the population is, for $i = 1, 2, \ldots, M$,*

$$\frac{\nu}{M}\binom{\nu+M-1}{i}^{-1}\binom{M}{i} \tag{7.9}$$

Proof. Let the random variable x be the frequency of the oldest allele. Theorem 7.2 implies that x has the same distribution as the frequency of the allele which will survive the longest, but this in turn has the same distribution as the frequency of the allelic type of a randomly chosen individual from the population. Thus, using equation (7.6),

$$\begin{aligned}\text{Prob}\{x = i\} &= \sum_{\mathbf{M}} \text{Prob}\{x = i \mid \mathbf{M}\}\pi_M(\mathbf{M}) \\ &= \frac{\nu}{M}\binom{\nu+M-1}{M}^{-1}\binom{\nu+M-i-1}{M-i} \\ &= \frac{\nu}{M}\binom{\nu+M-1}{i}^{-1}\binom{M}{i}\end{aligned}$$

Suppose a sample of size m ($\leq M$) is chosen from the population. We have seen in Theorem 7.1 that the sample has the description $\mathbf{m}$ with probability $\pi_m(\mathbf{m})$. Given the description $\mathbf{m}$ of the sample it is possible to make some deductions about the relative ages of the alleles represented in the sample.

Theorem 7.6. *An allele represented by i individuals in a sample of size m is the oldest allele in the population with probability*

$$\frac{i(\nu+M)}{M(\nu+m)}$$

Proof. We must calculate the probability that an individual chosen at random from the population is of the given allelic type. With probability m/M the randomly chosen individual will belong to the sample, and if this

happens the probability that it is of the given allelic type is i/m. With probability $(M-m)/M$ the randomly chosen individual will be in the $M-m$ members of the population outside the sample, and if this happens the probability that it is of the given allelic type is $i/(m+\nu)$ (Exercise 7.1.4). Thus the probability we are seeking is

$$\frac{m}{M}\frac{i}{m}+\frac{M-m}{M}\frac{i}{m+\nu}=\frac{i(\nu+M)}{M(\nu+m)}$$

Corollary 7.7. *The probability a sample of size m contains the oldest allele in the population is*

$$\frac{m(\nu+M)}{M(\nu+m)}$$

Proof. This follows directly from Theorem 7.6 by summing over all the alleles represented in the sample.

Corollary 7.8. *An allele represented by i individuals in a sample of size m is the oldest allele in the sample with probability i/m.*

Proof. The probability in question is just the probability that an individual chosen at random from the population is of the given allelic type, conditional on the individual chosen having an allelic type which is represented in the sample. The probability that a randomly chosen individual is of the given allelic type is

$$\frac{i(\nu+M)}{M(\nu+m)}$$

and the probability that it is of an allelic type represented in the sample is

$$\frac{m(\nu+M)}{M(\nu+m)}$$

Hence the conditional probability sought is i/m.

Exercises 7.2

1. If an allele is represented in the population by j individuals show that the probability the age of the allele is less than x, $P_j(x)$, satisfies

$$\frac{dP_j(x)}{dx}=q(j,j-1)[P_{j-1}(x)-P_j(x)]-q(j,j+1)[P_j(x)-P_{j+1}(x)]$$

where $q(j,j-1)$ and $q(j,j+1)$ are given by equations (7.8). Obtain a recursion for the probability that there have been exactly a births since the allele first appeared.

2. Suppose that $(f_1, f_2, \ldots, f_K)$ gives the frequencies of the alleles present in the population in order of the ages of the alleles, so that f_1 and f_K are the frequencies of the newest and oldest alleles respectively. Thus $(f_1, f_2, \ldots, f_K)$ contains rather more information than $\mathbf{M}$. Show that the age of the oldest allele is independent of $(f_1, f_2, \ldots, f_K)$.
3. Show that expression (7.9) is increasing or decreasing in i according to whether ν is less than or greater than unity. Show that in equilibrium the expected number of individuals of the oldest allelic type is $(\nu+M)/(\nu+1)$.
4. Establish Corollary 7.8 for a sample from the model of Exercise 7.1.10.
5. If alleles $1, 2, \ldots, k$ are represented in a sample by $n_1, n_2, \ldots, n_k$ individuals respectively, show that the probability allele r is older than allele $r+1$, for $r = 1, 2, \ldots k-1$, is

$$\prod_{r=1}^{k} \frac{n_r}{\sum_{i=r}^{k} n_i}$$

6. A sample of size m is to be taken from a population in equilibrium. Deduce from Corollary 7.8 that the probability the oldest allele in the sample will be represented by i individuals in the sample is given by expression (7.9) with M replaced by m.
7. We have seen that, for the reversible model considered, if we observe the state of the process at a fixed point in time the age of an allele present in the population has the same distribution as the time to extinction of that same allele. Consider now the labelled process $(\mathbf{M}, \mathbf{l})$ arising from the non-reversible genetic model of Exercise 7.1.14; $2M$ labels are required to ensure an interval between successive uses of the same label for different alleles, and it is simplest to suppose that at each point in (discrete) time the labels $l(j, 1), l(j, 2), \ldots, l(j, M_j)$ are the labels of those alleles represented by j individuals in the population arranged in a random order. Use Exercise 1.4.3 to show that if we do *not* observe the state of the process then at a fixed point in time the age of an allele given to be present in the population has the same distribution as the time to extinction of that same allele. Observe that this result follows from the stationarity of the model; the stronger property of reversibility is only required if we are given information about the state of the process at the fixed point in time.

7.3 FIXATION TIMES

If the mutation rate u is low it is quite likely that every individual in the population will be of the same allelic type. Recurrent mutation ensures, however, that no allele can become permanently fixed in the population. Call an allele *quasi-fixed* if it is the only allele present in the population. We

shall begin this section by calculating the probability that an allele becomes quasi-fixed and also the mean time between quasi-fixations of different alleles.

Theorem 7.9. *The probability that a new allele will become quasi-fixed is* Q *where*

$$Q^{-1} = \sum_{i=0}^{M-1} \binom{M-1}{i}^{-1} \binom{\nu+M-1}{i}$$

Proof. When a new allele appears it will be represented by one individual in a population of M. The future frequency of the allele in the population will follow a random walk with transition intensities (7.8); we are interested in the probability that this random walk reaches M before it reaches zero. This can be determined by the standard means for obtaining absorption probabilities or by using the electrical analogue described in Section 5.2. In the electrical analogue the resistance between nodes i and $i+1$ is proportional to

$$\binom{M-1}{i}^{-1} \binom{\nu+M-1}{i}$$

and so if nodes 0 and M are held at potentials of 0 and 1 respectively the potential of node 1 will be Q, establishing the theorem.

When $\nu = 1$ the expression for Q^{-1} becomes $M(1+\frac{1}{2}+\cdots+1/M) \simeq M \log M$. As ν approaches zero, Q^{-1} approaches M.

Corollary 7.10. *The intervals between first quasi-fixations of different alleles are independent and identically distributed with mean* $(QM\mu u)^{-1}$.

Proof. The same allele may become quasi-fixed more than once during its existence, but if we consider only those points in time at which an allele becomes quasi-fixed for the first time then it is clear that the intervening intervals are independent and identically distributed. Let the mean interval length be x. Now the mean period between successive mutations is $(M\mu u)^{-1}$, and so the long-run proportion of alleles which become quasi-fixed must be $(M\mu ux)^{-1}$. But this long-run proportion is Q by the previous theorem, and hence $x = (QM\mu u)^{-1}$.

When a mutation occurs the new allele will generally bear a close resemblance to the allelic type of the mutant's parent. This is because an allele is made up of a large number of smaller units, called nucleotides, and when a mutation occurs it is unlikely to affect more than one of these units.

Suppose then that an allele is made up of an infinite number of nucleotides and that a mutation affects just one of these nucleotides, with no nucleotide ever affected more than once by a mutation. Thus a mutation gives rise to a new allele, which will eventually disappear from the population, and a new nucleotide, which may or may not disappear from the population. Indeed, since the population will eventually consist entirely of the descendents of one of the M individuals alive at a given time the probability that a new nucleotide will eventually be fixed in the population is M^{-1}. We shall devote the rest of this section to the elucidation of this process of gene substitution.

Call the birth of a mutant individual a *determining mutation* if the entire population will eventually be descended from that individual. Thus a determining mutation gives rise to a nucleotide which will eventually become fixed in the population.

Theorem 7.11. *Determining mutations form a Poisson process of rate μu.*

Proof. We started this chapter by obtaining the infinite alleles model as a limiting case of the reversible migration processes of the last chapter. It is also possible to view the infinite alleles model as an invasion process of the form discussed in Chapter 5: regard the M individuals as sites and when an individual gives birth to an offspring regard this as one site invading another. Each time an invasion occurs there is a probability u that the invaded site becomes a completely new colour, corresponding to a mutation. Viewed in this light it is natural to choose an individual alive at time 0 and trace the ancestry of this individual. As we move backwards through time the points in time at which his ancestors were born form a Poisson process of rate μ. Further, the points in time at which an ancestor of his was born a mutant form a Poisson process of rate μu. This is true for each of the M individuals alive at time 0. We thus have M (dependent) Poisson processes; one of them is the sequence of determining mutations up until time 0, but we do not know which one since we cannot tell from which of the M individuals alive at time 0 the population will eventually be descended. Nevertheless, it will be one of these processes and since they are all Poisson processes of rate μu the theorem is proved.

Thus nucleotides destined to be fixed in the population arise as a Poisson process, and the number of nucleotides by which an individual alive at time t differs from his ancestor at time 0 is Poisson with mean μut. The points in time at which nucleotides actually become fixed form a much more complicated point process (Exercise 7.3.4).

Exercises 7.3

1. Show that the mean time between first quasi-fixations tends to infinity when the mutation rate u tends to either zero or infinity (with the population size M held fixed).

2. Show that the expected time till a new nucleotide is fixed or lost is

$$\frac{M-1}{\mu}\sum_{i=1}^{M-1}\frac{1}{i}$$

3. Show that the frequency of a nucleotide destined to be fixed performs a random walk with transition intensities

$$q(i, i-1) = \mu\frac{(i-1)(M-i)}{M(M-1)}$$

$$q(i, i+1) = \mu\frac{(i+1)(M-i)}{M(M-1)}$$

Prove that for a new nucleotide destined to be fixed the expected time till fixation is

$$\frac{(M-1)^2}{\mu}$$

and for a new nucleotide destined to be lost the expected time till loss is

$$\frac{M}{\mu}\sum_{i=2}^{M}\frac{1}{i}$$

4. Observe that arbitrarily many nucleotides may become fixed in the population at the same moment. Show that if $X(t)$ is the number of nucleotides which become fixed in the population in the interval $(0, t)$ then

$$E[X(t)] = \mu u t$$

and

$$\mathrm{Var}[X(t)] = \mu u t + 0(1)$$

where $0(1)$ is a term which remains bounded as $t \to \infty$.

5. Suppose that two distinct individuals are chosen at random from the population. Show that the number of nucleotides by which they differ has the same distribution as $X = Y + Z$ where Y is geometric with mean $(M-2)u$, Z is Bernoulli with mean u, and Y and Z are independent.
6. Suppose that at time 0 a population of size M is subdivided into a number of colonies, which henceforth evolve separately from each other with the original values of μ and u. Show that two individuals chosen at random from distinct colonies at time t differ by a number of nucleotides which has the same distribution as $W + X$ where W has a Poisson distribution with mean $2\mu u t$, X is as in the previous exercise, and W and X are independent.

7. The proof of Theorem 7.11 does not depend upon the reversibility of the genetic model considered. Show that in the non-reversible genetic model of Exercise 7.1.14 the intervals between successive determining mutations are independent and have the same distribution as $Y+1$ where Y has a geometric distribution.