was based on 95 human sequences; the conservative central portion was excluded from the analysis. Obviously, all eight types of transversion occurred with very low frequencies and the average transition/transversion ratio is 15.7, similar to the value (15.0) estimated by Vigilant et al. (1991). Moreover, the transitional rate between pyrimidines (C and T) is higher than that between purines (A and G), and f_{TC} and f_{GA} are, respectively, higher than f_{CT} and f_{AG}, which suggests asymmetry in forward and backward mutation.

PROBLEMS

1. Find a complete mRNA sequence from a mammalian species. Delete the first AT you encounter in the coding region. Determine whether a premature termination codon comes into the reading frame or whether translation will extend beyond the original stop codon.
2. Explain why a frameshift mutation is likely to be more serious if it occurs in a 5′ part than if it occurs in a 3′ part of the coding region.
3. Compute the proportion of synonymous changes among mutations involving single nucleotide changes in the codon ACT under (a) the assumption that mutation is random, i.e., no difference in rate or in preference in the direction of mutation among the four nucleotides, (b) the mutation pattern in Table 1.5 (the values in parentheses), and (c) the mutation pattern in Table 1.6. Note that in parts (b) and (c), the differences in mutability among nucleotides should be taken into account.
4. Repeat problem 3 for the codon GAT.

From: Molecular Evolution
Wen-Hsiung Li, 1997, Sinauer Associates

CHAPTER 2

Dynamics of Genes in Populations

ALL EVOLUTIONARY CHANGES START WITH CHANGES WITHIN POPULATIONS. The study of genetic changes that occur in populations belongs to the domain of population genetics. This chapter reviews some basic principles of population genetics that are essential for understanding molecular evolution; further theoretical background of population genetics will be provided in Chapter 9, in which the maintenance of molecular polymorphism is discussed.

A basic problem in population genetics is to determine how the frequency of a mutant gene will change with time under the influence of various evolutionary forces. Another basic problem is how genetic variability is maintained in natural populations. In addition, from the long-term point of view, it is important to determine the probability that a new mutant will completely replace existing variants in the population and to estimate how fast the replacement will take place. Unlike morphological changes, many molecular changes are likely to have only a small effect on the phenotype of an organism, so the frequencies of molecular variants are subject to strong chance effects. Therefore, chance elements should be taken into account when dealing with molecular evolution. The role of chance effects in evolution is controversial, however, and will be discussed in this chapter and Chapter 9.

CHANGES IN ALLELE FREQUENCIES

The chromosomal or genomic location of a gene is called a **locus**, and alternative forms of the gene at a given locus are called **alleles**. In a population, more than one allele may be present at a locus, and their relative proportions are referred to as the **allele frequencies** or **gene frequencies**. For example, assume that there are two alleles with n_1 and n_2 copies at a certain locus, in a haploid population of size N. Then, their allele frequencies are equal to n_1/N and n_2/N, respectively. Note that $n_1 + n_2 = N$, and $n_1/N + n_2/N = 1$.

Evolution is a process of change in the genetic makeup of populations, with the most basic component being change in allele frequencies with time. In fact, from the evolutionary point of view, a new mutation becomes significant only if

its frequency increases with time and ultimately reaches 1; the mutant gene is then said to have become **fixed** in the population. Without increasing its frequency, a mutation will have but a passing effect on the evolutionary history of the species; the only exception is that it is maintained in the population for a long time by balancing selection. For a mutant allele to increase in frequency, factors other than mutation must come into play. These factors include natural selection, random genetic drift, recombination, and migration.

To understand the process of evolution, we must study how the above factors govern the changes of allele frequencies. In this book, we discuss only natural selection and random genetic drift. In classical evolutionary studies involving morphological traits, natural selection has been considered as the major driving force of evolution. In contrast, random genetic drift is thought to have played an important role in evolution at the molecular level.

There are two mathematical approaches to studying genetic changes in populations: deterministic and stochastic. The **deterministic model** is simpler. It assumes that changes in the frequencies of alleles in a population from generation to generation occur in a unique manner and can be unambiguously predicted from knowledge of initial conditions. Strictly speaking, this approach applies only when two conditions are met: (1) the population is infinite in size and (2) the environment either remains constant with time or changes according to deterministic rules. These conditions are obviously never met in nature, and therefore a purely deterministic approach may not be sufficient to describe the temporal changes in allele frequencies in populations. Random or unpredictable fluctuations in allele frequencies must also be taken into account.

Dealing with random fluctuations requires a different mathematical approach. **Stochastic models** assume that changes in allele frequencies occur in a probabilistic manner, that is, from knowledge of the conditions in one generation we cannot predict unambiguously the allele frequencies in the next generation, but can only determine the probabilities with which certain allele frequencies will be attained. Obviously, stochastic models are preferable over deterministic ones, since they are based on more realistic assumptions. However, deterministic models are much easier to treat mathematically and, under certain circumstances, they yield sufficiently accurate approximations. In the following, we shall deal with natural selection in a deterministic fashion.

NATURAL SELECTION

Natural selection is defined as the differential reproduction of genetically distinct individuals or genotypes within a population. Differential reproduction is caused by differences among individuals in such factors as mortality, fertility, fecundity, mating success, and the viability of offspring. Natural selection is predicated on the availability of genetic variation among individuals in characters related to reproduction. It cannot occur in a population that consists of individuals that do not differ from one another in such traits. Selection leads to changes in allele frequencies over time. However, a mere change in allele frequencies from generation to generation does not necessarily indicate that natural selection is at work. Other processes, such as random genetic drift, can bring about temporal changes in allele frequencies as well (see below).

The **fitness** of a genotype, commonly denoted as w, is a measure of the individual's ability to survive and reproduce. Since the size of a population is usually constrained by the carrying capacity of the environment in which the population resides, the evolutionary success of an individual is determined not by its **absolute fitness**, but by its **relative fitness** in comparison with the other genotypes in the population. In nature, the fitness of a genotype is not expected to remain constant for all generations and under all environmental circumstances. However, by assigning a constant value of fitness to each genotype, we are able to formulate simple theories that are useful for understanding the dynamics of change in the genetic structure of populations brought about by natural selection. In the simplest class of models, we assume that the fitness of an organism is determined solely by its genetic makeup. We also assume that all loci contribute independently to the fitness of the individual, so that each locus can be treated separately.

Most new mutants arising in a population reduce the fitness of their carriers. Such mutations will be selected against and most will be eventually removed from the population. This type of selection is called **negative** or **purifying selection**. Occasionally, a new mutation may be as fit as the best allele in the population. Such a mutation is selectively **neutral**, and its fate is determined not by selection but by chance events. Rarely, a mutant that confers a selective advantage on its carriers may arise. Such a mutation will be subjected to **positive selection**. If the new mutant is advantageous only in heterozygotes but not in homozygotes, the resulting selective regime will be **overdominant selection**.

In the following, we shall consider the case of one locus with two alleles, A_1 and A_2. Each allele can be assigned an intrinsic fitness value; it can be advantageous, deleterious, or neutral. However, this assignment is only applicable to haploid organisms. In diploid organisms the fitness is determined by the interaction between the two alleles at the locus. With two alleles at a locus, there are three possible diploid genotypes: A_1A_1, A_1A_2, and A_2A_2, and their fitnesses can be denoted by w_{11}, w_{12}, and w_{22}, respectively.

Given that the frequency of allele A_1 in a population is p, and the frequency of the complementary allele, A_2, is $q = 1 - p$, we can show that under random mating, the frequencies of A_1A_1, A_1A_2, and A_2A_2 are p^2, $2pq$, and q^2, respectively. A population in which such genotypic ratios are maintained is said to be at **Hardy–Weinberg equilibrium**.

In the general case, the three genotypes are assigned the following fitness values and initial frequencies:

Genotype:	A_1A_1	A_1A_2	A_2A_2
Fitness:	w_{11}	w_{12}	w_{22}
Frequency:	p^2	$2pq$	q^2

Let us now consider the dynamics of allele frequency changes following selection. Given the frequencies of the three genotypes and their fitnesses as above, the relative frequencies of the three genotypes in the next generation will become p^2w_{11}, $2pqw_{12}$, and q^2w_{22}, for A_1A_1, A_1A_2, and A_2A_2, respectively. Therefore, the frequency of allele A_2 in the next generation will become:

$$q' = \frac{pqw_{12} + q^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \qquad (2.1)$$

The extent of change in the frequency of allele A_2 per generation is denoted as $\Delta q = q' - q$. We can show that:

$$\Delta q = \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}} \tag{2.2}$$

In the following, we shall assume that A_1 is the original or "old" allele in the population. We shall then consider the dynamics of change in allele frequencies following the appearance of a new allele, A_2. For mathematical convenience, we shall assign a fitness value of 1 to the A_1A_1 genotype. The fitness of the newly created genotypes, A_1A_2 and A_2A_2, will depend on the mode of interaction between A_1 and A_2. For example, if A_2 is completely dominant to A_1, then w_{11}, w_{12}, and w_{22} can be written as 1, $1 + s$, and $1 + s$, respectively, where s is the difference between the fitness of an A_2-carrying genotype and the fitness of A_1A_1. A positive value of s means an increase in fitness in comparison with A_1A_1, while a negative value means a decrease in fitness. If A_2 is completely recessive, the fitnesses of the three genotypes become 1, 1, and $1 + s$, respectively.

Two common modes of interaction will be considered: (1) codominance, or genic selection, and (2) overdominance. Codominance represents a case of directional selection and is mathematically the simplest mode of interaction, while overdominance represents a type of balancing selection.

Codominance

In the **codominant mode** of selection, or **genic selection**, the two homozygotes have different fitness values, while the fitness of the heterozygote is the mean of the fitnesses of the two homozygous genotypes. The relative fitness values for the three genotypes can be written as:

Genotype:	A_1A_1	A_1A_2	A_2A_2
Fitness:	1	$1 + s$	$1 + 2s$

From Equation 2.2 we obtain the following change in the frequency of allele A_2 per generation under codominance:

$$\Delta q = \frac{spq}{1 + 2spq + 2sq^2} \tag{2.3}$$

By iteration, this equation can be used to compute the frequency (q) of A_2 at any generation. However, the following approximation leads to a much more convenient formula. If s is small, as is usually the case, the denominator of Equation 2.3 is approximately equal to 1 and the equation becomes approximately

$$\Delta q = spq$$

This difference equation can be approximated by the following differential equation

$$\frac{dq}{dt} = spq = sq(1 - q) \tag{2.4}$$

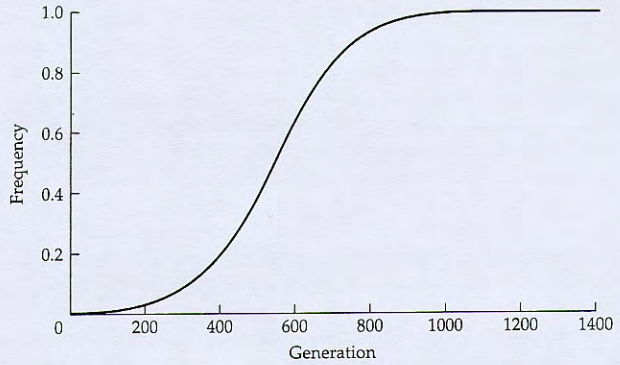


Figure 2.1 Frequency of a codominant advantageous allele with $s = 0.01$ following its appearance as a result of mutation in generation 0. From Li and Graur (1991).

The solution of this equation is given by

$$q_t = \frac{1}{1 + \left(\frac{1 - q_0}{q_0}\right)e^{-st}} \tag{2.5}$$

where q_0 and q_t are the frequencies of A_2 at time 0 and t, respectively.

Figure 2.1 illustrates the increase in the frequency of allele A_2 for $s = 0.01$. Clearly, the frequency of A_2 always increases with time. For this reason, genic selection is a type of **directional selection**. Note, however, that at low frequencies, selection for a codominant allele is not very efficient (i.e., the change in allele frequencies is slow). The reason is that, at low frequencies of A_2, the proportion of A_2 alleles residing in heterozygotes is large. For example, when the frequency of A_2 is 0.5, 50% of A_2 alleles will be carried by heterozygotes, whereas when the frequency of A_2 is 0.01, 99% of all such alleles reside in heterozygotes. Because heterozygotes, which contain both alleles, have a smaller selective advantage than do A_2A_2 homozygotes (i.e., s versus $2s$), the overall change in allele frequencies at low values of q is small.

In Equation 2.5 the frequency q_t is expressed as a function of time t. Alternatively, t can be expressed as a function of the frequency q as follows:

$$t = \frac{1}{s} \ell n \frac{q_t(1 - q_0)}{q_0(1 - q_t)} \tag{2.6}$$

From this equation, one can calculate the number of generations required for the frequency of A_2 to change from one value (q_0) to another (q_t).

Overdominance

In the **overdominant mode of selection**, the heterozygote has the highest fitness. Thus:

Genotype:	A_1A_1	A_1A_2	A_2A_2
Fitness:	1	$1+s_1$	$1+s_2$

In this case, $s_1 > 0$ and $s_1 > s_2$. Depending on whether the fitness of A_2A_2 is greater than, equal to, or less than that of A_1A_1, s_2 can be positive, zero, or negative. The change in allele frequencies is expressed as:

$$\Delta q = \frac{-pq(2s_1q - s_2q - s_1)}{1 + 2s_1pq + s_2q^2} \tag{2.7}$$

Figure 2.2 illustrates the changes in the frequency of an allele subject to overdominant selection. In contrast to the codominant selection regime, in which one of the alleles is eventually eliminated from the population, under overdominant selection the population sooner or later will reach an equilibrium in which the two alleles coexist. After equilibrium is reached, no further change in allele frequencies will be observed (i.e., $\Delta q = 0$). Thus, overdominant selection belongs to a class of selection regimes called **balancing** or **stabilizing selection**.

The frequency of allele A_2 at equilibrium is obtained by solving Equation 2.7 for $\Delta q = 0$:

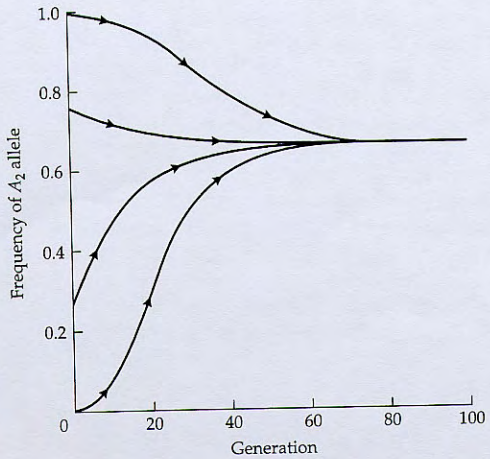


Figure 2.2 Changes in the frequency of an allele subject to overdominant selection. Initial frequencies from top to bottom: 0.99, 0.75, 0.25, and 0.01; $s_1 = 0.250$ and $s_2 = 0.125$. Since the s values are exceptionally large, the change in allele frequency is rapid. Note that there is a stable eqilibrium at $q = 0.667$. Modified from Hartl and Clark (1989).

$$\hat{q} = \frac{s_1}{2s_1 - s_2} \tag{2.8}$$

When $s_2 = 0$ (i.e., both homozygotes have identical fitness values), the equilibrium frequencies of both alleles will be 50%.

RANDOM GENETIC DRIFT

As noted above, natural selection is not the only factor that can cause changes in allele frequency. Allele frequency changes can also occur by chance, though in this case the changes are not directional but random. An important factor in producing random fluctuations in allele frequencies is the random sampling of gametes in the process of reproduction (Figure 2.3). Sampling occurs because, in

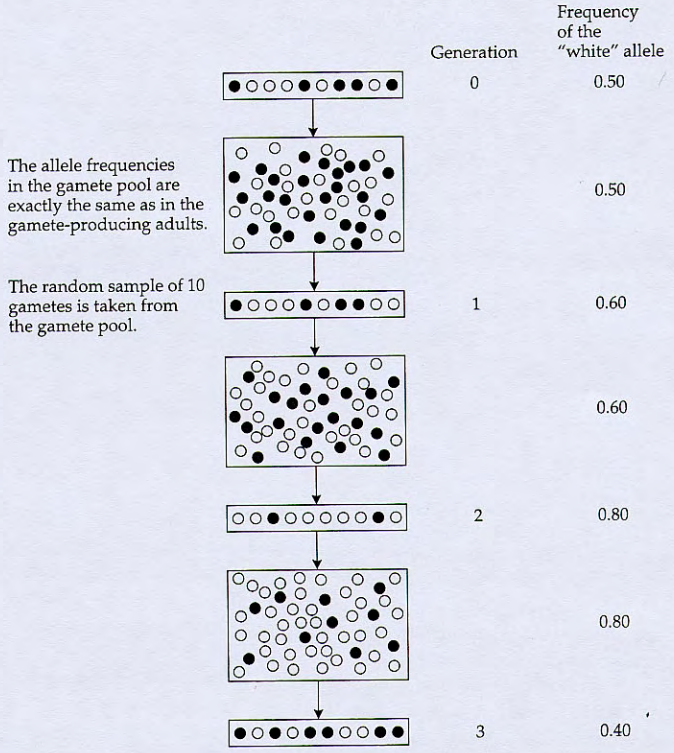


Figure 2.3 Random sampling of gametes. Allele frequencies in the gamete pools (large boxes) in each generation are assumed to reflect exactly the allele frequencies in the adults of the parental generation (small boxes). Since the population size is finite, allele frequencies fluctuate up and down. Modified from Bodmer and Cavalli-Sforza (1976).

the vast majority of cases in nature, the number of gametes available in any generation is much larger than the number of adult individuals produced in the next generation. In other words, only a minute fraction of gametes succeed in developing into adults. In a diploid population under Mendelian segregation, sampling can still occur even if there is no excess of gametes, i.e., even if each individual contributes exactly two gametes to the next generation. The reason is that heterozygotes can produce two types of gametes, but the two gametes passing on to the next generation may by chance be of the same type.

To see the effect of sampling, consider an idealized situation in which all the individuals in the population have the same fitness and selection does not operate. We further simplify the problem by considering a population with nonoverlapping generations (i.e., a group of individuals that reproduce simultaneously), such that any given generation can be unambiguously distinguished from both previous and subsequent generations. The population under consideration is diploid and consists of N individuals, so that the population contains $2N$ genes at each locus. Let us again consider the simple case of one locus with two alleles, A_1 and A_2, with frequencies p and $q = 1 - p$, respectively. When $2N$ gametes are sampled from the infinite gamete pool, the probability, P_i, that the sample contains exactly i genes of type A_1 is given by the binomial probability function:

$$P_i = \frac{(2N)!}{i!(2N-i)!} p^i q^{2N-i} \tag{2.9}$$

Since P_i is always greater than 0 for populations in which $0 < p < 1$, the allele frequencies may change from generation to generation without the aid of selection.

The frequency of allele A_1 at generation t, denoted by p_t, is a random variable. The mean and variance of p_t are given by

$$E(p_t) = p_0 \tag{2.10}$$

$$V(p_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right] \approx p_0(1-p_0)(1-e^{-t/(2N)}) \tag{2.11}$$

where p_0 denotes the initial frequency and is assumed to be known (see Crow and Kimura 1970). Note that although the expectation of p_t stays the same as the intitial frequency, the variance of p_t increases with time. Thus, although random sampling of gametes produces no systematic change in allele frequency, it causes random fluctuations in allele frequency.

The process of change in allele frequency due solely to chance effects is called **random genetic drift**. One should note, however, that random genetic drift can also be caused by processes other than the sampling of gametes. For example, stochastic changes in selection intensity can also bring about random changes in allele frequencies (see Gillespie 1991).

Figure 2.4 illustrates the effects of random sampling on the frequencies of alleles in populations of different sizes. In the figure each curve represents the result of a computer simulation of the random sampling process; in the simulation, each generation is formed by random sampling of $2N$ genes (with replacement) from

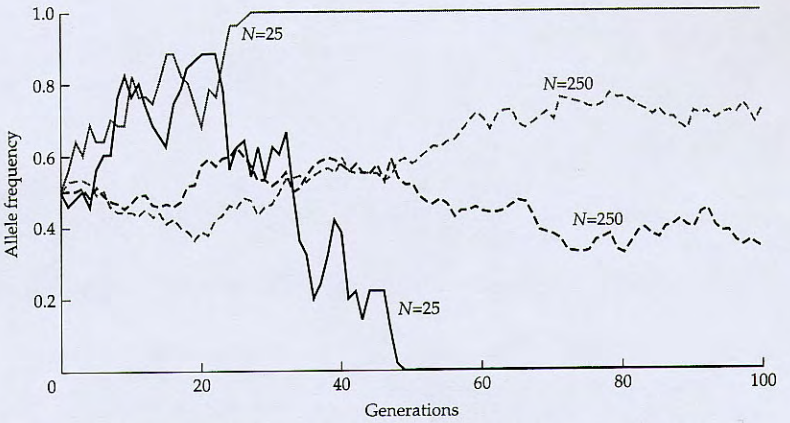
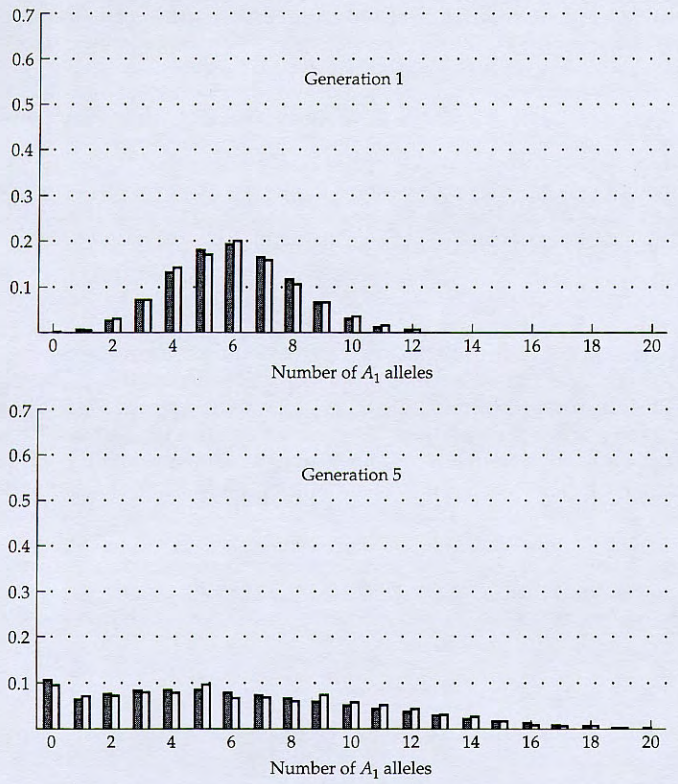


Figure 2.4 Changes in frequencies of alleles subject to random genetic drift in populations of different sizes (N). In each generation, $2N$ genes were sampled with replacement from the previous generation. For each population size, two replicates are presented. It is assumed that the effective population size N_e is equal to the actual size N.

the previous generation. The allele frequencies change from generation to generation, but the direction of change is random at any point in time. The fluctuations in allele frequency are conspicuous in the case of $N = 25$; in one simulation (replicate) the allele became fixed in the population at generation 27 and in the other the allele became lost from the population at generation 49. In the case of $N = 250$ the fluctuations in allele frequency are less pronounced and in both simulations the allele remained at an intermediate frequency at generation 100.

In the stochastic theory of allele frequency changes, one imagines an infinite array of identical populations, all of which have the same population size, are subject to the same sampling procedure, and start with the same initial frequency. At any time point t the allele frequencies in different populations represent the distribution of allele frequencies at time t. For example, the probability that the frequency of allele A_1 is in a certain frequency interval is equal to the proportion of populations in which the frequency of allele A_1 is in that interval. In computer simulation, of course, one cannot simulate an infinite array of populations. However, if the size of the array of simulated populations is large, the frequency distribution obtained will be close to the true distribution. For example, in Figure 2.5 the computer simulation was conducted with an array of 1000 populations. (In practice, it is difficult to simulate 1000 populations at the same time, and so instead one population is simulated at one time and the simulation is repeated 1000 times. Each repeat is known as one replicate.) In the simulation each population consists of $N = 10$ diploid individuals, or 20 genes, and starts with the frequencies 0.3 and 0.7 for the A_1 and A_2 alleles, respectively (a small N is used to save computer time). Since the number of genes is only 20, the exact distribution



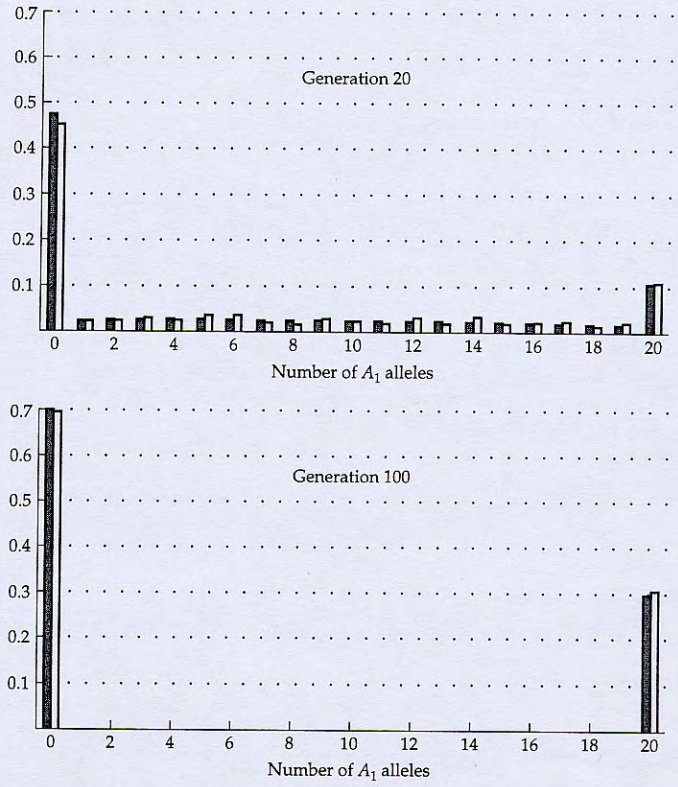


Figure 2.5 Probability distributions of allele frequencies in a diploid population of $N = 10$. The population is under random mating and is not subject to natural selection. The initial frequency of allele A_1 is 0.3, i.e., 6 A_1 alleles in the initial population. The white histograms are theoretical (exact) values and the shaded histograms are values obtained from computer simulation with 1000 replicates. The ordinate denotes the probability that the number of A_1 alleles is equal to a particular number. The heights at the two terminal classes represent the probabilities of loss and fixation of allele A_1, respectively.

can be obtained by the Markov chain method (see, e.g., Ewens 1979; Nei 1987). It is noted that the simulated distributions (shaded histograms) in different generations are very close to the exact distributions (white histograms).

Some interesting features emerge from Figure 2.5. At generation 1, the distribution of the frequency of A_1 follows a binomial distribution, with $p = 0.3$ and $2N = 20$. Since the variance is small, i.e., only $0.3(1 - 0.3)/20 = 0.01$, the distribution is concentrated around the mean (0.3). At generation 5, the distribution becomes flatter. The A_1 allele has become lost in 10.8% of the populations and has become fixed in 0.1% of the populations (these percentages correspond to the heights of the distribution at frequencies 0 and 1, respectively). At generation 20, the distribution becomes fairly uniform for frequencies between 0 and 1, and the probabilites of loss and fixation become 47.4% and 10.4%, respectively. At generation 100, the A_1 allele has become lost in 69.7% of the populations and fixed in 29.7% of the populations; that is, it has become either lost or fixed in 99.4% (almost 100%) of the populations. If t increases further, the A_1 allele will eventually become lost in 70% of the populations and fixed in 30% of the populations; since the initial frequency of A_1 is 0.3 and since there is no factor favoring either of the two alleles, the probability that A_1 will become fixed in the population is equal to the initial frequency of A_1 (see page 47).

The time required for most of the populations to become fixed for the A_1 or A_2 allele can be roughly estimated from Equation 2.11. In this equation the only term that changes with time is $e^{-t/(2N)}$, which, for $N = 10$, reduces to only 0.007 as t increases to 100. Thus, for $t \geq 100$, $V(p_t)$ is close to $p_0(1 - p_0)$, which is the maximum value for $V(p_t)$ and is obtained when all populations have become fixed for either

A_1 or A_2. This means that when $t \geq 100$, almost all populations should have already become fixed for either A_1 or A_2. This is in agreement with the simulation result.

EFFECTIVE POPULATION SIZE

The above mathematical formulation of random genetic drift assumed an idealized population in which all individuals contribute gametes to the next generation with equal probability, generations are nonoverlapping and population size is constant over time. In practice, it is unlikely that all these conditions hold. Furthermore, as will be discussed below, there are often other factors that complicate the mathematical treatment. To simplify the mathematical formulation, Wright (1931) introduced the concept of **effective population size**, N_e, which is the size of an idealized population that would have the same effect of random sampling on gene frequency as that in the actual population. Consider a population with actual size N and assume that the frequency of allele A_1 at the present generation is p. If any of the above conditions is violated, then the variance of the frequency of allele A_1 (p') in the next generation is expected to be larger than the following binomial variance

$$V(p') = p(1-p)/(2N) \tag{2.12}$$

which can be obtained from Equation 2.11 by putting $t = 1$. The concept of effective population size is to define N_e in such a way that the actual variance is given by

$$V(p') = p(1-p)/(2N_e) \tag{2.13}$$

In general, N_e is smaller, sometimes much smaller, than N, the actual population size. Various factors can contribute to this difference. For example, in a population with overlapping generations, at any given time, part of the population will consist of individuals in either their pre- or postreproductive stage. Due to this developmental stratification, the effective size can be considerably smaller than the census size, N. For example, according to Nei and Imaizumi (1966), in humans N_e is only slightly larger than $N/3$.

Reduction in the effective population size in comparison to the census size can also occur if the number of males involved in reproduction is different from the number of females. This disparity is especially pronounced in polygamous species, such as social mammals and territorial birds, or in species in which a nonreproducing caste exists (e.g., bees, ants, and termites). If a population consists of N_m males and N_f females, N_e is given by:

$$N_e = \frac{4N_m N_f}{N_m + N_f} \tag{2.14}$$

Note that, unless the number of females equals the number of males, N_e will always be smaller than N. As an extreme example, let us assume that in a population of size N, all the females ($N/2$) and only one male take part in the reproductive process. By using Equation 2.14, we see that $N_e = 2N / (1 + N/2)$. If N is considerably larger than 1, N_e becomes 4, regardless of the census population size.

The effective population size can also be much reduced due to long-term variations in the population size, which in turn are caused by such factors as environmental catastrophes, cyclical modes of reproduction, and local extinction and recolonization events. For example, the **long-term effective population size** in a species for a period of n generations is given by:

$$N_e = n/(1/N_1 + 1/N_2 + \cdots + 1/N_n) \tag{2.15}$$

where N_i is the population size of the ith generation. In other words, N_e equals the harmonic mean of the N_i values, and consequently it is closer to the smallest value of N_i than to the largest one. Similarly, if a population goes through a bottleneck, the effective population size is greatly reduced.

GENE SUBSTITUTION

Gene substitution is defined as the process whereby a mutant allele completely replaces the predominant or **wild type** allele in a population. In this process, a mutant allele arises in a population, usually as a single copy, and becomes **fixed** after a certain number of generations. The time it takes for a new allele to become fixed is called the **fixation time**. Not all new mutants, however, reach fixation. In fact, the majority of them are lost after a few generations. Thus, we also need to address the issue of **fixation probability** and discuss the factors affecting the chance that a new mutant allele will reach fixation in a population. New mutations arise continuously within populations. Consequently, gene substitutions occur in succession, with one allele replacing another and being itself replaced in time by a new allele. Thus, we can speak of the **rate of gene substitution**, i.e., the number of substitutions or fixations per unit time.

Fixation Probability

The probability that a particular allele will become fixed in a population depends on (1) its initial frequency, (2) its selective advantage or disadvantage, s, and (3) the effective population size, N_e. In the following, we shall consider the case of genic selection and assume that the relative fitness of the three genotypes A_1A_1, A_1A_2, and A_2A_2 are 1, $1 + s$, and $1 + 2s$, respectively.

Kimura (1962) showed that the probability of fixation of A_2 is given by

$$P = (1 - e^{-4N_e sq})/(1 - e^{-4N_e s}) \tag{2.16}$$

where q is the initial frequency of allele A_2. Since $e^{-x} \approx 1 - x$ when x is small, Equation 2.16 reduces to $P \approx q$ as s approaches 0. Thus, for a neutral allele, the fixation probability equals its frequency in the population. For example, in Figure 2.5 the initial frequency of allele A_1 is 30% and so it will eventually become fixed in 30% of the cases and become lost in 70% of the cases. This is intuitively understandable because in the case of neutral alleles, fixation occurs by random genetic drift, which favors neither allele.

We note that a new mutant arising as a single copy in a diploid population of size N has an initial frequency of $1/(2N)$. The probability of fixation of an individual mutant allele, P, is thus obtained by replacing q with $1/(2N)$ in Equation 2.16. When $s \neq 0$,

$$P = [1 - e^{-(2N_e s)/N}]/(1 - e^{-4N_e s}) \tag{2.17}$$

For a neutral mutation, Equation 2.17 becomes

$$P = 1/(2N) \tag{2.18}$$

If the population size is equal to the effective population size, Equation 2.17 reduces to

$$P = (1 - e^{-2s})/(1 - e^{-4Ns}) \tag{2.19}$$

If the absolute value of s is small, we obtain

$$P = 2s/(1 - e^{-4Ns}) \tag{2.20}$$

For positive values of s and large values of N, Equation 2.20 reduces to

$$P = 2s \tag{2.21}$$

Thus, if an advantageous mutation arises in a large population and its selective advantage over the rest of the alleles is small, say < 5%, the probability of its fixation is approximately twice its selective advantage. For example, if a new mutation with $s = 0.01$ arises in a population, the probability of its eventual fixation is 2%.

Let us now consider a numerical example. A new mutant arises in a population of 1000 individuals. For simplicity, we assume that $N = N_e$. The probability that this allele will become fixed in the population is $1/(2N) = 0.05\%$ if it is neutral, 2% if it confers a selective advantage of 0.01, and 0.004% if it has a selective disadvantage of 0.001 (the last two cases are computed from Equation 2.19). These results are quite noteworthy, for they mean that an advantageous mutation does not always become fixed in the population. In fact, 98% of all the mutations with a selective advantage of $s = 0.01$ will be lost by chance. On the other hand, even slightly deleterious mutations have a finite probability of becoming fixed in a population, albeit a small one. The mere fact that a deleterious allele may become fixed in a population illustrates in a powerful way the importance of chance effects in determining the fate of mutations during evolution. If the population size becomes larger, the chance effect, of course, becomes smaller. For instance, in the above example if the population size is $N = N_e = 10{,}000$ instead of 1,000, then the fixation probabilities become 0.005%, 2%, and $\approx 10^{-20}$, respectively. While the fixation probability for the advantageous mutation remains approximately the same, that for the deleterious allele has become very small when N_e increases from 1000 to 10,000. Therefore, in a large population, it is almost impossible for a deleterious mutation to become fixed in the population and the chance for a neutral mutation to become fixed in the population is very small; however, see below for the rate of substitution for neutral mutation.

Fixation Time

The time required for the fixation or loss of an allele depends on the frequency of the allele and the size of the population. The mean time to fixation or loss becomes shorter as the frequency of the allele approaches 1 or 0, respectively.

In terms of evolution, we are more interested in the chance of fixation of new mutations. Thus, in the following we shall deal with the mean fixation time of those mutants that will eventually become fixed in the population. This variable

is called the **conditional fixation time**. In the case of a new mutation [$q = 1/(2N)$], the mean conditional fixation time, $\bar{t}$, was calculated by Kimura and Ohta (1969). For a neutral mutation, it is approximated by

$$\bar{t} = 4N \text{ generations} \tag{2.22}$$

and, for a mutation with a selective advantage of s, it is approximated by

$$\bar{t} = (2/s)\ell n(2N) \text{ generations} \tag{2.23}$$

To illustrate the difference between different types of mutation, let us assume a mammalian species with an effective population size of about 10^6 and a mean generation time of 2 years. Under these conditions, it will take a neutral mutation, on average, $4 \times 10^6 \times 2 = 8$ million years to become fixed in the population. In comparison, a mutation with a selective advantage of 1% will become fixed in the same population in only about 5800 years. Interestingly, the conditional fixation time for a deleterious allele with a selective disadvantage s is exactly the same as that for an advantageous allele with a selective advantage s (Maruyama and Kimura 1974). This is intuitively understandable given the high probability of loss for a deleterious allele. That is, for a deleterious allele to become fixed in a population, fixation must occur very quickly.

Figure 2.6 schematically depicts the dynamics of gene substitution for advantageous and neutral mutations. Advantageous mutations are either rapidly lost or rapidly fixed in the population. In contrast, the frequency changes for neutral alleles are slow, and the fixation time is much longer than for advantageous mutants.

Rate of Gene Substitution

Let us now consider the **rate of substitution**, defined as the number of mutants reaching fixation per unit time. First, consider neutral mutations. If neutral mutations occur at a rate of u per gene per generation, then the number of mutants arising at a locus in a diploid population of size N is $2Nu$ per generation. Since the probability of fixation for each of these mutations is $P = 1/(2N)$ and since the rate of substitution is $K = 2NuP$, we have

$$K = u \tag{2.24}$$

Thus, for neutral mutations, the rate of substitution is equal to the rate of mutation, a remarkably simple result (Kimura 1968a). This result can be intuitively understood by noting that, in a large population, the number of mutations arising every generation is high, but the fixation probability of each mutation is low. In comparison, in a small population, the number of mutations arising every generation is low, but the fixation probability of each mutation is high. As a consequence, the rate of substitution for neutral mutations is independent of population size.

For advantageous mutations, the rate of substitution can also be obtained by multiplying the number of mutations arising every generation (i.e., $2Nu$) by the probability of fixation for such alleles as given in Equation 2.21. For genic selection with $s > 0$, we obtain

$$K = 4Nsu \tag{2.25}$$

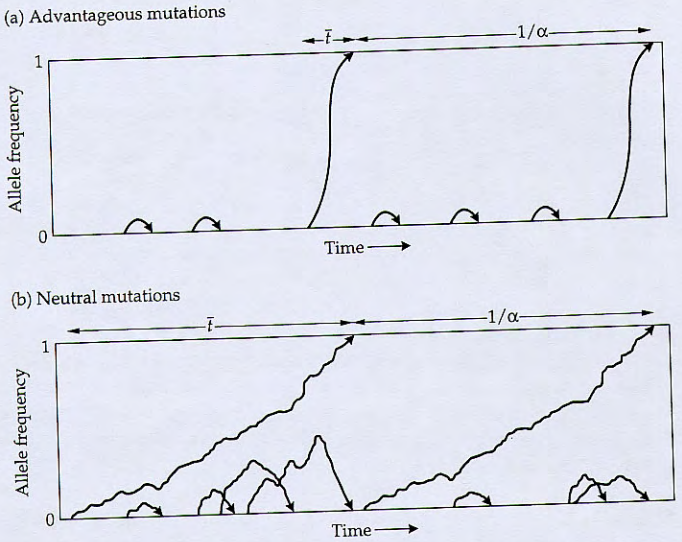


Figure 2.6 Dynamics of gene substitution for (a) advantageous and (b) neutral mutations. Advantageous mutations are either quickly lost from the population or quickly fixed, so their contribution to genetic polymorphism is small. The frequency of neutral alleles, on the other hand, changes very slowly by comparison, so a large amount of transient polymorphism is generated. $\bar{t}$ is the conditional fixation time and $1/\alpha$ is the mean time between consecutive fixation events. From Nei (1987).

In other words, the rate of substitution for the case of genic selection depends on the population size (N) and the selective advantage (s), as well as on the rate of mutation (u).

EXTINCTION OF AN ALLELE UNDER MUTATION PRESSURE

At the DNA (protein) level an allele represents a nucleotide (protein) sequence. The allele will become a different allele if a mutation occurs at any nucleotide site (amino acid residue). Once it mutates to a different allele, the chance for the new allele to mutate back to the original allele is very small because the mutation must occur at the mutated site and in a specific direction. To see this, consider a gene consisting of 300 nucleotides, a fairly small gene. Suppose that a mutation occurs at the tenth nucleotide in allele A_1, changing the nucleotide from T to C, so that a new allele, A_2, is created. If a new mutation occurs in A_2, the probability for it to mutate back to A_1 is less than $1/300$ because the mutation has a probability of $1/300$ to occur at the tenth nucleotide, and even if it occurs at that site, the nucleotide may change to A or G rather than back to T. Therefore, practically every allele in a population is subject to irreversible mutation and will eventually become extinct from the population.

The question then is, "How long will it take for an allele to become extinct from a population under irreversible mutation?" Many authors have studied this problem (e.g., Ewens 1964; Crow and Kimura 1970; Nagylaki 1974; Nei 1976). The problem can be formulated as follows. Let A be the allele under consideration and put all other possible allelic forms into one single class and denote it by a. Let u be the mutation rate per gene per generation. Then, A mutates to a irreversibly at the rate of u per generation. For simplicity, let us consider only genic selection and assume that the relative fitnesses of genotypes AA, Aa, and aa are 1, $1+s$, and $1+2s$, respectively. Thus, $s>0$ ($s<0$) means that all new mutations have a selective advantage (disadvantage) of s over A, whereas $s=0$ means that new mutations are all neutral. Let the initial frequency of A be p and effective population size be N_e. Li and Nei (1977) developed a general formula for the mean extinction time of an allele, $T(p)$, which is not presented here because it is rather complicated in the presence of selection. For neutral mutation, if $p=1$, the formula becomes

$$T(1)=\sum_{i=1}^{\infty}\frac{4N_e}{i(\theta+i-1)} \tag{2.26}$$

where $\theta=4N_eu$. If $\theta\le 0.1$, Equation 2.26 becomes approximately

$$T(1)=4N_e\left(1+\frac{1}{\theta}\right)=4N_e+\frac{1}{u} \tag{2.27}$$

Table 2.1 shows the mean extinction times for neutral mutations ($S=4N_es=0$), advantageous mutations ($S>0$), and disadvantageous mutations ($S<0$). The case

TABLE 2.1 **Mean extinction time of an allele under mutation pressure[a]**

		θ			
S	**p**	**1**	**0.1**	**0.01**	**0.001**
0	1	1.65	10.94	101.0	1001.0
	0.5	1.06	5.80	50.8	500.8
10	1	0.54	1.53	10.4	99.1
	0.5	0.29	0.32	0.39	0.99
100	1	0.10	0.20	1.05	9.61
	0.5	0.05	0.05	0.05	0.05
−5	1	9.93	251	—	—
−50	1	2×10^{18}	7×10^{20}	—	—

From Li and Nei (1977).

[a]Time is measured in units of $4N_e$ generations, where N_e is the effective population size. $\theta=4N_eu$, $S=4N_es$, and p is the initial frequency of the allele (u is the mutation rate and s is the selective advantage or disadvantage). The mutant alleles are selectively neutral if $S=0$, advantageous if $S>0$, and disadvantageous if $S<0$.

of $p = 1$ is of especial interest because if we consider all alleles currently existing in the population as a single allele and denote it by A, then $p = 1$ and $T(1)$ is the expected time until all presently existing alleles become extinct from the population.

First, let us consider neutral mutations. For $\theta \leq 0.1$, the mean extinction time is roughly inversely proportional to $\theta = 4N_e u$ or, in other words, it is roughly proportional to $1/u$ if N_e is fixed; note that in Table 2.1, time is measured in units of $4N_e$ generations. The dependence of the mean extinction time on u is easily seen from Equation 2.27, which shows that if $4N_e < 1/u$, the mean extinction time is roughly proportional to $1/u$. For a given θ, the mean extinction time decreases with the initial frequency p. For example, for $\theta \leq 0.1$, the mean extinction time for $p = 0.5$ is only about half of that for $p = 1$. This can be understood by noting that when $\theta \leq 0.1$, mutations occur rarely, so that for $p = 0.5$, roughly 50% of the cases the allele will become lost in a relatively short time (less than $4N_e$ generations) and in the other 50% of the cases the allele will become fixed or nearly fixed in the population, and the mean extinction time is then similar to that for $p = 1$.

Next, let us consider advantageous mutations. Clearly, if all new mutations are advantageous, the mean extinction time for A is greatly reduced. For example, for $p = 1$ and $\theta = 0.01$, the mean extinction times (in units of $4N_e$ generations) are 101.0 for $S = 0$, 10.4 for $S = 10$, and 1.1 for $S = 100$. Note also that for the case of $p = 0.5$ and $S = 100$, the mean extinction time is independent of mutation rate. This is because the selection intensity is strong so that allele A is quickly replaced by the mutant alleles already existing in the population.

Finally, if all mutations are disadvantageous, the mean extinction time is greatly increased. For example, if $p = 1$, $\theta = 1$, and $S = -50$, the mean extinction time is $2 \times 10^{18} \times 4N_e$ generations, which is extremely long. Note that if $N_e = 12{,}500$, then $S = -50$ implies $s = -0.001$, which is fairly small. Thus, if the effective population size is fairly large, disadvantageous mutations have practically no chance of becoming fixed in the population. The assumption that all mutations are disadvantageous implies that the present allele is the optimal allele and that it is subject to purifying selection arising from functional or structural requirements of the sequence. The above result implies that if such requirements are stringent, then the optimal allele can persist in the population for an extremely long time. As we shall see in Chapter 7, a protein sequence with stringent structural requirements can indeed persist for millions of years without change.

GENETIC POLYMORPHISM

A locus is said to be **polymorphic** if two or more alleles coexist in the population. However, if one of the alleles has a very high frequency, say, 99% or more, then none of the other alleles is likely to be observed in a sample unless the sample size is large. Thus, for practical purposes, a locus is commonly defined as polymorphic if the frequency of the most common allele is less than 99%. This definition is obviously arbitrary and other criteria have been used in the literature.

One of the simplest ways to measure the extent of polymorphism in a population is to compute the proportion of polymorphic loci (P) by dividing the number of polymorphic loci by the total number of loci sampled. For example, if 5 of the 20 loci studied are polymorphic, then $P = 5/20 = 25\%$. This measure, however, is dependent on the number of individuals studied. A more appropriate measure of genetic variability within populations is **gene diversity**. This measure does not depend on an arbitrary delineation of polymorphism, can be computed directly from knowledge of the gene frequencies, and is less affected by sampling effects. Gene diversity at a locus is defined as:

$$h = 1 - \sum_{i=1}^{m} x_i^2 \qquad (2.28)$$

where x_i is the frequency of allele i and m is the number of alleles observed at the locus. For any given locus, h is the probability that two alleles chosen at random from the population are different from each other. In a randomly mating population, h is also the **expected heterozygosity**, i.e., the expected frequency of heterozygotes in the population for a locus with allele frequencies x_i, $i = 1,\ldots, m$. The average of h values over all the loci studied can be used as a measure of genetic variability in the population.

The introduction of electrophoresis into population genetics in the early 1960s provided a convenient, powerful tool for studying protein polymorphism in natural populations. It was then discovered that natural populations such as humans and *Drosophila* contain a large amount of genetic variability (Harris 1966; Lewontin and Hubby 1966). In fact, early surveys revealed that the proportion of polymorphic loci is about 30% in mammalian species and can be more than 50% in *Drosophila* species (Table 2.2). The question was then how such high genetic variabilities are maintained in natural populations. This issue will be discussed in the next section and in Chapter 9.

Since the 1960s there has been much interest in developing mathematical models for studying genetic variability. A commonly used one is the infinite-allele

TABLE 2.2 Surveys of protein polymorphism in a number of organisms[a]

Species	Number of populations	Number of loci	P	H
Homo sapiens	1	71	0.28	0.067
Mus musculus musculus	4	41	0.29	0.091
M. m. brevirostris	1	40	0.30	0.110
M. m. domesticus	2	41	0.20	0.056
Peromyscus polionotus	7 (regions)	32	0.23	0.057
Drosophila pseudoobscura	10	24	0.43	0.128
D. persimilis	1	24	0.25	0.106
D. obscura	3 (regions)	30	0.53	0.108
D. subobscura	6	31	0.47	0.076
D. willistoni	10	20	0.81	0.175
D. melanogaster	1	19	0.42	0.119
D. simulans	1	18	0.61	0.160

From Lewontin (1974).

[a] P = proportion of polymorphic loci; H = average heterozygosity.

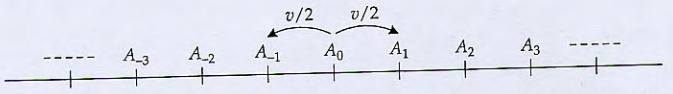


Figure 2.7 The stepwise mutation model. A_0, A_1, etc. denote the possible allelic states and v denotes the mutation rate per gene per generation. From Ohta and Kimura (1973).

model (Wright 1949; Kimura and Crow 1964), which assumes that every mutation creates a new allele or an allele that is not currently existing in the population. Under this model and the assumption of selective neutrality, the expected heterozygosity at equilibrium is given by

$$h = \frac{4N_e u}{1+4N_e u} \tag{2.29}$$

where N_e is the effective population size and u is the mutation rate per gene per generation (Kimura and Crow 1964). This simple formula has been heavily used as a reference point for the neutral expectation. There are of course many factors, such as selection and migration, that can cause deviations from this expectation. For example, overdominant selection can greatly increase the heterozygosity, whereas puryfing selection can reduce it (see Kimura and Crow 1964; Li 1977; Watterson 1977).

It was questioned whether the infinite-allele model is suitable for analyzing electrophoretic data, because the electrophoretic mobility of a protein may change in a stepwise manner so that recurrent and backward mutation can occur. In other words, a mutation may not create a new allele. To make the model more realistic, Ohta and Kimura (1973) proposed the stepwise-mutation model as shown in Figure 2.7. In this model the possible allelic states are visualized as the integers on a line, and a mutation causes the state of the allele to move one step either to the right or to the left. Under this model and the assumption of neutrality, the expected heterozygosity at equilibrium is given by

$$h = 1 - 1/\sqrt{1+8N_e u} \tag{2.30}$$

The h value under the step-mutation model can be substantially lower than that under the infinite-allele model. This model has been extended to include the possibility of two-step changes (Wehrhahn 1975; Li 1976). Although electrophoresis is now only occasionally used in polymorphism study, the stepwise model is introduced here, because it will be used when we discuss the population genetics of tamdem repeats of DNA (Chapter 13).

THE NEO-DARWINIAN THEORY AND THE NEUTRAL MUTATION HYPOTHESIS

Darwin proposed his theory of evolution by natural selection without knowledge of the sources of variation in populations. After Mendel's laws were rediscovered and genetic variation was shown to be generated by mutation, Darwinism and

Mendelism were used as the framework of what came to be called the synthetic theory of evolution, or neo-Darwinism. According to this theory, mutation is recognized as the ultimate source of genetic variation, but natural selection is given the dominant or "creative" role in shaping the genetic makeup of populations and in the process of gene substitution.

In time, neo-Darwinism became a dogma in evolutionary biology, and selection came to be considered the only force capable of driving the evolutionary process, while other factors such as mutation and random drift were thought of as minor contributors at best. This particular brand of neo-Darwinism was called **selectionism**.

According to the selectionist or neo-Darwinian perception of the evolutionary process, gene substitutions occur as a consequence of selection for advantageous mutations. Polymorphism, on the other hand, is maintained by balancing selection. Thus, neo-Darwinists regard substitution and polymorphism as two separate phenomena driven by different evolutionary forces. Gene substitution is the end result of a positive adaptive process, whereby a new allele takes over future generations of the population if and only if it improves the fitness of the organism, while polymorphism is maintained when the coexistence of two or more alleles at a locus is advantageous for the organism or the population. Neo-Darwinian theories maintain that most genetic polymorphisms in nature are stable.

The 1960s witnessed a revolution in population genetics. The introduction of electrophoresis into population genetics studies soon led to the discovery of the existence of large amounts of genetic variability in natural populations such as human and *Drosophila* populations (Harris 1966; Lewontin and Hubby 1966). The availability of protein sequence data removed the species boundary in population genetics studies and for the first time provided adequate empirical data for examining theories pertaining to the process of gene substitution. In 1968, Kimura postulated that the majority of molecular changes in evolution are due to the random fixation of neutral or nearly neutral mutations (Kimura 1968a); it was also independently proposed by King and Jukes (1969). This hypothesis, now known as the **neutral theory of molecular evolution**, contends that at the molecular level the majority of evolutionary changes and much of the variability within species are caused neither by positive selection of advantageous alleles nor by balancing selection, but by random genetic drift of mutant alleles that are selectively neutral or nearly so. Neutrality, in the sense of the theory, does not imply strict equality in fitness for all alleles. It only means that the fate of alleles is determined largely by random genetic drift. In other words, selection may operate, but its intensity is too weak to offset the influences of chance effects. For this to be true, the absolute value of the selective advantage or disadvantage of an allele must be smaller than $1/(2N_e)$.

According to the neutral theory, the frequency of alleles is determined largely by stochastic rules, and the picture that we obtain at any given time is merely a transient state representing a temporary frame from an ongoing dynamic process. Consequently, polymorphic loci consist of alleles that are either on their way to fixation or on their way to extinction. Viewed from this perspective, all molecular manifestations that are relevant to the evolutionary process should be regarded as the result of a continuous process of a mutational input and a concomitant random extinction or fixation of alleles. Thus, the neutral theory regards substitution

and polymorphism as two facets of the same phenomenon. Substitution is a long and gradual process, whereby the frequencies of mutant alleles increase or decrease randomly, until the alleles are ultimately fixed or lost by chance. At any given time, some loci will possess alleles at frequencies that are neither 0% nor 100%. These are the polymorphic loci. According to the neutral theory, most genetic polymorphism in populations is transient in nature.

The essence of the dispute between neutralists and selectionists essentially concerns the distribution of fitness values of mutant alleles. Both schools agree that most new mutations in proteins are deleterious and that these mutations are quickly removed from the population so that they contribute neither to the rate of substitution nor to the amount of polymorphism within populations. The difference concerns the relative proportion of neutral mutations among nondeleterious mutations. While selectionists maintain that very few mutations are selectively neutral, neutralists maintain that most nondeleterious mutations are effectively neutral. Of course, not all selectionists hold the same view of evolution nor do all neutralists (see reviews by Lewontin 1974; Kimura 1983; Nei 1987; Gillespie 1991). For example, Ohta's (1973, 1974) hypothesis of slightly deleterious mutation emphasizes the importance of slightly deleterious mutations in gene substitution and molecular polymorphism, whereas in Nei's (1987) view such mutations do not play an important role.

The heated controversy over the neutral-mutation hypothesis during the last two decades has had a strong impact on molecular evolution. First, it has led to the general recognition that the effect of random drift cannot be neglected when considering the evolutionary dynamics of molecular changes. Second, the synthesis between molecular biology and population genetics has been greatly strengthened by the introduction of the concept that molecular evolution and genetic polymorphism are but two facets of the same phenomenon (Kimura and Ohta 1971a). Although the controversy still continues, it is now recognized that any adequate theory of evolution must be consistent with both of these aspects of the evolutionary process at the molecular level.

PROBLEMS

1. If A_2 is completely dominant to A_1, what will be the change in the frequency of allele A_2 per generation?
2. Derive Equation 2.5 from Equation 2.4. Hint: Equation 2.4 can be rewritten as

$$\frac{dq}{q(1-q)} = sdt$$

3. Use Equation 2.6 to compute the times required for the allele frequency to increase (1) from 0.01 to 0.201, (2) from 0.3 to 0.5, and (3) from 0.8 to 0.99. From the results discuss the effect of allele frequency on the efficiency of natural selection.
4. Derive the equilibrium frequency in Equation 2.8 from Equation 2.7.
5. Use the binomial probability function (2.9) to compute the distribution of allele frequencies at generation 1 in Figure 2.5, assuming $N = 10$ and $p = 0.3$.

6. In an idealized population in which the effective size N_e is the same as the actual size N, show that the variance of allele frequency in the next generation is given by Equation 2.12.
7. What is the ratio of the effective population size to census population size in a population in which females outnumber males by 2:1?
8. A population runs through a bottleneck such that in six consecutive generations its population size is 10^4, 10^4, 10^4, 10, 10^4, and 10^4. What is its long-term effective population size?
9. What is the fixation probability of a new mutation with a selective disadvantage of 0.01 in a population in which the effective population size is 100 and $N_e = N$?
10. Discuss the differences between the neutral mutation hypothesis and the hypothesis that all mutations are neutral.