4. Populationsgenetik

Populations are never uniform, but individuals differ genetically and phenotypically. Population genetics is concerned with the study of the genetic composition of populations and how it evolves. This composition may be changed by segregation, selection, mutation, recombination, mating behavior, migration, and other genetic, ecological, and evolutionary factors. Therefore, in population genetics these mechanisms and their interactions and evolutionary consequences are investigated. Traditionally, population genetics has been applied to animal and plant breeding, to human genetics, and more recently to ecology and conservation biology. It also has important interfaces with molecular biology, systematics, natural history, mathematics, statistics, and computing. One of the main subjects is the investigation of the mechanisms that generate and maintain genetic variability in populations, and the study of how this genetic variation, shaped by environmental influences, leads to evolutionary change, adaptation, and speciation. Therefore, research in population genetics relies on empirical observations, on experiments, and on theoretical considerations. In particular, population genetics provides the basis for understanding the evolutionary processes that have led to the diversity of life we encounter and admire.

Since so many factors interact and determine the evolutionary fate of a population, a proper understanding of the relevant processes requires a good deal of abstraction in planning experiments and in devising mathematical models. A good mathematical model, as well as a good experiment, takes into account the relevant biological mechanisms for studying a particular phenomenon and disregards the less relevant ones. As in other sciences, good model building must rest on an adequate knowledge of the basic biological reality and requires a clear formulation of the underlying hypotheses. The process of abstraction that is involved entails generality which, sometimes, may appear to be unnecessary. However, general methods or models, devised to study a particular phenomenon, may reveal the essence and the underlying structure more clearly and can often be applied to questions not anticipated before.

Mathematical models and methods have a long history in population genetics, tracing back to Gregor Mendel, who used elementary mathematics to calculate the expected frequencies of the genes in his experiments. Francis Galton and the biometricians, notably Karl Pearson, developed new statistical methods to describe the distribution of trait values in populations and to predict their change between generations. The foundations

of modern population genetics were laid by the work of Ronald A. Fisher, J.B.S. Haldane, and Sewall Wright, who reconciled Mendelism with Darwinism during the second and third decades of the twentieth century. They demonstrated that the theory of evolution by natural selection, proposed by Charles Darwin (1859), can be justified on the basis of genetics as governed by Mendel's laws. The work of Fisher, Haldane, and Wright was highly mathematical for the biology of that time and was properly understood by only a small number of people. Nevertheless, their influence was enormous and they set the standards for mathematical modeling and for rigor of theoretical investigations for the subsequent decades.

Prior to 1900, the year when Mendel's work was rediscovered and then rapidly accepted, the hereditary mechanisms were unknown. Darwin believed in blending inheritance, according to which the hereditary material itself blended. However, as already noted by Darwin, blending inheritance produces uniformity and destroys variation that is so ubiquitous. In modern terms, heritable variance would be halved in each generation of random mating with blending inheritance (Fisher 1930). Therefore, one half of the heritable variance maintained in a population would have to arise anew in each generation. There were controversial lines of thought about the nature of this huge amount of new variation and its consequences for evolution. The 'gradualists', to which Darwin and the biometricians adhered, considered the changes across generations as gradual and incremental, whereas the 'saltationists' (e.g., T.H. Huxley and Galton) held that evolutionary changes occurred in 'jumps' of considerable magnitude. Much of the scientific dispute about Darwin's theory of evolution originated from the ignorance of the true hereditary mechanisms.

Despite the early work of Yule (1902), Hardy (1908), and Weinberg (1908), who showed that under the particulate mode of inheritance proposed by Mendel (1866), genetic variability is preserved under random mating, it was not before 1918 that the synthesis between genetics and the theory of evolution through natural selection began to take shape through Fisher's (1918) work (see Provine (1971) for a detailed account of the history of population genetics).

Today, the hereditary mechanisms have been firmly established and our knowledge about the molecular biology of the genes is rapidly increasing. Mutations are known to be the ultimate source of genetic variability, and many different processes at the chromosomal and molecular level have been identified that generate mutations. On the phenotypic level, the role of selection in shaping evolutionary change has been amply documented, whereas

on the molecular level, a significant amount of neutral evolution appears to take place, its extent still being disputed. Nevertheless, there remain many open problems, some of which are qualitative in nature and some quantitative. Questions concerning the processes involved in speciation events or in the evolution of sex belong to the first class, whereas questions concerning the prediction of the expected evolutionary change of a population subject to selection belong to the second class. Such predictions are highly nontrivial, unless confined to one or a few generations, because there exist many different forms of selection and the response to selection depends on the pattern and amount of genetic variability in the population. This variation, however, is a function of many genetic details (such as number of genes determining a trait, mutational properties, degree of linkage), of the demography (population size, mating structure), and of the selective forces acting. Therefore, the genetic variability may change from one generation to the next.

Mendel's (1866) prime achievement was the recognition of the particulate nature of the hereditary determinants, now called genes. A gene may have different forms, called *alleles*. From his experiments with peas he concluded that genes are present in pairs, one member of each pair having been inherited from the maternal parent, the other from the paternal. The allelic composition is called the *genotype*, and the set of observable properties derived from the genotype is called the *phenotype*. Thus, supposing that there are two alleles $\mathcal{A}_1$ and $\mathcal{A}_2$, there are three possible genotypes, $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$. In the first and third case, the organism's genotype is *homozygous* (for $\mathcal{A}_1$ or $\mathcal{A}_2$, respectively), in the second case it is *heterozygous*. In general, the genotypes $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_1$ cannot be distinguished. When the phenotype of the heterozygote $\mathcal{A}_1\mathcal{A}_2$ is the same as one of the homozygotes, say $\mathcal{A}_1\mathcal{A}_1$, allele $\mathcal{A}_1$ is called *dominant* and $\mathcal{A}_2$ is called *recessive*.

Mendel's first law states that when pure-bred (homozygous) strains are crossed, the hybrid progeny constituting the F_1 generation (the letter F stands for filial) are uniform (their genotype being $\mathcal{A}_1\mathcal{A}_2$ if the parents were $\mathcal{A}_1\mathcal{A}_1$ and $\mathcal{A}_2\mathcal{A}_2$) and usually express one of the two phenotypes (the one controlled by the dominant allele). According to Mendel's second law, recessive characters, which are masked in the heterozygous F_1, reappear in the F_2 in the proportion $1:3$ of the dominant character. This leads to the *Principle of Segregation*, stating that each reproductive cell (*gamete*) contains only one of the two alleles and that each gamete is equally likely to contain either one. The separation of the paired alleles from one another and their distribution to different cells, the gametes, is called *segregation* and occurs during *meiosis*. Meiosis is the process of formation of gametes from somatic cells. At mating, two reproductive cells fuse and form a *zygote*

(fertilized egg), which contains the full (diploid) genetic information.

Mendel also performed experiments with pure-bred lines that differed in two characters, round versus wrinkled seed shape and yellow versus green color. From previous experiments he knew that 'round' was dominant over 'wrinkled' and 'yellow' dominant over 'green' because their F_2 ratios were $3:1$ each. The F_1 seeds, from crosses of lines having round and yellow seeds with lines having wrinkled and green seeds, were all round and yellow. In F_2 progeny from the dihybrid cross, all four phenotypes reappeared, approximately in the proportions 9/16 'round yellow', 3/16 'wrinkled yellow', 3/16 'round green', and 1/16 'wrinkled green'. Thus, the proportion of the four phenotypes is $9:3:3:1$, as expected when two pairs of alleles segregate independently, so that the $3:1$ ratios are combined at random. This is called Mendel's third law or the *Principle of Independent Assortment*.

Since the 1940s it has been known that the genetic material is *deoxyribonucleic acid* (*DNA*). It consists of four bases: adenine (A), guanine (G), thymine (T), and cytosine (C). Each base is linked to a sugar and a phosphate group, yielding a *nucleotide*. The nucleotides are arranged along two chains to form a double-stranded helix in which the pairings A–T and G–C between the strands are formed. Therefore, all the genetic information is contained in each of the two strands. Three bases code for one amino acid, which are the building blocks of polypeptide chains and proteins. A gene typically represents a contiguous region of DNA coding for one polypeptide chain. Its position along the DNA is called the *locus*, and a particular sequence there is called an allele. Thus, two genes at the same locus, sampled from a population, may or may not be of the same allelic type. A double-stranded helix of DNA forms the backbones of the *chromosomes*, which are contained in the nucleus of each cell. In *diploid* organisms (higher plants and animals) chromosomes form homologous pairs, each one inherited from one parent. The exceptions are the *sex chromosomes*, which are involved in the genetic determination of sex. Usually, this is one pair of chromosomes which differ from each other, one called the X-chromosome, the other the Y-chromosome.

Any heritable change in the genetic material is called a *mutation*. Mutations are the ultimate source of genetic variability, and form the raw material upon which selection can act. Although the term mutation includes changes in chromosome structure and number, the vast majority of genetic variation is caused by gene mutations. Modern genetics has revealed that at the molecular level (gene) mutations occur in many different ways, for instance as base substitutions, in which one pair of nucleotides is replaced by

another, as insertions or deletions of DNA, as inversions of sequences of nucleotides, or as transpositions. The latter are mainly caused by transposable elements changing their position from one site to another. For many population-genetic models, however, the molecular origin of a mutant is not necessarily of relevance. What often counts is only the rate at which mutations occur and a mutant's effect on fitness or, more generally, on the character under consideration. Typically, spontaneous mutation rates per locus per generation are of the order of 10^{-4} to 10^{-6}, and genomic mutation rates summed over all loci may be on the order of one per generation, but can vary substantially between species.

During meiosis, different chromosomes assort independently and *crossing over* between two homologous chromosomes may occur. Consequently, the newly formed gamete contains maternal alleles at one set of loci and paternal alleles at the complementary set. This process is called *recombination*.

The mating pattern may have a substantial influence on the evolution of gene frequencies. The simplest and most important mode is *random mating*. This means that matings take place without regard to ancestry or the genotype under consideration. It seems to occur frequently in nature. For example, among humans, matings within a population appear to be random with respect to blood groups and allozyme phenotypes, but are non-random with respect to height. Random mating conserves allele frequencies and, after one generation, genotypic frequencies.

Selection occurs when individuals of different genotype leave different numbers of progeny because they differ in their probability to survive to reproductive age (*viability*), in their mating success, or in their average number of produced offspring (*fertility*). Darwin (1859) recognized and documented the central importance of selection as the driving force for adaptation and evolution. Since selection affects the entire genome, its consequences for the genetic composition of a population may be complex. Selection is measured in terms of *fitness* of individuals, i.e., by the number of progeny contributed to the next generation. There are different measures of fitness, and it consists of several components because selection may act on each stage of the life cycle.

4.1 The Hardy–Weinberg law

With the blending theory of inheritance variation in a population declines rapidly, and this was one of the arguments against Darwin's theory of evolution. With Mendelian inheritance there is no such dilution of variation, as was shown independently by the

famous British mathematician Hardy (1908) and the German physician Weinberg (1908, 1909).

Throughout this chapter, we consider a randomly mating population with discrete, nonoverlapping generations such that the genotype frequencies are the same in both sexes. We assume that the population is so large that gene and genotype frequencies may be treated as deterministic, and relative frequency can be identified with probability.

We consider a single locus with I possible alleles $\mathcal{A}_i$ and write $\mathsf{I} = \{1, \ldots, I\}$ for the set of all alleles. We denote the frequency of the ordered genotype $\mathcal{A}_i\mathcal{A}_j$ by P_{ij}, so that the frequency of the unordered genotype $\mathcal{A}_i\mathcal{A}_j$ is $P_{ij} + P_{ji} = 2P_{ij}$. Subscripts i and j always refer to alleles. Then the frequency of allele $\mathcal{A}_i$ in the population is

$$p_i = \sum_{j=1}^{I} P_{ij} \,.^{1} \tag{4.1}$$

After one generation of random mating the zygotic proportions satisfy

$$P'_{ij} = p_i p_j \qquad \text{for every } i \text{ and } j \,. \tag{4.2}$$

If there is random union of gametes, as in some marine organisms, (4.2) simply reflects its definition. If diploid individuals mate, (4.2) needs to be proved (see below).

A mathematically trivial, but biologically important, consequence is that (in the absence of other forces) gene frequencies remain constant across generations, i.e.,

$$p'_i = p_i \quad \text{for every } i \,. \tag{4.3}$$

In other words, in a (sufficiently large) randomly mating population reproduction does not change allele frequencies. A population is said to be in *Hardy–Weinberg equilibrium* if

$$P_{ij} = p_i p_j \,. \tag{4.4}$$

In a (sufficiently large) randomly mating population, this relation is always satisfied among zygotes. Thus, the Hardy–Weinberg Law states that after one generation of random mating, the genotype frequencies remain constant and can be expressed in terms of the allele frequencies according to (4.4). In particular, the allele (gene) frequencies remain constant and no genetic variability is lost by random mating.

Evolutionary mechanisms such as selection, migration, mutation, or random genetic drift distort Hardy-Weinberg proportions, but reproduction restores them among zygotes if there is random mating.

[1] If no summation range is indicated, it is assumed to be over all admissible values; e.g., $\sum_i = \sum_{i \in \mathsf{I}}$

Table 4.1: Mating table

Mating	Mating prob.	Cond. prob. of progeny		
		$\mathcal{A}_1\mathcal{A}_1$	$\mathcal{A}_1\mathcal{A}_2$	$\mathcal{A}_2\mathcal{A}_2$
$\mathcal{A}_1\mathcal{A}_1 \times \mathcal{A}_1\mathcal{A}_1$	P^2	1	0	0
$\mathcal{A}_1\mathcal{A}_1 \times \mathcal{A}_1\mathcal{A}_2$	$4PQ$	$\frac{1}{2}$	$\frac{1}{2}$	0
$\mathcal{A}_1\mathcal{A}_1 \times \mathcal{A}_2\mathcal{A}_2$	$2PR$	0	1	0
$\mathcal{A}_1\mathcal{A}_2 \times \mathcal{A}_1\mathcal{A}_2$	$4Q^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$\mathcal{A}_1\mathcal{A}_2 \times \mathcal{A}_2\mathcal{A}_2$	$4QR$	0	$\frac{1}{2}$	$\frac{1}{2}$
$\mathcal{A}_2\mathcal{A}_2 \times \mathcal{A}_2\mathcal{A}_2$	R^2	0	0	1

Proof of (4.2) **for a diallelic locus**

In most higher organisms adult individuals mate. Then we need a more elaborate approach. We label the relative frequencies of the genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$ in the population can by P, $2Q$, and R, respectively, and $P + 2Q + R = 1$. The heterozygous genotype $\mathcal{A}_1\mathcal{A}_2$ has been assumed to be unordered, so that $2Q$ is the combined frequency of the ordered genotypes $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_1$. We assume, furthermore, that the population is so large that gene and genotype frequencies may be treated as deterministic, and relative frequency can be identified with probability.

We want to derive the frequencies of the three genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$ in the next generation. This can be achieved by calculating the frequencies of all possible matings and their offspring produced. For example, with random mating (with respect to the locus under consideration), the probability of the mating $\mathcal{A}_1\mathcal{A}_1 \times \mathcal{A}_1\mathcal{A}_2$ is $4PQ$, because $\mathcal{A}_1\mathcal{A}_1$ can be male or female (and $\mathcal{A}_1\mathcal{A}_2$, thus, female or male), and the probabilities of the genotypes $\mathcal{A}_1\mathcal{A}_1$ and $\mathcal{A}_1\mathcal{A}_2$ are P and $2Q$, respectively. According to Mendel's laws, half of the progeny of such a mating are $\mathcal{A}_1\mathcal{A}_1$ and half are $\mathcal{A}_1\mathcal{A}_2$. Table 4.1 summarizes all possibilities.

Therefore, the frequency of $\mathcal{A}_1\mathcal{A}_1$ homozygotes in the next generation is[2]

$$\begin{aligned} P' &= P^2 \cdot 1 + 4PQ \cdot \tfrac{1}{2} + 2PR \cdot 0 + 4Q^2 \cdot \tfrac{1}{4} + 4QR \cdot 0 + R^2 \cdot 0 \\ &= P^2 + 2PQ + Q^2 = (P + Q)^2 \end{aligned} \tag{4.5a}$$

[2]Unless stated otherwise, a prime will always signify the next generation.

and, similarly,

$$2Q' = 2PQ + 2PR + 2Q^2 + 2QR = 2(P+Q)(Q+R) \tag{4.5b}$$

and

$$R' = Q^2 + 2QR + R^2 = (Q+R)^2 \,. \tag{4.5c}$$

Here we have assumed that no mutation occurs and that no evolutionary forces, such as viability selection, differential fertility, geographical dispersal, or separate sexes, change the genotype frequencies. By substituting P', Q', and R' into the right-hand sides of (4.5a) - (4.5c), and observing the fact that $P + 2Q + R = 1$, we obtain after another generation of random mating

$$P'' = (P' + Q')^2 = (P+Q)^2 = P' \tag{4.6a}$$

and, similarly,

$$Q'' = Q' \qquad \text{and} \qquad R'' = R' \,. \tag{4.6b}$$

Thus, the genotype frequencies established after one generation of random mating are maintained under random mating in all subsequent generations.

Now let us consider the gene frequencies p and $q = 1 - p$ of the alleles $\mathcal{A}_1$ and $\mathcal{A}_2$. Since all the genes in $\mathcal{A}_1\mathcal{A}_1$ individuals, and half the genes in $\mathcal{A}_1\mathcal{A}_2$ individuals, are $\mathcal{A}_1$ genes, therefore $p = \frac{1}{2}(2P + 2Q)$ and similarly $q = Q + R$. Hence, we can rewrite the equations (4.5) as

$$P' = p^2, \quad 2Q' = 2pq, \quad R' = q^2 \,, \tag{4.7}$$

which is the desired special case of (4.2).

With a bit more algebra, this proof can be generalized to multiple alleles. The Hardy-Weinberg law can be extended to more general situations, such as separate sexes (then it takes two generations to reach Hardy-Weinberg proportions) or X-linked loci (then geometric convergence to HW occurs).

4.2 Selection

Selection occurs when genotypes in a population differ in their fitnesses, i.e., in their viability, mating success, or fertility and, therefore, leave different numbers of progeny. As already mentioned, selection is one of the major driving forces of evolution. Therefore, it is essential to understand its consequences well. The basic models of selection were developed and investigated in the 1920s and early 1930s by Fisher, Wright, and Haldane.

4.2.1 The model

We shall be concerned with the evolutionary consequences of selection caused by differential viabilities (i.e., the probability that an offspring survives to reproductive age). We suppose that at an autosomal locus the alleles $\mathcal{A}_1, \ldots, \mathcal{A}_I$ occur. We count individuals at the zygote stage and denote the (relative) frequency of the ordered genotype $\mathcal{A}_i\mathcal{A}_j$ by $P_{ij} (= P_{ji})$.

Since mating is at random, the genotype frequencies P_{ij} are in Hardy-Weinberg proportions. Let us suppose that selection acts solely through differential viabilities, and denote the fitness (viability) of $\mathcal{A}_i\mathcal{A}_j$ individuals by $w_{ij} \geq 0$. We assume that the fitnesses satisfy $w_{ij} = w_{ji}$. Then the frequency of $\mathcal{A}_i\mathcal{A}_j$ genotypes among adults that have survived selection is

$$P_{ij}^* = \frac{w_{ij} P_{ij}}{\bar{w}} = \frac{w_{ij} p_i p_j}{\bar{w}} \,,$$

where we have used (4.4). Here,

$$\bar{w} = \sum_{ij} w_{ij} P_{ij} = \sum_{ij} w_{ij} p_i p_j = \sum_i w_i p_i \tag{4.8}$$

is the *mean fitness* of the population and

$$w_i = \sum_j w_{ij} p_j \tag{4.9}$$

is the *marginal fitness* of allele $\mathcal{A}_i$.

Therefore, the frequency of $\mathcal{A}_i$ after selection is

$$p_i^* = \sum_j P_{ij}^* = p_i \frac{w_i}{\bar{w}} \,. \tag{4.10}$$

Because of random mating, the allele frequency p_i' among zygotes of the next generation is also p_i^* (4.3), so that allele frequencies evolve according to the *selection equation*

$$p_i' = p_i \, \frac{w_i}{\bar{w}} \,, \quad i \in \mathsf{I} \,. \tag{4.11}$$

This recursion equation preserves the relation

$$\sum_i p_i = 1$$

and describes the evolution of allele frequencies at a single autosomal locus in a diploid population. We view the selection dynamics (4.11) as a (discrete) dynamical system on the simplex

$$\Delta_I = \left\{ p = (p_1, \ldots, p_I)^T \in \mathbb{R}^I : p_i \geq 0 \text{ for every } i \in \mathsf{I}\,,\ \sum_i p_i = 1 \right\},^3 \tag{4.12}$$

The selection dynamics is a *replicator equation* (see Hofbauer and Sigmund 1998).

Fitnesses are said to be *multiplicative* if constants v_i exist such that

$$w_{ij} = v_i v_j \tag{4.13}$$

for every i, j. Then $w_i = v_i \bar{v}$, where $\bar{v} = \sum_i v_i p_i$, and $\bar{w} = \bar{v}^2$. Therefore, (4.11) simplifies to

$$p_i' = p_i \frac{v_i}{\bar{v}}\,, \quad i \in \mathsf{I}\,. \tag{4.14}$$

This can be solved explicitly because it is equivalent to the linear system $x_i' = v_i x_i$. The solution is

$$p_i(t) = \frac{p_i(0) v_i^t}{\sum_j p_j(0) v_j^t}\,. \tag{4.15}$$

(4.14) also describes the dynamics of an asexual haploid population under selection.

Example 4.1 (Selection is very efficient)**.** Suppose there are only two alleles, $\mathcal{A}_1$ and $\mathcal{A}_2$. If $\mathcal{A}_1$ is the wild type and $\mathcal{A}_2$ is a new beneficial mutant, we may set (without loss of generality!) $v_1 = 1$ and $v_2 = 1 + s$. Then we obtain from (4.15):

$$\frac{p_2(t)}{p_1(t)} = \frac{p_2(0)}{p_1(0)} \left(\frac{v_2}{v_1} \right)^t = \frac{p_2(0)}{p_1(0)} (1+s)^t\,. \tag{4.16}$$

Thus, $\mathcal{A}_2$ increases exponentially relative to $\mathcal{A}_1$.

For instance, if $s = 0.5$, then after 10 generations the frequency of $\mathcal{A}_2$ has increased by a factor of $(1+s)^t = 1.5^{10} \approx 57.7$ relative to $\mathcal{A}_1$. If $s = 0.05$ and $t = 100$, this factor is $(1+s)^t = 1.05^{100} \approx 131.5$.

Thus, *slight fitness differences may have a big long-term effect.* Also note that 100 generations are short on an evolutionary time scale.

An important property of (4.11) is that mean fitness is nondecreasing along trajectories (solutions), i.e.,

$$\bar{w}' = \bar{w}(p') \geq \bar{w}(p) = \bar{w}\,, \tag{4.17}$$

[3]Throughout, the superscript ^{T} denotes vector or matrix transposition.

and equality holds if and only if p is an equilibrium.[4] This statement is a special case of *Fisher's Fundamental Theorem of Natural Selection* (see below). The proof of (4.17) is not trivial and uses inequalities related to convexity (see, e.g., Chap. I.9 of Bürger 2000, or Hofbauer and Sigmund 1988).

In mathematical terms, (4.17) shows that $\bar{w}$ is a Lyapunov function. This has a number of important consequences. For instance, complex dynamical behavior such as limit cycles or chaos can be excluded. All trajectories approach the set of points $p \in \Delta_I$ that are maxima of $\bar{w}$. This is a subset of the set of equilibria. From (4.11) it is obvious that the equilibria are precisely the solutions of

$$p_i(w_i - \bar{w}) = 0 \quad \text{for every } i\,. \tag{4.18}$$

We call an equilibrium internal, or fully polymorphic, if $p_i > 0$ for every i (all alleles are present). The I equilibria defined by $p_i = 1$ for some i are called monomorphic (only allele $\mathcal{A}_i$ is present). In general, it is difficult or even impossible to determine all equilibria.

4.2.2 Two alleles

Here, we specialize to two alleles. We write p and $1-p$ instead of p_1 and p_2. Further, we use relative fitnesses and assume

$$w_{11} = 1\,,\ w_{12} = 1 - hs\,,\ w_{22} = 1 - s\,, \tag{4.19}$$

where s is called the *selection coefficient* and h describes the degree of dominance. We assume $s > 0$.

The allele $\mathcal{A}_1$ is called *dominant* if $h = 0$, partially dominant if $0 < h < \frac{1}{2}$, *recessive* if $h = 1$, and partially recessive if $\frac{1}{2} < h < 1$. The terms *additivity*, or *no dominance*, refer to $h = \frac{1}{2}$. If $h < 0$, there is *overdominance* or *heterozygote advantage*. If $h > 1$, there is *underdominance* or *heterozygote inferiority*.

From (4.9), the marginal fitnesses of the two alleles are

$$w_1 = 1 - hs + hsp \quad \text{and} \quad w_2 = 1 - s + s(1-h)p \tag{4.20}$$

and, from (4.8), the mean fitness is

$$\bar{w} = 1 - s + 2s(1-h)p - s(1-2h)p^2\,. \tag{4.21}$$

[4] p is called an equilibrium, or fixed point, of the recursion relation $p' = f(p)$ if $f(p) = p$. We use the term equilibrium point to emphasize that we consider an equilibrium that is a single point.

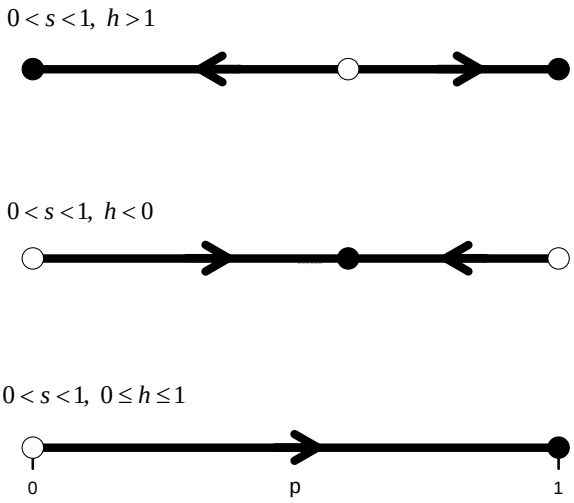

Figure 4.1: Convergence patterns for selection with two alleles.

It is easily verified that the allele-frequency change from one generation to the next can be written as

$$\Delta p = p' - p = \frac{p(1-p)}{2\bar{w}} \frac{\mathrm{d}\bar{w}}{\mathrm{d}p} \tag{4.22a}$$

$$= \frac{p(1-p)s}{\bar{w}} \left[1 - h - (1-2h)p\right]. \tag{4.22b}$$

There exists an internal equilibrium if and only if $h < 0$ (overdominance) or if $h > 1$ (underdominance). It is given by

$$\hat{p} = \frac{1-h}{1-2h}. \tag{4.23}$$

Because we can write (4.22) in the form

$$\Delta p = \frac{sp(1-p)}{\bar{w}}(1-2h)(\hat{p}-p), \tag{4.24}$$

and because $0 < sp(1-p)(1-2h)/\bar{w} < 1$ if $0 < p < 1$, $\hat{p}$ is globally asymptotically stable if and only if $h < 0$, and convergence is monotonic. If $h > 1$, then the monomorphic equilibria $p = 0$ and $p = 1$ each are asymptotically stable and $\hat{p}$ is unstable. For intermediate dominance, $0 \leq h \leq 1$, $p = 1$ is globally asymptotically stable.

The three possible convergence patterns are shown in Figure 4.1. Figure 4.2 displays the influence of the degree of (intermediate) dominance on the rate of adaptation of an advantageous allele.

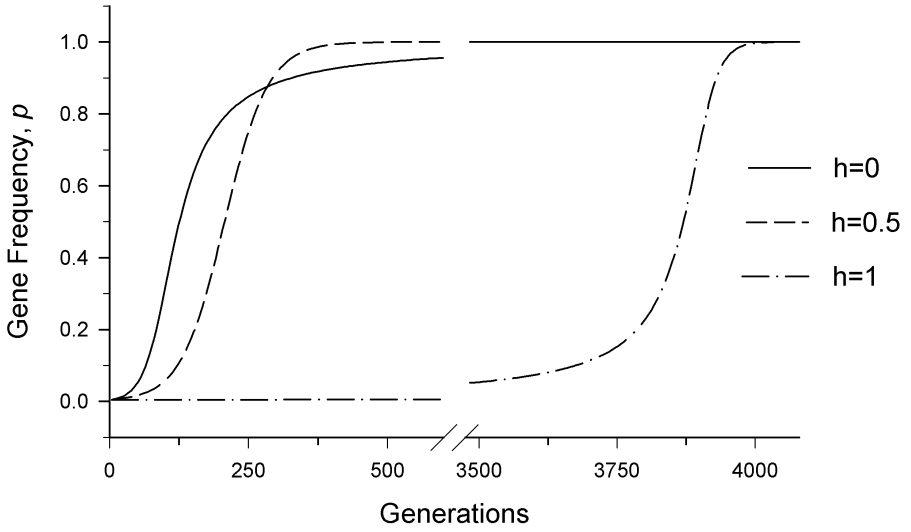

Figure 4.2: Selection of a dominant ($h = 0$, solid line), intermediate ($h = 1/2$, dashed), and recessive ($h = 1$, dash-dotted) allele. The initial frequency is $p_0 = 0.005$ and the selective advantage is $s = 0.05$.

4.2.3 The continuous-time selection model

Most higher animal species have overlapping generations because birth and death occurs continuously in time. This, however, may lead to substantial complications if one wishes to derive a continuous-time model from biological principles. By contrast, discrete-time models can frequently be derived straightforwardly from simple biological assumptions. If evolutionary forces are weak, a continuous-time version can often be derived as an approximation to the discrete-time model.

A rigorous derivation of the differential equations describing gene-frequency change under selection in a diploid population with overlapping generations is a formidable task and requires a complex model involving age structure. Here, we just state the system of differential equations and justify it in an alternative way.

In a continuous-time model, the fitness of a genotype, $\mathcal{A}_i\mathcal{A}_j$, is defined as its birth rate minus death rate. We denote it by m_{ij}. Then, the marginal fitness of allele $\mathcal{A}_i$ is

$$m_i = \sum_j m_{ij} p_j \,,$$

the mean fitness of the population is

$$\bar{m} = \sum_i m_i p_i = \sum_{i,j} m_{ij} p_i p_j \,,$$

and the dynamics of allele frequencies becomes

$$\dot{p}_i = p_i(m_i - \bar{m}) \,, \quad i \in \mathsf{I} \,. \tag{4.25}$$

This is the analogue of the discrete-time selection dynamics (4.11). Its state space is again the simplex S_I. The equilibria are obtained from the condition $\dot{p}_i = 0$ for every i.

Example 4.2 (Two alleles)**.** For two alleles, (4.25) simplifies considerably because it is sufficient to track the allele frequency $p = p_1$. In addition, we write $q = 1 - p$. Scaling the Malthusian parameters in the following way

$$\begin{array}{ccc} \mathcal{A}_1\mathcal{A}_1 & \mathcal{A}_1\mathcal{A}_2 & \mathcal{A}_2\mathcal{A}_2 \\ 0 & -hs & -s \end{array} \,,$$

we obtain the simple representations

$$\dot{p} = \tfrac{1}{2} spq \quad \text{if } h = \tfrac{1}{2} \text{ (no dominance)} \tag{4.26}$$

and

$$\dot{p} = spq^2 \quad \text{if } h = 0 \text{ ($\mathcal{A}_1$ is dominant)} \,. \tag{4.27}$$

Equation (4.26) is also obtained for a haploid population in which $\mathcal{A}_2$ has a selective disadvantage of $\frac{1}{2}s$ relative to $\mathcal{A}_1$.

Derivation of the continuous-time model from the discrete-time model. First, observe that the difference equation (4.11) and the differential equation (4.25) have the same equilibria if

$$w_{ij} = 1 + s m_{ij} \quad \text{for every } i, j \in \mathsf{I} \,. \tag{4.28}$$

This is obvious upon noting that (4.28) implies $w_i = a + s m_i$ and $\bar{w} = a + s\bar{m}$.

For weak selection the discrete model (4.11) can be approximated by the continuous model (4.25) as follows. Assume that w_{ij} is given by (4.28), rescale time according to $t = \lfloor \tau/s \rfloor$, where $\lfloor\ \rfloor$ denotes the closest smaller integer. Then s may be interpreted as generation length and, for $p_i(t)$ satisfying the difference equation (4.11), we define $q_i = q_i(\tau) = p_i(t)$. Then we have

$$\frac{d}{d\tau} q_i = \lim_{s \downarrow 0} \frac{1}{s} \left[q_i(\tau + s) - q_i(\tau) \right] = \lim_{s \downarrow 0} \frac{1}{s} \left[p_i(t+1) - p_i(t) \right] \,.$$

From (4.11) and (4.28), we obtain $p_i(t+1)-p_i(t) = sp_i(t)(m_i-\bar{m})/(1+s\bar{m})$ and, therefore, $\dot{q}_i = q_i(m_i - \bar{m})$. This proves the assertion because $\Delta p_i \approx s\dot{q}_i = sp_i(m_i - \bar{m})$.

One of the advantages of models in continuous time is that they lead to differential equations, and usually these are easier to analyze because the formalism of calculus is available. An example for this is that, in continuous time, (4.17) simplifies to

$$\dot{\bar{m}} = \frac{\mathrm{d}\bar{m}}{\mathrm{d}t} \geq 0\,. \tag{4.29}$$

This is *much* easier to prove than (4.17).

The exact continuous-time model reduces to (4.11) only if the mathematically inconsistent assumption is imposed that Hardy-Weinberg proportions apply at every time which is generally not true. Under weak selection, however, deviations from Hardy-Weinberg decay to order $O(s)$ after a short period of time.

4.2.4 Important general results

As already stated in (4.17), the selection dynamics (4.11) has the important property that mean fitness is nondecreasing along trajectories. More precisely, it has been shown that

$$\Delta\bar{w} \geq 0 \quad \text{and} \quad \Delta\bar{w} = 0 \quad \text{only at equilibria}\,, \tag{4.30}$$

where $\Delta\bar{w} = \bar{w}' - \bar{w}$.

In his famous *Fundamental Theorem of Natural Selection*, Fisher (1930) not only stated that mean fitness is nondecreasing but that its rate of change is equal to the additive genetic variance in fitness,

$$\sigma_A^2 = 2\sum_i p_i(w_i - \bar{w})^2\,. \tag{4.31}$$

In general, σ_A^2 is strictly smaller than the total genetic variance

$$\sigma_G^2 = \sum_{i,j} p_i p_j (w_{ij} - \bar{w})^2, \tag{4.32}$$

and $\sigma_A^2 = \sigma_G^2$ if there is no dominance.

The classical interpretation of the Fisher's Fundamental Theorem has been that

$$\Delta\bar{w} = \sigma_A^2/\bar{w}\,, \tag{4.33}$$

at least approximately. Unless there is no dominance, (4.33) does generally not hold exactly. However, it can be shown to hold to a very close approximation if selection is weak (s small); e.g. Nagylaki (1991).

The following result summarizes a number of further important properties of the selection dynamics. Proofs and references to the original literature may be found in Bürger (2000).

Theorem 4.3. *1. If an isolated internal equilibrium exists, then it is uniquely determined.*

2. $\hat{p}$ is an equilibrium if and only if $\hat{p}$ is a critical point of the restriction of mean fitness $\bar{w}(p)$ to the minimal subsimplex of S_I that contains the positive components of $\hat{p}$.

3. If the number of equilibria is finite, then it is bounded above by $2^I - 1$.

4. An internal equilibrium is asymptotically stable if and only if it is an isolated local maximum of $\bar{w}$. Moreover, it is isolated if and only if it is hyperbolic (i.e., the Jacobian has no eigenvalues of modulus 1).

5. An equilibrium point is stable if and only if it is a local, not necessarily isolated, maximum of $\bar{w}$.

6. If an asymptotically stable internal equilibrium exists, then every orbit starting in the interior of S_I converges to that equilibrium.

7. If an internal equilibrium exists, it is stable if and only if, counting multiplicities, the fitness matrix $\mathsf{W} = (w_{ij})$ has exactly one positive eigenvalue.

8. If the matrix W has i positive eigenvalues, at least $(i-1)$ alleles will be absent at a stable equilibrium.

9. Every orbit converges to one of the equilibrium points (even if they are not isolated).

4.3 Mutation and selection

Natural selection and mutation are two central factors guiding biological evolution: mutation generates the genetic variability upon which selection can act. This was clearly recognized by the pioneers of population genetics, Fisher, Haldane, and Wright, who developed mathematical models quantifying the relative importance of selection and mutation in maintaining genetic variation. To understand the patterns and amount of genetic variation within a population is essential because, as we have seen, they determine the response to selection. In particular, in the absence of genetic variation, evolution is impossible.

In traditional models, two alleles per locus are considered, the wild type and a mutant, and the equilibrium frequencies of the alleles can be calculated under recurrent mutation and various assumptions on the selective values of the genotypes. Usually, however, more than two alleles per locus may occur. For the type of model treated below, the molecular origin of the mutants is irrelevant; only their selective properties are used.

As above, we assume that populations are sufficiently large that random genetic drift can be ignored, that mating is at random if they are sexual, and genotypic fitnesses are constant.

4.3.1 Mutation only

We shall employ a simple concept of mutation, sufficient for most purposes in population genetics theory, by designating any change from one allelic type to another a mutation. Here, we assume that all mutations are neutral, i.e., all have the same fitness. Let us consider I alleles, $\mathcal{A}_1, \ldots, \mathcal{A}_I$, at a gene locus and label their frequencies by $p_1, \ldots, p_I$. For $i \neq j$ we denote the probability that an $\mathcal{A}_i$ gene has an $\mathcal{A}_j$ offspring by the mutation rate μ_{ij}. We shall use the convention $\mu_{ii} = 0$ for every i. Then the fraction of $\mathcal{A}_i$ genes that do not mutate is $1 - \sum_j \mu_{ij}$, and $\mathcal{A}_j$ genes give rise to a mutant $\mathcal{A}_i$ with probability μ_{ji}. Therefore, the frequency p_i' of $\mathcal{A}_i$ in the next generation is

$$p_i' = p_i \left(1 - \sum_j \mu_{ij} \right) + \sum_j p_j \mu_{ji} \,. \tag{4.34}$$

We call (4.34) the (*pure*) *mutation equation*. Due to the convention $\mu_{ii} = 0$, the index i may or may not be excluded in the above summations.

Linear algebra shows that there exists a unique equilibrium if all mutation rates are positive, and that convergence to this equilibrium occurs at a geometric rate (see below).

Example 4.4. Let us illustrate this for the simple case of two alleles. Denoting the mutation rate from $\mathcal{A}_1$ to $\mathcal{A}_2$ by μ, the reverse mutation rate by ν, and the frequency of $\mathcal{A}_1$ by p, the recursion (4.34) reduces to

$$p' = p(1 - \mu - \nu) + \nu \,.$$

If μ or ν is positive, there exists a unique equilibrium frequency (obtained from the condition $p' = p$). It is given by

$$\hat{p} = \frac{\nu}{\mu + \nu} \,.$$

The above recursion equation can be solved explicitly and, using $\hat{p}$, its solution can be expressed as

$$p(t) - \hat{p} = (p_0 - \hat{p})(1 - \mu - \nu)^t \,,$$

where $p_0 = p(0)$ is the initial frequency of $\mathcal{A}_1$. This shows that convergence to equilibrium occurs at a geometric rate, but is very slow because $\mu + \nu$ is typically very small.

4.3.2 The mutation-selection model

First, we set up the mutation-selection model for a diploid population and one locus with an arbitrary number I of alleles. Then we confine attention to the diallelic case.

Let μ_{ij} denote the mutation rate from $\mathcal{A}_i$ to $\mathcal{A}_j$ and $\mu_{ii} = 0$. Allele frequencies are measured among zygotes before selection, and the life cycle begins with selection, which is followed by the production of germ cells, during which mutation occurs, and the formation of zygotes. We assume a population with discrete generations. Therefore, applying (4.34) to the allele frequencies p_i^* after selection, as given by (4.11), we obtain

$$p_i' = p_i^* \left(1 - \sum_j \mu_{ij} \right) + \sum_j p_j^* \mu_{ji} \,. \tag{4.35}$$

This can be rewritten in the form

$$p_i' = p_i \frac{w_i}{\bar{w}} + \frac{1}{\bar{w}} \sum_j \left(p_j w_j \mu_{ji} - p_i w_i \mu_{ij} \right) \,, \tag{4.36}$$

where w_i is the marginal fitness of the allele $\mathcal{A}_i$. (4.36) provides the diploid mutation-selection dynamics.

For a population with overlapping generations, the differential equation

$$\dot{p}_i = p_i (m_i - \bar{m}) + \sum_j (p_j \mu_{ji} - p_i \mu_{ij}) \tag{4.37}$$

is used, where m_i is the marginal Malthusian fitness of $\mathcal{A}_i$.

Obviously, the dynamics of (4.36) remains unchanged if all fitnesses w_{ij} are multiplied by the same positive constant, and (4.37) remains unchanged if the same constant is added to every m_{ij}. For multiplicative fitnesses, $w_{ij} = w_i w_j$, the discrete-time recursion (4.36) reduces to the haploid recursion (in which the w_i are constants and $\bar{w}$ is linear), and for additive fitnesses, $m_{ij} = m_i + m_j$, the continuous-time equation (4.37) reduces to the corresponding haploid equation.

4.3.3 Two alleles

For the present purpose it is convenient to parameterize the fitness values of the genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, $\mathcal{A}_2\mathcal{A}_2$ as $w_{11} = 1$, $w_{12} = 1 - hs$, $w_{22} = 1 - s$. Instead of p_1 and p_2, we write p and $q = 1 - p$. Then the marginal fitnesses and the mean fitness are given by (4.20) and (4.21), respectively. For the mutation rates we write $\mu = \mu_{12}$ and $\nu = \mu_{21}$

and require $\mu + \nu < 1$. A straightforward calculation shows that the equilibria of the mutation-selection equation (4.36) are the solutions p of

$$p^3 s(2h-1) + p^2 s[2 - 3h + \mu h + \nu(1-h)] \\ + p[-s(1-h) + \mu(1-hs) + \nu(1-2s+hs)] - \nu(1-s) = 0 \qquad (4.38)$$

in the interval $[0,1]$. As we shall see below, there may be one, two, or three such solutions, depending on the parameters. Some elementary, but lengthy, algebra shows the following (Norman 1974, Nagylaki 1992):

Theorem 4.5. *If $0 < s < 1$ and $h \le \frac{1}{2}$, or $s < 0$ and $h \ge \frac{1}{2}$, then (4.37) has a unique solution in $[0,1]$. Because $\mu + \nu < 1$, this equilibrium is globally asymptotically stable. Convergence is monotonic.*

This result includes several important special cases, such as no dominance ($h = \frac{1}{2}$), complete dominance of $\mathcal{A}_1$ ($h = 0$), and overdominance ($h < 0$) (in all these cases $s > 0$ is assumed).

The equilibrium solutions are simple only in special cases. We restrict our attention to the case $\nu = 0$, in which back mutations from the deleterious (and thus rare) allele $\mathcal{A}_2$ to $\mathcal{A}_1$ are ignored. It will be convenient to give the precise formulas in terms of $q = 1 - p$. Obviously, $\hat{q}^{(0)} = 1$ is always an equilibrium, because if $\mathcal{A}_1$ is initially not present in the population, it will not arise by mutation. Since $\nu = 0$, (4.37) reduces to a quadratic equation which, if $4\mu/s \le 1$, has the following solutions in $[0,1]$:

$$\hat{q}^{(1)} = \frac{h(1+\mu)}{2(2h-1)}\left[1 - \sqrt{1 - \frac{4\mu(2h-1)}{(1+\mu)^2 h^2 s}}\right] \qquad \text{if } h \ne 0, \tfrac{1}{2}\,, \qquad (4.39a)$$

$$\hat{q}^{(1)} = \frac{2\mu}{s(1+\mu)} \qquad \text{if } h = \tfrac{1}{2}\,, \qquad (4.39b)$$

and

$$\hat{q}^{(2)} = \frac{h(1+\mu)}{2(2h-1)}\left[1 + \sqrt{1 - \frac{4\mu(2h-1)}{(1+\mu)^2 h^2 s}}\right] \qquad \text{if } h > h_c\,, \qquad (4.40)$$

where

$$h_c = \frac{1 - \mu/s}{1 - \mu}\,. \qquad (4.41)$$

Note that the case $h > h_c$ includes underdominance, i.e., $h > 1$. If $h < h_c$, then $\hat{q}^{(2)}$ is biologically not meaningful because $\hat{q}^{(2)} > 1$. If this holds, then the equilibrium $\hat{q}^{(1)}$

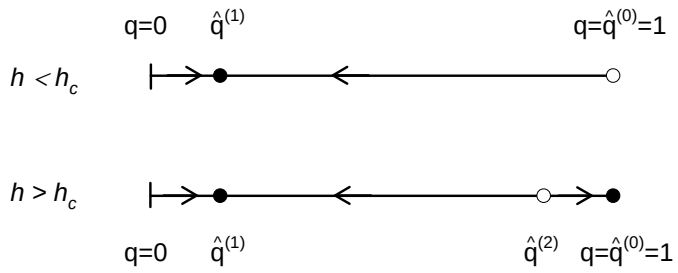

Figure 4.3: Equilibria and dynamics for the diallelic mutation-selection equation with one-way mutation. The drawing on top displays the case $h \leq h_c$, that on bottom is for $h > h_c$. Stable equilibria are indicated by $\bullet$, unstable ones by $\circ$.

is globally asymptotically stable. If $h > h_c$, then three equilibria coexist. They satisfy $0 < \hat{q}^{(1)} < \hat{q}^{(2)} < \hat{q}^{(0)} = 1$, and $\hat{q}^{(1)}$ and $\hat{q}^{(0)}$ are asymptotically stable, whereas $\hat{q}^{(2)}$ is unstable (see Figure 4.3). Thus, for one-way mutation, a simple and explicit classification of the stability of equilibria is available. It can be shown that this is also valid for the case $0 < \nu \ll \mu$. Then, of course, $\hat{q}^{(0)} < 1$ and $\hat{q}^{(0)} \approx 1$. Analogous results hold for the differential equation (4.36).

We point out that for $h_c \leq h \leq 1$ the pure selection model has one globally asymptotically stable boundary equilibrium ($\hat{q} = 0$), but the introduction of mutation, however weak, leads to two stable and one unstable equilibria. Thus, already with two alleles, the diploid mutation-selection dynamics may be qualitatively different from the haploid dynamics.

Assuming that μ is of smaller order than s, simple approximations for the equilibrium frequencies can be derived in the following cases:

If $h = 0$, then

$$\hat{q}^{(1)} = \sqrt{\frac{\mu}{s}}\,. \tag{4.42a}$$

If $h \gg \sqrt{\mu/s}$, then

$$\hat{q}^{(1)} \approx \frac{\mu}{hs}\,. \tag{4.42b}$$

If $h > h_c$, then the unstable interior equilibrium is admissible and satisfies

$$\hat{q}^{(2)} \approx \frac{h}{2h-1} - \frac{\mu}{hs}\,. \tag{4.42c}$$

For multiplicative selection coefficients, $w_{11} = 1$, $w_{12} = 1 - t$, $w_{22} = (1 - t)^2$, one obtains (exactly)

$$\hat{q}^{(1)} = \frac{\mu}{t}\,. \tag{4.42d}$$

The equilibrium frequency of a recessive deleterious mutant ($h = 0$) is much higher than that of an intermediate or dominant deleterious mutant ($h \gg \sqrt{\mu/s}$), because it occurs mostly in heterozygotes, against which selection is ineffective.

The case of weak mutation can also be treated by perturbation arguments. If $s > 0$ and $h < 1$, then introduction of sufficiently weak mutation [such that $h < (1 - \mu/s)/(1 - \mu)$; cf. (4.41)] leads only to very small disturbances of the equilibria which, in particular, maintain their stability properties. Then a stable equilibrium at the boundary will move inwards, and only the situation displayed in the upper drawing of Figure 4.3 can occur. The case $h = 1$ is not regular and this simple perturbation argument does not apply. It is generally held that the case of weak mutation is biologically the most important.

Remark 4.6. 1. It is known that the majority of mutations is (slightly) deleterious. Therefore, in general, mutations decrease the mean fitness of a population. This decrease is called the mutation load. Interestingly, the mutation load is, to leading order, independent of the selective coefficient of the mutations. It is twice the total mutation rate to deleterious alleles. This is called Haldane's principle.

2. Already the above results show that with diploidy the dynamics under mutation and selection can be more complex than with haploidy. If at least three alleles occur at a locus, it has been shown that stable limit cycles may occur. Essentially, this requires that the strength of mutation and selection are of comparable magnitude and interact in specific ways.

3. If fitnesses are multiplicative or the population is haploid, then global convergence to a unique fully polymorphic equilibrium can be proved, provided the mutation matrix is primitive, or ergodic. This result is a consequence of the Perron-Frobenius Theorem for positive matrices.

4.4 Recombination

The process of crossing between two homologous chromosomes during meiosis leads to recombination between genes on the same chromosome. Naturally, recombination also occurs between genes on different chromosomes because chromosomes are passed independently to daughter cells. Therefore recombination has the potential to combine favorable

alleles of different ancestry in one gamete and to break up combinations of deleterious alleles. These properties may confer a substantial evolutionary advantage to sexual species relative to asexuals. The interaction of recombination with selection and mutation also has other important evolutionary consequences. Here, we treat only the simplest model and study recombination in isolation.

According to the Hardy–Weinberg Law, the genotype frequencies attain an equilibrium value after one generation of random mating if gene loci are considered separately. This is no longer true for genotypes with respect to two or more loci considered jointly. Consider two loci, $\mathcal{A}$ and $\mathcal{B}$, each with two alleles, $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}_1$, $\mathcal{B}_2$. Then there are ten possible genotypes. If, for instance, in the initial generation only the genotypes $\mathcal{A}_1\mathcal{B}_1/\mathcal{A}_1\mathcal{B}_1$ and $\mathcal{A}_2\mathcal{B}_2/\mathcal{A}_2\mathcal{B}_2$ are present, then in the next generation only these double homozygotes, as well as the two double heterozygotes $\mathcal{A}_1\mathcal{B}_1/\mathcal{A}_2\mathcal{B}_2$ and $\mathcal{A}_1\mathcal{B}_2/\mathcal{A}_2\mathcal{B}_1$ will be present. After further generations of random mating, all other genotypes will occur, but not immediately at their equilibrium frequencies. Of course, the formation of gametic types other than $\mathcal{A}_1\mathcal{B}_1$ or $\mathcal{A}_2\mathcal{B}_2$ requires that recombination between the two loci occur. Disequilibrium with respect to two or more loci is called *linkage disequilibrium*, or gametic phase disequilibrium. It is equivalent to statistical dependence of allele frequencies between loci.

For a rigorous treatment, we consider more generally two loci, each with an arbitrary number of alleles. Let the frequencies of the alleles $\mathcal{A}_i$ at the $\mathcal{A}$ locus be denoted by p_i and those of the alleles $\mathcal{B}_j$ at the $\mathcal{B}$ locus by q_j. Let the frequency of the gamete $\mathcal{A}_i\mathcal{B}_j$ be P_{ij}, so that $p_i = \sum_j P_{ij}$ and $q_j = \sum_i P_{ij}$. In general, these allele frequencies are no longer sufficient to describe the genetic composition of the population. *Linkage equilibrium* is defined as the state in which

$$P_{ij} = p_i q_j \tag{4.43}$$

holds for every i and j. Otherwise the population is said to be in linkage disequilibrium.

Let the parameter r denote the *recombination frequency*, or *recombination rate*, between the two loci. This is the probability that a recombination event (crossing over) occurs between them. The value of r usually depends on the distance between the two loci along the chromosome. Loci with $r = 0$ are called completely linked (and may be treated as a single locus) and loci with $r = \frac{1}{2}$ are called unlinked. The maximum value of $r = \frac{1}{2}$ typically occurs for loci on different chromosomes, because then all four gametes are produced with equal frequency $\frac{1}{4}$. Thus, the recombination rate satisfies $0 \le r \le \frac{1}{2}$.

Given P_{ij}, we want to find the gametic frequencies P'_{ij} in the next generation after random mating. The derivation of the recursion equation is based on the following basic

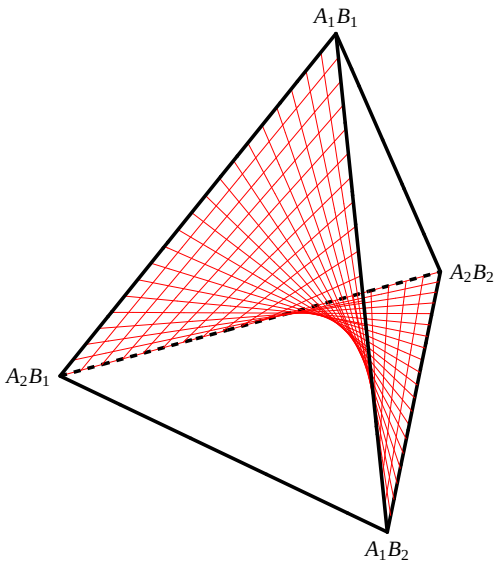

Figure 4.4: The tetrahedron represents the state space of the two-locus two-allele model. The vertices correspond to fixation of the labeled gamete, and frequencies are measured by the (orthogonal) distance from the opposite boundary face. At the center of the simplex all gametes have frequency $\frac{1}{4}$. The two-dimensional surface is the linkage-equilibrium manifold corresponding to the states in linkage equilibrium, $D = 0$. The states of maximum linkage disequilibrium, $D = \pm\frac{1}{4}$, are the centers of the edges connecting $\mathcal{A}_1\mathcal{B}_2$ to $\mathcal{A}_2\mathcal{B}_1$ and $\mathcal{A}_1\mathcal{B}_1$ to $\mathcal{A}_2\mathcal{B}_2$.

fact of Mendelian genetics: an individual with genotype $\mathcal{A}_i\mathcal{B}_j/\mathcal{A}_k\mathcal{B}_l$ produces gametes of parental type if no recombination occurs (with probability $1 - r$), and recombinant gametes if recombination between the two loci occurs (with probability r). Therefore, the fraction of gametes $\mathcal{A}_i\mathcal{B}_j$ and $\mathcal{A}_k\mathcal{B}_l$ is $\frac{1}{2}(1 - r)$ each, and that of $\mathcal{A}_i\mathcal{B}_l$ and $\mathcal{A}_k\mathcal{B}_j$ is $\frac{1}{2}r$ each. From these considerations, we see that the frequency of gametes of type $\mathcal{A}_i\mathcal{B}_j$ in generation $t + 1$ produced without recombination is $(1 - r)P_{ij}$, and that produced with recombination is rp_iq_j because of random mating. Thus,

$$P'_{ij} = (1 - r)P_{ij} + rp_iq_j \ . \tag{4.44}$$

This shows that the gene frequencies are conserved, but the gamete frequencies are not, unless the population is in linkage equilibrium, (4.43). Commonly, linkage disequilibrium between alleles $\mathcal{A}_i$ and B_j is measured by the parameter

$$D_{ij} = P_{ij} - p_iq_j \ . \tag{4.45}$$

The D_{ij} are often called simply linkage disequilibria, although no single D_{ij} is a complete measure of linkage disequilibrium. From (4.44) and (4.45) we infer that

$$D'_{ij} = (1-r)D_{ij} \tag{4.46}$$

and, hence,

$$D_{ij}(t) = (1-r)^t D_{ij}(0) \ . \tag{4.47}$$

Therefore, unless $r = 0$, linkage disequilibria decay at the geometric rate $1-r$ and linkage equilibrium is approached gradually without oscillation. With unlinked loci, $r = \frac{1}{2}$, linkage disequilibrium is halved each generation.

For two alleles at each locus, it is more convenient to label the frequencies of the gametes $\mathcal{A}_1\mathcal{B}_1$, $\mathcal{A}_1\mathcal{B}_2$, $\mathcal{A}_2\mathcal{B}_1$, and $\mathcal{A}_2\mathcal{B}_2$ by x_1, x_2, x_3, and x_4, respectively. A simple calculation reveals that in this case the difference of the frequency of coupling genotypes, $\mathcal{A}_1\mathcal{B}_1/\mathcal{A}_2\mathcal{B}_2$, and repulsion genotypes, $\mathcal{A}_1\mathcal{B}_2/\mathcal{A}_2\mathcal{B}_1$,

$$D = x_1x_4 - x_2x_3 \ , \tag{4.48}$$

satisfies

$$D = D_{11} = -D_{12} = -D_{21} = D_{22} \ . \tag{4.49}$$

Thus, the recursion equations for the gamete frequencies, (4.44), may be rewritten as

$$x'_1 = x_1 - rD \ , \tag{4.50a}$$

$$x'_2 = x_2 + rD \ , \tag{4.50b}$$

$$x'_3 = x_3 + rD \ , \tag{4.50c}$$

$$x'_4 = x_4 - rD \ . \tag{4.50d}$$

The two-locus gametic frequencies may be represented geometrically by the points in a tetrahedron, because $x_1 + x_2 + x_3 + x_4 = 1$. The set of quadruples (x_1, x_2, x_3, x_4), $x_i \geq 0$, satisfying this constraint is called the three-dimensional *simplex*, and denoted by S_4. The subset where $D = 0$ forms a two-dimensional manifold and is called the linkage equilibrium, or Wright, manifold. It is displayed in Figure 4.4.

It follows from (4.47) that, if $r > 0$, all solutions of (4.50) converge to the linkage-equilibrium manifold along straight lines, because the allele frequencies, $x_1 + x_2$ and $x_1 + x_3$, remain constant, and sets of the form $x_1 + x_2 = \text{const.}$ represent planes in this geometric picture. In the present simple model, the linkage-equilibrium manifold is

invariant under the dynamics (4.50). With selection or mutation, this is generally not the case.

If there are more than two loci, linkage disequilibria among any group of at least two loci have to be considered. The model becomes much more complex then. However, it can again be proved that under random mating all gametic combinations eventually reach equilibrium proportions.

4.5 Random genetic drift

So far, we assumed that population sizes are large enough to equate the probability of sampling an allele with its relative frequency. Here we shall briefly explore the consequences of finite population size. In a finite population, changes in allele frequencies must be viewed as a stochastic process, because of variation in the number of offspring produced by different individuals, and because of the stochastic nature of segregation. The stochasticity introduced through such random effects is called *random genetic drift*.

In contrast to the importance attributed to stochastic models in this course, they play a central role in population genetics. The reason is that in finite populations the fate of alleles may be very different from that predicted by an analogous deterministic model. In particular, advantageous alleles that occur as a single mutant (or in very low frequency) may be lost with high probability as long as they are rare. By contrast, slightly deleterious alleles have a non-negligible probability of becoming fixed in a small population.

The most widely used model for studying finite populations is the so-called *Wright–Fisher model*. We introduce it in its simplest form by assuming that there is only one sex and no mutation, selection, geographic dispersal, or else.

Consider a diploid, monoecious population of fixed size with two selectively neutral alleles, $\mathcal{A}_1$ and $\mathcal{A}_2$, at a certain locus. Then there are $2N$ genes in the population and we denote the number of $\mathcal{A}_1$ genes in generation t by $X(t)$. In the Wright–Fisher model, it is assumed that the $2N$ genes in generation $t+1$ are obtained from the $2N$ parental genes in generation t by sampling with replacement. This will be a good approximation if allelic proportions are preserved under reproduction and the number of gametes produced is sufficiently high that removing $2N$ gametes randomly does not change the relative frequencies in the gamete pool. Then $X(t+1)$ is a binomial random variable with index $2N$ and parameter $X(t)/(2N)$. More precisely, given that $X(t) = i$, the probability π_{ij}

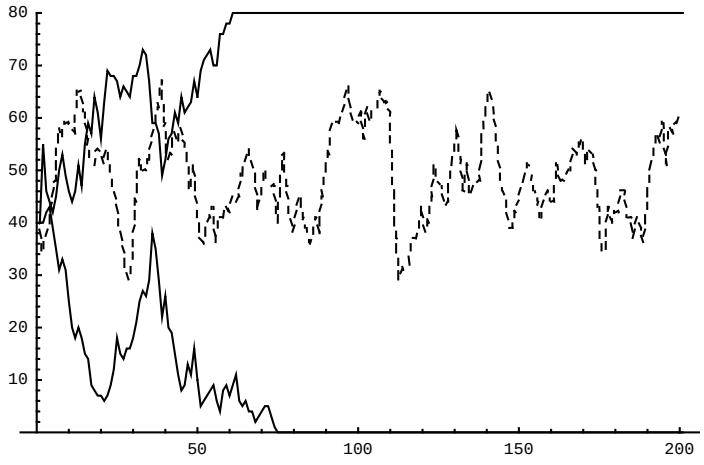

Figure 4.5: Evolution under the Wright-Fisher in a diploid population of size $2N = 80$. Three trajectories are shown, each starting at a frequency of $i = 40$.

that $X(t+1) = j$ is

$$\pi_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} , \tag{4.51}$$

for $i, j = 0, 1, 2, \ldots, 2N$. The π_{ij} are called the transition probabilities and the matrix (p_{ij}) is the transition matrix of the associated Markov chain $X(\,\cdot\,)$. Knowledge of the transition matrix allows one to calculate the probability distribution of $X(t)$ for every generation t if the (probability distribution of the) initial state $X(0) = X_0$ is known, because

$$\Pr[X(t+1) = j] = \sum_{i=0}^{2N} \Pr[X(t) = i]\pi_{ij} . \tag{4.52}$$

For this model let us derive some simple facts about the evolution of a finite population. First, we obtain from (4.52) by using $\sum_j j\pi_{ij} = i$,

$$\mathrm{E}[X(t+1)] = \mathrm{E}[X(t)] = \cdots = \mathrm{E}[X_0] , \tag{4.53}$$

where E denotes the expectation. (We write $\mathrm{E}[X_0]$ because we do not need to know the initial state with certainty, only its distribution.) Similarly, using $\sum_j j^2\pi_{ij} = i + (1 - 1/(2N))i^2$, we get

$$\mathrm{E}[(X(t+1))^2] = \mathrm{E}[X(t)] + \left(1 - \frac{1}{2N}\right) \mathrm{E}[(X(t))^2] . \tag{4.54}$$

Equation (4.53) shows that, on average, allele frequencies remain constant. However, because of random fluctuations, any given population will not maintain a constant frequency of $\mathcal{A}_1$. Indeed, (4.53) implies that the expected heterozygosity decreases geometrically to zero, i.e., if $p(t) = X(t)/(2N)$ and $H(t) = 2p(t)[1 - p(t)]$, then

$$\mathrm{E}[H(t+1)] = \left(1 - \frac{1}{2N}\right) \mathrm{E}[H(t)] \; . \tag{4.55}$$

Thus, random genetic drift eliminates all heterozygotes from the population. Since the population is random mating, this implies that one of the alleles becomes fixed. Once an allelic type is lost, it cannot be reintroduced into the population because this model ignores mutation.

It is easy to calculate the probability of fixation of, say, allele $\mathcal{A}_1$. Obviously, (4.55) implies that $\lim_{t\to\infty} \Pr[X(t) = i] = 0$ for $i = 1, \ldots, 2N-1$. Therefore,

$$\mathrm{E}[X_0] = \lim_{t\to\infty} \mathrm{E}[X(t)] = \lim_{t\to\infty} \sum_{i=0}^{2N} i \Pr[X(t) = i] \tag{4.56}$$

$$= \lim_{t\to\infty} 2N \Pr[X(t) = 2N] \; , \tag{4.57}$$

which shows that the probability of fixation of $\mathcal{A}_1$ is

$$\Pr[\mathcal{A}_1 \text{becomes fixed}] = \frac{\mathrm{E}[X_0]}{2N} \; . \tag{4.58}$$

Thus, if p_0 is the initial frequency of an allele, its fixation probability is also p_0. Much of this theory can be generalized to include mutation, selection, and other evolutionary forces.

There is also another, more intuitive approach to find the fixation probability of $\mathcal{A}_1$. Note that eventually every gene in a population is descended from one unique gene in the initial generation. The probability that such a gene is $\mathcal{A}_1$ is simply the initial fraction of such genes. This must, therefore, be the fixation probability of $\mathcal{A}_1$.

This idea, of going backwards in time, turned out to be essential for modern population genetics because it allows to draw inferences about evolutionary events in the history of a population from observations of their 'footprints' in the genomes of extant populations. The fundamental concept on which such inference is based is the coalescent process (Kingman 1980) which traces the genealogical relationship between genes from a sample backward in time until the most recent common ancestor is found. This is a beautiful mathematical theory of great importance in evolutionary genetics.