10

Genes in Populations: Forward in Time

> *The very small range of selective intensity in which a factor may be regarded as effectively neutral suggests that such a condition must in general be extremely transient.*
> *Ronald A. Fisher [76, p. 95]*

> *The principal evolutionary mechanism in the origin of species must thus be an essentially nonadaptive one.*
> *Sewall Wright [263, p. 364]*

Genetic diversity is ubiquitous and some of the best-known examples in humans include the different sex chromosomes, the different blood groups, and the presence or absence of genetic diseases such as cystic fibrosis. In fact, except for monozygotic twins, all humans are genetically distinct. At the molecular level this observation corresponds to the knowledge that no two humans have the same genome sequence. Although there is evidence for abundant large scale genetic variation between humans [126], a major fraction of the hitherto studied human variation concerns *single nucleotide polymorphisms* (SNPs). When comparing two homologous sequences from humans, there is approximately one such SNP per kilobase [244, 11]. However, this number can vary widely between different genomic regions, between different populations and it can be very different in other species.

As Gillespie has put it, it is the “great obsession” of population geneticists to account for the causes and consequences of genetic diversity found in natural populations [87, p. 4]. A *population* is a reproductive community of organisms belonging to the same species. Figure 10.1 illustrates that a genealogical tree of organisms taken from the same species is generally much shallower than a species tree. As a consequence, the number of mismatches and indels found in intra-specific alignments is generally much smaller than in inter-species comparisons. This, in turn, has important implications for the way we interpret and model sequences and their polymorphisms.

10.1 Polymorphism and Genetic Diversity

In Chapter 9 we have stressed the fact that new mutations originate as a single copy. The chromosome carrying the lone new variant or *allele* can be passed on to multiple descendants in subsequent generations. In this way, the frequencies of the novel and of the previously existing allele, also called *wild-type* allele, may change. Depending on factors such as chance and reproductive success of their carriers, both alleles will

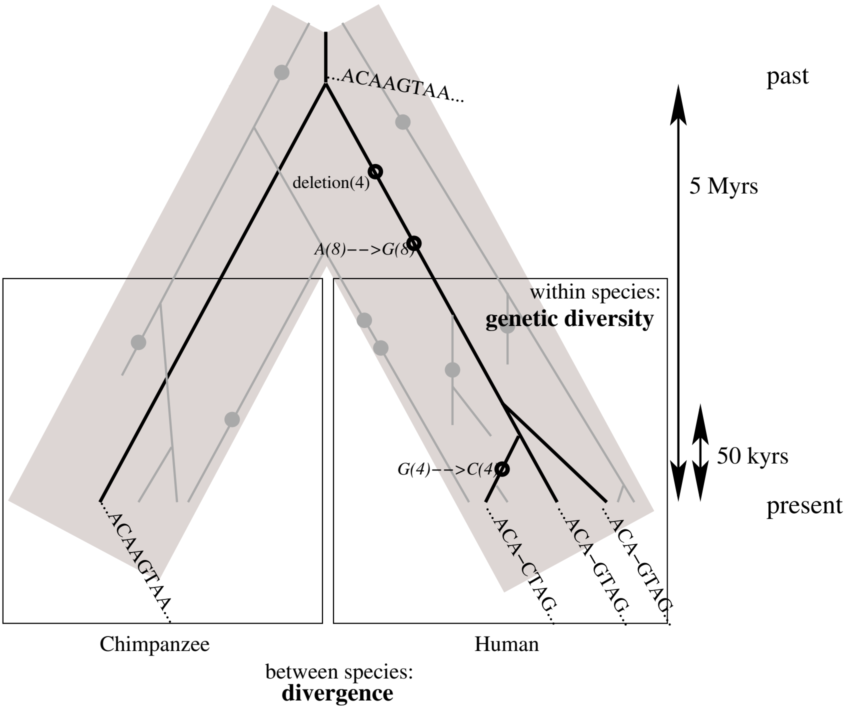

Fig. 10.1. A genealogical tree of DNA sequences. The black lines indicate the genealogical history of a sample of four homologous DNA sequences. While the most recent common ancestor of human and chimpanzee existed about 5 million years ago, the last common ancestor of the three human sequences belonged to an individual who lived approximately 50,000 years ago. The gray lines indicate lineages which are not present in the sample or which became extinct. The amount of variation is generally much smaller within species (genetic diversity) than between species (divergence).

be present for some time in the population. However, eventually only one allele will survive and there is some chance that the new allele will have *substituted* the previous wild-type allele (Fig. 10.2). The fluctuation of allele frequencies due to chance and random fixation of one or the other allele is called *random genetic drift*. The simultaneous presence in a population of two or more alleles at a defined position in the genome is called a *polymorphism*. A single nucleotide polymorphism, or *SNP*, refers to a polymorphism at a single nucleotide site in the genome. Its cause is a point-mutation, i.e. a single base exchange. The vast majority of SNPs are *bi-allelic*. An example is the $\mathtt{G}_4/\mathtt{C}_4$ polymorphism in the human lineage in Figure 10.1, with the two alleles G and C at position 4. Another type of frequently occurring polymorphism is due to replication slippage and becomes manifest as *length polymorphism*,

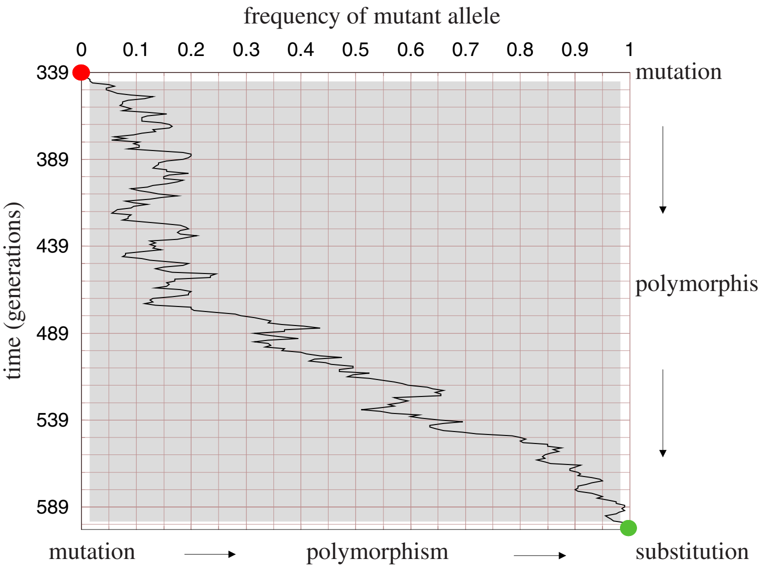

Fig. 10.2. Polymorphisms originate as mutations and may turn into substitutions due to random genetic drift. A computer simulation was performed under a two-allele neutral Wright-Fisher model with parameters $N = 100$ and $\mu = 0.001$. The trajectory shown and originating at generation $t = 339$ eventually reaches frequency $f = 1$. In practice, the term "polymorphism" is often used only for those cases in which the frequency of the minor allele is $f \geq 0.01$; this corresponds to the area shaded in gray.

in particular in tandem repetitive DNA. Stretches of tandemly repeated DNA are also called *micro-* or *minisatellites*, depending on the size of the repeat unit (Fig. 10.3A). Typical microsatellites have repeat units of length 2-3 bp, minisatellites of up to several 100 bp. Microsatellites usually are multi-allelic and constitute important genetic markers because of their high variability between organisms.

The average amount of polymorphisms measured within a population and across some genomic region is called *genetic diversity*. The first study of genetic diversity in *D. melanogaster* at the nucleotide sequence level was published by Kreitman in 1983 [152]. He sequenced a stretch of 2,721 bp at the *Adh* locus in a sample of 11 flies. Kreitman found 43 single nucleotide polymorphisms among the 2,721 sites. In addition, his data set contained six indels, yielding a total of 49 variable sites. Two out of the 11 sequences were identical, of the remaining nine each occurred only once in the sample (Table 10.1). Notice, however, that *Drosophila* strain Wa-F is distinguished from strain Af-F by a single nucleotide in the length of the insertion ∇_3 located in the 3' untranslated region. In our subsequent treatment of this data set we will ignore this polymorphism and say that it contains a total of nine distinct alleles.

We can also combine the analysis of genetic variation found within a population or species with that found between species. For example, we may compare two se-

Table 10.1. Polymorphism at the *Adh* locus of *Drosophila melanogaster*. The boxed strains are identical (ignoring the length polymorphism at insertion ∇_3). The *F* and *S* in the strain designations refer to the *fast* and *slow* allozyme variants of the *Adh* enzyme. ∇/Δ: insertion/deletions numbered from left to right; numbers in the insertion/deletion columns refer to length differences compared to the consensus sequence. The underlined polymorphism in exon 4 causes the Thr/Lys amino acid replacement that underlies the fast/slow allozyme polymorphism. Data taken from [152].

Strain	5' Flanking	Exon 1	Intron 1	Exon 2	Intron 2	Exon 3	Intron 3	Exon 4, translated	3' Untranslated	3' Flanking
	CCG		CAATATGGG∇_1C∇_2GC	T	AC	CCCC	GGAAT	CTCC$\underline{\text{A}}$CTAG	A∇_3C	AGC∇_4C∇_5TΔ_6
Wa-S	...		.....AT..	.	..	TT.A	CA.TA	AC.......	. . .	........Δ
Fl-1S	..C		....	.	..	TT.A	CA.TA	AC.......	. . .	........Δ
Af-S	...		....	.	..	....	.....	........A	. . .	..T ∇ . 1 A .
Fr-S	...		....	.	GT	....	.....	........A	. -1 .	TA. . . . . . .
Fl-2S	...		AG...A.TC . . . .A	G	GT	....	.....	....	C 3 .	....
Ja-S	..C		....	G	..	....	.....	...T.T.CA	C 4 .	.... . T . . .
Fl-F	..C		....	G	..	....	.....	..GTCTCC.	C 4 .	....
Fr-F	TGC		AG...A.TC ∇ G ∇ ..	G	..	....	.....	..GTCTCC.	C 4 G	....
Wa-F	TGC		AG...A.TC ∇ G ∇ ..	G	..	....	.....	..GTCTCC.	C 4 G	....
Af-F	TGC		AG...A.TC ∇ G ∇ ..	G	..	....	.....	..GTCTCC.	C 5 G	....
Ja-F	TGC		AGGGGA... ∇ . . T.	G	..	..A.	..G..	..GTCTCC.	C 4 .	...... -1 . .
Polym. Sites	3	0	12	1	2	4	5	9	2	5
Nucleotides[1]	63	87	690	99	65	405	70	264	178	767
% polymorphic	4.7	0	1.7	1.0	3.1	1.0	7.1	3.5	1.1	0.6

[1] Average number of nucleotides compared

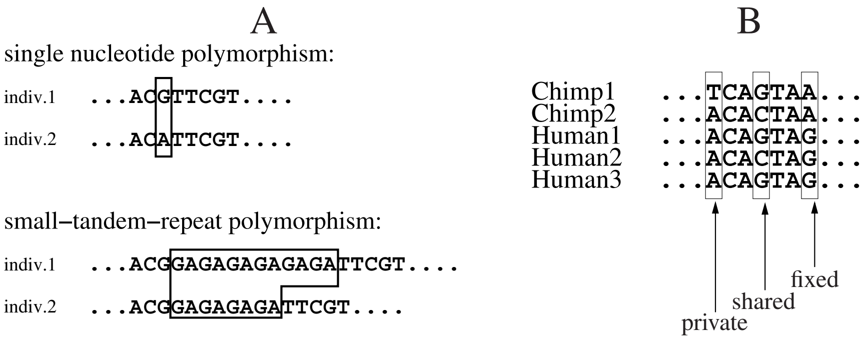

Fig. 10.3. A: Comparisons within a species or a population. The sketch shows the difference between single nucleotide *vs.* length polymorphisms; the lower example depicts a dinucleotide tandem repeat as found in microsatellites. The two alleles differ in their numbers of dinucleotide repeats. **B**: Comparison within and between species or populations. According to their distribution, the polymorphisms found in the five individuals are shared, private, or fixed differences.

quences of a particular gene from chimpanzee with three sequences of the same gene from human (Fig. 10.3). A *private* polymorphism is found only within one of the two species. In our example the private polymorphism only occurs in the chimpanzee. In contrast, a *shared* polymorphism is variable in both species. A *fixed difference*, finally, is monomorphic within a species but polymorphic between (Fig. 10.3B).

10.2 The Neutral Theory

The contribution of natural selection to shaping the genetic diversity around us has been debated for over a century. Darwin himself wrote [44, p. 103]

> I am inclined to suspect that we see in these polymorphic genera variations in points of structure which are of no service or disservice to the species.

In the 1920s the American population geneticist Sewall Wright and his English colleague Ronald A. Fisher started a controversy over the relative importance of adaptive and non-adaptive evolution that was to last until Fisher's death in 1962 [204]. Fisher saw little role for non-adaptive evolution, while this was an important component of Wright's models of evolution. However, by the 1950s Fisher's view that essentially all heritable differences were adaptive had gained an ascendancy that is rather surprising, given Darwin's skepticism in this matter.

In 1966, two studies on genetic diversity in natural populations were published that challenged the prevailing view. One study focused on genetic diversity in the fruit fly *Drosophila melanogaster* [118, 167], the other on human genetic diversity [104] and both were based on allozyme data. In order to collect these, total protein is extracted from an organism, run on a non-denaturing gel, and then stained for specific enzyme activity. The result is one or more bands on the gel, whose positions indicate the allele of the gene encoding the enzyme [210]. In *Drosophila*, 39% of the loci investigated were polymorphic, leading to the estimate that 8–15% of all loci in an individual fruit fly were heterozygous [167]. In humans, the amount of genetic diversity uncovered was also surprisingly high [104]. This lead Kimura [139] and, independently, King and Jukes [145] to propose that the unexpectedly high levels of polymorphism at the molecular level were best explained by assuming that the vast majority of them had no influence on an organism's fitness. Under this neutral hypothesis of molecular evolution, changes in allele frequencies between generations are due to chance, that is genetic drift, alone and not due to selection. As we will see in more detail below, the main effects of drift are:

1. Undirected change of allele frequencies.
2. Removal of established genetic diversity. The speed with which this happens is inversely proportional to population size. Loss of diversity through drift is therefore mainly relevant in small populations.
3. Removal of new mutations. This is important both in small as well as in large populations because mutations are the only source of genetic diversity and hence constitute the "raw material" of evolution.

The birth of new alleles through mutation and their death through drift tend to come to a mutation drift equilibrium. According to the neutral theory, most genetic diversity, which can be seen today, is a reflection of this equilibrium [141].

Two lines of reasoning were used to support the neutral theory. The first was put forward earlier by Haldane [102] who proposed that allele substitution is the result of positive selection acting on favorable alleles. However, each substitution is inevitably linked to a number of genetic deaths of those individuals that do not carry the favorable allele. This effect was called the *genetic load* or the *cost of natural selection*. Haldane [102] had estimated that a species could tolerate at most one substitution every 300 generations in order to cope with the *cost of natural selection*. Kimura examined globins (see Table 9.1) and other protein sequences from various vertebrates and calculated an average rate of 2.8 amino acid substitution per 10^7 years per 100 amino acids [139]. Assuming a size of the mammalian genome of $4 \cdot 10^9$ nucleotides, he estimated that this rate would correspond to 0.57 nucleotide replacements per year, or one substitution per 1.75 years. For the average mammalian generation size of four years, this amounts to 2.3 substitutions per generation [139]. This figure contrasts strongly with Haldane's result, unless the assumption that all substitutions are selective is dropped.

The second line of reasoning, put forward by King and Jukes in 1969, stressed the fact that the genetic code allowed for synonymous substitutions which were not affecting the encoded protein and were therefore invisible to selection (Table C.4). The discovery only a few years earlier of the genetic code and its degeneracy [191] was an essential prerequisite to this argument.

In this chapter we consider neutral models of evolution that move forward in time, while backward in time models are introduced in Chapter 11. We start by explaining how population genetic quantities are simulated forward in time, before approaching evolutionary models of increasing complexity and hence realism.

10.3 Modeling Evolution Forward in Time

Evolutionary processes tend to unfold over long periods of time. There are exceptions to this rule, such as the evolution of pathogenic viruses and bacteria, which is quick enough to generate vaccine-resistant or antibiotic-resistant strains within a few years—a very short period on an evolutionary time scale. However, even with microbes it takes dedication to observe evolution directly [61].

A popular substitute for direct observation of complex dynamical processes are computer simulations. Figure 10.4 demonstrates how simple simulations of genetic quantities can be carried out forward in time: the population, usually consisting of a large set of haplotypes or genotypes, is represented in its entirety and reproduces from one generation to the next by resampling with replacement. Genetic quantities of interest, for example the frequency of an allele, are then computed either from the entire population, or from a random sample.

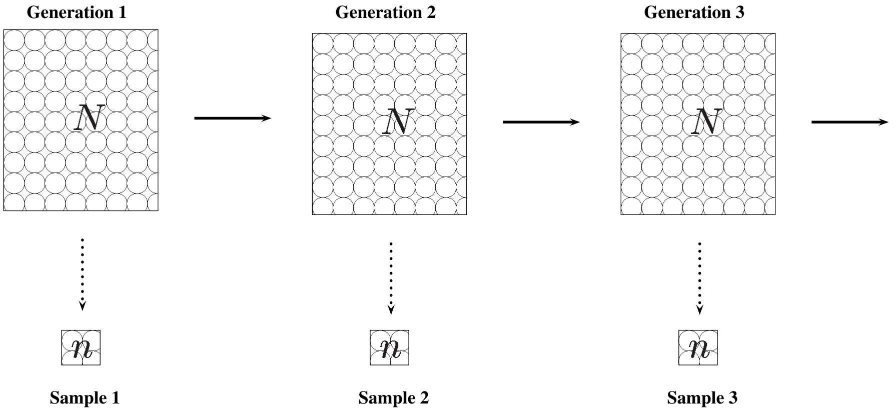

Fig. 10.4. Simulating the evolution of a population (size N) on a computer by moving forward in time. As the simulation moves from one generation to the next, samples (size n) are drawn from which the statistic of interest, e.g. the frequency of a particular allele, is calculated.

Consider one generation of a population of size $N = 6$ consisting of two alleles, A and a: $g_1 = \{a, a, A, A, A, A\}$. The frequency of allele a, p_a, is equal to $1/3$ and $p_\mathrm{A} = 1 - p_\mathrm{a} = 2/3$. Now simulate the process of evolving from the present generation to the next by rolling a die six times returning, say, $1, 3, 6, 4, 3, 6$. For the 1 we draw the first allele (a), for the 3 the third (A), and so on. This leads to the allele configuration $g_2 = \{a, A, A, A, A, A\}$ in the second generation. The allele frequencies have changed and we repeat the random drawing of alleles to produce the next generation. This time the result is $g_3 = \{A, A, A, A, A, A\}$; in other words, A has become fixed and a extinct, as in the absence of mutation there is no way of regenerating a. Thus, genetic diversity has been lost.

A standard measure of genetic diversity is the *heterozygosity*, H, which is the probability that two randomly drawn alleles are different. So for g_1 we have

$$H(g_1) = \frac{\binom{2}{1}\binom{4}{1}}{\binom{6}{2}} = 0.533$$

and for g_2 the diversity is

$$H(g_2) = \frac{\binom{1}{1}\binom{5}{1}}{\binom{6}{2}} = 0.333.$$

In g_3 finally, $H(g_3)$ drops to 0 and remains stuck at this value in all subsequent generations. Our simple experiment reconfirms two important points about drift: (i) it changes allele frequencies and (ii) it removes alleles from the population irreversibly.

10.4 The Neutral Wright-Fisher Model

The neutral *Wright-Fisher model* is the simplest population genetic model and makes the following assumptions. A population is represented as a set of genes. When considering diploid organisms, i.e. those with a double set of chromosomes such as humans, there are $2N$ genes in a population of N organisms. The model further assumes that population size is finite and remains constant over time. Generations are discrete and non-overlapping; therefore, they can be indexed with integers t, $t+1$ and so forth. The genes of generation $t+1$ are drawn randomly from the *gene pool*, i.e. the genes present in generation t (Fig. 10.5). Drawing of genes is carried out with replacement since any particular gene may be passed on to more than one offspring. The biological counterpart to random sampling is the assumption of *random mating* or *panmixis* in sexually reproducing organisms. Randomly mating individuals have no preference for a specific partner based on phenotypic traits or geographic proximity.

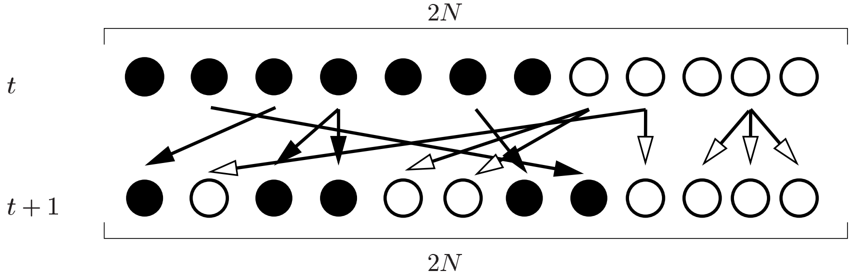

Fig. 10.5. Evolution in a Wright-Fisher population. Arrows indicate the transition from generation t to $t+1$ by drawing with replacement $2N$ genes from generation t. Note that the frequency of alleles (represented as filled and empty circles) may change due to genetic drift. In this example, the frequency of black alleles decreases from $p(t) = 7/12$ to $p(t+1) = 5/12$.

10.4.1 Fixation and Loss of Alleles

To formalize the ideas just presented, assume there are two versions of a given gene, the alleles A and a. Let them be present in a finite diploid population of size N in generation t with relative frequencies

$$p_A(t) = p(t) = \frac{i}{2N}$$

and

$$p_a(t) = 1 - p(t) = \frac{2N - i}{2N},$$

where i is a number between 0 and $2N$. The frequency $p(t+1)$ depends on the outcome of a binomially distributed random variable with parameters $2N$ and $p(t)$. The possible states $j = 0, ..., 2N$ in generation $t+1$ have therefore the conditional probabilities

$$\text{Prob}(j, t+1|i, t) = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1 - \frac{i}{2N}\right)^{2N-j}. \tag{10.1}$$

The righthand side of Equation (10.1) is independent of t. This property is called *homogeneity*. The associated stochastic process $(X(t))_t$ is Markovian with stationary transition probabilities and transition matrix

$$\mathcal{P} = \left(\binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1 - \frac{i}{2N}\right)^{2N-j}\right)_{ij} \quad i, j \in 0, ..., 2N.$$

Furthermore, the probabilities Prob$(j, t+u|i, t)$ are obtained from the matrix product $\mathcal{P}^u$, for any t and u. The boundaries $j = 0$ and $j = 2N$ are *absorbing* states. This property corresponds to our observation made in the die rolling experiment that once an allele is lost, it cannot be restored. The complementary allele is *fixed* and remains so forever. Formally, if $p(t^*) = 1$ for some time t^* then $p(t) = 1$ for all $t > t^*$. Two realizations of this process are depicted in Figure 10.6.

Claim 10.1 *Either one of the two alleles will eventually be fixed, i.e. there is a time t^*, such that*

$$\textit{Prob}(X(t) = 0 \vee X(t) = 2N) = 1 \tag{10.2}$$

for all times $t > t^*$.

PROOF. Let $X(t)$ be the number of A-alleles at time t and $B(n,p)(x)$ the cumulative binomial probability with parameters n and p, i.e.

$$B(n,p)(x) = \sum_{y=0}^{x}\binom{n}{y}p^y(1-p)^{n-y}.$$

For any time t and independently of any initial conditions it holds that

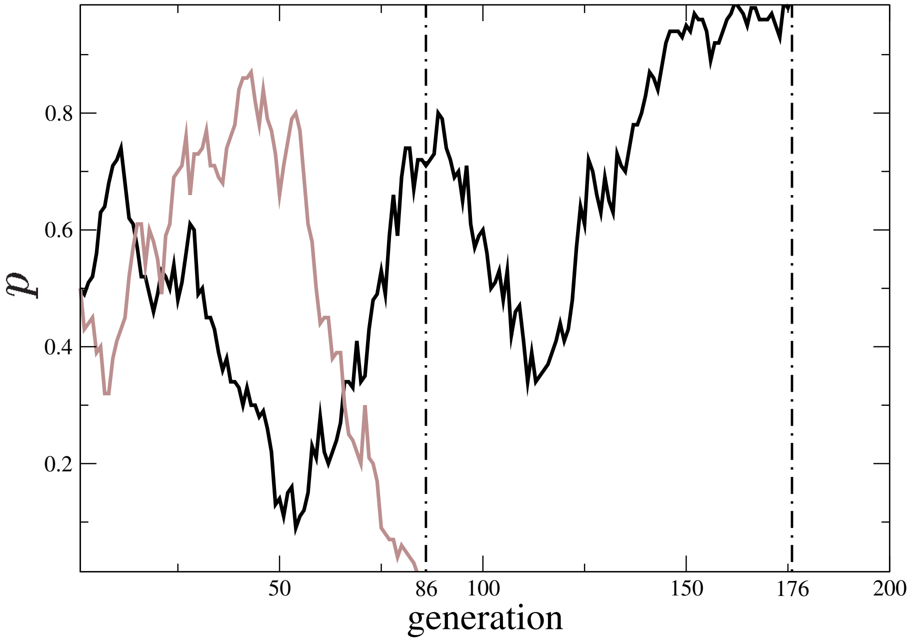

Fig. 10.6. Simulation of genetic drift over 200 generations in a two-allele system. In the gray trajectory allele A is lost at time $t = 86$, in the black trajectory allele A is fixed at $t = 176$. The initial frequency of allele A in both cases is $p(0) = 0.5$ and population size is $N = 100$.

$$\begin{aligned}\text{Prob}(0 < X(t+1) < 2N | p(t)) &= B(2N, p(t))(2N-1) - B(2N, p(t))(0)\\ &= 1 - p(t)^{2N} - (1-p(t))^{2N}\\ &\leq 1 - \left(\frac{1}{2N}\right)^{2N} .\end{aligned}$$

The last inequality is valid for any value of $p(t)$. Thus, the probability

$$\text{Prob}\,(0 < X(\tau) < 2N, \text{for all } \tau \leq t) \leq \left(1 - \left(\frac{1}{2N}\right)^{2N}\right)^{t} . \tag{10.3}$$

The limit of the right side of Equation (10.3) is 0 as t grows and N remains constant. The probability that X does not hit one of the boundaries before time t can be made arbitrarily small if t is allowed to be sufficiently large. □

Loss or fixation of an allele by drift is certain, but in large populations either outcome may take very long to reach. For large N the expression $(2N)^{-2N}$ is close to 0, thus the righthand side of Equation (10.3) is close to one. In the limit of infinitely large populations, there is no drift and allele frequencies stay constant in time. This is one of the two assertions of the Hardy-Weinberg law.

10.4.2 The Hardy-Weinberg Law

Consider again two alleles, A and a. Let their frequencies in an infinitely large, panmictic population be denoted by p_A and p_a. The three possible genotypes, the two homozygotes *AA* and *aa* and the heterozygote *Aa*, have frequencies p_{AA}, p_{Aa}, and p_{aa}. The Hardy-Weinberg law states that (i) from the second generation onwards the allele frequencies remain constant forever and that (ii) the genotype frequencies are uniquely determined by the allele frequencies and *vice versa* by the relationship

$$\begin{aligned} p_{AA} &= p_A^2 \\ p_{Aa} &= 2p_A p_a \\ p_{aa} &= p_a^2 . \end{aligned}$$

In other words, alleles are assembled independently into genotypes. The law is named after George Hardy and Wilhelm Weinberg, who formulated it independently in 1908. In practice, one often finds the Hardy-Weinberg law to be violated. This is mainly due to the fact that natural populations are not infinitely large. Other possible factors that may lead to deviations from the Hardy-Weinberg frequencies are differential action of natural selection upon different genotypes, non-random mating between individuals, or population substructure instead of panmixis.

10.4.3 Fixation Probability and Time to Fixation

We return now to a population of finite size N. In contrast to Claim 10.1, stating that either of two alleles will eventually be fixed, the following statement concerns the fixation probability of a particular allele.

Claim 10.2 *The fixation probability of an allele equals its current frequency* $p(t_0)$. *Formally,*

$$P_{\text{fix}} = \text{Prob}(X(t^*) = 2N \text{ for some } t^* > t_0 | p(t_0)) = p(t_0) . \tag{10.4}$$

PROOF. There are several proofs of this claim. Recurrence theory of finite state Markov chains provides the adequate framework and a detailed treatment of this can be found in a monograph by Ewens [65]. Here, we only note the following. As a consequence of Claim 10.1

$$P_{\text{fix}} = \text{Prob}(\lim_{t\to\infty} X(t) = 2N | p(t_0)) .$$

Furthermore, the limiting matrix $\mathcal{P}^\infty = \lim_{u\to\infty} \mathcal{P}^u$ has the form

$$\mathcal{P}^\infty = \begin{pmatrix} 1 & 0 \cdots 0 & 0 \\ \frac{2N-1}{2N} & 0 \cdots 0 & \frac{1}{2N} \\ \vdots & \vdots \;\vdots\; \vdots & \vdots \\ \frac{1}{2N} & 0 \cdots 0 & \frac{2N-1}{2N} \\ 0 & 0 \cdots 0 & 1 \end{pmatrix} .$$

Given that $p(t_0) = i/(2N)$ for some i, the initial distribution in the state space $\{0, 1, ..., 2N\}$ is $(0, ..., 0, 1, 0, ..., 0) = \pi_i$, where entry 1 is the i-th entry in this vector. By evaluating the product $\pi_i \cdot \mathcal{P}^\infty$ at its $2N$-th entry, one obtains

$$(\pi_i \cdot \mathcal{P}^\infty)_{2N} = \frac{i}{2N} \, .$$

□

In fact, this is true for any time t: given the process is in state i at time t, the fixation probability for allele A is equal to

$$\frac{i}{2N} .$$

The reason is, once again, the time-homogeneity of the underlying Markov process.

How long does it take for an allele to be fixed? To answer this question, the mean absorption time $\bar{t}_{\text{absorb}}(i)$ for, say, allele A with current frequency $p(t_0) = i/(2N)$, has to be determined. Note, that the conditional mean absorption time $\bar{t}_{\text{absorb}}(i)$ is different from the conditional mean fixation time, $\bar{t}_{\text{fix}}(i)$. For the latter, the additional condition that A will be fixed (and not lost) is imposed (Fig. 10.7). A rigorous treatment of this problem either within the theory of discrete Markov chains or con-

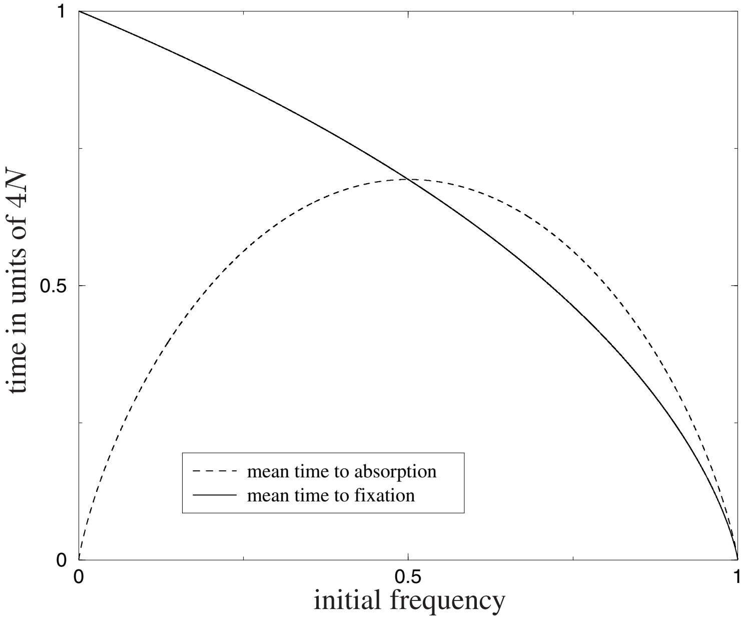

Fig. 10.7. Time to absorption and time to fixation of an allele as a function of its current frequency. Time is measured in units of $4N$ generations.

tinuous diffusion processes can be found in the book by Ewens [65]. However, there is a very instructive approximate solution to this problem. Let $x = i/(2N)$ and $x + \delta x = j/(2N)$. From Equation (10.1) one obtains for the expected change in allele frequency

$$E(\delta x) = E(\frac{j-i}{2N}) = \frac{1}{2N}\left(\sum_{k=0}^{2N} k\,\mathrm{Prob}(k, t+\delta t | i, t)\right) - \frac{i}{2N} = 0$$

and, similarly, for the second moment

$$E(\delta x)^2 = \frac{1}{4N^2}E((j-i)^2) = ... = \frac{x(1-x)}{2N}\,.$$

Treating allele frequencies as a continuous variable x, we write $t(x)$ instead of $t(i)$. Mean absorption time is $\bar{t}_{\text{absorb}}(x)$, given the process is currently in state x. At the next point in time, the process is expected to be in some state $E(x + \delta x)$ and the counter of the time units which the process has spent before reaching $x = 0$ or $x = 1$ is increased by one. Formally,

$$t(x) = 1 + E(t(x + \delta x))\,.$$

Assuming that $t(x)$ is twice differentiable we approximate the above equation by its Taylor series up to order two and obtain

$$\begin{aligned} t(x) &\approx 1 + E(\delta x)\left(t(x)' + \frac{1}{2}E(\delta x)^2(t(x))''\right) \\ &= 1 + t(x) + \frac{x(1-x)}{4N}(t(x))''\,. \end{aligned}$$

Thus, one has to solve

$$x(1-x)(t(x))'' = -4N \tag{10.5}$$

subject to the boundary conditions $t(0) = t(1) = 0$. The solution is

$$t(x) = -4N\left(x\log(x) + (1-x)\log(1-x)\right)\,. \tag{10.6}$$

With the above choice of boundary conditions, Equation (10.6) represents the mean time to absorption, $\bar{t}_{\text{absorb}}(x)$, given the current allele frequency is x. In particular, for a newly arising allele with frequency $x = 1/(2N)$, the mean absorption time is

$$\bar{t}_{\text{absorb}}(1/(2N)) \approx 2 + 2\log(2N)\,.$$

For the mean conditional fixation time $\bar{t}_{\text{fix}}(x)$ essentially the same arguments can be invoked. One only has to note that the transition probabilities p_{ij} (Eq. (10.1)) have to be replaced by the conditional transition probabilities p^*_{ij} that allele A will not be lost. At every generation, it has to be ensured that allele A leaves at least one offspring. Therefore, only $2N - 1$ alleles can be drawn at random and p^*_{ij} is the probability that among $2N - 1$ there are exactly $j - 1$ alleles of type A. Therefore, these transition probabilities are

$$p^*_{ij} = \binom{2N-1}{j-1} \left(\frac{i}{2N}\right)^{j-1} \left(1 - \frac{i}{2N}\right)^{2N-j} . \tag{10.7}$$

The differential equation, analogous to Equation (10.5), now reads

$$(1-x)(t(x))' + \frac{1}{2}x(1-x)(t(x))'' = -2N .$$

Its solution, subject to the boundary condition $t(1) = 0$, is

$$t(x) = -4N \left(\frac{1-x}{x} \log(1-x)\right) . \tag{10.8}$$

In particular, for a newly arising allele with frequency $x = 1/(2N)$, one obtains

$$\bar{t}_{\text{fix}}(1/(2N)) \approx 4N . \tag{10.9}$$

Finally we mention the average conditional time to loss of an allele; it is

$$\bar{t}_{\text{loss}}(x) = -\frac{4Nx}{1-x} \log(x) . \tag{10.10}$$

Given that the allele is lost and has frequency $x = 1/(2N)$, it takes only

$$\bar{t}_{\text{loss}}(1/(2N)) \approx 2\log(2N)$$

generations on average until it is lost.

An exact formulation of these arguments was provided by Kimura and Ohta [143] based on the theory of diffusion processes.

10.4.4 Loss of Genetic Diversity

We now return to the study of genetic diversity under a two-allele model. Allele A is sampled with probability p, allele a with probability $1 - p$. There are two ways in which two different alleles may be sampled: choose allele A first and then allele a or *vice versa*. Therefore, genetic diversity in the two-allele model is

$$H(t) = 2p(t)(1-p(t)) . \tag{10.11}$$

Note that $H(t)$ can also be interpreted as the expected *heterozygosity* (hence the H) of a diploid individual, that is, the probability that a diploid individual carries two different alleles. Let $X(t)$ be a binomially distributed random variable with parameters $2N$ and

$$p(t-1) = \frac{i}{2N}$$

and note that the second moment may be written in terms of the variance and the first moment as

$$E(X^2) = V(X) + (E(X))^2 \,.$$

The dynamics of H under random drift can then be determined as follows:

$$\begin{aligned}
H(t) &= 2p(t)(1-p(t)) \\
&= 2E\left(\frac{X(t)}{2N}\left(1-\frac{X(t)}{2N}\right)\right) \\
&= 2\left(\frac{E(X(t)}{2N} - \frac{E(X^2(t)}{(2N)^2}\right) \\
&= 2\left(\frac{i}{2N} - \frac{1}{(2N)^2}\left(2N\left(\frac{i}{2N}\left(1-\frac{i}{2N}\right)\right) + \left(2N\frac{i}{2N}\right)^2\right)\right) \\
&= 2\left(p(t-1) - \frac{1}{2N}(p(t-1)(1-p(t-1))\right) \\
&= 2\left(\left(1-\frac{1}{2N}\right)(p(t-1)(1-p(t-1))\right) \\
&= H(t-1)\left(1-\frac{1}{2N}\right) .
\end{aligned}$$

Iterating the above procedure, one finds

$$H(t) = H(0)\left(1-\frac{1}{2N}\right)^t \approx H(0)\exp\left(-\frac{t}{2N}\right) . \tag{10.12}$$

Thus, heterozygosity decays under random genetic drift at a rate of $1/(2N)$ per generation. This is another way of stating that drift reduces genetic diversity. As an immediate consequence of Equation (10.12), one finds the "half life" of heterozygosity

$$t_{h/2} = -\frac{\log(2)}{\log(1-1/(2N))} . \tag{10.13}$$

With the approximation $\log(1-1/(2N)) \approx -1/(2N)$, it follows that

$$t_{h/2} \approx 2N\ln(2) \,. \tag{10.14}$$

The half life of heterozygosity depends linearly on the population size. The above arguments all rest upon the assumption that there are no new mutations entering the population. But in reality, there is continuous influx of new alleles into a population, created by mutation. The key arguments of the neutral theory of molecular evolution rest upon the interplay of drift and mutation, which we consider next.

10.5 Adding Mutation to the Model

Depending on the degree of resolution with which molecular evolution is studied, different models of the mutation process are employed. Here, we describe the finite alleles model, the infinite alleles model, and the infinite sites model.

10.5.1 Finite Alleles Model

The finite alleles model, in particular the two-allele model, has originally been used to describe the evolutionary dynamics in situations where the alleles represent macroscopically observable phenotypes, such as eye or body color. The mutation dynamics in this case is modeled as switching between a given finite number of states. In particular, mutations can also be reversible and there is no influx of new alleles.

In fact, one-locus two-allele models with back mutation belong to the classical repertoire of theoretical population genetics and have been treated by all three founders of the subject, Haldane, Wright, and Fisher. In this case there exist nontrivial equilibria of the allele frequencies. In the presence of drift, there is a stationary density of the allele frequency distribution [262]: given a diploid population of size N, two alleles A_1 and A_2, and mutation rates μ_1 (for mutation from A_1 to A_2) and μ_2 (for mutation from A_2 to A_1) per chromosome per generation, the stationary distribution of the frequency x of allele A_1 is

$$f(x) = \frac{\Gamma(4N\mu_1 + 4N\mu_2)}{\Gamma(4N\mu_1)\Gamma(4N\mu_2)} x^{4N\mu_2 - 1}(1-x)^{4N\mu_1 - 1} .$$

The mean and variance are

$$E(f(x)) = \frac{\mu_2}{\mu_1 + \mu_2}$$

and

$$V(f(x)) = \frac{\mu_1\mu_2}{(\mu_1 + \mu_2)^2(4N\mu_1 + 4N\mu_2 + 1)} .$$

If $\mu_1 = \mu_2 = \mu$, then

$$f(x) = \frac{\Gamma(8N\mu)}{(\Gamma(4N\mu))^2} x^{4N\mu - 1}(1-x)^{4N\mu - 1} .$$

Abbreviating $4N\mu$ by θ, the probability that two chromosomes, drawn at random from the population, are of the same allelic type, is

$$\int_0^1 (x^2 + (1-x)^2) df(x) = \frac{1+\theta}{1+2\theta} .$$

This expression can also be interpreted as the expected fraction of homozygotes in the population. The expected fraction of heterozygotes is then

$$E(H) = \frac{\theta}{1+2\theta} .$$

With the advent of protein electrophoresis in the 1960s and the observation of unexpectedly high diversity within species [167, 165], the finite alleles model started to be superseded by the infinite alleles model.

10.5.2 Infinite Alleles Model

With the infinite alleles model [142] it is possible to account for continuous mutational influx into a population. In fact, its name derives from the assumption that any mutation event creates a new allele, which was previously not seen in the population. There is no back mutation to previous states. As a consequence, two alleles can be identical only due to shared ancestry rather than chance mutation to the same allele. Hence identical alleles are said to be *identical by descent*. In the framework of the infinite alleles model it is possible to decide whether any two alleles are identical or not. However, it is not possible to quantify the difference, for instance in terms of the number of nucleotide mismatches between two different alleles. Put more formally, allele space is a metric space with the trivial metric

$$d(A_i, A_j) = \begin{cases} 0, & \text{if } A_i = A_j \\ 1, & \text{if } A_i \neq A_j. \end{cases}$$

The infinite alleles model is reasonable under many real-world scenarios, when the level of resolution is fine enough. As an example, consider again the data from Kreitman shown in Table 10.1. While he found only two electrophoretic variants in his sample of eleven genes, the number of alleles which are distinguishable at the level of nucleotides was nine. To convince yourself why back mutation can be neglected when dealing with (infinitely) many alleles, consider a sequence of 180 nucleotides. Once this sequence has started mutating, it is unlikely to ever return to its original state, as there are $4^{180} \approx 2 \cdot 10^{108}$ distinct sequences it can reach. Compare this to the estimate that our universe has a volume of 10^{108} Å^3 [59, p.35] and you can appreciate that the space of possible nucleotide sequences is truly vast even for sequences of only moderate length.

10.5.3 Infinite Sites Model

Under the infinite sites model each mutation affects a different position along a stretch of DNA that has never mutated before. This model is realistic when describing and interpreting the evolution of sets of DNA sequences, where the mutation rate per site is low. Allele space in the infinite sites model is equipped with a non-trivial metric. The natural distance between any two sequences is the number of sites at which the two aligned sequences differ. If all sequences have the same length and are alignable without gaps, this metric is also called Hamming distance.

10.6 Mutation Drift Balance

10.6.1 The Rate of Fixation

How fast and how often new alleles are fixed is determined by the time to fixation and the rate of fixation. Figure 10.8 shows sample trajectories of newly arising alleles which eventually reach fixation. Under the infinite alleles or infinite sites models

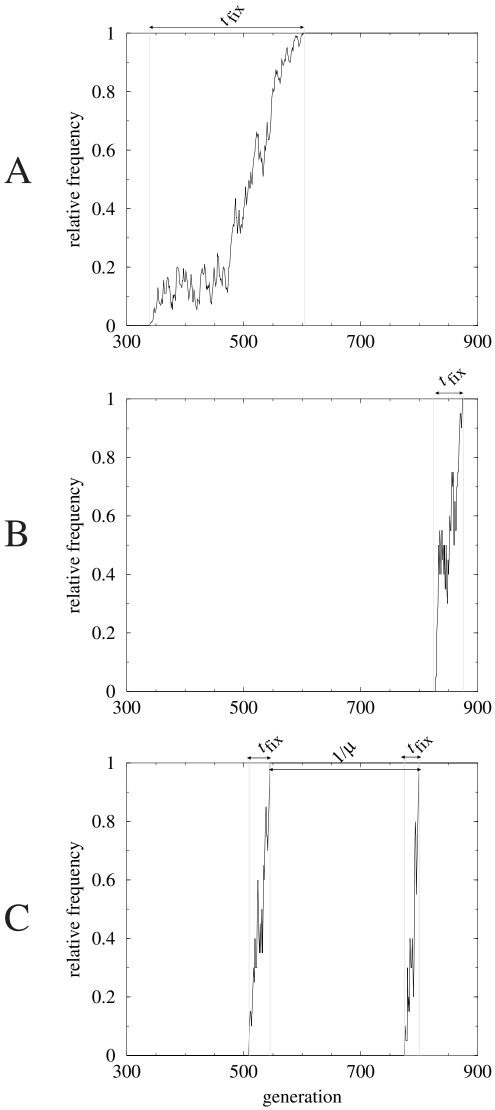

Fig. 10.8. Fixation of new alleles under an infinite alleles or infinite sites model. X-axis: time in generations, Y-axis: relative frequency of mutant alleles. Shown are only the trajectories of those alleles which eventually reach fixation. **A**: $N = 100$, mutation rate $\mu = 10^{-3}$; **B**: $N = 10$, mutation rate $\mu = 10^{-3}$; **C**: $N = 10$, $\mu = 10^{-2}$. The time to fixation (t_{fix}) depends on N (compare **A** and **B**), while the rate of fixation depends on μ (compare **B** and **C**). The average time to fixation is $t_{\text{fix}} = 4N$ generations. At any given time a random number κ of alleles is present in the population. This number depends on N and μ. Its lower bound estimate is $\underline{\kappa} = 1 + 4N\mu$ (Eq (10.20)). For our examples $\kappa = 1.4$ (**A**), 1.04 (**B**) and 1.4 (**C**). Note that the figures do not display the trajectories that do not reach fixation, which are nevertheless included in this estimate.

and in a diploid population, $2N\mu$ new alleles are generated independently in each generation. Any individual allele is eventually either lost or fixed. There is no equilibrium frequency of individual alleles different from 0 or 1. The rate of fixation is the product of newly generated mutants per generation times their probability of fixation,

$$2N\mu \cdot \frac{1}{2N} = \mu.$$

This is the rate of molecular evolution, or rate of substitution, introduced in Chapter 9.

The equality of substitution and mutation rates under the neutral infinite alleles model is an immediate, but nevertheless surprising result given our intuition that the mutation rate is some sort of universal constant, while the rate of substitution should be a function of the population size. Both intuitions are correct and under neutrality the corresponding terms cancel out. This calculation does not hold if mutations are non-neutral.

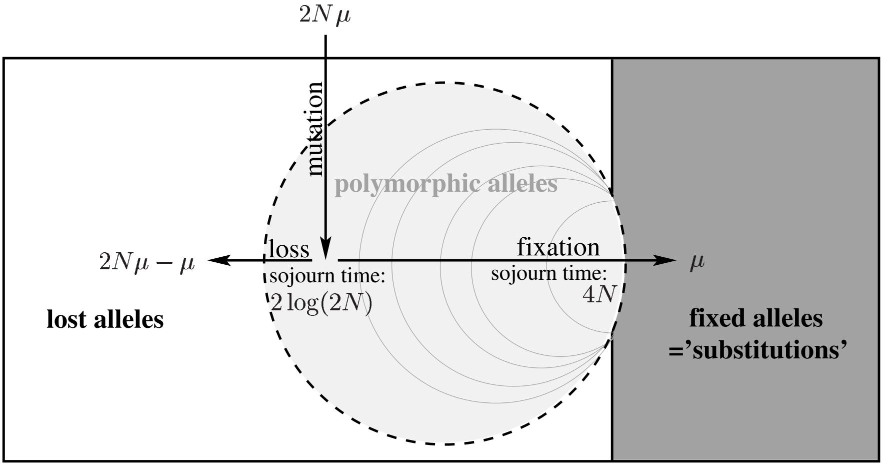

Fig. 10.9. Mutation drift balance may be visualized as a balance of influx and outflux of alleles. The mean sojourn time of a new allele destined to be lost is $2\log(2N)$ (Eq. (10.10)). The mean time to fixation is $4N$ (Eq. (10.8)).

10.6.2 Number of Alleles

We have seen that each individual allele is either lost or goes to fixation. However, as genetic diversity is lost through drift, new diversity is generated through mutation and these two factors lead to an equilibrium between the influx of mutations into the gene pool and their subsequent outflux (Fig. 10.9). We can now ask, is there an equilibrium under mutation and drift for the number of alleles which are simultaneously present in the population at any given point in time? A lower bound estimate of this number was derived by Kimura and Crow [142]. Consider two randomly chosen alleles in generation t. Let $F(t)$ be the probability that they are identical by descent, i.e. derived from either the same parent or from two alleles which were already identical by descent in generation $t-1$. The quantity F is also called *homozygosity*. The probability that two alleles are identical by descent in generation t is derived as follows. Either they have descended from the same allele in generation $t-1$. The probability of this is $p = 1/(2N)$ in a diploid population of N individuals; or they have not descended from the same allele in generation $t-1$ ($q = 1 - 1/(2N)$), but have descended from the same allele in some previous generation; by definition this is $F(t-1)$. In both cases neither of the two alleles is allowed to mutate while being passed from generation $t-1$ to generation t. The probability of this is $(1-\mu)^2$. Summarizing, we can write

$$F(t) = (1-\mu)^2 \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F(t-1) \right). \tag{10.15}$$

In order to solve Equation (10.15) for $F(t)$, we simplify the algebra by remembering that biologically reasonable mutation rates are small and populations large. Hence, we ignore terms multiplied by μ^2 and μ/N and write

$$F(t) \approx \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F(t-1) - 2\mu F(t-1). \tag{10.16}$$

In an equilibrated population the fraction of alleles which are identical by descent remains constant between generations, i.e. $F(t-1) = F(t) = \bar{F}$. Thus, when replacing $F(t-1)$ and $F(t)$ in Equation (10.16) by $\bar{F}$, equilibrium homozygosity can be written as

$$\bar{F} = \frac{1}{1 + 4N\mu}. \tag{10.17}$$

In case the number of alleles and their frequencies are known, homozygosity is calculated as

$$F = \sum_{i=1}^{\kappa} x_i^2, \tag{10.18}$$

where κ is the number of alleles and x_i is the frequency of allele i. The proportion of homozygotes (F) in the population is minimal if all alleles are equally frequent,

$$x_i = \frac{1}{\kappa}$$

for all i. In this situation, one has

$$F = \frac{1}{\kappa} = \frac{1}{1 + 4N\mu} \tag{10.19}$$

and a lower bound estimate for the number of alleles in an equilibrated population therefore is

$$\underline{\kappa} = 1 + 4N\mu\,. \tag{10.20}$$

This number has been called the *effective number of alleles* by Kimura and Crow [142] and numerical values for this are shown in Figure 10.8. Note that the scaled mutation rate, $\theta = 4N\mu$, which plays a central role in population genetics theory, is featuring again in this formula. For non-uniformly distributed allele frequencies the number of alleles is larger than $\underline{\kappa}$. In Section 10.7 we will come back to a formula for the number of alleles in a sample drawn from an equilibrated population.

10.6.3 Genetic Diversity

There are several measures of genetic diversity. One of them is the number of different alleles in a population. This statistic has the advantage that it is easy to measure. However, its disadvantage is that it is dependent on the size of the sample from which diversity is estimated. This makes it hard to compare measurements which are based on different samples with different sizes.

Another measure of genetic diversity is the proportion of heterozygotes in a population or, equivalently, the probability with which two randomly chosen alleles are different. In case of known alleles and allele frequencies, heterozygosity is

$$H = 1 - \sum_i x_i^2\,.$$

On the other hand, and as an immediate consequence of Equation (10.17), equilibrium heterozygosity is

$$\bar{H} = 1 - \bar{F} = \frac{4N\mu}{1 + 4N\mu} = \frac{\theta}{1+\theta}\,. \tag{10.21}$$

This result entails one of the corner stones of the neutral theory of evolution [139]. Diversity in neutrally evolving populations is monotonically increasing with the scaled mutation rate θ. In fact, if $\theta \ll 1$, then $H \approx \theta$. Equation (10.21) applies to average heterozygosity. Under the action of mutation and drift, heterozygosity is a Markov process with stationary transition probabilities. For each point in time, heterozygosity is a random variable, $H(t)$. Three realizations of this stochastic process under the infinite alleles model are shown in Figure 10.10. It demonstrates that heterozygosity as a random variable in time fluctuates around the value calculated in Equation (10.21). The entire distribution of H at a single locus and under the infinite alleles model has been derived by Fuerst and colleagues [82]. In particular, for the variance, $V(H)$, they obtained

$$V(H) = \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}\,. \tag{10.22}$$

Equation (10.21) plays a central role and will be rederived with different arguments in Chapter 11. We consider further its implications. Notice first that Equation (10.21) establishes a relationship between heterozygosity H, which can be

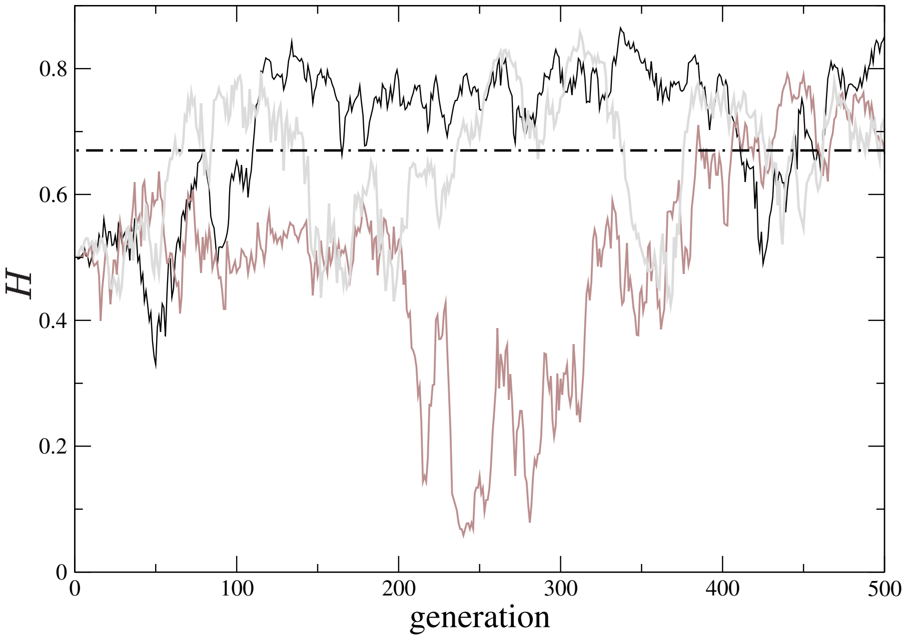

Fig. 10.10. Simulation of heterozygosity at three independent loci evolving under the infinite alleles model. Each trajectory represents the heterozygosity $H(t)$ at one locus. Input parameters are $N = 50$ and $\mu = 0.01$, which implies a mean heterozygosity of $H = \frac{4N\mu}{4N\mu+1} = 0.67$ (dotted line). Notice the fluctuations around the expected heterozygosity. For the parameters considered the standard deviation is 0.15 (Eq. (10.22)).

measured, and the unknown quantity $4N\mu$. As an application consider the data set shown in Table 10.1. If we ignore indels, we observe eight unique alleles and one allele which occurs three times (Fr-F, Wa-F, Af-F). Hence, the observed heterozygosity is $H_{\text{obs}} = (1 - 8 \cdot (1/11)^2 - 1 \cdot (3/11)^2) \cdot 11/10 = 0.95$, where the factor $n/(n-1) = 11/10$ corrects for the fact that we are using sample frequencies rather than population frequencies [121]. Based on Equation (10.21) one would estimate the scaled mutation rate to be $\hat{\theta} = 17.33$, where the hat indicates that the value is an estimate. Second, there are estimates of the (unscaled) mutation rate μ available [51], which means that Equation (10.21) can be used to calculate the population size. For *Drosophila*, Drake and his colleagues reported a mutation rate of $\mu = 8.5 \cdot 10^{-9}$ per bp per generation. Thus, the mutation rate for the 2,721 bp investigated in the *Adh* region would be $2.31 \cdot 10^{-5}$ per generation. This yields an estimate of the population size $\hat{N} \approx 187,000$. We need to stress that population size refers here, as always in our book, to the *effective* population size, not the census population size. Roughly, effective population size is the number of reproducing individuals. This number is generally much smaller than the actual number of individuals. Finally, Equation (10.21) predicts that heterozygosity approaches unity with increasing population size. This prediction was central to attempts in the 1970s to falsify the neutral theory by measuring H in the largest populations known, those of bacteria. The first of these studies came to the conclusion that for *Escherichia coli* $H \approx 0.2$, which is far from unity [182]. However, the estimation of N from H is fraught with difficulties. One of these follows from the shape of the graph of H as a function of $4N\mu$ shown in Figure 10.11. For large values of H, N cannot be predicted very accurately as the curve flattens out. Moreover, Equation (10.21) is based on the assumption of free recombination, which implies that all loci evolve independently of each other. This assumption is not valid for asexual organisms such as *E. coli* [222, 175].

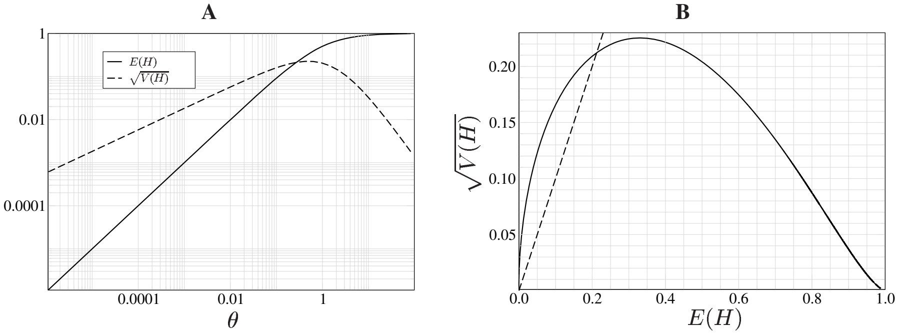

Fig. 10.11. A: Expected heterozygosity $E(H)$ and standard deviation $\sqrt{V(H)}$ as function of the scaled mutation rate $\theta = 4N\mu$ under the neutral infinite alleles model (Eqs. (10.21) and (10.22)). **B:** Standard deviation as function of expected heterozygosity. The dashed line is a plot of $x = y$ and shows that $\sqrt{V(H)} > E(H)$ for small values of $E(H)$.

10.7 Sampling Alleles from Populations

10.7.1 Ewens' Sampling Formula

A cornerstone of the study of the infinite alleles model is Ewens' sampling formula [64]. Consider a population in equilibrium under the neutral infinite alleles model and a sample of size n drawn from the population. The sample configuration is a vector $\mathbf{C}(n) = (C_1(n), ..., C_n(n))$, where the components $C_j(n)$ denote the number of alleles represented exactly j times in the sample. Then, $\sum_{j=1}^{n} jC_j(n) = n$. Ewens' sampling formula gives the probability distribution for the sample configuration. The probability for a particular configuration $\mathbf{c}(n) = (c_1(n), c_2(n), ..., c_n(n))$, where $c_j(n)$ are non-negative integers with the property $n = \sum_{j=1}^{n} jc_j(n)$, is

$$P(\mathbf{C}(n) = \mathbf{c}(n)) = \frac{n!\theta^{\sum_j c_j(n)}}{\prod_j j^{c_j(n)} \prod_j c_j(n)!\theta_{(n)}}, \tag{10.23}$$

where $\theta_{(n)} = \theta(\theta+1)...(\theta+n-1)$. Let further $\mathcal{K}_n$ be the number of distinct alleles in a sample of n. Then, $\sum_{j=1}^{n} C_j(n) = \mathcal{K}_n$. The distribution of $\mathcal{K}_n$ is

$$P(\mathcal{K}_n = \kappa) = \frac{\text{coeff}(\theta(n), \kappa)\theta^{\kappa}}{\theta_{(n)}} ,$$

where $\text{coeff}(\theta(n), \kappa)$ is the coefficient of θ^{κ} in the expansion of $\theta_{(n)}$.

Ewens' sampling formula is the basis for one possible test of the null hypothesis of neutral evolution. More tests of this hypothesis are discussed in Chapter 12. Here, we apply Equation (10.23) to the experimental data shown in Table 10.1. In this data set sample size is $n = 11$; therefore, we are dealing with a vector $\mathbf{c}(11)$ of the form $\mathbf{c}^*(11) = (8, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$. The probability for this configuration is

$$P(\mathbf{C}(11) = \mathbf{c}^*(11)) = \frac{330 \cdot \theta^8}{\theta(11)} .$$

For sample size $n = 11$, there are altogether 56 possible configurations. The number of possible configurations is equal to the number of decompositions of n into integer summands without regard of order [1, sec. 24.2.1]. The probabilities for each configuration depend only on θ. Figure 10.12 shows the probability distribution for different numerical values of θ. The observed configuration is indicated by a vertical black line. Crucial for deciding whether the observation is compatible with the model of neutral evolution is an accurate estimate of θ. One can show that all information

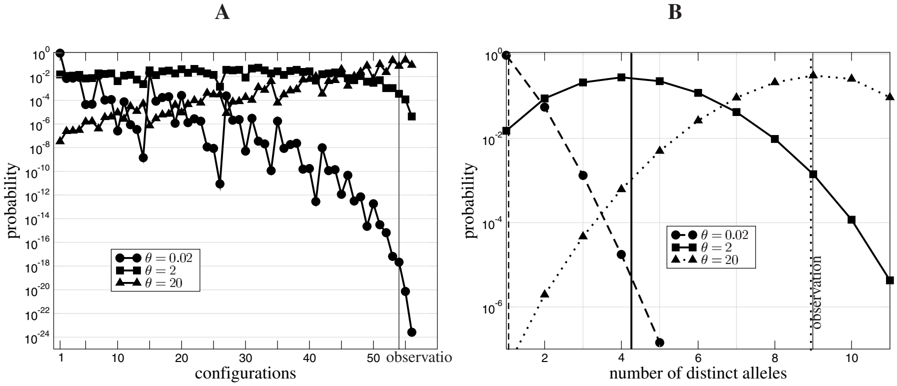

Fig. 10.12. **A**: Probability distribution of the possible configurations for a sample of size $n = 11$ and different values of θ. Configurations are ordered lexicographically. For instance, configuration 1 corresponds to the vector $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$ (i.e. all alleles are identical), configuration 2 to $(0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0)$, and configuration 54 to the observed configuration $(8, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$ (i.e. there are eight singletons and one cluster containing three alleles). **B**: Probability distribution for $\mathcal{K}_{11}$ and different values for θ. The observed number of distinct alleles (9) is indicated by the black vertical line. Vertical lines represent the mean values $E(\mathcal{K}_{11})$ for the three choices of θ.

about θ which is deducible from a configuration is already contained in the number of distinct alleles $\mathcal{K}_n$, such that explicit knowledge of the configuration is not needed. In statistical terms, $\mathcal{K}_n$ is a sufficient statistic for θ. Indeed, the conditional probabilities $\mathrm{Prob}(\mathbf{C}(n) = \mathbf{c}(n)|\mathcal{K}_n = \kappa)$ are independent of θ. It holds that

$$\mathrm{Prob}(\mathbf{C}(n) = \mathbf{c}(n)|\mathcal{K}_n = \kappa) = \frac{n!}{\mathrm{coeff}(\theta(n), \kappa) \prod_j j^{c_j(n)} \prod_j c_j(n)!} . \tag{10.24}$$

Thus, when writing the probability $\mathrm{Prob}(\mathbf{C}(n) = \mathbf{c}(n) \wedge \mathcal{K}_n = \kappa)$ as a product of $\mathrm{Prob}(\mathbf{C}(n) = \mathbf{c}(n)|\mathcal{K}_n = \kappa)$ and $\mathrm{Prob}(\mathcal{K}_n = \kappa)$, only the last factor depends on θ. This implies that a maximum-likelihood estimate of θ can be based on the probability distribution of $\mathcal{K}_n$. Given an observation $\mathcal{K}_n = \kappa$, it can be shown that this maximum likelihood estimate is the solution of

$$\kappa = \sum_{i=0}^{n-1} \frac{\hat{\theta}}{\hat{\theta} + i} . \tag{10.25}$$

Furthermore, using the fact that under neutral evolution $\mathrm{Prob}(\mathbf{C}(n) = \mathbf{c}(n)|\mathcal{K}_n = \kappa)$ is independent of θ, one can construct a statistical test for the null hypothesis of neutral evolution. Unlikely configurations of $\mathbf{C}(n)$ given $\mathcal{K}_n$ would lead to a rejection of the null hypothesis. This idea is formalized in the Ewens-Watterson test.

Apart from the computation of the probability of an observed allele configuration under the infinite alleles model, Equation (10.23) allows one to deduce the expected number of singletons and the expected number of different alleles. The expected number of singletons in a sample of n genes is

$$E(C_1(n)) = \sum_{\text{all configurations } \mathbf{c}(n)} C_1(n) P(\mathbf{C}(n) = \mathbf{c}(n)) = \frac{n\theta}{n - 1 + \theta} . \tag{10.26}$$

The expected number of different alleles in a sample of n is

$$E(\mathcal{K}_n) = \sum_{\kappa=1}^{n} \kappa P(\mathcal{K}_n = \kappa) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} , \tag{10.27}$$

where the last equality requires some arithmetical manipulations. In contrast, the formula given by Kimura and Crow (Eq. (10.20)) is a lower bound estimate for the number of simultaneously existing alleles in the population. In fact, $E(\mathcal{K}_n)$ is larger than $\underline{\kappa}$ only if n is sufficiently large. Furthermore, for large values of θ the expected number of alleles in a sample, $E(\mathcal{K}_n)$, may be smaller than the lower bound estimate for the number of alleles in the population, $\underline{\kappa}$ (see Table 10.2).

Finally, we note the variance of $\mathcal{K}_n$. It is

$$V(\mathcal{K}_n) = \sum_{i=0}^{n-1} \frac{\theta i}{(\theta + i)^2} .$$

10.7.2 Application

For the data presented in Table 10.1, some values of $E(C_1(n))$ and $E(\mathcal{K}_n)$ are shown in Table 10.2. The observed number of alleles is $\kappa = 9$. There are eight singletons and one set of three identical alleles. There are two possible configurations with $\kappa = 9$. One of them—the one observed—is $c_1(11) = (8, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$, the other possible one is $c_2(11) = (7, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0)$. From Equation (10.24) the conditional probabilities are

$$\text{Prob}(\mathbf{C}(11) = c_1(11)|\mathcal{K}_{11} = 9) = 0.25\,,$$
$$\text{Prob}(\mathbf{C}(11) = c_2(11)|\mathcal{K}_{11} = 9) = 0.75\,.$$

The maximum likelihood estimator for θ derived from Equation (10.25) is

$$\hat{\theta} = 20.73\,.$$

The observed allele configuration appears to agree well with the expected results for $E(\mathcal{K}_{11})$ and $E(C_1(11))$ for the parameter $\theta = 20$ (Fig. 10.12). On the other hand, 43 out of a total of 2,721 sites were polymorphic (or "segregating"), which is somewhat less than the 58.6 segregating sites expected under neutrality (Table 10.1). Is $\theta = 20$ a reasonable estimate for *D. melanogaster* and a stretch of 2.7 kb of its genome or do the data have to be explained by other evolutionary mechanisms than mutation and drift alone? After all, the two estimates of θ, obtained from Equation (10.21), which was $\hat{\theta} = 17.33$, and from Equation (10.27), which was $\hat{\theta} = 20.73$, are similar but not identical.

Table 10.2. Numerical values for the expected number of alleles, number of segregating sites and singletons of the data presented in Table 10.1.

θ	$\underline{\kappa}$†	$E(\mathcal{K}_{11})$	$V(\mathcal{K}_{11})$	S‡	$E(C_1(11))$
0.02	1.020	1.058	0.057	0.586	0.022
2.00	3.000	4.206	1.947	5.858	1.833
6.12	7.120	6.629	2.234	17.925	4.176
20.0	21.00	8.945	1.550	58.579	7.333

†lower bound estimate according to Equation (10.20).
‡expected number of segregating sites according to Equation (11.8).

We will come back to the problem of testing the significance of the difference between the neutral expectation and observation in Chapters 11 and 12.

10.8 Selection

The neutral theory of molecular evolution should not detract us for too long from the really interesting question, namely how is adaptation of a species to certain environmental conditions possible? Most answers contain in one way or the other the concept of an “advantageous mutation” [76]. An organism carrying an advantageous mutation, i.e. a genetic constitution which makes it better adapted to the given environmental conditions, will on average leave more offspring to the next generation than its competitors. Such an organism is said to have an increased fitness. In a very large population, which is the case envisaged by Fisher, the dynamics of such an advantageous mutation can essentially be treated in a deterministic framework. The following model captures the salient features of the spread of an advantageous mutation within a population.

Let x and y be the relative frequencies of the two alleles “advantageous mutation” (allele B) and “wild-type” (allele b), respectively. As always in a two-allele model, $x + y = 1$. For diploid organisms, i.e. organisms which have two copies of each chromosome, let the fitnesses of the three genotypes bb, Bb and BB be

BB	Bb	bb
$1+2s$	$1+s$	1

In other words, organisms with two copies of the advantageous mutation have a fitness difference of $2s$ compared to those which have two wild-type chromosomes. The dynamics of the B-type chromosomes is described by a difference equation which accounts for the change in frequency of type B when passing from one generation to the next

$$\Delta x_t = x_{t+1} - x_t = \frac{(1+2s)x_t^2 + (1+s)x_t y_t}{(1+2s)x_t^2 + 2(1+s)x_t y_t + y_t^2} - x_t \,. \tag{10.28}$$

Note, that the denominator on the righthand side represents the average fitness of the population, where the average is taken over the three possible genotypes. In fact, this term is called *mean fitness*. The difference equation (10.28) is approximated by an ordinary differential equation, the so-called deterministic selection equation

$$\frac{dx}{dt} = sx(t)(1 - x(t)) \tag{10.29}$$

with the initial condition

$$x(0) = \epsilon \,.$$

The solution of this differential equation is

$$x(t) = \frac{\epsilon}{\epsilon + (1-\epsilon)\exp(-st)} \,. \tag{10.30}$$

Since the advantageous mutation is initially present only on a single chromosome, one typically assumes $\epsilon = 1/(2N)$, where N is the (effective) population size. The time which the advantageous mutation requires to increase from frequency ϵ to $1-\epsilon$ is obtained by solving

$$1 - \epsilon = \frac{\epsilon}{\epsilon + (1-\epsilon)\exp(-st)}$$

for t. This is

$$t_{(\epsilon)}^{(1-\epsilon)} = \frac{2}{s} \log \frac{1-\epsilon}{\epsilon} \approx \frac{2}{s} \log(2N)\,. \tag{10.31}$$

It can be shown that the righthand side of Equation (10.31) is a good approximation of the fixation time, t_{fix}, for an advantageous allele also in a non-deterministic framework.

For the above fitness regime, mean fitness simplifies to

$$\bar{w}(t) = 1 + 2sx(t)\,.$$

Since $x(t)$ is a monotonically increasing function in t, mean fitness also has this property. In fact, Fisher's fundamental theorem of natural selection states that "mean fitness of a population increases". In other words, evolution is a hill-climbing process towards a fitness peak. However, it has to be emphasized that this is true only for the deterministic dynamics of infinitely large populations. In finite sized populations, positive selection and genetic drift act antagonistically. While a positively selected, newly arisen, allele is still rare in a population, there is a non-zero probability that this allele will be lost again by drift. It can be shown [65] that the fixation probability of an advantageous allele is goverend by its selective advantage and that it is approximately

$$\mathrm{Prob}_{\mathrm{fix}} = 2s\,. \tag{10.32}$$

A positively selected allele initially increases at a rate s per generation (see Eq. (10.30)), while diversity is lost due to drift at a rate of $1/(2N)$ per generation (see Eq. (10.12)). Crucial for positive selection to prevail as an evolutionary force over drift is the relation

$$2Ns > 1\,. \tag{10.33}$$

Thus, rather than the net fitness advantage alone, it is its relation to population size that is decisive for the possibility of adaptation. In sufficiently large populations, there is a chance even for weakly selected alleles to become fixed. The issue whether the "normal" mode of genome evolution is evolution under weak positive selection is the matter of an old debate [153, 115]. One of the problems, implicated by Equation (10.33), is that the action of weak selection is hard to prove in laboratory experiments. The quest remains to distinguish the action of selection from "molecular noise". This difficulty has been called the "uncertainty principle of molecular evolution" [235, 242].

The combined action of mutation, selection, and drift introduces rich dynamics into the evolutionary system which allows for a variety of non-trivial equilibria depending on the relative magnitudes of the selection coefficient, dominance effects, mutation rate, and population size [32, 14]. A good part of the mathematical foundation for these models has been laid out by Haldane, Fisher, and Wright in the first half of the past century [100, 101, 76, 262].

10.9 Summary

This chapter is concerned with the dynamics of genes in populations. The neutral Wright-Fisher model provides a convenient framework for describing such dynamics. This model is based on a population of constant size that evolves from one generation to the next by sampling the genes with replacement. Neutrality then simply corresponds to the fact that all genes are equally likely to get transferred to the next generation. Changes in allele frequencies are random and said to be due entirely to genetic drift. Given such a model without mutation, any existing genetic diversity will eventually be lost from the population. In other words, there will be only one allele for each gene. The probability that a certain allele replaces all others, i.e. is fixed, is equal to the initial frequency of the allele. If we add mutation to the model, the mutation process can be described in three different ways: (i) under the finite alleles model each mutation leads to a random switch between k alleles; (ii) under the infinite alleles model each mutation generates a new allele; (iii) under the infinite sites model each mutation affects a different site along a stretch of sequence. Mutations create new alleles, while drift removes alleles from the gene pool and these two factors can come into a mutation drift balance. If mutation is balanced by drift, it is possible to compute the number of alleles expected to occur in a population as well as the probability of drawing a pair of distinct alleles, also known as the genetic diversity. Since the number of alleles can easily be observed, a comparison between observed and expected allele configurations is used as a test of neutral evolution.

10.10 Further Reading

This chapter is mainly based on the comprehensive textbook by Hartl and Clark [106] as well as on the concise primer by Gillespie [87]. Either one of these books should be consulted for further details and references on the foundations of population genetics. Ewens [65] gives a mathematically advanced treatment of Markov chain theory and diffusion models in population genetics. Finally, Provine describes the history of population genetics by way of an illuminating biography of Sewall Wright [204].

10.11 Exercises and Software Demonstration

10.1. The `bioinformer` software contains under `Evolution` → `Drift` a program that shows forward in time simulations. Work through the tutorial for that program in Section A.4.2.

10.2. In the earliest attempt to test the validity of the neutral hypothesis using bacterial populations, a *virtual* heterozygosity of $H = 0.242$ was measured in *E. coli* [182]. Why was the genetic diversity in *E. coli* referred to as "virtual" heterozygosity?

10.3. In the *E. coli* study a mutation rate of $\mu = 10^{-8}$ was assumed. Which population size follows from this under the neutral infinite alleles model?

10.4. Based on considerations independent of the genetic diversity, the population size of *E. coli* was estimated to be 10^{10}. Given this estimate, what is the genetic diversity, H, expected under the neutral infinite alleles model?

10.5. The population size relevant for population genetics is usually referred to as *effective* population size, N_e. It may differ widely from a population's head count. Essentially, N_e is the size an idealized population would have, given the effect of drift on its allele distribution [106, p. 289]. A population study of *E. coli* has come to the conclusion that $N_e = 1.8 \cdot 10^8$, which is surprisingly small [107]. What is the value of H expected from this estimate of population size?

10.6. Show that allele frequencies do not change between generations when $N \to \infty$, i.e. show that the Hardy-Weinberg law holds if the population size is infinitely large. Hint: Start from Equation (10.1).

10.7. Consider a haploid population with $N = 50$. What is the probability that two randomly selected alleles had their last common ancestor exactly 1,000 generations ago?

10.8. In Section 10.7.1 a value of $\hat{\theta} = 20.73$ was found based on the assumption that there are nine alleles in the sample of eleven *Adh* sequences of *D. melanogaster*. Remember that there are ten alleles, if indel ∇_3 is included in the analysis. Estimate θ on the basis of ten alleles.

10.9. Consider a diploid population with k alleles. Show that the proportion of homozygotes is minimal if all alleles are equally frequent.