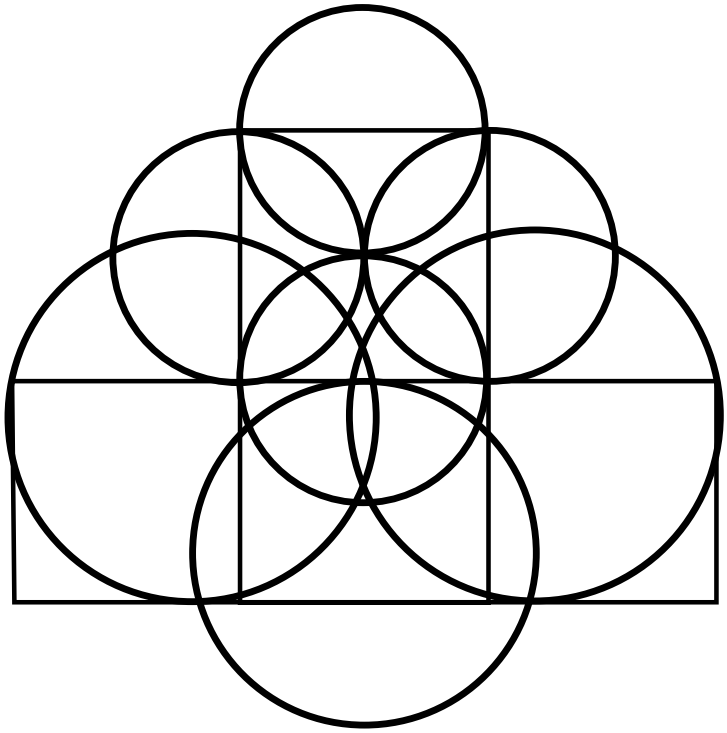
THEORETICAL EVOLUTIONARY GENETICS

JOSEPH FELSENSTEIN



Theoretical Evolutionary Genetics

GENOME 562

Joseph Felsenstein

Department of Genome Sciences
and Department of Biology
University of Washington
Box 355065
Seattle, Washington 98195-5065

December, 2019

Contents

PREFACE

This book was developed as a textbook for my course Genome 562 (Population Genetics). It is nearly in a final form.

Each chapter ends with two sets of problems. Those labeled Exercises are intended to be relatively straightforward application of principles given in the text. They usually involve numerical calculation or simple algebra. The set labeled Problems/Complements are more algebraic, and often involve extension or re-examination of the material in the text.

The level of mathematics required to read this text is not high, although the volume of algebra is sometimes heavy. It is probably sufficient to know elementary calculus, and parts of elementary statistics and probability. Matrix algebra is used in several places, but these can be skipped without much loss. The most relevant mathematical technique for population genetics is probably factorization of simple polynomial expressions, which most people are taught in high school (and then, unfortunately, forget).

The text can be criticized for not introducing the reader to empirical population genetics. That would roughly double the length of the book. Keep in mind that this an introduction to *theoretical* population genetics.

These notes have been developed over the last 34 years. They were not finished rapidly and published primarily because I got interested in phylogenies and was less interested in theoretical population genetics. Nevertheless I needed these to teach my theoretical population genetics course, and so they were gradually expanded. At first they were encoded on magnetic card storage for an IBM word processor. Later we had them transferred to magnetic tape, and hand-edited that into text for the Runoff family of text-formatting programs. They finally became a LaTeX file with Postscript figures.

Many of the references are from the 1970s and earlier. Population genetics theory had its major development in the 1920s-1940s (at the hands of Fisher, Wright, and Haldane) and was finally rigorized in the 1960s and 1970s under the influence of people like Richard Lewontin, James Crow, Motoo Kimura, Sam Karlin, Geoff Watterson, and Warren Ewens. I have been bringing the references up to date, but still find that much of the basic work in theoretical population genetics was done before 1980.

Many people have contributed to the production of these notes, particularly students in earlier years of the course who caught many errors in earlier versions. The presenta-

tions were heavily influenced by lecture notes and courses on this subject by J. F. Crow and R. C. Lewontin. The cover illustration is adapted from an original by Helen Leung. Sean Lamont wrote the plotting program that produced the originals of many of the figures. I am indebted to many people for suggestions and corrections, particularly to Jarle Tufto and his students, and to Eric Anderson, Max Robinson, Weiva Sieh, Tim Reluga, Marissa La Madrid, Norman Ehrentreich, Rich Neapolitan, Phil Hedrick, Eric Rynes, Pui Yee Fong, Fred Allendorf, Sterling Sawaya, Alirio Rosales, Jeff Staples, Benjamin Vernot, Leonard Jones, Rachel Gittelman, Qian Sophia Zhang, and Richard Sharpe. I am especially grateful, for many corrections, to Jeff Thorne and his students, and to Wenying Shou. But most of all, I must thank Nancy Gamble and Martha Katz for doing the enormous job of typing out the early versions of these notes, and Nancy Gamble for drawing some of the figures for earlier editions.

I am in the process of completing this book. The most recent revisions, from 2014 onwards, were made at the Helen R. Whiteley Center at the Friday Harbor Laboratories of the School of the Environment at the University of Washington. I am grateful to the Whiteley Center for providing such a wonderful environment for these revisions. As I retired in 2017, I will have more time to get the book into reasonable shape. I hope to keep it available in a downloadable PDF version. I also will make it available from a print-on-demand publisher for those who do not want to print their own copy and have it bound. But it is unlikely to ever be a conventional printed book – the market for textbooks in theoretical population genetics is small, as there are only a few schools offering courses in this subject. Having it available as a freely downloadable document is effective, and makes it available to students who would have difficulty affording a book.

Joe Felsenstein
Department of Genome Sciences
and Department of Biology
University of Washington
Seattle

joe (at) gs.washington.edu

Chapter I

RANDOM MATING POPULATIONS

Theoretical population genetics (or theoretical evolutionary genetics) is arguably the area of biology in which mathematics has been most successfully applied. Other areas such as theoretical ecology model phenomena which are more immediately important to human welfare, but are nevertheless not as successfully modeled. The major reason why theory is more readily applied to population genetics is that there is a precise framework – Mendelian segregation – on which to hang it. The Mendelian mechanism is a highly regular process with strong geometric and algebraic overtones. The other reason why Mendelian segregation is particularly important to population genetics is that it occurs whether or not natural selection is present, whether or not mutation is present, and whether or not migration is present.

Another mechanism important to population genetics is random mating. This is an approximate model for a messy and complex process, but it is successful for a simple reason – many aspects of the genome do not have the type of effects that influence who mates with who. Random mating is a good default assumption in those cases.

In this chapter we examine the consequences of Mendelian segregation and random mating for the genetic composition of a population. That there can be consequences that are not intuitively obvious follows from one property of Mendelian segregation – that the composition of offspring for some matings differs from the composition of the parents. For example, a cross of $AA \times aa$ yields, not half AA and half aa, but instead Aa.

"Normal" Mendelian segregation is diploid and sexual. To understand it we must start with an examination of the simpler cases in which populations are asexual or haploid. In doing so we hope to make the results of this chapter intuitively obvious – after the fact.

I.1 Asexual inheritance.

TWO GENOTYPES. The first case we cover is one so simple that there is virtually noth-

ing to report. Consider a mixed population of two strains which reproduce asexually (as do many bacteria, dandelions, and maybe bdelloid rotifers). The offspring of this form of uniparental inheritance have genotypes which are exact copies of their parents' genotypes (we are deliberately ignoring the possibility of mutation). Suppose that the population is undergoing synchronous reproduction with nonoverlapping generations. Let the two strains be numbered 1 and 2, and suppose that the number of strain i in some generation t is N_i, for $i = 1$ or 2. Now if each individual has W_t offspring in generation t, irrespective of its genotype, and we denote the number of strain i in the next generation as N_i', then

$$\begin{aligned} N_1' &= W_t N_1, \\ \text{and} & \\ N_2' &= W_t N_2. \end{aligned} \tag{I-1}$$

(In this book, putting a prime (an apostrophe) on a variable such as N_1 will usually indicate the value in the next generation. When it instead means the first derivative, I will specify that.)

The number of offspring of each genotype is simply the number of parents of that genotype, multiplied by the number of offspring each has. (For single-celled organisms that reproduce by cell division, W_t would be 2 in each generation, unless some of the offspring do not survive to reproduce themselves).

Note that we have assumed that the individuals of type 1 have *exactly* the same number of offspring as the individuals of type 2. If the populations are small this is very unlikely to be true, since random environmental circumstances will cause some individuals to have more surviving offspring than others. If there are a very large number of individuals, these circumstances should average out, and the average number of offspring per parent from each strain will be nearly equal.

Consider the fraction of all individuals that are of genotype 1. This is, in generation $t + 1$,

$$\frac{N_1'}{N_1' + N_2'} = \frac{W_t N_1}{W_t N_1 + W_t N_2} = \frac{W_t}{W_t} \frac{N_1}{N_1 + N_2} = \frac{N_1}{N_1 + N_2} \tag{I-2}$$

This establishes the fact that when different genotypes reproduce equally well, the proportion of any one of them does not change. We can make the same point by calculating the ratio of the numbers of one genotype to the other, and noting that the factors W_t cancel:

$$\frac{N_1'}{N_2'} = \frac{W_t N_1}{W_t N_2} = \frac{N_1}{N_2} \tag{I-3}$$

Thus, the proportions and ratios of different genotypes are not changed by asexual reproduction in a large population.

MULTIPLE GENOTYPES. If we had not two, but k different genotypes, the picture is the same. If we denote by p_i the frequency of the i-th genotype in generation t, then if

$$N \;=\; N_1 + N_2 + \cdots + N_k,$$

we find that

$$N' \;=\; W_tN_1 + W_tN_2 + \cdots + W_tN_k \;=\; \sum_{i=1}^{k} W_tN_i \;=\; W_t \sum_{i=1}^{k} N_i \;=\; W_tN,$$

and we have

$$p_i' \;=\; \frac{N_i'}{N'} \;=\; \frac{W_tN_i}{W_tN} \;=\; \frac{N_i}{N} \;=\; p_i, \tag{I-4}$$

so that the frequencies of the different genotypes do not change, even though their numbers may increase or decrease (depending on whether W_t is greater or less than 1).

We will have frequent recourse to the conclusions of this section. In sexual diploids the effect of Mendelian segregation is felt only as one moves from one generation to the next. Within a generation the population is effectively asexual. Thus the logic of this section applies perfectly to the genotypic composition of a single generation in which each individual has probability W_t of surviving to adulthood. From now on we will leave out the factor W_t and simply assume that genotypic compositions are not changed by random survival in infinite populations, provided that survival is unaffected by genotype.

Similarly, when we have a set of sexual offspring and ask who their parents were, we will assume that the composition of the parents is unaffected by differences between individuals in the amount of reproduction they do, provided that the differences in reproduction are independent of genotype, and provided that there are an infinite number of parents.

I.2 Some cautionary notes

I should make two cautionary notes. First, we will frequently use the word "frequency", as in gene frequency, genotype frequency, or haplotype frequency. When we do, we mean *relative* frequency. If in a sample of 200 rabbits, a genotype that makes the fur color white occurs in 6 of them, the genotype frequency is 6/200, or 0.03. More correct statistical usage would say that the frequency of that genotype is 6. Its relative frequency would be 0.03. But in population genetics we use the word "frequency" to refer to the relative frequency, not the number of occurrences.

There is also what appears to be a major self-contradiction in our models. In the previous section we had a model with a finite population, but we assume that the number of offspring is exactly W_1N_1. As mentioned in that section, this implicitly assumes that N_1 is very large, so that the randomness of events such as births and deaths averages out,

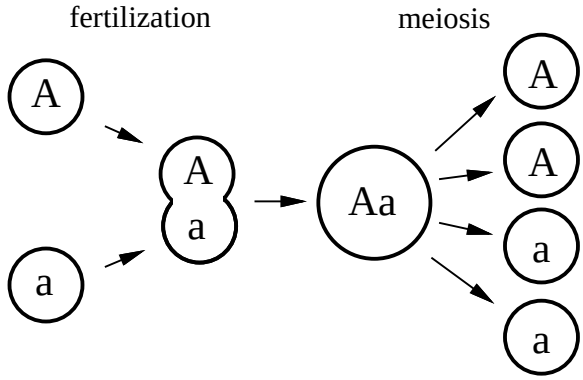


Figure 1.1: Diploid stage of a predominantly haploid organism.

and we can treat the actual numbers of offspring as if they were the same as the expected numbers. Then we can predict what will happen to the total number of individuals N, even though we are in the limit where N is effectively infinite. One way to have it both ways is to have N be the mean population density per unit space, with the population spread over a vast area, so the the actual population number is effectively infinite.

I.3 Haploid inheritance

There are many cases, particularly among microorganisms, of organisms which are haploid during most of their life cycle, having only the briefest of diploid phases. Figure 1.1 shows a typical generation in such an organism.

Suppose that we have a population of haploid organisms of two genotypes, A and a. Let the proportions of these genotypes be p and $1-p$ in generation t. If the organisms mate at random, we can easily compute the proportions of the three resulting diploid genotypes. When mating is random, the genotypes of the two mates are independent of one another. For an AA diploid to be formed the first parent must be an A, which will be true p of the time. The second parent must also be an A. If mating is at random, then given that the first parent is A, the probability that the second parent is also A is unaffected by that, so it too is p. Thus a fraction p of the time the first parent is A and a fraction p of those cases also have the second parent being A.

So an AA diploid is formed in $p \times p = p^2$ of the matings. An aa will be formed $(1-p) \times (1-p)$ of the time. There will be two ways of forming heterozygotes: Aa, with probability $p \times (1-p)$, and aA, with probability $(1-p) \times p$. Since we cannot normally tell these apart, we combine these cases, so that the proportions of the diploid genotypes

are:

$$\begin{array}{ll} AA & p^2 \\ Aa & 2p(1-p) \\ aa & (1-p)^2. \end{array} \tag{I-5}$$

These are the so-called Hardy-Weinberg proportions, actually a simple case of a binomial expansion, having the same probabilities as tossing a coin twice that has probability p of coming up Heads each time. To obtain the proportions of A and a in the next generation, we must consider the results of meiosis in these diploids. It is, of course, assumed that all three genotypes are equally likely to undergo meiosis. Then p^2 of the haploids in the next generation come from AA diploids. All of these haploids must be A, since there is no mutation in this idealized case. $2p(1-p)$ of the haploids will come from Aa diploids, and half of these will be A. All of the $(1-p)^2$ of the gametes which come from aa diploids will be a. The total proportions of A and a among the offspring generation are then

$$\begin{array}{lllllllll} A: & p^2 + 1/2 \times 2p(1-p) & = & p^2 + p(1-p) & = & p\,[p + (1-p)] & = & p, \\ a: & (1-p)^2 + 1/2 \times 2p(1-p) & = & (1-p)^2 + p(1-p) & = & (1-p)\,[(1-p) + p] & = & 1-p. \end{array} \tag{I-6}$$

So we once again, if we denote the gene frequency in generation t by p, and the frequency in generation $t+1$ as p',

$$p' = p, \tag{I-7}$$

so that genotype frequencies remain unchanged from their initial values. It is tempting to consider haploids as exactly equivalent to asexuals. But this is not true when we consider recombination, as we shall see later. We have ignored sex determination. It has been implicitly assumed that, even if there is a mating type system as in yeast, where two alleles, a and α determine the mating types, that the genotype frequencies are the same among both a and α haploids, so that we need not take mating types into account. We will shortly see the consequences of relaxing this assumption. Many of the phenomena of population genetics can be seen most clearly in haploid cases, and we will return to the haploid case more frequently than its biological importance alone warrants.

I.4 Diploids with two alleles: Hardy-Weinberg laws.

DERIVATION. We now consider a random-mating population of diploids in which two alleles are segregating. We assume that there is no difference in genotype proportions between the sexes. Suppose that in generation t the population contains the three genotypes AA, Aa, and aa in proportions P_{AA}, P_{Aa}, P_{aa}. These we henceforth call the *genotype*

frequencies. Consider a haploid gamete produced by one individual chosen at random. The individual has chance P_{AA} of being an AA, and P_{Aa} of being an Aa. In the latter case, the gamete is A one half of the time. The chance that the gamete produced by a randomly chosen individual is A is then p_1 and the chance that it is a is p_2 where

$$\begin{aligned} p_1 &= P_{AA} + \tfrac{1}{2}P_{Aa}, \\ p_2 &= \tfrac{1}{2}P_{Aa} + P_{aa}. \end{aligned} \tag{I-8}$$

p_1 and p_2 will be referred to as the *gene frequencies* of the two alleles. (*Allele frequencies* would be a more consistent term, but *gene frequencies* is solidly entrenched in the literature). They are not only the frequencies of the two types of gametes, but also the proportion of all genes in generation t which are each of the two alleles. We can see this by indirect argument, as follows: P_{AA} of all copies of this gene are in AA individuals, and all of these are A. P_{Aa} of the copies are in Aa individuals, and half of these are A alleles. So the total fraction of all copies which are A is $P_{AA} + \frac{1}{2}P_{Aa}$, which is just the gene frequency p_1. More directly, a randomly chosen haploid gamete contains a copy of a gene chosen at random from the parental diploids. So the probability that such a gamete is A is just the gene frequency, p_1. An alternative approach to this point, involving direct counting of A and a alleles, is given in the next section.

If it happened to be true that random mating of individuals gave the same results as random combination of the pool of gametes, then the following would be true, as a consequence of the results of the previous section:

1. The diploid genotypes in the next generation would occur in the frequencies p_1^2, $2p_1p_2$, p_2^2.
2. The gametes which they produce would be in the same frequencies as the gametes of generation t. So if we use the argument (t) to indicate which generation a gene frequency is from, $p_1^{(t+1)} = p_1^{(t)} = p_1^{(t-1)} = \cdots = p_1^{(0)}$.

It follows from these two principles that not only will the gene frequencies remain constant from one generation to the next, so will the genotype frequencies, with the exception of the initial generation. In fact, it turns out to be true that *random mating is equivalent to random union of gametes*. This is simply the result of the fact that choosing a gamete at random from the pool of gametes is equivalent to sampling a parent at random, and then having it produce a gamete containing one of its two genes (at this locus), chosen at random by the mechanism of Mendelian segregation. The reader who doubts that this is so can consult Table 1.1, which enumerates the possible matings, their probabilities, and the resulting offspring genotype frequencies. The Table makes use of the independence of the genotypes of the two mates under random mating, so that the probability of an $AA \times AA$ mating is $P_{AA} \times P_{AA}$.

Table 1.1: Mating types, their frequencies, their contribution to the offspring genotype frequencies, and the resulting genotype frequencies under random mating.

	Mating Type	Contribution to Offspring Generation		
Mating	Frequency	AA	Aa	aa
$AA \times AA$	$P_{AA} \times P_{AA}$	P_{AA}^2	—	—
$AA \times Aa$	$P_{AA} \times P_{Aa}$	$\frac{1}{2}P_{AA}P_{Aa}$	$\frac{1}{2}P_{AA}P_{Aa}$	—
$AA \times aa$	$P_{AA} \times P_{aa}$	—	$P_{AA}P_{aa}$	—
$Aa \times AA$	$P_{Aa} \times P_{AA}$	$\frac{1}{2}P_{Aa}P_{AA}$	$\frac{1}{2}P_{Aa}P_{AA}$	—
$Aa \times Aa$	$P_{Aa} \times P_{Aa}$	$\frac{1}{4}P_{Aa}^2$	$\frac{1}{2}P_{Aa}^2$	$\frac{1}{4}P_{Aa}^2$
$Aa \times aa$	$P_{Aa} \times P_{aa}$	—	$\frac{1}{2}P_{Aa}P_{aa}$	$\frac{1}{2}P_{Aa}P_{aa}$
$aa \times AA$	$P_{aa} \times P_{AA}$	—	$P_{aa}P_{AA}$	—
$aa \times Aa$	$P_{aa} \times P_{Aa}$	—	$\frac{1}{2}P_{AA}P_{Aa}$	$\frac{1}{2}P_{aa}P_{Aa}$
$aa \times aa$	P_{aa}^2	—	—	P_{aa}^2

The genotype frequencies from Table 1.1 are:

$$\begin{aligned} AA: &\quad P_{AA}^2 + P_{AA}P_{Aa} + 1/4\, P_{Aa}^2 &&= (P_{AA} + 1/2\, P_{Aa})^2 \\ Aa: &\quad P_{AA}P_{Aa} + 1/2\, P_{Aa}^2 + 2P_{AA}P_{aa} + P_{Aa}P_{aa} &&= 2\,(P_{AA} + 1/2\, P_{Aa})\,(1/2\, P_{Aa} + P_{aa}) \\ aa: &\quad 1/4\,(P_{Aa})^2 + P_{Aa}P_{aa} + (P_{aa})^2 &&= (1/2\, P_{Aa} + P_{aa})^2 \end{aligned} \tag{I-9}$$

MEANING. The two principles given above are often known as the Hardy-Weinberg Law. They have two important impacts on population genetics. The first implies that genotype frequencies can (under appropriate conditions) be predicted from gene frequencies. Together with the second, it implies that we can carry through an analysis in terms of gene frequencies instead of genotype frequencies. The second part of the Hardy-Weinberg Law implies that Mendelian reproduction in a random-mating population has no inherent tendency to favor one allele or the other: it will not tend to lose genotypic variability. This is a dramatic difference from the pre-Mendelian scheme of blending inheritance, in which the offspring's genotype (supposed to be contained in its

blood) was a mixture of the parents', without any mechanism of segregation. Blending inheritance would tend to lose half of the genotypic variability each generation, with dramatic consequences for evolution. A professor of engineering in Scotland, Fleeming Jenkin (1867), made this point in response to Darwin's *Origin of Species*. It led him to the conclusion that the response to natural selection would shortly stall for lack of variation. Darwin was unable to convincingly rebut Jenkin. In later editions of the *Origin*, he raised the origin of new variation by direct effects of the environment to a greater importance than he had hitherto assigned it, in order to provide the continuous torrent of new variation necessary to keep evolution operating. With the rise of Mendelian genetics, and the realization of its consequences, the problem vanished.

HISTORY. The Hardy-Weinberg law was discovered by the famous English mathematician G. H. Hardy (1908), and simultaneously and independently in a paper by the German obstetrician and human geneticist Wilhelm Weinberg (1908), whose proof was more generalized. Hardy seems to have deliberately buried his paper in an obscure American journal so that his mathematical colleagues would not realize that he had strayed into applied mathematics. It has sometimes been claimed that William Ernest Castle made use of it in an earlier paper (1903), but a careful reading of that paper will show that Castle worked in terms of genotypes rather than gene frequencies. The Hardy-Weinberg Law is as close to being trivially obvious as it can be, but it had a major impact on the practice of population genetics. Before it, calculations of the effect of natural selection required one to keep track of three variables, the genotype frequencies, and the algebra required to do even simple cases was quite complicated. By focusing attention on the gene frequencies, and establishing the constancy of gene frequencies in the absence of perturbing forces, the Hardy-Weinberg Law greatly simplified calculations. The advances of the next two decades would have come much more slowly and tortuously if it had not been understood. The history of Hardy and Weinberg's work has been well-explained by Crow (1988, 1999) and by Edwards (2008). For a more detailed history of population genetics during the decade of the 1900s, the reader should consult the book by Provine (1968).

EQUILIBRIUM?. The Hardy-Weinberg Law is sometimes referred to as the Hardy-Weinberg Equilibrium. It is an equilibrium in only a restricted sense. If we change the gene frequency of a population, there is nothing inherent in the Law which will restore the gene frequency to its original value. It will remain indefinitely at the new gene frequency. But if we perturb the genotype frequencies *in such a way that the gene frequency is not changed*, then in the next generation Hardy-Weinberg proportions will be restored. If we take a population in Hardy-Weinberg proportions 0.81 AA : 0.18 Aa : 0.01 aa, and alter the genotype frequencies to 0.88 AA : 0.04 Aa : 0.08 aa, then the gamete frequencies will be 0.9 A : 0.1 a, and the offspring generation will once again have genotype frequencies 0.81 AA : 0.18 Aa : 0.01 aa. But had we altered the gene frequency,

the genotype frequencies of the offspring would be in Hardy-Weinberg proportions, but in those dictated by the new gene frequency.

ASSUMPTIONS. To maintain the Hardy-Weinberg principles, we have made many assumptions. Among these are:

1. **Random mating.**
2. **No differential fertility of the genotypes**, so that the contribution a mating type makes to the next generation is simply its frequency among all mating types.
3. **Equal genotype frequencies in the two sexes,** which we have assumed since we use the same three genotype frequencies for both parents.
4. **No mutation**, so that the offspring of any mating are simply those expected from Mendel's laws.
5. **No immigration**, so that all members of the next generation come from the present generation. It is also assumed that there is
6. **No differential emigration**, so that any emigration which occurs does not change the genotype frequencies.
7. **No differential viability**, so that any mortality between newly fertilized zygote and adult stages does not alter the genotype frequencies.
8. **Infinite population size**, so that the proportions of mating types expected from random mating, as well as the proportions of offspring expected from Mendelian segregation are exactly achieved.

Much of the remainder of these notes will be devoted to the consequences of relaxing one or more of these assumptions. We will not be able to cover all possibilities, even superficially, but we should be able to arrive at some intuitive understanding of the effects, singly and in combination, of these various evolutionary forces.

I.5 Where the rare alleles are found.

Hardy-Weinberg proportions imply that homozygotes for rare alleles will be uncommon. This must be emphasized, since it makes it much easier for us to intuitively understand the behavior of natural selection in diploids. The algebra is simple if we calculate the proportion of copies of a rare allele that are found in homzygotes. If the gene frequency of allele A is p, we know that p^2 of all individuals in the population are expected to be homozygotes for that allele. If there are N individuals in the population, we expect that Np^2 of them will be AA homozygotes, and in these there will be a total of $2Np^2$ copies

of the A allele. Overall, there are $2N$ copies of this gene, of which a fraction p are copies of A, so that there are $2Np$ copies of that allele.

The fraction of all copies of A that are expected to be found in AA homozygotes is then

$$\frac{2Np^2}{2Np} \;=\; p$$

which is a dramatic and simple result. If an allele has gene frequency 0.0003, only a fraction 0.0003 of the copies of that allele will occur in homozygotes. Fully 0.9997 of them will be found in heterozygotes. This has strong implications for the effectiveness of selection for or against recessive alleles, and also for the relative importance of the fitness effects of a rare allele when it is heterozygous and when it is homozygous.

It will be helpful to keep in mind that

> Rare alleles occur mostly in heterozygotes; common alleles occur mostly in homozygotes.

because these will help us understand the results of natural selection on rare alleles.

There is an even simpler way to obtain the result of equation (I.5). Imagine that you are a copy of a rare allele, and you have been segregated into a gamete (say an egg). What is the probability that you will end up in a homozygote, paired with another allele like yourself? That is simply the probability that the sperm will contain that rare allele. If mating is at random, the probability of this is your allele frequency, p.

I.6 Multiple alleles.

If, instead of 2 alleles, a population contains n alleles, the principles stated in the previous section either apply or generalize naturally. In a haploid population, we have n different haploid genotypes $A_1, A_2, \ldots, A_n$, whose frequencies in generation t we call $p_1, p_2, \ldots p_n$. When diploids are formed by random mating, the frequencies of the diploid genotypes are simply the products of the respective haploid frequencies. Thus the frequency of the A_1A_1 diploid genotypes is p_1^2 since each of the two haploid genotypes independently has probability p_1 of being A_1. In general (if we count genotype A_iA_j as being distinct from genotype A_jA_i for $i \neq j$),

$$\begin{aligned} A_iA_i: \quad & P_{ii} = p_i^2 \quad && i = 1, 2, \ldots, n \\ A_iA_j: \quad & P_{ij} = p_i\, p_j \quad && i = 1, 2, \ldots, n, \\ & && j = 1, 2, \ldots, n, \\ & && (i \neq j). \end{aligned} \tag{I-10}$$

To keep the notation straight, you must keep in mind that, although we cannot tell A_iA_j and A_jA_i genotypes apart, we count their genotype frequencies P_{ij} and P_{ji} separately, as if we could distinguish them in practice. Thus, the total genotype frequency of A_iA_j and A_jA_i heterozygotes is

$$p_i\,p_j + p_j\,p_i \;=\; 2\,p_i\,p_j. \tag{I-11}$$

If we had a population of diploid genotypes, in which we knew the numbers N_{ii} of A_iA_i homozygotes, and the numbers $N_{ij} + N_{ji}$ of A_iA_j or A_jA_i heterozygotes, we could compute the genotype frequencies directly, by counting A_i genes. There are two A_i genes in each A_iA_i homozygote and one in each A_iA_j heterozygote. If we have N individuals in all, there are $2N$ copies of the A gene, so that the fraction of them which are A_i is

$$\begin{aligned} p_i^* &= [2N_{ii} + (N_{1i} + N_{2i} + \cdots + N_{i-1,i} + N_{i+1,i} + \cdots + N_{ni}) \\ &\qquad +(N_{i1} + N_{i2} + \cdots + N_{i,i-1} + N_{i,i+1} + \cdots + N_{in})]\,/(2N) \\ &= [(N_{1i} + N_{2i} + \cdots + N_{ni}) + (N_{i1} + N_{i2} + \cdots + N_{in})]\,/(2N). \end{aligned} \tag{I-12}$$

Dividing each term of the numerator by $2N$, and noticing that $N_{ij}/N = P_{ij}$,

$$\begin{aligned} p_i^* &= 1/2\;(P_{1i} + \cdots + P_{ni}) \;+\; 1/2\;(P_{i1} + \cdots + P_{in}) \\ &= 1/2\sum_{j=1}^{n} P_{ji} \;+\; 1/2\sum_{j=1}^{n} P_{ij}. \end{aligned} \tag{I-13}$$

This is half the total frequency of genotypes in the ith column of the table, plus half the total frequency of genotypes in the ith row.

In producing the next generation of haploids from a diploid generation with genotype frequencies P_{ij}, the proportion of haploid offspring of genotype A_i is just the gene frequency of A_i in the diploids of the previous generation:

$$p_i' \;=\; p_i^* \;=\; 1/2\;\sum_{j=1}^{n}(P_{ji} + P_{ij}). \tag{I-14}$$

If generation t was itself formed by random mating, then $P_{ij} = p_ip_j$, so if we denote by p_i' the gene frequency in the next generation,

$$\begin{aligned} p_i' &= 1/2\sum_{j=1}^{n}(2\,p_i\,p_j) \\ &= \sum_{j=1}^{n} p_i\,p_j \\ &= p_i\,p_1 \;+\; \ldots \;+\; p_i\,p_n \;=\; p_i\,(p_1 + p_2 + \cdots + p_n), \end{aligned} \tag{I-15}$$

which clearly equals p_i, since the sum of all of the haploid genotype frequencies is 1. So if $p_i^{(t)}$ is the gene frequency in generation t,

$$p_i^{(t+1)} = p_i^{(t)} = \ldots = p_i^{(0)}, \tag{I-16}$$

for all n values of i. Thus the gene frequencies of all n alleles remain constant through time and, by equations (I-10), the diploid genotype frequencies can be predicted from the gene frequencies.

All of the above has been for a haploid organism. The results for diploids are identical. All we need to do is note that the principle that *random mating is equivalent to random union of gametes* is still valid, unaffected by the number of alleles present. Therefore, under the assumptions of the Hardy-Weinberg Law (random mating, no differential fertilities, no sex differences, no mutation, no migration, no differential viabilities, infinite population size), the Hardy-Weinberg Laws still hold. In fact, Weinberg (1908) made his derivation in terms of multiple alleles at the outset.

AN INTUITIVE ARGUMENT. At least part of the results of this section can be seen intuitively. If we classify alleles into two classes, one containing the A_1 allele and the other containing all other alleles, we can consider the resulting population as having two-alleles. The gene frequency of A_1 cannot depend on whether or not the geneticist can perceive differences among the other alleles. Neither can the frequency of A_1A_1 homozygotes. It follows immediately that the gene frequency of A_1 (or of any other allele we choose) must remain constant through time, and that the genotype frequency of A_1A_1 must become the square of the frequency of the A_1 allele. Only the genotype frequencies of the heterozygotes are not predicted by this analogy between two and many alleles.

I.7 Overlapping generations.

So far, the generations have been discrete. One generation gives rise to another, whereupon the parents do not reproduce again, and are no longer counted as part of the population. In that case, the population moves into Hardy-Weinberg proportions in one generation. This life cycle is reasonable only for organisms which breed synchronously and only once in their lifetime (such as annual plants). If there is repeated reproduction and overlapping generations it is not a good representation of the life cycle. A realistic model for continuous reproduction and/or overlapping generations would be quite complex. As a start towards considering such cases, in this section we consider a very simple continuous-time model.

We assume overlapping generations, continuous time, but *not* age-dependent reproduction. The discrete-generation model is one with perfect memory: organisms "remember" exactly when they were born, and reproduce exactly on schedule. But the present

model is the opposite: in each small interval of time, a small fraction of the population, chosen *irrespective* of age, dies. These individuals are replaced by newborns formed by random mating among all existing individuals, again irrespective of age. Since we wish to consider a case parallel to the Hardy-Weinberg situation, we here assume that deaths and births occur irrespective of genotype, that there is no difference in genotype frequencies between sexes, no mutation, no migration, and an infinite population size. The relationship between clock time and generation time is set once we know what fraction of individuals die in a given amount of time, and therefore how rapidly the population turns over. To equate one unit of time with one generation, we assume that during an amount δt of time (assumed to be short), a fraction δt of the population dies and is replaced. This scales the situation so that the probability that an organism survives t units of time is $(1-\delta t)^{t/\delta t}$ which as δt is made small approaches e^{-t}. (You may remember from a calculus course that $(1+1/n)^n$ approaches e as $n \to \infty$, and this is a variant on that result). So lifespan has an exponential distribution, which turns out to have a mean (average) of 1. The process of allowing δt to approach zero is justified by the fact that if the process of death and replacement occurs continuously with constant death rates the probability of survival for δt units of time is $1-\delta t$ only approximately, the approximation improving as δt becomes small.

The newborns who replace the deaths constitute a fraction δt of the population (again approximately: exactly if we let $\delta t \to 0$). They are the result of random mating in the population under Hardy-Weinberg assumptions, so if the current population gene frequency of A is $p_A(t)$, the newborns are of genotype AA with probability $[p_A(t)]^2$. The AA individuals after δt units of time are a mixture of a fraction δt of newborns and $1-\delta t$ of survivors, so if $P_{AA}(t)$ is the frequency of genotype AA at time t:

$$P_{AA}(t+\delta t) \;=\; P_{AA}(t)\;(1-\delta t) \;+\; \delta t\;[p_A(t)]^2 \tag{I-17}$$

and (rearranging)

$$\frac{P_{AA}(t+\delta t) \;-\; P_{AA}(t)}{\delta t} \;=\; [p_A(t)]^2 \;-\; P_{AA}(t). \tag{I-18}$$

Taking the limit as $\delta t \to 0$, the left side of (I-18) is simply the derivative of $P_{AA}(t)$:

$$\frac{dP_{AA}(t)}{dt} \;=\; [p_A(t)]^2 \;-\; P_{AA}(t). \tag{I-19}$$

Similarly, it is easy to show that if $P_{Aa}(t)$ is the frequency of heterozygotes Aa (and aA)

$$\frac{dP_{Aa}(t)}{dt} \;=\; 2\,p_A(t)\;p_a(t) \;-\; P_{Aa}(t). \tag{I-20}$$

Before attempting to solve these equations to find the way $P_{AA}(t)$ changes through time, it will be instructive to look at the gene frequency $p_A(t)$. This is equal to $P_{AA}(t) + \frac{1}{2}P_{Aa}(t)$. We can add together equations (I-19) and (I-20), after multiplying (I-20) by one-half. We get

$$\frac{d(P_{AA}(t) + \frac{1}{2}P_{Aa}(t))}{dt} \;=\; [p_A(t)]^2 \;+\; p_A(t)\; p_a(t) \;-\; P_{AA}(t) \;-\; 1/2\; P_{Aa}(t), \qquad \text{(I-21)}$$

so

$$\frac{dp_A(t)}{dt} \;=\; p_A(t)\; [p_A(t) + p_a(t)] \;-\; p_A(t) \;=\; 0. \qquad \text{(I-22)}$$

So $p_A(t) = p_A(0) = p_A$: the gene frequency does not change, just as we might have expected. Knowing that p_A remains constant, as does p_a, means that we can solve equations (I-19) and (I-20) by treating $p_A(t)$ as a constant.

Before going through any algebraic details, we can see from (I-17) and (I-19) what the result will be. Equation (I-17) shows what is happening: as the initial generation of individuals dies out, it is replaced by newborns who are in Hardy-Weinberg proportions at the constant gene frequency p_A. Ultimately, when the last of the original individuals has died, the population will be in Hardy-Weinberg proportions. Equation (I-19) verifies this conclusion. If $P_{AA}(t) > p_A^2$, then we have more AA individuals than Hardy-Weinberg proportions would predict. Then the right side of (I-19) is negative, so that $P_{AA}(t)$ decreases. Likewise, when $P_{AA}(t) < p_A^2$, it will increase. Ultimately $P_{AA}(t) = p_A^2$, and P_{AA} will not change further.

We can solve (I-19) by elementary separation of variables and integration. It first becomes

$$\frac{dP_{AA}(t)}{[p_A(t)]^2 \;-\; P_{AA}(t)} \;=\; dt. \qquad \text{(I-23)}$$

Then (remembering that $p_A(t) = p_A$ is constant) we can integrate both sides:

$$\int \frac{1}{p_A^2 \;-\; P_{AA}(t)}\, dP_{AA}(t) \;=\; \int dt, \qquad \text{(I-24)}$$

which yields on taking the natural logarithm

$$-\ln [p_A^2 \;-\; P_{AA}(t)] \;=\; t + C. \qquad \text{(I-25)}$$

(In this book I will use $\ln$ rather than $\log_e$ to denote the natural logarithm).

We can determine the value of the unknown constant C by setting $t = 0$. Then

$$C \;=\; -\ln (p_A^2 \;-\; P_{AA}(0)). \qquad \text{(I-26)}$$

So

$$\ln (p_A^2 - P_{AA}(t)) \;=\; -t \;+\; \ln (p_A^2 \;-\; P_{AA}(0)). \qquad \text{(I-27)}$$

Taking the exponential function (e^x) of both sides of this equation:

$$p_A^2 \; - \; P_{AA}(t) \; = \; [p_A^2 \; - \; P_{AA}(0)] \; e^{-t}. \tag{I-28}$$

which shows that the deviation of $P_{AA}(t)$ from the Hardy-Weinberg proportion p_A^2 decays exponentially with time. Solving for $P_{AA}(t)$:

$$P_{AA}(t) \; = \; P_{AA}(0) \; (e^{-t}) \; + \; p_A^2 \; (1 - e^{-t}). \tag{I-29}$$

This confirms precisely the explanation already given. As time passes, a fraction e^{-t} of the population consists of survivors of the original population. A fraction $P_{AA}(0)$ of these are AA. All individuals born later are in Hardy-Weinberg, proportions, so that a fraction p_A^2 of them are AA. Analogous equations hold for P_{Aa} and P_{aa}. While $P_{AA}(t)$ approaches its limiting value exponentially, and never quite reaches it, all newborns are in Hardy-Weinberg proportions. In that sense, Hardy-Weinberg proportions are reached in one generation.

In the remainder of this book we will rarely make use of the overlapping-generations models, but you should keep in mind that there are overlapping-generations versions of some of the models treated here. However, overlapping-generations models are generally far less tractable than discrete-generations models. This is mostly because Hardy-Weinberg proportions cannot be assumed. As we have seen, they are approached only asymptotically even with random mating. If there is any evolutionary force, such as natural selection, making the population continually depart from Hardy-Weinberg proportions, we will have to follow genotype frequencies rather than gene frequencies, which makes life harder. In discrete-generations models one is usually in Hardy-Weinberg proportions once per generation, when the new generation of zygotes is produced.

The monograph by Charlesworth (1980) should be consulted for a clear review of the problems involved in extending overlapping-generations models to cases in which birth and death rates are age-dependent. The paper by Nagylaki and Crow (1974) should also be consulted on this issue.

I.8 Different Gene Frequencies in the Two Sexes

We have been assuming that the genotype frequencies are the same in both sexes. We now relax that assumption, in a discrete generations model which otherwise obeys all of the Hardy-Weinberg assumptions. We follow a population in which two alleles segregate. Suppose that in the initial generation the gene frequencies of A in females and in males are, respectively p_f and p_m. Random mating is equivalent to the combination of a random female gamete with a random male gamete. Table 1.2 shows the resulting genotypes:

Table 1.2: Genotype frequencies when gene frequencies differ in the sexes.

		Female Gametes:	
		A	a
Male Gametes:		p_f	$1-p_f$
A	p_m	$p_f p_m$	$p_m(1-p_f)$
a	$1-p_m$	$p_f(1-p_m)$	$(1-p_f)(1-p_m)$

which give the genotype frequencies:

$$\begin{array}{ll} AA & p_f\, p_m \\ Aa & p_f\,(1-p_m)\ +\ p_m\,(1-p_f) \\ aa & (1-p_f)(1-p_m) \end{array} \qquad \text{(I-30)}$$

We are assuming that the gene A is unlinked to the sex chromosome or sex-determining locus. Thus in the offspring generation the genotypes AA, Aa, and aa are distributed independently of the sex of the offspring. So in that generation, although the genotypes may not be in Hardy-Weinberg proportions, they are the same in both sexes. Therefore the next offspring generation is produced by parents with equal gene frequencies in both sexes, and it will therefore be in Hardy-Weinberg proportions, as will all subsequent generations. Putting primes on the p_f's and p_m's to denote the next generation, the gene frequency in the gametes forming the offspring generation is

$$\begin{aligned} p'_m\ =\ p'_f\ &=\ p_f\, p_m\ +\ \tfrac{1}{2}[p_f(1-p_m)\ +\ p_m(1-p_f)] \\ &=\ p_f\, p_m\ +\ \tfrac{1}{2}\, p_f\ -\ \tfrac{1}{2}\, p_f\, p_m\ +\ \tfrac{1}{2}\, p_m\ -\ \tfrac{1}{2} p_f\, p_m \\ &=\ \tfrac{1}{2}\, p_f\ +\ \tfrac{1}{2}\, p_m. \end{aligned} \qquad \text{(I-31)}$$

It is entirely intuitively obvious why this must be so. The gametes produced by the first offspring generation contain in half of them genes coming from the initial female generation, and in half of them genes coming from the initial males. This is true even if there is a great inequality of the sex ratio: even if there are very few females (say), the symmetry of mating - the fact that each mating consists of one male and one female - ensures that (I-31) will hold. The totality of male genes is copied into the next generation as many times as the totality of female genes.

The picture we get from all this is that after starting with unequal male and female gene frequencies, we do not reach Hardy-Weinberg proportions in the offspring. But we *do* achieve equal gene frequencies in the two sexes of the offspring. In the second

generation Hardy-Weinberg proportions are achieved. So the effect of unequal gene frequencies in the two sexes is to delay achievement of Hardy-Weinberg proportions by one generation. We can still say that the overall gene frequency of the population does not change. But we can only say this if we define it as $p = \frac{1}{2}p_f + \frac{1}{2}p_m$, *irrespective of the actual numbers of the two sexes.* In other words, we must count the aggregate of all females as contributing as much to the population gene frequency as the aggregate of all males. Any other weighting system - such as counting each individual as equivalent - will lead to the population gene frequency changing during the first generation.

In this presentation, p has been the frequency of an allele A, and $1 - p$ of a. But we could as easily have designated $1 - p$ as being the frequency of all other alleles than A. So the above argument applies to the frequency of an allele A irrespective of how many other alleles there are. Having multiple alleles in a population will not alter the conclusions.

GENOTYPE FREQUENCIES. Finally, we verify the direction of departure of genotype frequencies from Hardy-Weinberg proportions. Suppose that, instead of having variables p_f and p_m, we measure the gene frequency in each sex as the average gene frequency plus (or minus) a deviation from that quantity, so that

$$\begin{aligned} p_f &= p + \delta \\ p_m &= p - \delta. \end{aligned} \tag{I-32}$$

Then the genotype frequencies in the next generation are:

$$\begin{aligned} AA &\quad (p+\delta)(p-\delta) \\ Aa &\quad (p+\delta)(1-p+\delta) \ + \ (p-\delta)(1-p-\delta) \\ aa &\quad (1-p-\delta)(1-p+\delta). \end{aligned} \tag{I-33}$$

or (collecting terms)

$$\begin{aligned} AA &\quad p^2 \ - \ \delta^2 \\ Aa &\quad 2\,p\,(1-p) \ + \ 2\delta^2 \\ aa &\quad (1-p)^2 \ - \ \delta^2. \end{aligned} \tag{I-34}$$

This demonstrates that in the two allele case, if there is any difference between gene frequencies in the sexes, if $\delta \neq 0$, there will be a departure from Hardy-Weinberg proportions in the next generation. Furthermore, whether δ is positive or negative, the result is the same: there are fewer homozygotes and more heterozygotes than we would expect from Hardy-Weinberg proportions.

With multiple alleles, there must also be a deficit of each homozygote class, and also an average excess of heterozygotes compensating for this. But specific heterozygote classes can be in deficit, despite the fact that there is an overall excess of heterozygotes.

Biologically, the main implication of the results of this section is that for autosomal loci, we would not expect to see gene frequency differences between the sexes unless some evolutionary force continually created such differences. This has an interesting implication for differentiation of the sexes: it will be difficult to explain it by genotypic differences at loci that are not linked to the sex-determining loci.

I.9 Sex linkage.

HAPLOIDS. We get quite different results when the locus in question is on the sex chromosome. In the haploid case, the results are a bit trivial. If the system resembles yeast, we may have two sex-determining alleles (say S and s). Each mating must be between an S and an s haploid, producing heterozygous diploids. The "sexes" of the offspring are determined by which of the two alleles the haploid receives in the segregation of the diploid. If we follow another allele which is completely linked to the sex-determining locus, the results are rather obvious. If we have an allele *(A)* which has gene frequency p_S among the S haploids, and p_s among the s haploids, neither of these gene frequencies will change. The allele linked to the S haploid in any mating will show up only in the S haploid offspring. The same, of course, holds for s. Figure 1.2 may help you see this.

If the sexes are determined by a region of the chromosome instead of by a single locus, there will be powerful natural selection to prevent this region of the sex chromosome from having any crossing-over that might cause recombination. Typically a recombination event will create a chromosome that causes the individual to be an intersex, capable of successfully mating with neither of the existing sexes. If natural selection suppresses crossing-over in that region, the two forms of the sex-determining segment will segregate as if they were alternative alleles.

DIPLOIDS. When the organism is diploid, with an X-Y chromosome sex-determination, the situation is both more complex and more interesting. Now we assume that a sex-linked locus is carried on the X chromosome, with no counterpart on the Y. Suppose that allele A has gene frequency p_f among X-bearing gametes from females, and frequency p_m among X-bearing gametes from males. Since female offspring contain one X from their male parent, and one from their female parent, then under Hardy-Weinberg

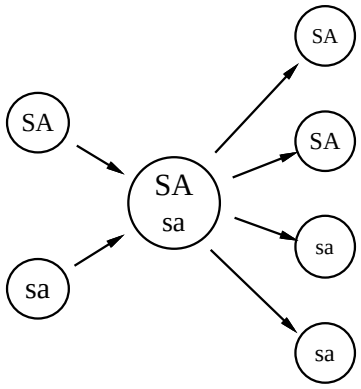


Figure 1.2: Segregation of an allele completely linked to a sex-determining locus in a haploid organism

.

conditions the genotype frequencies in the female offspring are

$$\begin{array}{ll} AA & p_f\ p_m \\ Aa & p_f(1-p_m)\ +\ p_m(1-p_f) \\ aa & (1-p_f)(1-p_m), \end{array} \tag{I-35}$$

which are exactly the same as these genotype frequencies would be for an autosomal locus in which gene frequencies differ between the sexes. But we cannot expect to see Hardy-Weinberg proportions in only two generations. After the first generation we do not have equal gene frequencies in both sexes in this case, because the locus is linked to the sex-determining chromosome. In male offspring, the genotype frequencies are:

$$\begin{array}{ll} AY: & p_f \\ aY: & 1-p_f. \end{array} \tag{I-36}$$

We can easily calculate the gene frequency of A among the gametes coming from these offspring. In males there is no algebra to do. In females the algebra is identical to that in Equation (I-31) of the previous section. Placing primes on the p's to indicate the next generation, the results are:

$$\begin{aligned} p'_f &= \tfrac{1}{2}p_f + \tfrac{1}{2}p_m \\ p'_m &= p_f. \end{aligned} \tag{I-37}$$

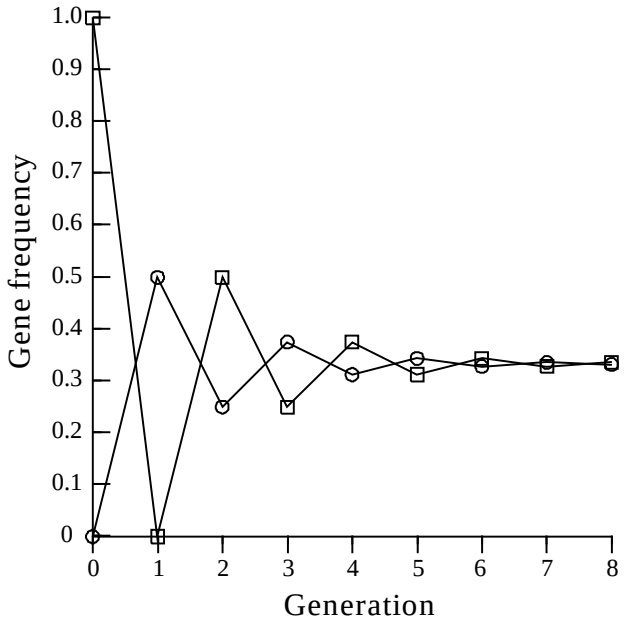


Figure 1.3: Gene frequency changes resulting from initial sex differences of gene frequencies in the two sexes (circles are for females, squares for males) at a sex-linked locus (with initial gene frequencies $p_m = 1$, $p_f = 0$).

LONG-TERM BEHAVIOR. It is not immediately obvious what are the long term implications of these relations. Figure 1.3 shows the results if we start with $p_m = 1$ and $p_f = 0$. Clearly the gene frequencies do not settle down immediately, but oscillate to an equilibrium.

There are methods available for the complete solution of simultaneous difference equations such as (I-37). But here we will take a short cut which we can only do once we know the answer in advance. Suppose that we arbitrarily decided to look at the quantity $\frac{2}{3}p_f + \frac{1}{3}p_m = p$. Then from (I-37),

$$\begin{aligned} p' &= \tfrac{2}{3}p'_f + \tfrac{1}{3}p'_m &&= \tfrac{2}{3}(\tfrac{1}{2}p_f + \tfrac{1}{2}p_m) + \tfrac{1}{3}p_f \\ &= \tfrac{2}{3}p_f + \tfrac{1}{3}p_m &&= p, \end{aligned} \tag{I-38}$$

so this quantity does not change through time. It is a weighted average of the gene frequencies in females and in males. The weighting assigns twice as much weight to females as to males. This may seem to be straightforward: that each X chromosome is being counted once. But notice that it is irrespective of the sex ratio: the males as a whole are given half as much weight as the aggregate of all females. As in the previous section, if there are very few males, this is compensated for by the fact that each male

will then mate more times than each female (on average). This is a simple consequence of the fact that each mating involves one male and one female.

If the gene frequencies of the two sexes converge to the same value, then since at that point $p_f = p_m$, from (I-37) if the initial gene frequencies are $p_f(0)$ and $p_m(0)$

$$p_f \;=\; p_m \;=\; p \;=\; \frac{2}{3}\, p_f(0) \;+\; \frac{1}{3}\, p_m(0). \tag{I-39}$$

But will this equilibrium value always be approached? We can examine this by computing the difference between the female and male gene frequencies, and seeing how this changes in successive generations:

$$\begin{aligned} p'_f - p'_m &= (\tfrac{1}{2}p_f + \tfrac{1}{2}p_m) - p_f \\ &= \tfrac{1}{2}p_m - \tfrac{1}{2}p_f \\ &= (-\tfrac{1}{2})(p_f - p_m). \end{aligned} \tag{I-40}$$

So the magnitude of the differences between the gene frequencies in the two sexes decreases by half every generation, and it changes sign every generation. Convergence of the male and female gene frequency is certain, irrespective of their initial values.

GENOTYPE FREQUENCIES. When both gene frequencies are equal, (I-35) and (I-36) are:

	Females		Males	
AA	p^2	AY	p	(I-41)
Aa	$2\,p\,(1-p)$	aY	$1-p$	
aa	$(1-p)^2$.			

When p is small, this has the interesting property that male hemizygotes for A will be much more common than female homozygotes: if $p = 0.01$, then $p^2 = 0.0001$. This is of course reasonable: to get a hemizygote for A we need only one copy of the rare allele, but to have an AA female two rare alleles must be present in the same individual.

If we have multiple alleles, the results are the same: the frequency of each allele oscillates to an equilibrium value which is $\frac{2}{3}p_f(0) + \frac{1}{3}p_m(0)$, the oscillations being reduced in magnitude by one-half in each generation. But if we have a model of continuous overlapping generations without age effects (analogous to Section I.7), there are no oscillations! Nagylaki (1975b) has demonstrated that in such cases the gene frequencies in the two sexes approach each other smoothly from their initial values, reaching the same equilibrium values as calculated above.

Although our calculations have been stated in terms of an X-Y system, we may make the following comments about other systems of sex determination:

1. An XX-XO system will behave like an XX-XY system in this respect.

2. A ZW-ZZ system (as in birds or lepidoptera, where the female is the heterogametic sex) will behave like an XX-XY system with sex labeling reversed.

3. A haplo-diploid sex determination system, as in Hymenoptera (males coming from unfertilized haploids and females from fertilized eggs) will have every locus in the organism segregating as if sex-linked.

The oscillating approach to equilibrium genotype frequencies was first shown by H. S. Jennings, a pioneer protozoan geneticist, in 1916.

I.10 Linkage.

INDEPENDENCE OF GENOTYPES AT TWO LOCI. Let us consider two linked loci, each with two alleles. The gene frequency of allele A will be p_A, the frequency of a being $1 - p_A$. Likewise, the gene frequency of B will be p_B, and of b, $1 - p_B$. It is a basic property of the Mendelian system that the segregation of one locus is not affected by the genotypes of neighboring loci. So each locus will individually follow the Hardy-Weinberg laws if the assumptions underlying those laws apply, as we now assume. Then p_A and p_B will each remain constant through time. The genotype frequency of AA will be p_A^2 after the first generation, and similarly the frequency of BB will be p_B^2. But what about the frequency of $AABB$? Can we assume that the genotypes at the two loci are independent, and compute the genotype frequency of $AA\ BB$ as $p_A^2 p_B^2$? If so, is this situation reached after one, two or many generations of random mating?

A RETROSPECTIVE DERIVATION. To investigate this we must compute gamete frequencies. An $AA\ BB$ individual is the product of the fusion of two AB-bearing gametes. In thinking about gamete frequencies, we discover that they cannot simply be computed from gene frequencies. They have a life of their own. Consider two populations, each having $p_A = \frac{1}{2}$ and $p_B = \frac{1}{2}$. The first consists of half $AA\ BB$ individuals and half *aa bb*. There are only two gamete types produced by this population: AB and *ab*, in equal frequencies. On the other hand, the population might consist of half *AB/ab* and half *Ab/aB* individuals (it is necessary in this case for us to know the phase of the double heterozygotes). Then whatever the recombination fraction between the loci, one-quarter of all gametes will be AB. So we must consider gamete frequencies as well as gene frequencies. Different haploid genomes, such as found in gametes, are often referred to as *haplotypes*.

Let P_{AB} be the frequency of AB among the gametes that formed generation t. We want to compute P'_{AB}, the frequency in the next generation. There are two ways in which this could be done. One is to enumerate all possible matings. The other makes use of

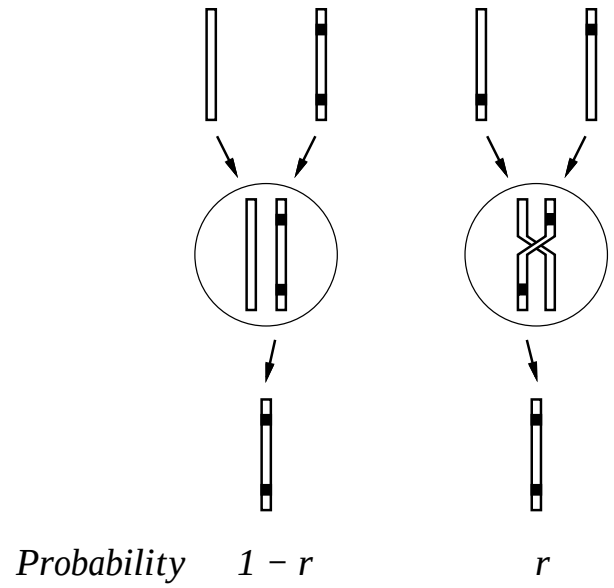


Figure 1.4: A two-locus haplotype in a gamete is derived either from a single haplotype one generation earlier or from the different loci in the two haplotypes of the parent, depending on whether there has been a recombination between those loci. The probabilities of these two origins are respectively $1 - r$ and r.

a shortcut. Consider a gamete of the next generation, and let r be the recombination fraction between these two loci. We need not restrict ourselves to the case where the two loci are on the same chromosome: if they are not, $r = \frac{1}{2}$. In the next generation, $1 - r$ of the gametes will not have suffered any fresh recombination between these two loci. The gamete frequency of AB in these gametes will be the same as in the previous generation. But r of the time, there will have been a recombination. Then the gamete will be AB only if one gamete coming into the parent carried an A, and the other a B. But we have assumed random mating, so that the two gametes which go to make up an individual are chosen randomly and independently of one another. Then the chance that one is A, and the other B, is simply $p_A p_B$. We do not need to inquire about the other gene copy at either of these two loci, since we are not concerned with the genes which are not copied into the gamete. Putting all of this together,

$$P'_{AB} \;=\; (1-r)\; P_{AB} \;+\; r\; p_A\; p_B. \tag{I-42}$$

The result in the long run can be seen by subtracting $p_A p_B$ from both sides of (I-42). Then

$$\begin{aligned} P'_{AB} - p_A\, p_B &= (1-r)\, P_{AB} \;+\; r\, p_A\, p_B \;-\; p_A\, p_B \\ &= (1-r)\,(P_{AB} \;-\; p_A\, p_B). \end{aligned} \tag{I-43}$$

A MEASURE OF NONINDEPENDENCE. The quantity $P_{AB} - p_A p_B$ measures the difference between the actual gamete frequency of AB and the hypothetical frequency which would obtain if the presence of A in a gamete (an event with probability p_A) were independent of the presence of B. Let us call this difference $D_{AB}(t)$ in generation t. Then

$$\begin{aligned} D_{AB}(t) &= (1-r)\; D_{AB}(t-1) \\ &= (1-r)^t\; D_{AB}(0). \end{aligned} \tag{I-44}$$

Provided there is any recombination between the two loci $(1-r)$ is less than unity, so that as $t \to \infty$, D_{AB} approaches zero. When D_{AB} is zero, not only does

$$P_{AB} \;=\; p_A\; p_B, \tag{I-45}$$

but the genotype frequency of $AA\ BB$, being P^2_{AB}, is then $p^2_A p^2_B$. So ultimately we end up in a state where each locus is in Hardy-Weinberg proportions and the occurrence of genotypes at the two loci is independent of each other. This latter state is usually called *linkage equilibrium*, and the measure D_{AB} is the amount of *linkage disequilibrium*. The name is somewhat misleading. It seems to imply that there will be no linkage disequilibrium if there is no linkage. But equations (I-43) and (I-44) show that this is not so. If there is no linkage $r = \frac{1}{2}$. Then D_{AB} declines by half each generation. It will rapidly become quite small, but will not be exactly zero if it is initially nonzero. In fact, there is little difference between two loci being far apart on the same chromosome, or being unlinked. Some authors have preferred "gametic phase imbalance" instead of "linkage disequilibrium," but the latter phrase seems impossible to dislodge from the literature. Linkage disequilibrium is commonly referred to as "LD", and it is not clear that all users of the initials know what name it stands for.

The decline of D_{AB} at the rate $(1-r)^t$ has a straightforward interpretation. Note that we can give a general expression for the chromosome frequency $P_{AB}(t)$:

$$\begin{aligned} P_{AB}(t) &= p_A\, p_B \;+\; (P_{AB}(0) \;-\; p_A\, p_B)\; (1-r)^t \\ &= P_{AB}(0)\,(1-r)^t \;+\; [1-(1-r)^t]\; p_A\, p_B. \end{aligned} \tag{I-46}$$

Note that $(1-r)^t$ is the probability that a gamete passes through t generations without suffering a recombination. The first term on the right side represents the contribution to the gamete frequency of AB from those gametes which have never suffered recombination between these loci since the initial generation. The persistence of part

these unrecombined gametes is the reason for the persistence of part of the initial linkage disequilibrium. Note that the right-hand term on the right side of (I-46) implies that any gamete which has ever suffered a recombination has an expected frequency of $p_A p_B$ for AB, irrespective of the initial gamete frequency $P_{AB}(0)$. It is the fact of random mating each generation which allows us to reach this conclusion. In particular, for (I-43) and hence (I-44) to hold, the initial generation must itself have been formed by random mating. Otherwise we could only write $D_{AB}(t) = (1-r)^{t-1} D_{AB}(1)$. In either case, D tends to zero. Recombination gradually scrambles the initial associations of alleles at different loci, until a state of complete randomness is obtained, in which each chromosome is a patchwork of segments derived from different ancestors.

The implications of linkage equilibrium go unnoticed by many geneticists. Suppose the population is in linkage equilibrium. Then if a plague carries off all but the AA individuals, what will happen to the gene and genotype frequencies at the unselected B locus? Precisely nothing! Among the A-bearing gametes, the fraction which are B is simply p_B. And among AA individuals, the fraction which are BB is simply p_B^2. This illustrates a general principle, that if linkage equilibrium is maintained, natural selection at one locus will not affect another.

SIMPLIFYING POPULATION GENETICS: THE GENE POOL. It should not go unmentioned that linkage equilibrium among all combinations of loci allows a vast reduction in the number of variables required to describe genotype frequencies. Consider a genotype at twenty loci, each of which can have two alleles. There are 2^{20} different gametes possible, so that there are $2^{20} \times 2^{20} = 2^{40}$ possible genotypes. Of course, we usually cannot tell coupling from repulsion double heterozygotes, or which alleles came from the maternal and which from the paternal gamete. Since we can observe at each locus only three distinct genotypes, there are merely 3^{20} distinguishable genotypes. But this is still 3,486,784,401 genotypes! We can predict the genotype frequencies from the gamete frequencies, of which there are 1,048,576. We can discard one of these as an independent quantity, since the sum of all gamete frequencies must be unity. This does not help much. But linkage equilibrium does. At one stroke, it allows us to compute all genotype and gamete frequencies from only 20 quantities, the gene frequencies! It is this simplification which allows us to speak of the evolving population in terms of changes in its "gene pool," the collection of its gene frequencies. If linkage equilibrium does not hold, the best we could do would be to consider it as a "gamete pool".

A MORE DIRECT DERIVATION. Now let us briefly consider the other, more exhaustive proof of the approach to linkage equilibrium. We consider the four types of gametes: AB, Ab, aB, and ab, designating their frequencies P_{AB}, P_{Ab}, P_{aB}, and P_{ab}. Consider all of the parents from which a AB gamete might emerge. These are given in Table 1.3, along with their genotype frequencies and the proportion of their gametes which are AB.

Table 1.3: Genotype frequencies of genotypes giving rise to AB gametes, and the frequencies with which they do so.

Genotype	Frequency (assuming random mating)	Proportion of AB (among gametes)
AB/AB	P_{AB}^2	1
AB/Ab or Ab/AB	$2P_{AB}\,P_{Ab}$	$\frac{1}{2}$
AB/aB or aB/AB	$2P_{AB}\,P_{aB}$	$\frac{1}{2}$
AB/ab or ab/AB	$2P_{AB}\,P_{ab}$	$\frac{1}{2}(1-r)$
Ab/aB or aB/Ab	$2P_{Ab}\,P_{aB}$	$\frac{1}{2}r$

The resulting frequency of AB is

$$\begin{aligned} P'_{AB} &= (P_{AB})^2 + (P_{AB}P_{Ab}) + (P_{AB}P_{aB}) + (P_{AB}P_{ab}) - r(P_{AB}\,P_{ab} - P_{Ab}\,P_{aB}) \\ &= P_{AB}[P_{AB} + P_{Ab} + P_{aB} + P_{ab}] - r(P_{AB}P_{ab} - P_{Ab}P_{aB}) \\ &= P_{AB} - r(P_{AB}\,P_{ab} - P_{Ab}\,P_{aB}). \end{aligned} \tag{I-47}$$

This does not look familiar, until we consider the quantity in parentheses on the last line of the equation. Note that

$$\begin{aligned} P_{AB} + P_{Ab} &= p_A, \\ P_{AB} + P_{aB} &= p_B, \\ P_{ab} &= 1 - P_{AB} - P_{Ab} - P_{aB}. \end{aligned} \tag{I-48}$$

Substituting (I-48c) into (I-47),

$$\begin{aligned} P'_{AB} &= P_{AB} - r\,[P_{AB} - P_{AB}\,P_{AB} - P_{AB}\,P_{Ab} - P_{AB}\,P_{aB} - P_{Ab}\,P_{aB}] \\ &= P_{AB} - r\,[P_{AB} - (P_{AB} + P_{Ab})\,(P_{AB} + P_{aB})] \\ &= P_{AB} - r\,[P_{AB} - p_A\,p_B] \\ &= (1-r)\,P_{AB} + r\,p_A\,p_B, \end{aligned} \tag{I-49}$$

which is simply (I-42). Comparing the quantity in parentheses in the last line of (I-47) to the quantity in brackets in the third line of (I-49) demonstrates an interesting fact:

$$P_{AB}\,P_{ab} - P_{Ab}\,P_{aB} = P_{AB} - p_A\,p_B. \tag{I-50}$$

This implies that the linkage disequilibrium D_{AB} is half the difference between the frequencies of coupling and repulsion double heterozygote genotype frequencies.

HAPLOTYPE FREQUENCIES IN TERMS OF D. Since D_{AB} is defined as $P_{AB} - p_A\, p_B$, by simple rearrangement we can write P_{AB} in terms of $p_A\, p_B$ and D_{AB}. We can do the same for the other three haplotypes:

$$\begin{array}{rcl} P_{AB} &=& p_A\, p_B \;+\; D_{AB} \\ P_{Ab} &=& p_A\, p_b \;+\; D_{Ab} \\ P_{aB} &=& p_a\, p_B \;+\; D_{aB} \\ P_{ab} &=& p_a\, p_b \;+\; D_{ab} \end{array} \tag{I-51}$$

For this case with two alleles at each locus, we might think that we need two gene frequencies and four linkage disequilibrium parameters to predict the four haplotype frequencies. That seems like too many. In fact if we add the first two equations, we get

$$P_{AB} \;+\; P_{Ab} \;=\; p_A\,(p_B \;+\; p_b) \;+\; D_{AB} \;+\; D_{Ab} \tag{I-52}$$

But the quantity in parentheses is the sum of the frequencies of all alleles at the B locus, and so it must be 1. Making that substitution we recognize that in view of equation (I-48) we must the have

$$D_{Ab} \;=\; -D_{AB}. \tag{I-53}$$

A similar derivation using the first and third equations shows that

$$D_{aB} \;=\; -D_{AB}. \tag{I-54}$$

I leave it to you to persuade yourself that, since the frequencies of the four haplotypes sum to 1, it is also true that

$$D_{ab} \;=\; D_{AB}. \tag{I-55}$$

Then we can write the four haplotype frequencies in terms of only three quantities (as we know that $p_a \;=\; 1 - p_A$ and $p_b \;=\; 1 - p_B$):

$$\begin{array}{rcll} P_{AB} &=& p_A\, p_B & +\; D \\ P_{Ab} &=& p_A\,(1 - p_B) & -\; D \\ P_{aB} &=& (1 - p_A)\, p_B & -\; D \\ P_{ab} &=& (1 - p_A)\,(1 - p_B) & +\; D \end{array} \tag{I-56}$$

This is more comforting: we knew that we had four quantities that must add to 1. Now we have predicted them from three variables: two gene frequencies and one linkage disequilibrium parameter.

All of the above proofs have been for the case of two alleles at each locus. The first proof did not refer to a or b at all. It would not have altered things at all had there been several alternatives instead of just a and b. The principle of approach to linkage equilibrium proportions at a rate $(1-r)^t$ holds for any number of alleles, and for each gamete type (say A_6B_4) we can compute a linkage disequilibrium measure $D_{A_6B_4} = P_{A_6B_4} - p_{A_6}p_{B_4}$, which will gradually decline to zero (or, if initially negative, will rise to zero).

HISTORY. Although Weinberg (1909) was aware that with linkage, random association would be approached only gradually, the algebraic treatment for two loci was first given by Jennings (1917). The linkage disequilibrium quantity was first used by Robbins (1918), though that name was not given it until the paper of Lewontin and Kojima (1960). Hilda Geiringer (1944, 1948) was the first to prove convergence to random association for multiple loci; a more abstract proof was given by Reiersøl (1962), who used incomprehensible genetic algebras.

WHAT CAUSES LINKAGE DISEQUILIBRIUM. We will see when we discuss migration that genetic admixture from parent populations that have different gene frequencies at more that one locus will cause linkage disequilibrium. In Chapter VIII there is an extensive discussion of the ways in which natural selection at linked loci can produce linkage disequilibrium, especially if the fitnesses at the two loci show epistasis (that is, interact). Later in that chapter there is also a discussion of the formation of random disequilibria as a result of genetic drift at sites that are nearby in the genome. An understanding of the ways LD can arise is essential to anyone wanting to work on population genomics.

I.11 Other Measures of Linkage Disequilibrium

Linkage disequilibrium, usually simply referred to as "LD", is has become more widely computed as population-level studies have become more common in genomics. But using the quantity D to describe the extent of linkage disequilibrium between two alleles at different loci, such as A and B, can be misleading, because there is less possibility of large values of D the nearer the gene frequencies are to 0 or to 1. This has led to two alternative measures that try to rescale it, D' and r^2.

The first of these was invented by Lewontin (1964a). It computes the largest value that D could have, and divides it by the absolute value of that value. If D is positive, it cannot exceed $(1-p_A)p_B$, and also cannot exceed $p_A(1-p_B)$, because past those values one or more of the frequencies of haplotypes Ab or aB would be negative, which is impossible. Likewise, if D is negative, it cannot be more negative than either $-p_A\,p_B$ or $-(1-p_A)(1-p_B)$, for beyond those values one or more of the frequencies of haplotypes

AB or ab would be negative. Lewontin suggested scaling D by dividing by the absolute value of the relevant limit, so that the scaled value could not be above 1 or below -1:

$$D' = \begin{cases} \dfrac{D}{\min\left(p_A(1-p_B),\ (1-p_A)p_B\right)} & \text{if } D \geq 0 \\ \dfrac{D}{\min\left(p_A\, p_B,\ (1-p_A)(1-p_B)\right)} & \text{if } D < 0 \end{cases} \tag{I-57}$$

Although D' is widely used as a descriptive statistic, it has what appears to be a disadvantage. If $p_A = 0.1$ and $p_B = 0.2$, then a value of (say) $D = 0.01$ is to be scaled by dividing by the largest possible positive value of D, which will be 0.08, to give $D' = 0.125$. But if instead we had $D = -0.01$, it should instead be scaled by dividing by 0.02, which gives $D' = -0.5$. The sudden change of the scaling factor as D passes through 0 seems arbitrary and unnatural.

An alternative that does not have this problem is the r measure of Hill and Robertson (1968). In effect, this codes the two alleles at each locus as 0 and 1, and computes the correlation coefficient of these two numbers. The result is

$$r = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}. \tag{I-58}$$

This measure is often used by giving its square, r^2. As r cannot lie outside the interval between -1 and 1, r^2 cannot lie outside of $[0,1]$. This would seem to solve the problem, but alas, it too has a seeming limitation. If gene frequencies at one locus are less extreme than at the other, the value of r may not be able to reach 1 or -1. Thus in our example, if $p_A = 0.1$ and $p_B = 0.2$, it turns out that the value of r cannot be greater than $2/3$ or less than $-1/6$. Only when the gene frequencies of the two alleles are equal at the two loci will r be able to get as small as -1 or as large as 1.

Thus, as useful as these are as descriptive measures, each seems to have distinct limitations. But the problem is that we have not formulated the problem in terms of estimating or testing anything – instead we relied on intuitive feel as to which properties were "natural". When we try to formulate a well-posed statistical problem that these measures solve, we are in fact led away from using any of them, and into the wonderland of coalescent methods, which we visit in Chapter X.

I.12 Estimating Gene Frequencies

MAXIMUM LIKELIHOOD ESTIMATION OF GENE FREQUENCIES.

If we draw a sample of n diploid individuals from a random-mating population, and wish to estimate the gene frequency p_A in the population, there would seem to be several

courses of action possible. Suppose that we sampled 100 individuals, and found 49 AA, 26 Aa, and 25 aa. We could estimate the gene frequency in the population by simply taking the gene frequency in the sample. This gives $p_A = (98 + 26)/200 = 0.62$. But we could also consider that we expect the proportion of AA individuals in the sample to be (on the average) the same as the population genotype frequency p_A^2. So we could take the observed frequency of AA, 0.49, and take its square root to get an estimate of the gene frequency, 0.7. We could also take the square root of the observed frequency (0.25) of aa, which gives an estimate of 0.5 for the frequency of a, and hence 0.5 for the frequency of A. Now we have three different estimates (0.5, 0.62, and 0.7) for the same quantity. All share one justification: as the sample size increases, the observed genotype frequencies in the sample will approach those in the population. Thus all three of these methods will give a gene frequency close to that in the population, if the sample size is large. But which estimate is to be preferred when it is not?

To get an answer to this problem, we must pose the problem as a statistical one, and use a standard statistical approach. There are a variety of these (e.g., minimum variance unbiased estimates, minimum mean square error methods, Bayesian and empirical Bayesian approaches). But one method exceeds the others in general applicability and widespread acceptance by statisticians. This is R. A. Fisher's *method of maximum likelihood*. This is discussed in Box 1.

In this case, the data are the numbers of the genotypes observed in the sample. Suppose that these are n_{AA}, n_{Aa}, n_{aa}. The role of θ is played by the unknown gene frequency p. We need to know how to compute $\text{Prob}(n_{AA}, n_{Aa}, n_{aa} \mid p)$. We have a sample of n individuals, drawn from a population in which the true genotype frequencies are p^2, $2p(1-p)$, $(1-p)^2$. The probability of the observed numbers n_{AA}, n_{Aa}, n_{aa} is the multinomial probability

$$\text{Prob}\ (n_{AA}, n_{Aa}, n_{aa} \mid p) \ = \ \binom{n}{n_{AA}\ n_{Aa}\ n_{aa}} \ (p^2)^{n_{AA}} \ [2p(1-p)]^{n_{Aa}} \ [(1-p)^2]^{n_{aa}}. \quad \text{(I-59)}$$

This can be rewritten as

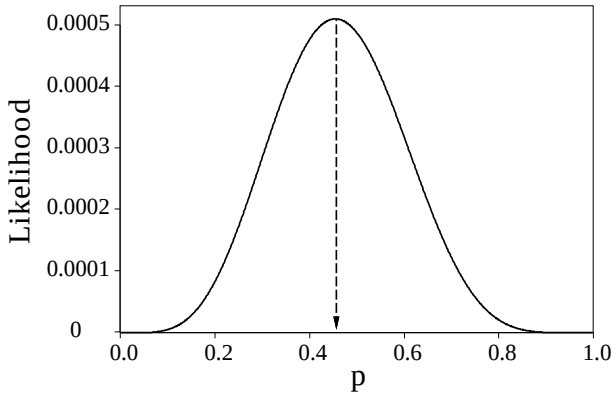
$$\text{Prob}\ (n_{AA}, n_{Aa}, n_{aa} \mid p) \ = \ C\ p^{2n_{AA}+n_{Aa}}\ (1-p)^{n_{Aa}+2n_{aa}}, \quad \text{(I-60)}$$

where C incorporates the constant terms and the factorials which depend on the n's but not on p. We want to vary p to maximize the likelihood. It will turn out to be easier to work in terms of the natural logarithm of the likelihood. Since the logarithm of a quantity increases as the quantity increases, the value of p which maximizes one maximizes the other.

The logarithm of the likelihood is:

$$\ln L \ = \ \ln C \ + \ (2n_{AA} + n_{Aa})\ \ln p \ + \ (n_{Aa} + 2n_{aa})\ \ln(1-p). \quad \text{(I-61)}$$

Suppose that we want to estimate a parameter, θ, and are given some data. If we have a probabilistic model for the generation of the data, we could compute for a given value of θ, the probability $\text{Prob}(\text{Data} \mid \theta)$ that the observed set of data would have arisen. (When the data are values of continuous variables, we use instead the probability density rather than the probability). This is *not* to be confused with $\text{Prob}(\theta \mid \text{Data})$, which would be the probability of a particular value of θ, given the data. We usually do not have enough information to find that. (For more on this distinction, consult a text of mathematical statistics concerning the distinction between Bayesian and maximum likelihood methods).



The likelihood curve for the case of tossing a coin with probability p of Heads, when we toss 11 times and get 5 Heads. The likelihood curve is the product of 5 factors of p and 6 factors of $(1-p)$, so that it is $p^5(1-p)^6$. The maximum likelihood is achieved at $p = 5/11$, which is 0.454545, indicated by the dashed arrow. Note that this is not a distribution: the area under the curve is not 1, as the curve never gets much higher than 0.000511.

The method of maximum likelihood is to vary θ until we find that value which maximizes $\text{Prob}(\text{Data} \mid \theta)$, the probability of the data, given θ. $\text{Prob}(\text{Data} \mid \theta)$ is referred to as the *likelihood* of θ. Considered as a function of the data, it is a *probability*. But for a fixed set of data, as a function of θ, it is called a *likelihood*. The two terms (probability and likelihood) that are barely different in English usage, become distinct and specific in statistical use. The maximum likelihood method has a number of desirable properties. In most well-behaved cases, as the sample size increases, the estimate will approach the true value of θ (a property called *consistency*). For typical cases with a large sample size, the variance of the estimate of θ around the true value is less under the ML method than under any other (this property is called *efficiency*). The estimate is not necessarily *unbiased* (that is, the average estimate of θ on repeated sampling may not be exactly θ), but the amount of bias declines as sample size increases.

Box 1: The method of maximum likelihood

If we plot ln L as a function of p, when it reaches the maximum, the slope of the curve will be zero. Trying to find the value of p at this point, we take the derivative of (I-61) and equate it to zero. The term C does not contain p, so:

$$\frac{d \ln L}{dp} \;=\; \frac{2n_{AA} + n_{Aa}}{p} - \frac{n_{Aa} + 2n_{aa}}{1-p} \;=\; 0. \tag{I-62}$$

The value of p which solves this equation is, after a few straightforward algebraic rearrangements,

$$\begin{aligned} \hat{p} \;&=\; \frac{2n_{AA} + n_{Aa}}{2n_{AA} + 2n_{Aa} + 2n_{aa}} \\ &=\; \frac{2n_{AA} + n_{Aa}}{2n}. \end{aligned} \tag{I-63}$$

The numerator is simply the number of A genes observed in the sample. The denominator is the total number of genes. So the estimate is the observed fraction of A genes. In the example just given, this was 0.62, so that the maximum likelihood method selects one of the three methods as preferable. This selection is not arbitrary: R. A. Fisher showed that maximum likelihood estimates make more efficient use of data than do others. For large amounts of data, they have at least as small a variance as do other estimators.

CONFIDENCE INTERVALS. The maximum likelihood estimate is a point estimate; it gives you a single number, but we really want an interval estimate giving upper and lower bounds on p. If we want to put confidence limits on p, there are several possible approaches:

Using the curvature of the log-likelihood surface. If we can compute the second derivative of the likelihood, and evaluate it at the point $\hat{p}$, there is a well-known formula which estimates the variance of $\hat{p}$, from the second derivative of the likelihood:

$$\mathrm{Var}\,(\hat{p}) \;=\; -1 \Bigg/ \left[\frac{d^2 \ln L}{dp^2}\right]_{p \,=\, \hat{p}} \tag{I-64}$$

The 95% confidence limits on p will be approximately found by taking the standard deviation $\sigma = [\mathrm{Var}\,(\hat{p})]^{1/2}$, with the limits being $\pm 1.96\sigma$. The logic of this formula, derived by Fisher, involves approximating the binomial distribution by a normal distribution. It will be inaccurate when p is near 0 or 1, since then the confidence limits it calculates on p can exceed 0 or 1.

Using the distribution of $\hat{p}$. A second, and simpler approach looks directly at the formula for the estimate $\hat{p}$, and finds its variance from the multinomial distribution (I-59) of n_{AA}, n_{Aa}, and n_{aa}. Here we are helped by a simplification: $\hat{p}$ is simply the fraction of

the $2n$ genes in the sample which are A. If the population is in Hardy-Weinberg proportions (which we assume) each gene sampled independently has probability p of being A. In estimating p we are simply estimating the parameter of a Binomial distribution, based on a sample of $2n$ genes. If we are willing to approximate the binomial distribution by a normal distribution, we can obtain 95% confidence limits from $\hat{p} \pm 1.96\sigma$, where σ is the standard deviation of the underlying binomial distribution. This is obtained from:

$$\sigma^2 = \frac{p(1-p)}{2n}. \tag{I-65}$$

Of course, this can only be calculated once we know the true underlying value of p. But this is precisely what we are trying to estimate! We can use our estimate $\hat{p}$ in (I-65) to get an approximate confidence interval of p. The interval will sometimes exceed 0 or 1. If the observed $\hat{p}$ is zero, we estimate $\sigma = 0$ from (I-65), and find that (apparently) $\hat{p}$ is not an estimate, but is exact! This cannot really be so: we are being betrayed by the inaccuracy of the normal approximation, and by the fact that we are using an estimate $\hat{p}$ rather than the true p in (I-65). An improved approximation is

$$\sin^2 \left[(\sin^{-1} \sqrt{\hat{p}}) \pm 1.95996 \sqrt{\frac{1}{8n}} \right], \tag{I-66}$$

with the quantity in brackets being kept confined to the interval $(0, \pi/4)$. (Note that the angles in (I-66) are expressed in radians rather than degrees). This is still an approximation. For a truly correct confidence interval, we can either make use of published tables of confidence limits in statistical tables (using $2n$ as the sample size) or can use tables of the binomial distribution, as follows. For the upper confidence limit, we find a value of p such that only 2.5% of the binomial distribution will lie at or below the observed sample gene frequency $\hat{p}$. The lower limit will be the value of p such that only 2.5% of the binomial distribution is at or above the observed $\hat{p}$. No approximation is then involved.

The exact confidence interval. A more exact confidence statement can be made using the binomial distribution. Suppose that we can exactly compute tail probabilities for the binomial, using a computer program. Given the observed numbers of A and a alleles sampled, we can calculate the smallest value of p such that the probability of getting less than this number of A alleles in a sample of that size is less than or equal to 2.5%. We can also calculate the largest value of p such that the probability of getting more than the observed number of A alleles is less than or equal to 2.5%. These are the bounds of a confidence interval on p.

For example, if we sample 100 copies at a locus and find that 96 of them are A, these exact confidence limits are 0.914824 and 0.983568. We can contrast this to the normal approximation, which gives limits of 0.93284 and 0.98716. The arc sine approximation

(I-66) comes closer, giving 0.928518 and 0.982661. The exact and arc sine approximations cannot get outside the allowable range $[0, 1]$; the normal approximation can.

This is only one of the ways to construct an exact confidence interval. One other one tries to find the shortest such interval, by allowing a probability of slightly more than 2.5% in one tail and slightly less than 2.5% in the other. We will not go further into that here.

If more than two alleles are involved, the situation is more complex. When all genotypes can be identified, as above, the procedures parallel the above ones. The estimate of each allele is simply its fraction among the $2n$ genes in the sample. The estimated variance of allele A is simply

$$\sigma_A^2 \;=\; \frac{p_A(1-p_A)}{2n}. \tag{I-67}$$

Its covariance with allele A' is

$$\mathrm{Cov}\,(p_A, p_{A'}) \;=\; -p_A p_{A'}/(2n). \tag{I-68}$$

It is also possible to compute joint confidence intervals on the frequencies of two or more alleles, but we will not consider that further here.

GENE COUNTING (EM ALGORITHM). When not all genotypes can be distinguished, we can use a general technique known as *gene-counting*. This can be illustrated by using the ABO blood group alleles as examples. The relationship between genotypes and phenotypes is (if we ignore the two types of A alleles):

Genotypes	Blood Type	Number	Phenotype Frequency
AA, AO	A	n_A	$p_A^2 + 2p_A p_O$
BB, BO	B	n_B	$p_B^2 + 2p_B p_O$
AB	AB	n_{AB}	$2p_A p_B$
OO	O	n_O	p_O^2

(I-69)

If we somehow knew how many of the n_A individuals in our sample of blood type A were AA and how many AO, our estimate of p_A would be the observed frequency of A

$$p_A \;=\; \frac{2n_{AA} + n_{AO} + n_{AB}}{2n}. \tag{I-70}$$

But we do not know n_{AA} and n_{AO} separately: we cannot tell these genotypes apart. If we knew p_A and p_O, then we expect, from the relative Hardy-Weinberg frequencies that on the average $p_A^2/(p_A^2 + 2p_A p_O)$ of all type A individuals are really AA, and the remainder, $2p_A p_O/(p_A^2 + 2p_A p_O)$, will be AO. These are only expectations, and will not

necessarily apply in any given sample. In any case we do not know p_A and p_O. Note that we can remove a factor of p_A from the numerator and denominator of these fractions, which simplify to $p_A/(p_A+2p_O)$ and $2p_O/(p_A+2p_O)$. The gene-counting method takes the seemingly senseless approach of using these expectations, themselves based on the gene frequencies that we do not know and are trying to estimate, to divide the type A individuals into AA and AO according to the above expressions, and doing the same for BB and BO. Having done so, we then pretend that these numbers (such as the number of type A individuals that are inferred to really be AA) are observed numbers, and estimate the gene frequencies by counting of alleles. The estimates are

$$\begin{aligned} p_A &= \left[2\left(\frac{p_A+p_O}{p_A+2p_O}\right)n_A + n_{AB}\right] \Big/ (2n), \\ p_B &= \left[2\left(\frac{p_B+p_O}{p_B+2p_O}\right)n_B + n_{AB}\right] \Big/ (2n), \\ p_O &= \left[\left(\frac{p_O}{p_A+2p_O}\right)n_A + \left(\frac{p_O}{p_B+2p_O}\right)n_B + n_O\right] \Big/ n. \end{aligned} \tag{I-71}$$

These equations have one major problem: we cannot compute p_A, p_B, and p_O (on the left side of the equations) until we know them (to use on the right side)! One way to resolve this difficulty would be to consider (I-71) as a set of equations whose unknowns are p_A, p_B, and p_O, and solve them. An easier (and equivalent) technique is to start with a set of guesses at p_A, p_B, and p_O, then use them on the right side of (I-71) to compute new estimates of p_A, p_B, and p_O. These are then used to compute newer estimates, and so on until the process converges. This is a relatively easy process, which can be carried out on a small computer.

This procedure seems to be merely another exercise in ad-hocery, of equating variables with their expectations. Normally, such techniques are recipes for confusion, uninformed by valid statistical principles. In this case, and in analogous ones, it turns out that the estimates of p_A, p_B, and p_O obtained are actually the maximum likelihood estimates! In fact, the more general gene-counting technique usually has this property. This technique consists of using estimates of the gene frequencies to divide up phenotype classes into their underlying genotypes, according to expected fractions computed using the guesses of the gene frequencies. These reconstructed genotype numbers are then used as if they were observed data to count genes and obtain thereby new estimates of the gene frequencies. The process is then repeated until it converges.

This technique was introduced by C. A. B. Smith (Ceppelini, Siniscalco, and Smith, 1955). For a general treatment see the paper of Smith (1957). Dempster, Laird, and Rubin (1977) have introduced a more general version of gene counting called the "EM Algorithm" which has become widely-used in statistics. The gene counting technique often converges slowly, but is much less vulnerable to bad choices of initial guesses than

are other iterative methods of finding maximum likelihood estimates. It is a good way of starting the search for the maximum likelihood estimate.

I.13 Testing Hypotheses about Frequencies

The preceding section considered the estimation of gene frequencies. The natural statistical counterpart of estimation is testing. Some of the hypotheses we may be most interested in testing include Hardy-Weinberg proportions, linkage equilibrium, and equality of gene frequencies in different populations. This section will briefly cover the first two of these. The third will be covered in a later chapter when we consider the effects of migration.

In testing for departure from Hardy-Weinberg proportions, we have a sample of individuals from a population and have scored their phenotypes. We have a genetic model which generates expected phenotype frequencies from gene frequencies under the assumption of Hardy-Weinberg proportions. The problem reduces to comparing observed and expected frequencies in a sample from a multinomial distribution (such as I-59), where the gene frequencies are not known but must be estimated. Two closely related methods, which should give nearly the same result, are the chi-square test of goodness of fit and the likelihood ratio test. The chi-square test can in fact be shown to be an approximation to the likelihood ratio test.

CHI-SQUARE GOODNESS-OF-FIT TEST. To do a chi-square test of goodness of fit we first estimate gene frequencies, then use them to generate expected numbers of the different observed phenotypes. We then compute the chi-square statistic:

$$\chi^2 = \sum_i \frac{(n_i - N_i)^2}{N_i}, \tag{I-72}$$

where the observed number in class i is n_i, the expected number is N_i, and summation is over all classes i. If the number of classes is k and the number of independent gene frequencies estimated is m, this chi-square statistic should have (to good approximation) a Chi-Square distribution with $k - 1 - m$ degrees of freedom. We can use standard tables of this distribution to test whether the value of χ^2 is too large to be the result of sampling error. In doing so we are, of course, doing a one-tailed test. It is unfortunate that the statistic and the distribution with which we compare it have both come to be known as "chi-square". It is important to distinguish between them. Here the one will be called the "chi-square statistic" and the other the "Chi-Square distribution."

THE LIKELIHOOD RATIO TEST. This proceeds similarly, starting with the estimation of the gene frequencies and the computation of the expected numbers N_i. In principle,

what it computes is the ratio of the likelihood of the sample allowing the expected genotype frequencies to be completely arbitrary (and to be estimated directly from the sample), L_1, and the likelihood L_0 when the expected genotype frequencies are constrained to be in Hardy-Weinberg proportions. The likelihood ratio test, which you will find described in mathematical statistics textbooks, but in all too few introductory statistical "cookbooks", tests whether the likelihood is significantly higher under the hypothesis of no constraint than under the null hypothesis of Hardy-Weinberg proportions. To do it one calculates the the statistic $2\ln(L_1/L_0)$. This should approximately have a Chi-Square distribution, the number of degrees of freedom being the difference between the number of parameters estimated under the alternative and the null hypotheses.

In practice, this turns out to be rather easy to do. If there are k observed classes then there are $k-1$ parameters being estimated under the alternative hypothesis; these are the $k-1$ for the genotype frequencies. We do not have k parameters because they must sum to 1. Under the null hypothesis we have m parameters being estimated. The difference between these is $k-1-m$, which is the same number of degrees of freedom we used when we computed the chi-square statistic (I-72). Twice the log of the likelihood ratio turns out to be simply

$$G \;=\; 2\,\ln(L_1/L_0) \;=\; -2\sum_i n_i\,\ln(n_i/N_i), \tag{I-73}$$

which is just as easy to compute as the chi-square statistic (I-72). We compare its value with the significance levels of the Chi-Square distribution in a one-tailed test. Thus both statistics, the chi-square statistic and the likelihood ratio test statistic, are expected to have approximately the same distribution.

One difficulty that arises is that if expected numbers in some of the classes are small, the approximation starts to break down. The usual rule of thumb is that it cannot be trusted if the expected number in any class is less than 5. This seems to be an overly conservative value; both tests usually do not break down until expected numbers approach 1. If you encounter small expected numbers in any class, you can combine that class (adding up the observed numbers and also the expected numbers) with some other class. This reduces the number of observed classes k.

Here is a sample data set and an example of both tests. Suppose that we had observed genotypes *AA*, *Aa*, and *aa* in a sample of 1000 individuals in numbers 520, 426, and 54. Our best estimate of the gene frequency of A is the observed frequency 0.733. With that gene frequency the expected Hardy-Weinberg frequencies are 0.537, 0.391, and 0.066.

The observed and expected numbers are

Genotype	Observed number	Expected number
AA	520	537.29
Aa	426	391.42
aa	54	71.29
Total :	1000	1000.00

The chi-squared statistic is $\chi^2 = (520 - 537.29)^2/537.29 + (426 - 391.42)^2/391.42 + (54 - 71.29)^2/71.29 = 0.556 + 3.055 + 4.193 = 7.804$. The number of degrees of freedom is $3 - 1 - 1 = 1$. The 95% significance level of the Chi-Square distribution for a one-tailed test with one degree of freedom is 3.841, so that we can reject the null hypothesis of Hardy-Weinberg proportions; the observed excess of heterozygotes is significant.

The likelihood ratio test uses the same observed and expected numbers, computing instead $-2 \times [520 \ln(520/537.29) + 426 \ln(426/391.42) + 54 \ln(54/71.29)] = -[-17.01 + 36.06 - 15.00] = 8.11$. The number of degrees of freedom is again $3 - 1 - 1 = 1$. The one-tailed 95% value 3.841 is again exceeded. The two tests give very similar numbers in this case, and they reach the same conclusion, that the excess of heterozygotes is significant.

TESTING LINKAGE DISEQUILIBRIUM. When we test linkage disequilibrium, there are a number of cases that have to be considered. If we can observe haploid gametes, the test is quite simple. For the two-allele case, we have four observed numbers, n_{AB}, n_{Ab}, n_{aB}, and n_{ab}. We can estimate the gene frequencies of A and a by direct counting, and generate expected values for the numbers of the four gametes. As in the case of a single locus, the data is assumed to be a sample from an infinite population, so that the observed numbers follow a multinomial distribution with some expectations. Computing the four expectations under the null hypothesis of no linkage disequilibrium, we have four observeds and four expecteds, and can compute either the chi-square statistic or the likelihood ratio statistic. The number of degrees of freedom is $4 - 1 - 2 = 1$, since we have estimated two parameters, the gene frequencies.

Alternatively, we could imagine ourselves making a 2×2 table, placing each gamete in a row according to whether or not is has the A allele, and in a column according to whether or not it has the B allele:

	B	b
A	n_{AB}	n_{Ab}
a	n_{aB}	n_{ab}

(I-74)

If we employ a standard chi-square heterogeneity test on this table, it will in fact be exactly the same test as the chi-square test for linkage disequilibrium! If we compute the statistic G (e.g. Sokal and Rohlf, 1969, Chap. 16) instead of χ^2, we will simply be using the likelihood ratio statistic instead of the chi-square statistic.

Often we will not observe the gametes directly, but instead will have to infer their identities from diploid zygotes in which we cannot tell an *AB/ab* double heterozygote from an *Ab/aB*. If we could distinguish these, then we could reconstruct the gametes from which each individual arose. For example, an *AABB* arose from two *AB* gametes, and an *AaBB* from one *AB* and one *aB*. Each sample of n diploid individuals is then exactly equivalent to a sample of $2n$ haploid gametes, and we can test those to see whether there is evidence for $D \neq 0$. If we cannot divide the double heterozygotes into the coupling and repulsion classes we have nine observable phenotypes, which we can regard as being arranged in a 3×3 table:

$$\begin{array}{c|ccc} & BB & Bb & bb \\ \hline AA & n_{AABB} & n_{AABb} & n_{AAbb} \\ Aa & n_{AaBB} & n_{AaBb} & n_{Aabb} \\ aa & n_{aaBB} & n_{aaBb} & n_{aabb} \end{array} \tag{I-75}$$

On seeing this arrangement, it is tempting to test linkage disequilibrium by testing independence of rows and columns in this table. In doing so we would in effect be assuming arbitrary genotype frequencies at both loci, while testing linkage disequilibrium. However, we would be testing more than we intended. For example, if heterozygosity at locus A were not independent of heterozygosity at locus B (for example, if locus A were heterozygous only when locus B was not), the test could be significant.

The matter is complex; there are many possible hypotheses that could be tested with these data. The reader is referred to the papers by Hill (1974) and Weir (1979). A solid grasp of the theory of likelihood ratio tests will be helpful to anyone setting out to test for the presence of linkage disequilibrium.

Exercises

1. Suppose that at a two-allele locus in a random-mating diploid population we find 32% of the individuals to be of the *aa* phenotype. What fraction of the individuals are *Aa* ?

2. In a population where the frequency of *a* is 0.4, what proportion of *aa* individuals have neither parent *aa*? One parent *aa*? Both parents *aa* ? Assume that both parents and offspring were produced by random mating.

3. An obtuse researcher is investigating a locus with two alleles in a random-mating population, with no selection, migration, etc. (i.e., Hardy-Weinberg proportions are expected). The researcher finds in a population 44% heterozygotes and 56% homozygotes, but forgets to distinguish between the two kinds of homozygote. What can the researcher say about the gene frequency of allele *A*?

4. Suppose there are two populations that have genotype frequencies

	AA	Aa	aa
Pop. 1	0.64	0.32	0.04
Pop. 2	0.09	0.42	0.49

If a researcher draws a very large sample, thinking it is coming from a single population, but it is actually composed of individuals two-thirds of whom came from population 1, and one-third from population 2,

(i) Are the two original populations each in Hardy-Weinberg proportions?

(ii) If these individuals are simply collected together, but have no time to interbreed, what will the genotype frequencies in the sample expected to be?

(iii) What will the gene frequencies be in that sample?

(iv) If we mistakenly assume that the sample is from a single random-mating population, and want to see whether it is in Hardy-Weinberg proportions, if we use the sample gene frequency to examine that, what proportion of heterozygotes will we expect to see?

5. A locus has three alleles, B', B, and b. B' is completely dominant to B, and both of these are completely dominant to b. What are the frequencies of the three alleles in a random-mating population which has these phenotype frequencies: 50% B'-, 30% B-, and 20% bb ?

6. We have a sample of 190 individuals from a diploid population and have genotyoped them at a locus which has three alleles, A_1, A_2, and A_3. We find the following numbers in our sample:

A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3
20	35	47	62	14	12

(i) What are the gene frequencies of these three alleles in the sample?

(ii) At these gene frequencies, what genotype frequencies will be expected to result from random mating?

(iii) Are there more or fewer heterozygotes in this sample than expected?

7. In a sample of 200 individuals from a population which is expected to be at Hardy-Weinberg equilibrium for a locus with 3 alleles, the numbers of the 6 possible genotypes found are

Genotype	Number
A_1A_1	76
A_1A_2	54
A_1A_3	33
A_2A_2	18
A_2A_3	16
A_3A_3	3

Calculate the gene frequencies of the three alleles, and what numbers of the six genotypes you would have expected from those frequencies. Without doing a formal statistical test, can you see any apparent discrepancies between the numbers in the sample and Hardy-Weinberg proportions? Where?

8. Suppose that at a sex-linked locus, the frequency of *a* hemizygotes among males is 0.2 and the frequency of *aa* homozygotes among females is 0.1. Assuming that random mating with different gene frequencies in the two sexes produced the current generation, figure out what the gene frequencies were in those two sexes. What will the genotype frequencies be in the next generation if it too is produced by random mating?

9. Suppose that a sex-linked locus has two alleles, *A* and *a*. We look at a population and find among females:

AA	*Aa*	*aa*
0.95	0.04	0.01

while among males:

A	*a*
0.94	0.06

Note that we have not said that the gene frequencies in the parents of these individuals were the same in both sexes – they might be different.

(i) Does the population appear to be the result of at least one generation of random mating? *(Hint: consider whether random mating with different gene frequencies in the two sexes can produce exactly those frequencies)*.

(ii) If it reproduces for one more generation by random mating of these females with these males, what genotype frequencies do we expect to see in the offspring?

10. Suppose that we have two populations, each at linkage equilibrium for two unlinked loci. Suppose that the gene frequencies are:

	Allele			
Population	*A*	*a*	*B*	*b*
1	0.6	0.4	0.3	0.7
2	0.3	0.7	0.5	0.5

Suppose we produce an F_1 population by crossing the two populations (mating males from one population with females from the other), and an F_2 by mating at random among the offspring of all of those F_1 individuals. What will the linkage disequilibrium value D_{AB} be in gametes produced by the F_1 individuals and in gametes produced by the F_2 individuals?

11. Suppose that in a population produced by random mating, we have two alleles at each of two loci, with $p_A = p_B = 0.5$, and $D_{AB} = 0.2$. Let half of the individuals be females and half males. The recombination fraction between the loci is 0.3 in females and 0.1 in males. What will D_{AB} be in the offspring generation in terms of D_{AB} in the current generation? What will be the frequency of genotype *AA BB* in the offspring generation?

12. Suppose that we have a large diploid population at linkage equilibrium at two loci, each with two alleles. At the first locus the frequency of allele *A* is 0.1, and of allele *a* 0.9. At the second locus the frequencies of alleles *B* and *b* are respectively 0.2 and 0.8. A strongly favorable mutation at a nearby locus occurs in an *AB* haplotype, and natural selection brings its descendants rapidly up to a haplotype frequency of 0.5. Suppose that there has been no recombination among these haplotypes during this rise.

    (i) What will the haplotype frequencies in the population be at that point? *(Hint: simply make an equally-weighted average of the haplotype frequencies in a population that is all AB, and the original population which had gene frequencies 0.1 and 0.2, and had no linkage disequilibrium).*

    (ii) In this population, what is the value of D for alleles *A* and *B*?

    (iii) What is the value of D' ?

    (iv) Now suppose that this "selective sweep" continues until we reach the point where the favorable mutation is now at frequency 0.9, still with no recombination having occurred. What are the haplotype frequencies? *(Hint: now it's a weighted average with weights 0.9 : 0.1). What is D ? Has it increased? What is D' ? Has it increased?*

13. When we sample 100 individuals from a random mating population, we observe 63 *AA*, 27 *Aa*, and 10 *aa*. Put 95% confidence limits on the frequency of *A*. What have you had to assume?

14. Among 100 individuals, we observe 10 *aa*'s. Assuming random mating, how do you place 95% confidence limits on the frequency of *A*?

15. We sample 200 individuals from a diploid population and find 89 *AA*, 57 *Aa*, and 54 *aa* individuals. Test the hypothesis that this is a sample from a population that

is Hardy-Weinberg proportions.

16. We sample 100 individuals from a diploid population and find the following numbers of genotypes at two two-allele loci:

	BB	Bb	bb
AA	0	25	0
Aa	25	0	25
aa	0	25	0

Use a 3×3 heterogeneity chi-square test to test whether the genotypes at these two loci are distributed independently of each other. See if you can also make an estimate of the linkage disequilibrium D_{AB} between these loci. Is there a discrepancy between these two conclusions? Why or why not?

17. In a computer simulation of mutation and genetic drift in a stretch of 100,000 bases of DNA, with rates of mutation and recombination typical of human populations, I found 304 SNP loci (single nucleotide polymorphisms). Here are 10 sampled diploid genotypes, each a row, each scored at the same set of four of the SNP loci. The alleles are called 0 and 1 – these are actually two different bases. The genotypes are unphased – we do not know which parent each copy came from:

11 00 00 01
11 01 01 01
01 00 00 01
11 01 00 01
01 00 00 00
11 00 00 01
01 01 00 01
00 00 01 01
01 00 00 00
00 00 00 00

(i) Resolve these 10 diploid genotypes into haplotypes as well as you can. (First, resolve those that have one or fewer heterozygous loci, as you can always do that. Then try to resolve the others so as to come up with more of those same haplotypes, to the extent you can). Do this "by eyeball" and list the haplotypes you get, and how often each one occurs in this sample which must be of 20 haplotypes.

(ii) Compute the gene frequencies at the 4 SNP loci.

(iii) Make a table of values of D between each pair of loci. (It should be lower-triangular). It will have 6 entries.

(iv) Also make tables of the relative linkage disequilibrium measures D' and r.

(v) The loci are in this order left-to-right on the genome. Do you see any sign (looking at the D' values for extreme LD) that there are "haplotype blocks"?

Complements/Problems

1. Imagine a multiple-allele locus with gene frequencies p_1, p_2, ..., p_n. In terms of these quantities, after random mating, what fraction of copies of allele A_i occur in heterozygotes? What is the overall fraction of all copies that occur in heterozygotes? (Be sure to consider what is supposed to be in the denominator when answering each of these questions).

2. Suppose that there are n equally-frequent alleles. In terms of n, what will be the proportion of individuals in the population that are homozygotes? Heterozygotes?

3. Suppose we draw a "panel" of 20 diploid individuals from a population and proceed to discover SNPs (single nucleotide polymorphisms) by looking for sites that vary and have more than one copy of the rare allele. We do that to exclude sites that might simply have a sequencing error. Suppose that about one site in every 1000 has a true SNP, and the rate of error in sequencing is 0.002 per copy sequenced. Further, suppose that the true SNPs all happen to have their minor allele gene frequency be p, and sequencing error makes a copy change to the other one of the two alleles.

   (i) For a site which has only one allele in the population, so that all the variation in the sample is due to sequencing error, what is the probability, for a single copy, that this copy is scored as having the other allele?

   (ii) Using that, in the sample of 20 diploid individuals, what is the probability of seeing more than one copy of the less common allele? *Hint: compute the probability of seeing none, and also the probability of seeing one.*

   (iii) At a site which has a true SNP, whose rarer allele has population frequency p, when sequencing error is taken into account, what is the probability that one copy is scored as having that rarer allele?

   (iv) At that true SNP site, what is that probability of seeing more than one copy of that rarer allele?

   (v) Taking into account that it is 999 times as likely that any site is not a true SNP than that it is a true SNP site, what fraction of the putative SNPs we find will be true ones?

(vi) What would that fraction be if we included as putative SNPs also sites where the minor allele only occurred once?

(v) For $p = 0.001$ and for $p = 0.01$ what is the fraction computed in each of these two cases?

Hint for all these: If an allele has expected frequency q, in a sample of 20 diploids the chance that it shows up k times is

$$\binom{40}{k} q^k (1-q)^{40-k}.$$

Note that you will need to compute the probability of 0 or 1 occurrences of an allele (or an error), and also the probability of 39 or 40 copies of that allele, and then subtract all these from 1, in order to get the probability that more than there are more than one occurrence. The combinatorial factor for each of these terms will just be 1 or 40.

4. Suppose that we have a large diploid population which is in Hardy-Weinberg proportions for two alleles A and a, with gene frequency p of A. We draw a sample consisting of n individuals. We estimate the gene frequency from the gene frequency in that sample. From it, we are going to predict the proportion of heterozygotes in that sample.

   - Suppose that the sample has $n = 1$, that is, it consists only of a single individual. Are we as likely to see an excess of heterozygotes as a deficit of heterozygotes? *Hint: work out all possibilities – there are only 3 of them.*
   - Now consider samples of size $n = 2$. There are only 9 possibilities, which are not equally probable. Calculate the excess of heterozygotes over the prediction from the sample gene frequency for each case, and make a weighted average of this, using the sample probabilities computed from the true gene frequency (which the researcher would not know). Is the average excess of heterozygotes zero?
   - In a sample of n diploid individuals, in which of the $2n$ copies, k of them are allele A, what will be the predicted proportion of Aa individuals, computed from the observed gene frequency using Hardy-Weinberg proportions, as a function of n and k? If we draw two copies without replacement from that sample, what is the probability that the resulting individual is Aa? Are these the same? Does this predict the results you found for the two previous cases?

   (This problem is based on a paper published in 1949 by Howard Levene).

5. Suppose that we have a locus with two alleles, linked to the sex-determining locus in a haploid organism with random mating. (The sex-determining locus has two alleles, a female-determining allele and a male-determining allele, and each generation females and males mate, form a diploid, and then the sex-determining alleles segregate out in the resulting haploids). The recombination fraction between our locus and the sex locus is r. If the initial gene frequency of A in one sex is p_1 and in the other p_2,

   (a) What will be the value of p_1 and p_2 in the next generation?

   (b) What will the value of the average of p_1 and p_2 be? How does it change from generation to generation?

   (c) What will the value of the difference between p_1 and p_2 be? How will it change from generation to generation?

   (d) From these, work out what will be the values of p_1 and p_2, t generations from now.

   (e) What will be the ultimate values of p_1 and p_2 in terms of their initial values.

6. Suppose we have a haploid population with two alleles, A and a, whose frequencies are p and $1 - p$. If a fraction s of the gametes mate only with others having the same allele, the remaining $1 - s$ combining at random

   (i) What will be the genotype frequencies in the diploid stage?

   (ii) What will be the gene frequencies in the next haploid stage?

7. If we have two populations, with a three-allele locus, find two sets of gene frequencies such that if we cross males from one population randomly with females of the other, there will be *fewer* A_1A_2 heterozygotes in the first-generation cross than in a simple mixture of the two populations.

8. In a population with overlapping generations, in which the males are initially in Hardy-Weinberg proportions at gene frequency p_m, and females are in Hardy-Weinberg proportions at gene frequency p_f,

   (i) What are the equations for change in p_m and p_f?

   (ii) What will be the departure from Hardy-Weinberg proportions in the whole population at time t?

9. If we have a two-allele locus and two populations, one at Hardy-Weinberg proportions with gene frequency p_1 and the other at Hardy-Weinberg proportions at gene frequency p_2,

   What is the frequency of Aa heterozygotes

(i) in a simple 50:50 mixture of individuals from population 1 and population 2?

(ii) in a offspring of a cross between males of population 1 and females of population 2?

(iii) in offspring produced by random mating of those offspring?

What is the algebraic relationship between these three quantities (i.e., can one of them be predicted given only the other two and not the gene frequencies)?

10. Suppose that a chromosome has been duplicated so that where there was once one locus, there are now two unlinked loci, each with two alleles A and a. We cannot distinguish which locus contributed an A or an a to a genotype. The two loci are each diploid and they are in linkage equilibrium with each other. At the first locus the gene frequency of A is p_1, and at the second locus the gene frequency of A is p_2. In terms of those two quantities, what are the expected frequencies of genotypes with 4, 3, 2, 1 and 0 A's? Note that we cannot tell the difference between, for example, $AAaa$ and $AaAa$, so that they both contribute to the genotypes that have 2 A's.

11. Suppose that we have partial sex-linkage in a diploid organism, with a sex locus with two alleles (X and Y), and a nearby locus with two alleles (A and a) which has recombination fraction r with the sex locus (gene copies at this locus are present on both the X and the Y), what will be the equation of change of frequencies from one generation to the next? *(Hint: you will need to follow two gene frequencies).* Do you need to assume random mating for this calculation?

12. With a recessive sex-linked gene (in an ordinary XX-XY system) with gene frequency p, what fraction of affected individuals (homozygous AA females or hemizygous AY males) are female? (Assume random mating, and that the population has reached an equilibrium genetic state and has 1:1 sex ratio).

13. In an autosomal locus with an allele a whose frequency is p, what fraction of all extant copies of the a allele in a random mating population are located in aa homozygotes?

14. For an autosomal locus in a random-mating population, where aa individuals are affected and the gene frequency is p, what fraction of aa individuals have both of their parents affected? One affected? Neither affected?

15. Suppose that we have two unlinked autosomal loci with two alleles each, so that their recombination fraction is $r = 0.5$. Suppose that the initial population consists entirely of AB/ab double heterozygotes and thereafter reproduces by random mating.

What will D_{AB} be in the initial population? In its offspring? How does this compare with what we expect from the formula $D_{AB}(t) = (1-r)^t D_{AB}(0)$? Why the discrepancy?

16. With three loci, each with two alleles, see if you can find a set of gamete frequencies (there are 8 possible gametes) which has $D_{AB} = 0$, $D_{BC} = 0$, and $D_{AC} = 0$, but where the gametes are not in linkage equilibrium, in that, for example, $P_{ABC} \neq p_A p_B p_C$. Can we completely determine the gamete frequencies by specifying p_A, p_B, p_C, D_{AB}, D_{BC}, and D_{AC}? Show by examples whether or not this is possible.

17. Suppose that we have two autosomal loci, with no gene or gamete frequency differences between the sexes, but with different recombination fractions r_f and r_m, in the two sexes. Assuming random mating, how will D_{AB} change with t?

18. *(Harder)* What are the equations for the decline of linkage disequilibrium at a sex-linked locus, if we have the same initial gamete frequencies in all X-bearing gametes, whether male or female (assume that there is no crossing-over between X and Y chromosomes, and that the genes are not present on the Y chromosome) ?

19. If we have two populations each at linkage equilibrium with gene frequencies (respectively) of p_A, p_B, and $p_{A'}$, $p_{B'}$, what will D_{AB} be in the gametes arising from cross between males from one population and females from another, in terms of the recombination fraction?

20. Take a population of gametes segregating for two alleles at each of two loci. For each gamete compute two numbers, x and y, where

    $x = 0$ if the gamete is a, 1 if it is A.
    $y = 0$ if the gamete is b, 1 if it is B.

    In terms of p_A, p_B, and D_{AB}, what is the variance (over gametes) of x? of y? What is the covariance of x and y? The correlation between x and y?

21. If we have two loci, one with n_1 alleles and the other with n_2 alleles, there are $n_1 n_2$ different pairs of alleles, and there will be a linkage disequilibrium parameter for each such pair. In view of the fact that the frequencies of the haplotypes containing each allele must sum to the frequency of the allele, and in view of the fact that the frequencies of all haplotypes must sum to 1, how many of these $n_1 n_2$ linkage disequilibrium parameters can really be varied independently. *(Hint: make sure your logic works for the two-allele case).*

22. In a population with N haploid individuals, suppose that at a locus there is one copy of allele A, the rest of the copies being of allele a. At another locus, there is

only one copy of of allele B, the rest being allele b. If allele B is placed on one of the haploid genomes at random,

- What is the probability that the linkage disequilibrium value D is positive? What is its value in that case?
- What is the probability that the linkage disequilibrium value D is negative? What is its value in that case?
- Can D be zero?
- What is the average value of D (weighting the positive and negative values by their probabilities).
- What is the variance of D?

23. For three loci, each with two alleles possible, can we find a set of four (out of the eight possible) haplotypes, such that when these four are equally frequent all three pairwise values of D (namely D_{AB}, D_{AC} and D_{BC}) are zero. Compare this to the case where all eight haplotypes are equally frequent. Do these differ in gene frequencies? In values of D? In view of the answers to these, is it possible to write a formula for a haplotype frequency in terms of the gene frequencies at the three loci, together with the values of D for the three pairs of loci?

24. In a case with 10 two-allele loci, there are 1024 different possible haplotypes. If

    - the frequencies of all of these sum to 1, and
    - for each of the ten loci, we can add up the frequencies of all haplotypes that contain one allele (at that locus) and it will sum to the gene frequency, and
    - For each of the 45 pairs of two loci, we can add up the frequencies of all of the haplotypes that have the one given allele at each of the two loci (such as A and g), and this will be predicted by an equation analogous to (I-56).

    How many quantities does that give us to predict the 1024 haplotype frequencies? By counting "degrees of freedom", how many more quantities (in fact, higher-order linkage disequilibrium parameters) will we need to predict them all? How does this compare to the sum of the number of triples of 10 loci, plus the number of quadruples, plus the number of quintuples, and so on up to 10-tuples?

25. Suppose we take a sample from a random-mating population, where we can detect a recessive phenotype, so that there are two phenotypes, which we can call A- and aa. We find n_{aa} individuals with the aa phenotype, and n_{A-} individuals of the A- phenotype. What is the maximum likelihood estimate of p_a? What are the equations we would use to estimate p_a by the gene counting technique? Do they lead to the maximum likelihood estimate?

26. Show that the likelihood ratio test of the hypothesis that the genotypes at a two-allele locus are in Hardy-Weinberg proportions is in fact identical to the G statistic of goodness of fit (I-73).

27. How would we construct a likelihood ratio test of Hardy-Weinberg proportions at the ABO blood group locus? How many degrees of freedom would the test have?

28. Given the numbers of the nine genotypes in a sample from a diploid population with two two-allele loci, and assuming that the two loci are unlinked, so that the the recombination fraction is $r = 0.5$, what are the frequencies of the four gamete types among the haploid gametes produced by this sample? Compute D_{AB} for these gametes in terms of the nine genotype numbers. If the genotypes were sampled from a population produced by random mating, with an unknown true value of D_{AB}, what is the expectation of this estimate of D_{AB} in terms of the true unknown value? If we try to estimate D_{AB} in the population of gametes that gave rise to our diploid individuals by doubling the D_{AB} in the gametic output of our sample, will we be making a biased or an unbiased estimate?

Chapter II

NATURAL SELECTION

II.1 Introduction

Natural selection can be viewed either narrowly or broadly. Narrowly conceived, it is simply one class of violations of the assumptions of the Hardy-Weinberg Laws, namely the cases in which viability or fertility depends on genotype. Broadly conceived, it is the primary force which causes evolution to be adaptive, the creative and progressive element in the evolutionary process. A comprehensive theory of evolution, one which does not yet exist, would integrate ecological processes (which determine the range of environments and the fitnesses of phenotypes), developmental processes (which determine the effect of genotype on phenotype), and population genetics (which tells us the changes in genetic composition of a population when the fitnesses of the genotypes are known). Lacking the other elements of this future theory, we concentrate here on the population genetics.

We first examine the mathematics of gene frequency change, to see how much change of gene frequency is caused by a given pattern of differential viability and fertility, to see what the pattern of gene frequency change through time will be, and to see how selection acting on diploid genotypes affects gene frequencies. There are nontrivial evolutionary questions which are addressed by this part of the chapter. It is not obvious in advance how to quantitate fitness, nor how effective small differences in fitness can be. Diploidy is a major complication whose effects are also not obvious, and much of the effort in this part of the chapter is devoted to explaining its effects. The second part of the chapter discusses situations in which selection with constant fitnesses causes an equilibrium genetic composition to be maintained. The third part concerns the effect of natural selection on the mean fitness of individuals in the population. It was a central tenet of Darwin's thinking that natural selection had an average tendency to increase adaptedness, but it is a nontrivial matter to investigate whether or not it actually does so in simple model situations.

All of this development takes place under the assumption that fitnesses are constants which do not change either as a function of time or asn a function of the composition of the population. The final part of the chapter examines frequency-dependent selection (including cases of fitnesses based on social interaction among genotypes), temporally-varying fitnesses, and fitnesses depending on population density. These cause interesting alterations to the rules concerning the adaptive effects of natural selection.

Some important situations are deferred until later in this book. The effects of spatial variation in fitness are covered in the chapter on migration, the effect of mutation on fitness in the chapter on mutation, and the effects of linkage on selection are given a separate chapter. The more realistic situation in which the phenotype under selection is controlled by many loci is covered in a very rough and approximate fashion in the final chapter.

II.2 Selection in Asexuals - Discrete Generations

Let us consider a population consisting of two asexual strains, A and a. The numbers of the two strains in generation t are $N_A(t)$ and $N_a(t)$. We are interested in computing the numbers of the two strains in the next generation. For this we need to know how many offspring individuals of each genotype will have. If an individual of strain A has probability v_A of surviving to adulthood, and if each survivor has an average of f_A offspring, then the expected number of offspring left by a newborn A individual is $v_A f_A$. This quantity, a composite of viability (v_A) and fertility (f_A), is the *absolute fitness* (or Darwinian fitness) of genotype A. We call it W_A, and its counterpart for a is W_a. Throughout this chapter, we will assume that the numbers of individuals $N_A(t)$ and $N_a(t)$, and their counterparts in other cases, are sufficiently large that we can ignore the random fluctuations which will arise from the randomness of birth and death. Thus we will have a completely deterministic mathematical system, in which $N_A(t)$ newborn individuals of genotype A are assumed to leave *exactly* $W_A N_A(t)$ offspring.

In the next generation, the numbers of newborn A and a individuals are given by

$$\begin{aligned} N_A(t+1) &= W_A\, N_A(t) \\ N_a(t+1) &= W_a\, N_a(t). \end{aligned} \tag{II-1}$$

We will also be interested in the relative frequencies of the two genotypes. These are defined in straightforward fashion to be

$$\begin{aligned} p_A^{(t)} &= \frac{N_A(t)}{N_A(t)+N_a(t)} \\ \text{and} & \\ p_a^{(t)} &= \frac{N_a(t)}{N_A(t)+N_a(t)}\,, \end{aligned} \tag{II-2}$$

the fractions of all individuals who are A (or a). Taking the definitions II-2 in generation $t+1$, and substituting from II-1 for the quantities $N_A(t+1)$ and $N_a(t+1)$, we find that

$$p_A^{(t+1)} \;=\; \frac{N_A(t+1)}{N_A(t+1)+N_a(t+1)} \;=\; \frac{N_A(t)\,W_A}{N_A(t)\,W_A+N_a(t)\,W_a}. \tag{II-3}$$

The total number of individuals in the population is $N(t)=N_A(t)+N_a(t)$. Dividing each of the N's in the fraction on the right-hand side of (II-3) by $N(t)$, and using (II-2), we find that

$$p_A^{(t+1)} \;=\; \frac{p_A^{(t)}\,W_A}{p_A^{(t)}\,W_A+p_a^{(t)}\,W_a}\,, \tag{II-4}$$

and correspondingly

$$p_a^{(t+1)} \;=\; 1-p_A^{(t+1)} \;=\; \frac{p_a^{(t)}\,W_a}{p_A^{(t)}\,W_A+p_a^{(t)}\,W_a}. \tag{II-5}$$

The denominator of both of these fractions is $p_A^{(t)}W_A+p_a^{(t)}W_a$. This is the weighted average

$$\frac{N_A(t)\,W_A+N_a(t)\,W_a}{N_A(t)+N_a(t)} \;=\; \frac{N(t+1)}{N(t)}, \tag{II-6}$$

the average absolute fitness of all newborn individuals at the beginning of generation t. Thus this mean absolute fitness, which we denote $\widehat{W}$, also tells us the factor by which the population increases from this generation to the next. $\bar{W}$ is dependent on t through the numbers $N_A(t)$ and $N_a(t)$, but from here on we will not indicate this explicitly.

Equation (II-4) can therefore be rewritten as

$$p_A^{(t+1)} \;=\; \frac{p_A^{(t)}\,W_A}{\bar{W}}. \tag{II-7}$$

and there is a similar equation for $p_a^{(t+1)}$,

$$p_a^{(t+1)} \;=\; \frac{p_a^{(t)}\,W_a}{\bar{W}}. \tag{II-8}$$

Taking the ratio of these equations (or alternatively taking the ratio of the two equations in (II-1),

$$\frac{p_A^{(t+1)}}{p_a^{(t+1)}} \;=\; \frac{W_A}{W_a}\frac{p_A^{(t)}}{p_a^{(t)}} \tag{II-9}$$

The ratio of p_A to p_a gives us the same information as does p_A alone, for we can obtain the one from the other.

$$p_A^{(t)} = \frac{p_A^{(t)}}{p_A^{(t)} + p_a^{(t)}} = \frac{\left(p_A^{(t)}/p_a^{(t)}\right)}{\left(p_A^{(t)}/p_a^{(t)}\right) + 1} \tag{II-10}$$

We will frequently pass back and forth between equations which compute frequency, such as (II-7), and those which compute frequency ratios.

FITNESS AND POPULATION DENSITY. We can make use of the frequency ratio equation (II-9) to reveal an important property of natural selection. If we are interested in population composition, as reflected by $p_A^{(t)}$ and $p_a^{(t)}$ (or by their ratio), and if we are not interested in population size or population density, as reflected by $N_A(t)$ and $N_a(t)$, then (II-9) shows that the changes in composition of the population depend on W_A and W_a only through their ratio. This is also true in equations (II-7) and (II-8), though it is less transparent from the way they are written. This dependence on W_A/W_a can have dramatic implications.

The effects of selection in changing population composition depend on the relative sizes of the fitnesses of different genotypes, not on their absolute sizes. For instance, if $W_A = 1.01$ and $W_a = 1$, then equation (II-9) is

$$\frac{p_A^{(t+1)}}{p_a^{(t+1)}} = \frac{1.01}{1}\frac{p_A^{(t)}}{p_a^{(t)}} \tag{II-11}$$

so that the ratio of A to a is multiplied by 1.01 every generation. In this example, the A population is growing very slowly (by 1% per generation), while the a population is remaining constant in size. Now consider another case: $W_A = 101$ and $W_a = 100$. In that case both A and a populations are growing very rapidly, with the A population growing 1% more each generation than the a population. Equation (II-9) becomes

$$\frac{p_A^{(t+1)}}{p_a^{(t+1)}} = \frac{101}{100}\frac{p_A^{(t)}}{p_a^{(t)}} \tag{II-12}$$

which is exactly the same as (II-11), the ratio of A to a being multiplied by 1.01 per generation. Of course, in the latter case, the numbers of A and a individuals around will be growing very rapidly, and if we concentrate on population numbers the two cases will look very different. But as long as we are interested only in population composition rather than population size, we will find that the genetic composition of the two populations undergoes the same changes in both cases. In general, if we take W_A and W_a and multiply them by the same number, we will obtain a case which will still show the same

sequence of changes in $p_A^{(t)}$ and $p_a^{(t)}$. This should be clear from (II-9), where W_A and W_a enter only through their ratio.

The same property can be seen in (II-4), though less easily. If we (say) double both W_A and W_a, we will double $\widehat{W}$, doubling both terms in the denominator. So both the numerators (involving W_A) and the denominator (involving W_A and W_a) are doubled, and the fraction is left unchanged. There is then no effect on $p_A^{(t+1)}$. We have the same property whether we consider the effects of selection on the frequency p_A or on the ratio of p_A to p_a, so all is as it should be.

Relative and abolute fitness. Since we will only be interested in the ratios of the absolute fitnesses, we can pick some particular genotype as our standard, and measure the ratios of the fitnesses of other genotypes to the fitness of the standard. These ratios we call the *relative fitnesses* of the genotypes. Hereafter when the word "fitness" is used, it will mean relative fitness. We denote the relative fitnesses by w_A and w_a. If we take a to be the standard, then $w_a = 1$. If $W_A = 101$ and $W_a = 100$, then $w_A = 1.01$ and $w_a = 1$. In both numerical examples given above, $w_A = 1.01$.

The biologically relevant aspect of relative fitnesses is that an extra source of mortality or fertility may change absolute fitnesses, but may leave relative fitnesses unaltered, provided that it falls on all genotypes equally. If it does not change the relative fitnesses, it will not change the ratio of W_A to W_a. Thus it will have an effect on population size without in any way affecting population composition. This is particularly important when we consider population size regulation. Suppose that the population's size is regulated naturally by a drop in fertility under crowded conditions. When population density (or equivalently, size) was low we might have:

Genotype	A	a
Viability as larva	0.5	0.4
Fertility as adult	6	6
Absolute fitness (viability $\times$ fertility)	3	2.4
Relative fitness	1.25	1

while when the population has reached a high density and is crowded:

Genotype	A	a
Viability as larva	0.5	0.4
Fertility as adult	2.2	2.2
Absolute fitness (viability $\times$ fertility)	1.1	0.88
Relative fitness	1.25	1

The drop in fertility affected both genotypes equally, and the relative fitnesses are unaffected by the population density regulation. This is enormously convenient for the algebraic treatment of the consequences of natural selection: even though the density regulation means that absolute fitnesses will not remain constant through time, in this

case the relative fitnesses will, so that we can simply ignore all considerations of population density in our treatment.

The same phenomenon appears if we were instead to have genotypes which differed in fertility, and had viability be affected by population density:

	Low Density		High Density	
Genotype	A	a	A	a
Viability	0.5	0.5	0.1	0.1
Fertility	6	5	6	5
Absolute Fitness	3.0	2.5	0.6	0.5
Relative Fitness	1.2	1	1.2	1

Things are not quite so simple when the population density affects the same life stage as the selection. Suppose that at low population density the situation is

Genotype	A	a
Viability	0.5	0.4
Fertility	6	6
Absolute Fitness	3.0	2.4
Relative Fitness	1.25	1

If high population density imposes an extra mortality on the population, we must specify how the viabilities (0.5 and 0.4) are affected. A natural, but hardly inevitable, assumption is to specify that after the individuals have passed the life stage at which they are at risk of dying as a consequence of their genotype, a completely independent source of mortality occurs. This mortality does not depend on genotype, so that the organisms have a probability (say) 0.25 of surviving this mortality irrespective of genotype. There is an overall chance 0.5×0.25 that an A survives to adulthood, and 0.4×0.25 for an a. The resulting fitness table is then at high population density:

Genotype	A	a
Viability	0.125	0.1
Fertility	6	6
Absolute Fitness	0.75	0.6
Relative Fitness	1.25	1

You may wish to draw up the corresponding tables for the case where both genotype and population density affect fertility. The conclusion is similar: the relative fitnesses remain constant provided that population density multiplies all the fertilities by the same factor. In general, if population density multiplies both viabilities by a factor V, and both fertilities by a factor F, then the absolute fitnesses will be multiplied by VF and the relative fitnesses will be unaffected.

Multiplicative combination of forces affecting fertility is not as reasonable a null hypothesis as is multiplication of viabilities. It is not obvious why having a high population

density must decrease the number of offspring per adult by 20% instead of decreasing it by a fixed amount (say 2). This latter will not keep relative fitnesses constant, but will reduce the relative fitness of genotypes whose relative fitness is low to start with. Even with viabilities, physiologically independent forces may act nonmultiplicatively. If the survivors of a genotype which has viability 0.4 are weaker and less vigorous than survivors of a genotype whose viability is 0.5, then the population-density-dependent effects would be expected to take a greater toll of the weaker genotype. Such models are not by any means irrelevant. For the moment, we assume that relative fitnesses remain constant, but it should be kept in mind that this may not be the case at all.

SELECTION COEFFICIENTS. Rather than representing relative fitnesses as $w_A : w_a$, it is frequently convenient to write them as $1 + s : 1$. The quantity s will be zero when there is no natural selection, and is referred to as the *selection coefficient* favoring A. It can take any value from -1 to ∞. Equation (II-9) becomes

$$\frac{p_A^{(t+1)}}{p_a^{(t+1)}} \;=\; (1+s)\,\frac{p_A^{(t)}}{p_a^{(t)}} \tag{II-13}$$

Alternatively, we could take the relative fitnesses $w_A : w_a$ to be $1 : 1 - s$, so that the standard genotype is A. Then s would be the selection coefficient against a. It can take any value from $-\infty$ to 1. The exact meaning of s will thus depend on which genotype is taken as the standard, and whether the selection coefficient measures selection for or against the genotype. Note particularly that the ratio $1 + s : 1$ is *not* the same as $1 : 1 - s$ unless s is zero. A selection coefficient of 0.01 in favor of A is not exactly the same as a selection coefficient of 0.01 against a. In fact, it *is* equivalent to a selection coefficient of 0.00990099... against a. When s is small this is not a great difference, but it is well to be aware of it.

Equation (II-7) can also be rewritten in terms of s. The relative fitnesses w_A and w_a can be used in place of the absolute fitnesses W_A and W_a, as we have seen. In the present case w_A and w_a are $1 + s$ and 1, so that we can replace $\bar{w}$ by

$$\bar{w} \;=\; (1+s)\,p_A^{(t)} + p_a^{(t)}. \tag{II-14}$$

If we replace $p_A^{(t)}$ by p_t for simplicity of notation, and note that $p_a^{(t)} = 1 - p_t$, we have

$$\bar{w} \;=\; 1 + s\,p_t, \tag{II-15}$$

so that

$$p_{t+1} \;=\; \frac{p_t\,(1+s)}{1+s\,p_t}. \tag{II-16}$$

The relative fitnesses of A and a are in the ratio $1 + s : 1$. We might think that this will be the same as taking A as the standard genotype, and using fitnesses $1 : 1 - s$.

After all, isn't having a selection coefficient s in favor of A the same thing as having a selection coefficient of s against a? No, it isn't. The quantities $(1 + s)/1$ and $1/(1 - s)$ are not the same. For a dramatic example, try $s = 1$.

CHANGE OF GENOTYPE (OR GENE) FREQUENCY. Using relative fitnesses in place of absolute fitnesses in equations (II-7) and (II-8), we can easily write formulas for the change in the frequency of A. Using primes for the next generation, (II-7) and (II-8) are

$$p'_A = \frac{p_A\, w_A}{\bar{w}} \tag{II-17}$$

and

$$p'_a = \frac{p_a\, w_a}{\bar{w}} \tag{II-18}$$

From (II-17), if we drop the subscript A from p_A, and replace p_a by $1 - p$, we can calculate Δp, the change in p:

$$\begin{aligned} \Delta p &= p' - p \\ &= p\, w_A/\bar{w} - p \\ &= p\,(w_A - \bar{w})/\bar{w} \end{aligned} \tag{II-19}$$

Alternatively, since $\bar{w} = p\, w_A + (1 - p)\, w_a$,

$$\begin{aligned} \Delta p &= p\,[w_A - p\, w_A - (1 - p)\, w_a]/\bar{w} \\ &= p\,[(1 - p)\, w_A - (1 - p)\, w_a] \\ &= p(1 - p)\,(w_A - w_a)/\bar{w} \end{aligned} \tag{II-20}$$

We will use close analogues of these equations for Δp when we consider diploids.

HAPLOIDS. We have been discussing cases of asexual inheritance. The results are no different if we consider sexual haploids. Suppose that selection precedes meiosis. Selection will change the proportions of genotypes (and in this case, of genes) from $p_A^{(t)}$ and $p_a^{(t)}$ to $p_A^{(t+1)}$ and $p_a^{(t+1)}$ in exactly the same way as in asexuals, following equations (II-7) or (II-9) exactly. The subsequent fertilization, followed by a meiosis, will not alter the gene frequencies of A and a, provided that both genotypes are equally able to participate in meiosis and provided there is no violation of the Mendelian rules. In this case, a sexual haploid behaves exactly like an asexual. Only when we consider more than one locus will we find asexuals and haploids behaving differently.

HISTORY. The basic equations of this section are due to J. B. S. Haldane (1924), in the first of his classic series of papers which outlined the deterministic theory of gene frequency changes due to selection. Despite early numerical computations of Castle (1903), analysis of cases of complete selection by Warren (1917), Punnett (1917), Norton (1915, 1928), and Robbins (1922), as well as analysis of gene frequency equilibria under selection by Fisher (1922), Haldane's work forms the basis of modern selection theory. His series of papers covered a great many cases of interest. His work is summarized in the Appendix to his 1932 book, *The Causes of Evolution*, which has been reprinted in paperback by two different publishers.

II.3 Selection in Asexuals - Continuous Reproduction

When generations overlap, we must take a somewhat different route to obtain the effects of selection. The most extreme model of continuous reproduction is the one we used in the previous chapter, one in which the probabilities of birth and death do not depend at all on age. This model is not only the polar opposite of the discrete-generations model, but also is mathematically simple.

Suppose that in a very short time interval of length Δt, the probability that a particular individual of genotype A dies is $d_A \Delta t$, and the probability that it gives birth to a single offspring is $b_A \Delta t$. The corresponding probabilities for genotype a are $d_a \Delta t$ and $b_a \Delta t$. The number of individuals of genotype A at time t is $n_A(t)$, and the number of individuals of genotype a is $n_a(t)$. For genotype A we obtain the number of individuals expected to exist at the end of the time interval by adding the births and subtracting the deaths:

$$n_A(t+\Delta t) \;=\; n_A(t) \;+\; n_A(t)\, b_A\, \Delta t \;-\; n_A(t)\, d_A\, \Delta t \tag{II-21}$$

with a similar equation for a. Equation (II-21) can be rearranged to give

$$\frac{n_A(t+\Delta t) - n_A(t)}{\Delta t} \;=\; n_A(t)\;(b_A - d_A)\,. \tag{II-22}$$

This equation is approximate rather than exact, because in (II-21) we have ignored the effects of births early in the interval Δt on the number of deaths later in that interval, and of the number of deaths early in the interval on the number of births later in it. However this error is proportional in size to $(\Delta t)^2$, and as we shrink Δt these terms disappear faster than the terms in Δt and the equation becomes more exact. As $\Delta t \to 0$, the left side of equation (II-22) becomes a derivative and we have (dropping the argument (t) of n_A to simplify the appearance of the expressions):

$$\frac{dn_A}{dt} \;=\; n_A\,(b_A - d_A)\,. \tag{II-23}$$

If we let $r_A = b_A - d_A$ this is the familiar exponential growth equation

$$\frac{dn_A}{dt} \;=\; r_A\, n_A. \tag{II-24}$$

There is an exactly similar equation for a, with $r_a = b_a - d_a$. The quantity r for a genotype is sometimes called the "Malthusian parameter" of a population consisting only of that genotype. We are interested in following the frequency of genotype A, $p = n_A/(n_A + n_a)$. We can simply differentiate p with respect to time:

$$\begin{aligned} \frac{dp}{dt} \;&=\; \frac{d}{dt}\left[\frac{n_A}{n_A+n_a}\right] \;=\; \left[(n_A+n_a)\frac{dn_A}{dt} - n_A\frac{d(n_A+n_a)}{dt}\right] \Big/ (n_A+n_a)^2 \\ &=\; \frac{1}{(n_A+n_a)}\frac{dn_A}{dt} \;-\; \frac{n_A}{(n_A+n_a)^2}\left[\frac{dn_A}{dt} + \frac{dn_a}{dt}\right]. \end{aligned} \tag{II-25}$$

We can substitute into this equation (II-24) and the analogous equation for a, and get

$$\frac{dp}{dt} \;=\; \frac{n_A}{(n_A+n_a)}\, r_a \;-\; \frac{n_A}{(n_A+n_a)}\left[\frac{n_A\, r_A + n_a\, r_a}{n_A+n_a}\right], \tag{II-26}$$

which can be rewritten as

$$\begin{aligned} \frac{dp}{dt} \;&=\; p\, r_A \;-\; p\,(p\, r_A \;+\; (1-p)\, r_a) \\ &=\; p\,(r_A \;-\; \bar{r}), \end{aligned} \tag{II-27}$$

where $\bar{r} = p\, r_A + (1-p)\, r_a$ is the average growth rate of the whole population. We could alternatively rewrite (II-27), by collecting terms differently, as

$$\frac{dp}{dt} \;=\; (r_A - r_a)\; p(1-p). \tag{II-28}$$

Note that equation (II-27) shows a similar structure to equation (II-19 in the discrete-generations case, with r playing the role of w and without any denominator like $\bar{w}$ being present. Equation (II-28) is analogous to equation (II-20) in the same way.

These continuous-time equations for the change of p will show an independence of population density effects similar to that invoked in the previous section, but relative fitnesses must be defined differently. In equation (II-28) we can see that the dependence of p on the birth and death rates b_A, b_a, d_A, and d_a is entirely through the quantity

$$r_A - r_a \;=\; b_A - b_a - d_A + d_a. \tag{II-29}$$

If density effects act by adding the same amount to d_A and to d_a, and/or by subtracting equal amounts from b_A and b_a, they will affect the growth rate of the population

without altering the value of $r_A - r_a$. Thus the counterpart to ratios of fitnesses in this model is differences of the intrinsic rates of increase r_A and r_a. The counterpart to standardizing $w_a = 1$ is to set $r_a = 0$ and to set r_A so that the difference $r_A - r_a$ is the correct value. If two genotypes have growth rates 2.7 and 2.5, then the standardization makes their relative growth rates 0.2 and 0. A population density effect which subtracted (say) 2.6 from both growth rates would leave the relative growth rates, and the formulas for genotype frequency change, unaltered.

The analysis in this section has depended heavily on the assumption that the number of births expected (say $b_A n_A \Delta t$) is exactly the number seen. As in the case of discrete-generations model, this implicitly assumes that n_A and n_a are such large numbers that random fluctuations of births and deaths are averaged out. If they are not, then the phenomenon of genetic drift occurs. The analysis of genetic drift is far more difficult, and will be taken up in a later chapter. For the time being, we continue to treat our models as deterministic.

II.4 Selection in Diploids

The results for asexual or haploid models are inherently simple, because each genotype's offspring are of the same genotype (if mutation is ignored). In an outcrossing diploid population matters are strikingly different. An AA parent may have either AA or Aa offspring, and even in the simplest models the relative numbers of these two types of offspring will depend on the composition of the rest of the population. Nevertheless the complexities introduced by diploidy are generally manageable. We first examine the model with discrete generations.

VIABILITIES AND FERTILITIES. In this section, we obtain formulas for the change of gene frequency in simple diploid selection models. To do so, we make a simple model of the life history of the organism. We assume a discrete-generations model and a single locus with two alleles. All of the standard Hardy-Weinberg assumptions apply, with two exceptions. The viabilities of individuals are assumed to depend on their genotype, and also their fertilities depend on their genotype. With respect to fertilities a very particular assumption is being made. Suppose that in a particular generation AA has a fertility of 2, and Aa a fertility of 1. It is reasonable to suppose from these numbers that an $AA \times Aa$ mating has on average twice as many offspring as an $Aa \times Aa$ offspring. Since we will be considering an effectively unisexual population, it seems reasonable to suppose that $Aa \times AA$ matings have the same expected number of offspring as $AA \times Aa$ matings. But what about $AA \times AA$? Does the presence of two AA parents lead to two, three, or four times the number of offspring? In order to have the mathematics come out simply, we will obtain the fertility of each mating from the products of the fertilities of the genotypes of the two individuals. This requires that, in the present case, $AA \times$

AA have a fertility four times that of *Aa* × *Aa.* Various models involving this and other ways fertilities could combine have been examined by Bodmer (1965), who was first to call attention to this assumption of standard diploid selection models.

If we start a generation with a population of N newly formed zygotes in Hardy-Weinberg proportions, we will have Np^2 individuals of genotype AA. If the viability of individuals of genotype AA is v_{AA} and their fertility is f_{AA}, then the contribution of AA individuals to the gamete pool which forms the next generation will be $Np^2 v_{AA} f_{AA}$. We continue to assume that N is sufficiently large that stochastic effects average out, so that we can use a deterministic treatment. In letting the contribution of AA to the gamete pool be $Np^2 v_{AA} f_{AA}$, and in assuming random mating, we are specifying that when gametes combine at random the number of resulting zygotes which come from two AA parents is proportional to $(Np^2 v_{AA} f_{AA})^2$. Since there are $Np^2 v_{AA}$ adults of genotype AA and these mate at random, the number of $AA \times AA$ matings which occur will be proportional to the product $(Np^2 v_{AA})(Np^2 v_{AA})$. Each of these matings must then have $f_{AA} f_{AA}$ offspring. Thus our assumption that f is a factor which affects a genotype's contribution to a gamete pool is precisely equivalent to a product rule for the fertilities of matings. The product rule is also what allows us to have Hardy-Weinberg proportions among the newborns, and we will assume that.

THE BASIC TWO-ALLELE SELECTION FORMULA. If we define the absolute fitness of genotype AA as $W_{AA} = \frac{1}{2} v_{AA} f_{AA}$ (half because a gamete contributed to the gamete pool is only half an offspring), then since the numbers of gametes coming from AA, Aa, and aa parents will be (respectively) $Np^2 v_{AA} f_{AA}$, $N2p(1-p) v_{Aa} f_{Aa}$, and $N(1-p)^2 v_{aa} f_{aa}$, these can also be written as $\frac{1}{2} Np^2 W_{AA}$, $\frac{1}{2} N2p(1-p) W_{Aa}$, and $\frac{1}{2} N(1-p)^2 W_{aa}$. The A gametes will be coming from AA parents, plus half those from Aa parents. Thus (dropping the factors of $\frac{1}{2}$ from every term) the gene frequency of A among these gametes which form the next generation will be:

$$p' = \frac{Np^2 W_{AA} + Np(1-p) W_{Aa}}{Np^2 W_{AA} + 2Np(1-p) W_{Aa} + N(1-p)^2 W_{aa}} \qquad \text{(II-30)}$$

Note that every term in both the numerator and denominator of (II-30) has an N and a W in it. The Ns are all the same, so that we can cancel out the Ns top and bottom. If we divide W_{AA}, W_{Aa}, and W_{aa} by the same number, we will leave the fraction unchanged. This means we can put relative fitnesses in place of the absolute fitnesses. Doing this:

$$p' = \frac{p^2 w_{AA} + p(1-p) w_{Aa}}{p^2 w_{AA} + 2p(1-p) w_{Aa} + (1-p)^2 w_{aa}} \qquad \text{(II-31)}$$

The denominator is the average relative fitness of a randomly chosen individual. We will call it $\bar{w}$. This equation gives the frequency of A in the gametes which make up the next generation. Since those gametes in effect combine at random (as a result of

random mating), the resulting newly formed zygotes will again be in Hardy-Weinberg proportions at the new gene frequency p'. The full Hardy-Weinberg law will not apply, in that the new gene frequency p' will not necessarily be the same as the old, but the newly fertilized zygotes in each generation are in Hardy-Weinberg proportions, except possibly for the initial population which may not have been produced by this process.

Other forms of the equation. Three alternate forms of (II-31) will be of use. The first is the parallel to (II-9):

$$\frac{p'}{1-p'} \;=\; \frac{p}{1-p} \times \frac{p\,w_{AA} + (1-p)\,w_{Aa}}{p\,w_{Aa} + (1-p)\,w_{aa}} \tag{II-32}$$

which can be obtained from (II-31) and the corresponding expression for $1 - p'$. The second form factors p out of the numerator of (II-31) to obtain:

$$p' \;=\; \frac{p\,(pw_{AA} + (1-p)w_{Aa})}{\bar{w}} \;=\; \frac{p\,\bar{w}_A}{\bar{w}} \tag{II-33}$$

where $\bar{w}_A$ is the mean relative fitness of those individuals which carry A genes, weighted by the number of A genes they carry.

That this is so is seen by computing that in a population of size N, there are $2Np^2$ A genes contained in AA individuals, and $2Np(1-p)$ contained in Aa heterozygotes. It follows that a fraction p of all A genes are in AA's, and $(1-p)$ in Aa's. The denominator of (II-33) is the mean relative fitness of the population, so (II-33) shows p being multiplied by a factor which is the ratio of the mean fitness of individuals carrying A to the mean fitness of all individuals in the population.

The third form of (II-31) expresses the change in gene frequency, Δp, as a function of p. Subtracting p from both sides of (II-31):

$$\begin{aligned}
\Delta p &= p' - p \\
&= \Big(p^2 w_{AA} + p(1-p)w_{Aa} \;-\; p\,[p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa}]\,\Big) \\
&\quad /\;\Big(p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa}\Big) \\
&= \Big((p^2 - p^3)w_{AA} \;+\; p(1-p)(1-2p)w_{Aa} - \;p(1-p)^2 w_{aa}\Big)\,/\bar{w} \\
&= \Big(p^2(1-p)w_{AA} \;+\; p(1-p)((1-p)-p)w_{Aa} \;-\; p(1-p)^2 w_{aa}\Big)\,/\bar{w} \\
&= p(1-p)\;[\;p(w_{AA} - w_{Aa}) \;+\; (1-p)(w_{Aa} - w_{aa})\;]/\bar{w} \\
&= p(1-p)(\bar{w}_A - \bar{w}_a)/\bar{w}.
\end{aligned} \tag{II-34}$$

Alternatively you can do it more simply as

$$\begin{aligned}\Delta p &= p' - p \\ &= \frac{p^2\, w_{AA} \;+\; p(1-p)\, w_{Aa} \;-\; p\, \bar{w}}{\bar{w}} \\ &= \frac{p\,(p\, w_{AA} \;+\; (1-p)\, w_{Aa}) - p\, \bar{w}}{\bar{w}} \\ &= \frac{p\,(\bar{w}_A - \bar{w})}{\bar{w}}\,. \end{aligned} \tag{II-35}$$

Note the close analogies between the diploid and the asexual (or haploid) cases. Equation (II-35) is the analogue of (II-19), equation (II-34) of (II-20), and (II-33) of (II-17). Equation (II-32) can be rewritten as

$$\frac{p'}{1-p'} \;=\; \frac{p}{1-p}\,\frac{\bar{w}_A}{\bar{w}_a} \tag{II-36}$$

in which form it is closely analogous to the relative fitness version of equation (II-9). In each case the analogy is the same: replacing w_A by $\bar{w}_A$ and w_a by $\bar{w}_a$ converts asexual or haploid equations into diploid equations. But note that $\bar{w}_A$ and $\bar{w}_a$ are not constant – they change with gene frequency. Thus we cannot simply take their ratio at the start and keep multiplying by that ratio, so that we raise it to the power of the number of generations.

We are now in a position to examine some special cases of importance:

MULTIPLICATIVE (GEOMETRIC) FITNESSES. Suppose that the fitnesses are:

AA	Aa	aa
$(1+s)^2$	$1+s$	1

In this case when we alter a genotype by replacing one a by an A, we multiply the fitness by $1+s$. Then

$$\begin{aligned}\bar{w}_A &= (1+s)^2\, p \;+\; (1+s)\,(1-p) \\ &= (1+s)\,[\,p(1+s) \;+\; 1-p\,] \\ &= (1+s)\,[1+s\,p], \qquad \text{(II-37)} \\ \bar{w}_a &= 1 \;+\; sp, \qquad \text{(II-38)}\end{aligned}$$

and

$$\begin{aligned}\bar{w} &= p^2(1+s)^2 + 2p(1-p)(1+s) + (1-p)^2 \\ &= [p(1+s)+1-p]^2 \qquad \text{(II-39)} \\ &= (1+sp)^2.\end{aligned}$$

The equations for gene frequencies in the next generation become

$$\begin{aligned} p' &= \frac{p(1+s)(1+sp)}{(1+sp)^2} \\ &= \frac{p(1+s)}{1+sp}, \end{aligned} \tag{II-40}$$

$$\begin{aligned} \frac{p'}{1-p'} &= \frac{p}{1-p}\,\frac{(1+s)(1+sp)}{1+sp} \\ &= \frac{p}{1-p}\,(1+s), \end{aligned} \tag{II-41}$$

and

$$\begin{aligned} \Delta p &= \frac{p(1-p)\,[(1+s)(1+sp)-(1+sp)]}{(1+sp)^2} \\ &= \frac{s\,p(1-p)}{1+sp}. \end{aligned} \tag{II-42}$$

A comparison of these equations with the asexual case will show that (II-41) is precisely the same as (II-13). This is the particular utility of the multiplicative case: it is the counterpart to the asexual case. In both cases replacement of an a gene by an A gene multiplies fitness by $1+s$, and in both cases the change in gene frequency is the same, provided we are willing to consider cases with equal values of the selection coefficient s.

As in the haploid case, a selection coefficient of s in favor of A is not exactly the same as a selection coefficient of s against a. So switching alleles while replacing s by $-s$ does not result in exactly the same changes of gene frequency.

ADDITIVE FITNESSES. Many people have a dogmatic belief that additivity is always simpler than multiplicativity. When fitnesses are additive:

AA	Aa	aa
$1+2s$	$1+s$	1

the heterozygote fitness is the arithmetic mean of the fitnesses of the two homozygotes (in the multiplicative case it was the geometric mean). Now

$$\begin{aligned} \bar{w}_A &= p(1+2s)+(1-p)(1+s) \\ &= 1+s+sp, \end{aligned} \tag{II-43}$$

$$\begin{aligned} \bar{w}_a &= p(1+s) + (1-p) \\ &= 1 + sp, \end{aligned} \tag{II-44}$$

and

$$\bar{w} = 1 + 2sp, \tag{II-45}$$

so that

$$p' = \frac{p\,(1+s+sp)}{1+2sp}. \tag{II-46}$$

and

$$\Delta p = \frac{s\,p(1-p)}{1+2sp}. \tag{II-47}$$

This last equation has a relatively simple numerator and denominator, but unlike (II-42) it is not identical to the haploid case. If s is taken to be small, both additive and multiplicative cases will behave similarly, as we will see later in this chapter. For the moment we need only note that the numerators of the first line of equation (II-42) and of (II-47) are the same, and the denominators $1 + 2sp + s^2p^2$ and $1 + 2sp$ are nearly the same if s is small. But the two cases are not identical.

A RECESSIVE GENE. If the A allele is recessive, so that fitnesses are:

AA	Aa	aa
$1+s$	1	1

then

$$\bar{w}_A = p(1+s) + (1-p) \times 1 \tag{II-48}$$

while $\bar{w}_a$ is simply 1, so that, collecting terms,

$$\bar{w}_A = 1 + sp, \tag{II-49}$$

$$\bar{w}_a = 1, \tag{II-50}$$

and

$$\bar{w} = 1 + sp^2. \tag{II-51}$$

The formulas for change of gene frequency are

$$p' = \frac{p\,(1+sp)}{1+s\,p^2}, \tag{II-52}$$

$$\Delta p = \frac{s\,p^2\,(1-p)}{1+s\,p^2}, \tag{II-53}$$

and

$$\frac{p'}{1-p'} \;=\; (1+sp)\,\frac{p}{1-p}. \tag{II-54}$$

One can see from these formulas, especially the latter two, that selection will be relatively weak if the recessive allele is rare. This follows from the p^2 term in (II-53), and (II-54) shows directly that the effective selection coefficient of allele A is not s but sp. To see this, compare (II-54) with (II-13).

A DOMINANT GENE. When the A allele is dominant, so that fitnesses are:

$$\begin{array}{ccc} AA & Aa & aa \\ 1+s & 1+s & 1 \end{array}$$

then

$$\bar{w}_A \;=\; 1+s, \tag{II-55}$$

$$\bar{w}_a \;=\; 1+sp \tag{II-56}$$

and

$$\bar{w} \;=\; 1+2sp(1-p)+sp^2, \tag{II-57}$$

so that

$$p' \;=\; \frac{p(1+s)}{1+2sp(1-p)+sp^2} \tag{II-58}$$

and

$$\Delta p \;=\; \frac{sp(1-p)^2}{1+2sp(1-p)+sp^2}. \tag{II-59}$$

The counterpart to (II-54) is:

$$\frac{p'}{1-p'} \;=\; \frac{(1+s)}{(1+sp)}\frac{p}{1-p}. \tag{II-60}$$

Again the effective selection coefficient depends on p. When p is small the selection coefficient for A is s, but when p is nearly 1 there will be hardly any selection.

OVERDOMINANCE AND UNDERDOMINANCE. Two cases of particular interest will be those in which the fitness of the heterozygote exceeds that of either homozygote, and in which the fitness of the heterozygote is lower than that of either homozygote. A particularly convenient parameterization of the fitnesses is:

$$AA \quad Aa \quad aa$$

$$1-s \quad 1 \quad 1-t$$

For the case of overdominance s and t are both taken to be positive, so that the heterozygote fitness is highest. For underdominance, s and t are taken to be negative.

In this case,

$$\bar{w}_A = 1 - s\,p, \tag{II-61}$$

$$\bar{w}_a = 1 - t\,(1-p), \tag{II-62}$$

and

$$\bar{w} = 1 - s\,p^2 - t\,(1-p)^2, \tag{II-63}$$

so that, from equation (II-33) and from equation (II-34),

$$p' = \frac{p(1-s\,p)}{1 - s\,p^2 - t\,(1-p)^2}, \tag{II-64}$$

$$\Delta p = \frac{p(1-p)\,[\,t-(s+t)p\,]}{1-s\,p^2-t\,(1-p)^2}, \tag{II-65}$$

and

$$\frac{p'}{1-p'} = \left(\frac{1-s\,p}{1-t\,(1-p)}\right)\frac{p}{1-p}. \tag{II-66}$$

In later sections of this chapter we will return to each of these cases to develop the implications of all of the formulas just given. First, however, we investigate whether there is an analogous model in which gene and genotype frequencies change in continuous time.

II.5 Diploid models with continuous time

We saw a continuous-time model for asexual or haploid organisms. Are there such models for diploid organisms? There are, but one has to be very careful in defining and using them. There are serious potential problems with getting out of Hardy-Weinberg proportions.

A simple continuous-time model has adult genotype frequencies that are possibly not in Hardy-Weinberg proportions. In each small increment of time of length Δt, a fraction Δt of the population is replaced by offspring. The members of the population produce

a cloud of gametes, and those combine at random, to produce diploid individuals in Hardy-Weinberg proportions. Those immediately are subject to viability selection, with individuals of some genotypes surviving better than others. The survivors are then added to the population of adults.

In that model, the gene frequency in the cloud of gametes at time t is $p(t)$, the gene frequency in the adult genotypes. The genotype frequencies in the diploids formed by the cloud of gametes are simply the Hardy-Weinberg proportions implied by the gene frequency p. The gene frequency p^* among the survivors is then given by the discrete-generations formula, such as

$$p^*_A \;=\; p_A \, \bar{w}_A \;/\; \bar{w}. \tag{II-67}$$

The resulting population of adults after a time Δt is a mixture of a fraction Δt of individuals who have gene frequency p^* and $1 - \Delta t$ who have gene frequency p. That mixture is the weighted average

$$p(t + \Delta t) \;=\; (\Delta t)\, p^*(t) \;+\; (1 - \Delta t)\, p(t). \tag{II-68}$$

rearranging,

$$p(t + \Delta t) - p(t) \;=\; \Delta t \, (p^*(t) - p(t)) \tag{II-69}$$

so that, dividing by Δt, in the limit of small Δt, the left-hand side becomes the derivative of $p(t)$ so that

$$\frac{dp(t)}{dt} \;=\; p^*(t) - p(t) \tag{II-70}$$

This will be a function of $p(t)$ as in equations (II-34) or (II-35). The formula for Δp for a discrete generations model becomes the derivative of p for this continuous-generations model.

LIMITATIONS OF THE CONTINUOUS DIPLOID MODEL.

It all seems very straightforward, but it really isn't. First, note that the population of adults is now a mixture of populations, each of which is itself possibly not in Hardy-Weinberg proportions, and we have seen that when we mix populations, the result is usually out of Hardy-Weinberg proportions. So we cannot rely on the adults being in Hardy-Weinberg proportions.

We have also invoked viability selection, but how could we also take differential fertilities into account? That would require us to assign fertilities to the different diploid genotypes in the adults, and calculate the gene frequency in the cloud of gametes by weighting each adult genotype by its genotype frequency, times its fertility. But we don't know those genotype frequencies. A similar problem arises if we allow different adult genotypes to have different rates of mortality, instead of having all death occur in the newborn individuals formed from the cloud of gametes. There is also the issue of whether the weights in the mixture would depend, not only on Δt, but on the mean fertilities and mean death rates.

Even if we don't allow fertilities and viabilities to depend on individuals' ages, we could only keep track of all this by writing equations for the genotype frequencies, not just the gene frequencies. A conventional discrete-generations model has the great advantage that there are Hardy-Weinberg proportions at the start of each generation. The continuous-time model does not have this advantage. We will not develop this model further. Charlesworth (1980) has written a particularly well-thought-out monograph on the complications that arise in age-dependent selection, which is an even worse case. For now, we go back to the case of discrete generations.

II.6 Rates of Change of Gene Frequency

When the relative fitnesses of the genotypes do not change from generation to generation, we can use the formulas for change in gene frequency to examine the speed to gene frequency change. Among the questions which can be answered this way is: how effective will weak selection be?

ASEXUALS AND HAPLOIDS. Gene frequency change through time is easiest to follow in the asexual (or haploid) case. Here time will always be measured in generations. Equation (II-13) shows us immediately that

$$\frac{p_A^{(t)}}{p_a^{(t)}} = (1+s)^t \frac{p_A^{(0)}}{p_a^{(0)}}, \tag{II-71}$$

since the ratio of gene frequencies is multiplied by the same amount $(1+s)$ in each generation. We can take this equation and solve it for t, given the value of s and the initial and final gene frequencies. We obtain, taking natural logarithms of both sides in (II-71):

$$t = \left[\ln\left(\frac{p_A^{(t)}}{p_a^{(t)}}\right) - \ln\left(\frac{p_A^{(0)}}{p_a^{(0)}}\right)\right] / \ln(1+s). \tag{II-72}$$

(As I mentioned before, in this book I will use ln rather than $\log_e$ for the natural logarithm).

This allows us to calculate how many generations it will take for a given gene frequency change. For example, if a population starts at gene frequency 0.01 for A and ends at 0.99, with $s = 0.01$, then we can substitute into (II-72), keeping in mind that $p_a = 1 - p_A$

$$\begin{aligned} t &= \left[\ln\left(\frac{0.99}{0.01}\right) - \ln\left(\frac{0.01}{0.99}\right)\right] \Big/ \ln(1.01) \\ &= [\ln 99 - \ln(1/99)] / \ln(1.01) \\ &= 923.6115 \text{ generations} \end{aligned} \tag{II-73}$$

Table 2.1: Time required to change gene frequency of A from 0.01 to 0.99 when the relative fitness of A is $1 + s$.

s	time (generations)
1.0	13.26
0.5	22.67
0.2	50.41
0.1	96.42
0.05	188.36
0.02	464.09
0.01	923.61
0.001	9194.83

We do not get a whole number of generations in this case, which simply means that the gene frequency p_A will be below 0.99 after 923 generations, but above 0.99 after 924 generations.

An interesting comparison is obtained by doubling the selection coefficient s to 0.02. Then the same gene frequency change (from 0.01 to 0.99) requires 464.09 generations, a bit more than half the time required before. Table 2.1 shows this calculation for a variety of selection coefficients. Roughly speaking, the time required for a given change of gene frequency is inversely proportional to the selection coefficient. This proportionality is particularly accurate when the selection coefficients are small. The comparison of $s = 0.01$ with $s = 0.001$ shows that with a small selection coefficient one-tenth the size, it will take about ten times as long to make a given change of gene frequency. In fact, if we note that $(1.01)^2 = 1.0201$, Equation (II-71) shows that two generations of change at $s = 0.01$ will cause the same shift of gene frequency as one generation at $s = 0.0201$. This reflects both the fact that the proportionality between time required for a change and $1/s$ is not exact, and that when s is small it is nearly exact. This reflects an evolutionary principle of some significance: in a totally deterministic system (an infinitely large population under constant selection), a very small selection coefficient will still be effective in causing gene frequency changes. If s is reduced by a factor of (say) 1,000, the same gene frequency changes will still occur, but will take about 1,000 times as long.

Another important kind of information we can get from this calculation is about the time course of gene frequency changes, the amount of time necessary to change through different gene frequency ranges. We can obtain the gene frequency in any given generation by noting that $p_a^{(t)} = 1 - p_A^{(t)}$, substituting this into (II-71), and solving it for

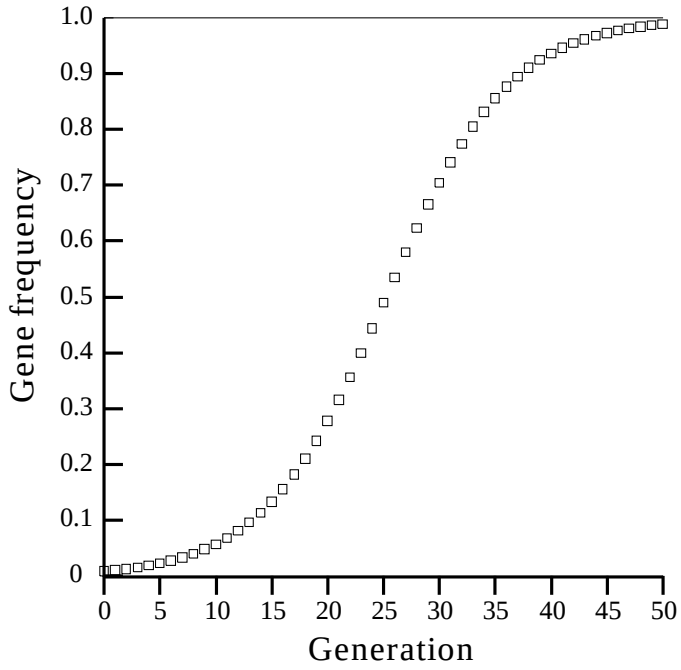


Figure 2.1: Course of gene frequency change in haploid selection with initial gene frequency $p_0 = 0.01$ and relative fitness 1.2 of the A genotype.

$p_A^{(t)}$. We find that

$$p_A^{(t)} = \frac{p_A^{(0)}(1+s)^t}{p_A^{(0)}(1+s)^t + p_a^{(0)}} \tag{II-74}$$

which gives the full course of gene frequency change when $p_A^{(0)} = 0.01$ and $s = 0.25$. Note that when it is plotted (in Figure 2.1) the curve is symmetrical about the point where it passes through $p_A = 0.5$

The same phenomenon can be seen using Equation (II-72) to compute the time needed to change the gene frequency through various ranges. Table 2.2 demonstrates this: it takes exactly as much time to change the gene frequency of A from p to 0.5 as it does to change it from 0.5 to $1 - p$.

This table also illustrates another feature of gene frequency changes: it takes far longer for natural selection to change the gene frequency by a given amount when the gene frequency is extreme than it does when the gene frequency is intermediate. This is the counterpart of the observation that the curve in Figure 2.1 rises slowly at first, then rapidly through the intermediate gene frequencies, then slowly again when gene frequencies are extreme.

ASEXUALS AND HAPLOIDS - CONTINUOUS GENERATIONS. We can do the same sort of calculations when generations overlap. In the model of continuous reproduction used in section II.4, things are particularly simple. Equation (II-28) can be solved for p as a function of t (and vice versa) in the following way: dividing both sides of (II-28) by $p(1-p)$ and multiplying both by dt, which is illegal but not immoral,

$$\frac{dp}{p(1-p)} \;=\; (r_A - r_a)\, dt. \tag{II-75}$$

Noting that $1/[p(1-p)] = 1/p + 1/(1-p)$, we can integrate both sides to obtain

$$\ln p \;-\; \ln(1-p) \;=\; (r_A - r_a)\, t + C, \tag{II-76}$$

where C is the constant of integration which depends on the initial conditions. Initially, $t = 0$ and $p = p_0$, so that

$$\ln p_0 - \ln(1-p_0) \;=\; C \tag{II-77}$$

from which

$$\ln p_t - \ln(1-p_t) \;=\; (r_A - r_a)\, t + \ln p_0 - \ln(1-p_0). \tag{II-78}$$

We can obtain t as a function of p_0 and p_t by rearranging this to give

$$t \;=\; \left[\ln\left(\frac{p_t}{1-p_t}\right) - \ln\left(\frac{p_0}{1-p_0}\right)\right] \Big/ (r_A - r_a) \tag{II-79}$$

Comparison of this equation with (II-72) shows immediately that $\ln(1+s)$ has been replaced by $r_A - r_a$. Table 2.1 is replaced by Table 2.3. It shows the property that when $r_A - r_a$ is halved, a given gene frequency change takes *exactly* twice as long. This comes from the appearance of $r_A - r_a$ in the denominator of (II-79). Thus in this respect the continuous-generations case is simpler than the discrete-generations case.

We can also proceed from (II-78) to solve for p_t as a function of p_0 and t. Taking the exponential (the antilogarithm) of both sides of (II-78) and solving for p_t:

$$p_t \;=\; \frac{p_0\, e^{(r_A - r_a)t}}{p_0\, e^{(r_A - r_a)t} + (1 - p_0)} \tag{II-80}$$

Table 2.2: Times needed for various gene frequency changes.

s	0.01 - 0.1	0.1 - 0.5	0.5 - 0.9	0.9 - 0.9
1.0	3.46	3.17	3.17	3.46
0.1	25.16	23.05	23.05	25.16
0.01	240.99	220.82	220.82	240.99
0.001	2399.09	2198.32	2198.32	2399.09

Table 2.3: Time required to change gene frequency from 0.01 to 0.99 in a continuous-generations model.

$r_A - r_a$	time (generations)
1.0	9.19
0.5	18.38
0.2	45.95
0.1	91.90
0.05	183.80
0.02	459.51
0.01	919.02
0.001	9190.24

Table 2.4: Times required for various gene frequency changes in the case of continuous reproduction.

$r_A - r_a$	0.01 - 0.1	0.1 - 0.5	0.5 - 0.9	0.9 - 0.99
1.0	2.40	2.20	2.20	2.40
0.1	23.98	21.97	21.97	23.98
0.01	239.79	219.72	219.72	239.79
0.001	2397.90	2197.22	2197.22	2397.90

This also has a direct analogy to the corresponding expression for discrete generations: if $1 + s$ is replaced by $\exp(r_A - r_a)$ in equation (II-74) we obtain (II-80). This is in effect the same substitution we use to compare (II-72) and (II-79). Equation (II-80) is a logistic growth curve whose limits are 0 and 1. It will show the same symmetry around $p = 0.5$ as in the discrete-generations case. The correspondence of discrete and continuous cases allows us to use Figure 2.1: instead of $s = 0.2$ if we assume $r_A - r_a = \ln(1 + s) = 0.182$ we will have exactly the same Figure, except that the curve of gene frequency change will be a continuous curve rather than a discrete set of points. But this continuous curve will pass through the discrete points shown in Figure 2.1.

Using equation (II-79) we once again find that rates of gene frequency change are slower near the extremes of gene frequency than at intermediate frequencies, but now the approximate inverse proportionality between the strength of selection and the time required for gene frequency change is an exact inverse proportionality. It is well not to be blinded by the analogies between the discrete and continuous cases. There is no sense in which $r_A - r_a$ "is" $\ln(1 + s)$: these are different models. One has gene frequencies which change in discrete jumps, the other continuously.

MULTIPLICATIVE FITNESSES. We have already seen that when fitnesses of the three diploid genotypes are $(1+s)^2 : 1+s : 1$ we have exactly the same gene frequency change as in the haploid case. It follows that equations (II-71), (II-72), and (II-74) can all be directly applied. Once again the gene frequency follows a sigmoid logistic curve, or more properly a discrete set of points which is interpolated by such a curve. Once again, gene frequency change is slowest when p is near zero or near 1. Once again, there is an approximate inverse proportionality between s and the time taken for a gene frequency change. The ease with which this multiplicative case can be solved is one of the most compelling reasons for studying it.

ADDITIVE FITNESSES. It may come as a surprise that when we have the additive fitnesses $1+2s : 1+s : 1$ there is no explicit formula for p_t as a function of p_0 and t. We cannot compute the gene frequency in a future generation except by the strong arm method of repeatedly evaluating the iteration (II-46) t times. There is no simple formula for evaluating the number of generations needed to change from one given gene frequency to another. With the advent of desktop or hand-held computers, this direct approach has become feasible.

Aside from the trivial case $s = 0$, there is one case in which we can obtain an exact solution. When $s = -1/2$, so that the fitnesses are $0 : 1/2 : 1$, substitution in (II-46) will give

$$p' \;=\; p/2, \tag{II-81}$$

from which, taking natural logarithms, we can see that $\ln(p)$ decreases by $\ln 2$ each generation, from which it is easy to show that

$$t \;=\; (\,\ln(p_0) - \ln(p_t)\,)/\ln 2. \tag{II-82}$$

This equation is not correct for $p = 1$, which is a degenerate case because then the relative fitnesses of all existing individuals are zero! Note that we can never have s be more negative than $s = -1/2$, since that would imply a negative fitness for AA.

AN APPROXIMATION. Although we cannot solve the additive case exactly, there is an approximation which can be used when s is small. Equation (II-47) has $\bar{w} \;=\; 1+2sp$ in its denominator. We can expand $1/(1 \;+\; 2sp)$ as a geometric power series in s:

$$1/(1+2sp) \;=\; 1 - 2sp + 4s^2p^2 - \ldots. \tag{II-83}$$

This series will converge if $|2sp| < 1$. We are interested in small values of s. Using this series, equation (II-47) becomes

$$\begin{aligned} \Delta p \;&=\; s\,p(1-p)\,(1 - 2sp + \text{terms in } s^2) \\ &=\; s\,p(1-p) \;-\; \text{terms in } s^2 \end{aligned} \tag{II-84}$$

Our approach will be to approximate the discrete process of gene frequency change by a continuous process. This will be a good approximation if s is small. We replace the difference Δp by a derivative, and since s is assumed small, we ignore the terms involving s^2 and higher powers. The result is the approximation

$$\frac{dp}{dt} \;=\; s\,p(1-p). \tag{II-85}$$

This differential equation is really the same as (II-28), with $r_A - r_a$ replaced by s. So its solution is given by (II-80), which becomes

$$p \;=\; \frac{p_0\; e^{st}}{p_0\; e^{st} \;+\; (1-p_0)}, \tag{II-86}$$

and of course the time for a gene frequency change is

$$t \;=\; \left[\ln\left(\frac{p_t}{1-p_t}\right) - \ln\left(\frac{p_0}{1-p_0}\right)\right] \Big/ s. \tag{II-87}$$

Suppose that we try the same approximation for the multiplicative case. The denominator $1 + sp$ in (II-42) yields

$$\frac{1}{(1+sp)} \;=\; 1 - sp + s^2p^2 - \ldots \tag{II-88}$$

which once again results in

$$\frac{dp}{dt} \;=\; s\,p(1-p). \tag{II-89}$$

We immediately see that, at this level of approximation, *the additive and multiplicative cases have the same approximation.* Equations (II-86) and (II-87) apply to the multiplicative case as well. That is one for which we also have an exact solution, so we can use it to test the adequacy of this approximation technique. Comparing (II-86) and (II-87) to the exact solutions (II-74) and (II-72), we find that they differ only in replacing $1 + s$ by e^s (or, correspondingly, $\ln(1+s)$ by s). As seen in Table 2.5, these are approximations which will be quite good if s is small.

The approximations are good to within 10% when s is as high as 0.2. We can therefore have some confidence in the results for small s, and we can also be confident that the additive and multiplicative cases give nearly the same results. We have already discussed the pattern and speed of gene frequency change, and we can simply note that those patterns will be nearly exactly applicable to the case of additive fitnesses.

THE RECESSIVE CASE. For the case where A is recessive (recall that the size of the letter means nothing), there is no general solution of the recurrence equation (II-52),

Table 2.5: Accuracy of approximations for $\ln(1+s)$ and e^s.

s	$\ln(1+s)$	e^s	$1+s$
0.001	0.0009995	1.0010005	1.001
0.01	0.0099503	1.01005	1.01
0.1	0.09531	1.10517	1.1
0.2	0.18232	1.2214	1.2
0.5	0.4054	1.648	1.5
1.0	0.693	2.718	2

which gives p' as a function of p. There is one specific case which can be exactly solved, and that is the case of a recessive lethal, where $s = -1$. Then (II-53) becomes

$$\begin{aligned} p' &= \frac{p(1-p)}{1-p^2} \\ &= \frac{p}{1+p}. \end{aligned} \tag{II-90}$$

This may not look promising, but if we take $1/p'$, we get

$$1/p' = 1/p + 1. \tag{II-91}$$

This tells us that the reciprocal of the gene frequency increases by one each generation. Then

$$1/p_t = 1/p_0 + t \tag{II-92}$$

so that

$$p_t = \frac{p_0}{1+p_0\,t}. \tag{II-93}$$

The time needed to change from p_0 to p_t will, from (II-92) be simply

$$t = 1/p_t - 1/p_0. \tag{II-94}$$

An interesting subcase is when, for some n, $p = 1/n$. Then, from (II-52),

$$p' = \frac{1/n}{1/n+1} = \frac{1}{n+1} \tag{II-95}$$

Thus if $p_0 = 1/2$, then $p_1 = 1/3$, $p_2 = 1/4$, etc. When $p = 1/1000$, in the next generation $p = 1/1001$. The successive gene frequencies form a harmonic progression. Table 2.6 shows the times needed to change through various gene frequency ranges.

For comparison, in a case of multiplicative diploid selection with $s = -0.2$ (so that fitnesses are $0.64 : 0.8 : 1$) it takes 9.85 generations to reduce the gene frequency from 0.5

Table 2.6: Time needed to make various changes of gene frequency in the case of a recessive lethal.

From	To	Time (generations)
0.5	0.1	8
0.1	0.01	90
0.01	0.001	900
0.001	0.0001	9000
0.0001	0.00001	90000

to 0.1. But to reduce it from there to 0.01 takes only another 10.75 generations, and to reduce it to 0.001 takes another 10.35 generations. Thereafter each reduction by a factor of ten takes only an additional 10.32 generations. In fact, in multiplicative selection with negative s, when gene frequency is low each reduction by a factor of 10 takes $\ln 0.1/\ln(1+s) \simeq -2.3/s$ generations. In contrast, Table 2.6 shows that selection against a deleterious recessive allele gets progressively less effective as the gene becomes rare. This occurs for a straightforward reason: the lethal allele is only lethal in homozygotes. As the allele becomes rare, a progressively smaller fraction of the extant copies of the allele are found in homozygotes (in fact, p of them). Only this fraction of the deleterious allele is eliminated by selection, so that the fractional decrease of the gene frequency becomes smaller and smaller. The result is that it can take an astronomical amount of time to eliminate a recessive: 999,998 generations to reduce the gene frequency from 0.5 to 0.000001. By contrast, with multiplicative fitnesses and s as weak as $s = -0.01$, this would take only 1375 generations!

I am not aware of any other exact solutions to equations (II-52), (II-53), or (II-54). To gain further insight into the behavior of the recessive case, we must resort to approximations. Jarle Tufto (personal communication) has pointed out that we can start from the difference equation (II-54) and replace the Δp by dp/dt, getting

$$\frac{dp}{dt} \;=\; \frac{s\,p^2(1-p)}{1+s\,p^2}. \tag{II-96}$$

This differential equation can be solved by separating the variables:

$$\frac{1+s\,p^2}{p^2(1-p)}\,dp \;=\; s\,dt, \tag{II-97}$$

and then integrating. The left side can be integrated by partial fractions since

$$\frac{1+s\,p^2}{p^2(1-p)} \;=\; \frac{1}{p}+\frac{1}{p^2}+\frac{1+s}{1-p}. \tag{II-98}$$

Doing the integrals and removing the constant of integration by requiring that $p = p_0$ when $t = 0$, we finally obtain for the time required for a gene frequency change:

$$t \;=\; \frac{1}{s}\left[-\frac{1}{p_t} + \ln\left(\frac{p_t}{1-p_t}\right) + \frac{1}{p_0} - \ln\left(\frac{p_0}{1-p_0}\right) - s\;\ln\left(\frac{1-p_t}{1-p_0}\right)\right] \qquad \text{(II-99)}$$

There remains only to solve for p_t as a function of p_0 and t. Alas, this turns out to be impossible. Of course, numerical iteration of the basic recurrence equation (II-52) will give considerable insight. For the moment we defer examining numerical values from (II-99) until we can compare multiplicative, recessive, and dominant cases.

THE DOMINANT CASE. When the fitnesses of *AA*, *Aa*, and *aa* are respectively $1 + s$, $1 + s$, and 1, so that *A* is dominant, there are two cases for which we can solve exactly the relationship between gene frequency and time. One is trivial: it is the case of $s = -1$, where *A* is a dominant lethal. Consideration of the gene frequency recursion formula (II-58), or for that matter simple common sense, will show that $p' = 0$, since all the *A*'s are killed in one generation. The other case that can be solved is $s = \infty$. If we divide the fitnesses $1 + s : 1 + s : 1$ by $1 + s$ we get $1 : 1 : 1/(1 + s)$, and $s = \infty$ can then be seen to be the case where *A* is the dominant wild-type allele and *a* is a recessive lethal. We have already solved this case. We can take equation (II-90), which gives the frequency of the recessive lethal allele, and substitute $1 - p$ for p to follow the fate of the dominant allele. This gives us

$$p' \;=\; \frac{1}{2-p} \qquad \text{(II-100)}$$

p now being the frequency of the dominant normal allele *A*. The equation (II-93) which relates gene frequency to generation number can similarly be altered to follow the fate of the dominant allele.

All other cases of complete dominance cannot be solved exactly. The approximation method can be employed, and in fact the mathematics is exactly that of the recessive case with p and $1 - p$ substituted for each other. So the time taken to change from p_0 to p_t in the dominant case is exactly the time taken to change from $1 - p_t$ to $1 - p_0$ in the recessive case. As before, this cannot be solved for p_t as a function of p_0 and t. But it can be solved numerically by holding p_0 and t constant and adjusting p_t until the equation is satisfied.

DOMINANCE, RECESSIVENESS, AND GENE FREQUENCY CHANGE. We now have approximate formulas for the times taken to change through gene frequency ranges for the recessive and dominant cases, and we have an exact formula for the multiplicative case. Table 2.7 shows the times to change through various gene frequency ranges, calculated from these formulas. We can use it to get a feel for the effects of the degree of dominance:

Table 2.7: Times required to change through various gene frequency ranges when $s = 0.01$. The recessive and dominant cases are approximate values from equation (II-99).

		Favored Allele		
From	To	Dominant	Multiplicative	Recessive
0.001	0.01	234.38	231.32	90,231.2
0.01	0.1	252.19	240.99	9,239.89
0.1	0.5	310.22	220.82	1,020.31
0.5	0.9	1,020.31	220.82	310.22
0.9	0.99	9,239.89	240.89	252.19
0.99	0.999	90,231.2	231.32	234.38

This table shows times for only one value of s, 0.01, but from it we can easily find the times for other values of s. The approximation formula (II-99) shows an approximate inverse relationship between s and t. When s is cut by a factor of ten, t will increase by almost the same factor. If $s = 0.001$, it will take approximately 2320.7 generations to change from $p = 0.001$ to $p = 0.01$ instead of the 232.07 generations shown in the table. Although this exact inverse proportionality holds only for the approximations, they are good approximations, and the proportionality is nearly exact for small s. For the multiplicative case, the numbers shown in the table are exact. While formula (II-72) does not show an exact inverse proportionality of s and t, we have already seen that the closeness of $\ln(1+s)$ to s when s is small makes the proportionality nearly exact. Thus for the dominant and recessive cases we have exact proportionality in an approximate formula, and for the multiplicative case we have approximate proportionality in an exact formula. In short, the times are nearly proportional to $1/s$.

Two properties of the numbers in Table 2.7 are immediately noticeable. First, we can see that it takes a very long time to change the gene frequency of a rare recessive allele. This is true irrespective of whether that allele is advantageous or deleterious. The top part of the Recessive column shows the long times needed to increase the frequency of the advantageous recessive A allele. The numbers at the bottom of the Dominant column show a similar phenomenon. The a allele is now the recessive allele, and it is deleterious and in the process of being eliminated. The slowness of change is associated with rareness of the recessive allele rather than whether it is advantageous or deleterious.

The second feature of Table 2.7 which is striking is the similarity of the top ends of the Dominant and Multiplicative columns, and the similarity of the bottom ends of the Multiplicative and Recessive columns. This is no mere numerical accident. The fitnesses of the genotypes are:

	AA	Aa	aa
Dominant	$1+s$	$1+s$	1
Multiplicative	$(1+s)^2$	$1+s$	1

These differ only in the fitness of *AA*. When *A* is rare, almost all the *A* genes in the population will occur in *Aa* heterozygotes, and almost all *a* genes in *aa* homozygotes. This is guaranteed by the fact that in each generation the zygotes start out in Hardy-Weinberg proportions. The relative mean fitnesses of *A*- and *a*- bearing individuals will be nearly $1+s:1$ in both cases. As long as *A* is rare, we expect the course of gene frequency change to be nearly the same in both cases, and this is precisely what Table 2.7 shows.

It is less easy to see why the Multiplicative and Recessive cases behave similarly, but the same principle is involved. Superficially, the genotypic fitnesses look different:

	AA	Aa	aa
Multiplicative	$(1+s)^2$	$1+s$	1
Recessive	$1+s$	1	1

We are interested in the case when *a* is rare, so we want to compare the fitnesses of *AA* and *Aa*. Since these are relative fitnesses, we can use *Aa* as the standard genotype whose fitness we set to 1. Now the Multiplicative case changes to

AA	Aa	aa
$1+s$	1	$1/(1+s)$

which differs from the Recessive case only in the fitness of the very rare *aa* genotype. Thus when *A* is common we expect the Recessive and Multiplicative cases to behave similarly, as in fact is the case in Table 2.7.

These two phenomena stem from the same cause. The extreme rareness of the homozygote when one allele is rare means that only the fitnesses of the other two genotypes are relevant to the rate of gene frequency change. That in turn means that fitness schemes which differ only in the fitness of the rare homozygote are nearly identical in their consequences as long as the homozygotes remain rare. The slowness of change of the rare allele when it is recessive is because the fitness scheme is

AA	Aa	aa
1	1	something else

Then if *a* is rare, the fitnesses of individuals carrying *A* and those carrying *a* are both nearly 1, except that a very few of the individuals carrying *a* are *aa*. The rarer *a* is, the more similar are the fitnesses of *A*-bearing and *a*-bearing genotypes, so response to selection will slow down as *a* is made rarer. This is not the case if *a* is not recessive. When fitnesses are multiplicative, then no matter how rare *a* is, the fitnesses of *A*-bearing and *a*-bearing genotypes are in the ratio $1+s:1$.

While we have relied on approximations for the numbers in Table 2.7, the same pattern appears when exact changes of gene frequency, Δp, are computed from equations

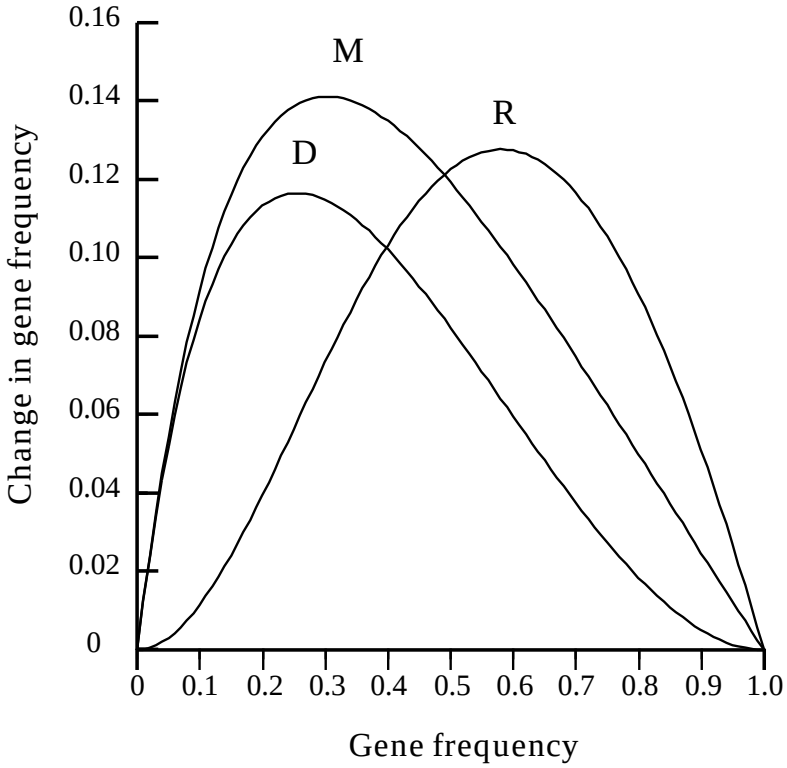


Figure 2.2: Change of the gene frequency plotted against gene frequency of A for cases in which the favored allele is dominant (D), multiplicative (M) and recessive (R). Fitnesses of $AA : Aa : aa$ genotypes were respectively 2.3 : 2.3 : 1, 5.29 : 2.3 : 1, and 2.3 : 1 : 1.

(II-42), (II-53), and (II-59). Figure 2.2 shows Δp as a function of p for the three cases considered above when $s = 1.3$. We can see both phenomena: when p is small, the Δp for the recessive case is much closer to zero than the Δp for the other cases, and in this same region the Dominant and Multiplicative curves are nearly the same. Figure 2.3 shows the resulting course of gene frequency change for the two cases:

	AA	Aa	aa
Dominant	2.3	2.3	1
Recessive	2.3	1	1

where the initial gene frequency is 0.1. The slowness with which A increases when it is recessive (squares) and rare, and the slowness with which a is eliminated when it is recessive (circles) and rare, are both evident. The fundamental principle behind all this is simple: to get a qualitative idea of how selection is operating when one allele is rare,

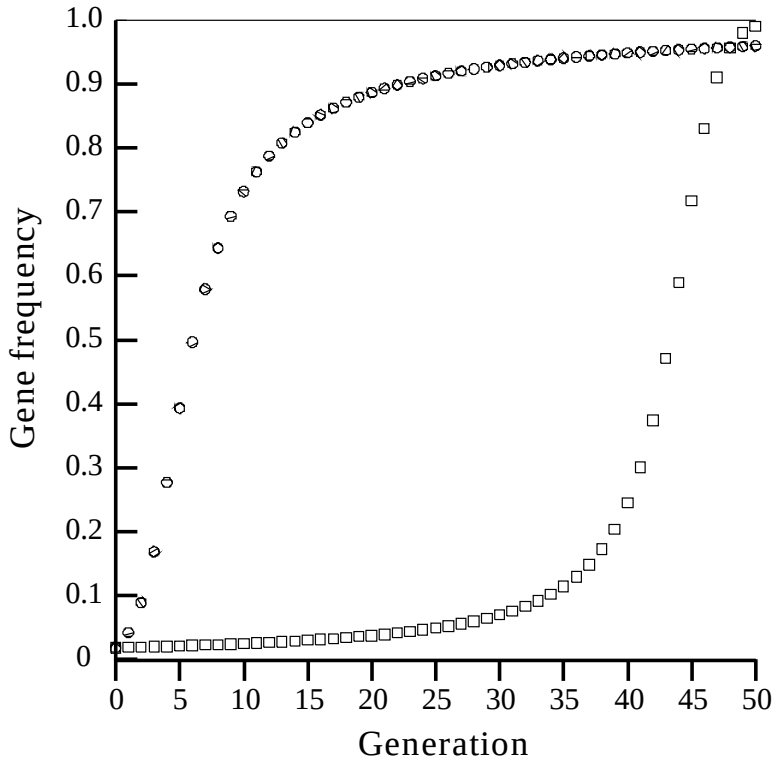


Figure 2.3: The course of gene frequency change over 50 generations when fitnesses of *AA, Aa,* and *aa* are 2.3 : 2.3 : 1 (circles) and 2.3 : 1 : 1 (squares). Initial frequency of *A* is 0.02.

compare the fitness of heterozygotes for the rare allele with the fitness of homozygotes for the common allele.

HISTORY. The harmonic series and the general method of calculation for recessive lethals appear to have been known to Castle (1903). Norton (1915) introduced the slow selection approximation and pioneered the analysis of overlapping generations, though the details of his work were not available until later (1928). Jennings (1916) and Wentworth and Remick (1916) examined the elimination of recessive lethals. Punnett (1917) used the harmonic series for recessive lethals as a powerful argument for the ineffectiveness of negative eugenics measures. This was ignored for decades by public advocates for eugenics. Warren (1917) obtained gene frequency recursions for a case where fitnesses depend on the gene frequency. Though it is evident that Fisher and Wright were familiar with the mathematics of natural selection (see especially Fisher, 1922), most of the modern work on this subject is descended from the extensive work of Haldane (1924,

1926a, 1927, 1932) who considered dozens of cases by exact, approximate, and numerical methods. He was the first author to convey a fairly comprehensive picture of the quantitative effects of natural selection. The case of overlapping generations was first treated by Norton (1915, 1928) and Haldane (1926b), who were primarily interested in the more general case in which the genotypes have different age-specific birth and death schedules.

II.7 Overdominance and Underdominance

In all cases considered in the previous section, we were dealing with patterns of fitness which resulted in the substitution of one allele for another, so that the only questions of interest are the rates of change of gene frequency through various ranges. When the fitness of the heterozygote lies outside the range of the homozygote fitnesses, the situation is altered dramatically. When the heterozygote fitness exceeds that of either homozygote, selection can maintain a stable polymorphism, and when the heterozygote has the lowest fitness, the outcome of selection can depend on the initial composition of the population. Both of these behaviors are of great biological interest.

With a few trivial exceptions, we cannot solve exactly for future gene frequencies in either of these cases, but we can gain much insight by looking at the change Δp of gene frequency as a function of the gene frequency, p. When the fitnesses are

AA	Aa	aa
$1-s$	1	$1-t$

we may recall that the change in gene frequency is

$$\Delta p \;=\; \frac{p(1-p)\,[\,t-(s+t)p\,]}{1\;-\;s\,p^2\;-\;t\,(1-p)^2} \tag{II-101}$$

which we have already seen as equation (II-65). We can start by inquiring whether there are any gene frequencies p for which Δp is zero. There are four possible ways Δp could be zero:

1. The denominator $1-sp^2-t(1-p)^2$ could be infinite,
2. p could be zero,
3. $1-p$ could be zero,
4. $t-(s+t)p$ could be zero.

The first is impossible, as t and s are finite and are not larger than 1. The second and third represent the cases where A or a are absent from the population. They reflect the rather obvious fact that, in the absence of mutation or immigration, selection acting

by itself cannot re-introduce an allele which has been lost. The fourth possibility is of interest. It establishes that $\Delta p = 0$ when p has the value

$$p_e = \frac{t}{s+t}. \tag{II-102}$$

This value of the gene frequency will represent an equilibrium gene frequency, in that if the population achieves that gene frequency it will not be expected to change further as a result of this natural selection. Our interest is in two cases, both of which have $t/(s+t)$ a realistic gene frequency, in the sense that it lies between zero and one. The cases are the one in which s and t are both positive, and the one in which they are both negative.

OVERDOMINANCE. When s and t are both positive, the heterozygote has the highest fitness. This is known as *overdominance*. It is sometimes miscalled *heterosis*, a term which refers to the observation that a cross between two populations yields offspring whose average phenotype lies above the average of either parent population. As we will see in the chapter on quantitative genetics, heterosis may or may not be due to overdominance.

When we have a case of overdominance, we have an equilibrium gene frequency at the value $t/(s+t)$. But will this equilibrium be achieved? Will it be a stable equilibrium? Some insight can be gained by noting the sign of the quantity Δp for various gene frequencies. The terms p and $1-p$ in the numerator of (II-101) are always positive, and the denominator is always positive. The sign of Δp will therefore always be the same as that of the term $t-(s+t)p$. For the case of overdominance, this will be the same as the sign of $p_e - p$. When $p < p_e$, Δp will be positive. When $p > p_e$, it will be negative. The change of gene frequency is thus always pushing the gene frequency back towards its equilibrium value. This would seem to show that the equilibrium at p_e is a stable equilibrium, that whenever the gene frequency is perturbed from this value it returns to it. But this is a hasty conclusion. We have ignored the possibility that each change in p is so great as to overshoot the equilibrium, and by such a large amount that the gene frequency oscillates wildly and gets farther and farther from the equilibrium. The process which then resemble a drunken golfer who is trying to make a small putt, but succeeds only in getting farther and farther from the hole.

ANALYZING STABILITY. When we are faced with a dynamic system which changes in discrete jumps, there is a way of determining whether an equilibrium is stable which takes the possibility of oscillation into account. The logic behind it is easily seen when, as in the present case, there is only one variable, p, changing. Suppose that we have a formula, $f(p)$, for the gene frequency in the next generation, so that

$$p_{t+1} = f(p_t), \tag{II-103}$$

and from this we can obtain another formula, Δp, which gives the change in p as a function of p:

$$\Delta p = f(p) - p. \tag{II-104}$$

There are two types of stability which we could investigate. An equilibrium will be *globally stable* if, no matter how far from the equilibrium we move the gene frequency, it always ultimately returns to that equilibrium. Global stability is quite difficult to investigate. We will concentrate instead on determining whether or not the equilibrium is *locally stable.* It is locally stable if there is some region, however small, enclosing the equilibrium, such that any perturbation which keeps the equilibrium within that region resulted in the gene frequency returning to the equilibrium. Thus if the gene frequency will return to its equilibrium when changed by less than (say) 1%, we describe the equilibrium as locally stable. If a perturbation of (say) 20% would result in no return to the equilibrium, then this equilibrium is not globally stable. If an equilibrium is not locally stable, we say that it is *unstable*.

To investigate local stability, it is sufficient to consider what happens when the gene frequency is moved an infinitesimal amount. If it always returns, it is necessarily locally stable, if not it is unstable. At the equilibrium point, $\Delta p = 0$. Figure 2.4 shows a plot of Δp against p for an overdominant case in which fitnesses are $0.85 : 1 : 0.7$. There are three equilibrium points, at $p = 0$, $p = 1$, and $p = 0.667$. It seems that $p = 0$ and $p = 1$ must be unstable equilibria. When p is perturbed just above $p = 0$, Δp is positive in that region. Thus p will continue to increase away from the equilibrium. By much the same reasoning $p = 1$ also seems unstable. Any change of gene frequency which makes p a bit less than 1 puts it in a region where p continues to decrease away from 1. The equilibrium at $p = 0.333$ looks locally stable, but a casual glance is not enough to determine its stability.

If we assume (as is true in our example) that $f(p)$, and hence also Δp, are continuous functions of p, we can make a simple algebraic analysis of local stability. In the vicinity of an equilibrium let us assume that the Δp curve can be approximated by a straight line. If x is the distance between p and p_e, so that $p = p_e + x$, then we will approximate Δp by ax. The quantity a will be the slope of the Δp curve as it passes through $p = p_e$. In the next generation, the deviation x' from the equilibrium will be

$$\begin{aligned} x' &= p + \Delta p - p_e &&= p_e + x + \Delta p - p_e \\ &= x + \Delta p &&\simeq x + a\,x &&= x\,(1 + a) \end{aligned} \tag{II-105}$$

When we are close to the equilibrium, the value of x is thus multiplied by $1 + a$ each generation. After t generations, it will be $(1 + a)^t$ times its current value.

When a is positive $(1 + a)^t$ is a positive number greater than 1, and it will grow with t. This is the situation near the equilibria $p = 0$ and $p = 1$, where the slope a of the Δp curve is positive. Any movement of p from $p = 0$ to a very small positive quantity will create a positive deviation x which then grows until p leaves the immediate vicinity of $p = 0$. Near $p = 1$, if p is set just below 1, this is a negative value of x which also becomes steadily more negative until p departs from the region near 1. Thus the algebra confirms our suspicions about the lack of stability of $p = 0$ and $p = 1$.

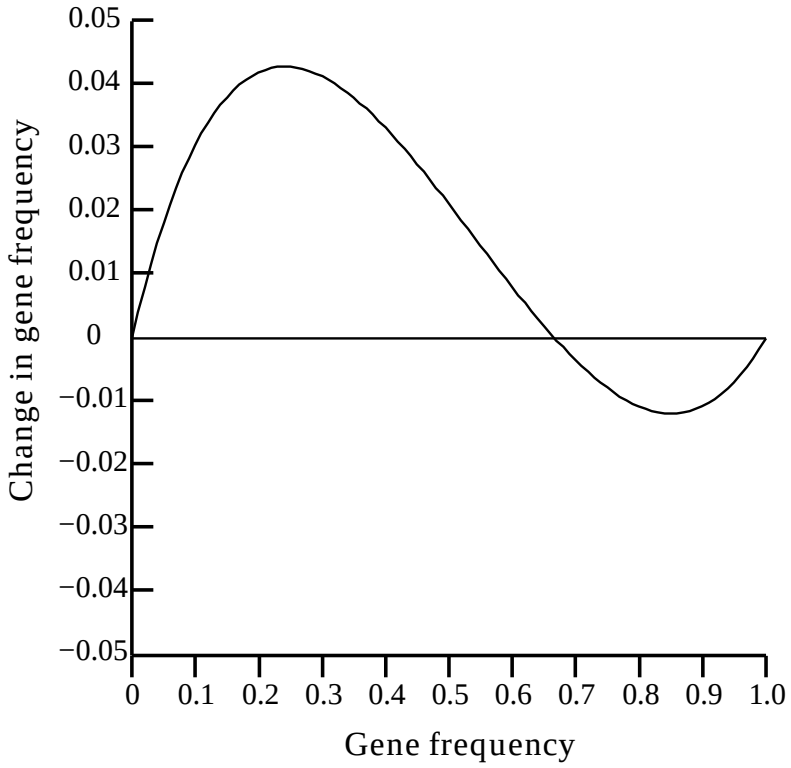


Figure 2.4: The change in gene frequency (Δp) plotted against the gene frequency in a case of overdominance where fitnesses of $AA : Aa : aa$ are 0.85 : 1 : 0.7.

When $-1 < a < 0$, $1 + a$ lies between 0 and 1. Raising $1 + a$ to the t-th power makes it approach zero without ever becoming negative. This is the case in which p approaches the equilibrium smoothly without ever overshooting. Whatever the initial sign of the deviation x, it remains of the same sign but goes to zero.

When $-2 < a < -1$, $1 + a$ lies between -1 and 0. Multiplying x by $1 + a$ will change its sign but reduce its magnitude. That corresponds to the case where there is overshooting of the equilibrium, but the overshoot leaves the gene frequency each time closer to the equilibrium than it was. The gene frequency oscillates, but with decreasing amplitude, and ultimately converges to the equilibrium.

Finally, when $a < -2$, $1 + a < -1$ so that the deviation x changes sign each generation and grows in amplitude. The overshoot is so great as to leave the population farther from the equilibrium each time. It oscillates away from the equilibrium. Extrapolation of this behavior would lead to an absurdity: the gene frequency would ultimately be greater than 1 or less than 0. This need not trouble us, since the multiplier $(1 + a)$ is only

relevant in a region small enough to allow us to approximate the Δp curve as a straight line. Farther from the equilibrium the higher-order derivatives of Δp become relevant, inevitably in such a way as to keep the gene frequency between 0 and 1.

Our criterion for local stability is this:

$$-2 \; < \; \left[\frac{d(\Delta p)}{dp}\right]_{p=p_e} \; < \; 0, \tag{II-106}$$

the brackets indicating evaluation at $p = p_e$. There are two sorts of qualification of this picture necessary. We have not investigated what will happen when a is exactly equal to 0, -1, or -2. In each case the exact behavior depends on the higher-order terms in Δp. The results do not modify (II-106) in any essential way. The second qualification is a more serious one. When generations are continuous instead of discrete, oscillation is no longer a possibility. In that case the stability is simply determined by the sign of dp/dt (the quantity analogous to Δp). If it is positive below the equilibrium and negative above it, the equilibrium is stable, and not otherwise. Overshooting is impossible because the gene frequency would have to pass smoothly through p_e in order to overshoot, and once it reached p_e it would not change further. There is an analogous damping of oscillations in discrete-time overlapping-generation models, but their analysis requires more than one variable x.

Analyzing stability is more difficult when $a = 0$. In that case second-order terms will determine the stability. For general theorems for such cases see the paper by Lessard and Karlin (1982).

STABILITY OF OVERDOMINANT EQUILIBRIA. We can now apply the local stability criterion to overdominance. After tedious algebra, the derivative of Δp at $p = p_e$ turns out to be

$$\left[\frac{d(\Delta p)}{dp}\right]_{p \, = \, t/s(s+t)} \; = \; \frac{-s\,t}{s+t-s\,t}. \tag{II-107}$$

Since in overdominant cases s and t are both positive and necessarily < 1, it is easy to show that expression (II-107) is negative and never smaller than -1. This puts all cases of overdominance in the category which do not overshoot, but approach the equilibrium smoothly from one side. In fact, unless $p = 0$ or $p = 1$ the equilibrium is always globally stable as well, never overshooting the equilibrium.

Figure 2.5 shows the course of gene frequency change starting from $p = 0.01$ and from $p = 0.99$ and proceeding to near the equilibrium when the fitnesses are

AA	Aa	aa
0.85	1	0.70

The gene frequencies converge relatively smoothly on the equilibrium value $p_A = 0.30/(0.15 + 0.30) = 0.667$.

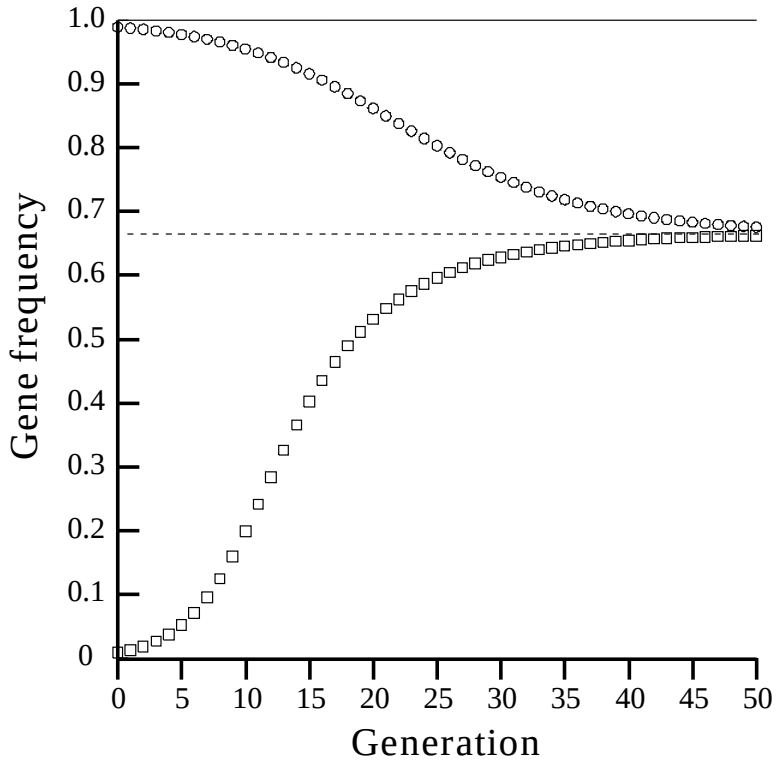


Figure 2.5: Convergence of initial gene frequencies from $p_A = 0.99$ and $p_a = 0.01$ to equilibrium when the fitnesses of *AA*, *Aa*, and *aa* are $0.85 : 1 : 0.70$

.

UNDERDOMINANCE. When s and t are both negative, the heterozygote has the lowest fitness of the three genotypes, and we refer to these cases as exhibiting *underdominance.* Once again, we have an equilibrium at $p = t/(s+t)$. The sign of Δp will be the same as that of $t-(s+t)p$, which is the same as the sign of $p - p_e$. Above p_e, Δp will be positive, and below p_e it will be negative. The change of gene frequency is now always away from the equilibrium. The equilibrium is unstable (the slope of Δp is positive). Both of the terminal equilibria at $p = 0$ and $p = 1$ are now stable. Though these conclusions are based on local arguments, they give a correct indication of the global pattern, which is that any initial departure from $p_e = t/(s+t)$ is amplified by selection until the gene frequency reaches 0 or 1. Figure 2.6 shows the course of gene frequency change in such a case, when the fitnesses are

AA	*Aa*	*aa*
1.15	1	1.3

which has $s = -0.15$ and $t = -0.3$. The small initial departures from the equilibrium gene frequency are amplified by selection until A is fixed or is lost.

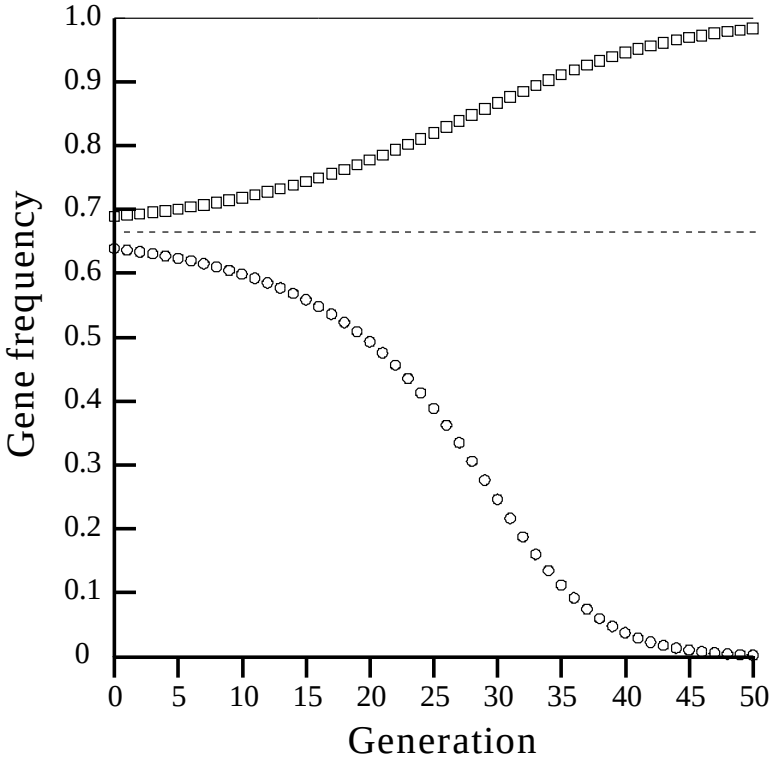
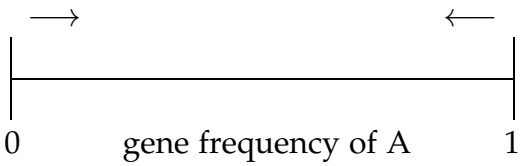


Figure 2.6: Gene frequencies in successive generations when fitnesses of AA, Aa, and aa are underdominant ($1.15 : 1 : 1.3$) and the initial gene frequency is 0.65 (circles) or 0.68 (squares).

The case of underdominance is relevant for two reasons. Practical examples are known, since chromosome rearrangements will show underdominance if there is a loss of fertility in the inversion (or translocation) heterozygote. Underdominance is also the simplest case in which the outcome of natural selection depends strongly on the initial composition of the population. In the case of Figure 2.6, an initial gene frequency were 0.67, A becomes fixed in the population. If it were 0.66, A becomes lost. This is rather dramatic behavior, and we shall see that it provides a counterexample to a widely-used biological principle.

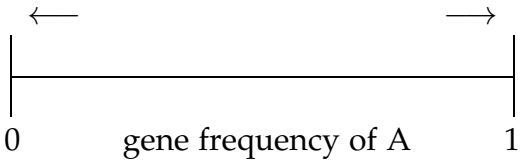
PROTECTED POLYMORPHISM. We can get a rough but useful idea of the behavior of gene frequencies in over- or underdominant cases by examining what happens when one allele or the other is rare. We make use of the rule that rare alleles appear mostly in heterozygotes, common alleles mostly in homozygotes. Consider overdominance. If A is rare, it will appear mostly in Aa heterozygotes. The common allele a occurs mostly in aa homozygotes, which are less fit. We can immediately see that A will increase in

frequency when rare. The terminal equilibrium point ($p = 0$) is therefore unstable. A completely analogous argument applies when A is common. Then a occurs mostly in heterozygotes and A in homozygotes, so a increases in frequency. So $p = 1$ is also an unstable equilibrium. We have a situation like this:



The quantity Δp is a continuous function of p, one which by the above simple arguments is positive near $p = 0$ and negative near $p = 1$. It must therefore be zero at some point in between. Although strictly speaking we have not excluded various strange kinds of instability and cycling, it turns out that the crude qualitative picture we get from these rough arguments gives us the correct impression: there is one equilibrium gene frequency, and it is stable.

For the case of underdominance, *Aa* is less fit than either *AA* or *aa*. The same rough argument applied to the situations where A is rare or A is common give us the picture.



which once again conveys the correct information: that there is an equilibrium gene frequency between 0 and 1, and that it is unstable.

In more complex patterns of selection, this kind of analysis-by-endpoints is often all that can be done. It often enables us to establish that both alleles will increase when rare. This establishes that there is a *protected polymorphism*. Whatever the behavior of the gene frequency when it is in the interior of the 0-1 scale, having a protected polymorphism guarantees that it will return to the interior of the scale. Only if the gene frequency is pushed all the way to 0 or 1 will it fail to rebound. Strictly speaking, we have not ruled out oscillations of increasing amplitude, but in most cases the slowness of gene frequency change will guarantee that this behavior cannot occur. In cases of constant relative fitnesses at one locus such as the overdominant cases we have been discussing, growing oscillations are impossible.

While the one case we have seen (overdominance) which has a locally stable polymorphic equilibrium of gene frequency is a case in which the polymorphism is protected, this will not always be the case with other, more complicated patterns of selection. Figure 2.7 shows a physical analogy to illustrate this possibility. In the case of protected polymorphism the ball will always return to the center, but when the polymorphism is unprotected, it may be locally stable, but cannot be globally stable to sufficiently large perturbations in the right direction. There are two aspects of the physical analogy which

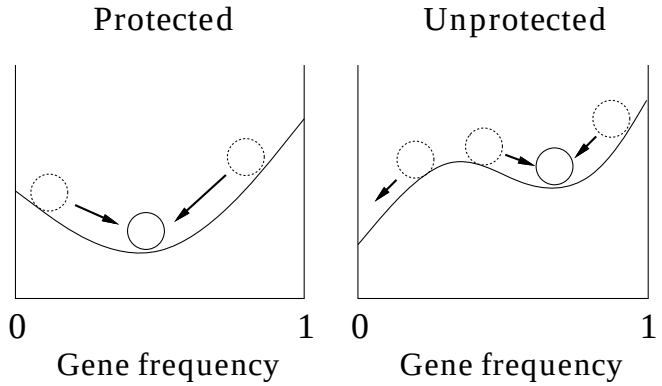


Figure 2.7: Physical analogy to protected and unprotected polymorphisms, using balls rolling on surfaces.

can be misleading. One is momentum, which has no analogue in biology. A ball which rapidly rolls to the equilibrium will continue beyond it as a result of its momentum - the gene frequency will not. The other misleading aspect of these pictures is the behavior at the two walls. When a gene frequency reaches 0 or 1 it becomes stuck and cannot change further until the other allele is reintroduced by mutation or by migration. A ball placed at one of the walls depicted in the Figure simply rolls away if the local slope leads away downhill.

HISTORY. The equilibrium gene frequencies in an overdominant polymorphism were first derived by Fisher (1922), and more detail on the dynamics in over- and underdominance was provided by Haldane (1926b). Muller (1918) had previously pointed out the properties of balanced lethal factors, when only the heterozygote can survive selection.

II.8 Selection and Fitness

This is a convenient point at which to undertake an examination of the effects of selection on the average fitness of the population. We would like to know whether natural selection does, as expected, increase the adaptedness of the population. In the scheme we have developed in this chapter, the only available measure of the extent of adaptation is the mean fitness. It would be nice if we could show that the mean absolute fitness of the population increased under natural selection, but a moment's reflection will show that this cannot be so. The absolute (Darwinian) fitness of each genotype depends on the population density. Generally, it will fall as population density rises. If the population reaches a stable size, at that point the mean absolute fitness must be 1, so that in this

sense natural selection will make no progress, since it will always result in a population which has a mean fitness of unity. There are two senses in which the population might be making progress. It may come to its equilibrium at higher and higher population density. It may come to consist of those genotypes whose fitness, relative to some standard genotype, is higher. In this section we will explore this latter suggestion.

ASEXUALS AND HAPLOIDS. In the asexual (or the one-locus haploid) case, matters are particularly simple and general results easy to obtain. Suppose that we have k different genotypes, the relative fitness of the i-th of these being w_i. Suppose that in some generation the proportions of these genotypes are $p_1 : p_2 : \cdots : p_k$. After selection the genotypes will be in proportions $p_1 w_1 : p_2 w_2 : \cdots : p_k w_k$. The frequency of the i-th genotype will be

$$p_i' \;=\; \frac{p_i\, w_i}{\bar{w}}. \tag{II-108}$$

This is the k-genotype (or in the haploid case, k-allele) version of (II-17). The mean relative fitness of the initial population is

$$\bar{w} \;=\; \sum_{i=1}^{k} p_i\, w_i, \tag{II-109}$$

the weighted average of the fitnesses. After selection, at the start of the next generation, the mean relative fitness of the population is

$$\bar{w}' \;=\; \sum_{i=1}^{k} p_i'\, w_i. \tag{II-110}$$

Substituting the right-hand side of (II-108) for p_i', we get

$$\bar{w}' \;=\; \sum_{i=1}^{k} \frac{p_i\, w_i^2}{\bar{w}}. \tag{II-111}$$

The question of immediate interest is whether $\bar{w}' > \bar{w}$, and if so, how quickly $\bar{w}$ increases. The difference between $\bar{w}$ in successive generations is

$$\bar{w}' - \bar{w} \;=\; \frac{1}{\bar{w}} \left[\sum_{i=1}^{k} p_i\, w_i^2 \right] \;-\; \bar{w}. \tag{II-112}$$

Now note that $\sum p_i w_i^2$ is the weighted mean of w^2 over genotypes in the initial generation. The variance of w over genotypes will be the difference between this mean square and the square of the mean, $\bar{w}^2$, so that

$$\begin{aligned} \bar{w}' - \bar{w} \;&=\; \frac{1}{\bar{w}} \left[\sum p_i\, w_i^2 \;-\; \bar{w}^2 \right] \\ &=\; \frac{\mathrm{Var}\,(w)}{\bar{w}}. \end{aligned} \tag{II-113}$$

Thus the increment of the mean population relative fitness is the ratio of the genetic variance in fitness (the variance among genotypes) to the mean fitness. This has two immediate implications. Since the variance can never be negative, *the mean relative fitness will never decrease* as a result of natural selection. This is a fairly reassuring result. Natural selection seems to be doing what it is supposed to - make the organisms better adapted.

The second implication of (II-113) is that the rate of progress in fitness is proportional to the square of the selection coefficient. If we double the differences in relative fitness between genotypes, increasing some fitnesses and decreasing others so as to keep $\bar{w}$ constant, the genotypic variance in fitness will not double, but will quadruple. The rate of change of the genotype frequencies will roughly double, but we will also be doubling the effect of those changes on $\bar{w}$, so that the net change in that quantity quadruples.

The tendency of natural selection to increase mean relative fitness is simple enough to explain: the more fit genotypes are increasing in frequency, so that ultimately only the single most fit genotype will exist in the population. This may seem entirely automatic and somewhat trivial. It is not: in only one other, more complex case - multiple alleles in diploids - does mean relative fitness necessarily increase. Beyond that, when we involve multiple loci, mean fitness can actually decrease as a net result of natural selection and recombination, as it can also with a single locus when fitnesses depend on gene frequency.

THE FUNDAMENTAL THEOREM OF NATURAL SELECTION. The haploid result tempts us to think that a general maximization principle must be possible. R. A. Fisher (1930) proposed what he called the Fundamental Theorem of Natural Selection according to which mean fitness would increase at a rate determined by the additive genetic variance of fitness (for more on additive genetic variances, see Chapter IX). There was also a term for deterioration of the environment. The model was of change in continuous time, leading to the suspicion that it was an approximation. In addition, there is some ambiguity about exactly which measure of fitness was the one which would increase. As will be seen below, mean fitness often, but not always, increases in population-genetic models. As a result of careful algebraic work the meaning of Fisher's formula has been made clear, but at the same time it has been made clear that it is not-so-fundamental. For more on close analysis of Fisher's results, papers by Edwards (1967, 1990, 1994, 2014), Price (1972b), Ewens (1989), and Lessard (1997) will be a good starting point.

WHERE FITNESS IS MAXIMIZED. We cannot be absolutely certain that mean population fitness is always maximized by natural selection in each generation. Computer simulations show that the net change in fitness over many generations is frequently positive when relative fitnesses of genotypes are constant. But the mathematics of change of fitness in the asexual case does establish one principle – that the mean fitness increases within the generation. When viabilities differ, if we compare the newborns in one generation with the survivors, the mean fitness of the survivors will have increased according

to equation (II-113). If there are also differences in fertility, we can get the same result if we weight genotypes by their contributions to the next generation instead of simply counting survivors. Fitness is changing as Darwin expected.

But then, alas, Gregor Mendel intervenes, the genes segregate, and the matter is cast in doubt. When we are not dealing with asexual clones, offspring are no longer of the same genotype as their parents. We cannot then use equation (II-113) to establish a net increase of fitness. It turns out that there the genetic system is not optimized to result in increase of fitness by natural selection. We would only expect such optimization of the genetic system if multiple genetic systems had competed, with choice among them according to their results. The system of inheritance that we have need only be good enough that there is a net increase if fitness most of the time. We are fit enough to sit here and read this book, but not we are not optimal organisms.

DIPLOIDY: TWO ALLELES. For the case of two alleles, there is a simple result which is suggestive. We can show that the gene frequency will always move in the direction of the peak in the plot of $\bar{w}$ versus gene frequency. Consider the general expression (II-34) for gene frequency change. In particular, consider $p(w_{AA} - w_{Aa}) + (1-p)(w_{Aa} - w_{aa})$, the quantity in brackets. Suppose that we were to plot $\bar{w}$ against gene frequency, and consider the slope of the resulting curve. This would be the derivative $d\bar{w}/dp$. Since

$$\bar{w} \;=\; p^2\, w_{AA} \;+\; 2p(1-p)\, w_{Aa} \;+\; (1-p)^2\, w_{aa} \tag{II-114}$$

we can take the derivative with respect to p and obtain

$$\begin{aligned} \frac{d\bar{w}}{dp} \;&=\; 2p\, w_{AA} \;+\; 2(1-p)\, w_{Aa} \;-\; 2p\, w_{Aa} \;-\; 2(1-p)\, w_{aa} \\ &=\; 2[\, p\,(w_{AA} - w_{Aa}) \;+\; (1-p)(w_{Aa} - w_{aa})\,]. \end{aligned} \tag{II-115}$$

The quantity in brackets in (II-115) is precisely the quantity in brackets in the next-to-last line of(II-34). Substituting from (II-115) in (II-34) we obtain

$$\Delta p \;=\; \frac{p(1-p)}{2\bar{w}}\, \frac{d\bar{w}}{dp}. \tag{II-116}$$

Of the factors on the right-hand side of this equation, p, $(1-p)$, and $\bar{w}$ can never be negative. The sign of Δp will be controlled by the sign of $d\bar{w}/dp$. When that slope is positive, $\bar{w}$ is rising as p is increased, and in that situation p increases. When $\bar{w}$ rises with a decrease in p, equation (II-116) shows us that p will decrease.

The equilibria of the system are the values of p at which $\Delta p \;=\; 0$. These occur when $p = 0$, $1 - p = 0$, or $d\bar{w}/dp \;=\; 0$. This establishes that $d\bar{w}/dp \;=\; 0$ at the polymorphic equilibrium points in overdominant and underdominant cases. This can be verified by

taking the expression for $d\bar{w}/dp$ in (II-115), equating it to zero, and solving for p. The value of p obtained is

$$p_e = \frac{w_{Aa} - w_{aa}}{(w_{Aa} - w_{AA}) + (w_{Aa} - w_{aa})} \tag{II-117}$$

which is precisely the equilibrium gene frequency with over- or underdominance. This gene frequency is also a stationary point (a maximum or a minimum) of the curve relating $\bar{w}$ to p. In a continuous curve, such a point is either a relative maximum or a relative minimum. In fact, the curve is a quadratic function of p, which immediately tells us that p_e is either the overall maximum or the overall minimum. Taking the second derivative by further differentiating (II-115),

$$\frac{d^2\bar{w}}{dp^2} = 2[w_{AA} - 2w_{Aa} + w_{aa}]. \tag{II-118}$$

We can immediately see that the curvature of $\bar{w}$ is a constant, not depending on p. If $w_{Aa} > w_{AA}, w_{aa}$ this constant is negative, and if $w_{Aa} < w_{AA}, w_{aa}$ it is positive. This establishes that in the overdominant case, the $\bar{w}$ curve has a maximum at $p = p_e$, and in the underdominant case it has a minimum there. Figure 2.8 shows $\bar{w}$ plotted against p for the case

AA	Aa	aa
0.55	1	0.25

Adaptive topography. Looking at Figure (2.8), we see that it is a hill: we can think of it as "fitness surface", "fitness landscape", or "adaptive topography". The maximum of $\bar{w}$ at the equilibrium gene frequency $p_e = 0.75/(0.45 + 0.75) = 0.625$ is evident. Equation (II-116) establishes that the gene frequency always moves in that direction which is uphill on the fitness surface, the curve relating $\bar{w}$ to the gene frequency. In cases of directional selection, where the peak p_e lies outside the $(0, 1)$ interval, this establishes that each change of p moves the population higher on the $\bar{w}$ curve. In the underdominant case there is also a continual increase of $\bar{w}$ as the population gene frequency moves away from $p = p_e$. However, we must be careful in interpreting the meaning of (II-116) in the case of overdominance. It shows us that change is in the uphill direction, but it does not, of itself, allow us to rule out the possibility that p overshoots the equilibrium, possibly even by so much as to end up farther down from the $\bar{w}$ peak than it started. As we have seen, this is not the case. There is in fact a continual increase of $\bar{w}$ until the population comes to rest on the peak of the adaptive surface, but (II-116) alone is not sufficient to establish this.

FITNESS OPTIMIZATION. We now have the pleasing picture of the population changing so as to continually increase $\bar{w}$, until it comes to rest at a peak of the adaptive surface. This would seem to provide a basis for the use of "fitness optimization" arguments in

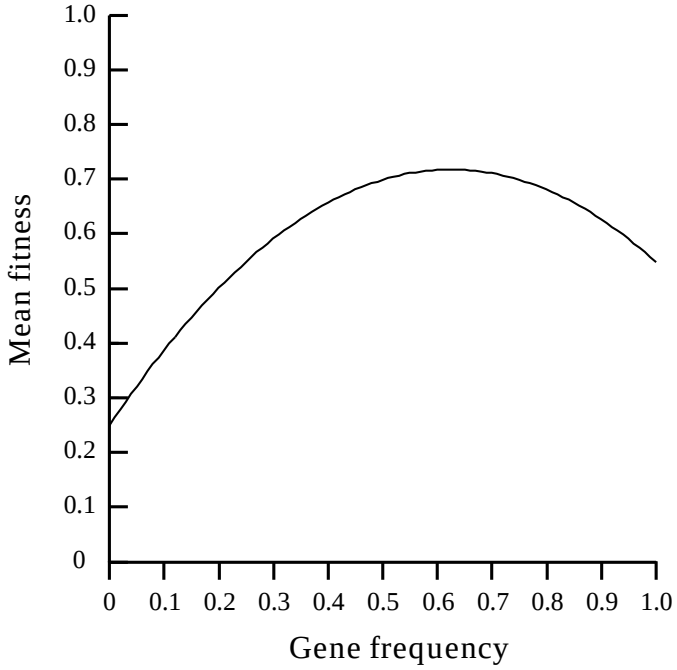


Figure 2.8: Mean fitness plotted as a function of gene frequency when fitnesses are: *AA* 0.55, *Aa* 1, *aa* 0.25.

ecology and animal behavior. In those arguments it is assumed that the population will evolve to that collection of phenotypes which maximizes the mean fitness. The picture we have developed above is only partly consistent with this notion. In the first place, genetic constraints may prevent the population from achieving this optimum configuration. In a case of overdominance, the highest mean population fitness would be achieved if all individuals were to be heterozygotes. Mendelian segregation makes this impossible: a population of heterozygotes will not be stable - it will immediately segregate out some homozygous offspring. In the second place, the peak of mean fitness which is achieved need not be the highest available peak.

When we have a case of underdominance, the final equilibrium achieved depends on the initial gene frequencies. Since the mean fitness at the equilibrium will be either $1 - s$ or $1 - t$, depending on initial position, it is entirely possible that a population will fail to find the best solution to its adaptive problems. Although its fitness cannot decrease, it may be climbing the smaller of the two peaks of the adaptive surface. If it starts out at the smaller peak, it will never find its way to the higher peak if natural selection is the only force changing gene frequencies. Fitness optimization arguments implicitly assume that a global optimization is carried out. The actual process of natural selection involves

a very narrow view of the adaptive surface. As we have seen, the local slope of the surface is all that the process of natural selection can "see" - it cannot know that there is a higher peak in another direction. If the *A* allele occurs mostly in *Aa* heterozygotes and these are not very fit, the frequency of *A* will decline, even though the *AA* homozygote may be the best genotype. Setting out from your present location and proceeding always uphill is perhaps a good recipe for escaping a small flood, but it is not the best route to the top of Mt. Everest.

The analogy between the fitness curve and a landscape is due to Sewall Wright (1932, 1935a, b), who discussed forces other than selection, particularly genetic drift, as means of moving a population across valleys in the surface.

SEGREGATIONAL LOAD. If the genetic system were asexual, reproducing by apomictic parthenogenesis, then natural selection would result in the increase to fixation of the most fit genotype. The failure of this to happen in outcrossing diploid populations is a weakness of the Mendelian genetic system. If we consider only the portion of a generation from fertilization to meiosis, the effect of natural selection will be to increase fitness by an amount which equals the genetic variance (variance among genotypes) of fitness divided by the mean fitness. This follows from the argument which we gave for asexuals, for in this portion of the generation even an outcrossing diploid population is effectively asexual. Meiosis, by disrupting genotypic combinations and reshuffling the genes, will lower the fitness, though we have seen that in the case of the two alleles it can never lower it past its initial value.

A numerical example will be useful here. Suppose that we have a diploid population with fitnesses

AA	Aa	aa
0.4	1	0.8

and a gene frequency of 0.2 in the population at the beginning of a generation. After fertilization, when the genotype frequencies are in their Hardy-Weinberg proportions $0.04 : 0.32 : 0.64$, the mean fitness will be

$$0.04 \times 0.4 \; + \; 0.32 \times 1 \; + \; 0.8 \times 0.64 \; = \; 0.848.$$

Natural selection (which in this example is most easily conceived of as differential viability) will alter the genotype frequencies to

$$0.04 \times 0.4/0.848 : 0.32 \times 1/0.848 : 0.8 \times 0.64/0.848$$

or

$$0.0189 : 0.3774 : 0.6038$$

These genotype frequencies are not in Hardy-Weinberg proportions: there is an excess of heterozygotes as a result of their high viability. If the genetic system were asexual, these would be the genotype frequencies at the start of the next generation. The

mean fitness would then be $0.0188 \times 0.4 + 0.3774 \times 1 + 0.6038 \times 0.8 \;=\; 0.862$, an increase of 0.02 over the value before selection. However meiosis intervenes and makes the next generation start in Hardy-Weinberg proportions at the new gene frequency of $0.0189 + 0.3774/2 = 0.2076$. The genotype frequencies are then

$$0.0431 : 0.3290 : 0.6279$$

which gives a mean fitness at the start of that generation of $0.0431 \times 0.4 + 0.3290 \times 1 + 0.6279 \times 0.8 \;=\; 0.84856$. This is considerably lower than 0.868 but still above the initial mean fitness of 0.848. The increase of mean fitness by 0.02 due to natural selection has been rolled back by meiosis to a net increase of 0.00056. The disruptive effect of meiosis on genotypic combinations is apparent.

If the population lacked meiosis, how much higher could its fitness be? This question was first posed by Morton, Crow, and Muller (1956), who defined and calculated the *segregational load*. This they defined as the fractional reduction in the fitness of a population as a result of the existence of Mendelian segregation. In the presence of overdominance, an asexual population could come to consist entirely of heterozygotes. In our parameterization, it would then have a mean fitness of 1. An outcrossing population would come to equilibrium at a gene frequency of $p_e = t/(s+t)$ for allele A, which would result in a mean fitness of

$$\begin{aligned} \bar{w} &= 1 - s p_e^2 - t(1-p_e)^2 \\ &= 1 - st^2/(s+t)^2 - ts^2/(s+t)^2 \\ &= 1 - st(s+t)/(s+t)^2 \\ &= 1 - st/(s+t) \end{aligned} \qquad \text{(II-119)}$$

The fraction by which fitness is reduced by the presence of segregation, relative to the fitness of Aa, is $st/(s+t)$. This is half the harmonic mean of s and t.

Is it a burden? The segregational load calculation is often misinterpreted as meaning that a population segregating at an overdominant locus somehow suffers a loss in fitness. Keep in mind that we have been calculating in terms of relative, not absolute fitnesses. These depend on which genotype is taken as the standard. We have taken Aa as the standard, so that $\bar{w}$ is necessarily less than 1. To see that an overdominant polymorphism need not impose a burden, imagine a population initially all aa, into which A alleles are introduced. Initially, the mean relative fitness of the population is $1-t$. As we have seen, A will increase when rare, until it reaches the equilibrium frequency of $p = t/(s+t)$. At that point, the mean fitness will be, from (II-119),

$$\begin{aligned} 1 - st/(s+t) &= 1 - s\,p_A &> 1 - s \\ &= 1 - t\,p_a &> 1 - t \end{aligned} \qquad \text{(II-120)}$$

so that the final mean fitness exceeds that in populations fixed for either *a* or *A*. The net effect of introducing either allele into a population fixed for the other is to bring about an increase in mean relative fitness. We can calculate a segregational load, but there is no sign that this load causes any more difficulty to the population than it would experience if it were all *aa* or all *AA*. In a purely technical sense there is no segregational load in a population which is (say) all *AA*. This is because the standard of comparison in such a population is the *AA* genotype, the genotype which would also comprise the population if it had no Mendelian segregation (as there are no *a* alleles in the population). As soon as we introduce *Aa* heterozygotes, the standard of comparison becomes *Aa*, the type which would take over in the absence of segregation, and there is now a positive segregational load. This increase in segregational load is purely a matter of changing the standard from *AA* to *Aa*. As (II-120) shows, it poses no threat to the population, whose mean relative fitness has increased.

It may be hard to imagine that introduction of an allele whose homozygote has low fitness could increase mean population fitness. A numerical example may be useful. Consider the sickle-cell hemoglobin polymorphism. Calling the two alleles *A* and *S* (not the correct notation for hemoglobin variants but good enough for our purposes), the relative fitnesses in the presence of *falciparum* malaria and inadequate medical care are thought to be roughly:

AA	AS	SS
0.8	1	0

where the heterozygote is taken to be the standard. This gives an equilibrium gene frequency for *S* of

$$\begin{aligned} p_S &= \frac{(1-0.8)}{(1-0.8)+(1-0)} \\ &= 0.2/1.2 = 0.1667. \end{aligned}$$

Before introduction of the *S* allele, the mean relative fitness of the population was 0.8. After the polymorphism reaches its equilibrium frequency, the mean relative fitness is

$$0.8 \times (0.8333)^2 + 1 \times 2 \times (0.8333) \times (0.1667) + 0 \times (0.1667)^2 = 0.8333,$$

for an increase of fitness of 0.0333. The introduction of *AS* homozygotes at a genotype frequency of 28% has more than compensated for the accompanying presence of 3% of *SS* individuals. You may want to verify that (II-119) correctly predicts the mean fitness in this case. It was not the intention of Morton, Crow, and Muller to argue that a high segregational load creates a problem for the population. Their computation was part of a rather sophisticated attempt to determine whether natural variation in viability in humans is maintained by recurrent mutation to deleterious alleles or by the presence of

overdominant polymorphism. The segregational load computation is part of the analysis of the so-called "B/A Ratio". The results are equivocal, and this method has fallen from use. The interested reader will find accounts of this controversy in Lewontin (1974, pp. 74-82) and two books by Wallace (1970, chap. 9; 1968, chap. 15).

II.9 Selection and Fitness : Multiple Alleles

When we have multiple alleles in an outcrossing diploid population with constant relative fitnesses, the principle that $\bar{w}$ increases as a result of natural selection becomes an essential part of predicting the equilibrium gene frequencies and analyzing the stability of these equilibria. We will only sketch the method in this section. We start by formulating the equations of change of gene frequencies.

EFFECT OF SELECTION. We can directly generalize equations (II-31) and (II-35), the basic equations for gene frequency change with two alleles. The extension is straightforward. The genotype frequency of the (ordered) genotype A_iA_j immediately after fertilization is p_ip_j. This holds for all i and all j, including the case where $i = j$. The contribution from these individuals to the pool of A_i gametes will be proportional to $\frac{1}{2}p_ip_jw_{ij}$, w_{ij} being the fitness of A_iA_j. The total frequency of A_i copies in the gene pool, those A_i copies that come from the left-hand gene, is the sum over j of $\frac{1}{2}p_ip_jw_{ij}$. Of course, an A_i in the gene pool could also have come from an individual of genotype A_jA_i, from which it is a copy of the right-hand gene, so there is a similar sum of $\frac{1}{2}p_jp_iw_{ij}$. The total number of genes in the gamete pool is the sum of these expressions over all i and j, so the gene frequency of A_i in the gene pool is:

$$\begin{aligned} p_i' &= \frac{\frac{1}{2}\sum_{j=1}^{n} p_i\, p_j\, w_{ij} \;+\; \frac{1}{2}\sum_{j=1}^{n} p_j\, p_i\, w_{ij}}{2 \times \frac{1}{2}\sum_i \sum_j p_i\, p_j\, w_{ij}} \\ &= \frac{p_i\left(\sum_j p_j\, w_{ij}\right)}{\bar{w}}, \end{aligned} \tag{II-121}$$

where $\bar{w} = \sum_i \sum_j p_ip_jw_{ij}$, the mean fitness of the population. In this equation $\sum_j p_jw_{ij}$ is a quite straightforward quantity: the mean fitness of the organisms in which A_i alleles find themselves, weighted by the numbers of A_i alleles they contain. We call it $\bar{w}_i$. These are direct parallels to the quantities $\bar{w}_A$ and $\bar{w}_a$ which we used in the two-allele argument in section II.4. Equation (II-121) can now be rewritten as

$$p_i' = \frac{p_i\, \bar{w}_i}{\bar{w}} \tag{II-122}$$

which also leads us directly to

$$\begin{aligned} \Delta p_i &= p_i' - p_i = p_i \frac{\bar{w}_i}{\bar{w}} - p_i \\ &= p_i \frac{(\bar{w}_i - \bar{w})}{\bar{w}}. \end{aligned} \tag{II-123}$$

The two-allele argument we presented earlier is a special case of this n-allele case.

EQUILIBRIUM. Either of these last equations readily yield the conditions for equilibrium. At equilibrium, for allele i either $p_i = 0$ or $\bar{w}_i = \bar{w}$. Thus if we want to find an equilibrium at which, out of 8 alleles, $p_1 = p_2 = p_5 = p_6 = 0$, and p_3, p_4, p_7, and p_8 are non-zero, the conditions at equilibrium are $\bar{w}_3 = \bar{w}$, $\bar{w}_4 = \bar{w}$, $\bar{w}_7 = \bar{w}$, and $\bar{w}_8 = \bar{w}$. Since $\bar{w}$ is a quadratic expression in n variables, these seem unpromising candidates for exact solution. However, we can eliminate $\bar{w}$ from these equations, so that they become: $\bar{w}_3 = \bar{w}_4$, $\bar{w}_4 = \bar{w}_7$, and $\bar{w}_7 = \bar{w}_8$. Each of these is a linear equation in the gene frequencies, since

$$\bar{w}_i = \sum_{j=1}^{n} p_j w_{ij}. \tag{II-124}$$

The equations are then, since $p_1 = p_2 = p_5 = p_6 = 0$,

$$\begin{aligned} p_3 (w_{33} - w_{43}) + p_4 (w_{34} - w_{44}) + p_7 (w_{37} - w_{47}) + p_8 (w_{38} - w_{48}) &= 0, \\ p_3 (w_{43} - w_{73}) + p_4 (w_{44} - w_{74}) + p_7 (w_{47} - w_{77}) + p_8 (w_{48} - w_{78}) &= 0, \\ p_3 (w_{73} - w_{83}) + p_4 (w_{74} - w_{84}) + p_7 (w_{77} - w_{87}) + p_8 (w_{78} - w_{88}) &= 0. \end{aligned} \tag{II-125}$$

These are three linear equations in four unknowns, but the four gene frequencies are not independent variables, as they must add up to 1. We can add a fourth equation,

$$p_3 + p_4 + p_7 + p_8 = 1. \tag{II-126}$$

We have four linear equations in four unknowns. These can be solved by standard matrix methods. The solution can then be checked as to whether it has all four gene frequencies positive, for if not the solution is irrelevant.

This pattern can be followed to find all equilibria of the n-allele system prescribed by the fitnesses w_{ij}. For each subset of the alleles, we can set up the equations corresponding to (II-125) and see whether there is an equilibrium containing only those alleles. This works, but is a rather gloomy prospect. There are in all $2^n - 1$ subsets of a set of n alleles, counting the set itself but not the empty set. All of these would have to be checked for equilibria. Each equilibrium which is found would need to be checked to determine stability. With large numbers of alleles, this can be a lot of work.

STABILITY AND MEAN FITNESS. We are presumably interested in finding all stable equilibria. Fortunately, there is a property of the population mean fitness which saves most of this work and allows us to picture the matter relatively simply. It turns out that the result of selection is always that $\bar{w}' \geq \bar{w}$. Mean population fitness never decreases. This is not particularly easy to prove, and the proof will not be given here. It was established in a series of papers by Scheuer and Mandel (1959), Atkinson, Watterson, and Moran (1960) and Kingman (1961a,b). These papers contain successively simpler proofs of the same basic result. The interested reader will also find Kingman's proof in the books by Ewens (2004, section 2.4) and Nagylaki (1977b), with a particularly detailed presentation of stability conditions in the latter.

The nondecreasing nature of mean fitness immediately lets us prove that certain points are stable equilibria. Among all combinations of gene frequencies, one will have the highest value of $\bar{w}$. (It is possible that the maxima of $\bar{w}$ be a line or plane of points, all with the same value of $\bar{w}$, but in most cases this does not arise, and we ignore it). Suppose that a point $(p_1, \ldots, p_n)$ is the maximum. In the immediate vicinity of this point there is a region in which the fitness falls smoothly off as one moves away from the equilibrium. If we perturb the gene frequencies to a point near the equilibrium, we will find that $\bar{w}$ has decreased slightly to a new value, $\bar{w}^*$. Around the equilibrium is an elliptical contour of height $\bar{w}^*$. The population must move to a point within that contour, since $\bar{w}$ cannot decrease further. Thus it will climb back to the equilibrium. We have not established this rigorously, but it is true. If a point is a local maximum of $\bar{w}$, or at least has higher $\bar{w}$ than any nearby point with no negative gene frequencies, it will be a stable equilibrium. All equilibria with more than one allele present are stationary points (maxima, minima, or saddle points) with respect to variation in the frequencies of the alleles present. Among those, the stable equilibria are the maxima. With respect to the alleles not present, the stable equilibria (and only those) will have for each allele k which is absent, $\bar{w}_k < \bar{w}$. We can check the stability of an equilibrium by verifying whether these conditions are true. Kimura (1956b) gave the conditions necessary to determine stability of an equilibrium with respect to variation in the frequencies of the alleles present, and Kingman (1961a) the condition for the alleles absent.

An example. Figure 2.9 shows an example with three alleles involving estimated fitnesses in the human hemoglobin-β polymorphism with alleles A, S and C. Note how much information can be gleaned from simply plotting the "fitness surface", $\bar{w}$ plotted against the gene frequencies. It is convenient to plot the gene frequency as points in an equilateral triangle, the distances from any point to the three sides being proportional to the gene frequencies. Since the sum of these three altitudes of the triangle must be equal, we can take this sum to be 1. In Figure 2.9, there are two equilibria. One is the familiar polymorphism of hemoglobins A and S, the other fixation for hemoglobin C. To which peak the population will climb depends on the starting gene frequencies.

Considerable insight is obtained by looking at the sides of the triangle. When only the

	A	S	C
A	0.685	0.763	0.679
S	0.763	0.151	0.545
C	0.679	0.545	1.000

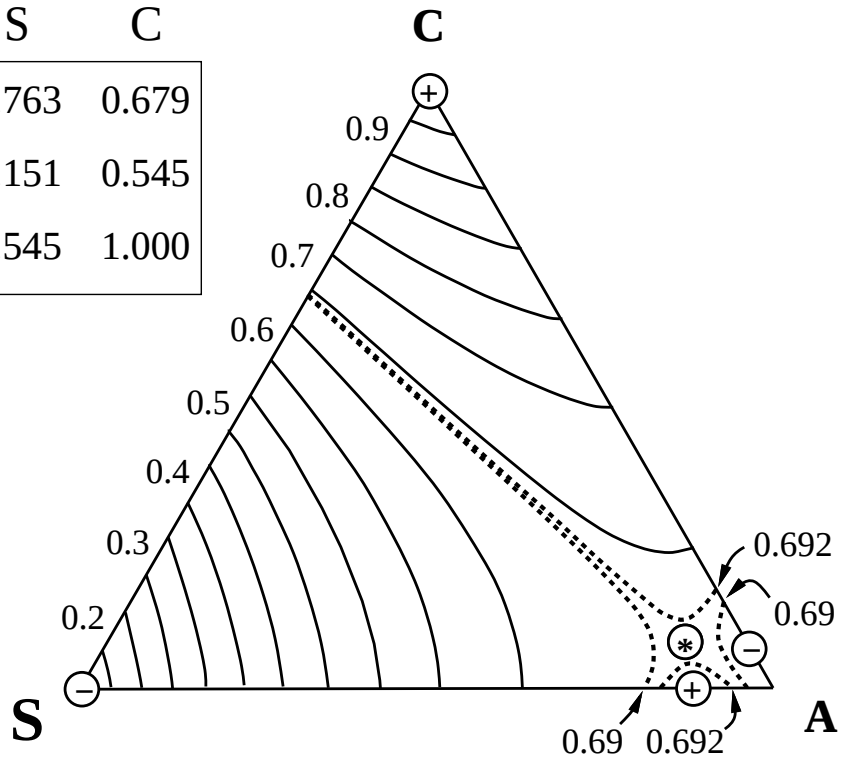


Figure 2.9: Contours of fitness plotted against allele frequencies in a three-allele β-hemoglobin polymorphism. At any point, the frequency of each allele is proportional to the attitude to the side opposite the corner labeled with that allele's symbol. Minima (-), Maxima (+) and a saddle-point (*) of the fitness surface are indicated.

S and *C* alleles are present the fitnesses of *SS : SC : CC* are 0.151 : 0.545 : 1.0, predicting an unstable equilibrium when *S* is fixed and a possibly stable equilibrium when *C* is fixed. The side corresponding to *A* and *C* has fitnesses of *AA : AC : CC* of 0.685 : 0.679 : 1.0, which is underdominance. There will be an unstable equilibrium when $p_A = 0.982$, $p_C = 0.018$, and $p_S = 0$. The equilibria at either end of the *A*-*C* side have a chance to be stable. In fact, knowing that the *C* corner of the triangle is indicated as possibly stable on analysis of both the *S*-*C* and the *A*-*C* sides is sufficient to establish its stability in general when all three alleles are considered. The *A*-*S* side shows fitnesses 0.685 : 0.763 : 0.151, the familiar sickle-cell overdominant polymorphism. This has an equilibrium at $p_S = 0.113$, $p_A = 0.887$, $p_C = 0$. The instability of fixation for *A* in the side is also proof that fixation for *A* is unstable when all three alleles are considered: it is only necessary to add a small frequency of S to move away from that equilibrium. We already know

that the state of fixation for S is unstable, from our examination of the S-C side, but this is confirmed by analysis of the A-S side.

It still remains to evaluate the stability of the polymorphism for alleles A and S. For that equilibrium to be stable, we must know that C alleles introduced at low frequency will not increase in frequency. The mean fitness of these alleles at this equilibrium will be $p_A w_{AC} + p_S w_{SC}$, which is $(0.018)(0.545) + (0.982)(0.679) = 0.676$, which is to be compared with the mean fitness of the population at the equilibrium, which is $(0.113)^2(0.151) + 2(0.113)(0.887)(0.763) + (0.887)^2(0.685) = 0.6938$. So the mean fitness of C alleles is below the population average, and they will not increase in frequency. The only part of the triangle not yet investigated for stable equilibria is the interior, where all three alleles are present. There is an equilibrium there, as the appropriate linear equations show, but it is not stable. In fact, if there is a stable equilibrium with one fewer allele (including stability to introduction of the missing allele), the interior equilibrium must be unstable. This fact follows from the identification of peaks of $\bar{w}$ with stable equilibria, and from the quadratic nature of $\bar{w}$, though it will not be proven here. Since there is a stable equilibrium with only alleles A and S, there cannot be a stable equilibrium with all three alleles.

Note that the simple plotting of $\bar{w}$ over the triangle, as done in Figure 2.9, immediately shows the locations of the two peaks, and shows that there is an interior equilibrium which is unstable, as it is a saddle of the fitness surface. Such a plot of $\bar{w}$ will always convey the full picture in multiple-allele cases, and is by far the easiest and most accessible way of analyzing these situations. Of course, the plot must be done to sufficient accuracy: both the saddle point and the A-S polymorphic equilibrium would have been missed if only the contours 0.1 apart (the coarsely dashed curves) were plotted.

The small size of the peak for the A-S polymorphism would seem to indicate that West African populations are proceeding to fixation for allele C. This is not necessarily so, because these populations are starting from the vicinity of this small peak. There is a lack of good information, but C is believed to have reached sufficient frequency in some localities. You may care to consult the more extensive discussion of this example by Cavalli-Sforza and Bodmer (1971). There have been many papers on the conditions for maintaining multiple alleles at a locus. We will not explore this literature here because overdominance seems to be very rare in nature.

II.10 Selection Dependent on Population Density

We have so far been assuming that relative fitnesses of genotypes are constant, so that we need not consider absolute fitnesses or population sizes. This amounts to the assumption that the absolute fitness of genotype A_iA_j is a product of two factors, so that $W_{ij} = w_{ij}f$, where w_{ij} is the relative fitness, which depends on the genotype, and f is a

factor which is the same for all genotypes, though it may depend on population density, gene frequency, or time. This constancy of relative fitnesses allowed us to make our analysis purely in terms of them, provided we were only interested in knowing the gene frequency, and not the population density. Although selection turned out to maximize mean *relative* fitness, the presence of the factor f allows us to concoct situations in which f is very sensitive to the gene frequency, declining rapidly as the gene frequency approaches its equilibrium. In this way one can even make a model in which a population evolves its way to extinction. In the coming sections we relax the assumption that the relative fitnesses are constant and allow them to depend on population density, time, or gene frequency.

ASEXUALS AND HAPLOIDS. If the fitnesses are functions only of population density (which we consider equivalent to population size), there are some regularities in the outcome of natural selection which provide us with partial assurance that the population does not evolve towards extinction. It is simplest to consider the absolute fitness is a function of population size, N. Each genotype may have a different dependence of N, but they all respond to the overall density N. It is important to realize that this is not the case in which each genotype's fitness responds only to its own numbers.

In the two-genotype (or two-allele) case, we have absolute fitnesses $W_1(N)$ and $W_2(N)$. The numbers of the two genotypes follow the dynamics

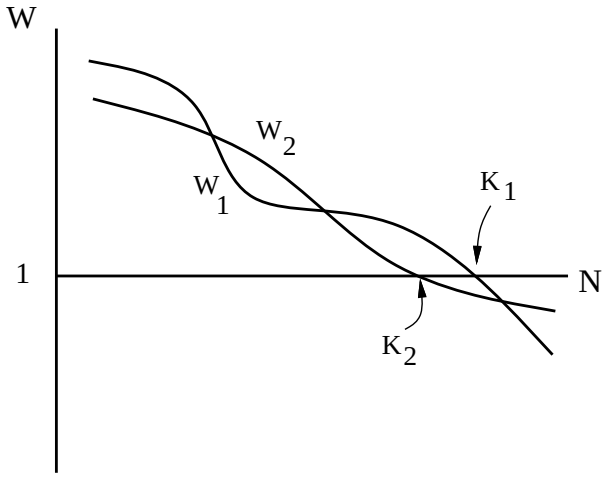
$$\begin{aligned} N_1' &= N_1\, W_1(N) \\ N_2' &= N_2\, W_2(N), \end{aligned} \tag{II-127}$$

where $N = N_1 + N_2$. Charlesworth (1971) has made a particularly penetrating analysis of density-dependent selection. In the asexual case his argument works out particularly simply, and is the basis for what is presented here. Suppose that each genotype's absolute fitness declines strictly monotonically as N increases. We can trace out curves of W_1 and W_2 as functions of N:

There will be particular points K_1 and K_2, the values of N at which w_1 and w_2, respectively, reach unity:

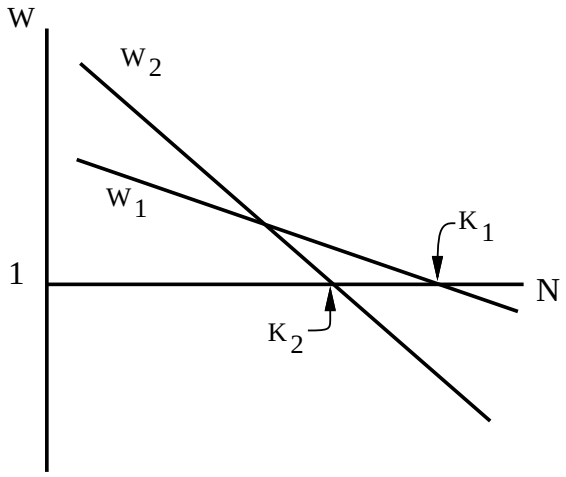
$$\begin{aligned} W_1(K_1) &= 1, \\ W_2(K_2) &= 1. \end{aligned} \tag{II-128}$$

Suppose (arbitrarily) that $K_2 < K_1$. If the population size N is below K_2, then equations (II-127) show that both subpopulations are growing. Which is growing faster is not known without considering the exact shapes of the dependence of W_1 and W_2 on N, but the total population size must continue increasing until it reaches K_2. At that stage, unless the genotype A_1 has been completely eliminated, it will have the advantage, for when N is above K_2 but below K_1, $W_1 > 1 > W_2$. The population of A_1 genotypes



continues increasing and that of A_2 decreases. Ultimately A_1 comes to constitute the entire population, at which point the population size N has risen to K_1. If the initial population contains both types and has $N > K_1$, a similar argument applies and the population size falls into the range $K_2 < N < K_1$, after which type A_1 will win out. A similar argument applies with multiple strains (or multiple alleles).

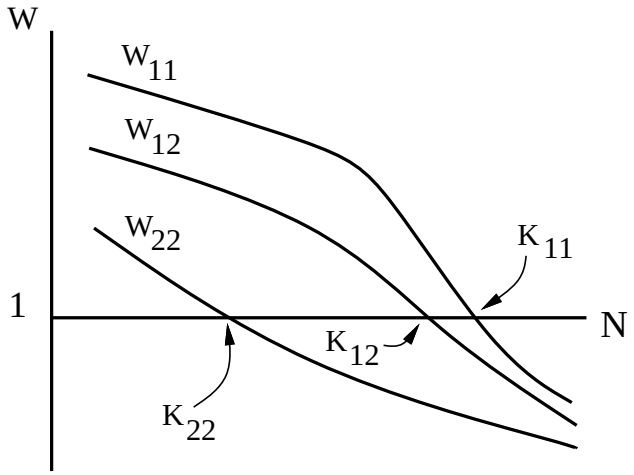
We thus have some assurance that the genotype which will maintain the highest equilibrium population size will be favored by natural selection. This would seem to guarantee that natural selection will act so as to reduce the probability of extinction of the population, but a moment's reflection will show that this is not necessarily the case. The $W_2(N)$ curve may pass through $N = K_2$ at a very steep angle, while $W_1(N)$ passes through $N = K_1$ at a shallow angle:



If this happens to be the case, a population of A_1 individuals may recover less quickly from environmental fluctuations than would an A_2 population. If a fluctuation of the environment or a fluctuation due to the randomness of birth and of death events were

to carry the population size below the point of intersection of the W_1 and W_2 curves, at those low population sizes a population consisting of the A_1 genotype will grow more slowly than one consisting of A_2. The favored A_1 genotype will then be more susceptible to extinction, provided that population size is sufficiently strongly fluctuating to carry it often into this region.

DIPLOIDS. The diploid version of density-dependent selection has, for the two-allele case, three curves, one for each genotype. The discussion here will assume that $W = 1$ is the value when the population exactly replaces itself (if it were actually $W = 2$ the expressions and figures would have to be adjusted accordingly). In certain simple cases the outcome is straightforward. If the three curves do not cross in the region between the point where the lowest one reaches $W = 1$ and the point where the highest one reaches $W = 1$, then the population is guaranteed to grow into this region, and the outcome of natural selection can be qualitatively predicted from the ordering of the fitness curves. If we have the heterozygote intermediate then natural selection will favor the A_1 allele.



As it reaches high frequencies the population will come to an equilibrium at $N = K_{11}$.

When the curves are allowed to cross in the relevant region things can be rather complicated. The exact equilibrium gene frequency depends on the population size, whose growth rate in turn depends on the gene frequency. The equilibrium of the two variables requires solution of two simultaneous equations. Charlesworth (1971) has made a general examination of the matter. He proved that polymorphism can only be maintained provided that there is overdominance in the K_{ij}. He also showed that natural selection maximizes the quantity $N^*(p)$, where $N^*(p)$ is the equilibrium population size which would be achieved if p were held fixed, namely the root of

$$p^2\, W_{11}(N) \;+\; 2p(1-p)\, W_{12}(N) \;+\; (1-p)^2\, W_{22}(N) \;\;=\;\; 1, \qquad \text{(II-129)}$$

where p is given and N is the quantity which we vary. In particular, if $K_{11} > K_{12} > K_{22}$, then allele A_1 will be fixed. Natural selection acts as if it is trying to maximize

the equilibrium population size. This does not necessarily mean that the population size increases steadily throughout the course of natural selection. Ultimately, natural selection finds that value of p which allows it to reach the highest equilibrium population size. Of course, this maximum is a local maximum - in underdominant cases the result depends on the initial gene frequency and the global maximum need not be reached.

OSCILLATIONS AND CHAOS. The entire foregoing discussion rests on a premise that the population settles down to a single equilibrium size. This will not be true if the $W_{ij}(N)$ pass through 1 too rapidly. If population density regulation causes too great a decrease in absolute fitness, the equilibrium population size can be unstable. The result may be either cyclical oscillation of population size or a pattern known as "chaos," in which the population size remains within a fixed interval of sizes and fluctuates without ever achieving the same size twice (May, 1974, 1976). In such cases the preceding analysis fails to hold because we cannot argue that the population comes to rest at that size N at which $\bar{W}(N) = 1$.

SOME PARTICULAR GROWTH LAWS. Various workers have analyzed particular functions $W(N)$ which seem to do a good job of approximating biological reality, but which allow some exact analysis. The simplest model has $W(N)$ a linear function of N:

$$W_{ij}(N) = 1 + r_{ij} - r_{ij}N/K_{ij}, \tag{II-130}$$

which is often referred to as the discrete logistic growth law. It was first worked on in a genetic context by Roughgarden (1971). Roughgarden showed numerically, and Charlesworth (1971) analytically, that natural selection acts as if trying to maximize K, in that when $K_{11} > K_{12} > K_{22}$ allele A_1 is favored, with polymorphism only when there is overdominance in K. This is only true when the equilibrium value of N is approached. Where $r_{ij} > 2$, oscillations ($r < 2.83$) or chaos ($r > 2.83$) occurs, (May, 1976) and it is not known what governs the outcome of natural selection. One of the disadvantages of the growth law (II-130) is that when N is large the population can go negative. This tends to occur if $r > 3$.

A second growth law,

$$W_{ij}(N) = \frac{1 + r_{ij}}{1 + (1 + r_{ij})N/K_{ij}} \tag{II-131}$$

cannot yield oscillation or chaos no matter how large is r. This growth pattern has the advantage of never yielding a negative fitness. It has been introduced in a genetic context by Clarke (1972). Roughgarden (1979) shows that the sequence of population sizes for a population with a single genotype lie on a logistic growth curve, so that in this sense this growth curve is a truer analog of the logistic growth law than (II-130). Little explicit analysis of this growth curve has been done, but since it never yields

oscillations or chaos, Charlesworth's (1971) results predict that natural selection will favor the genotype with the highest K if there is neither over- nor underdominance, will bring about a stable polymorphism if there is overdominance in K, and an unstable polymorphism if K is underdominant.

A third growth law,

$$W_{ij} = \exp\left[r_{ij}\left(1 - \frac{N}{K_{ij}}\right)\right], \tag{II-132}$$

has been extensively investigated by May (1972, 1974; May and Oster, 1976). In a population consisting of a single genotype, there will be a stable equilibrium population size if $0 < r < 2$, cyclic behavior if $2 < r < 2.692$, and chaos if $r > 2.692$. For $r < 2$ Charlesworth's analysis predicts maximization of K. For the oscillating and chaotic realm Michael Turelli and the late Timothy Prout, in unpublished work, have proven, by applying rules similar to those we shall develop in the next section, that in that case as well selection favors genotypes with high values of K.

In both of these last two cases there has been no tradeoff between r and K: the outcome of selection is predicted qualitatively by the K values, although the r's will influence the exact position of a polymorphic equilibrium. On intuitive grounds, we would expect that sometimes there would be a compromise between r- and K-selection, the outcome depending on both parameters. I have analyzed (1979) a simple (if rather extreme) model due to Williamson (1974) which shows effects of both r and K. It is

$$W_{ij} = \begin{cases} R_{1,ij}, & \text{if } N \leq K_{ij} \\ R_{2,ij}, & \text{if } N > K_{ij} \end{cases}. \tag{II-133}$$

This model is almost always chaotic, except for very special values of R_1 and R_2 which allow it to be cyclic. The outcome of selection is complex: if the genotypes differ only in R's, the effect of natural selection is to favor the genotype with the highest value of $(\ln R_1)/(-\ln R_2)$. When K alone varies, the genotype with the highest value of K is favored. When both quantities differ among genotypes the outcome depends on both. Interestingly enough, when r- and K selection are counterposed it is possible in this growth model to have protected polymorphism under certain conditions *even in an asexual (or haploid) population.* Roughgarden (1971) introduced a model in which seasonality causes the population size to cycle, and in that case as well found that there were effects of both r and K, with polymorphism possible when r- and K-selection were appropriately counterposed.

ADDITIONAL WORK. The investigation of density-dependent selection was pioneered by MacArthur (1962). Roughgarden has extended the conditions for maximization of $N^*(p)$ to multiple interacting species - an account of the results and further references will be found in that book (1979). Other investigations of particular interest are those of

Anderson (1971), León and Charlesworth (1978), Asmussen and Feldman (1979), Asmussen (1979), and Hansen (1992). Turelli and Petry (1980) have made a thorough analysis of a model which has both density-dependence and temporally-varying environments.

II.11 Temporal Variation in Fitnesses

It is unlikely that relative fitnesses of genotypes will remain constant through time, since the environmental conditions, population density, and densities of other species will fluctuate, and these will in many cases affect the strength of natural selection. There has been a certain amount of work on cases in which fitnesses vary randomly or cyclically from one generation to the next. It leads to surprisingly simple conclusions.

ASEXUALS AND HAPLOIDS. Dempster (1955) first considered a haploid case with two alleles. If the fitnesses in generation t for alleles A and a are $1 + s_t : 1$, we can invoke equation (II-13), altering it only by subscripting s:

$$\frac{p_A^{(t+1)}}{p_a^{(t+1)}} \;=\; (1 + s_t)\,\frac{p_A^{(t)}}{p_a^{(t)}}. \tag{II-134}$$

The ratio p_A/p_a after t generations of selection will be (using $\prod$ as the symbol for repeated multiplication)

$$\frac{p_A^{(t)}}{p_a^{(t)}} \;=\; \left[\prod_{u=0}^{t-1}(1 + s_u)\right]\frac{p_A^{(0)}}{p_a^{(0)}}. \tag{II-135}$$

in direct analogy to equation (II-71). We can immediately see what happens in the case of cyclic selection. This is the case where there are T different values of s, which are repeated cyclically, so that the fitnesses are $1 + s_1, 1 + s_2, 1 + s_3, \ldots, 1 + s_T, 1 + s_1, \ldots, 1 + s_T, \ldots$. If we follow the change of gene or genotype frequency over one cycle, we must take the product of these fitnesses over the cycle. If that product (in square brackets in the equation) is greater than one, the ratio of gene frequencies is increased by each cycle of generations. If it is less than one, the ratio is decreased, and if it is exactly one, it does not change. Whether the A allele increases to ultimate fixation depends on whether this product of relative fitnesses exceeds one. Notice that if we take the T-th root of the product, that is the *geometric mean* of the T quantities $1 + s_1, \ldots, 1 + s_T$. Taking the T-th root of a quantity does not change the fact of whether it exceeds 1, since T is positive. We can succinctly state the result by saying that *the allele with higher geometric mean relative fitness takes over*. This follows because the geometric mean of $1 + s_t$ is the same as the

ratio of the geometric means of $w_A^{(t)}$ and $w_a^{(t)}$, since

$$\left[\prod_u (1+s_u)\right]^{1/T} = \left[\prod_u \left(\frac{w_A^{(u)}}{w_a^{(u)}}\right)\right]^{1/T} = \left[\frac{\prod_u w_A^{(u)}}{\prod_u w_a^{(u)}}\right]^{1/T} = \left[\prod_u w_A^{(u)}\right]^{1/T} \Big/ \left[\prod_u w_a^{(u)}\right]^{1/T} \tag{II-136}$$

Randomly varying fitnesses. When the fitnesses do not go through an exact cycle, but vary randomly with time, the mathematics is more complex but the result is essentially the same. In the short run we may encounter a run of generations favorable to A or to a, so we must look to the long run for more exact predictions. The result will depend on the value of the product $\prod(1+s_u)$, that is on $\prod(w_A^{(u)}/w_a^{(u)})$. Though there is no simple generalization covering all cases, we need only place very mild restrictions on the way in which fitnesses vary to get a simple result. If we can assume that there is very little long-term correlation of fitnesses (if we cannot predict future fitnesses far ahead of time) then we need only take the logarithm of this product and get a sum,

$$\ln\left[\prod_u (1+s_u)\right] = \sum_u \ln(1+s_u) \tag{II-137}$$

and we can apply the Strong Law of Large Numbers, from probability theory, to this sum. As we consider large numbers of generations, this sum will approach t times its expectation, and its variance will also rise proportionally to t.

If a number's expectation rises with t, but its variance also rises similarly, that means that its standard deviation rises as the square-root of t. So if we ask how many standard deviations the number is from zero, this rises proportionately to $t/\sqrt{t}$, which is $\sqrt{t}$. Sooner or later it will be 10 standard deviations above (or below) zero, so that we can be certain that the net effect of selection is as expected.

Thus if the expectation of each term is positive, it will tend to ∞, and if the expectation is negative, it will tend to $-\infty$. The expectation of the logarithm of $1+s_u$ will be positive when the geometric mean of $1+s_u$ exceeds 1. If the geometric mean relative fitness of A exceeds that of a, the sum of logarithms becomes positive, and increasingly so. If the geometric mean relative fitness of A is less than of a, the sum goes negative and becomes increasingly so, without limit, as time passes. Since we are looking at the logarithm of the original product of fitnesses, these correspond to the product becoming infinite or going to zero.

After all the probability theory, we again have the result that the allele with higher geometric mean fitness wins out. Within broad limits the pattern of correlation of fitnesses through time does not affect the ultimate outcome, a fact which is not obvious in advance. The reader to whom numerical examples are more illuminating may care to ponder these two sets of fitnesses:

Type of year	Genotype		
	A		a
Wet	1.1	:	1
Dry	0.8	:	1

If wet and dry years occur in a cycle of two wet years followed by one dry year, in regular succession, WWDWWDWWD..., then since $1.1 \times 1.1 \times 0.8 = 0.968 < 1$ allele A will decrease with each cycle of two generations. If instead we have some other (possibly random) pattern of wet and dry years, with wet years 2/3 of the time and dry years 1/3 of the time, we can predict the outcome of natural selection by computing the geometric means:

$$A: \qquad (1.1)^{2/3}\,(0.8)^{1/3} \;=\; 0.968^{1/3} \;=\; 0.989217$$

$$a: \qquad 1.0^{2/3}\,1.0^{1/3} \;=\; 1,$$

again predicting a long-run decrease in the frequency of A. The outcome essentially does not depend on the pattern of correlations or the length of cycles. The short term variation in fitnesses may be dramatically affected by whether dry years tend to come in runs, but the long-term result depends only on the relative frequencies of wet and dry years. This is perhaps counterintuitive.

Figure 2.10 shows a cyclic case and a random case. Other random sequences of Wet and Dry may rise to higher frequencies or drop quickly to lower ones, but all will ultimately lose the wet-adapted allele. Note that despite the continued variation in environments, there is no general principle that both wet-adapted and dry-adapted alleles will persist in the long run.

Note also that the allele with higher geometric mean fitness wins out, but this is not necessarily the allele with higher arithmetic mean fitness! In the numerical example, the *arithmetic mean* relative fitness of A is $(2 \times 1.1 + 1 \times 0.8)/3 = 1$, which is tied with the arithmetic mean of a's relative fitness, which is 1. Nevertheless A, the allele with equal mean fitness, is certain to ultimately be eliminated. In the case of temporal variation in fitness, $\bar{w}$ is not maximized. In the numerical example in Figure 2.10, the arithmetic mean fitness of A is higher than that of a, and yet it is certain to ultimately be eliminated. However, the dependence on the geometric mean does give us some assurance that evolution will not be generally maladaptive. For example, if in all years $w_A < w_a$, it is impossible that the frequency of A should ever increase.

DIPLOIDS. We have concentrated so much attention on the asexual or haploid case because this is the key to the analysis of the diploid case. A complete analysis of the diploid case is forbiddingly difficult, but if we confine attention to the conditions for protected polymorphism, we can find these easily from the asexual conditions. We want conditions under which both A and a increase when rare. We start with the effect of selection on the ratio of gene frequencies. If we take equation (II-36) and put superscripts

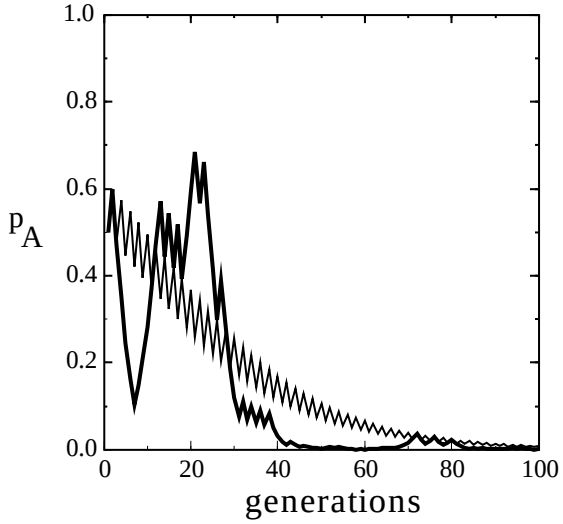


Figure 2.10: Course of gene frequency change in a haploid organism in a numerical example of a case of alternating Wet and Dry years (lighter lines) and when there are random Wet and Dry years, independently drawn with equal probabilities. In the two cases the relative fitnesses of A are 1.5 and 0.6 in Wet and Dry years, respectively. The starting gene frequency in both cases is 0.5.

on the fitnesses to indicate their dependence on the generation number, it becomes

$$\frac{p'}{1-p'} \;=\; \frac{p}{1-p} \;\times\; \frac{\mathrm{p}\; w_{AA}^{(t)} \;+\; (1-p)\; w_{Aa}^{(t)}}{p\; w_{Aa}^{(t)} \;+\; (1-p)\; w_{aa}^{(t)}}. \tag{II-138}$$

One of the cases in which we are interested is when A is rare. Then p is near zero, so that equation (II-138) is well approximated by

$$\frac{p'}{1-p'} \;=\; \frac{p}{1-p}\; \frac{w_{Aa}^{(t)}}{w_{aa}^{(t)}}. \tag{II-139}$$

This is precisely the asexual or haploid formula, with $w_A^{(t)}$ and $w_a^{(t)}$ replaced by $w_{Aa}^{(t)}$ and $w_{aa}^{(t)}$. The reason for this concordance is straightforward. We are interested in the behavior of A when it is rare. When it is, almost all A alleles occur in heterozygotes, and almost all a alleles in aa homozygotes. The inheritance of the genotype is then effectively haploid, since $Aa \times aa \rightarrow 1/2\ Aa + 1/2\ aa$, just as in haploids $A \times a \rightarrow 1/2\ A + 1/2\ a$. Of course, an $Aa \times Aa$ mating does not follow the haploid analogy, but such matings

essentially never occur if A is very rare. A rare allele can thus be treated as if haploid, or even asexual. From (II-139) the conditions for ultimate increase of A are immediate: it will increase if the geometric mean over time of w_{Aa} exceeds that of w_{aa}. This condition is actually the condition for the product $\prod(w_{Aa}/w_{aa})$ to rise to infinity. As the gene frequency of A increases away from zero, it leaves the region where the approximation (II-139) holds. So we have obtained a condition for A to increase away from zero, not for A to go to fixation.

A similar analysis can be carried out when a is rare, and the results are completely analogous. The a allele increases when rare when the geometric mean of w_{Aa} exceeds that of w_{AA}. Putting these together, *the condition for protected polymorphism is overdominance of geometric mean fitnesses of the genotypes*. If Aa is the genotype with highest geometric mean fitness, we will have a protected polymorphism. As in the asexual or haploid case, the pattern of correlations essentially does not affect this condition, which holds for both cyclic and random variation in fitnesses. Of course, this geometric mean overdominance condition is only part of the story. It tells us whether we have a protected polymorphism, but not the distribution of gene frequencies over time or the amount of short-term fluctuation in gene frequencies. These *are* affected by the pattern of temporal correlation in fitnesses.

In the case of random temporal variation in fitnesses, the conditions for protected polymorphism are also the conditions for existence of a polymorphism (except in the case where one allele is exactly recessive in geometric mean fitness). The equivalence of polymorphism and protected polymorphism is easily motivated. Random variation of fitnesses will cause the frequency of each allele to occasionally wander close to zero. When this occurs, to retain the polymorphism it must have probability one of returning to the interior of the $(0,1)$ interval. The conditions for certainty of increase of a rare allele thus are not only sufficient to ensure a polymorphism, but necessary for its maintenance as well.

In the case of cyclic variation of fitness we cannot use this argument. It is possible to have a stable cycle of gene frequencies in the interior of the $(0,1)$ interval, even though one or both alleles will be lost if made rare. Since cyclic variation in fitness need not ever cause either allele to become rare, a polymorphism need not be protected in order to exist and be stable. An example is a two-generation cycle of fitnesses in which the fitnesses of the genotypes are alternatively $0.9 : 1 : 0.9$ and $1.115 : 1 : 1.115$. It turns out that if the gene frequency starts below about $p = 0.18$, A will be lost. If it starts above $p = 0.82$, a will become lost and A fixed. The geometric means of the genotypes are $1.0017 : 1 : 1.0017$, which is geometric mean underdominance, so that protected polymorphism is not guaranteed. Yet there is a stable polymorphism, for any starting point between 0.18 and 0.82 results in the gene frequency converging to a stable equilibrium at $p = 0.5$. So stable polymorphism can exist in the absence of protected polymorphism.

FITNESSES VARYING WITHIN A GENERATION. When fitnesses are different for

different life stages, the result is the same - yet different. Strobeck (1975) considered this case, and showed that there will be protected polymorphism if the fitnesses in different life stages show geometric mean overdominance. In the random case this result is not very simple, but in the "cyclic" case it is obvious. Suppose that the fitnesses (viabilities) were $0.8 : 0.9 : 1$ in the larval stage, and were $1 : 0.915 : 0.82$ in the adult stage before reproduction. As we saw when we first discussed fitness, the overall fitness involves the product of these quantities, assuming that the causes of mortality act independently. The result is fitnesses which show overdominance: $0.8 : 0.8235 : 0.82$. If these sets of viabilities repeat every generation, this is a case of simple overdominance, and causes a stable polymorphism.

We can also treat this as a case of fitnesses cycling within a generation. At no single life stage is there overdominance. The geometric mean fitnesses among life stages show overdominance, being the square roots of the overall fitnesses: $0.8^{1/2} : 0.8235^{1/2} : 0.82^{1/2}$ or $0.894 : 0.9075 : 0.906$. So the results follow both the rules for constant fitnesses and those for varying fitnesses. The mean fitness is maximized by selection, but this is the mean of the overall fitness of each genotype. If we were misguided enough to average fitnesses within each generation among life stages, we find arithmetic means of $0.9 : 0.9075 : 0.91$, which do not show overdominance.

We can sum up the case of variation of fitness within a generation by saying that there will be protected polymorphism if the geometric mean fitness among life stages is overdominant. In the cyclic case, where the cycle repeats every generation, this is simply the requirement that overall fitnesses be overdominant, in which case overdominance is necessary as well as sufficient for stable polymorphism to exist.

There need not be overdominance in any life stage for there to be overdominance in net fitness. In this sense conflicting directional selection in different parts of the life cycle can cause polymorphism. It is important to note that it is the net overdominance which is necessary: in the haploid or asexual case there is no pattern of conflicting directional selection in different life stages which can cause polymorphism.

THE SAS-CFF MODEL. Gillespie and Langley (1974) have made temporal variation the centerpiece of a general hypothesis for the maintenance of protein polymorphism. They argue that geometric mean overdominance can arise from standard enzyme kinetics. If the heterozygote has an enzyme activity which is the arithmetic mean of the activities of the two homozygotes, and if the curve relating fitness to enzyme activity is concave downwards, then sufficient variation in the enzyme activities of alleles over time can result in geometric mean overdominance. This is biologically plausible, which is not the same as saying that we know that it is important as a cause of real polymorphisms. For further work generalizing the SAS-CFF model, see the papers by Gillespie (1978, 1979, 1982), and particularly his remarkable book *The Causes of Molecular Evolution* (1991).

REFERENCES. Haldane and Jayakar (1962) were the first to give the geometric mean

overdominance condition. The necessary side conditions for stable polymorphism when one allele is completely recessive were given by Haldane and Jayakar (1962) and, more generally, by Hoekstra (1975). Gillespie (1973) developed the geometric mean conditions further, and Norman (1975b) gave a general proof that if fitnesses vary independently from generation to generation, overdominance of geometric means is both necessary and sufficient for maintenance of a polymorphism.

For early work on varying selection, the reader may wish to consult my review article (1976) or the reviews by Hedrick, Ginevan, and Ewing (1976) and Hedrick (1986). The much more recent paper by Cvijović et al. (2015), on the fixation probability of a new mutant in a fluctuating environment cites many more recent papers.

II.12 Frequency-Dependent Fitnesses

Fitnesses can also vary as a function of the genetic composition of the population. When they depend on the gene frequencies or the genotype frequencies, various complex outcomes are possible, including oscillation of the gene frequency and chaotic fluctuation. We are most interested in a simpler outcome, stable polymorphism. A natural condition to examine is frequency-dependent selection in which the rare allele is at an advantage. There are a number of biological mechanisms which have been proposed which would lead to frequency dependent selection:

1. *Specialization on different limiting resources*. If two genotypes eat different foods, then an individual of the rare genotype will have a more abundant source of food, by virtue of the rareness of other individuals who eat that food. The same argument will hold for many other limiting nonfood resources, such as breeding sites.
2. *Different diseases or parasites for different genotypes*. If each genotype has its own diseases and parasites, then whichever type is rarer will be less likely to come into contact with carriers of its own particular pests.
3. *Specialization of different predators on different genotypes*. When each genotype has its own predators, then the genotype which is rare will presumably sustain a lower population density of predators, and hence might suffer a lower mortality rate from predation.

4. *Predator search images: apostatic selection*. Many intelligent visual predators form "search images" of the desired appearance of their prey. They tend to reject potential prey which do not fit this image. The search image depends on the last few prey eaten. Thus the predators may tend to avoid taking the rare genotypes, which they have not encountered recently.

5. *Rare male advantage*. In some species, notably *Drosophila melanogaster*, males of a rare genotype seem to have an advantage in mating simply because they are rare. This pattern of female choice may be an adaptation to avoid inbreeding.
6. *Social Interactions*. In a social species, if the genotypes differ in their social behavior, the fitness of a genotype may depend on the frequencies of the genotypes among the individuals it encounters in the population.

The first four of these mechanisms involve ecological interactions, the last two behavioral interactions. In the first four, the interaction involves, in one way or another, a limiting resource for the population, or the members of the population being themselves a limiting resource for predators. This suggests that it may be difficult for many loci to be simultaneously unders frequency-dependent selection. Even with rare male advantage, if too many loci have their rare alleles being favored, the the preference for unusual males runs the risk of being spread among too many loci, with almost all males being judged as equally unusual.

In many of these scenarios the natural selection would be expected to be density-dependent as well as frequency-dependent. For example, when population density is low, the first mechanism (different resources) will not operate, since individuals of both genotypes will find an abundance of food. When population density is high, the fitnesses will depend on the genotype frequencies. To analyze the outcome of these kinds of frequency-dependent selection requires a model of the specific case, including variables for the numbers of predators or parasites, or the amount of each kind of food resource available. The details of the model will be strongly dependent on the specific biology involved. In this section, we will examine frequency-dependent selection without this biological specificity. We will allow the fitnesses to be arbitrarily chosen functions of the gene frequency, in order to see what types of evolutionary outcome are possible, and what the implications of frequency-dependence are for the mean population fitness.

ASEXUALS AND HAPLOIDS. that we have two genotypes, A and a, with the relative fitness of A depending on the genotype (or gene) frequency, p, in a simple linear fashion. Let

$$\begin{aligned} w_A &= 1 + t - s\,p \\ w_a &= 1. \end{aligned} \tag{II-140}$$

The equations for the evolution of genotype frequencies then become

$$\frac{p'}{1-p'} = (1 + t - s\,p)\,\frac{p}{1-p} \tag{II-141}$$

and

$$p' = \frac{p\,(1 + t - s\,p)}{1 + (t - s\,p)\,p}. \tag{II-142}$$

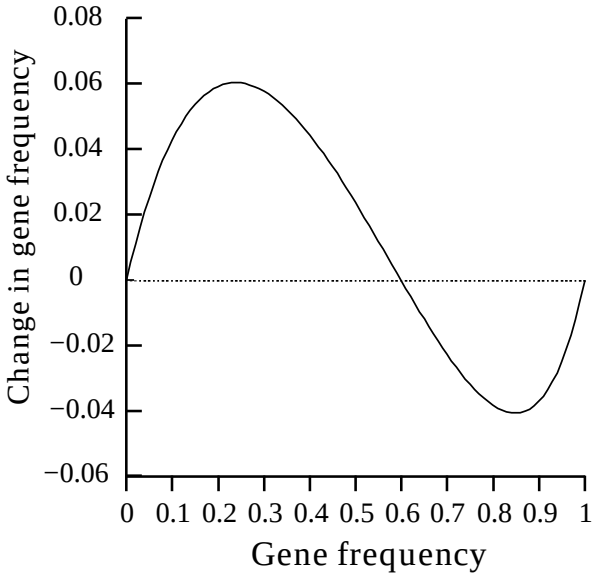


Figure 2.11: Δp as a function of p for a case of frequency-dependent selection. The relative fitness of genotype A is $1.6 - p$.

When we compute the change of gene frequency; it is, from (II-142)

$$\begin{aligned} \Delta p \;&=\; p' - p \;&=\; \big(p\,(1 + t - s\,p) \;-\; p\,[1 + (t - s\,p)\,p]\big) \big/ \big(1 \;+\; (t - s\,p)\,p\big) \\ &= && p(1-p)\,(t - s\,p) \;/\; \big(1 \;+\; (t - s\,p)\,p\big). \end{aligned} \tag{II-143}$$

The equilibria of the genotype frequency are the values of p at which $\Delta p = 0$. For this to occur, either the denominator of (II-143) must be infinite, which is not possible, or the numerator must be zero. The equilibria then occur at $p = 0$, $p = 1$, and $p = t/s$. This last equilibrium will lie in the $[0, 1]$ interval if $s > t > 0$ or if $0 > t > s$. Otherwise w_A will always be greater than (less than) w_a, and although selection will be frequency-dependent, it will nevertheless always be directional selection which leads to the substitution of one genotype for another.

Figure 2.11 shows Δp plotted for particular values of s and t. These values lead the relative fitness of A to be higher when it is rare and lower when a is rare. It would seem on intuitive grounds that this should lead to a stable polymorphism. The graph shows that Δp is positive below the equilibrium point and negative above it. This shows that $p = 0$ and $p = 1$ are both unstable equilibria. The equilibrium $p = t/s$ will be a stable one provided that (by the stability criterion developed above)

$$-2 \;<\; \left[\frac{d(\Delta p)}{dp}\right]_{p\,=\,t/s} \;<\; 0. \tag{II-144}$$

After differentiating (II-143), substituting in $p = t/s$, and doing some tedious collection of terms we find that

$$\left[\frac{d(\Delta p)}{dp}\right]_{p\,=\,t/s} \;=\; -t\left(1-\frac{t}{s}\right). \tag{II-145}$$

The equilibrium is only a relevant one if t/s is between zero and one. If t is negative, the equilibrium is unstable. This corresponds to frequency-dependent fitnesses in which $w_A < w_a$ when A is rare, and the opposite when A is common. It should be obvious that this will lead to an unstable equilibrium at the gene frequency at which $w_A = w_a$, with stable equilibria at $p = 0$ and $p = 1$. When t is positive, the quantity (II-145) will be negative (if $t/s < 1$, which we assume). There is one further restriction on t and s. It makes no sense to have negative fitnesses, so when t and s are positive we must have $1 + t - s > 0$, so that $s - t < 1$. Consideration of the right-hand side of equation (II-145) shows it to be $-(t/s)(s-t)$, so that it will never be below -1 in biologically relevant cases. In this case we always have a stable equilibrium with no overshooting. While it is quite possible that there will be oscillations or chaos in frequency-dependent cases, the particular linear dependence of w_A on p which we have used here has ruled this out.

Does frequency-dependent selection necessarily maximize some measure of the mean fitness? This is easily investigated in the present case. At the polymorphic equilibrium $p = t/s$, $w_A = 1 = w_a$, so that the mean relative fitness is 1, since

$$\bar{w} \;=\; p\,w_A \;+\; (1-p)\,w_a \;=\; p \;+\; (1-p) \;=\; 1. \tag{II-146}$$

The maximum value of $\bar{w}$ can be found by writing

$$\begin{aligned} \bar{w} \;&=\; p\,w_A \;+\; (1-p)\,w_a \\ &=\; p\,(1+t-s\,p) \;+\; (1-p) \\ &=\; 1 \;+\; p(t-s\,p). \end{aligned} \tag{II-147}$$

This is a quadratic function of p which can be maximized by equating its derivative to zero:

$$\frac{d\bar{w}}{dp} \;=\; t - 2s\,p \;=\; 0, \tag{II-148}$$

so that the maximum or minimum occurs at

$$p \;=\; \frac{t}{2s}$$

where

$$\bar{w} \;=\; 1 + \frac{t^2}{4s}. \tag{II-149}$$

If s is positive, which will be the case when we have a stable polymorphic equilibrium, the quadratic has a negative coefficient of p^2 so that the stationary point $p = t/(2s)$ is the maximum. There is thus no correspondence between the polymorphic equilibrium and the value of p which maximizes the mean relative fitness. In fact, maximum occurs at half the equilibrium gene frequency in this case. If the population approaches the equilibrium from above, it will have a continually increasing $\bar{w}$. But if instead it approaches from below, $\bar{w}$ at first increases, then decreases. For the particular example in the Figure, $t = 0.6$ and $s = 1$, so that the equilibrium lies at $p_e = 0.6/1 = 0.6$. The mean fitness there is 1. The maximum mean fitness is achieved at $p_{max} = 0.6/2 = 0.3$, where $\bar{w} = 1.09$.

Why is $\bar{w}$ not maximized? Since w_A is a function of p, the current fitness of the A genotype is not necessarily a good guide to its future fitness. Natural selection increases the frequency of whichever genotype has the higher fitness. In doing so it alters the fitness of A for the worse. Natural selection will maximize mean fitness only if current fitness is a good guide to future fitness.

DIPLOIDS. All of the phenomena which we see in haploids and asexual cases of frequency-dependence also occur in diploids.

When all three genotypes have fitnesses which are arbitrary functions of the gene frequency, there is hardly any limit to the complexity of the behavior of the model. The equations of change of the gene frequency are the usual ones, but now with the fitnesses being functions of p. The equilibria of the model are at the points $p = 0$, $p = 1$, and

$$p_e = \frac{w_{Aa}(p_e) - w_{aa}(p_e)}{[w_{Aa}(p_e) - w_{aa}(p_e)] + [w_{Aa}(p_e) - w_{AA}(p_e)]} \qquad \text{(II-150)}$$

This equation may have many roots, depending on the way in which the w's depend on p. The principle at work here is that in any generation the gene frequency changes according to the momentary fitnesses, so that a polymorphic equilibrium can only occur if the fitnesses at the value of p yield an equilibrium at that value of p. However it is now neither necessary nor sufficient for stability of a polymorphic equilibrium that $s(p_e)$ and $t(p_e)$ be positive at the equilibrium. It is quite possible for there to be underdominance of fitnesses at a stable polymorphic equilibrium!

REFERENCES. One of the early papers on selection, that of Warren (1917), described a pattern of selection whose intensity was frequency-dependent, even though the direction of selection was not. Haldane (1932) discussed the frequency-dependence of an altruistic trait. Wright (Wright and Dobzhansky, 1946) gave a startlingly modern discussion of frequency-dependent selection. Only later was much attention been focused on frequency-dependence. General discussions have been given by Lewontin (1958) and Wright (1969). Some specific models of note have included the competition models of Nei (1971), Mather (1969), Clarke and O'Donald (1964), Cockerham and Burrows (1971), and Cockerham, Burrows, Young, and Prout (1972). Sacks (1967) presents a case in

which selection leads to minimization of mean absolute fitness (which is not the same as minimization of mean relative fitness). The reader will find many further references on specific models of ecological interactions in the review of coevolution models by Slatkin and Maynard Smith (1979).

FREQUENCY-DEPENDENCE AND EVOLUTIONARILY STABLE STRATEGIES. A particularly important source of frequency-dependence is social interactions among unrelated members of a population. This has proven to be one of the best areas of application of Game Theory in biology. In cases where population members engage in ritual fighting and bluffing, as male birds and mammals do for access to mates, one can model the interactions as a game and apply Game Theory. The reward is in fitness. The interesting question is whether the resulting natural selection brings about a solution consistent with Game Theory.

For interactions between unrelated members of a population, where kin selection is not involved, Maynard Smith and Price (1973) introduced the concept of an evolutionarily stable strategy (ESS), which is closely related to the Nash Equilibrium in game theory. Maynard Smith (1974, 1982) developed the theory further. Maynard Smith (1981) and Eshel (1982) showed that genetic models in such cases reach the same conclusions as ESS methods. Nowak (1990) has criticized this conclusion as overly simple, and argued that the models must be analyzed in each case, in order to show whether an ESS will be achieved.

II.13 Kin selection: a case of frequency-dependence

.

One class of examples of frequency-dependent selection which has attracted wide attention is kin selection, in view of its usefulness as an explanation for the evolution of social behavior. Haldane (1932) pointed out that an altruistic behavior, one which benefited the recipient but was disadvantageous to the donor, would be selected against within populations, even though the existence of the trait benefited the population as a whole. Haldane proposed that subdivision of the species into groups could result in increase of the trait, if selection against the trait within groups were counterbalanced by selection for it by differential increase of those groups having the highest frequencies of the trait. Haldane's mechanism is often referred to as group selection. While it *is* group selection, it is also an example of a scheme proposed by Hamilton (1963, 1964a, 1964b), known as kin selection.

In kin selection the impact of the selection falls not only on the individual but on others who happen to be relatives. In the case of altruistic behavior these others are the recipients of the behavior. If they are kin, they have some chance of also carrying the alleles which caused the behavior in the original individual. If the increase in the fitness

of the kin as a result of the behavior is great enough and their relationship to the individual close enough, the resulting increase in the frequency of the alleles is enough to more than counterbalance the selection against these alleles in the individual displaying the altruistic behavior. Alternatively, the whole process may be viewed from the point of view of individual selection. The alleles predisposing an individual towards the altruistic behavior have a net advantage because they also predispose its kin towards that behavior. Thus they bring about a loss of fitness by causing an individual to engage in the behavior, but a compensating gain in fitness by causing the individual to be surrounded by altruistic relatives. This is the "personal fitness" approach to intuiting the effects of kin selection (Hamilton, 1964a, b; Orlove, 1979). The other approach Hamilton (1964a, b) involves computing an "inclusive fitness" which involves the effect of a gene on the fitness of its bearer, plus a fraction of its effect on the fitness of each relative.

HAMILTON'S RULE. Hamilton's rule is that an allele that incurs a cost (in fitness) c on its bearers and also confers a total benefit b on a set of individuals related to it, whose average coefficient of relatedness with the individual is r, will increase in the population if

$$c < r\,b \tag{II-151}$$

The coefficient of relatedness is the probability that a copy drawn at random from the one individual is identical by descent to one of the copies in the other individual. (Note that it is not IBD to a *random* copy drawn from the second individual, but to some copy). It is also important to understand that the quantity b is not the benefit to one of the beneficiaries, but the sum of benefits to all of them.

Hamilton's Rule is sensible, if we consider a rare allele which is acting in heterozygote. The behavior reduces the fitness of the actor by c, thus losing c copies of the allele. But if it benefits enough recipients to increase their total of their fitnesses by b, and if a fraction r of these recipients also contain copies of this allele, then $r\,b$ copies who would otherwise be lost are saved, at the cost of c copies lost in the altruist. If Hamilton's formula holds, there is then a net gain of copies.

This heuristic argument is forceful but not entirely convincing. Below we will see that the rule can be derived more rigorously in the particular case of a model of pairwise interactions.

KIN AND GROUP SELECTION. In the example used by Haldane (1932), different groups were assumed to have different frequencies of the gene for the altruistic behavior. This implies that each group contains individuals who are more related than average, so that the increase of groups containing large numbers of altruists is as a result of the benefit from each altruist tending to be conferred on its relatives. In such a case, kin and group selection are the same phenomena, as pointed out by Price (1970, 1972a). If one requires of group selection that it involve the mortality of whole groups, then one might

not want to call this group selection if the survival of a group is simply a consequence of the survival of individuals.

A MODEL OF PAIRWISE INTERACTION. The computation of inclusive fitnesses or of personal fitnesses is a shorthand for more complete modeling which discloses more precisely what is going on. Done properly, these heuristic methods provide a valuable tool, but it must be understood that they are summaries of a more detailed account that is done by conventional methods. Inclusive fitness is carried aloft by the humdrum of population genetic modelling: it cannot fly on its own.

Hamilton (1971) has given a model of pairwise interaction which specifies more clearly what is behind notions of inclusive fitness. The model presented below is an altered version of his model. It is in no sense a canonical model of kin selection, but only one of the simplest cases.

We consider a rare allele, A, in a diploid population, and ask for the conditions of its increase when rare. Because of this rareness and because the population is outbred, the AA genotype will be so rare that we ignore it (which amounts of ignoring terms of order p^2). Each generation, the individuals in the population are assumed to associate in pairs, the two members of which play different roles. These are not mating pairs: the two individuals dissociate before random mating ensues. In each associated pair of individuals, there may or may not be a certain social interaction (perhaps an altruistic behavior by the first individual). For the four possible ordered pairs: (*Aa, Aa*), (*Aa, aa*), (*aa, Aa*), (*aa, aa*) there will be probabilities c_{11}, c_{10}, c_{01}, and c_{00} that this behavior occurs. If it does occur, the fitness of the individuals is respectively $1+s$ and $1+t$. If it does not, their fitnesses are both 1. Thus the fitness of an *Aa* individual which is the first individual in the first sort of pair is

$$(1-c_{11})\times 1 \;+\; c_{11}\times(1+s) \;=\; 1 \;+\; c_{11}\,s. \qquad \text{(II-152)}$$

This pattern of selection is frequency-dependent because the overall fitness of an *Aa* individual depends not only on the frequency with which it assumes a certain role, but the identity of its partner, which will depend on the genotype frequency in the population. We are interested in the conditions for increase of the A allele when it is rare. In that case the frequency of *Aa* heterozygotes in the population will be $2p(1-p)\simeq 2p$, and the frequency of *aa* homozygotes will be $(1-p)^2\simeq 1-2p$. If each individual's role and partner were assigned at random, then the probability that a pair would be (*Aa, Aa*) would be $(2p)^2$. We are not going to make this assumption, but instead we will assume that the members of a pair tend to be kin. In particular, the probability that an *Aa* has a partner which also carries the rare A allele will be taken to be r, and we will examine the effects of different values of r. If this occurs as a result of kinship between the two individuals, r will be computed using the probabilities of identity by descent defined below in Chapter V. It is a quantity first defined by Sewall Wright (1922), known as the *coefficient of relationship.* Here are some values for various relatives:

Relative	r
self	1
full sib	1/2
parent	1/2
child	1/2
half-sib	1/4
aunt/uncle	1/4
niece/nephew	1/4
grandparent	1/4
grandchild	1/4

The model is summarized in Table 2.8. The frequencies of the four types of pair are set by the definition of r and the requirement that the overall frequency of *Aa* be $2p$.

Whether the A allele increases in frequency will be determined by the relative fitnesses of *Aa* and *aa*. A moment's consideration will show that all but an infinitesimal fraction of the *aa* individuals will find themselves in (*aa, aa*) pairs. The average fitness of *aa* will then be

$$\begin{aligned} \bar{w}_{aa} &= 1/2\,(1 + c_{00}\,s) \;+\; 1/2\,(1 + c_{00}\,t) \\ &= 1 + c_{00}\,(s+t)/2. \end{aligned} \tag{II-153}$$

Of the $2p$ of the population which are *Aa*, this can be divided into $2pr/2$ individuals playing the first role in an (*Aa, Aa*) pair, $2pr/2$ playing the second role in such a pair, $2p(1-r)/2$ playing the first role in an (*Aa, aa*) pair, and $2p(1-r)/2$ playing the second

Table 2.8: The pairwise interaction model.

	Probability of interaction	Frequency of this pair	Fitness	
			of first partner	of second partner
(Aa, Aa)	c_{11}	$2p\,r$	$1 + c_{11}\,s$	$1 + c_{11}\,t$
(Aa, aa)	c_{10}	$2p\,(1-r)$	$1 + c_{10}\,s$	$1 + c_{10}\,t$
(aa, Aa)	c_{01}	$2p\,(1-r)$	$1 + c_{01}\,s$	$1 + c_{01}\,t$
(aa, aa)	c_{00}	$1 - 4p + 2p\,r$	$1 + c_{00}\,s$	$1 + c_{00}\,t$

role in an (*aa, Aa*) pair. The mean fitness of *Aa* is then

$$\begin{aligned} \bar{w}_{Aa} &= [\, p\,r\,(1+c_{11}\,s) \;+\; p\,r\,(1+c_{11}\,t) \;+\; p\,(1-r)\,(1+c_{10}\,s) \\ &\qquad\qquad + \; p\,(1-r)\,(1+c_{01}\,t)\,] \;/\; (2p) \\ &= 1 \;+\; r\,c_{11}\,(s+t)/2 \;+\; (1-r)\,c_{10}\,s/2 \;+\; (1-r)\,c_{01}\,t/2. \end{aligned} \tag{II-154}$$

These fitnesses do not contain the gene frequency p, which would seem to give the lie to the assertion that the fitnesses are frequency-dependent. A more careful derivation would show that w_{AA} contains additional terms in p, but that these can be ignored since we are only interested in cases in which p is very small. The quantities $2p$ and $1-2p$ are, as we have seen, also approximations ignoring terms in p^2.

Allele A will increase if $w_{Aa} > w_{aa}$, or (discarding the constant 1 and a factor of $1/2$)

$$r\,c_{11}\,(s+t) \;+\; (1-r)\,c_{10}\,s \;+\; (1-r)\,c_{01}\,t \;>\; c_{00}\,(s+t), \tag{II-155}$$

which is easily rearranged as

$$r\,[c_{11}\,(s+t) \;-\; c_{10}\,s \;-\; c_{01}\,t] \;>\; c_{00}\,(s+t) \;-\; c_{10}\,s \;-\; c_{01}\,t. \tag{II-156}$$

Some cases. We are now in a position to look at some particular cases of interest:

1. *Altruistic behavior.* Suppose that the occurrence of a behavior in the pair depends only on the genotype of the first individual, and the behavior is deleterious to that individual and advantageous to its partner. Then $c_{11} = c_{10}$, and $c_{01} = c_{00}$, $s < 0$ and $t > 0$ so that (II-156) becomes

   $$r\,(c_{11} - c_{01})\,t \;>\; (c_{00} - c_{10})\,s. \tag{II-157}$$

   If allele A makes the behavior more likely, then $c_{11} > c_{01}$. We already know that $c_{11} - c_{01} = c_{10} - c_{00}$ so that (II-157) becomes

   $$r \;>\; (-s)/t. \tag{II-158}$$

   This is precisely Hamilton's basic result. It shows that allele predisposing toward the behavior will spread if the partners are sufficiently close kin. The greater the benefit the less closely they need be related. The greater the loss to the individual performing the behavior the more closely they need be related.

2. *Mutualism.* If the behavior is equally beneficial to both members of the pair, then $s = t > 0$. Then the inequality (II-156) becomes, after cancellation of s and t:

   $$r\;(2\,c_{11} - c_{10} - c_{01}) \;>\; (2\,c_{00} - c_{10} - c_{01}) \tag{II-159}$$

If the more A alleles a pair has the more likely it is to engage in mutualistic behavior, $c_{11} > c_{10}$ and $c_{01} > c_{00}$ so that the expression on the left of (II-159) is always positive and that on the right always negative. It is satisfied for all values of r (since r cannot, by its definition as a conditional probability, be negative).

The result is straightforward: an allele predisposing toward a mutualistic behavior will always spread, although consideration of the magnitudes of w_{Aa} and w_{aa} will show that its spread will be faster the greater is r.

3. *Complementary behaviors.* If the two genotypes are predisposed toward different behaviors which are complementary (e.g. in a cooperative hunting behavior, one tends to chase prey, the other to wait in ambush for the prey being chased), we would expect that $c_{11} < c_{10}$ and $c_{01} > c_{00}$, with $s = t > 0$. Inequality (II-159) then becomes

$$r < \frac{c_{10} + c_{01} - 2\,c_{00}}{c_{10} + c_{01} - 2\,c_{11}}. \qquad \text{(II-160)}$$

The allele A can spread only if r is sufficiently small, if the individuals are not too closely related! If $c_{00} < c_{11}$ there is no restriction on r, since the right-hand side of (II-160) is greater than 1. If $c_{00} \geq c_{11}$ the limit on r lies between 0 and 1 and is a relevant limit. Consideration of the magnitudes of w_{Aa} and w_{aa} shows that the increase of allele A is more rapid the lower is r.

4. *Narrow selfishness.* If the behavior is advantageous to the individual but harmful to the other member of the pair, then $s > 0$ and $t < 0$. In that case the condition is $r < s/(-t)$, which implies that if the harm done to the partner is greater than the gain to the individual, the partner ought not be too closely related, otherwise the loss of alleles in the partner will more than counterbalance their gain in the individual that shows the behavior.

The preceding ignores effects of the homozygote AA and is limited to consideration of the fate of the A allele when it is rare. As the allele becomes common, the approximations we have made, ignoring terms in p in the fitnesses, become invalid. The frequency-dependence of the fitness becomes important, and terms involving the effects of AA individuals enter as well. Some insight can be gained by considering the change of the frequency of a when it is rare, using the above approximation.

It must be borne in mind that this is but one possible model of social interactions, and that it is limited by its many assumptions (e.g., one pairwise interaction per generation, between individuals of the same generation). There is no single canonical model of the evolution of social behavior: this one is useful primarily for its simplicity.

PITFALLS. In making models of the evolution of social behavior, there are a number of traps into which it is easy to fall:

1. *Requiring Conscious Recognition of Kin.* In the above model, as in virtually every other kin selection model, there is no requirement that individuals who are kin be actively recognized from among the rest of the population. It need only be the case that the association of individuals into an interacting pair be such that the result is that the pairs have average coefficient of relationship r. The association could be based on geographical proximity, with nearby individuals being closer relatives than faraway individuals. The theory of kin selection does not necessarily presume mental or physical adaptations for recognizing kin from nonkin.

2. *Ignoring Mutualism.* Kin selection of an altruistic trait is one of the ways of explaining the spread of a behavior which is deleterious to the individual expressing it, but advantageous to others. There may be many social behaviors which are advantageous to both individuals. For these the mutualism mechanism mentioned above is a viable explanation. As we have seen, in the evolution of a mutualistic trait there is no strict requirement that the interacting individuals be close kin. Textbooks on the evolution of animal behavior commonly ignore or understress mutualism as an explanation for the evolution of social behaviors. This springs partly from a fascination with the paradoxes inherent in altruism, and partly from an ideological preference for "nature red in tooth and claw." (However, mutualism does have the weakness that there can easily be natural selection for one partner to cheat on the other, and this can lead in turn for natural selection for various means of deterrence or retaliation).

3. *Ignoring Cultural Inheritance.* In all of the above arguments, the trait spreads by the differential death and reproduction of individuals. As some animals (in particular, primates) evolved an ability to learn and to communicate learned information, it became possible for "cultural" information to survive and be transmitted. Most of the information we humans possess is culturally transmitted. The explanation of the existence and spread of a human behavior is not necessarily genetic variation or genetic transmission. Cultural information is evaluated subjectively by humans, rather than objectively by their survival and reproduction. It does not necessarily "mutate" in random directions, but can be consciously altered so as to solve a problem. It can spread laterally within a generation, and information from different sources can be chosen and recombined. As such, "cultural evolution" is capable of enormously greater speed of change than is genetic evolution. The amount of information transmitted culturally is enormous. This includes, among others, most of the behavior affected by libraries, universities, mass media, government, business, and religion (and it includes this book). The amount of recognizably culturally transmitted variations in human behavior is so great, compared to the amount of recognizably genetically transmitted variations in human behavior, that cultural transmission is a natural null hypothesis for any human behavior. To ignore it as a

possible explanation of specific changes of human behavior is silly, though this is often done by biologists who have a preference for genetic determinism of human behavior.

REFERENCES. After the fundamental work of Haldane and of Hamilton just mentioned, there has grown up a population genetic literature which has been in part dedicated to verifying Hamilton's inclusive fitness principle in specific, well analyzed genetic models. Among the models of this genre are those of Levitt (1975), Matessi and Jayakar (1976), Charlesworth (1978a), Wade (1978), and Cavalli-Sforza and Feldman (1978b). These are all single-gene models with various assumptions. A polygenic model is presented by Yokoyama and Felsenstein (1978). The mutualism mechanism for evolution of social behavior was advocated by Hamilton (1964a), Lin and Michener (1972) and West-Eberhard (1975), though not in terms of a quantitative model. Engels (1983) considered the effects of evolution of the cost/benefit ratio considered as a quantitative character.

A controversy of particular interest as an illustration of the usefulness of explicit population genetic models is that on conflict of parent and offspring, between Trivers (1972, 1974) and Alexander (1974). The quantitative model of Charlesworth (1978a) provided substantial support for Alexander's position that parents would win the conflict on a evolutionary scale.

Group selection can be brought into the same framework, as was shown by Crow and Aoki (1982), who showed that Hamilton's condition also applied to it. Group selection acts when this condition holds. Thus the same equation covers more than one level of selection. There is some recent controversy over the connection between group and kin selection, and whether group selection is of primary importance. A good entry into that literature is the review by Kerr, Godfrey-Smith, and Feldman (2004).

Recently, a dramatic controversy erupted when Nowak, Tarnita, and Wilson (2010) argued that group selection, rather then kin selection, explained the evolution of eusociality. This led to a number of strongly-worded replies, one (Abbot et al., 2011) signed by 137 authors. D. S. Wilson and E. O. Wilson, who are not related to each other, had previously (2007) argued that group selection should replace kin selection as a general explanation for the evolution of social behaviors. This argument is all the more surprising given the central role that E. O. Wilson played in popularizing kin selection, particularly in his 1975 book *Sociobiology*. The argument continues, with most researchers on the evolution of behavior continuing to defend kin selection and reject its replacement by group selection.

Exercises

1. Suppose that we have a haploid population with two alleles, and their absolute fitnesses are $W_A = 4$ and $W_a = 2$. If the initial frequency of A is 0.001, what will it

be after 20 generations?

2. In a haploid or asexual population with continuous reproduction in which a individuals die instantly after birth, what are the values of r_A and r_a? What do these imply about the change of gene frequencies?

3. In a haploid system with two alleles, A and a, with fitnesses $1+s:1$, how long will it take to change the frequency of A from 0.1 to 0.2 if $s = 0.01$? How long will it take to change the frequency of A from 0.9 to 0.8 if $s = -0.01$? Explain why these numbers are or are not the same.

4. How large must the selection coefficient favoring a dominant allele be in order to change it from a gene frequency of 0.5 to 0.51 in one generation? Compute this exactly. Compare the result with the proper approximation.

5. Suppose that we have a locus with 3 alleles in a haploid organism, and the three alleles have relative fitnesses $1.5 : 1.2 : 1$. If the initial gene frequencies (p_1, p_2, and p_3) are 0.01, 0.1, and 0.89, what will the gene frequencies be after 5, 10, 15, and 20 generations?

   (How to do it.: *Don't* just crank out the frequencies using a computer. Consider the ratio p_1/p_2. Can you derive a formula for how it changes in one generation of selection? You can just work out formulas for p_1' and for p_2', and take their ratio, which will show a helpful cancellation. Show the work. From this figure out what happens in 5 generations of selection Do the same for p_2/p_3. That should enable you to get the ratios of gene frequencies in each of those generations. From that, it is easy to work out the gene frequencies. Explain how you did that.)

   Someone argues that for this kind of selection, the gene frequency of each allele should either be continually rising or continually falling, because it will either be favorable or unfavorable. Is that true? Why?

6. In consideration of mutation versus selection, one type of selection we will consider is selection against a partially dominant allele. Let's just consider one generation of selection (with no mutation). Suppose we have three genotypes AA, Aa, and aa, with an initial gene frequency of a of 0.0001. Starting at Hardy-Weinberg proportions, if the viabilities of the genotypes are $1 : 0.99 : 0.1$, then ...

   ... of all of the deaths from this selection, what fraction of all individuals are *aa* individuals who die? What fraction are heterozygotes who die? Then what fraction of all deaths are in heterozygotes? (This problem does not require any algebra other than Hardy-Weinberg proportions, just calculation).

7. For a diploid population with absolute fitness $3 : 4 : 2$ of genotypes AA, Aa and aa, compute $\bar{W}$ as a function of gene frequency.

Find the maximum of this curve, and compare it to the equilibrium predicted from the relative fitnesses. Is mean absolute fitness being maximized?

8. What is the segregational load in a system of two balanced lethal alleles (i.e., a situation where both homozygotes are lethal, so only the heterozygotes survive)?

9. Suppose three genotypes A_1A_1, A_1A_2, and A_2A_2 have fitnesses 4, 0, and 3. What is the equilibrium gene frequency? Is it stable? Why can't we just use the formulas for the fitnesses $1-s \ : \ 1 \ : \ 1-t$?

10. Find all equilibria for the following three-allele case:

genotype	fitness
A_1A_1	4
A_1A_2	0
A_1A_3	5
A_2A_2	3
A_2A_3	5
A_3A_3	2

What are the mean fitnesses at these equilibria? What does this imply about their stability?

11. Suppose that we have three genotypes *AA*, *Aa*, and *aa* in a sexual population (where an absolute fitness of 1 denotes exactly enough offspring to replace the population). The fitnesses depend on population density (N) in the following way:

$$\begin{aligned} W_{AA} &= 2/\left(1+0.004N\right), \\ W_{Aa} &= 1.9/\left(1+0.003N\right), \\ W_{aa} &= 1.8/\left(1+0.002N\right), \end{aligned}$$

What will be the ultimate fate of the gene frequency if both alleles are initially present in the population? *(Hint: first compute what would be the equilibrium population density for a completely asexual clone of each genotype if it were present alone).*

12. Suppose that exactly once every ten years a haploid desert plant experiences a wet year. If genotype *A* has, relative to *a*, fitness 2 during wet years and 0.92 during dry ones,

    (i) what is the arithmetic mean relative fitness of *A?* The geometric mean relative fitness?

    (ii) what will happen to the frequency of *A* over the long run?

13. Suppose that in a diploid plant for which there is one chance in three that each year will be wet, with an independent chance each year, the fitnesses of genotypes are:

	AA	Aa	aa
Wet	1	2	3
Dry	1	0.8	0.6

What will happen to the frequency of A if we watch it for many years: what kind of behavior do we expect? Do we expect it to approach an equilibrium frequency?

14. An insect species has three genotypes at a locus AA, Aa, and aa. In the spring, the probability that individuals of these three genotypes survive until summer are 0.9, 0.8, and 0.5. If they do, then in the summer these survivors have probability of surviving until fall of 0.6, 0.7, and 0.8. After that they mate randomly and reproduce, the offspring going into hibernation until spring. There are no fertility differences of the genotypes.

(i) Start with gene frequency p of A, and derive an equation for the gene frequency in the next generation (p') in terms of these viabilities and p.

(ii) Suppose that after surviving the spring (or not), the insects mate randomly at the start of the summer, and the summer viabilities are for that new summer generation (which in turn then mate randomly in the fall, just as the previous case did). Derive the equations for the change of gene frequencies over the course of a year.

Are these two cases exactly equivalent? Check that by starting with some gene frequency and seeing whether they reach exactly the same gene frequency in the fall.

15. In case of a haploid frequency-dependent selection, suppose that the relative fitness of A is $3.5 - 3p$. What are the equilibrium gene frequencies of A? Which ones are or are not stable?

16. Suppose that we have a diploid frequency-dependent case of the following sort:

$$\begin{array}{cc} AA & 1/(1/2+p)^2 \\ Aa & 1/(1/2+p) \\ aa & 1 \end{array}$$

What are the equilibria of such a system? Can you say anything about their stability? What happens to A when it is rare?

Complements/Problems

1. J. B. S. Haldane preferred to work with the variable $u = \ln(p_A/p_a)$, instead of the gene frequencies or gene frequency ratios. Obtain the equation for u_t in terms of the fitnesses and u_0 in a haploid or asexual case. What attracted Haldane to this quantity?

2. Extend the approximation for change under selection in the additive model $1+2s : 1+s : 1$ by one more term, to terms in s^2. Solve the resulting differential equation. How does the result compare with the exact solution for the case $s = -1/2$?

3. Why aren't the multiplicative and dominant cases the same in Figure 2.2 when p is near 1?

4. For $s = -1$ in the case of a recessive lethal gene, obtain from the exact treatment in II.6 the equation for the number of generations it takes to change gene frequency from p_0 to p_t. Compare this to the continuous approximation formula for $s = -1$.

5. Can you obtain a set of equations similar to Section II.4 for diploids? Be sure to check your equations by trying to predict gene frequency changes in the case of a recessive lethal which dies as soon as it is born (so $d_{aa} = \infty$). Do your equations appear to work, or do they predict that the gene frequency of a will go to zero instantly? What is the main difficulty in setting up these equations? (Think).

6. Is the equilibrium $p = 1$ stable when the fitnesses of $AA : Aa : aa$ are $1 : 1 : 1-s$? When p is displaced *above* 1? Mathematically stable? Biologically stable? *Hint: you will need to consider terms that are quadratic in the departure from the equilibrium gene frequency.*

7. For a haploid population with continuous reproduction define a meaningful mean fitness. Obtain an equation for its value in an arbitrary generation t, given the initial gene frequency and the values of b_A, d_A, b_a, d_a in a two-allele case. Can the mean relative fitness ever decrease?

8. Suppose that in one multi-allele haploid population the relative fitnesses are $w_1 : w_2 : \cdots : w_n$, and in another they are $w_1^2 : w_2^2 : \ldots w_n^2$. Compare the change in gene frequencies and in fitness in the first population in two generations to their changes in the other population in one generation. In the case of weak selection, what does this tell us about the effects of doubling the selection coefficients? For this one it will help to use a haploid version of (II-122) and construct ratios of gene frequencies and ask how they change.

9. A researcher in experimental evolution has three haploid genotypes of yeast, which we call A_1, A_2, and A_3. She wants to estimate their fitnesses, which are called W_1, W_2, and W_3. She takes a mixture of strains of frequencies f_1, f_2, and f_3 (these of course add to 1).

   After growing them in culture for one generation the frequencies are g_1, g_2, and g_3 (which add to 1). They strains don't mate and form diploids, they just reproduce asexually.

   (i) What are the equations for these in terms of the f_i and the W_i? Also what happens to ratios between the frequencies of two genotypes?

   (ii) Do they allow us to make an estimate of the values of the absolute fitnesses, the W_i? Of the relative fitnesses?

   (iii) Show the calculations for estimating fitnesses if the f_i are 0.1, 0.3, and 0.6 and the g_i turn out to be 0.18, 0.3, and 0.52.

10. The *gametophytic* system of self-incompatibility in plants has the property that there is a multi-allele self-incompatibility locus, and if pollen falls on a plant that has alleles (say) A_1A_2 only pollen which is neither A_1 nor A_2 can fertilize the ovules. Assume that there is a three-allele gametophytic self-incompatibility system. Assume that all ovules get fertilized – there is never a shortage of pollen. So an A_1A_2 plant gets all its ovules fertilized by A_3 pollen. What are the equations for the change of frequencies, from one generation to the next, of the three possible genotypes (all are heterozygotes as homozygotes cannot form)? Do some numerical calculations for a few generations, for a case where one allele starts out rare. What do you think will happen if a fourth allele occurs by mutation?

11. In the above case, if we start out with two of the allele frequencies (say the frequencies of A_1 and A_2) equal to each other, what will happen to those gene frequencies? (A little staring at the equations should disclose the answer). If we start out with alleles A_1 through A_3 equal in frequency, but a small frequency of A_4, can you derive equations for the change of A_4?

12. Some plants reproduce by obligate self-fertilization, so that every offspring is the result of a random pollen grain and a random ovule from the same plant. Suppose that we have a locus with two alleles, A and a, in such a completely self-fertilizing plant. What are the equations for change of the three genotype frequencies (note that one cannot assume Hardy-Weinberg proportions so that we have to follow the three genotype frequencies)? If there is an overdominant locus with fitnesses $1-s, 1$, and $1-s$ what are the equations for the change of genotype frequencies from one generation to the next when one observes the genotype frequencies immediately after self-fertilization but before selection has had time to act? How

large a value of s is needed to prevent the heterozygotes from disappearing from the population?

13. For a sex-linked overdominant lethal whose fitnesses are:

females:			males:	
AA	Aa	aa	A	a
1	$1+h$	0	1	0

work out the equations for the change of gene frequencies. What are the equilibrium gene frequencies? For what values of h does a polymorphic equilibrium exist? (It will help to compute the genotype and gene frequencies in the newborns, immediately before the selection has had time to act).

14. Prove that in a two-allele diploid population the segregational load cannot be greater than $1/2$. Prove that in an n-allele diploid population it cannot be greater than $(n-1)/n$.

15. Prove that in two-allele diploid overdominant selection, gene frequency changes will never overshoot the equilibrium.

16. For a two-allele case, prove that $\bar{w}$ is never higher after selection and mating than it is immediately after selection but before mating.

17. Use the principle that $\bar{w}' \geq \bar{w}$ to prove that (in cases of multiple alleles) a minimum or a saddle point in the $\bar{w}$ surface cannot be a stable equilibrium.

18. If $\bar{w}$ is maximized with all alleles $A_1, \ldots, A_k$ present, prove that there is no stable equilibrium which has all but one of these alleles present.

19. *Meiotic drive* is a situation where, in a heterozygote, instead of A and a gametes being produced in equal numbers, there are more of one allele (say $\frac{1}{2}(1+\alpha)$ are A and $\frac{1}{2}(1-\alpha)$ are a). Although most cases of meiotic drive have this happening in one sex only, the mathematics is simpler if we assume it happens in both sexes. For that case, derive a formula for the change of the gene frequency of A as a function of its gene frequency and α. Does it look like the formula for the change of gene frequency by a simple form of natural selection? What plays the role of the selection coefficient? In what way is the formula not exactly similar to the one for selection?

20. Suppose that in a diploid species, among gametes produced by females and among gametes produced by males there is haploid selection with different fitnesses, so that among female gametes the fitnesses are A and a are $w_f : 1$ and among male

gametes the fitnesses are $w_m : 1$. What are the conditions for A to increase when rare? Are there conditions on w_f and w_m that allow maintenance of polymorphism by this selection, when the resulting diploid genotypes otherwise have no differences in fitness? *Be careful – this case will not maintain Hardy-Weinberg proportions. Nevertheless, you can follow just one variable, the gene frequency of* A *just before selection acts. Why?* For theory relevant to this see Gregorius (1982).

21. Suppose that a species has environmental sex determination, with 45% of individuals developing into females and 55% into males. Suppose that a rare allele A changes that to a fraction F becoming females and a fraction $1 - F$ becoming males. Be careful to think about whether there will be Hardy-Weinberg proportions. For what values of F will allele A increase in frequency when rare? (The verbal argument which this model validates is usually attributed to R. A. Fisher in his 1930 book, but that theory of sex ratio was actually first given by Carl Düsing in 1883 and 1884 and by Darwin in the first edition of *The Descent of Man*).

22. Suppose that in a haploid population the relative fitness of A (compared to a) is in alternate generations $1 + s$ and $1 - s$. Which allele will increase in frequency?

23. Suppose that in a diploid population, fitnesses vary randomly and independently each generation, being (in a two-allele case) for $AA : Aa : aa$ $1 + s : 1 : 1 - s$ half of the time, and $1 - s : 1 : 1 + s$ the other half of the time. What will happen to gene frequencies? What will be the difference in behavior between cases with different values of s?

24. What will happen in a diploid population if fitnesses are half of the time $1 : 1 : 1 + s$ and half of the time $1 + s : 1 : 1$? If these occur cyclically, in alternate generations? If they occur randomly, drawn independently in each generation? How does this behavior differ from having fitnesses $(1 + s)^{1/2} : 1 : (1 + s)^{1/2}$ all of the time?

25. Suppose that in a diploid population with temporal variation in fitnesses, a recessive allele a has relative fitness $1 + s_t$ in generation t. What are the conditions on the selection coefficients s_t such that allele a will increase in the long run when it is rare? (This case was treated by Haldane and Jayakar, 1963).

26. Suppose that we have the following case of haploid frequency-dependent selection: every generation, a constant fraction f of the individuals are discarded by natural selection, and fierce competition ensures that the individuals dying are never of the competitively superior of two genotypes (A) as long as there are individuals of the other genotype (a) available. Derive equations for change in gene frequency. Do these correspond to your intuition as to what the results of selection ought to be?

27. Suppose that an asexual or haploid population has frequency dependent selection with $w_A/w_a = [p/(1-p)]^B$. What is the behavior of the model for different values of B? *(Hint - work in terms of $p/(1-p)$ and take logs).*

28. Suppose that a diploid population has two resources available to eat. Suppose that each individual specializes on one or the other resource. All AA and half of the Aa's can eat only resource #1, all aa's and half of the Aa's will eat only resource #2. Suppose that there are a total of N individuals in all, N_1 of whom specialize on resource 1, N_2 on resource 2. Suppose that if there are N_1 specialists on resource 1, the fraction of survivors among them is given by $1/(1+0.001N_1)$, and that is $1/(1+0.001N_2)$ for the N_2 specialists on resource 2. This is a form of frequency-dependent selection (note that N_1 and N_2 are functions of the frequency of the A allele). What are the equilibrium points of this system? Are they stable?

29. Suppose that a tasty butterfly species has two color patterns, controlled by one locus with a dominant allele B that makes a brown spot on the wing, and a recessive allele b that has a red spot there instead, when the genotype is bb. The butterflies are preyed on by birds, who form a "search image" which is based on the most recent butterfly of this species that they have eaten: if it was brown-spotted, they will be eager to eat the next brown-spotted butterfly they see, if it was red-spotted, when they see a brown-spotted butterfly, they will eat it one 1/3 of the time. The reverse is true too: a bird that has a red-spotted search image will eat a brown-spotted butterfly only 1/3 of the time.

    (i) If each bird eats many butterflies and a fraction F of all butterflies of that species encountered have a brown spot, what will the fraction of birds that come to have a brown-spotted search image? *(Hint: Let Q be the fraction of birds with a Brown search image, and find an expression for Q', the fraction that will have a Brown search image after they have seen, and possibly eaten, the next butterfly. Be sure to include the cases of birds that have Brown and Red search images, see brown-spotted or red-spotted butterflies, and do or do not eat them. The frequencies of the two phenotypes of butterfly that they encounter depend on the current gene frequency p. Then equate Q' to Q and solve for Q.)* We're going to be assuming that each bird sees many butterflies during the course of a single butterfly generation, so that most bird/butterfly encounters will occur after the frequencies of the two search images in the birds has settled down to this equilibrium.

    (ii) Once you have Q (as a function of p), assume that each butterfly is seen by exactly one bird in its lifetime, a bird which has probability Q of having the Brown search image (and $1-Q$ of having the Red search image). What are the fitnesses of the two butterfly phenotypes, as a function of p? What are the fitnesses of the three butterfly genotypes (yes, this is easy)?

(iii) To what gene frequency p does this form of frequency-dependent natural selection lead? At that gene frequency, what are the fitnesses of the two phenotypes?

30. Suppose that in a haploid population with two alleles, each individual occupies a burrow with another chosen at random. Let the fitnesses of individuals depend on their genotypes and that of their burrow-partners:

An individual of type	whose partner is	has fitness
A	A	w_{AA}
A	a	w_{Aa}
a	A	w_{aA}
a	a	w_{aa}

What are the equations for change of gene frequency? Where are the equilibria of this system? Are these formulas the same as for overdominance?

Chapter III

MUTATION

III.1 Introduction

Natural selection is the evolutionary force responsible for the progressive adaptational aspects of evolution - for the fact that organisms are as good as they are at surviving and reproducing. If this were all there were to population genetics, it would be a dull subject indeed. The independent existence of population genetics as a field (as contrasted with evolutionary studies in general) comes from the interaction of the genetic system with natural selection. The mating system and the mechanism of recombination distribute genetic material in particular patterns which affect the rates and directions of responses to evolutionary forces. As we shall see in future chapters, the fact that populations are spread out in space affects the mating system so that migration may be considered as an evolutionary force in its own right. The very finiteness of natural populations introduces yet another force, called random genetic drift, which will be treated in chapters V, VI, and VII. In the present chapter we treat mutation. These three evolutionary forces are not responsible for creating the adaptive information content of living organisms. Rather, they set the context within which natural selection takes place, and to some extent they interfere with its operation.

Among these forces, mutation has a unique role. In a sense, it is a destructive force, making random changes in the genetic material. In any highly adapted organism such changes are overwhelmingly likely to be detrimental. The usual analogies we make in such cases involve making random adjustments in a finely constructed watch, or making random alterations of a carefully-written poem. While one will occasionally improve the timing of the watch or the effectiveness of the poem by random changes, with much greater probability one will make things worse. Migration may have a somewhat similar effect, in moving organisms into regions to whose environments they are ill-adapted. Genetic drift, which changes gene frequencies at random, may cause a favored allele to be lost. Yet mutation holds a special place among these, for without it the whole process

of evolution would grind to a halt. For natural selection favoring genotype AA over Aa and aa in the absence of any mutation will soon cause the gene frequency of A to reach unity. At that point, the population has lost genetic variability at this locus. If at some future time the fitness of aa were to rise (as a result of environmental changes) to exceed that of AA there would be no way to reverse the gene substitution. In the very act of altering of A alleles into a alleles, mutation both erodes contemporary adaptation and creates the variability which is the basis of future adaptation.

This suggests that there is some need for mutation, that there might be some natural selection favoring its existence in a species. But we do not need to explain its existence, for mutation is a thermodynamic inevitability. There can be little question that natural selection has acted to reduce rates of mutation. The very existence of a system of precise genetic replication testifies to this, as mutation has its evolutionary effect as incorrect replication. The question which remains to be answered by population geneticists is whether there are limits set by natural selection to lowering the mutation rate. Would a population having a very low mutation rate evolve to have a higher one? Or is the mutation rate as low as selection can make it, awaiting only genetic variability (paradoxically - awaiting the mutations) for further reduction of the amount of mutation.

The answer to this puzzle is not known. Existing models of selection for mutation rates are too crude, and too little is known about the availability of genetic variability which might allow a decrease in mutation rates.

III.2 Effect of Mutation on Gene Frequencies

One of the nicer aspects of mutation is that the mathematics of its effects on gene frequencies are very simple. The main complications come from the model of mutation itself. When the genetic scheme is simple, everything else comes easily.

TWO ALLELES.

The simplest possible mutational scheme has only two alleles. This is intended literally: there is imagined to be only one site at which the two alleles can differ, and only two possible nucleotides at this site. We can denote the two possibilities by A and a and the two types of mutational event which are possible by $A \rightarrow a$ and $a \rightarrow A$. This is obviously a wildly oversimplified model of mutation in a gene, but there is a large class of circumstances in which it is a reasonable approximation to reality. Often we may be considering a gene with a large number (say 500) of nucleotide sites, but we can only detect two phenotypically different proteins, those that are active as enzymes and those that are not. Thus mutation is in effect moving the gene back and forth between two different categories of nucleotide sequences: those which form active enzyme and those which do not. Of course, we are going to assume that all sequences in the A category have equal probabilities of mutating to sequences in the a category, and similarly for the

a sequences. This is at best only approximately true: some inactive sequences may be many base pairs removed from the "nearest" active sequence, while others may be able to mutate to an active sequence by several different routes, each involving only one base pair change.

These considerations aside, the mathematics is a straightforward exercise in elementary probability. In an infinite random-mating diploid population with discrete nonoverlapping generations, suppose that the gene frequency of A is p. We have two possible mutational events, $A \rightarrow a$ and $a \rightarrow A$. The rate of mutation for the first sort of event will be u, and for the second sort of event v. Keep in mind that the probability of each of these events is calculated *per copy of that allele*. A fraction u of all copies of A change into a each generation, and a fraction v of all copies of a change into A.

Suppose that the current gene frequency of A is p. In the next generation, the genes which are A will come from two sources. Some are copies of genes which were A in the last generation and which did not mutate to a's. Since p of the genes are A in this generation, the fraction of all copies in the offspring generation that are unmutated A's will be $p(1-u)$. The other source of A copies are genes which are mutated copies of a's. In the offspring generation there are expected to be a fraction $(1-p)v$ of these. So

$$p' \;=\; p\,(1-u) \;+\; (1-p)\,v. \tag{III-1}$$

One can immediately see one characteristic of gene frequency change by mutation: it is going to be *very* slow. Typical values of mutation rates for a single gene (summing over all sites able to mutate so as to inactivate a gene) are 10^{-7}. This means that $1-u$ will very nearly be 1, and $(1-p)v$ will be very small. So p will change little from one generation to the next. This point is made more clearly by computing the change in p from one generation to the next:

$$\Delta p \;=\; p' - p \;=\; -u\,p \;+\; v\,(1-p). \tag{III-2}$$

Note that every term on the right-hand side of (III-2) has a u or a v in it, so that the whole right-hand side will be very small (in fact, it can be no larger than the larger of u and v).

The direction of change contains a pattern of change which will be evident from (III-2). When all genes are A, so that $p = 1, \quad \Delta p \;=\; -u$. This reflects the obvious fact that when all genes are A the gene frequency will decrease in the next generation by the fraction of them which mutate to a. Likewise when $p = 0$ there are only a's to mutate to A's, and the gene frequency increases by $\Delta p \;=\; v$. So the frequency of A decreases when large (albeit by a very small amount) and increases (by a similarly small amount) when small. In between somewhere lies an equilibrium.

The equilibrium point is easily found by using (III-2) to inquire when $\Delta p \;=\; 0$, or by using (III-1) to ask when $p' = p$. Either way, the result is

$$p_e \;=\; \frac{v}{u+v} \tag{III-3}$$

At this mutational equilibrium the numbers of A's being converted by mutation into a's equals the number of a's being converted into A's.

APPROACH TO EQUILIBRIUM.

The rate at which the population approaches this state is easily found from (III-1) owing to the linear form of that equation. From (III-1), the equilibrium frequency must satisfy

$$p_e \;=\; (1-u)\; p_e \;+\; v\,(1-p_e) \tag{III-4}$$

and subtracting (III-4) from (III-1)

$$\begin{aligned} p' - p_e \;&=\; (1-u)\; p \;-\; (1-u)\; p_e \;+\; v\,(1-p) \;-\; v(1 \;-\; p_e) \\ &=\; (1-u)\,(p-p_e) \;+\; v\,(1 \;-\; p \;-\; 1 \;+\; p_e) \\ &=\; (1 \;-\; u \;-\; v)\,(p \;-\; p_e). \end{aligned} \tag{III-5}$$

Thus the deviation of p from its mutational equilibrium value p_e is multiplied by $(1-u-v)$ every generation. This is a number very near 1: a typical value might be 0.9999998. The distance from the equilibrium will decline very slowly, at the rate at which the powers of $1-u-v$ decline.

We can get some sense of exactly how slow is the approach to equilibrium by making an approximation. Since u and v are both very small,

$$1-u-v \;\approx\; e^{-(u+v)}, \tag{III-6}$$

the error in this approximation being terms in u^2, v^2 or uv, all of which may be safely ignored. After t generations, the original departure of the gene frequency from the mutational equilibrium will have been multiplied by $(1-u-v)^t$, which is nearly the same (by III-6) as $e^{-(u+v)t}$. If we ask at what generation the gene frequency will have moved half of the way to the equilibrium, this will be given by solving for t in

$$e^{-(u+v)\,t} \;=\; 0.5, \tag{III-7}$$

the solution to which is

$$t_{0.5} \;=\; -\frac{\ln 0.5}{u+v} \;=\; \frac{0.693147}{u+v}. \tag{III-8}$$

As a rough order of magnitude estimate, we can say that it takes about $1/(u+v)$ generations to move a substantial fraction of the way to the mutational equilibrium. Figure 3.1 illustrates this. It shows the whole course of approach of two populations, one started at $p=1$ and the other at $p=0$, to mutational equilibrium. Note the horizontal time scale, which is in *millions* of generations. All of which emphasizes just how weak a force mutation is, how slowly it will change gene frequencies. We will see the implications of this shortly. Figure 3.1 demonstrates the slowness of the approach to mutational equilibrium.

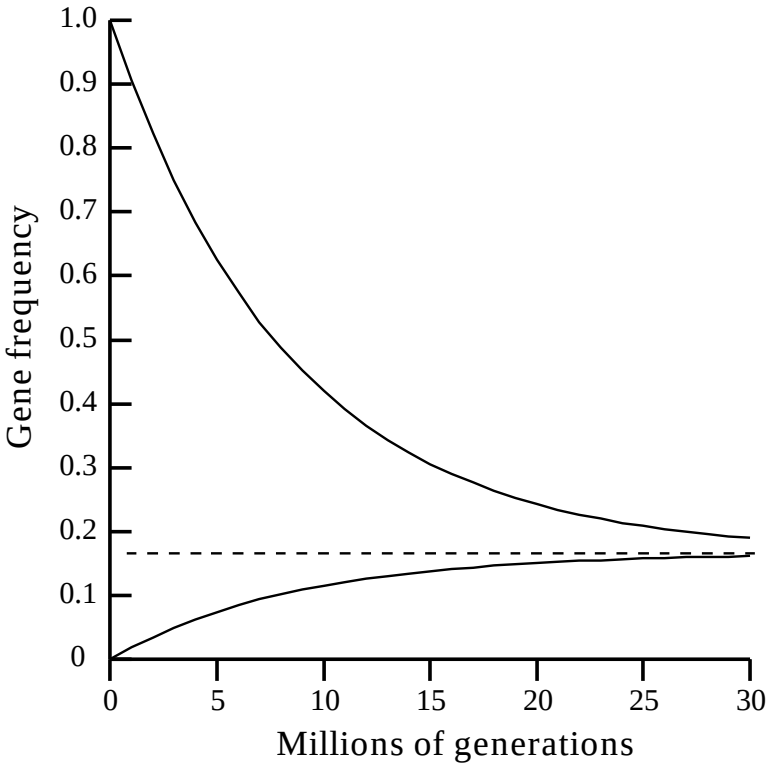


Figure 3.1: Approach of gene frequency to equilibrium in a two-allele case starting from fixation at either allele when $u = 5v$ with $u = 10^{-7}$. Note the large number of generations on the horizontal time scale.

III.3 Mutation with Multiple Alleles

FORWARD AND BACK MUTATION. In the above analysis, we did not comment on the relative sizes of u and v. There are good reasons for believing that u will commonly be many times larger than v. Usually we will denote the functional enzyme as allele A, and the nonfunctional enzyme as allele a. In that case, u is the rate of forward mutation and v the rate of back mutation. Underlying the fiction of two alleles, there is a reality of a very large number of possible base sequences, giving rise to a smaller, but still astronomical, number of possible protein sequences. Of these, only a tiny fraction could be functional enzymes (or structural proteins). While most changes in a functional sequence may inactivate it, few changes in a nonfunctioning sequence will restore it to function, particularly if it is a sequence far removed from the nearest functioning sequence.

In these circumstances u will be far larger than v. In fact, it is often a reasonable

approximation to let v be zero. We then have unidirectional mutation $A \rightarrow a$. Equation (III-1) is then simply

$$p' \;=\; (1-u)\; p \tag{III-9}$$

which predicts a mutational equilibrium at a zero frequency of A, and a slow approach to this equilibrium, it taking about $1/u$ generations to move a substantial fraction of the way to the equilibrium.

MULTIPLE ALLELES. For a more complete consideration of such a situation, we would have to consider mutation back and forth among a large (a *very* large) number of possible alleles. Suppose that there are n alleles $A_1, A_2, \ldots, A_n$, and that the frequency of the i-th allele is given by p_i. Let u_{ij} give the frequency of mutation of an A_i allele into an A_j allele. A simple counting-up of the possible origins of an A_i allele will give the equations of change of gene frequencies:

$$p_i' \;=\; p_i \left(1 - \sum_{j \neq i} u_{ij}\right) \;+\; \sum_{j \neq i} p_j u_{ji}, \qquad i \;=\; 1, 2, \ldots, n \tag{III-10}$$

the prime indicating the next generation. At the equilibrium, $p_i' = p_i$, so that this gives

$$0 \;=\; -p_i \sum_{j \neq i} u_{ij} \;+\; \sum_{j \neq i} p_j u_{ji} \qquad i \;=\; 1, 2, \ldots, n. \tag{III-11}$$

This is a set of linear equations in the p_i which can be solved for the mutational equilibrium gene frequencies once the u_{ij} are known, and keeping in mind that the p_i must sum to unity. The general formulas (which will be in matrix form) are not particularly enlightening, but in certain cases the results become simple. When all of the mutation rates u_{ij} are assumed to be equal, the equilibrium can readily be shown to be the situation

$$p_i \;=\; 1/n, \qquad i \;=\; 1, 2, \ldots, n \tag{III-12}$$

in which all alleles are equally frequent. The rate of approach to this equilibrium will depend on the total rate of mutation. Suppose that we define μ as the total rate of mutation away from one allele. Then $\sum_{j \neq i} u_{ij} \;=\; \mu$. It turns out that $\frac{n}{n-1}\mu$ is the fraction of the distance toward the equilibrium that will be covered each generation, so that it will take about $1/\mu$ generations to go a substantial fraction of the way toward equilibrium.

This situation of total symmetry is not a particularly good model of mutation at a protein locus. In a sequence with 500 sites, there are only 1500 among the 4^{500} possible base sequences which can be reached by single point mutations.

Nevertheless, if the rate of mutation at each site is equal, and if the three possible base changes which can occur have equal rates u (so that the probability that a C at a given site will mutate to a G is u, and the probabilities that it will mutate to A or T are also both u, for a total mutation rate of $3u$), the equilibrium can be demonstrated

to be the situation in which all 4^{500} base sequences occur with equal frequencies. This number is far greater than the number of elementary particles in the known universe! We can safely say that there is no population that has ever had that genetic composition. The rate at which convergence to this mutational equilibrium occurs is not so simple to discover, but will be the total mutation rate of the gene $3Su$, where S is the number of sites in the gene.

A DISTINCTION. The prospect of a population which has 4^{500} different alleles segregating at equal frequencies raises the issue of how we are to regard mutation. Is it a deterministic or a random force in evolution? In the mathematics above, mutation has appeared as a deterministic force, pushing gene frequencies slowly toward a set equilibrium gene frequency. In that view, mutation is not a force which will bring about different results in different populations. Although it will often act to increase the genetic variability within a population by reintroducing alleles which have become lost, it will have the same effect in all populations and in all generations (assuming given mutation rates).

But a population with 4^{500} equally frequent alleles is an impossibility in practice, as it would require a population size of at least 10^{300}. In an actual population at mutational equilibrium, only a tiny fraction of all possible alleles would be present. Two different populations at mutational equilibrium will contain different mixtures of DNA sequences. The same population followed through time will vary in gene frequencies and in the identities of the alleles present. Doesn't this mean that mutation is actually a random force acting to diversify populations? It does not. In such a situation the diversity is the result of the finiteness of the population. This is an effect of random genetic drift, which is an evolutionary force we will study in chapters V to VII. That the randomness occurs through finiteness of the population is easily seen by a thought experiment. Consider the mutational equilibrium in a series of populations of size N. The larger is N, the more chance that the same mutant alleles will be in existence in different populations. The amount of diversity between populations generated by mutation depends most critically on their size. By moving the population towards a state in which there would be a great number of alleles, mutation allows genetic drift to have a dramatic effect by eliminating most of the possible alleles from the population. It is genetic drift, not mutation, that is the random force. We shall see in Chapter VII a model which allows us to approximate how many common alleles we expect to see in a population of finite size at mutational equilibrium.

The above point may seem to be only a semantic distinction, but it is important to have a correct intuitive understanding of evolutionary forces, and confusing a deterministic force with a random one is a matter of no small consequence.

III.4 Mutation versus Selection: Haploids

All the above discussion has assumed that the genotypes are equally fit. Much of our interest in the phenomenon of mutation stems from situations where the genotypes created by mutation are less fit than normal genotypes. Mutation may be causing inactivation of a functional protein. As we saw in the previous section, if there were no selection, the population would move toward an equilibrium in which functional alleles would be the exception rather than the rule. This would seem to pose a problem for the continued existence of the organism, but the very differences in fitness which seem to threaten the extinction of the population will also act to keep the functional alleles at high frequency. This two-edged effect of selection results in a surprising cancellation of its two effects, a cancellation known as the Haldane-Muller principle, which we shall discuss shortly.

First let us show the effects of selection when it acts in opposition to the effects of mutation. The simplest case in which we can investigate this is a haploid population with discrete generations and two alleles. We assume that one of the two alleles, A, produces a functional protein, and the other, a, a nonfunctional protein. Here is the life cycle:

$$\begin{array}{ccccccccc} \text{Haploid} & \xrightarrow{\textit{Selection}} & \text{Haploid} & \xrightarrow{\textit{Mating}} & \text{Diploids} & \xrightarrow{\textit{Meiosis}} & \text{Haploid} & \xrightarrow{\textit{Mutation}} & \text{Haploid} \\ \text{newborns} & & \text{adults} & & & & \text{gametes} & & \text{Newborns} \end{array}$$

Suppose that the frequency of the A allele is p among newborns. If the fitnesses of the two haploid genotypes A and a are respectively 1 and $1 - s$, then after selection

$$p^* \;=\; \frac{p}{1 \;-\; (1-p)\,s} \tag{III-13}$$

which is a version of equation (II-16) in the previous chapter.

This is a haploid organism, which we are treating as if it were asexual, since mating and meiosis will have no effect on the gene frequency. So among "gametes" the gene frequency will still be p^*. Mutation does have an effect, one which will depend on the rates of forward and back mutation. In this case, we are primarily concerned with forward mutation $A \rightarrow a$. Let us take its rate to be u, setting the back mutation rate v to zero. As we shall see, the results will be little affected by whether back mutation is present or not. The effect of forward mutation on the gene frequency will be simple. Equation (III-1) will give us the gene frequency of A after mutation in terms of that before mutation:

$$p' \;=\; p^*(1-u). \tag{III-14}$$

Putting (III-13) and (III-14) together by using the former to substitute for p^* in the latter, we get the recursion formula for gene frequency from one generation to the next:

$$p' \;=\; \frac{p\,(1-u)}{1 \;-\; s\,(1-p)} \tag{III-15}$$

We want to find the equilibrium of the population under the two forces of mutation and selection. This is most readily done by setting $p' = p$ in (III-15) and solving for p. Multiplying (III-15) by the denominator of its right-hand side after removing primes, we obtain

$$p - s\,p(1-p) \;=\; p\,(1-u),$$

or

$$u\,p - s\,p(1-p) \;=\; 0,$$

which is

$$[u - s(1-p)]\,p \;=\; 0. \tag{III-16}$$

The system will be in equilibrium if $p = 0$, which is simply the situation where the normal allele has been lost. This will be an equilibrium because we have not allowed for back mutation. This is not the equilibrium in which we are most interested. That is

$$u \;=\; s\,(1-p). \tag{III-17}$$

Since we may prefer to follow the frequency of the nonfunctional mutants, we can replace their frequency $(1-p)$ by q and see that the mutant alleles at equilibrium have frequency

$$q_e \;=\; u/s. \tag{III-18}$$

We will discuss this simple result below, because it turns out that it can also be obtained in a simple diploid case, and many of the organisms we are interested in are diploids.

III.5 Mutation versus selection: Diploids

The haploid results can easily be recycled to analyze the diploid case in which there are multiplicate (geometric) fitnesses. Suppose that the life cycle is

$$\begin{array}{ccccccccc} \text{Diploid} & \xrightarrow{\textit{Selection}} & \text{Diploid} & \xrightarrow{\textit{Meiosis}} & \text{Haploid} & \xrightarrow{\textit{Mutation}} & \text{Haploid} & \xrightarrow{\textit{Random union}} & \text{Diploid} \\ \text{Newborns} & & \text{Adults} & & \text{Gametes} & & \text{Gametes} & & \text{Newborns} \end{array}$$

and the fitnesses of genotypes AA, Aa, and aa are the geometric series $1 : 1-s : (1-s)^2$. Then it is easy to show that equation (III-13) in the previous section gives the gene frequency p^* after selection. Mating and meiosis do not change the gene frequencies. Mutation changes them in the same way that it does for haploids, since mutation acts on single copies of the allele. Thus equation (III-13) also applies, and the mathematics of the change of gene frequencies is exactly the same as in the haploid case. We get the equilibrium gene frequency u/s (as in III-17) for the deleterious mutant allele.

IMPLICATIONS. This is a fairly simple result. It shows us immediately that the outcome of the interaction of mutation and natural selection is given by the ratio of their coefficients u and s. In many cases, we will be able to assume that $s \gg u$. In particular, we are often interested in specific loci at which there are mutations that cause a fairly drastic change in the phenotype. Since values of u are likely to be so small, in almost all such cases it is hard to imagine that s is not many orders of magnitude greater than u. Whenever a phenotypic difference is large enough for us to see, it is hard to imagine that it is not so large that $s > 10^{-6}$. The exception is protein and DNA sequence data. Molecular methods enable us to discern differences which may be so slight as to have little or no selection acting on them. Only in those cases is it a reasonable expectation that s and u are of the same order of magnitude. Even if s is not much greater than u, there is only a small range of values of s for which the equilibrium frequency of mutants is not small. If $s \gg u$, q_e is small. If $s < u$, q_e is predicted by (III-18) to be greater than 1. This is a strange result, to say the least. A closer examination of (III-15) will show that in such a case it is always true that $p' > p$, unless $p = 0$. So when $s < u$, the mutation to a is always a stronger force than the selection which opposes it, and the mutant becomes fixed in the population. The equilibrium with $p = 0$ is then the relevant one.

Except for a small range of values of s, we expect either that selection has little influence or that it is far stronger a force than mutation, and holds the equilibrium gene frequency of the mutant allele to a very low value. If the latter is the case, then back mutation will be a force of little consequence. There will be few a genes available to mutate back into A, even if the back mutation rate were as large as the forward mutation rate. For example, when $u = v = 10^{-7}$ and $s = 0.001$, $q_e = 0.0001$. A crude examination shows that, each generation, a fraction $u(1 - q) = 0.9999 \times 10^{-7}$ of all copies mutate from A to a. A fraction $uq = 10^{-11}$ mutate from a to A. By far the strongest effect raising the gene frequency of A is selection, which in this case causes the death (or reproductive failure) of $sq = 10^{-7}$ of the genes. A more careful analysis of the case of back mutation will lead to a quadratic equation for q_e instead of (III-17), and will give support to the practice used here of ignoring back mutation.

A useful way of intuiting (III-18) is to note that rare mutant a has, in effect, a risk s of being eliminated each generation. This leads to the prediction that each mutant will remain in the population an average of $1/s$ generations (this is the average number of tosses of a coin with probability s of Heads until Heads finally occurs). The population should contain as many copies of a as accumulated by mutation during the last $1/s$ generations. Since about a fraction u of mutants arises each generation, we should expect that $q_e = u \times 1/s = u/s$. This argument ignores a number of terms, but those terms are small and the resulting error is small. In fact, the resulting error is zero since the various approximations used happen to cancel each other's effects!

III.6 Mutation vs. Selection: Effects of Dominance

It is natural to wonder whether diploidy, dominance and recessiveness in particular, alter this picture in major ways. With diploidy the mathematics becomes slightly messier but is still not very difficult if we are willing to make certain approximations.

RECESSIVE MUTANTS. If the mutant alleles are completely recessive, an exact result is still possible. Once again the gene frequency among newborns will be taken to be p. First we need to know how much the gene frequency is changed by selection. The fitnesses are taken to be

$$\begin{array}{ccc} AA & Aa & aa \\ 1 & 1 & 1-s. \end{array}$$

This does not fit easily into the scheme of section II.6 unless we exchange p for q and change the sign of s. Rather than do that, we can fall back on the general formula for gene frequency change in diploids, equation (II-31). When the above fitnesses are substituted in, we get for the gene frequency after selection

$$p^* \;=\; \frac{p\,(p\times 1+(1-p)\times 1)}{p^2\times 1+2p(1-p)\times 1+(1-p)^2\times(1-s)} \tag{III-19}$$

and after a little algebra in the denominator, this gives

$$p^* \;=\; \frac{p}{1-s\,(1-p)^2} \tag{III-20}$$

We do not really need to know the genotype frequencies after selection, since all we are interested in is the effect selection and mutation will have on the gene frequencies. Mutation occurs to genes one at a time, without substantial regard to the identity of their homologue. As a result, we can follow the effect of mutation on gene frequencies without knowing how those gene frequencies are organized into genotypes. This was the basis of section III.2 above, where we derived the mutational equilibrium without in any way using the fact that the population was diploid. The same equations for mutational effects on gene frequencies hold in diploids as in haploids.

Equation (III-20) shows the effect of selection on the frequency of the "normal" allele A. If mutation is taken to be unidirectional $A \to a$ at a rate u, its effect will simply be to multiply p^* by $(1-u)$, as before. Then we will get for the gene frequency after mutation,

$$p' \;=\; \frac{p\,(1-u)}{1-s\,(1-p)^2} \tag{III-21}$$

Equating p' to p and solving for possible equilibrium values of p, we find that either $p_e = 0$, or

$$1-s(1-p_e)^2 \;=\; 1-u \tag{III-22}$$

which gives

$$(1 - p_e)^2 \;=\; u/s \tag{III-23}$$

so that the frequency q_e of the mutant allele a at equilibrium is

$$q_e \;=\; 1 - p_e \;=\; \sqrt{u/s}. \tag{III-24}$$

As in the haploid case, if $s < u$ the equilibrium frequency of the mutant exceeds 1. This is simply the situation where selection is so weak that it is never able to stem the increases in gene frequency caused by mutation, and in this case the other equilibrium $q_e = 1 - p_e = 1$ is the relevant one.

The gene frequency of the mutant allele is higher (for the same values of u and s) in the recessive case than in the haploid case. This can be seen by using the values of u and s from the numerical example in the previous section. If $u = 10^{-7}$ and $s = 10^{-3}$, we obtain $q_e = 0.01$, which is 100 times higher than in the haploid case. Note that a rather small mutation rate has resulted in a far higher gene frequency at equilibrium.

That this result is a reasonable one is seen by making a more intuitive argument along the same lines as in the haploid case. Each mutant has a probability sq_e of being eliminated in each generation. To be eliminated by natural selection, it must occur in a homozygote (an event of probability q_e given that we already know that the one gamete carries the mutant), and natural selection must kill (or sterilize) the resulting homozygote, an event with probability s. So the average mutant will persist in the population for $1/(sq_e)$ generations. The population will then contain $1/(sq_e)$ generations worth of mutants. Since a fraction u of the genes mutate each generation, the total frequency of mutants will be roughly

$$q_e \;=\; u \times 1/(s\,q_e), \tag{III-25}$$

We can solve this for q_e. When we do, we get exactly the result (III-23). It is remarkable that an imprecise argument such as this happens to give us exactly the correct result. It contains a number of approximations, such as the assertion that there are u mutants each generation when in fact the number is closer to $u(1 - q_e)$, as we cannot mutate copies which are already a. Apparently the different approximations we have made in this intuitive argument just happen to cancel each other.

Note from (III-24) that the equilibrium frequency of the mutant is far higher in the recessive case than in the haploid case, given that we compare cases with equal values of u and s. This is primarily due to the weakness of selection in the recessive cases. A mutant can only be eliminated by selection if it is in the company of another mutant. The fraction of mutants eliminated each generation is sq_e rather than s. With $u = 10^{-7}$ and $s = 0.001$, this means that in the haploid case 0.001 of all mutants are eliminated each generation, while in the recessive case the corresponding number is $0.001 \times 0.01 \;=\; 10^{-5}$, so that a given mutant will remain in the population 100 times longer. This slower rate

of loss of mutants raises their frequency in the population by a large factor (100 times in this case). This in turn compensates for the greater difficulty of forming the "affected" phenotype, which now only appears in homozygotes. The result is that in both cases we see exactly the same frequency, u/s, of affected individuals. We shall see later a remarkable consequence of this fact.

TURNOVER OF ALLELES. These calculations give us a picture of how rapidly the pool of deleterious mutant alleles at a locus "turns over". In the case of partial dominance, a fraction hs of the deleterious mutants are eliminated in each generation. They are replaced by new deleterious mutants. Thus the pool of deleterious mutants turns over on average every $1/(hs)$ generations. For a mutant with $h = 0.02$ and $s = 1$ this is 50 generations, or in humans about 1,250 years. At a locus with recessive deleterious mutations, the fraction of copies of them that are eliminated in each generation is sq_e, which is $\sqrt{u\,s}$. For a locus with recessive lethal mutants which has mutation rate 10^{-6}, a fraction 0.001 of the mutants are thus eliminated each generation, and the pool turns over on average every 1,000 generations, or in humans 25,000 years.

In chapter VII we will see that a simple calculation also allows us to predict how many different mutational events contributed to the pool of deleterious mutations at one locus in a finite population. Except in small populations, the pool will be surprisingly diverse.

RATE OF APPROACH TO EQUILIBRIUM. This intuitive argument also tells us much about the rate of approach of the system to equilibrium. If we were to start with no mutants, and were to wait until the system were near its equilibrium, then since the new mutants at equilibrium constitute a fraction sq_e of the mutant pool, we would in effect be waiting for there to be at least $1/(sq_e)$ generations of mutation. Otherwise it would be impossible for enough to have occurred. Let us call this number of generations G. In a few multiples of G generations, almost all of the existing mutants are eliminated by selection, and enough new mutants occur to replace them. So G is a natural time scale for the equilibrium of gene frequencies, because it tells us about how many generations are needed for selection to obliterate the history of the process. It takes about G generations for the pool of mutants present to "turn over" about once.

This also allows us to get a rough idea of how rapidly a population will respond to changes of mutation rate. If mutation were to suddenly cease, it would take a few multiples of G generations for mutants to disappear from the population. On the other hand, if mutation rates were suddenly to be doubled, it would take essentially the same length of time for the mutant gene frequencies to approach their new equilibrium frequencies, which in the recessive case will be $\sqrt{2}$ times their current equilibrium frequencies.

EFFECT OF BACK MUTATION. In all of the above, back mutation has been ignored. It is possible to incorporate it into the analysis, by changing (III-21) so as to replace

$p^*(1-u)$ by $p^*(1-u)+(1-p^*)v$, where p^* is given by (III-20). The result is a quadratic equation for the equilibrium gene frequency. When this is solved, it is found that, unless the equilibrium frequency of the mutant allele is large, the presence of back mutation makes hardly any difference to the equilibrium gene frequency of the mutant allele. An intuitive rationale for this is easily constructed. Aside from our expectation that back mutation rates will be smaller than forward mutation rates, the very rareness of *a* alleles makes back mutation an infrequent phenomenon. At equilibrium in the absence of back mutation, we may approximate by saying that a fraction u of genes mutate from A to a, and about an equal fraction of all genes are a's which are killed off by selection. So the decrease in frequency of a by selection (a decrease balanced by its increase from mutation) is about u. (Not fraction u of the a copies, but an absolute decrease of u in the gene frequency of *a).* Back mutation will also decrease the frequency of a, but by an absolute amount vq, that is, by converting to A a fraction v of the a copies. This will be an insignificant change in the frequency of a compared to the changes by selection or forward mutation. If the forward mutation rate is 10^{-7} and the back mutation rate is 10^{-8}, with an equilibrium gene frequency of 10^{-2}, forward mutation increases the gene frequency of a each generation by 10^{-7} (and selection decreases it by about the same amount), while the change due to back mutation is only a reduction by $10^{-8} \times 0.01 = 10^{-10}$, three orders of magnitude smaller.

A COMPUTATIONAL EXAMPLE. It seems at least a reasonable approximation to ignore back mutation in these calculations. The reader who is skeptical may wish to state and solve the quadratic equations for p, and to see how much difference back mutation makes in the equilibrium gene frequency of the mutant allele a.

The utility of the calculations of this section can be seen by consideration of the disease cystic fibrosis. This is a recessive disease, which until very recently was almost always fatal before reproductive age. The disease has an incidence at birth of about 1 in 2,500. If Hardy-Weinberg proportions are assumed to hold in the newborns (which will be the case if there was random mating among their parents), the gene frequency for the cystic fibrosis allele is 0.02. One hypothesis that can be made to explain the frequency of cystic fibrosis is that the alleles are introduced by mutation and are currently in a mutation-selection balance. If the selection is only on the affected homozygote and amounts to complete lethality (s = 1), then equation (III-23) shows us that $0.02 = \sqrt{u}$, so that the mutation rate would have to be 0.0004 per gene per generation. This is almost 1000 times higher than the admittedly imperfect estimates available to us of mutation rates per cistron. This renders it unlikely that we can explain the prevalence of cystic fibrosis as the equilibrium under a balance between mutation and selection. Either there is some other pattern of natural selection (perhaps heterozygote superiority) or the situation is not an equilibrium.

DOMINANCE. When the mutant allele is partially or completely dominant, exact

algebraic solution is not so simple. The same sort of equations as before can be used, but the counterpart to (III-21) now yields a quadratic equation of q_e. While its solution is not difficult, interpretation of the resulting formula is. Consequently we will limit ourselves to approximate treatment of this case, since the approximations are quite good ones.

One case of partial dominance can be solved exactly: we have seen that the equations for multiplicate (geometric) fitnesses are the same as those for haploids, and the equilibrium frequency can be obtained exactly in the case of mutation from A to a. It is convenient to have this one diploid case which can be exactly solved, for this allows us to compare the amount of selection which occurs in homozygotes and heterozygotes. When the frequency of the mutant allele is q, the fraction of all genes which are a's killed off (or sterilized) by selection in heterozygotes is $q(1-q)s$, keeping in mind that only half of the genes in heterozygotes are a. The fraction of all genes killed off as a's in homozygotes is $q^2[1-(1-s)^2]$ or $q^2(2s-s^2)$. The ratio of these two mortalities is $(1-q)/[q(2-s)]$, which will be somewhere between $(1-q)/q$ and $(1-q)/2q$, depending on s. Since q is expected to be small at equilibrium, we can conclude that far more a copies are killed off in heterozygotes than in homozygotes, simply because a rare allele occurs far more frequently in heterozygotes than in homozygotes.

Partial dominance. Now we can apply this to construct an approximate argument good for a wide range of dominance patterns. Suppose that the fitnesses of our three genotypes are

AA	Aa	aa
1	$1-hs$	$1-s$

Since the gene frequency of a will be small at equilibrium (at least, it will be for small u and larger s), we will be hard-pressed to distinguish between this pattern of selection and the multiplicative or geometric pattern

AA	Aa	aa
1	$1-hs$	$(1-hs)^2$.

In both cases the selection against heterozygotes is hs, while homozygous aa genotypes will account for hardly any of the mortality of a genes, as this genotype will be very rare. So the equilibrium gene frequency should be quite close to the multiplicative value, which would be

$$q_e = \frac{u}{hs}. \tag{III-26}$$

The intuitive interpretation of this result is straightforward. Since a mutant allele is exposed each generation to a probability hs of elimination (except for the very unlikely possibility that it will occur in a homozygote), it will be expected to remain in the population for $1/(hs)$ generations, and the current mutant frequency should be the fraction of genes which mutate during that time, $u/(hs)$.

Table 3.1: Exact and approximate equilibrium gene frequencies with partial dominance, $u = 10^{-7}$ and $s = 0.01$.

h	q_e $(approx.)$	q_e $(exact)$
0.001	0.01	0.0027
0.002	0.005	0.00232
0.005	0.002	0.001534
0.01	0.001	0.0009175
0.02	0.0005	0.0004885
0.05	0.0002	0.0001993
0.10	0.0001	0.0000999
0.20	5.0×10^{-5}	5.00392×10^{-5}
0.50	2.0×10^{-5}	2.0×10^{-5}
0.75	1.33×10^{-5}	1.33179×10^{-5}
1.00	1.0×10^{-5}	1.00583×10^{-5}

From the way we have obtained the result, it is expected to be accurate only when q_e is small, i.e. when $u \ll hs$. In fact, a more complete consideration of the solution of the full quadratic equation which we would get by using (II-33) together with (III-1) verifies our intuition (the equation would be cubic if we allowed back mutation as well). Here (in Table 3.1 and Figure 3.2) is a comparison of some values for $s = 0.01$ and $u = 10^{-7}$. We may use these exact solutions to verify the unimportance of selection against mutant homozygotes. When $h = 0.2$, we have $q_e = 5.00392 \times 10^{-5}$. The frequency of heterozygotes in the population among newborns will be $2(0.99995)(0.00005) = 0.0001$. Mutant homozygotes will be much rarer, being only 2.5×10^{-9} of all individuals. The fraction of genes being lost as a result of selection against heterozygotes will be $0.2 \times 0.01 \times 0.0001 = 2 \times 10^{-7}$. The fraction of genes eliminated as a result of selection against homozygotes is $0.01 \times 2.5 \times 10^{-9} = 2.5 \times 10^{-11}$. Even taking into account the fact that two mutant alleles are lost when a homozygote dies but only one is lost when a heterozygote dies, the loss of heterozygotes is by far the more severe effect on mutant gene frequencies. This helps justify our approach, which is based on more or less ignoring the selection which occurs in mutant homozygotes.

Note that the approximation $u/(hs)$ is quite good, even for as little dominance as $h = 0.01$, although it begins to degrade below that value. Even a very slight selection against heterozygotes will have more impact than a much stronger selection against the much rarer homozygotes. While this may seem a perfectly straightforward result, it is less obvious when we look at rare disorders in human populations, for it tells us that medically trivial effects in heterozygotes are likely to have more impact on gene frequen-

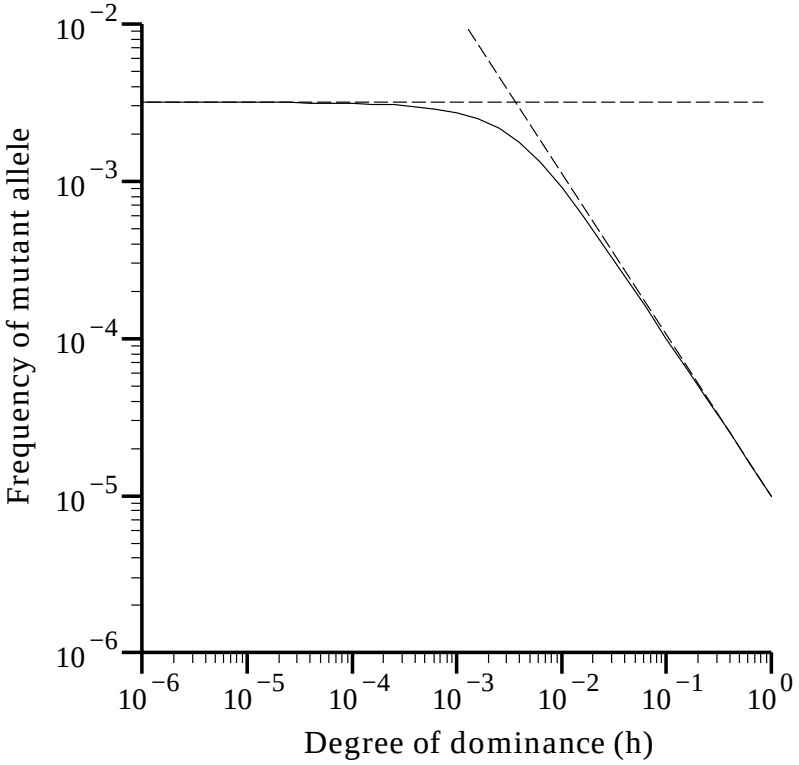


Figure 3.2: Equilibrium gene frequency as a function of h for a case in which $u = 10^{-7}$ and $s = 0.01$, with no back mutation. The horizontal and sloped dashed lines, respectively, show the recessive and partial-dominant approximations, from equation (III-24) and equation (III-26). Note the transition from the recessive case to the partial dominant case.

cies than the much better publicized effects which the gene may have in homozygotes.

Figure 3.2 shows the entire process of transition from validity of the partial-dominant approximation to the validity of the recessive approximation, for the case of Table 3.1.

POLYPLOIDY. The same sort of logic will serve as well with higher ploidy levels. If mutants are even partially dominant, most of the elimination of mutants will take place in genotypes carrying only one mutant. For instance, in a tetraploid with mutant gene frequency q, the frequency of $AAAa$ heterozygotes will be approximately $4q$. If q is small, most copies of a will occur in such genotypes. If the fitness of $AAAa$ is $1 - h_1 s$, then the reduction of a frequency by selection will be about $-h_1 sq$, while the increase due to mutation is about u (most genes being A and thus available for mutation to a) If

we equate these, we find that

$$q_e \;=\; \frac{u}{h_1 s} \tag{III-27}$$

which is the same formula as in the diploid case. It will be valid under the same sort of conditions, namely that $u \ll hs$ so that q_e is small and most elimination of mutants is in heterozygotes which carry a single copy.

III.7 Mutational Load.

If there were no mutation in a population, there would be no source of new variation allowing evolutionary progress, and the population would be worse off as a result. But there would also be no deleterious mutants occurring. How much advantage would accrue by the absence of mutation, owing to the lack of deleterious mutants? We are now in position to answer this question, although not the more general question of the effects of the absence of favorable mutants.

The simplest approach to this question, the one taken by J. B. S. Haldane in 1937 and by H. J. Muller in 1950, is to compute the effect of mutations on mean population fitness. This is easily done using the results of the previous section, and it leads to a rather surprising general result known as the Haldane-Muller principle. In the haploid case, the mean fitness of the population will be

$$\bar{w} \;=\; (1-q)\times 1 \;+\; q\times(1-s). \tag{III-28}$$

At equilibrium under mutation vs. selection, $q_e = u/s$ so that since

$$\begin{aligned} \bar{w} \;&=\; 1 - q_e\, s, \\ \bar{w} \;&=\; 1 - (u/s)\; s \;\;=\;\; 1-u. \end{aligned} \tag{III-29}$$

Without mutation, the gene frequency of the mutant allele will be zero, leading to $\bar{w} = 1$. So the presence of mutation depresses the mean fitness of the population by an amount equal to the rate of mutation to the deleterious allele. This is an unusual result, since it predicts that the amount by which a deleterious allele affects population fitness is independent of its fitness. So a mildly deleterious allele with $s = 0.01$. will have just as much effect in depressing population mean fitness as will a lethal which has $s = 1$!

In diploids, closely similar results are obtained. For recessive mutants,

$$\begin{aligned} \bar{w} \;&=\; (1-q)^2\times 1 \;+\; 2\,q(1-q)\times 1 \;+\; q^2\times(1-s) \\ &=\; 1 - q^2\, s \end{aligned} \tag{III-30}$$

and since $q_e = \sqrt{u/s}$,

$$\begin{aligned} \bar{w} &= 1 - \left(\sqrt{u/s}\,\right)^2 s \\ &= 1 - (u/s)\, s \\ &= 1 - u, \end{aligned} \tag{III-31}$$

so that at a locus with recessive mutant allele, the depression of fitness is again equal to the mutation rate and independent of the selective effect of the mutant.

A partially dominant mutant will have

$$\begin{aligned} \bar{w} &= (1-q)^2 \times 1 + 2q(1-q) \times (1-hs) + q^2 \times (1-s) \\ &= 1 - 2\,q(1-q)\,hs - q^2\,s. \end{aligned} \tag{III-32}$$

Since we have to very good approximation that $q_e = u/hs$,

$$\begin{aligned} \bar{w} &= 1 - 2(u/hs)\,(1-u/hs)\,hs - (u/hs)^2 s \\ &= 1 - 2u + 2u^2/hs - u^2/h^2 s \\ &\simeq 1 - 2u, \end{aligned} \tag{III-33}$$

The approximation involving dropping terms of size $q^2 s$ or $q^2 hs$, which are both expected to be very small since q is itself small. Once again we see that the decrease in fitness is dependent on u but not on either h or s (to the accuracy of the approximations used).

This reduction in fitness is known as the *mutational load*. Recall that we have computed it for a single locus. If population mean fitness is the product of mean fitnesses for the individual loci, as will be the case if there is multiplication of fitnesses between loci and no linkage disequilibrium, then for n partially dominant loci

$$\bar{w} \simeq (1-2u)^n \simeq e^{-2nu}. \tag{III-34}$$

If there are (say) 20,000 loci, each of which can mutate to a partially dominant deleterious allele, and if the mutation rate is 10^{-6} per locus, then the total mutation rate per diploid genome is 0.02. By (III-33) the fitness of the population is reduced to $e^{-0.02} = 0.9801986$, so that in this case the reduction of fitness is nearly the same as the total mutation rate per genome. If all mutations were instead recessive, the reduction in fitness would be only half as great. Figure 3.3 shows the full dependence of load at a locus on h, computed with the same equations that were used for figure 3.2.

A HEURISTIC APPROACH. The mutational load can also be derived in a heuristic fashion by a direct argument which does not utilize equilibrium gene frequencies. This

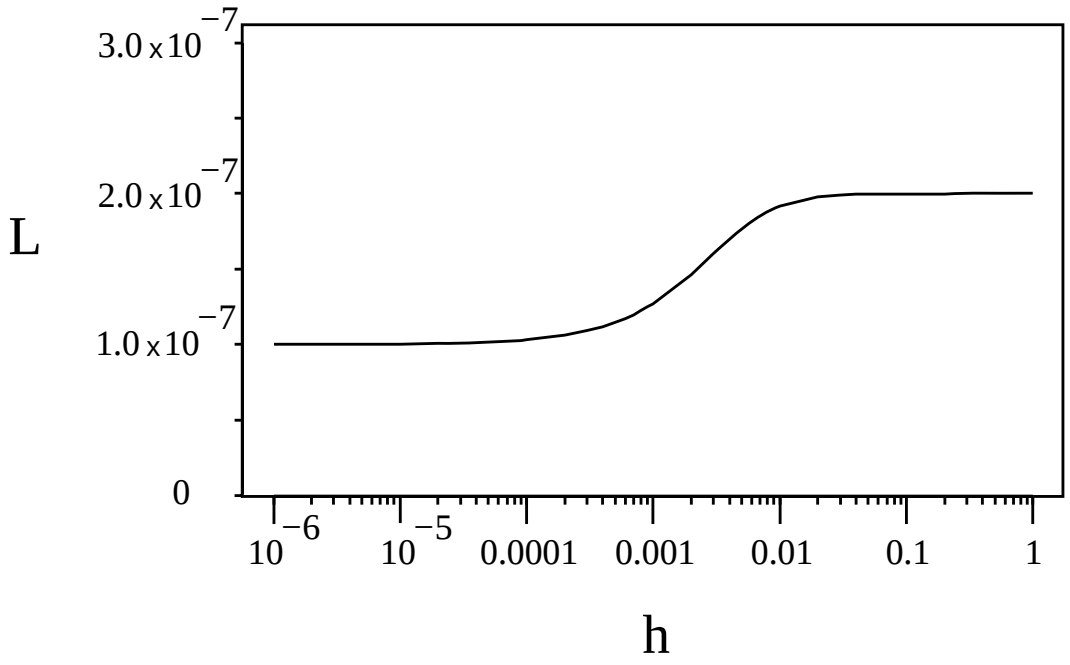


Figure 3.3: The mutational load as a function of dominance when $u = 10^{-7}$ and $s = 0.01$, for the case of no back mutation.

approach will make it a bit clearer why the load is the same for a weakly deleterious allele as for a strongly deleterious allele. Consider a (very large) population of N individuals. Every generation $2Nu$ new mutants will occur, since there are about $2N$ copies of the wild-type allele available to mutate. At equilibrium we require that the number of mutants eliminated by natural selection equal the number added by mutation. Let L be the fraction of individuals who die, so that L is our measure of the mutational load. If the mutants are partially dominant, then each selective death (with rare exceptions) kills one mutant. So requiring that the number of alleles killed equal the number that mutate amounts to requiring that

$$2Nu = NL \tag{III-35}$$

or

$$L = 2u. \tag{III-36}$$

If the mutants are recessive, each death by natural selection kills two mutants, so that

$$2Nu = 2NL \tag{III-37}$$

or

$$L = u. \tag{III-38}$$

Similar arguments easily compute the mutational load in haploids and in polyploids.

This deceptively simple argument is not as airtight as it seems. At equilibrium one should actually require that selection reduce the gene frequency by the same amount that mutation has increased it. This is not quite the same as decreasing the number of mutants by the number which have just mutated. A simple numerical example will serve. Suppose that a population of 1/2 million individuals has 10 mutants occur in the current generation, these being added to 100 copies of the mutant allele already present. If selection now restores the mutant frequency to its previous value of 0.0001 by killing homozygotes (let us assume that the mutants are recessive) it has to kill 10.001001 copies instead of 10 copies. The difference comes from the fact that if we killed exactly 10 copies of the mutant, we would not only reduce the number of mutants from 110 to 100, but would also reduce the total population number by 5 individuals (10 gene copies). This is hardly a dramatic inaccuracy, and it points up both the approximate nature of the argument and the essential accuracy of that approximation.

WEAK SELECTION AND MUTATIONAL LOAD. As the mutational load is said to be a function of the mutation rate, but not the selection coefficient, it is natural to wonder how a very weak selection could impose a load. Surely the Haldane-Muller principle cannot hold all the way to $s = 0$. Of course, it does not. In the haploid case, the mutational equilibrium gene frequency $q_e = u/s$ is only correct if $u < s$, otherwise the only equilibrium of the system (III-21) is $q_e = 1$. If $u \geq s$, so that $q_e = 1$, the load is

$$L = q_e s = s, \tag{III-39}$$

so that as we consider cases with progressively smaller values of s, the load will remain u until $s = u$, then below that point the load will smoothly decline to zero as s declines.

We have also been ignoring back mutation, and justifying this practice on the basis that the frequency of the mutant allele is very low. As $s \to 0$, we will be less and less able to make this approximation. As s becomes of the same order of magnitude as u, the equilibrium gene frequency will rise not to one, but to the mutational equilibrium frequency $q_e = u/(u+v)$, and the load will become $su/(u+v)$, which will approach zero.

Similar considerations apply in the diploid case and the polyploid case. The load will be a simple function of u unless s is so small that it is not substantially greater than u. Below that point the load will decline to zero as s does.

MEANING OF THE MUTATIONAL LOAD. We have considered the load as if it imposed a burden on the population, yet the reader may recall that in the case of the segregational load (in section II.8), serious reservations were expressed as to whether that load really imposed any burden. In that case, one might imagine that an overdominant mutant arose in a population previously fixed for one of the alleles. In the process

$\bar{w}$ would increase, even though a formal "load" would be created, through alteration of the standard against which $\bar{w}$ is judged from AA to Aa. In the present case, if we compare mean fitnesses in the population with or without mutation, we are using the same standard, namely the "wild-type" genotype AA. The presence of mutation is undoubtably making the population worse off, by introducing the genotypes Aa and aa which have lower fitnesses.

While there seem to be fewer difficulties with the notion of mutational load than with segregational load, matters are not quite as simple as they seem. We have carried out our computations in terms of relative, not absolute fitnesses. The existence of mutational load means that average population fitness will be below the fitness of the AA genotype. If the imposition of the mutational load were somehow to coincide with a raising of the absolute fitness of AA, then the "load" might actually benefit the population. It is also not necessarily true that the load will be visible to an ecologist as a lowering of population size. A population with density-dependent population size regulation might have an average of 1000 offspring per parent, with only two of those surviving to adulthood. In such a case, the reduction of the number of surviving offspring from 1000 to 500 will probably have only a small effect on the number of adults maintained in the population, since the population will still be pressing against the limits set by density-dependence. It would be a mistake to conclude from this that the mutational load has no effect whatsoever on population dynamics. The reproductive excess acts as a buffer allowing the population to survive various misfortunes, and a reduction of the size of this buffer, even if it has little effect on average population size, may make the population more vulnerable to extinction in time of crisis.

THE c PARADOX AND MUTATIONAL LOAD. The question of the reality of mutational load as a problem for the population is made more pressing when we consider what used to be called the "c paradox." This is the observation that eukaryotes have far more DNA in their genomes than we can account for based on estimates of the number of structural genes producing polypeptide chains. A typical mammal has about 5 billion base pairs in the DNA of its haploid chromosome complement. Taking an average structural gene to have about 1000 nucleotides in the part coding for the polypeptide chain, this is enough for 5 million loci. This is far more than could be estimated by extrapolation from the number of salivary gland chromosome bands in Drosophila, far more than was guessed from the numbers of different mRNA sequences expressed in the cell, and far more than turned out to be there in sequenced genomes.

Accounting for this "extra" DNA was a classical problem in molecular biology (i.e., it was a problem for more than 4 years). From the standpoint of population genetics, the mutational load calculation is relevant. If there were 20 million genes, each subject to mutation at a rate 10^{-7} to deleterious alleles, the fitness of the population would be reduced to $e^{-2} \simeq 0.15$ of its potential value. If we assign an absolute meaning to the mutational load, this would mean either a 15% probability that a newborn would survive

to adulthood (all risk coming from genetic disorders, with any death from environmental accident being on top of this) or else that fertility would be reduced by 85% by sterile or partially-sterile mutants, or an intermediate combination of these two. Clearly an organism with as much DNA as we have would be in severe trouble. Yet in humans well over 98% of all newborns survive to adulthood in most industrial countries.

WHY WE AREN'T ALL DEAD. There are several possible resolutions of the dilemma. If much of the DNA is simply "spacer" DNA whose sequence is irrelevant, then there will be a far smaller mutational load. But notice that the sequence must be truly irrelevant, not just of unknown function. If the "extra" DNA has regulatory or chromosome-pairing function requiring it to have a specific base sequence, then mutations in that sequence will still cause a mutational load, even if these loci are not producing polypeptide chains. Thus the mutational load argument seems to give weight to the notion that this DNA is nonspecific in sequence. That is now believed to be the case, and the mutational load must give pause to anyone who proposes to find important functions for most of the DNA in eukaryotic genomes, especially functions that constrain its sequence.

The other way out is to question whether the load is truly a burden. Surely an organism which increases its amount of DNA and evolves a new gene function cannot be making things worse for itself! It may be that in nominally increasing the load, organisms have at the same time increased the upper limit of their fitness (perhaps by increasing their reproductive excess) more than enough to compensate. Thus if a species starts out with a nonfunctional sequence at a given locus, and evolves a functional sequence, there is a nominal gain in mutational load as we change the "normal" standard from (say) a to A, but this is more than offset by the increase of fitness. The difficulty with accepting this view is that it seems to predict that as more loci enter the genome the discrepancy between maximum possible fitness and average fitness increases. This is hard to reconcile with the high viability of humans in industrial countries.

An alternative solution is to suggest that each selective death kills many mutant genes, so that one needs fewer selective deaths to balance the number of mutations. This implies that selection occurs in such a way that there is an interaction between mutants. If an individual only dies if it has ten or more mutants, then we cannot simply predict the fitness of a ten-fold heterozygote from the fitness of single heterozygotes. Models of selection of this sort were introduced by Sved, Reed, and Bodmer (1967), King (1967), and Milkman (1967) to deal with segregational load problems, and these papers may be consulted for more details. Hopf, Michod, and Sanderson (1988) have examined how the mutational load should vary among different mating systems. They find that the mutational load varies among them in easily predictable ways.

The mutational load calculation continues to be relevant to understanding whether most eukaryotic DNA has any function that is visible to natural selection. Recent announcements (Encode Project Consortium, 2012) that 80% of human DNA is "functional", based on finding some transcription or binding of transcription factors in it,

are very misleading. Junk DNA is still junk DNA, however often its demise has been announced.

III.8 Quasispecies

In an important paper that also put forward the hypercycle concept for the origin of life, Manfred Eigen and Peter Schuster (1971) also described as a *quasispecies* a population of genotypes with different fitnesses and with mutation among them. This has been applied primarily to viral populations. If the mutation rates are high enough relative to the fitness differences, there will be cloud of genotypes, with few individuals having the most fit genotype. This is the situation that Eigen and Schuster refer to. I have not been able to find any clear difference between a quasispecies and an ordinary population that has an equilibrium between mutation and natural selection against the deleterious alleles.

Eigen and Schuster gave equations for the selection/mutation process in haploid (or asexual) organisms, in a continuous-time model. In our terms, the fitnesses in Eigen and Schuster's model would be absolute fitnesses W_i. If there are n genotypes, and if the mutation rates between genotypes i and j is u_{ij}, Eigen pointed out that the equilibrium distribution under selection versus mutation would be proportional to the leading eigenvector of the matrix $\mathbf{WU}$, where $\mathbf{W}$ is a diagonal matrix of fitnesses of genotypes. This is an interesting general result, but not of much use except in particular cases where the equilibria can also be found by simpler methods. A more detailed account of their theory will be found in the paper by Eigen, McCaskill, and Schuster (1988). Wilke (2005) reviewed the concept of quasispecies, finding it essentially equivalent to a population with mutation-selection balance, and arguing that there is no disagreement between these two ways of describing viral evolution.

MUTATION IN A UNICELLULAR SPECIES.

When we have unicellular species, and reproduction is by cell division, when there are L sites that can mutate, then for the fully functional sequence the fitness $W = 2$. If there is no back mutation and if mutation occurs independently at each site, then by constructing the above matrices we find that the dominant eigenvalue is $2(1-u)^L$. The population will be able to survive if this quantity exceeds 1. Taking Lth roots and solving for u, the condition for this will be $u < 1 - (1/2)^{1/L}$ which is approximately $uL < \ln 2$. Thus the total mutation rate, summed over all sites, must be less than 0.693147. Eigen and Schuster noted that mutation rates high enough that each individual has on average 1 or more new mutant per generation were too high for continued existence of the species.

III.9 Mutation versus selection optimizing a phenotype

In his book *The Genetical Theory of Natural Selection* in 1930, R. A. Fisher argued that mutations of modest size will play the largest role in evolution. He supported this by a heuristic argument that has been influential, which has come to be called Fisher's Geometric Model. He imagines a phenotypic character that has an optimum value, with fitness falling away symmetrically on either side of the optimum. Suppose that the organism is haploid, and the allele at one locus determines the phenotype. If the current phenotype is a distance d from the optimum value, then a new mutant will have a higher fitness if it causes change x in the direction of the optimum and ends up at a distance less than d from the optimum.

If the mutant has a large enough effect that $x > 2d$, or $x < 0$, then it will have a lower fitness than the current allele.

III.10 Mutation and Linkage Disequilibrium

Mutation is a particularly simple evolutionary force in that (at least in our simple models) it occurs independently to each gene copy. This has enabled us to treat diploids as if haploid, and has generally kept things fairly simple. The amount of mutation at a locus, being independent of what are the genotypes at other loci, cannot be affected by the amount of linkage disequilibrium, but it remains to be seen whether mutation can in some way cause linkage disequilibrium. Will the presence of mutation cause a lack of independence between loci, and in this way complicate the analysis of other evolutionary forces?

Intuitively, one feels that an evolutionary force whose action at each gene is independent could not create disequilibrium. In fact, this intuition is correct, but to validate it requires a bit of algebra. Let us look at two loci in a population with discrete generations, where each locus has two alleles. Suppose that the population is in a state of linkage equilibrium, but does not necessarily have its gene frequencies at their mutational equilibrium values. Before mutation we have four chromosome types, AB, Ab, aB, and ab at frequencies given by the four quantities x_{AB}, x_{Ab}, x_{aB}, and x_{ab}, which we assume add to one. The forward and backward mutation rates at the A locus are given by u_1 and v_1, at the B locus by u_2 and v_2. After mutation the frequency of an AB chromosome (or haploid genome) is

$$x'_{AB} \;=\; x_{AB}(1-u_1)(1-u_2) \;+\; x_{Ab}(1-u_1)\,v_2 \;+\; x_{aB}\,v_1(1-u_2) \;+\; x_{ab}\,v_1v_2. \quad \text{(III-40)}$$

The justification of this formula is straightforward: to end up with an AB chromosome we must either start with an AB chromosome and have neither gene mutate, or start with an Ab chromosome and have the A gene remain unmutated while the b mutates to a B, and similarly for the other two chromosome types.

We have not yet used the fact that this generation started out in linkage equilibrium. We can write $x_{AB} = p_A p_B$, etc. so that

$$\begin{aligned} x'_{AB} &= p_A p_B (1-u_1)(1-u_2) + p_A p_b (1-u_1) v_2 + p_a p_B v_1 (1-u_2) + p_a p_b v_1 v_2 \\ &= [p_A(1-u_1) + p_a v_1]\ [p_B(1-u_2) + p_b v_2]. \end{aligned} \tag{III-41}$$

The fact that the chromosome frequency is a product of two terms, one corresponding to each locus, suggests that this may be a state of linkage equilibrium, and such is in fact the case. Comparison of (III-41) with formula (III-1) shows that (allowing for differences in notation) each of the terms in brackets is the new gene frequency of A or of B after mutation. So immediately after mutation we have linkage equilibrium:

$$x'_{AB} = p'_A\ p'_B. \tag{III-42}$$

We have considered part of a single generation, and seen that the changes in chromosome frequencies as a result of mutation are such as to alter the gene frequencies but leave the fact of linkage equilibrium unaltered. Although we have talked of chromosomes, the same calculation applies to two loci which are unlinked but in the same haploid genome or gamete genome.

Since recombination will have no effect on chromosome (gamete) frequencies which are in linkage equilibrium, we did not need to take it into account.

The principle that mutation alters gene frequencies without causing departure from linkage equilibrium can be applied to any number of loci and any number of alleles, but we are too lazy to do so here. In the more general case where we start in a state of linkage disequilibrium, mutation will act to move the population closer to linkage equilibrium by a very small amount, as is shown by Turner (1967).

III.11 History and References.

The mathematics of simple mutational equilibrium are so elementary that they were well-known to Fisher, Wright, and Haldane. Most of the early work on mutation was concerned with the fate of a single mutant (Fisher, 1922; Haldane, 1927) which we will treat in Chapter VII. The pioneer of work on mutation in populations seems to have been Danforth (1921), but full mathematical treatment was somewhat delayed. Recurrent mutation was not often considered because these authors held too realistic a view of population sizes and mutation rates: when Haldane (1927) finally treated the case of recurrent mutation, he did so in the more realistic and more general case of the balance between mutation and selection. There he presented a general treatment utilizing the full cubic equation generated by a model with forward and backward mutation and complete recessive inheritance of the mutant. A somewhat simplified version of this

argument was given in the Appendix to his book (1932). Fisher (1928, 1930) presented a verbal argument obtaining the equilibrium frequency of a dominant mutant opposed by selection. Wright (1929a) presented a more complete treatment of partial dominance and recessiveness, very similar in notation and conclusions to our section III.5 above.

The discussions of mutation-selection equilibrium by Haldane, Fisher, and Wright took place in the context of a debate over Fisher's theory of the origin of dominance of rare mutants. Fisher started from the well known fact that rare visible mutants tend to be recessive to wild-type far more frequently than they are dominant. Noting that the rare mutant would be present almost exclusively in heterozygotes, Fisher asked what would happen if alleles arose at another locus, a modifier locus, which altered the fitness of heterozygotes or homozygotes for the mutant. Alleles would be favored at the modifier locus that increased the fitness of those genotypes. Since few homozygotes would be present, the selection favoring increase of homozygote fitness would hardly ever be present, while modifier alleles whose effects were on the heterozygote would be subject to much larger selection. The result, Fisher argued (1928, 1930) would be a faster increase in the fitness of the mutant heterozygote than in the homozygote fitness, leading to the evolution of the degree of dominance of rare mutants toward complete recessiveness.

Wright (1929a) disagreed, arguing that in either case the selection on the modifiers was far too weak (owing to the rareness of mutants at the main locus) to allow it to be a significant evolutionary force in the face of genetic drift. In the ensuing controversy (Fisher, 1929; Wright, 1929b) debate centered around the effects of finite population size, with Wright citing results he was later (1931) to publish in his classic paper on the interaction of deterministic forces and genetic drift. Haldane (1930a) and Muller (1950b) sided with Wright in this dispute. These authors adduced other, more directly physiological reasons why deleterious mutants tended to be recessive, and saw no need for Fisher's theory of the evolution of dominance to be of real importance.

The mutational load was discovered independently by Haldane (1937, 1939, 1940) and Muller (1950a). Haldane's work of the 1930's (1935) followed the lead of Danforth (1921), who had made the first estimate of the rate of mutation of a human gene. In the 1950s and 1960s, the concept of mutational load played a role in the controversy over the genetic effects of atomic radiation and atmospheric testing of atomic weapons. Today it continues to be embroiled in controversy, as it validates a major argument that most of our genome must be junk DNA.

Exercises

1. Suppose that allele A is initially fixed in a population, and that it has a mutation rate of 10^{-5} to a, and that there is no back mutation. What will be the frequency of

allele a in the population (*exactly)* after 2 generations?

2. If there were back mutation in the above case, at the same mutation rate, what would the result be?

3. When $A \rightarrow a$ at rate 10^{-5} and $a \rightarrow A$ at rate 10^{-6}, what will be the equilibrium gene frequency in an infinitely large population?

4. In the above case, how long will it take the population to move half way to its equilibrium starting from fixation for A? Starting from equal frequencies of the two alleles?

5. Two alleles that are selectively neutral (have no difference in fitness) exist in mutation balance in a population. The rate of mutation from B to b is 10^{-5} and the rate of mutation from b to B is 5.0×10^{-6}.

   (a) What will the equilibrium frequencies of B and b be?

   (b) At that equilibrium which of these are correct:

   (i) The fraction of copies of B mutating to b is equal to the fraction of copies of b mutating to B.

   (ii) The number of copies of B mutating to b is equal to the number of copies of b mutating to B.

   (iii) Half of the mutation events occurring at this locus are to B, half to b.

   (c) If we start a new population, from individuals all of whom are BB, how many generations will go by before the frequency of b in that population rises to half of its equilibrium value?

6. DNA in eukaryotes ofter contains regions that have multiple tandem repeats of two or three nucleotides, such as CAGCAGCAGCAG .... The number of copies can either increase or decrease as a result of "slippage" in replication, and this can occur at a much higher rate than point mutation does.

   Suppose that we have a region with a trinucleotide repeat which has 30 copies in all members of the population. With a rate u there is slippage in replication of the region, and half of the time this leads to one more copy, and half of the time to one fewer copy.

   (i) What are the frequencies of 29, 30, and 31 copies after one generation?

   (ii) What are the frequencies of 28, 29, 30, 31, and 32 copies after 2 generations?

   (iii) What is the mean (expectation) of the change in copy number in producing that first offspring generation? (Yes, this is as easy as it seems).

(iv) What is the variance of the change in copy number in that generation? *[This will require you to know that the variance of a variable is the mean of its square, minus the square of its mean.]*

(v) In this case the variance after t generations will be the sum of the variances of change in each generation, so just the variance of change in that first offspring generation multiplied by t. If $u = 0.0001$, how many generations will be needed for the variance of copy number to become 1?

7. If a population reproduces apomictically (by completely asexual clonal reproduction) and has $A \rightarrow a$ at rate 10^{-5} and $a \rightarrow A$ at the same rate, what will be the gene frequencies at mutational equilibrium? What will be the genotype frequencies?

8. Suppose that we consider a haploid organism with lethal mutants occurring at rate u. How do the equations given in this chapter compare with our intuitive understanding of what the frequency of mutants at equilibrium will be? How rapid will be the return to equilibrium after the frequency of mutants changes?

9. What will be the frequency of mutants at equilibrium if the fitnesses of A (nonmutant) and a (the mutant allele) are $1 + s : 1$ instead of $1 : 1 - s$? (This can be solved without redoing all of the equations).

10. At Hiroshima and Nagasaki there can be little doubt (from the frequency of somatic effects such as leukemia) that many mutations occurred, yet studies of offspring of survivors have shown few genetic diseases among offspring of survivors. What might this imply about the values of h and s in such mutants?

11. A recessive deleterious mutant causes a syndrome S that reduces the fitness of the homozygote for the mutant by 20%. The heterozygote fitness is normal.

    (a) If we observe that the frequency of the allele (*note:* not the frequency of the homozygote) in the population is 0.01, and assume that the allele frequency is at equilibrium under selection versus mutation, what would the rate of mutation to the deleterious allele from the normal allele have to be? Why?

    (b) When in this state of equilibrium gene frequency, what fraction of copies of the allele are killed off each generation by the reduced fitness? Why?

    (c) This pool of copies of the mutant allele is about how many times larger than the number of copies that arise in one generation? Does this imply that the average copy arose very recently or not? Why or why not?

    (d) If we suddenly double the mutation rate, owing to some mutagen in the air or water, and it stays at that new value, by what fraction of its size will this pool of mutants increase in the next generation? Explain your reasoning.

12. What are some of the biases which we are likely to risk by taking observed human genetic disorders and using their frequencies and presumed fitnesses to calculate the average mutation rate per gene in humans?

13. Huntington's Disease is an autosomal dominant disorder with a frequency of about 1 in 100,000 in the population. It is a severe neurological disorder with onset in the mid-thirties, so that it is probable that the fitness of affected individuals is only reduced by 2% or so, since they have already had most of their offspring by this age. Estimate the mutation rate at this locus. How many generations ago did the average mutant allele in the population occur?

14. How would the frequency of Huntington's Disease be altered if (i) carriers now could be detected at birth and as a result of knowing their status they reduced their average number of offspring by 20%? (ii) if instead genetic counselling resulted in the offspring of affected individuals (who do not know their exact carrier status) having 20% fewer offspring themselves? How rapidly would the disease incidence change in these cases once those practices began?

15. Suppose we have a chromosome rearrangement that changes the gene order on the chromosome back or forth between two gene orders, which we will call I and II. The two gene orders are inherited as if they were alleles at a locus. Suppose that this change of gene order has no effects on fitness when the two chromosomes in a diploid are of the same gene order, but has 10% lower fitness when there is a "chromosome heterozygote" which has both of the two gene orders. If you have a rate u of rearrangement in both directions between these, what will be (approximately, using the formulas for equilibrium between mutation and selection) the equilibrium frequency of II in a population that starts from gene order I? What will be the equilibrium frequency of gene order I in a population that starts with gene order II? Give a formula in terms of u, and calculate the frequencies for $u = 10^{-6}$ and $u = 10^{-4}$

Complements/Problems

1. If we consider a population with partial self-fertilization (a randomly chosen fraction s of the population selfing in each generation, and the rest mating at random), in the absence of selection what will be the mutational equilibrium of gene frequency? of genotype frequencies?

2. In a diploid random mating population consider an underdominant locus with fitnesses of AA, Aa, and aa of $1 : 1 - s : 1$. If mutation rates for $A \to a$ and $a \to A$

are both μ, what will be the (approximate) stable equilibrium gene frequencies if μ is small?

3. In a sex-linked locus with two alleles, if the mutation rates for $A \rightarrow a$ are u_f and u_m in the two sexes, and the rates for $a \rightarrow A$ are v_f and v_m, what will be the equilibrium gene and genotype frequencies in the absence of selection?

4. For what values of s and u will the various equilibrium solutions of (III-15), the haploid case of mutation-selection balance, be stable or unstable?

5. Bacteria grow by doubling. Suppose that a bacterial population has its DNA have n sites that are under selection and need to be maintained in their current state. If fitnesses at the different sites are multiplicative (the assumption I typically use in class), what is the largest mutation rate per site per generation (to other base states, all of which are deleterious) you can have and still have the most-fit class just barely be able to reproduce itself (i.e. just be able to have the average bacterium have one fully-fit descendant). *Hint – you won't need to compute mutational loads, just try to figure out the expected number of nonmutant, fully-fit descendants and ensure that it is 1. That number would be 2 if there were no mutation.*

   What would this mutation rate be if $n = 5 \times 10^6$? Now look at the results of the the subsection on Mutation in a Unicellular Species in section III.8 and compare them to your result.

6. Make a simple analog to the 1971 argument of Eigen and Schuster (see section III.8 above) for two haploid genotypes with unidirectional mutation and selection against genotype 2. Imagine that we have mutation rate u from genotype 1 to genotype 2, no reverse mutation, and that the relative fitness of genotype 2 is $1 - s$. This is actually identical to the haploid model in section III.4. If we observe the genotype frequencies immediately after mutation, the matrices **W** and **U** that correspond to Eigen and Schuster's matrices are

$$\mathbf{U} = \begin{bmatrix} 1-u & 0 \\ u & 1 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} W_1 & 0 \\ 0 & W_1(1-s) \end{bmatrix}.$$

   Compute the product of matrices **UW**, find its eigenvalues and eigenvectors, and show that the proportions of the frequencies of the terms in the leading right eigenvector are exactly the equilibrium frequencies of the two alleles derived in section III.4.

7. Suppose that in a diploid there is a locus whose alleles include one normal allele and two mutant alleles. The mutation rate from the normal allele to the first mutant allele is μ_1, and that to the second mutant allele is μ_2. There is no back mutation or

mutation between the two mutant alleles. Suppose that any individual with two of the mutant alleles (whether or not they are the same mutant) dies. Heterozygotes between either mutant allele and the normal allele have no reduction in fitness. What are the equilibrium frequencies of the two alleles? What would they be if instead the heterozygote between the two mutant alleles had normal fitness?

8. We could approximate (III-15) by saying that the two evolutionary forces, selection and mutation, make changes of gene frequency of approximately $s\,p(1-p)$ and $-u\,p$ respectively. When we require that these cancel each other out by summing to zero, we can obtain an equilibrium gene frequency. Is it correct? Should it be? Why or why not?

9. If a recessive lethal mutant can occur at a sex-linked locus, with equal mutation rates in females and males, what will its equilibrium gene frequency be (assuming that male hemizygotes are equivalent to female homozygotes in fitness)? What will be the answer if the fitness of hemizygotes and homozygotes is $1-s$ rather than zero? How could we use sex-linked mutants to estimate mutation rates of human genes?

10. We have explained that You can work backwards from the formula for the equilibrium frequency of a deleterious allele and infer the mutation rate to it.

    (i) Imagine a deleterious recessive allele with selection coefficient s and observed to be at gene frequency q. If this is the equilibrium gene frequency under this pattern of selection, derive the formula for the estimate of the mutation rate to this allele, explaining as you go.

    (ii) But now suppose that the allele actually has a fitness $1-hs$ in the heterozygote as well, and that that is what affects its equilibrium frequency, and you just didn't know that. Instead of the correct mutation rate, what mutation rate would you get? (Use the above equations, but plug into them the equilibrium gene frequencies you expect with this amount of selection against the heterozygote. Write the result it as an expression in the true mutation rate u and the values of h and s). Will that be higher or lower than the truth?

11. If the fitnesses of AA, Aa, and aa are respectively, 1, $1-hs$, and $1-s$, with unidirectional mutation $A \rightarrow a$ at rate u, what is the exact quadratic equation for equilibrium gene frequency of a?

12. Suppose that we have fitnesses

Genotype	AA	Aa	aa
Fitness	0	1/2	1

at a locus.

(i) What is the formula for the gene frequency of the deleterious allele A in the next generation? (For this simply see subsection II.6 on Additive Fitnesses in the notes)

(ii) Now suppose that after selection there is mutation, with forward and backward mutation rates both being u. What is the formula for the gene frequency after that?

(iii) What is the equilibrium frequency? (This does not even require solving a quadratic equation).

(iv) How does this compare with the formula for the equilibrium frequency derived in the book in cases of partial dominance? Why the discrepancy? How serious is it likely to be? (Note that the deleterious allele is taken to be allele a in that derivation).

13. What equations must be solved to obtain the equilibrium gene frequencies in the sex-linked case corresponding to the autosomal case of the preceding problem? Assume hemizygotes have the same fitnesses as the corresponding homozygotes.

14. Based on problem 8, what is the exact mutational load as a function of u, s, and h?

15. Use the equations for selection at a multiplicative locus to obtain an expression for the mutational load in a case where fitnesses are geometric: $1 : 1-s : (1-s)^2$. Is the approximation $L = 2u$ a good one in this case?

16. When mutation is bidirectional, with $A \rightarrow a$ at rate u and the reverse at rate v, obtain an exact (quadratic) expression for the mutational load in the haploid case. Does it show results consistent with our intuitive discussion when s is allowed to be very small?

17. Epigenetic modifications of DNA such as methylation can produce changes of phenotype, though the modifications generally revert in one or a few generations. Consider a haploid organism with a site with two states, unmethylated and methylated, which we will treat as if they were alleles that mutated to each other. Suppose that the probability of change to the methylated state is u, and that the probability of reversion to the unmethylated state is $1/3$ per generation. Suppose that the fitness of the individuals with the unmethylated state is $1-s$. Use the mathematics developed in the answer to the previous question to find out what values of s and u will predict a frequency of the methylated state that is greater than 0.01. You may want to compare your answer to the conclusions of Slatkin (2009) and Tal, Kisdi and Jablonka (2010).

18. Using the results in problem 6, which asked you to compare Eigen and Schuster's equations to the results of the section on the mutation/selection equilibrium at a two-allele haploid locus, to show that if we had n loci, where fitnesses were multiplicative across loci, that the mean fitness in a population at equilibrium will be $(1-u)^n$. Consider for what values of n the mean relative fitness will be substantially lower than the fitness of the best genotype. In light of this, show that this is a counterpart to Eigen and Schuster's argument that if $u > 1/n$, the best genotype will be rare in the population.

19. In an infinite haploid population, suppose that there are n loci in the genome. Each mutant allele multiplies the fitness by $1-s$, where this selection coefficient is the same for each locus. Suppose that there is a mutation rate u from the normal allele to the mutant allele at each locus, and that there is no back mutation. Group the haplotypes according to how many mutant alleles they carry, so that all we know is the values of f_k, where $k = 0, 1, \ldots, n$.

    - What is the distribution of the number of new mutants that occur in offspring of a genome that already has k mutants?
    - Derive the equations for change of the f_k under mutation, in the production of newborn offspring.
    - Derive the equations for the change of the f_k between the newborn stage and the adult stage, if the selection is by differential viability.
    - What is the equilibrium distribution of k at the life stage immediately after mutation occurs and immediately before natural selection acts?
    - Is this a member of a known family of probability distributions?
    - Is it the distribution you would get if there were linkage equilibrium?

20. What are the equations for change of linkage disequilibrium when we start in linkage disequilibrium and with gene frequencies away from their mutational equilibrium? Can independent mutation at each locus ever increase the value of D?

Chapter IV

MIGRATION

IV.1 Introduction

Migration is a bit of an enigma. Although it is one of the evolutionary forces whose mathematics is simplest, it is not easy to see whether it is an adaptive or a maladaptive force. As we shall see, while migration works against adaptation to local environments, and in this sense is maladaptive, there may be situations in which adaptation to local environments is itself maladaptive with respect to future environments. Migration is of particular interest because its consideration enables us to take advantage of the existence of the spatial dimensionality of the environment. Alternatively, one may regard it as a complication, a violation of the simplicity of the single random-mating population, but even then the presence of this complication is accompanied by an increased amount of information in the form of gene frequencies or phenotypes at different locales.

IV.2 The Effect of Migration on Gene Frequencies

The mathematics of migration are only slightly less simple than those of mutation. Mutation occurs (roughly) independently to each gene, while if it is diploid individuals who are doing the migrating, the genes they introduce into a population come in packets of two. In terms of gene frequencies, the effects of migration are easily seen. Suppose that a population, immediately after some migration has occurred, consists of a fraction $1 - m$ of individuals who have not immigrated, plus a fraction m of new immigrants from another population. Suppose that the gene frequencies of an allele A were p_1 and p_2 in these two populations before the migration occurred. Finally, suppose that individuals migrate or stay at home without regard to their genotypes. Then among non-immigrants the frequency of the A allele will still be p_1, and among immigrants it will be p_2.

A simple calculation of a weighted average then immediately tells us that the new

gene frequency after migration will be

$$p_1' \;=\; (1-m)\; p_1 + m\; p_2. \tag{IV-1}$$

this result applies to a single allele. Since we have not specified how many other alleles there are, we have in effect done the computation for the general case of multiple alleles. But we have restricted ourselves to two populations. If there are a total of n populations, and if m_{1i} is the fraction of individuals in population 1, after a bout of migration, who come from population i, then a directly analogous formula holds:

$$p_1' \;=\; (1 - m_{12} - m_{13} - \cdots - m_{1n})\, p_1 \;+\; m_{12}\, p_2 + \ldots \;+\; m_{1n}\, p_n. \tag{IV-2}$$

We can make this expression a bit more compact by defining m_{11} to be the fraction of individuals in population 1 which are nonimmigrants. This replaces the expression in parentheses in (IV-2), yielding

$$p_1' \;=\; \sum_i m_{1i}\, p_i. \tag{IV-3}$$

We shall see later that these conventions enable a compact matrix notation.

IV.3 Migration and Genotype Frequencies: Gene Pools

We have been following gene frequencies rather than genotype frequencies. When we start keeping track of genotypes, things are not quite so simple. It starts to matter very much at what stage of the life cycle the migration takes place. Initially, suppose that adults migrate. If we have two populations, each fixed for a different allele. If population number 1 is fixed for A, and receives 20% immigration of diploid individuals from population 2 which is fixed for a, then the frequencies of genotypes after migration will be

AA	Aa	aa
0.80	0	0.20.

These are certainly not Hardy-Weinberg proportions. There is a strong deficit of heterozygotes.

We can compute the frequencies of genotypes in a mixture by a simple process of weighted averaging. If $P_{ij}^{(1)}$ and $P_{ij}^{(2)}$ are the genotype frequencies of A_iA_j respectively in populations 1 and 2, after migration, and $P_{ij}^{(m)}$ is the genotype frequency in the resulting mixture, we have straightforwardly

$$P_{ij}^{(m)} \;=\; (1-m)\; P_{ij}^{(1)} + m\; P_{ij}^{(2)}. \tag{IV-4}$$

In the numerical example both populations were fixed before migration. Note that fixation for an allele is also a state of being in Hardy-Weinberg proportions. So by mixing

two populations, each in Hardy-Weinberg proportions, we have created a mixture which is not in Hardy-Weinberg proportions.

THE WAHLUND EFFECT. Mixed populations are usually out of Hardy-Weinberg proportions immediately after the mixture, as a calculation due to Wahlund (1928) shows. Let us retreat to the case of two alleles. Let p_i be the gene frequency of A in population i (*not* the frequency of allele A_i). Before mixture let each population be in Hardy-Weinberg proportions. Immediately after the mixture of diploid individuals occurs, but before any mating can take place, the proportion of AA homozygotes in the population will be

$$P_{AA} \;=\; m_1\, p_1^2 \;+\; m_2\, p_2^2 \;+\; \cdots + \; m_n\, p_n^2, \tag{IV-5}$$

where m_i is the fraction of the mixture which was contributed by population i. This formula is just a version of (IV-4) for n ancestral populations, each in Hardy-Weinberg proportions, with two alleles. Note that (IV-5) is just the weighted average of p^2, taken over all n populations with weights (m_i) proportional to the contribution of each population to the mixture. It is simply $\mathbb{E}(p^2)$, the expectation of the random quantity p^2, where the randomness comes from having probability m_i of encountering an individual from population i, a population having $p = p_i$.

By contrast with (IV-5), if the mixture were in Hardy-Weinberg proportions we could predict its genotype frequencies from its gene frequencies in the usual way. The gene frequency in the mixture is given by (IV-2) and (IV-3). Note that these can be written as the expected value $\mathbb{E}(p)$, where the random quantity p has probability m_i of taking on the value p_i. The Hardy-Weinberg genotype frequency of AA is of course simply the square of this, $[\mathbb{E}(p)]^2$, so that the deviation of the frequency of AA from Hardy-Weinberg proportions is

$$P_{AA} \;-\; [\mathbb{E}(p)]^2 \;\;=\;\; \mathbb{E}(p^2) \;-\; [\mathbb{E}(p)]^2. \tag{IV-6}$$

The righthand side of (IV-6) is the standard expression for the variance $\mathrm{Var}\,(p)$ of the random variable p. So rearranging this equation

$$P_{AA} \;\;=\;\; [\mathbb{E}(p)]^2 \;+\; \mathrm{Var}\,(p). \tag{IV-7}$$

This is the Wahlund Effect: the frequency of homozygotes in a mixture of populations, each in Hardy-Weinberg proportions, is increased over Hardy-Weinberg proportions by the variance of the frequencies of the allele in the components of the mixture. For instance, if we draw a sample of individuals from a geographic area which comprises not one random-mating population but many populations, whose mean gene frequency is 0.3 with standard deviation of gene frequency 0.1, then the frequency of homozygotes in our sample is expected to be $(0.3)^2 + (0.1)^2 = 0.09 + 0.01 = 0.10$, so that the overall frequency of homozygotes for this allele is increased above its Hardy-Weinberg expectation by an amount equal to the variance of gene frequency among the populations.

If there are two alleles, Wahlund's Law applies to each homozygote, and since $p+q=1$, the variance of the frequency of a must be exactly the same as the variance of the frequency of A. So equal amounts are added to the frequencies of AA and aa, and there must be a corresponding subtraction from the frequency of Aa. Replacing $\mathbb{E}(p)$ by the simpler notation $\bar{p}$, we have

$$\begin{aligned} p_{AA} &= \bar{p}^2 + \mathrm{Var}\,(p) \\ p_{Aa} &= 2\,\bar{p}(1-\bar{p}) - 2\,\mathrm{Var}\,(p) \\ p_{aa} &= (1-\bar{p})^2 + \mathrm{Var}\,(p). \end{aligned} \tag{IV-8}$$

With multiple alleles the situation becomes more complex. The Wahlund Effect applies to each homozygote, although the variances of the frequencies of different alleles need no longer be equal. The heterozygote frequencies are increased by twice the covariance of the frequencies of the relevant alleles, the covariance being weighted, as was the variance, by the contributions of each population to the mixture. Most pairs of alleles have negative covariances, so that usually the heterozygote frequencies are decreased. But it is is possible for covariances to be negative, so it is possible for mixture to create an excess of certain heterozygotes relative to Hardy-Weinberg expectation. However, since each homozygote is present in excess, there must be an overall deficit of heterozygotes. The Wahlund Effect is worth keeping in mind: it is one of the most important sources of departure from Hardy-Weinberg proportions in samples from nature that may come from more than one population.

EFFECTS OF RANDOM MATING. The Wahlund Effect operates in a direct mixture of individuals who have not yet had an opportunity to intermate (indeed, the mixture may be entirely a product of our sampling methods). If a generation of random mating occurs after mixture, the results are entirely different. We have already seen that the offspring of random mating will be in Hardy-Weinberg proportions no matter what the genotype frequencies in the parents, provided that suitable conditions apply. We may conclude that, following random mating, the population which has received immigration will be in Hardy-Weinberg proportions at the new gene frequency given by (IV-3), and Wahlund's Law will no longer be relevant.

It will therefore be important in modelling the process of migration to keep careful track of the life stage of which the migration occurs. If haploid gametes (both eggs and sperm) migrate and thereafter all the gametes in the population combine at random, Hardy-Weinberg proportions will be achieved immediately, while migration of adults leads to a mixed population out of Hardy-Weinberg proportions, which does not achieve Hardy-Weinberg proportions until the next generation. But if migration recurs each generation, there may never be a generation that is truly in Hardy-Weinberg proportions.

LINKAGE DISEQUILIBRIUM CREATED BY MIGRATION. The foregoing discussion has been entirely in terms of single loci. Just as it creates departure from Hardy-Weinberg proportions, the process of mixture creates linkage disequilibrium as well. An example may be useful. Suppose that population 1 consists entirely of *AA BB* individuals, and population 2 entirely of *aa bb* individuals. In a mixture of these in equal proportions it is obvious that there will be departure from Hardy-Weinberg proportions at both loci, as the new gene frequencies are 0.5 A : 0.5 a and 0.5 B : 0.5 b while there are no heterozygotes at all. There will also be linkage disequilibrium: of the gametes contributed to the next generation, half will be *AB* and half *ab*, with no *Ab* or *aB*. Note that the two original populations are each in Hardy-Weinberg proportions and linkage equilibrium, since fixation for (say) *AA BB* is such a state. So it is possible to create linkage disequilibrium by mixing individuals (or gametes) from two or more populations, neither of which itself shows any linkage disequilibrium.

Will linkage disequilibrium always result from mixture? How much disequilibrium will be created? A simple derivation may shed some light on this. We have seen that when we make a mixture of individuals, a fraction m_i coming from population i, we can regard there as being a random variable p which takes on the value p_i with probability m_i. Let p be random variable corresponding to allele A, and let q be the random variable corresponding to allele B. Since these alleles are at different loci, we do *not* have that $p + q = 1$. Consider a randomly chosen individual from the mixture, and consider a gamete produced by it. With probability m_i the individual came from population i. We are assuming that it was randomly chosen from that population, and that the population is in linkage equilibrium. If so, the probability that the gamete is *AB* is simply $p_i q_i$. So the overall proportion of *AB* gametes produced by individuals in the mixture is

$$\text{Prob}\,(AB) \;=\; \sum_i m_i\, p_i\, q_i. \tag{IV-9}$$

We can subtract from this the product of the frequencies of A and B in the mixture. These are given by expressions like (IV-3), so that the amount of linkage disequilibrium in the mixture is given by

$$D \;=\; \sum_i m_i\, p_i\, q_i \;-\; \left(\sum_i m_i\, p_i\right)\left(\sum_i m_i\, q_i\right). \tag{IV-10}$$

This is formally the answer, but its meaning will be clearer with some further interpretation. Each of the parts of (IV-10) can be written in terms of the random variables p and q:

$$D \;=\; \mathbb{E}(p\,q) \;-\; \mathbb{E}(p)\,\mathbb{E}(q) \;=\; \text{Cov}\,(p, q). \tag{IV-11}$$

The linkage disequilibrium in the mixture is simply the covariance in the frequency of p and q over the populations contributing to the mixture, weighted by the proportions

in which they contribute. Only if the distributions of gene frequencies at the two loci are independent when examined across the original populations will there be no linkage disequilibrium created by mixture. On the other hand, if p and q do not vary much over populations, the amount of disequilibrium cannot be large (in fact, it cannot be greater than the product of the standard deviations of the allele frequencies at the loci).

Once the mixture occurs, the linkage disequilibrium starts to decay by recombination. If the loci are unlinked, then it rapidly disappears, leaving a population in linkage equilibrium at the new gene frequencies given by (IV-3).

GENE FLOW. Once the departures from linkage equilibrium and Hardy-Weinberg proportions are gone, what has been accomplished by the admixture is to alter the gene frequencies to new values, intermediate between those of the contributing populations. This is an example of the notion of the population as a *gene pool*, a mixture of genes rather than phenotypes, genotypes, or gametes. In a state of linkage equilibrium and random mating, all genotype frequencies can be constructed from the gene frequencies. In fact, we can predict the existence of genotypes whose frequencies may be so low that in most generations they do not occur in the population. The migration computations serve as a clear illustration of the fact that it is the gene frequencies, and not individual genotypes, which form the inheritance of a population.

This has often been expressed by referring to migration as *gene flow*, to emphasize that from the standpoint of evolution it is the genes which move, and not individuals or genotypes. Cases can be found in which the retention or creation of departures from random combination have qualitative effects on the outcome. More often, the vision of the population as a gene pool and migration as gene flow gives a true picture of the underlying dynamics of this evolutionary force.

IV.4 Estimating Admixture

Human populations are frequently composed of mixtures of individuals of different genetic backgrounds. When this admixture has occurred in past generations, it leaves its trace in the gene frequencies. Attempts have been made to use gene frequencies to estimate the degree of admixture of various populations.

In the simplest case, the computation simply reverses equation (IV-1). If we call the frequency of an allele in the putatively admixed population p, (IV-1) is

$$p \;=\; (1-m)\; p_1 + m\; p_2. \tag{IV-12}$$

we can use (IV-12) to estimate m if p, p_1, and p_2 are known, solving it to give

$$m \;=\; \frac{p - p_1}{p_2 - p_1} \tag{IV-13}$$

For example, in the town of Claxton, Georgia the gene frequency of the A blood type allele (actually a composite of two alleles, A_1 and A_2) was found by Cooper et. al. (1963) to be 0.05 among whites and 0.129 among African-Americans. In West African populations (Cavalli-Sforza and Bodmer, 1971, p. 493) the frequency of A is about 0.143. Our estimate of the fraction of European ancestry among African-Americans in Claxton will then be

$$\hat{m} \;=\; \frac{0.129 - 0.143}{0.05 - 0.143} \;=\; 0.1505$$

The pitfalls of such a computation are many. In the first place, each of these gene frequencies is itself an estimate, and the variances and covariances of these quantities must be taken into account in computing the variance of the resulting estimate. There is a presumption that the admixture of white ancestry is fairly represented by looking at gene frequencies of whites in Claxton. Even more dubious are the West African gene frequencies. Slaves were taken not only from West Africa, but from Central Africa and East Africa. Even within West Africa there are local differences in gene frequencies which render it dubious whether a given West African population is representative of the ancestry of African-Americans.

Granted these limitations, an interesting application can be made of this sort of calculation. When admixture estimates are made for the same set of populations using different loci, the results sometimes differ. For example, Cavalli-Sforza and Bodmer (1971) obtain an estimate of 0.296 for the European ancestry of African-Americans in Claxton when the allele for sickle-cell anemia is used. The most likely explanation for this discrepancy is that the gene frequency of the $Hb\beta^S$ allele reflects not only admixture but also selection. In a New World environment lacking *falciparum* malaria, the deaths of $Hb\beta^S\ Hb\beta^S$ individuals from sickle-cell disease will reduce the frequency of that allele, an effect which will make it look as if there has been greater admixture from the white population.

REFERENCES. Admixture studies led to some of the earliest work on migration effects. The basic formula (IV-1) was used by Bernstein (1931) in the form of equation (IV-13). The notion of using admixture studies to indicate which loci are under natural selection was first used by Workman, Blumberg, and Cooper (1963). Wahlund (1928) derived the effects of mixture on genotype frequencies, but it was only realized four decades later that such mixtures would also be a source of linkage disequilibrium (Cavalli-Sforza and Bodmer, 1971, p. 69). The formulation in terms of covariances is due to Timothy Prout (in his Appendix to Mitton and Koehn, 1973).

IV.5 Recurrent Migration: Models of Migration

.

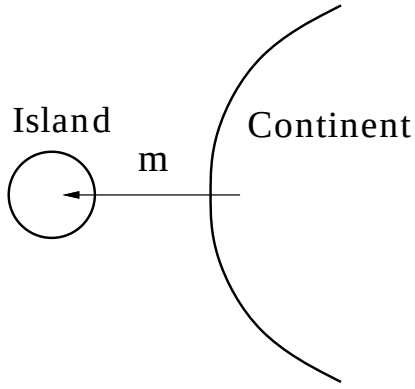


Figure 4.1: The one-island model.

When migration exerts a continual effect, the mathematics is a simple extension of that given above. There are many different situations possible, but a few patterns of migration have accounted for most of the studies in the literature of population genetics. Let us review a few of these models of migration.

THE ONE-ISLAND MODEL. One imagines a small island located near a large continent. Migration from the island to the continent is too tiny a fraction of the gene pool of the continent to be of any influence, and it is imagined that the gene frequencies on the continent remain unchanged. But a fraction m of the gene pool on the island comes from the continent each generation. The geographic situation looks like this: and the equation for gene frequency $p^{(t)}$ on the island in generation t is:

$$p^{(t)} \;=\; (1-m)\; p^{(t-1)} \;+\; m\; p_c \tag{IV-14}$$

where p_c is the constant gene frequency on the continent.

THE ISLAND MODEL. Suppose that we had n islands which exchanged migrants with each other. Each generation the fraction of genes which arrive from each given other island is taken to be $m/(n-1)$, so that m is the fraction of genes which come from outside each island. In this model there is not only a total symmetry, but even an absence of geography. No island is imagined to be closer to one neighbor than to another. This geography cannot be realized except in $n-1$ dimensions, but it may be represented as in Figure 4.2. Population i will have its gene frequency $p_i^{(t)}$ in generation t be determined by:

$$p_i^{(t)} \;=\; (1-m)\; p_i^{(t-1)} \;+\; \sum_{j\neq i} \frac{m}{(n-1)}\; p_j^{(t-1)} \tag{IV-15}$$

a similar equation holding for each population.

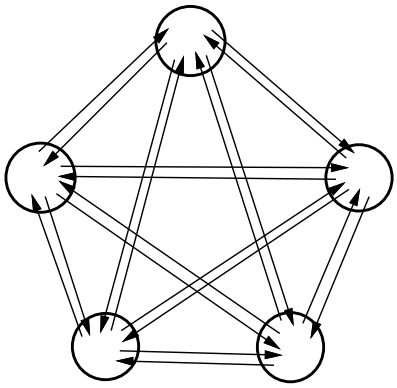


Figure 4.2: The island model.

THE STEPPING-STONE MODEL. Imagine that the populations are arrayed in a regular pattern in space, and let migration depend on the distance between them. The simplest patterns are a one dimensional string of equally-spaced populations, and a two dimensional rectangular lattice:

Migration is imagined to occur between neighboring populations. In the one-dimensional case a fraction $m/2$ of genes after migration are immigrant from each neighbor. In the two-dimensional case $m/4$ are immigrants from each neighbor. More general patterns of migration can be imagined in which the number of migrants depends on distance in a more complicated way, with some migrants being received from populations 2, 3, or more steps away along the chain of stepping stones.

In the Figure, the arrays of stepping stones are imagined to extend off to infinity in all directions. If the object is to study a population of finite geographical extent, it is easy to envision a line of populations of finite length, or a rectangle of populations. In these cases some special provision must be made for the pattern of migration at the boundaries. In the one-dimensional case, an end population may be imagined to receive $m/2$ of its genes from its one neighbor, so that only $m/2$ of its genes are new immigrants immediately after migration. Alternatively we may specify that m of its genes come from the neighbor, so that the fraction of genes in each population which are newly immigrant is kept the same.

To avoid the mathematical difficulties inherent in having some populations be more equal than others, we may consider a circle of populations instead of the one-dimensional line, or a toroidal pattern (instead of the rectangle), where the top and bottom rows of populations exchange migrants, and so do the leftmost and rightmost columns. If this is done properly, the model is seamless, if not terribly realistic.

THE GENERAL MIGRATION MATRIX MODEL. The most general possible pattern,

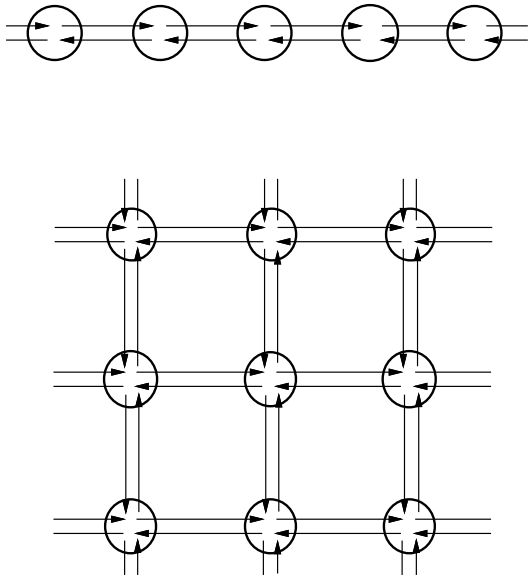


Figure 4.3: The one- and two-dimensional infinite stepping stone models.

one which subsumes all the others, simply states that we have n populations, and that after migration of fraction m_{ij} of the genes in population i are newly arrived from population j. Some care must be taken in interpreting the quantity m_{ij}. The table of the m_{ij} is referred to as the *backward migration matrix*, because it looks back in time from the recipient (i) to the source (j), measuring the number of immigrant individuals (or genes) as a fraction of the number of individuals (genes) in the population receiving the immigration, the measurement being conducted after the migration. One can also describe the migration rates by means of the *forward migration matrix* m^*_{ij}, which gives the fraction of individuals (or genes) in population i who end up in population j. Since the *number* of individuals migrating must be the same whether we measure it as immigrants or emigrants, we have the following relationship:

$$N_i \, m^*_{ij} \;=\; N'_j \, m_{ji} \tag{IV-16}$$

where N_i is the population size of population i before migration, and N'_j is the size of population j after migration. The two sides of (IV-16) simply compute the number of migrating individuals using the two different rates m^*_{ij} and m_{ji}.

The advantage of posing this general migration matrix model is that the computations can be recast in matrix form. If $\mathbf{M}$ is the matrix of the m_{ij}, and if $\mathbf{p}^{(t)}$ is the vector of the $p_i^{(t)}$, then the counterpart to (IV-1) is simply

$$\mathbf{p}^{(t)} \;=\; \mathbf{M}\,\mathbf{p}^{(t-1)}. \tag{IV-17}$$

A special case of particular interest is when the expected number of individuals arriving in population i from population j is equal to the expected number arriving in j from i. In that case

$$N_i\, m_{ij} \;=\; N_j\, m_{ji}. \tag{IV-18}$$

For the particular case of equal population sizes, this gives equal migration rates in both directions, so that the matrix **M** is symmetric. Whether or not the population sizes are equal, (IV-18) defines a situation where each population exports as many genes as it imports, so that there is no net outflow or inflow from any population. In this case, a gene gains no advantage from being in any particular population. As we shall see, this case yields particularly simple results.

IV.6 Recurrent Migration: Gene Frequencies

.

Having defined these models of migration, let us look at a few of them to see what will be the effect of migration on gene frequencies.

THE ONE-ISLAND MODEL. Equation (IV-14) has an equilibrium value which is particularly easy to find. Dropping the superscripts (t) and $(t-1)$ and solving it for p, we get $p = p_c$. More interesting is the rate of convergence to this equilibrium. Subtracting p_C from both sides of (IV-14).

$$\begin{aligned} p^{(t)} - p_c &= (1-m)\; p^{(t-1)} \;+\; m\; p_c - p_c \\ &= (1-m)\; (p^{(t-1)} \;-\; p_c). \end{aligned} \tag{IV-19}$$

The island population is approaching the continental gene frequency p_c as a result of recurrent immigration. Equation (IV-19) shows that the deviation of p from its final value p_c is multiplied by $(1-m)$ every generation. Thus the island population moves a fraction m of the way to its equilibrium each generation. The effect of migration is to make the gene frequencies p and p_c more similar (in this case by changing only p). the rate at which this happens is given by m.

THE ISLAND MODEL. This model has a particularly symmetrical structure which yields results easily. Equation (IV-15) can be rewritten (after a little algebra) as

$$\begin{aligned} p_i^{(t)} &= (1 - m + \tfrac{m}{n})\; p_i^{(t-1)} + \tfrac{m}{n} \textstyle\sum_j p_j^{(t-1)} \\ &= \left[1 - \left(\tfrac{n}{n-1}\right) m\right]\; p_i^{(t-1)} + \left(\tfrac{n}{n-1}\right) m\; \bar{p}^{(t-1)}, \end{aligned} \tag{IV-20}$$

where $\bar{p}^{(t)} = \sum p_i^{(t)}/n$ is the mean gene frequency in all colonies in generation t. This shows that in each generation, the gene frequency p_i is obtained by averaging its previous value with the overall mean gene frequency $\bar{p}$. If we take (IV-20), sum both sides over i and divide by n, we get:

$$\begin{aligned} \bar{p}^{(t)} &= \left[1 - \left(\tfrac{n}{n-1}\right) m\right] \bar{p}^{(t-1)} + \left(\tfrac{n}{n-1}\right) m\, \bar{p}^{(t-1)} \\ &= \bar{p}^{(t-1)} \end{aligned} \tag{IV-21}$$

which shows us that the mean gene frequency over all populations does not change through time. Note the quantity $nm/(n-1)$ which appears in all of these expressions. It is nearly equal to m if n is large (when $n = 20$ it is $20n/19$). Let us call this quantity m^*. Now we can rewrite (IV-21) as

$$p_i^{(t)} = (1 - m^*) p_i^{(t-1)} + m^* \bar{p}. \tag{IV-22}$$

This is really just (IV-14) with somewhat different notation, and we can immediately see what will be the behavior of the gene frequencies in the individual populations. Each population's gene frequency moves a fraction m^* of the way toward its ultimate equilibrium value each generation. The equilibrium value is given by the overall average gene frequency $\bar{p}$. So migration has an averaging effect, bringing all gene frequencies to a common value but not actually changing the overall gene frequency. As we shall see, this property of migration is general to a large class of migration schemes.

The rate at which gene frequencies approach their common equilibrium is nearly given by m, with the time to move a substantial fraction of the distance to equilibrium being roughly $1/m$. This has an intuitive interpretation as the time scale on which a large fraction of the population has been replaced by immigrants.

MORE GENERAL MODELS. Comparable derivations can be done in the stepping-stone models, although things become more difficult there. The rate of approach to equilibrium is then quite a bit more complicated to obtain. But the qualitative results are similar. Even in the most general scheme, the migration matrix, certain generalities emerge. We shall state them intuitively here but not give derivations:

1. *At equilibrium, all populations have the same gene frequency.* It is rather easy to use (IV-17) to show that a state in which there is no geographic differentiation is an equilibrium state. One uses the fact that each row of the matrix **M** must sum to 1, since each gene in the population must have come from somewhere. It is not so easy to show that this is the only kind of equilibrium state - in part because that is not so. If we are allowing full generality in the pattern of migration, we allow cases where one can easily see that other equilibria exist. For instance, there is the case in which there is no migration at all, so that $m_{ij} = 1$ and all other m_{ij} are zero!

In such a situation it is immediately apparent that whatever gene frequencies we start with will simply remain unchanged. Another case involves two subsets of populations. There is exchange of migrants within each subset, but no migration between different subsets. In that case each subset could have a uniform gene frequency, but the gene frequencies in the two subsets could differ substantially.

It is not even ruled out that there may be no equilibrium at all. If we have two populations, and $m_{12} = m_{21} = 1$, then in each generation the ancestry of each population is supplied entirely by the other. So if we start with different gene frequencies (say 0.6 and 0.3) the gene frequency of the first population will oscillate: 0.6, 0.3, 0.6, 0.3, ... In effect, the two populations change places every generation.

All of which establishes that the above "generality" requires some special conditions. We will not attempt to find the least restrictive possible conditions generating convergence to an equilibrium at which all gene frequencies are equal, but the following rather weak conditions will suffice: each population must have some individuals who are nonmigrant (so that all the m_{ii} must be nonzero), and it must be possible for any given population to receive immigration from any other if we wait a sufficient number of generations. The second condition simply requires that the populations form a connected set with regard to migration - that no subset of populations be isolated from the others. An example where we can readily verify that these conditions hold is the stepping stone model. There will be no isolated sets of populations (provided that $m > 0$) and some individuals will not migrate (provided that $m < 1$).

2. *The rate of approach to equilibrium is controlled by m*. Usually the rate of approach will be relatively close to m, where m is an appropriately defined migration parameter. This is a rather hazy principle which cannot be more precisely stated without making quite elaborate theorems. Suffice it to say that when all migration rates are small multiples of m, and we take m small, it is usually found that halving m halves the rate of approach to equilibrium. In many cases the rate of approach will itself be a small multiple of m.

Although perfect generalities are hard to come by, the reader will not be seriously misled by the following conclusion:

Migration tends to smooth out geographic differences in gene frequencies. The rate at which this occurs is given by the rate of migration.

IV.7 History and References

The earliest work on migration involved admixture computations, in the work of Bernstein (1931). Glass and Li (1953) applied the notion of recurrent migration to admixture computations. The one-island model was first used by Haldane (1930a), and the island model was invented by Sewall Wright (1931), who used it to investigate genetic drift effects on gene frequencies, which we will cover in Chapter VII. The great Russian mathematician A. N. Kolmogorov (1935) did some early computations involving means and variances of gene frequencies in island models. The stepping stone model was invented independently by Malécot (1950) and Kimura (1953), although both were apparently influenced by the model of a population continuously spread out in a spatial continuum, propounded by Wright (1940). The migration matrix approach has only been investigated rather recently (Cavalli-Sforza and Zei, 1967; Bodmer and Cavalli-Sforza, 1968). The constant in most of these investigations is that migration in and of itself is such a simple evolutionary force that these papers are largely investigations of its interaction with other evolutionary forces such as selection or genetic drift. We will cover one of these interactions later in this chapter and the other in a subsequent chapter.

Just as simple mixture creates linkage disequilibrium, so does recurrent migration. A model of this sort was first put forward by Li and Nei (1974). Feldman and Christiansen (1974) presented an interesting model in which migration maintained a gradient of gene frequencies (a *cline*) and in which linkage disequilibrium persisted as a result of recurrent migration.

IV.8 Migration vs. Selection: Patches of Adaptation

Migration as an evolutionary force continually works to make the genetic composition of different populations more similar. Under certain circumstances natural selection will be operating in the opposite direction, and it is clearly of interest to see to what extent one force will prevail over the other.

The simplest model available seems to be that in which a single population is subject to selection different from its neighbors. Selection will then work to differentiate its gene frequency from its neighbors', but migration will work against this. Let us start by examining a haploid model.

A ONE-ISLAND HAPLOID MODEL. A single island lies near a continent. On the continent natural selection keeps allele a fixed. On the island the A allele is favored. We consider a one-locus two-allele haploid (or asexual) population with discrete generations. The fitnesses of A and a individuals on the island are 1 and $1 - s$. As far as migration is concerned, the model is a simple one-island model. A fraction m of the individuals each generation are immigrants. There is no migration back to the continent,

or if there is it has no influence on the gene frequencies on the continent.

This model is erected as the simplest possible model in which a local patch of genetic adaptation to a local environment is continually in danger of being swamped by immigration. What we are interested in is the conditions under which local adaptation can be maintained in the face of gene flow. When it can be maintained, we also want to know how strong a genetic differentiation can be maintained.

The changes of gene frequency are readily computed under this model. If p is the gene frequency of A, after selection

$$p^* \;=\; \frac{p}{1-(1-p)s}, \tag{IV-23}$$

and after migration

$$\begin{aligned} p' \;&=\; (1-m)\; p^* + m \times 0 \\ &=\; (1-m)\; p^* \\ &=\; \frac{p\,(1-m)}{1-(1-p)s}. \end{aligned} \tag{IV-24}$$

The reader may have noticed that these are precisely the equations for the balance between mutation and selection in haploids, equations (III-13) through (III-15). The analogy between immigration and mutation is generally useful, and much of the mathematics is basically the same, with the replacement of the mutation rate μ by the immigration rate m.

From (IV-24) we can compute the change of gene frequency:

$$\Delta p \;=\; p' - p \;=\; \frac{p\;(1-m)}{1-(1-p)s} - p \;=\; \frac{-m\;p + p(1-p)s}{1-(1-p)s} \tag{IV-25}$$

There are two values of p for which $\Delta p \;=\; 0$. These are the values at which the numerator of (IV-26) is zero. They are when $p = 0$ or when

$$p \;=\; 1 - m/s. \tag{IV-26}$$

These correspond to the two possible fates of the patch of adaptation: it may be lost ($p = 0$) or it may be maintained in the face of continued gene flow.

Our interest is in which of these equilibria is stable. This will reflect the sign of Δp, and that in turn is solely a function of the numerator of (IV-25). The denominator, being the mean fitness, can never be negative. The numerator is $p[-m + (1-p)s]$. There are two cases of interest. If $s \leq m$, then a moment's consideration will show that $(1-p)s$ can never be greater than m, so that the numerator of (IV-26) will always be negative. This is exactly the case in which the equilibrium gene frequency $1 - m/s$ will be negative.

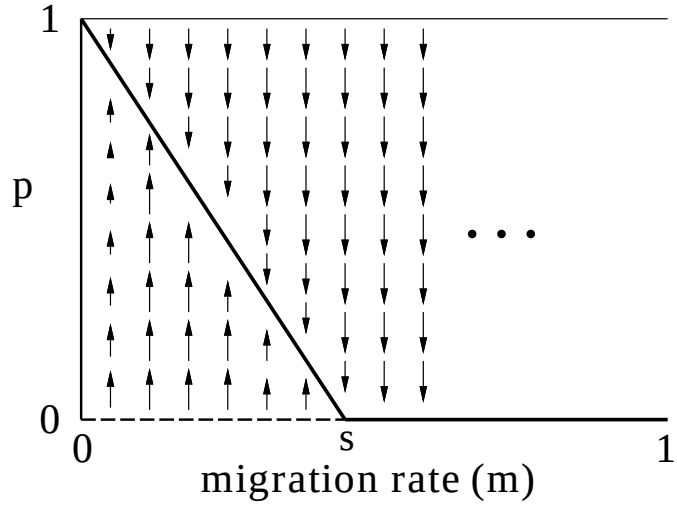


Figure 4.4: Equilibrium frequencies of a locally-favored allele on an island in the haploid case, with migration coming from a nearby continent at rate m. The equilibrium frequencies are shown in bold lines for different values of the migration rate. The arrows show the directions of gene frequency change for different migration rates. As discussed in the text and shown in equation (IV-26), the equilibrium frequency declines to zero as the migration rate is increased, and is zero for cases where the migration rate reaches or exceeds the selection coefficient against the continental allele.

So we get a picture which is consistent: only the equilibrium $p = 0$ makes any sense, and consideration of Δp shows that p will continually decrease toward that equilibrium. When $m \geq s$, migration which brings in a's overwhelms the selection which is trying to maintain some A's in the patch.

When $s > m$, the picture is different. The quantity $-m + (1 - p)s$ will sometimes be positive and sometimes negative. In particular, it will be positive when $p < 1 - m/s$ and negative when $p > 1 - m/s$. So in this case the gene frequency of A will rise toward the equilibrium value when it is below it and fall toward it when above it. To prove that the equilibrium is truly a stable one, we would have to also show that there is no overshooting of the equilibrium (or at least, not enough to allow any oscillations of increasing amplitude). This can be done without much difficulty, but we will not allow this matter to detain us here.

The picture which emerges is a fairly simple one. When $m \geq s$, migration overwhelms selection and a patch of adaptation to local conditions cannot be maintained. When $s > m$, local adaptation can be maintained, but the frequency of the locally favored allele will be only $1 - m/s$, so that some a alleles will always be present, in an

equilibrium between their introduction by migration and their elimination by selection. The behavior is shown in Figure 4.4.

This result forms a reasonably consistent picture with our intuitive feeling that migration as an evolutionary force operates with a speed given by m, and selection with a speed given by s, so that maintenance of a patch of local adaptation depends on whether selection can eliminate inappropriate alleles as fast as they come in. We can also make an intuitive analysis similar to the mutation models, using the fact that in each generation a fraction m of the copies at this locus are a alleles that have newly immigrated, and that each of these copies has descendants that persist an average of $1/s$ generations. These can be multiplied to predict that their equilibrium frequency will be m/s.

DIPLOID MODELS. The extension of this model to diploidy reveals new phenomena, so that it is important to look at that case. Suppose that migration occurs after selection, just before random mating within the patch. It will then alter the gene frequencies within the patch, but there will be Hardy-Weinberg proportions before selection. We will investigate only the two extreme cases - complete dominance and complete recessiveness of A. When A is dominant things are fairly straightforward. The equation for change of p is readily found to be

$$p' \;=\; \frac{p\,(1-m)}{1-s\,(1-p)^2} \tag{IV-27}$$

(s is the selection coefficient against aa). The condition for A to increase in the patch becomes simply

$$1-m > 1-s\,(1-p)^2 \tag{IV-28}$$

or

$$m \;<\; s\,(1-p)^2. \tag{IV-29}$$

As in the haploid case, this is essentially the same equation as we found in the case of mutation to a recessive deleterious allele (here the non-locally-adapted allele is recessive). Compare this equation to (III-21). A little examination of this condition fills in most details of the behavior of gene frequencies in this case. When $m > s$, there is no equilibrium frequency of A, which declines in frequency until it disappears. When $m < s$, migration does not overwhelm selection, but there is a stable equilibrium at

$$p_e \;=\; 1-\sqrt{m/s}. \tag{IV-30}$$

if A is introduced into the patch at low frequency it will increase to this equilibrium value. If the patch starts out all AA, the frequency of A will be reduced by immigration to this equilibrium.

The picture with dominance of A is much the same as with haploidy: the ratio between m and s controls the frequency of A. No A genes will persist if $m \geq s$. When s is much larger than m almost all the genes in the patch will be A.

Recessive adaptations. With the favored allele A being recessive things become more complex, and a bit strange. The fitnesses are

$$\begin{array}{ccc} AA & Aa & aa \\ 1 & 1-s & 1-s \end{array}$$

and the equation for change of p is

$$p' \;=\; \frac{p\,[1-(1-p)s]\,(1-m)}{1-[2p(1-p)+(1-p)^2]\,s} \tag{IV-31}$$

and the condition for p to increase is

$$[1-(1-p)s]\,(1-m) > 1-(1-p^2)s \tag{IV-32}$$

which reduces to

$$-p^2 s \;+\; p\,s(1-m) \;-\; m(1-s) > 0. \tag{IV-33}$$

We are only interested in cases in which $s > 0$, so we can immediately see that when p is near zero, it will not increase, since (IV-33) is not satisfied if we set $p = 0$. When p is near 1, the left-hand side of (IV-33) has the value $-m\,(1-s)$, which shows that p will decrease when large. This quadratic thus has a negative value at $p = 0$ and at $p = 1$. The coefficient of the p^2 term is negative, showing that the parabola opens downward. The point at which this parabola achieves its maximum can be found by equating its derivative (with respect to p) to zero:

$$-2\,p\,s + s\,(1-m) \;=\; 0 \tag{IV-34}$$

which shows that the maximum is achieved at $p = (1-m)/2$. Putting together this information, keeping in mind that (IV-33) is satisfied when p increases, we must have one of the two following circumstances: In the first case there are two equilibrium values (these are the points at which the parabola intersects the axis). The lower one is unstable, with p decreasing below it and increasing above it. The upper one is a stable equilibrium, though to complete the demonstration we would have to show that there is no oscillating overshooting of the equilibrium. In the second case, there are no equilibria, except for the equilibrium $p = 0$ which can be seen in (IV-32). The A allele will continually decline until it is lost. It remains for us to find out which of these cases applies. This can be done by examining the discriminant of the quadratic (the expression which you find under the square root sign when you solve the quadratic). For roots to exist which are not complex numbers, the discriminant must be positive. For the quadratic on the left side of (IV-33) this yields

$$s^2(1-m)^2 \;-\; 4m\,s(1-s) > 0 \tag{IV-35}$$

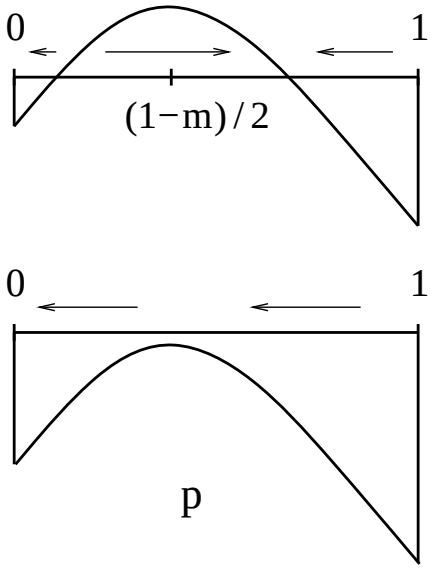


Figure 4.5: Change of gene frequency in the one-island model with migration opposed by selection, with the locally favored allele recessive. The horizontal line is zero change of gene frequency. In the upper case the patch of local adaptation can persist, in the lower case it is lost.

or since $s > 0$, provided s and m are both less than one,

$$\frac{s}{1-s} > \frac{4m}{(1-m)^2} \tag{IV-36}$$

When we consider weak selection versus weak migration, the terms in s^2 and m^2 generated by the denominators of both sides of this inequality are small compared to s and m, so that the condition for existence of a nonzero equilibrium frequency of A becomes essentially $s > 4m$, just as in all the other cases.

But case of A recessive is different. While it is true that an equilibrium frequency of A exists, and is a stable equilibrium, when s (roughly) exceeds m, the behavior of A when rare is very different. It can never increase when rare for there is an unstable equilibrium on the p-axis between 0 and the stable equilibrium. There is a simple interpretation of this behavior. When A is rare it has very little selective advantage, for hardly any of the A genes are in the advantageous AA homozygotes. There is a constant introduction of a fraction m of a's into the island or patch. For A to increase the effective selection coefficient favoring A must be large enough to overcome the continued dilution of A genes by migration. In the recessive case, when A is rare its selective advantage is nil, and it cannot invade the patch even though it might be able to persist if introduced in high enough

frequency. This is an unusual example of a type of behavior which in physics would be called hysteresis - the population can (if $s > m$) maintain a high frequency of A, but if that frequency is perturbed to near zero, it will not return. Recalling the terminology of our discussion in chapter II, this polymorphism is stable, but not protected.

One consequence of this phenomenon should be that if an allele arises which is locally advantageous, but that allele is recessive, there will be a reduced chance that the allele is able to become the basis of a patch of local adaptation. There should thus be a selection for the dominance of locally-advantageous alleles, by weeding out those mutants which are recessive.

Another, entirely different consequence of these results is that we should not expect to see locally-favored recessive alleles maintained at frequencies much below 1/2. Recall that our quadratic in p has a maximum at $p = (1 - m)/2$, which will often be slightly below 1/2. Consider what happens as we look at a series of cases with different values of m. When m is small A will be maintained at a high frequency. The unstable equilibrium will be near $p = 0$. To see this consider that when $m = 0$, one can see from (IV-33) that $p = 1$ is an equilibrium and so is $p = 0$. When we have m small, the small amount of migration will reduce the stable equilibrium a little below $p = 1$ and increase the unstable equilibrium to a bit above $p = 0$. As m is gradually increased, both equilibria move toward $p = (1 - m)/2$ from opposite sides. When m reaches the critical threshold value defined by (IV-36), both equilibria collide at $p = (1 - m)/2$, and thereafter there are no equilibria other than $p = 0$. So the stable equilibrium frequency of A is reduced from $p = 1$ down to just below 1/2 as m is increased, and then the pocket of local adaptation suddenly collapses, with natural selection no longer able to maintain it in the face of migration. This behavior may be contrasted with the dominant and the haploid cases, where the collapse of the patch occurs as the equilibrium gene frequency (as can be seen from (IV-26) and (IV-30)) reaches zero.

REFERENCES. The case of an island with locally-adapted alleles was explored rather fully by Haldane (1930b), who observed all of the above phenomena. Sewall Wright (1931) was aware of the same phenomena at about the same time. Nagylaki (1975a) has examined these phenomena more fully for a model with weak selection, in the context of consideration of the more complicated case of geographic continuum.

IV.9 Two-Population Models

When we considered the effect of migration on a single island or patch of adaptation, no account could be taken of the effect of the island on its neighbors. In most realistic models of the geographic structure of natural populations, a patch or local population will export some of its genes to its immediate neighbors. This in turn means that some of the genes which migration brings into a population are these recently-exported genes.

Since neighbors export part of their gene pools to each other, they thereby reduce the genetic impact of the genes they receive from each other. When two-way migration is allowed, we should expect that maintenance of a patch of local adaptation will not be quite so difficult as in the one-island model.

The simplest model which can show this phenomenon has two islands exchanging migrants. Let us consider the most symmetric and simplest form of this two-island model. Imagine two populations, each of which receives a fraction m of its genes from the other each generation. Each population is haploid (or asexual) with discrete generations. The fitnesses of alleles A and a in population I are $1+s:1$, and in population II they are $1:1+s$. Thus each allele is favored in its own population, and the selection coefficients are taken to be equal. Let us call the gene frequencies of A in the two populations p_1 and p_2. We assume that these frequencies are measured at the beginning of the generation, and that selection precedes migration. After selection

$$\begin{aligned} p_1^* &= \frac{p_1\,(1+s)}{1\ +\ sp_1} \\ \text{and} & \\ p_2^* &= \frac{p_2}{1\ +\ s(1-p_2)}. \end{aligned} \tag{IV-37}$$

After migration

$$\begin{aligned} p_1' &= (1-m)\,p_1^*\ +\ m\,p_2^* \\ &= \frac{p_1(1+s)(1-m)}{1+s\,p_1}\ +\ \frac{p_2\,m}{1+s\,(1-p_2)}, \end{aligned} \tag{IV-38}$$

with a similar equation for p_2'. We can look for equilibria of this system of equations by requiring that $p_1' = p_1$ and $p_2' = p_2$ and trying to solve this system of two equations for p_1 and p_2. But in this case we can exploit the symmetry of the model and simplify things.

A SYMMETRIC EQUILIBRIUM. It seems reasonable that the equilibria we are seeking should be of the form $p_1 = p, \quad p_2 = 1-p$. That is, if the equilibrium frequency of A in population I is p, the equilibrium frequency of a in population II is also p. This expectation comes from the symmetry of the model: if we exchange the names of populations I and II, and then exchange the names of alleles A and a, we still have the same model. It is not ruled out that there might be other kinds of polymorphic equilibria, but it seems worthwhile to see if there are any of this form. It is a simple matter to do this using (IV-38). If we substitute $p_1 = p, \quad p_2 = 1-p$, and require that also $p_1' = p_1 = p$, then the equation becomes

$$p = \frac{p\,(1+s)\,(1-m)}{1+sp}\ +\ \frac{(1-p)\,m}{1+sp} \tag{IV-39}$$

Multiplying both sides by the denominator $(1+sp)$ and rearranging to collect powers of p, we have the quadratic equation

$$s\ p^2\ +\ [2m + m\,s - s]\ p - m\ =\ 0 \qquad \text{(IV-40)}$$

This quadratic has the value $-m$ when $p = 0$ and $m(1+s)$ when $p = 1$, and since these are of opposite sign there must be a root (exactly one, since it is a quadratic) between 0 and 1.

The equation for equilibrium gene frequency is not a particularly revealing expression, being the solution to this quadratic. It is possible to make various approximations. The simplest is to notice that if we let m and s both be small, we can ignore the term ms as being smaller than terms in m or in s. The quadratic equation becomes (dividing by s)

$$p^2\ +\ (2m/s - 1)\ p\ -\ m/s\ \ =\ \ 0. \qquad \text{(IV-41)}$$

The salient fact about this equation is that its coefficients (and therefore its solution) depend on m and s only through their ratio. Figure 4.6 shows values of p obtained by solving (IV-40) for different values of s and m/s. The dependence of p chiefly on m/s when s is small is readily apparent.

When $m \ll s$, in the upper right corner of the diagram, the curve relating p to m/s is very close to $p = 1 - m/s$. That can be derived from the solution of the quadratic equation when s, m, and m/s are all taken to be small. It also makes good intuitive sense: when m is small the favored allele is nearly fixed in each population, and immigration brings in mostly the other allele, so that the situation is nearly the same as the one-island model, whose equilibrium frequency of A would be $1 - m/s$. When $m \gg s$, the frequency of A is nearly $1/2 + s/(8m)$, which can also be obtained as an approximation from the solution to the quadratic.

ASYMMETRY AND PATCH-SWAMPING. A stability analysis of this equilibrium can be done, although it involves two variables, p_1 and p_2, it involves matrix algebra and we will not reproduce it here. It can be shown that the equilibrium we have found is always stable. This may be a bit surprising, for in the one-island model the equilibrium disappeared when m was made large. In the two-island model the equilibrium never disappears or becomes unstable. In fact, there is in the two-island model a behavior which corresponds to the swamping of a patch of adaptation. It arises when the strength of selection in the two populations is unequal. This can be seen intuitively by imagining the case in which there is selection coefficient s_1 favoring A in population I, but in population II the selection favoring a is infinitely strong ($s_2 = \infty$). Then even if migration introduces some A's into population II, selection immediately kills them off and returns that population to consisting entirely of a. Migration from population II to population I thus always consists entirely of a's. What we have done is to make the dynamics of gene

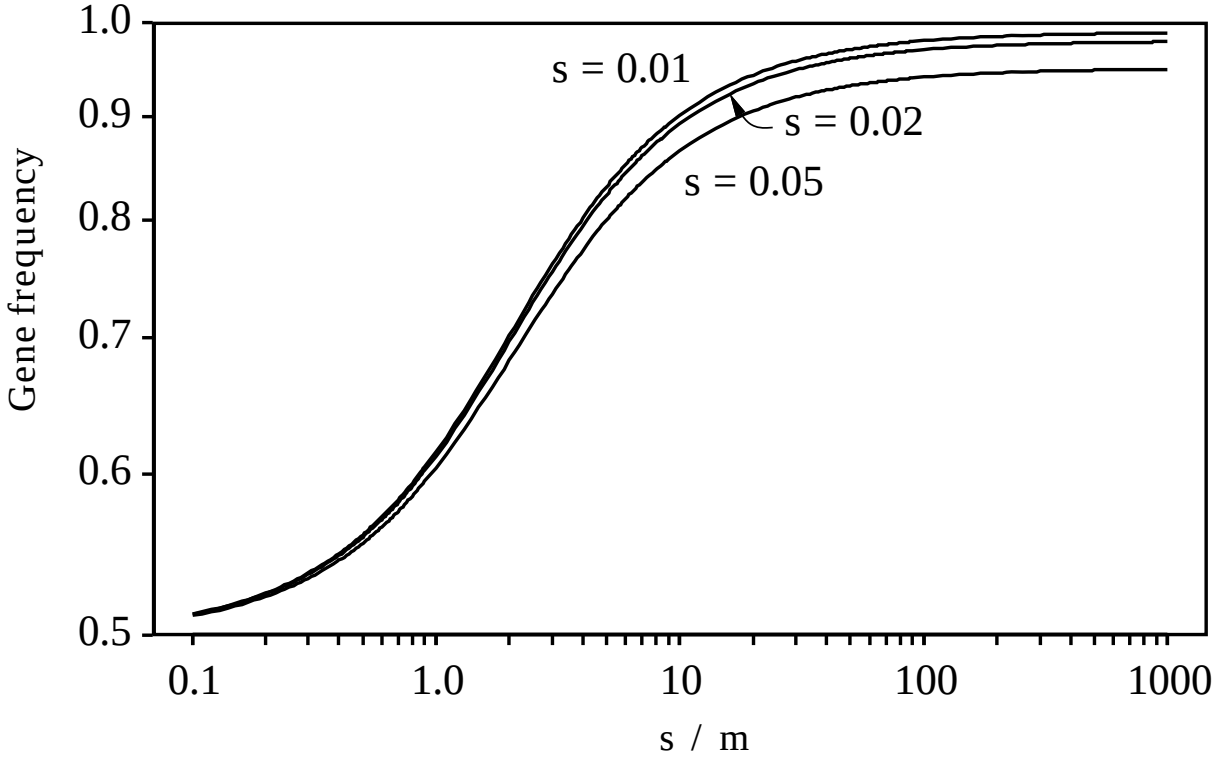


Figure 4.6: Gene frequency in the two-island model as a function of s/m. The upper two solid curves are for $s = 0.01$ and $s = 0.02$, the lower curve is for $s = 0.05$.

frequency in population I follow precisely a one-island model! When $m > s$, the A allele cannot persist in population I.

Asymmetry in the selection thus allows a sufficiently high rate of migration to abolish the patch of adaptation in the population with weaker selection. By choosing to examine a case with perfectly symmetric selection we missed this behavior. We will have a bit more information on this phenomenon in the next section.

All of our discussion has been in terms of haploid selection, but the equations are exactly the same if heterozygote fitnesses are the geometric mean of homozygote fitnesses. We could have considered partial dominance, but that would only have made life difficult by converting quadratic equations into cubics.

REFERENCES. The two-island model received little attention for many years. Moran (1959b) solved for equilibria in a diploid two-island model with weak selection and migration, and he also established the stability of this equilibrium and the instability of the other equilibria at which both populations are fixed for the same allele. Eyland (1971)

analyzed the case where selection is not symmetric, giving conditions for existence of the equilibrium when there is a possibility of one island's alleles swamping the patch of adaptation on the other island. Maynard Smith (1970) presented sufficient conditions for the maintenance of polymorphism in a two-island model.

IV.10 The Levene Model: High Migration

In 1953, Howard Levene set forth a model of selection and migration which it is useful to discuss at this point because it gives us some insight as to when the contradictory selection in different populations gives rise to adaptations which cannot coexist. Levene's model constitutes the extreme in migration, complete random mating among all populations. There is a single pool of mating individuals who mate at random. The i-th population contributes a fraction c_i of them. The resulting offspring are distributed at random among the populations (let us say that there are n populations). Because the same mating pool contributes all the offspring, in a diploid two-allele case each population starts out with frequencies p^2, $2p(1-p)$, and $(1-p)^2$ of the three genotypes. This is what makes the Levene Model so easy to analyze - the overall random mating leaves us with only one variable, p, to follow.

Within each population selection occurs. Suppose that the i-th population has fitnesses $w_i : 1 : v_i$ of the genotypes AA, Aa, and aa. Then after selection the gene frequency of A in the i-th population will be

$$p_i \;=\; \frac{p^2\, w_i \;+\; p(1-p)}{p^2\, w_i \;+\; 2p(1-p) \;+\; (1-p)^2 v_i} \tag{IV-42}$$

as a result of the usual selection formulae. Now each population contributes a given fraction c_i of the mating pool in the next generation. The overall gene frequency p' in that mating pool is simply the weighted average $\sum c_i\, p_i$, which is

$$p' \;=\; \sum_i \; c_i \; \frac{p^2\; w_i + p(1-p)}{p^2\; w_i + 2p(1-p) + (1-p)^2\; v_i} \tag{IV-43}$$

While this looks like a complex expression, it involves only one variable, p. This makes analysis quite simple.

If we look for equilibrium values of p, there are of course $p \;=\; 0$ and $p \;=\; 1$. The obvious way to find equilibrium values between 0 and 1 is to set $p' \;=\; p$ and try to solve (IV-43) as an equation in p. Alas, this gives us a polynomial in p of order $(2n+1)$ which has no known explicit solution. However, we can easily get conditions for there to be a protected polymorphism by asking whether p will increase near $p \;=\; 0$ and decrease near $p \;=\; 1$. When p is small, then we can ignore the rare AA homozygotes and write

(IV-43), ignoring terms in p^2, (terms in p^2 in the numerator or in p in the denominator)

$$p' = \sum_i c_i \frac{p_i}{v_i}. \tag{IV-44}$$

The condition for p to increase is just $p' > p$, which from this equation is seen to be

$$\sum_i \frac{c_i}{v_i} > 1. \tag{IV-45}$$

There is an exactly analogous equation for the case in which $(1 - p)$ is small, namely

$$\sum_i \frac{c_i}{w_i} > 1 \tag{IV-46}$$

If both of these conditions are satisfied, there will be protected polymorphism. If not, then there may be stable polymorphic equilibria between 0 and 1, but polymorphism may disappear if the gene frequencies are perturbed to values near fixation.

The conditions (IV-45) and (IV-46) require, in effect, that the weighted harmonic mean (the reciprocal of the mean of reciprocals) of the heterozygote fitnesses exceed the weighted harmonic means of both homozygote fitnesses. Why weren't the results dependent on weighted arithmetic means? If the newborns are being randomly distributed among the populations, why didn't the maintenance of polymorphism just depend on the weighted average fitnesses $\sum_i c_i w_i$, 1, and $\sum_i c_i v_i$? The answer to this points up an important property of the Levene model. If a fraction c_i of the newborns were exposed to the selection regime of population i, and the survivors (let's assume for the moment that the selection is by differences in viability) were simply taken and pooled to form the parents of the next generation, then we could simply use arithmetic weighted mean fitnesses. The average probability that an AA individual would survive would then be $c_1 w_1 + \cdots + c_n w_n$. The conditions for polymorphism would simply be overdominance of these mean fitnesses.

The probability that an AA individual survives and contributes to the mating pool is, however, not this arithmetic mean fitness. Recall that c_i is not the proportion of newborns allocated to population i, but the fraction of the mating pool which population i ends up contributing. The contribution of each population is fixed, independent of the fitness of individuals in that population. Even if most of the individuals in population i are of very low fitness, the contribution to the mating pool is still c_i. This in effect assumes that selection in each population precedes density-dependent population size regulation, which takes place *separately in each population*. The population size, and hence its contribution to the mating pool, is determined after selection, and essentially independently of it. Thus an individual of low fitness in a given population may be more or less likely to end up in the mating pool, depending on the fitnesses of the others in its population.

An example will help clarify the point. Let's consider for the moment a haploid two-allele case, with two populations. The two genotypes A and a are equally frequent. In one population A is twice as fit as a, in the other the reverse:

		fitness of	
i	c_i	A	a
1	1/2	0.5	1
2	1/2	2	1

If c_i represented the probability of an individual landing in population i, and there were no density regulation in each population separately, then the mean fitness of A would be $(0.5+2)/2 = 1.25$, and the mean fitness of a would be $(1+1)/2 = 1.0$. We expect A to increase. But with density regulation in each population with c_i being the contribution population i makes to the mating pool, we find that in population 1 the gene frequency of A after selection is $1/3$, and in population 2 it is $2/3$. The resulting genotype frequencies in the mating pool are $1/2 : 1/2$, the same as the initial genotype frequencies. The separate population regulation in each population has yielded a different outcome. That this is the critical way in which Levene's model differs from a single random-mating population was first pointed out by Dempster (1955).

The frequency-dependence of overall fitnesses which is induced by having densities regulated in each population makes it easier to maintain an overall polymorphism than if there were only one population. In the diploid case, this is a general phenomenon. Allele A changes when rare as if it had an overall fitness of 1 while the common aa homozygote had a fitness of $1/(\sum_i c_i/v_i)$, the weighted harmonic mean of the aa fitnesses. Likewise when a is rare, the frequency of A changes as if the fitnesses of $AA : Aa$ were $1/(\sum_i c_i/w_i) : 1$. It can be proven that harmonic means are never larger than the corresponding arithmetic mean. So the effect of having density regulation in each population is to lower the effective fitnesses of the homozygotes, making polymorphism easier to envisage. The reader may find it useful to compute some numerical examples.

In effect there are two things that we mean by the word "population". The first is a group of organisms in perfect competition. The second is a group of organisms which mate at random. Levene's model presents us with a divergence between the two definitions, with competition only within populations, but mating at random over all populations.

We can use Levene's model to find conditions under which polymorphism will be protected in the two-population model of the previous section, provided $m = 1/2$, which is equivalent to having one mating pool covering both populations. Our model of that section was a haploid model, but it should have exactly the same dynamics as if the

population were diploid with geometric fitnesses. So let us consider the following case:

		fitness of		
i	c_i	AA	Aa	aa
1	1/2	$1+s_1$	1	$1/(1+s_1)$
2	1/2	$1/(1+s_2)$	1	$1+s_2$

Each population shows geometric fitnesses with heterozygote relative fitness 1, and with selection acting in opposite directions in the two populations (assuming s_1 and s_2 are both positive). We have allowed the selection coefficients in the two populations to differ. The object is to investigate by how much s_1 and s_2 have to differ to eliminate the possibility of stable polymorphism. Of course our answer will only tell us about *protected* polymorphism, and then only when $m = 1/2$.

The conditions for increase of A when rare are, from (IV-45)

$$\frac{1}{2}\left(\frac{1}{1+s_1}\right)+\frac{1}{2}\left(\frac{1}{1/(1+s_2)}\right)>1 \tag{IV-47}$$

or

$$\frac{1}{2(1+s_1)}+\frac{1+s_2}{2}>1 \tag{IV-48}$$

and the corresponding condition for increase of a is the same with subscripts 1 and 2 interchanged:

$$\frac{1+s_1}{2}+\frac{1}{2(1+s_2)}>1 \tag{IV-49}$$

We want to know when both of these will hold. If we express each of these by finding the limits on s_2 as a function of s_1, we get from (IV-48)

$$s_2>\frac{s_1}{1+s_1} \tag{IV-50}$$

and from (IV-49)

$$s_2<\frac{s_2}{1-s_1} \tag{IV-51}$$

yielding the overall condition

$$\frac{s_1}{1-s_1}>s_2>\frac{s_1}{1+s_1} \tag{IV-52}$$

These are fairly tight limits. If $s = 0.1$, then they require that

$$0.1111>s_2>0.0909$$

so that s_2 must be (roughly) within 10% of the value of s_1 or one patch of adaptation will swamp the other out of existence. As the selection coefficients become smaller, the conditions become more restrictive. When $s_1 = 0.01$,

$$0.0101 > s_2 > 0.0099 \qquad \text{(IV-53)}$$

which means that s_2 must now be within 1% of s_1.

We can conclude that when selection in the two populations is of unequal strength, then at high migration rates one patch of adaptation swamps the other. At low migration rates it is much easier to maintain polymorphism: when $s_1 = \infty$ we are in a one-island model, and we will still have polymorphism (in the form of the maintenance of local adaptation to local conditions) no matter what the value of s_2, provided that $s_2 > m$. So both patches can maintain adaptation to local conditions, provided that either the selection coefficients in each population are nearly equal, or that migration is sufficiently restricted. Correspondingly, the swamping of a patch of local adaptation will nearly always occur if there is sufficient migration, unless the strengths of selection in different populations are rather precisely balanced.

We have thus been able to use Levene's conditions for protected polymorphism to get some insight into the behavior of the two-island model. Eyland's (1971) conditions for the local stability of polymorphism in a two-island model differ slightly from the above, but show the same general patterns. Eyland used an additive rather than a geometric fitness pattern, and made an analysis restricted to small values of s_1 and s_2, so it is not surprising that the results should differ, and it is comforting that the differences between Eyland's and our conditions disappears as s_1 and s_2 are made small. Karlin and McGregor (1972) have presented a method of small parameters, a method which is a version of one widely used in applied physics, which can easily be used to show that polymorphism will always be stable if m is sufficiently near zero.

REFERENCES. In addition to Levene's and Dempster's papers, there has been some further general work on the Levene Model. Li (1955) showed that under Levene's model gene frequency changes in the direction that moves uphill on a surface which is the weighted geometric mean of the individual population mean fitnesses (as a function of p). Cannings (1971, 1973) generalized Li's result to multiple alleles. Gliddon and Strobeck (1975) showed that the Levene Model could maintain polymorphism in haploids. Maynard Smith (1962) and Prout (1968) gave conditions for protected polymorphism when one allele is completely dominant, in which case Levene's result needs to be supplemented by other conditions. Karlin (1977) has presented a general mathematical analysis of some special cases of Levene's model, as well as intuitive speculations on more general patterns.

IV.11 Selection-Migration Clines

Having examined these simple geographic structures, we can more readily understand more realistic situations. Complete realism is unattainable, save at the sacrifice of mathematical tractability. The compromise we examine in this section is the simplest models yielding *clines*, situations in which a regular geometric arrangement of populations and a simple pattern of selection yields a smooth unidirectional pattern of change in gene frequency. Clines were defined by Julian Huxley as smooth gradients in the average value of a character. Here we take the character to be gene frequency. We will look at some simple arrangements which generate clines.

CLINES IN A STEPPING-STONE MODEL. We start with some numerical results. We consider an infinite one-dimensional stepping-stone model, with discrete generations. Each population receives a fraction $m/2$ of its mating pool from each of its two neighbors, and $1 - m$ from itself. Each generation consists of selection, followed by migration, and then random mating. The populations are numbered with consecutive integers -5, -4, -3, -2, -1, 0, 1, 2, ... Selection is assumed to involve geometrically intermediate heterozygotes: alternatively we may regard this as a haploid model. Each population has its own selection coefficient on A (as usual, we have two alleles).

All that we have done is to generalize to a stepping-stone model the two-population selection-migration model of previous sections. It will be convenient to alter the model of selection slightly to make it additive rather than geometric: the fitnesses of $A : a$ (if we consider the model haploid) are $1 + s_i/2 : 1 - s_i/2$. This has the merit of being symmetrical: a selection coefficient $+s$ in favor of A should be exactly as much selection as one of $-s$ in favor of a. The equations for change in gene frequencies are

$$p_i^* = \frac{p_i\,(1 + s_i/2)}{1 + s_i\,(p_i - 1/2)}, \tag{IV-54}$$

and

$$p_i' = (m/2)\,p_{i-1}^* + (1 - m)\,p_i^* + (m/2)\,p_{i+1}^*. \tag{IV-55}$$

No algebraic solution of these equations is known, for any other than biologically uninteresting cases (such as all s_i the same). There are, after all, an infinite number of nonlinear equations to solve, even to get equilibria by requiring that $p_i' = p_i$ for all i. A little examination will show that there are always the solutions in which all the p_i are 1 or all are zero. After all, neither selection nor migration can create an allele when it is absent. But the solutions we are most interested in would have p_i near 1 in regions where $s_i > 0$ and near zero when $s_i < 0$.

While finding solutions algebraically may be impossible, there is no difficulty in simply iterating them numerically. We are interested for the moment in two patterns of selection. In one, which I call the "step", the selection coefficient shifts abruptly as

an environmental boundary is crossed. For example, in populations 27, 28, 29, ... the selection coefficient favoring A might be s, and in populations 1, 2, ... 23, 24, 25, 26 it might be $-s$. Each allele is favored in its own region. The other pattern I call the "ramp". It represents the case where an environmental factor changes smoothly, with selection coefficients against A gradually weakening until A becomes favored, at first slightly, then more strongly. The selection coefficient s_i in this pattern might look like this:

$$-3.5s, -2.5s, -1.5s, -0.5s, \ 0.5s, \ 1.5s, \ 2.5s, \ 3.5s, \ldots .$$

Of course we cannot actually numerically iterate the full set of equations (IV-54) and (IV-55), for we would have to compute an infinite number of quantities every generation. Instead, we take a suitably large number of populations (in the present case 16), so that the terminal colonies are nearly fixed one way or the other. The finite stepping stone differs from the infinite stepping-stone model in that the end populations do not receive immigration from neighbors on one side. If all these populations are fixed for the same allele, then it makes no difference whether or not the end colony receives immigrants from one or both sides.

Figure 4.7 shows the equilibria which result when we iterate these equations. In both the step and the ramp patterns of selection clines are produced. The terminal populations are so near fixation that we can have some confidence that the infinite stepping-stone model would show a similar pattern.

The clines are similar in shape - in fact the strengths of selection were chosen so as to result in clines of similar slope near the center. Notice that a cline resulting from a step pattern of selection is not itself a step pattern of gene frequency. Migration introduces alleles from each side of the environmental transition into the other, rounding off the pattern of gene frequencies into a smooth cline. The "step-cline" beloved of some evolutionists - the sudden absolute change in a character as a boundary is crossed - is more of a myth than a reality. It is only possible when selection is infinitely strong, so that migrants crossing the boundary are instantly killed, or else when no migrants at all cross the boundary. A smooth cline of gene frequency can be the result of either a sudden transition of selection coefficients (a step) or a smooth change (ramp). Although there are quantitative differences in the shape of the cline in these two cases, a glance at the Figure should convince you that our chances of distinguishing between these patterns in the field is nil, given sampling errors, historical perturbations, and the geographical inhomogeneities of actual patterns of migration and selection.

Another sort of information we get from iterating these equations numerically is the rate at which the equilibrium cline shape is approached. We do not have space for an extensive treatment of this matter, but suffice it to say that the fraction of the distance toward the final cline shape which the collection of populations moves is roughly indicated by the sizes of the migration rate m and the selection coefficients s_i. It is a function of all of these but is about the size of the larger of m and s_i, at least, those s_i in the middle

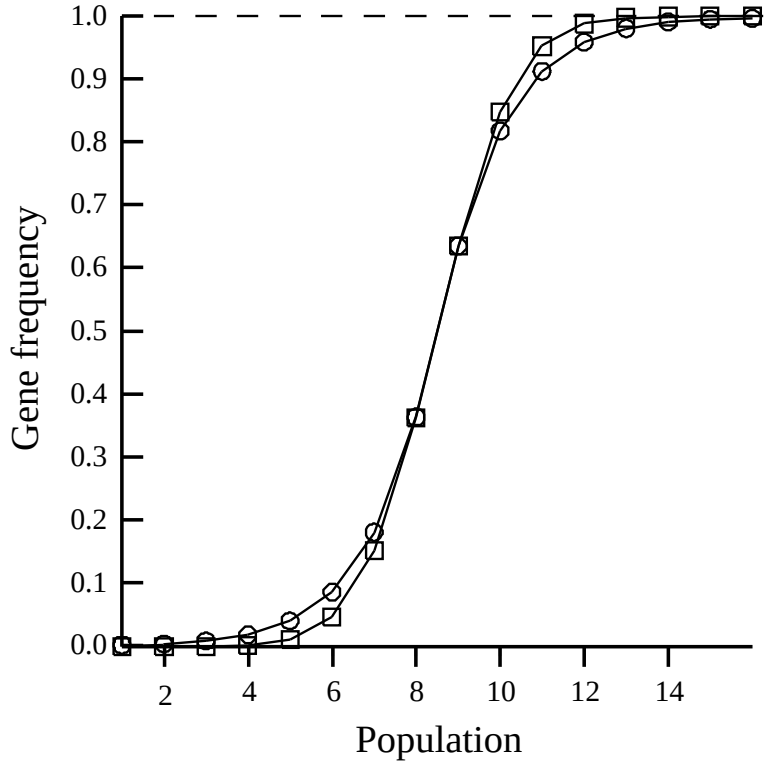


Figure 4.7: Clines in a ramp case (squares) and in a step case (circles). In both cases $m = 0.3$. In the ramp case $s = 0.1$ and in the step case $s = 0.09$. This approximately matches the slope in the center of the clines so that the differences in their shape are more easily seen.

of the cline. So if we have a step pattern of selection, with $s = m = 0.1$, it will take on the order of ten generations to move a substantial distance toward the ultimate cline, while if $s = m = 0.01$, it will take on the order of 100 generations. In the case of a step pattern, a gene from one region survives an average of $1/s$ generations in the other region before succumbing to selection. So the A alleles in the region in which a is favored have existed there no more than $1/s$ generations. The cline should re-establish itself in at most a few multiples of that time if we eliminate each allele from the "wrong" region.

There is only a modest amount of theory for clines in stepping-stone models. The small-parameter theory of Karlin and McGregor (1972a) shows that for sufficiently small values of the migration rate m, the cline is stable. In stepping-stone models, Bruce Walsh (1983) has given reasonably tight bounds on how much selection is needed to maintain a cline. Walsh's result takes into account the migration rate, and improves considerably on a much looser bound obtained by Karlin (1976, 1982) which applies for all migration

rates.

APPROXIMATE SOLUTIONS OF CLINES: A DIFFERENTIAL EQUATION. If we assume that m and the s_i are small, and alter the geography a bit, then we can convert the system of nonlinear equations (IV-54) and (IV-55) into a differential equation. From that equation we can get information on the slope of the cline, and in some cases we can solve for the entire shape of the cline. The change in geography is to imagine that populations are packed so closely in space that we in effect have a true continuum of populations. We can index position by a coordinate x which runs from $-\infty$ to $+\infty$. At position x the gene frequency is $p(x)$ at equilibrium, and the selection coefficient favoring A is $s(x)$. As before we consider a haploid (or geometrically-intermediate diploid) population with discrete generations.

Since in a continuum there is no notion of the "next" population, we must alter the migration scheme a bit. Let $M(y)\,dy$, for a very small interval of width dy, be the probability that an individual found at point x after migration came from the interval between $x+y$ and $x+y+dy$. The counterpart to equation (IV-1) is the integral (in effect a summation) over all possible displacements of the individual by migration:

$$p(x) \;=\; \int M(y)\; p^*(x+y)\; dy. \tag{IV-56}$$

If migration is weak, migrants come almost entirely from nearby locations (small values of y), so it is legitimate to approximate $p^*(x+y)$ by its Taylor series:

$$\begin{aligned} p(x) \;&\simeq\; \int M(y)\left[p^*(x) \;+\; y\,\frac{dp^*(x)}{dx} \;+\; \frac{y^2}{2}\,\frac{d^2p^*}{dx^2}\right] dy \\ &\simeq\; p^*(x)\int M(y)dy \;+\; \frac{dp^*(x)}{dx}\int yM(y)\;dy \;+\; \frac{1}{2}\frac{d^2p^*(x)}{dx}\int y^2M(y)\;dy. \end{aligned} \tag{IV-57}$$

Of the three integrals on the right-hand side of (IV-57), the first is simply the sum of all probabilities of different origins of an individual, and thus is 1. The second is $\mathbb{E}(y)$, the expectation of the displacement under migration. Note that the displacement is *not* the distance migrated, but has a sign indicating the direction of migration. We are primarily interested in the most straightforward cases, and these are cases in which migration by an amount $+y$ is just as probable as migration by an amount $-y$, so that there is no directional tendency to drift rightwards or leftwards on the average. In these symmetrical cases the average displacement $\mathbb{E}(y)$ is zero, so that the second term will vanish. The third integral is $\mathbb{E}(y^2)$, the mean square displacement. When $\mathbb{E}(y)=0$, this is also the variance of the distribution whose density function is $M(y)$. This quantity we use to measure the amount of migration, and call m. Note that in the discrete stepping stone model where migration involves $y=+1$ with probability $m/2$, $y=-1$ with probability $m/2$, and $y=0$ with probability $(1-m)$, $\mathbb{E}(y^2)$ does in fact turn out to be m.

The result of these changes in (IV-55) is

$$p'(x) \simeq p^*(x) + \frac{m}{2}\frac{d^2p^*(x)}{dx^2}, \tag{IV-58}$$

the terms ignored in the Taylor series being terms in m^2.

The quantities $p^*(x)$ are the gene frequencies before migration, after selection. Since $s(x)$ is small, we have approximately, from (II-47)

$$p^*(x) \simeq p(x) + s(x)\,p(x)[1-p(x)], \tag{IV-59}$$

ignoring terms containing $s^2(x)$. This can be substituted into (IV-58). If the product $m\,s(x)$ of two small quantities is also ignored, the result is

$$p'(x) \simeq p(x) + s(x)\,p(x)[1-p(x)] + \frac{m}{2}\frac{d^2p(x)}{dx^2}. \tag{IV-60}$$

At equilibrium, we can erase the prime on $p'(x)$ and make it $p(x)$. Then writing $p(x)$ as p and $s(x)$ as s (but keeping in mind that these are functions of x), we have the differential equation

$$\frac{m}{2}\frac{d^2p}{dx^2} + s\,p(1-p) = 0. \tag{IV-61}$$

It is on this equation that we concentrate.

An alternative derivation of the equation is to take (IV-59), substitute it into (IV-55) and ignore terms in $s_i m$ and s_i^2, getting

$$p_i' \simeq p_i + s_i\,p_i(1-p_i) + \frac{m}{2}[p_{i+1} - 2p_i + p_{i-1}]. \tag{IV-62}$$

The differential equation then arises when we approximate the second order difference in the rightmost term by d^2p/dx^2 and the difference between p_i' and p_i by dp/dx. This differential equation was first obtained by Fisher (1937) and was also given by Haldane (1948) and Fisher (1950). The latter two papers use it to try to solve for equilibrium positions of clines, but Fisher is forced to solve numerically rather than exactly. Fisher's 1950 paper is almost certainly the first application of computers to biology, as his cline computations were carried out on the Cambridge University's EDSAC, the first (or perhaps second) stored-program electronic computer operational, within a few months of its completion (see Wilkes, 1975).

APPROXIMATE SOLUTIONS OF CLINES: THEIR SHAPE. In one case, the full solution of equation (IV-61) is available. In his pioneering paper on clines, Haldane (1948) gave the solution for the case of a step pattern with symmetric selection. The solution of this differential equation is tedious, and since it is the only case of interest which can be

easily solved it is not particularly useful to go over the steps of the solution. The result is, for $x > 0$

$$p(x) \;=\; -\frac{1}{2} + \frac{3}{2}\left(\tanh\left[\left(\sqrt{s/2m}\right)x + \tanh^{-1}\left(\sqrt{2/3}\right)\right]\right)^2, \qquad \text{(IV-63)}$$

where $\tanh(y) \;=\; (e^y - e^{-y})/(e^y + e^{-y})$ and $\tanh^{-1}(\sqrt{2/3})$ turns out to be 1.1462158. When $x < 0$ there is a corresponding formula which has the same slope, but reflected so that $p(-x) = 1 - p(x)$. For details of the derivation see the step-by-step explanation by Roughgarden (1979, pp. 243-246).

In most other cases the only solutions to the basic differential equation (IV-61) are numerical. The earliest papers on equilibrium clines, those of Haldane (1948) and Fisher (1950), presented numerical solutions of different cases. Haldane found the equilibrium for a step pattern of selection and a completely dominant gene, Fisher for intermediate dominance and a smooth ramp pattern of selection. Slatkin (1973) presented numerical solutions for a number of different cases. Slatkin's solutions were not actually of the differential equation (IV-61), but of the integral equation which gives rise to it, (IV-57). Slatkin gives a number of numerical results validating the notion of the characteristic length. May, Endler, and McMurtrie (1975) have presented further numerical results and scaling arguments supporting Slatkin's generalizations.

APPROXIMATE SOLUTIONS OF CLINES: THEIR SLOPE. We can use the differential equation to solve rather simply for the slope of the cline in the case of the step pattern of selection. In that case, in the right-hand half of the cline $s(x) = s$, so that there the differential equation is

$$\frac{d^2p}{dx^2} \;=\; -\frac{2s}{m}\,p(1-p). \qquad \text{(IV-64)}$$

Let $y = dp/dx$ be the slope of the cline. Then this equation can be rewritten

$$\frac{dy}{dx} \;=\; -\frac{2s}{m}\,p(1-p). \qquad \text{(IV-65)}$$

For the moment, let us rename $2s/m$ by letting $a = 2s/m$. Now divide both sides of (IV-65) by $dp/dx = y$. We get

$$\frac{dy}{dx}\Big/\frac{dp}{dx} \;=\; -a\,p(1-p)/y, \qquad \text{(IV-66)}$$

or

$$\frac{dy}{dp} \;=\; -a\,p(1-p)/y, \qquad \text{(IV-67)}$$

so that

$$y\,dy \;=\; -a\,p(1-p)\,dp. \qquad \text{(IV-68)}$$

We have managed to obtain an equation in which the variables are separated, and we now simply integrate both sides, getting

$$\frac{y^2}{2} + C = -\frac{a\,p^2}{2} + \frac{a\,p^3}{3}, \tag{IV-69}$$

where C incorporates the undetermined constants of both integrations. We can determine C by using boundary conditions. As we move out to the right, the cline approaches $p = 1$ and is increasingly flat. This means that "at" $p = 1$ the slope $y = 0$. Requiring that this be true in (IV-69) by substituting in these values of p and y, we find that $C = -a/6$. So we can solve for y, the slope as

$$y = \left[\frac{a}{3} - ap^2 + \frac{2a}{3}p^3\right]^{1/2}. \tag{IV-70}$$

The specific case we are interested in has a symmetric step pattern of selection, so that $s(x) = s$ if $x > 0$, and $s(x) = -s$ if $x < 0$. The solution we are interested in runs through $p = 1/2$ at $x = 0$ (which is not to say that there might not be other solutions as well). Substituting in $p = 1/2$ and recalling the definition of a, we find that the slope in the center of the cline will be

$$\frac{dp}{dx} = \left(\frac{s}{3m}\right)^{1/2}. \tag{IV-71}$$

It can be verified that this is the slope of the curve in (IV-65) at $x = 0$.

The slope depends on s and m only through their ratio, but it is less obvious that it should be proportional to the square root of their ratio. This is a result different from the one-island case. There the slope, actually the difference in equilibrium gene frequencies between populations, was $1 - m/s$.

We can compare this approximation to the slope found when we iterate equations (IV-54) and (IV-55) to equilibrium and evaluate the difference between the two central populations. Table 4.1 shows the comparison between the two slopes. The results of the iteration are in reasonable agreement with our approximation. The agreement is better when s is small (and would presumably have been even better if m had been smaller). We do not expect the two numbers to be exactly the same because in the iteration the slope is measured at two points each $1/2$ unit from the center of the cline. Even if the approximate solution were to exactly interpolate the points of the discrete iteration, since the curve is sigmoid ("S-shaped") its slope would be higher in the exact center than if measured between two points each $1/2$ unit distance from the center. In the rightmost column of the table this is computed using the solution for the continuous cline with a step pattern of selection (equation IV-63 above).

The "exact" iterations in Table 4.1 are only exact for a region 16 populations in length: in a truly infinite array of stepping stones we expect the cline to have a slightly different (in fact slightly higher) slope, because of the absence of end effects.

Table 4.1: Comparison of slope of a stepping-stone model cline with $m = 0.1$, the corresponding approximation from the slope in the center of the gene frequency curve predicted from the differential equation, and a better approximation from the differential equation's predicted gene frequencies at ± 0.5

s	slope (exact)	slope (approx.)	slope(better)
0.005	0.123	0.129	0.123
0.01	0.176	0.183	0.170
0.02	0.246	0.258	0.233
0.05	0.374	0.408	0.347
0.1	0.499	0.577	0.458

THE CHARACTERISTIC LENGTH OF A CLINE. There are many variants on the cline which can be investigated by numerical solution of the differential equation (IV-61). One leads to recognition of a phenomenon of particular importance, first pointed out by Slatkin (1973). This is the existence of a *characteristic length* of a cline.

We know that for a step pattern of selection with parameters s and m, the slope of the cline at its center is $(s/3m)^{1/2}$. If the cline had this slope throughout its central region, then it would go from 0 to 1 in a distance of $\sqrt{3m/s}$. If we take the quantity $\sqrt{m/s}$, this will be slightly more than half this distance. We call this quantity ℓ_c the characteristic length of the cline. We now show that it has some meaning beyond this interpretation. Slatkin showed that a variation in fitness which is substantially shorter in extent than ℓ_c is too short for selection to respond to. He did this by placing in the middle of the cline a region with no selection, and asking how long it has to be before it has a noticeable effect on the shape of the cline.

CHARACTERISTIC LENGTH AND SWAMPING OF PATCHES. We saw in Moran's two-island model that if selection was asymmetric, the export of alleles from one island could overwhelm selection against them on the other. A similar phenomenon occurs in clines, and the notion of characteristic length plays a role, one which rather neatly fits in with its other effects. Each end of a cline of finite extent can be regarded as in a battle with the other. Selection in a given region reduces the frequency of the ill-adapted allele in that region. The export of alleles by migration reduces the frequency of those alleles in their own region of the cline, thus making the first region less susceptible to swamping by immigration. The second region is engaged in the same "activities". The stronger is selection in each region, the less likely it is to be overwhelmed by an influx of unfavorable alleles.

Another factor favoring retention of locally adapted alleles in a region is the length

Table 4.2: Numerical solution of stepping-stone cline with step pattern of selection. Details of case explained in text. The table shows the maximum gene frequency in the smaller patch as function of s, for $m = 0.1$. The characteristic length of the cline is also given.

s	p_{max}	ℓ_c
0.1	0.994	1
0.05	0.974	1.41
0.02	0.895	2.24
0.01	0.769	3.16
0.005	0.557	4.47
0.003	0.320	5.77
0.002	0.058	7.07
0.0019	0.019	7.25
0.0018	0	7.45
0.0015	0	8.16
0.001	0	10

of the region. The shorter a region is (in terms of distance from the boundary), the larger is the fraction of its gene pool which consists of new immigrants. For a given rate of migration, whether each patch persists depends on both the strength of selection in that patch and the length of the patch. As migration rates are increased, one patch of adaptation will be lost before it can swamp out the other patch.

There is then an amount of migration which does not allow local adaptation in both regions. It is particularly interesting that this seems to depend on the characteristic length of the cline. Table 4.2 shows results from numerical solutions of the stepping-stone model (IV-54) and (IV-55). The length of the species range is 16 populations, the first 4 of which have A favored with selection coefficient s, and the last 12 of which have a favored to the same extent. It can be seen from the Table that the value of s at which the smaller patch disappears corresponds to a characteristic length which is about twice the length of the patch. This is a fairly general rule. If the patch is being eroded by immigration from both sides the critical threshold is reached when ℓ_c is roughly equal to half the length of the patch.

It might be thought that this poses a problem for the species: if adaptation to a local environmental variation cannot occur if it is smaller than ℓ_c, surely this means that there would be regions which were small but in which the environment was very unfavorable to the nearby genotypes. In such cases local adaptation to the region would not occur, and the organism would never be able to adapt to the region well enough to survive

there. This problem does not occur. If selection is strong in that region, then that will reduce ℓ_c (which has s in its denominator). The result will be that if s is large enough ℓ_c will be small enough to allow local adaptation to occur.

Hanson (1966) found the approximate collapse of a pocket of adaptation, but was misled by his numerical methods into believing that in such a case a small frequency of the locally-adapted allele would exist at equilibrium. The first person to point out that the collapse of a patch of adaptation would be absolute was Nagylaki (1975). Fleming (1975) and Conley (1975) have applied various analytical mathematical techniques to prove theorems concerning the existence of stable clines in cases of habitats of finite length.

IV.12 The Wave of Advance of an Advantageous Allele

Fisher (1937) and Kolmogorov, Petrovsky, and Piskunov (1937) posed an ingenious problem which has led to much interesting mathematical work. Imagine a large region in which allele A is at an advantage over a, the advantage being the same everywhere. If we introduce a few copies of A into one area, they will increase in frequency and then descendants will also begin to spread out horizontally. After a while A will be nearly fixed in a patch, and at each end of the patch the gene frequency will be increasing and the patch will be widening. At a patch's end, there will form a "wave" of A individuals (or genes) which will propagate outwards, ultimately ending in fixation of A everywhere. These authors posed the problem of determining how fast that wave travels, and what its shape is. This problem happens to lead to fascinating mathematics, and has become a favorite exercise for mathematicians specializing in differential equations, although these subsequent investigations have added little of substance to Fisher's treatment. All we shall do here is to present the basic differential equation, and cite a few of the conclusions.

With haploid selection and selection coefficient s, an argument similar to that which leads to equation (IV-60) yields

$$p(x,t+1) \;\simeq\; p(x,t) \;+\; s\,p(x,t)\,[1-p(x,t)] \;+\; \frac{m}{2}\,\frac{\partial^2 p(x,t)}{\partial x^2} \qquad \text{(IV-72)}$$

In this case, since we are not dealing with an equilibrium cline, we have indexed p by t as well as x, in which case the second derivative must be a partial derivative. The argument implicitly assumes that s and m are small. In that case, we will not go far wrong by approximating $p(x,t+1)$ by its Taylor series expansion around $p(x,t)$,

$$p(x,t+1) \;\simeq\; p(x,t) \;+\; \frac{\partial p(x,t)}{\partial t}, \qquad \text{(IV-73)}$$

so that we get (dropping the arguments of p but keeping in mind that it is a function of x and t),

$$\frac{\partial p}{\partial t} \;=\; s\,p(1-p) + \frac{m}{2}\,\frac{\partial^2 p}{\partial x^2} \tag{IV-74}$$

which is the partial differential equation of the system. In particular, we are interested in the solution curves which propagate unchanged in shape at a constant velocity. If the velocity rightwards along the x axis is v, then moving rightwards by an amount $v\,dt$ at time t should involve as much change in gene frequency as moving back in time by an amount dt, so that we have the wave condition that the velocity of the wave, times the negative of its slope, is the rate at which the water rises

$$-v\,\frac{\partial p}{\partial x} \;=\; \frac{\partial p}{\partial t}, \tag{IV-75}$$

This common-sense condition can be substituted into (IV-72) to get

$$\frac{m}{2}\,\frac{\partial^2 p}{\partial x^2} \;+\; v\,\frac{\partial p}{\partial x} \;+\; s\,p(1-p) \;=\; 0 \tag{IV-76}$$

This equation has no explicit solution, but Fisher obtained information regarding the velocity v. It turns out that depending on the initial pattern of gene frequencies, there may be waves of different velocities. But the wave of greatest biological interest corresponds to the slowest possible velocity, which is

$$v \;=\; \sqrt{2ms}. \tag{IV-77}$$

The methods by which this is established are discussed in some detail by Moran (1962). References to further work are given by Hadeler (1976).

While the mathematics involved is no doubt challenging, some caveats are necessary. In some cases (e.g. a recessive advantageous gene) a wave does not even exist. If the environment has inhomogeneities, as in a stepping-stone model, they can severely affect the speed of the wave, so that the result (IV-75) may not accurately approximate more realistic spatial distributions. The speed of propagation of a wave depends critically on the exact shape of the leading edge of the wave, so that genetic drift may also have an effect on wave speed. Slatkin and Charlesworth (1978) give a numerical simulation in which a stepping-stone model with finite populations achieved only half the wave speed predicted by the Fisher theory.

There is thus reason for skepticism of the relevance of the theoretical result. The question is of some importance, because a wave of advance would be difficult to distinguish from a stationary cline in practice, and we would like to know how easily we may be thus misled about the type of selection present. In chapter VII we will see that genetic drift can also mimic a cline if conditions are right.

Exercises

1. With three alleles, find a set of gene frequencies in two populations which give *more* heterozygotes than expected under Hardy-Weinberg proportions for at least one of the heterozygous genotypes, if we sample from a mixture of the two populations.

2. Here are the gene frequencies of one allele at each of four loci in four (imaginary) populations:

	gene A, allele A_1	gene B, allele B_3	gene C, allele C_6	gene D, allele D_3
Sedro-Woolley	0.1137	0.4521	0.0438	0.2311
Burlington	0.1442	0.4799	0.0626	0.2216
Mount Vernon	0.1800	0.5309	0.0843	0.1850
Anacortes	0.1527	0.4276	0.0692	0.3017

   Someone suggests that the Burlington population may be entirely the result of people who came from the other three towns. What would then be the three fractions of the genes in Burlington that came from each of those three towns? *(Hint – the three fractions have to add up to 1).* Show how you figured this out.

3. When linkage disequilibrium is created by an initial admixture of two populations, each in linkage equilibrium, but with both loci having different gene frequencies in the two populations, what will be the formula for the decay of D with time?

4. In a two island model with immigration rates 0.1 and 0.2 into the islands, whose initial gene frequencies of an allele are respectively 0 and 1, what will be the equilibrium gene frequency? *(Hint – find a quantity* a *which has the property that if* p_1 *and* p_2 *are the frequencies of the allele in the two islands that the weighted average* $ap_1 + (1-a)p_2$ *stays unchanged from one generation to the next).*

5. Suppose that there are three populations with gene frequencies for allele A of 0.1, 0.2, and 0.5. In each generation a fraction 0.9 of each population comes from the population and the rest come equally from the other two. (This is true for all three populations). What is the ultimate gene frequency in the populations? How quickly is this attained (i.e. how long does it take to get halfway to the equilibrium gene frequencies?). *Hint: this looks hard but all that is necessary is that you look through the various models in chapter IV until you find one that handles this case – then figure out how to use the equations that are given for that model.*

6. Can you construct a case with two populations exchanging migrants at a constant rate, in which the gene frequency in a population oscillate but ultimately settle down to an equilibrium? A numerical example will suffice.

7. In the one-island model, what will be the "migrational load" as a function of m and s? Investigate the cases where the allele favored on the island is dominant, and where fitnesses on the island are geometric. You can assume that $m > s$.

8. For different degrees of dominance of A in a one-island model, obtain from an intuitive argument the conditions for A to increase in its island if initially present in very low frequency. How does this compare with the conditions for A to have an equilibrium gene frequency which is nonzero? What does the comparison of these two sets of conditions tell us about the patterns of dominance will we see among locally adapted alleles (as compared to the dominance of a random sample of locally favorable new mutants)?

9. For the *haploid* Levene model of section (IV.10), with two equal-sized patches having fitnesses of $A : a$ of $0.5 : 1$ and $2 : 1$, what will be the behavior of the gene frequency of A as a function of its frequency p?

10. What is the expression for the "slope" of the gene frequencies in the two-island model with haploid selection? How does it compare numerically with the slope from (IV-69) for an infinitely long cline in cases with small s and small m? What are the intuitive explanations for any discrepancy?

11. On intuitive grounds, do we expect that a cline of finite length will have a greater or a lesser slope than a cline in a habitat of infinite length? Why?

Complements/Problems

1. In a mixture of populations, express the general formula for $\mathbb{E}[P_{A_iA_j}]$ in terms of covariances of allele frequencies. In a two-allele case, can $\mathbb{E}[P_{A_iA_j}]$ ever be less than the product of the mean gene frequencies of each allele?

2. In a mixture of gametes from populations which are themselves *not* in linkage equilibrium, what is the expression for D in terms of m_i, p_i, q_i, and D_i? In terms of covariances between and within populations?

3. With two parental populations and an admixed population, with observed (sample) gene frequencies p_1, p_2, and p_{ad}, and with sample sizes n_1, n_2, and n_{ad}, what is the maximum likelihood estimate of the degree of admixture? Be sure to consider all cases.

4. Suppose that admixture occurs by male sailors from population 2 settling down in population 1. How will the effects on gene frequency differ if we compare autosomal vs. sex-linked loci?

5. In the general two-island model with immigration rates m_1 and m_2, what is the equilibrium gene frequency as a function of the initial gene frequencies $p_1(0)$ and $p_2(0)$? Now suppose we assume that You can use matrices. What is the rate of approach to equilibrium as a function of m_1 and m_2?

6. In the above case, the different values of m_1 and m_2 are due to there being an equal *number* of migrant individuals M in each direction, but different population sizes N_1 and N_2 in the islands. Express the above result in terms of N_1, N_2 and M rather than m_1 and m_2. Why does the result make good intuitive sense?

7. Two populations exchange migrants. They are of equal size and have migration rate m between them. There is a locus which has fitnesses of AA, Aa, and aa which are $1 : 1 - s : 1$. Suppose that allele A is at initial frequency 0.1 in population 1 and 0.9 in population 2. If m is considerably smaller than s, what will be the approximate gene frequencies in each population when everything settles down to equilibrium? *Hint: there is an analogy here of migration to mutation. This one can be done exactly by equations using the symmetry of the situation, or approximately using the analogy – whichever you feel you can handle.*

8. Suppose that we have a stepping-stone population structure with migration rate m and no selection. Imagine a region in which gene frequency is initially a linear function of position. Will the gene frequencies change? What is wrong with this result?

9. Why didn't we use 20 rather than 16 populations in the numerical calculations of equilibrium gene frequencies for the ramp pattern of selection? (It wasn't just the numerical difficulty - it was a cover-up. Unmask this dastardly deed.)

10. Suppose that we have two continents, one with a gene frequency of 1.0 for allele A, and the other with a gene frequency of 0. Between them stretch a chain of islands that form a perfect stepping stone model, with migration rate $m/2$ between adjacent islands, and migration rate $m/2$ into each of the terminal islands from the nearby mainland. The mainland gene frequencies are unaffected by the gene flow across the islands, because the continents are so big. There is no selection or mutation. What is the equilibrium array of gene frequencies in the islands? *(Hint – the pattern is the same no matter what the number of islands. Don't bother to derive the equilibrium from first principles, you will probably succeed if you make a good guess and then verify that it is the equilibrium).*

11. Suppose that a stream has a large resident population of rainbow trout (that remain there and do not run to the sea). They are fixed for a locally-favored allele A at a locus that would have fitnesses of AA, Aa, and aa of $1 : 0.9 : 0.81$ in that stream. A

hatchery is suddenly set up next door and as a result of straying from the hatchery in each generation trout that are all *aa* enter the stream and breed with the locals, the newly-arrived hatchery fish arriving adults and constituting 5% of all parents in each generation.

(i) What will be the ultimate fate of allele A?

(ii) Make some calculation that gives us a good sense for how rapidly this ultimate state is approached, and describe why the calculation conveys that.

12. What would be the "migrational load" in a cline if we roughly approximate the cline by saying that gene frequency is linear from 0 to 1 with slope $\sqrt{s/3m}$ in the middle of the cline and either 0 or 1 beyond there?

13. A (non-recessive) favorable allele expanding in a two-dimensional environment will form a nearly fixed patch which becomes circular in shape, and at its edge there is a wave of advance outwards that moves at the same speed as in the one-dimensional case. Can you see why? Think about the fact that our planet is spherical, but to those who live on it, its surface looks flat.

Chapter V

INBREEDING

V.1 Introduction

We now deal with the consequences of non-random mating. There are two important kinds of non-random mating: assortative mating and inbreeding. The first is preferential mating of individuals with similar phenotypes. For example, at a locus with two alleles, *aa* individuals might mate only with other *aa* individuals, and *A*- individuals (*AA* or *Aa*) might mate only with other *A*- individuals. The second type of non-random mating is inbreeding. Inbreeding is the preferential mating of relatives, where the probability of mating depends only on the degree of relationship, with the genotype or phenotype not further affecting the chance of occurrence of a particular mating, once the degree of relationship is known. In this chapter we are concerned only with inbreeding.

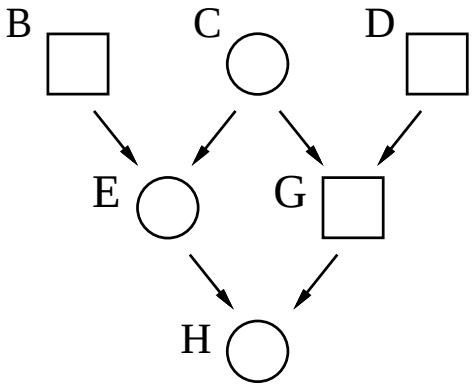


Figure 5.1: Pedigree example. Circles are females, squares males. Arrows point from parents to offspring.

Nonrandom mating seems at first to be a prohibitively difficult phenomenon to deal

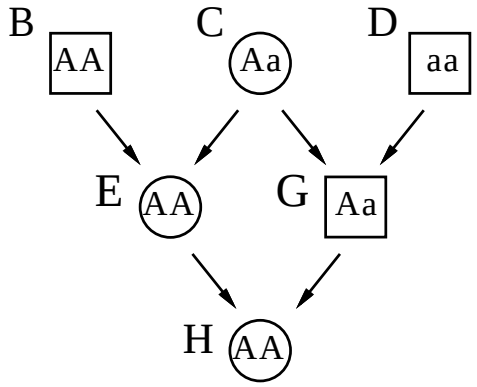


Figure 5.2: A particular assignment of genotypes compatible with the pedigree of Figure 5.1 and with the Mendelian rules. There are 155 such assignments, 45 of which have individual H as AA.

with. Consider the situation shown in Figure 5.1. Two half-sibs, E and G, have mated to produce an offspring, H. Suppose that a great many such half-sib matings occurred, and that in each mating, the three original parents B, C, and D were taken at random from a random-mating population with two alleles A and a, at gene frequencies p and $1-p$. What would be the fraction of the resulting individuals (those designated H) who were (say) AA? The straightforward approach is to consider all possibilities. Each of the three founding individuals B, C, and D could be any of the three genotypes, so that we start with $3 \times 3 \times 3 = 27$ possibilities. For each of these there are further possibilities for the genotypes of E and G. Thus, if B, C, and D were respectively AA, Aa, and aa, E could be either AA or Aa, and G could be Aa or aa. There are then further possibilities for H, so that if E is AA and G is Aa, H could be either AA or Aa.

The particular assignment of genotypes to individuals shown in Figure 5.2 has a probability of

$$p^2 \times 2p(1-p) \times (1-p)^2 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \qquad \text{(V-1)}$$

Summing this sort of probability over all possibilities in which H is AA, we could obtain the probability of this event. But it would be very tedious. There are $3^6 = 729$ possible assignments of genotypes to individuals. Of these, 155 do not violate the Mendelian rules in the assignment of genotypes to phenotypes. Of those, 45 have individual H being AA. Of course, as the individual cases are not equiprobable, the actual probability of H being AA is not 45/155. Rather, it is a sum of 45 terms, each a probability of the same sort as (V-1).

Clearly, this approach is difficult to carry out, even on a simple pedigree such as this one. Fortunately a much simpler, although subtler method is available. It was invented

and developed by Sewall Wright (1921a, b, c), although the version presented here is Malécot's (1948, 1969), who re-worked Wright's methods in terms of probabilities rather than partial regression coefficients.

V.2 Inbreeding Coefficients and Genotype Frequencies

.

The solution to the problem posed in Figure 5.1 will be easy if we can calculate the inbreeding coefficient f_H of individual H. The inbreeding coefficient of an individual is the probability that the two gene copies present at a locus in that individual are *identical by descent*, relative to an appropriate base population. Two genes are identical by descent if, and only if, they are descended from the same individual gene copy. Now of course we must stop somewhere as we trace back the ancestry of the two genes. Otherwise any two gene copies would be certain of being identical by descent, provided that life has a monophyletic origin. The function of the base population is to set the context of the problem. In the base population, all gene copies are assumed not to be identical by descent.

Once we know f for an individual, it is not hard to calculate the expected genotype frequencies. There are two cases. A fraction f of the time, the two gene copies in the individual are identical by descent. If so, then in the two-allele case both will be A if the gene from which they were copied in the base population was an A. This will be so p of the time, where p is the frequency of A in the base population. If the two copies that are identical by descent are descended from a copy that is a, then both copies will be a, which will happen $1 - p$ of the time. A fraction $1 - f$ of the time, the two gene copies in the individual will not be identical by descent. They are then descended from different copies in the base population. Under the particular assumption that the base population was formed by random mating and is in Hardy-Weinberg proportions, p^2 of the time both gene copies will be A. Putting all of this together, the expected genotype frequencies will be

$$\begin{array}{ll} AA & p^2\,(1-f) \;+\; pf \\ Aa & 2p(1-p)\,(1-f) \\ aa & (1-p)^2\,(1-f) \;+\; (1-p)f. \end{array} \qquad \text{(V-2)}$$

Note that we have not only assumed that the base population is in Hardy-Weinberg proportions, we have implicitly assumed that there is no mutation, since we assume that in the time since both gene copies originated from the same copy, there has been no further mutation. The assumption that two gene copies not identical by descent have, in effect, been drawn at random from the base population implicitly assumes that there are no differential viabilities or fertilities. These formulas are implicit in the classic and

pioneering treatment of the population genetics of inbreeding by Sewall Wright (1921a, 1921b, 1921c). They seem to have been first stated explicitly by him 12 years later (Wright, 1933).

Extension of (V-2) to multiple alleles is straightforward. If both gene copies are identical by descent, the probability that the individual is an A_iA_i homozygote is p_i. If the gene copies are not identical by descent, then the probability of any genotype is simply its Hardy-Weinberg probability. Then the genotype frequencies will be:

$$\begin{array}{ll} A_iA_i & p_i^2\,(1-f)\ +\ p_i\,f \\ A_iA_j & 2\,p_ip_j\,(1-f) \qquad\qquad (\text{where } i \neq j). \end{array} \tag{V-3}$$

In the case of the pedigree in Figure 5.1, it will turn out that $f_H = 1/8$. The base population is the population from which the individuals B, C, and D were drawn. In that population, the alleles A and a had gene frequencies p and $1-p$. So the probability that H is AA is simply

$$\frac{7}{8}\,p^2\ +\ \frac{1}{8}\,p, \tag{V-4}$$

the chance that it is Aa is $14/8\,p(1-p)$, and the probability that it is aa is $7/8\,(1-p)^2 + 1/8\,(1-p)$, which when expanded becomes

$$1\ -\ \frac{15}{8}\,p\ +\ \frac{7}{8}\,p^2. \tag{V-5}$$

Notice that the average gene frequency of A among individuals produced by the same mating scheme leading to H is simply p, since

$$\frac{7}{8}\,p^2\ +\ \frac{1}{8}\,p\ +\ \frac{1}{2}\,\frac{14}{8}\,p(1-p)\ \ =\ \ p. \tag{V-6}$$

More generally, the gene frequency of A among all individuals having inbreeding coefficient f is

$$p^2(1-f)\ +\ p\,f\ +\ \frac{1}{2}\times 2p(1-p)\,(1-f)\ \ =\ \ p\,(1-f)+p\,f\ \ =\ \ p. \tag{V-7}$$

The reader may verify that the same relationship holds for multiple alleles. Inbreeding does not affect gene frequencies, on average. But it does affect the probability of co-occurrence of two A or two a genes in the same individual.

Expressions (V-4) show how easy it is to compute genotype frequencies once we know f. If there is a simple method for computing f itself, then the inbreeding-coefficients approach will be decidedly superior to direct enumeration.

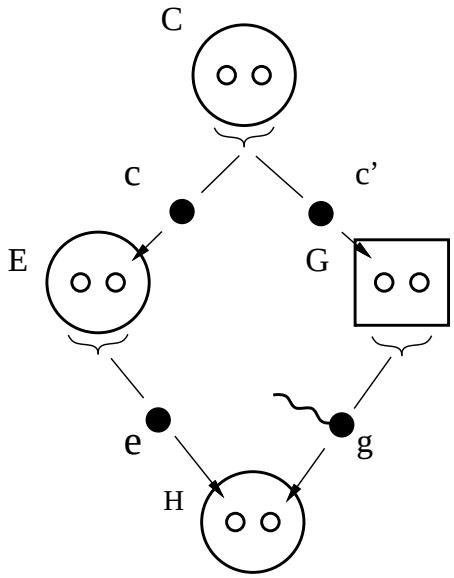


Figure 5.3: A simple example of inbreeding. The same as Figure 5.1, except that the gametes and the origin of the gene copies are shown, and extraneous individuals are deleted from the pedigree. The braces under individuals emphasize that each gamete contains at random one of the two gene copies at the locus in the parent individual.

V.3 The Loop Calculus: A Simple Example.

In the pedigree in Figure 5.1, the fact that H is partly inbred is the consequence of the fact that both the mother of H (namely E) and the father of H (namely G) share a common ancestor. Thus it is possible to trace back from H through E to the common ancestor, C, and then from C through G back to H. A common ancestor creates at least one loop in the pedigree. A method of calculating f was developed by Wright (1922) which works by finding and examining loops in the pedigree of the individual whose f must be computed. Figure 5.3 shows the same pedigree, redrawn in the form we will use it. If H has two gene copies which are identical by descent, this must reflect the fact that the copy in gamete e is descended from the copy in gamete c, that the copy in c is copied from the same gene copy in C as is the copy in c', and the copy in g is descended from the copy in G. These events have easily-computed probabilities, dependent only on the Mendelian rules of inheritance. Let us denote the event that "the gene copies c and c' are identical by descent" by $(c \equiv c')$. Suppose that we denote the event "the gene copy in e is a copy of the gene copy in c" by $(e \leftarrow c)$. We distinguish between $\leftarrow$ and $\equiv$ because one gene copy may be identical by descent to another without being a direct

copy of it. Note that we have no use for individuals B and D in this computation, so that they can be omitted from the pedigree. We want to compute

$$\begin{aligned} f_H &= \text{Prob } (e \equiv g) \\ &= \text{Prob } [(e \leftarrow c) \text{ and } (c \equiv c') \text{ and } (c' \rightarrow g)]. \end{aligned} \tag{V-8}$$

But the events $(e \leftarrow c)$, $(c \equiv c')$, and $(c' \rightarrow g)$ are the results of meioses in two different individuals, so that they are independent of one another. So we can multiply their probabilities:

$$f_H = \text{Prob } (e \leftarrow c) \times \text{Prob } (c \equiv c') \times \text{Prob } (c' \rightarrow g). \tag{V-9}$$

The probability that the gene copy in e is descended from that in c is clearly $1/2$, as that is the fraction of time that the gene in e arises from the maternally-derived gene in E. The event $(c \equiv c')$ is a bit more complex. If c and c' are both descended from the left-hand (maternally-derived) gene in C, they will be identical by descent. This event has probability $1/2 \times 1/2 = 1/4$. But c and c' could also be identical by descent through both being copies of the right-hand (paternally-derived) gene in C, an event which also has probability $1/4$. If they are descended from different copies in C, they cannot be identical by descent. So $\text{Prob}(c \equiv c') = 1/4 + 1/4 = 1/2$. The event $(c \equiv c')$ is clearly independent in its occurrence from $(e \leftarrow c)$, as whether c and c' are copies from the same gene in C has to do with the random alignment of chromosomes in two successive meioses in C, and whether $(e \leftarrow c)$ depends on the random alignment of chromosomes in a meiosis in E. The event $(c' \rightarrow g)$ is clearly of the same nature as $(e \leftarrow c)$, and has probability $1/2$, and is independent of the other two events. So

$$f_H = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = 1/8. \tag{V-10}$$

This is the value previously stated, and justifies the formulas (V-4).

The computation of f makes the same assumptions as the direct enumeration method. For the stated probabilities to be correct, there can have been no mutation. The assumption of no natural selection plays a more hidden role: if there is natural selection, an individual receiving (say) the left-hand gene copy in E may be more likely to survive to adulthood than if it received the right-hand gene copy. This makes that individual more likely to be included in the pedigree, and biases the probability $\text{Prob}(e \leftarrow c)$ away from $1/2$. Interestingly, we must not only assume that there is no natural selection acting at the locus with which we are concerned, but we must also assume that there is no natural selection acting at any locus linked to it.

We have not discussed the way in which the definition of the base population has entered into the computation of f_H. The base population is the initial population from

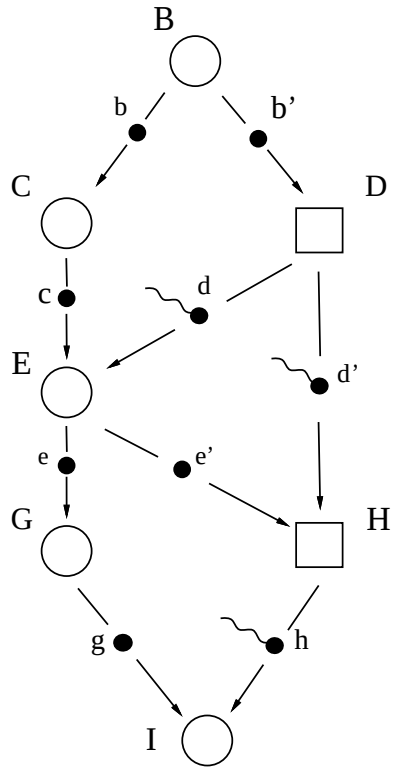


Figure 5.4: A more complex pedigree.

which B, C, and D were drawn. We assumed that if the gene in e came from B, or the gene in g came from D, or if the genes in e and g came from different genes in C, then the two genes in H could be regarded as drawn at random from the base population. This implicitly defines the base population as the one from which B, C, and D were drawn, and assumes that population was in Hardy-Weinberg proportions. If B, C, and D actually came from a different situation, for example a population of F_1 offspring from a cross of two pure lines, then the method of computation given here will not work.

V.4 The Loop Calculus: A Pedigree With Several Loops.

The pedigree in Figure 5.4 will serve as an example of the use of the loop calculus in its fullest form. Extraneous individuals have again been omitted (one parent each of G, C, and D).

Three loops lead from g to h which are relevant. The first loop is $IGECBDHI$. This corresponds to the events $(g \leftarrow e \leftarrow c \leftarrow b \equiv b' \rightarrow d' \rightarrow h)$. The probability that the

copy in g is identical by descent to the copy in h through this route is $(1/2)^6 = 1/64$. The second loop is $IGEDHI$. This corresponds to the events $(g \leftarrow e \leftarrow d \equiv d' \rightarrow h)$. The probability that $(g \equiv h)$ by this route is $(1/2)^4 = 1/16$. The problem arises as to how to combine these probabilities. The key to this problem is to notice that it is impossible for both $(e \leftarrow d)$ and $(e \leftarrow c)$ to be true at the same time. So I cannot be inbred through both of these loops simultaneously. These are mutually exclusive events, so their probabilities should be summed.

The third loop, *IGEHI*, presents a thornier problem. It corresponds to the events $(g \leftarrow e \equiv e' \rightarrow h)$. The problem arises with the event $(e \equiv e')$. If individual E, at the top of this loop, were one of the original ancestors of the pedigree, there would be no problem. Then we could have e be identical by descent to e' only if both were copies of the same gene copy in E. Otherwise they would be copied from different genes in the base population, and by definition could not be identical by descent. However, E is itself a partially inbred individual. It is at the bottom of the loop *ECBDE*. Half of the time, e and e' are copied from the same gene copy in E. The other half of the time, they represent both gene copies in E, and then their probability of being identical by descent is f_E. So

$$\text{Prob}\ (e \equiv e') \ = \ \frac{1}{2} + \frac{1}{2} f_E \ = \ \frac{1}{2}(1 + f_E). \tag{V-11}$$

Then the contribution of loop *IGEHI* to f_I is $(1/2)^3(1 + f_E)$. Now the same arguments tell us that $f_E = (1/2)^3$, so the loop *IGEHI* contributes $(1/2)^3[1 + (1/2)^3]$ to f_I. This loop represents an event which is mutually exclusive with the events of the other two loops, so we can add the three probabilities, getting

$$\begin{aligned} f_I \ &= \ \tfrac{1}{64} + \tfrac{1}{16} + \tfrac{1}{8}\left(1 + \tfrac{1}{8}\right) \\ &= \ \tfrac{1}{64} + \tfrac{4}{64} + \tfrac{8}{64} + \tfrac{1}{64} \\ &= \ \tfrac{14}{64} \ = \ 0.21875. \end{aligned} \tag{V-12}$$

The reader may have noticed some loops in Figure 5.4 which we have not counted. The loop *HEDH* is not relevant. It is useful in part of the computation of f_H, but that quantity is not needed in the computation of f_I. The events $(h \rightarrow e')$ and $(h \rightarrow d')$ have probabilities which do not depend on f_H, only on the Mendelian rules. Loops such as *IGECBDEGI* look relevant, but are not. They would be useful only to calculate the probability that the gene copy in g is identical by descent to itself! This we could obtain without necessity of following loops, and the answer would tell us nothing of interest. Loop *IHDBCEGI* looks very relevant. It is: we have already counted it. It is the same loop as *IGECBDHI*, which we already counted. Since one corresponds to one of the ways that the gene copy in h could be identical to that in g, and the other to the analogous way that the gene copy in g could be identical to that in h, these are the *same event*, and we

should not count the event twice. To avoid this duplication, we should always start from one parent (say G) and proceed through the loop to the other (H). Finally, what about loop *IGECBDEHI*? This loop passes through E twice, requiring that $(e \leftarrow c)$ and $(e' \leftarrow d)$. The only thing wrong with this loop is that we have already taken it into account when computing the contribution from loop *IGEHI*. For in that case we included a contribution of $1/4\ f_E$ which covered the event that the two gene copies in E were identical by descent (the event $(c \leftarrow b \equiv b' \rightarrow d)$ and that e came from the left-hand gene copy in E and e' from the right-hand copy. The other $1/4\ f_E$ made the loop *IGEDBCEHI* redundant.

Loops must start from one parent of the individual and end at the other, must never pass through the same individual twice, and may not change direction (up to down or down to up) more than once. If the common ancestor at the top of a loop is itself inbred, its own inbreeding coefficient must be taken into account by multiplying by $1/2\,(1+f)$ rather than $1/2$. We will not try to develop the rules of the loop calculus further: what is important is that the reader understand the logic of the procedure.

V.5 The Loop Calculus: Sex Linkage.

For computing the inbreeding coefficient at a sex-linked locus, the logic followed is the same, except for a change in the Mendelian rules. We will never, in a standard X-Y sex determination system, want to compute the inbreeding coefficient of a male for a sex-linked locus. Knowing that the lone gene copy at this locus on the X-chromosome is identical to itself is of no interest. We will therefore be confined to computing f for females.

The pedigree in Figure 5.4 will again serve as our example. The three loops originating from I are again the relevant ones to consider, plus the subsidiary loop above E. Let us consider first the loop *IGEDHI*. Keep in mind that the locus under consideration is on the X-chromosome. Clearly, the probability that $(g \leftarrow e)$ is one-half, since an X-linked gene copy in g could have been copied from the X-chromosome in e or from the X donated by the male parent of G. Likewise, the probability that $(e \leftarrow d)$ is also one-half, as d carries one of the two X-chromosomes which end up in E. But we cannot compute the probability that $(d \equiv d')$, as it cannot occur. The gamete d' carries a Y-chromosome from D to H, not a copy of the gene in question. Immediately, we can see that this loop cannot be the cause of any identity between g and h. All loops containing two males in succession must be regarded as broken, for the purposes of computing the inbreeding coefficient at a sex-linked locus. The next of our three main loops, *IGECBDHI*, has the same two males, B and D, so that it too cannot contribute anything to f_H.

We are left with *IGEHI*. As already mentioned, the probability that $(g \leftarrow e)$ is one-half. The probability that $(e \equiv e')$ depends on f_E, and is $(1/2)(1+f_E)$, as before. The chance that $(e' \rightarrow h)$ is one of those affected by sex-linkage. Since e' is the X-

chromosome going into male H, and since h carries a copy of the same X-chromosome, $\text{Prob}(e' \to h) = 1$. So

$$f_I = \frac{1}{2}\frac{1}{2}(1+f_E)(1) = \frac{1}{4}(1+f_E). \tag{V-13}$$

We now compute f_E in the same fashion. It is essentially the same kind of loop as *IGEHI*, except that B is not inbred. So $f_E = 1/4$, giving

$$\begin{aligned} f_I &= \tfrac{1}{4}\left[1+\tfrac{1}{4}\right] \\ &= \tfrac{1}{4}+\tfrac{1}{16} = \tfrac{5}{16} \\ &= 0.3125. \end{aligned} \tag{V-14}$$

Note that a sex-linked locus in I has a greater inbreeding coefficient than does an autosomal locus. This will often, but not always, be true. In general, the breakage of the male links in the pedigree will produce fewer loops relevant to a sex-linked locus, with each of these loops contributing more heavily to f because of the higher probability of transmission of X-linked genes through males. As in the case of an autosomal locus, understanding the logic of the computation is more useful than rote memorization of the rules.

V.6 The Method of Coefficients of Kinship.

A more systematic method of calculating inbreeding coefficients makes use of *coefficients of kinship*. This is a translation of Malécot's (1948) term "coefficient de parente," which is also often rendered: *coefficient of parentage* or *coefficient of consanguinity*. The coefficient of kinship of two individuals, B and C, may be roughly defined as the inbreeding coefficient of the offspring of a mating between B and C. The coefficient is defined whether or not B and C ever actually mate: it may even be the case that they are the same sex. Given that possibility, it is perhaps better to redefine the coefficient of kinship of B and C as the probability that a randomly chosen gene copy from B is identical by descent to a randomly chosen gene copy from C. Let us call this quantity F_{BC}. Clearly, if D is the offspring of a mating between B and C,

$$f_D = F_{BC}. \tag{V-15}$$

To compute inbreeding coefficients in a pedigree, we therefore need only know the coefficients of kinship of the pairs of individuals in the pedigree. It will turn out that these can be computed in a systematic fashion, using a few simple rules. Figure 5.5

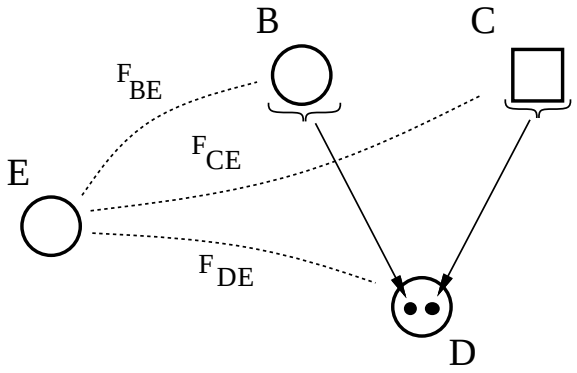


Figure 5.5: Diagram showing the logic of computing a coefficient of kinship involving an offspring from those involving its parents.

shows one of the two situations we need consider. B and C have mated and given rise to an offspring, D. Suppose that we know F_{EB} and F_{EC}, and wish to compute F_{ED}. Notice that the maternal (left-hand) gene in D is a copy of a randomly chosen gene copy from B, and the paternal (right-hand) gene is a copy of a randomly chosen gene copy from C. In choosing a gene copy from D, we therefore will get a random gene copy from B half of the time, and a random gene copy from C half of the time. Then

$$F_{ED} = \frac{1}{2}F_{EB} + \frac{1}{2}F_{EC}. \tag{V-16}$$

The other case we must consider is when E is the same individual as B. Formula (V-16), applied mechanically, would yield

$$F_{BD} = \frac{1}{2}F_{BB} + \frac{1}{2}F_{BC}. \tag{V-17}$$

But what is F_{BB}? It must be the probability that two gene copies, drawn independently from B, are identical by descent. Since these gene copies will represent the same copy half of the time, and different copies half of the time,

$$F_{BB} = \frac{1}{2}(1 + f_B). \tag{V-18}$$

f_B can be computed from the coefficient of kinship of the two parents of B, as given by a relation like (V-15).

For the initial individuals founding a pedigree, they are assumed to be non-inbred and have coefficients of kinship zero. We can then use (V-13), (V-15), and (V-12) to compute those coefficients of kinship which involve their immediate offspring, then those

which involve *their* offspring, and so on. Whenever one of the individuals involved is new to the pedigree, drawn from the base population, we assume that its coefficient of kinship with all pre-existing individuals in the pedigree is zero, and its inbreeding coefficient is zero. It will be useful to keep in mind that $F_{BC} = F_{CB}$ for all individuals B and C. Table 5.1 shows the computation of the coefficients of kinship for the pedigree in Figure 5.4. The computation is simple but repetitious, lending itself easily to programming for a computer. There are, however, some pitfalls possible. In the pedigree of Figure 5.4, if we compute F_{EC} by looking at the kinship coefficient of E with the ancestors of C (rather than of C with the ancestors of E) the result may be incorrect. To see this, try it both ways. This method was first suggested by Cruden (1949) and Emik and Terrill (1949). It is expounded in some detail by Kempthorne (1957) and by Falconer (1989). It is efficient enough for medium-size pedigrees.

For very large pedigrees (tens of thousands of individuals) the method will generate too many coefficients of kinship to be practical. In such cases, loop-finding programs such as that of Mange (1964) will be more efficient if the inbreeding coefficients of only a few individuals are desired. Alternatively, one may use random sampling techniques such as those of Wright and McPhee (1925) or Edwards (1968) for a useful rough estimate of f.

The method of coefficients of kinship is easily extended to sex-linkage.

V.7 The Complication of Linkage.

So far, we have considered only a single locus. Suppose that we wanted to know the probabilities of genotypes at two loci in an inbred individual. We can develop expressions for the genotype frequencies in terms of the probabilities that the two loci are or are not inbred. If F_{00} is the probability that neither locus is identical by descent, if F_{10} is the probability that locus A is identical by descent but B is not, if F_{01} is the probability that A is not identical by descent but B is, and if F_{11} is the probability that A and B are both identical by descent, then the genotype probabilities can be written straightforwardly in terms of these quantities. If the base population is at linkage equilibrium with frequencies p of A and q of B, then

$$\text{Prob}\,(AABB) \;=\; p^2q^2\,F_{00} \;+\; pq^2\,F_{10} \;+\; p^2q\,F_{01} \;+\; pq\,F_{11}. \qquad \text{(V-19)}$$

The logic of (V-19) is the same as for (V-2). I will not explain it in detail here. Similar expressions can be worked out for the other genotypes. The four coefficients F_{00}, F_{01}, F_{10}, and F_{11} actually require only one new quantity to be computed. Note that $F_{10} + F_{11} = f_A$, the probability that locus A is identical by descent irrespective of the status of locus B. Likewise $F_{01} + F_{11} = f_B$, and if both A and B are autosomal (or both sex-linked),

Table 5.1: Sequence of computations of the coefficients of kinship of the individuals in the pedigree in Figure 5.4. Whenever used, O denotes an individual outside the pedigree.

Generation 1	F_{BB}			$=$	$\frac{1}{2}(1+F_{OO})$	$=$ 0.5
Generation 2	F_{BC}	$=$	F_{CB}	$=$	$\frac{1}{2}F_{BB}+\frac{1}{2}F_{BO}$	$=$ 0.25
	F_{CC}			$=$	$\frac{1}{2}(1+F_{BO})$	$=$ 0.5
	F_{BD}	$=$	F_{DB}	$=$	$\frac{1}{2}F_{BB}+\frac{1}{2}F_{BO}$	$=$ 0.25
	F_{CD}	$=$	F_{DC}	$=$	$\frac{1}{2}F_{BC}+\frac{1}{2}F_{OC}$	$=$ 0.125
	F_{DD}			$=$	$\frac{1}{2}(1+F_{BO})$	$=$ 0.5
Generation 3	F_{EB}	$=$	F_{BE}	$=$	$\frac{1}{2}F_{CB}+\frac{1}{2}F_{DB}$	$=$ 0.25
	F_{EC}	$=$	F_{CE}	$=$	$\frac{1}{2}F_{CC}+\frac{1}{2}F_{DC}$	$=$ 0.3125
	F_{ED}	$=$	F_{DE}	$=$	$\frac{1}{2}F_{CD}+\frac{1}{2}F_{DD}$	$=$ 0.3125
	F_{EE}			$=$	$\frac{1}{2}(1+F_{CD})$	$=$ 0.5625
Generation 4	F_{GB}	$=$	F_{BG}	$=$	$\frac{1}{2}F_{EB}+\frac{1}{2}F_{OB}$	$=$ 0.125
	F_{GC}	$=$	F_{CG}	$=$	$\frac{1}{2}F_{EC}+\frac{1}{2}F_{OC}$	$=$ 0.15625
	F_{GD}	$=$	F_{DG}	$=$	$\frac{1}{2}F_{ED}+\frac{1}{2}F_{OD}$	$=$ 0.15625
	F_{GE}	$=$	F_{EG}	$=$	$\frac{1}{2}F_{EE}+\frac{1}{2}F_{OE}$	$=$ 0.28125
	F_{GG}			$=$	$\frac{1}{2}(1+F_{EO})$	$=$ 0.5
	F_{HB}	$=$	F_{BH}	$=$	$\frac{1}{2}F_{EB}+\frac{1}{2}F_{DB}$	$=$ 0.25
	F_{HC}	$=$	F_{CH}	$=$	$\frac{1}{2}F_{EC}+\frac{1}{2}F_{DC}$	$=$ 0.21875
	F_{HD}	$=$	F_{DH}	$=$	$\frac{1}{2}F_{ED}+\frac{1}{2}F_{DD}$	$=$ 0.40625
	F_{HE}	$=$	F_{EH}	$=$	$\frac{1}{2}F_{EE}+\frac{1}{2}F_{DE}$	$=$ 0.4375
	F_{HG}	$=$	F_{GH}	$=$	$\frac{1}{2}F_{EH}+\frac{1}{2}F_{OH}$	$=$ 0.21875

Then

$$f_I = F_{GH} = 0.21875$$
$$f_H = F_{DE} = 0.3125$$
$$f_E = F_{CD} = 0.125$$

$f_A = f_B = f$. Furthermore, all four of the F's can be written in terms of f and F_{11}:

$$\begin{aligned} F_{10} &= f - F_{11} \\ F_{01} &= f - F_{11} \\ F_{00} &= 1 - F_{11} - F_{10} - F_{01} \;=\; 1 - 2f + F_{11}. \end{aligned} \tag{V-20}$$

Once we compute f for an individual, by standard methods, we need only know F_{11} to be able to obtain genotype probabilities jointly at two loci.

There are two simple cases. If the two loci A and B are so tightly linked that no recombination ever occurs between them, then whenever one locus is identical by descent, so is the other. In that case, $F_{11} = f$, and $F_{01} = F_{10} = 0$ with $F_{00} = 1 - f$. Of course, in such a case the base population is unlikely to be at linkage equilibrium, and we must revise equation (V-19) to take gamete frequencies in the base population into account. On the other hand, if the two loci are unlinked, so that $r = 1/2$, then it turns out that $F_{11} = f^2$. The events of identity by descent at the two loci are independent, once the pedigree is specified. This is not a fact which will be immediately apparent to the reader of these notes. If you consider the various logical steps involved in inquiring whether locus A *and* locus B are identical by descent, you will soon convince yourself that in all cases, the ancestor from which a copy at locus A is descended is determined independently of the ancestor from which a copy at locus B is descended. I leave this to the reader as an exercise.

Unhappily, there is no simple method for computing F_{11} by a loop-calculus approach when the recombination fraction r is neither 0 nor $1/2$. The complication comes from a fact which was implicit in the discussion of the unlinked case: it is possible for A to be identical by descent through one loop and for B to be simultaneously identical by descent through another. To use the loop approach one has to enumerate over all pairs of loops above the given individual, and also compute as well the probabilities of joint identity of descent through the same loop. The steps involved are tedious but straightforward, and are practical in small pedigrees. Denniston (1975) shows how to do this in the case where no other individual in the pedigree, above the particular individual we are interested in, is inbred. It is also possible to develop methods based on coefficients of kinship. However, the number of quantities which we must keep track of increases greatly. We must compute probabilities such as $F_{IJ;KL}$, the probability that locus A is identical by descent in random gametes from I and J *and* that locus B is identical by descent in gametes from K and L, plus a number of other sorts of quantities. Using this approach with even a moderate-sized pedigree will require much computation.

V.8 More Elaborate Probabilities of Identity.

The standard coefficient of kinship tells us the probability that two genes drawn at random, one each from individuals I and J, will be identical by descent. But it does not answer more complex questions, such as the probability that the maternally-derived copies in I and J are identical by descent and *at the same time* the two paternally-derived copies are identical by descent. Such quantities are of more than academic interest: they enable us to compute quantities such as the probability that I is of genotype *aa* given that J is *aa*. This is of importance in genetic counselling. We may want to know the probability that a relative of an affected individual will be affected.

To compute the joint genotype probabilities of pairs of individuals, we can use the set of 15 coefficients developed by Gillois (1964, 1965) and expounded by Jacquard (1974). Other expositions of similar approaches include those of Cockerham (1971) and Denniston (1974). In simple cases one may use the matrix-computation methods of Li and Sacks (1954), which were independently derived earlier by Geppert and Koller (1938). Thompson (1974) has developed a general algebraic approach to the calculation of simultaneous genotype probabilities of a number of individuals in a pedigree, suitable for computation in small pedigrees. See Karigl (1982) for further generalizations of these methods.

V.9 Regular Systems of Inbreeding: Selfing.

When the same pattern of mating is repeated in each generation of a pedigree, we can take advantage of this regularity in computing the inbreeding coefficient. In the next few sections, we will see how this is done. Although regular systems of inbreeding are fun and of some importance, our underlying objective will be to explain the machinery that we will then use to analyze inbreeding in random-mating finite populations.

The simplest possible regular system of inbreeding is repeated self-fertilization, shown in Figure 5.6. In each generation, the single individual in the line self-fertilizes to produce the single individual of the next generation. Suppose that we want to know the inbreeding coefficient of the individual in generation t in a self-fertilizing line, where generation 0 is drawn from the base population. Let this inbreeding coefficient be f_t. Since f_t is also the coefficient of kinship of the two gene copies in the gametes of generation $t-1$,

$$f_t \;=\; \frac{1}{2}(1+f_{t-1}). \tag{V-21}$$

This is simple enough, and it can readily be used to find a formula for f_t in terms of t and f_0. But an even simpler approach is to follow $h_t = 1 - f_t$. This is the probability that the two gene copies in the individual of generation t are *not* identical by descent. A direct argument which gives us a recurrence relation for h_t is as follows. Half of the time, the

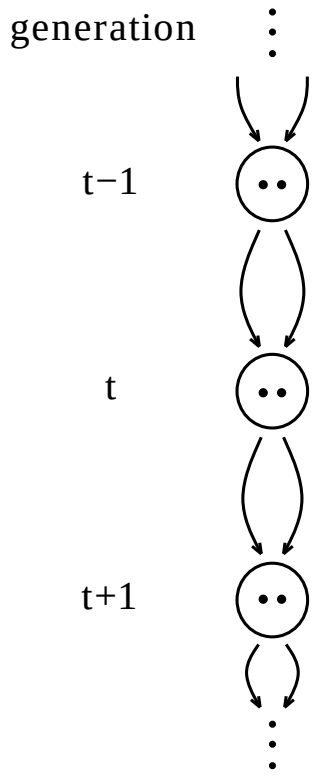


Figure 5.6: The system of repeated self-fertilization.

two gene copies are descended from the same gene in the individual of generation $t-1$. If so, they cannot be non-identical. Half of the time, they are descended from different copies in generation $t-1$, in which case the probability that they are not identical by descent is simply h_{t-1}. So

$$h_t \;=\; \frac{1}{2}\, h_{t-1}. \tag{V-22}$$

We immediately see that since h is being multiplied by $1/2$ every generation,

$$h_t \;=\; \left(\frac{1}{2}\right)^t h_0 \;=\; \left(\frac{1}{2}\right)^t \tag{V-23}$$

Clearly

$$f_t \;=\; 1 - h_t \;=\; 1 - \left(\frac{1}{2}\right)^t. \tag{V-24}$$

As h declines to 0, f rises to 1.

In this particular example, we can also analyze the results by enumerating genotypes. Simple consideration of the Mendelian rules will show that if the individual of genera-

Table 5.2: Proportion of self-fertilizing lines having various genotypes, when the base population is in Hardy-Weinberg proportions with two alleles whose frequencies are 0.6 and 0.4.

	Proportion of lines which are:		
	AA	*Aa*	*aa*
generation			
0	0.36	0.48	0.16
1	0.48	0.24	0.28
2	0.54	0.12	0.34
3	0.57	0.06	0.37
4	0.585	0.03	0.385
5	0.5925	0.015	0.3925
6	0.59625	0.0075	0.39625
7	0.598125	0.00375	0.398125
⋮	⋮	⋮	⋮
∞	0.60000	0	0.4000

tion t is homozygous, so must be the individuals in all subsequent generations. But if the individual in generation t is a heterozygote *Aa*, the individual of the next generation will be *AA* one-quarter of the time, *aa* one-quarter of the time, and *Aa* half of the time. Table 5.2 shows the results we expect if we set up a large number of self-fertilizing lines from a random-mating population initially at Hardy-Weinberg proportions. Every generation, half of the heterozygous lines are converted into homozygotes. Note that the overall gene frequency of A is not changed by inbreeding, although the proportion of A within any one line tends to zero or one. The results fit equation (V-2), as they must, the proportion of heterozygotes in generation t being $2p(1-p)(1-f_t) = 2p(1-p)h_t$.

Thus the relative proportion of heterozygotes in generation 7, compared to the initial generation, is

$$\frac{0.00375}{0.48} \;=\; 0.0078125 \;=\; (0.5)^7. \qquad \text{(V-25)}$$

So h_t may be regarded either as the probability of non-identity-by-descent, or as the fraction of initial heterozygosity still remaining among replicate inbreeding lines.

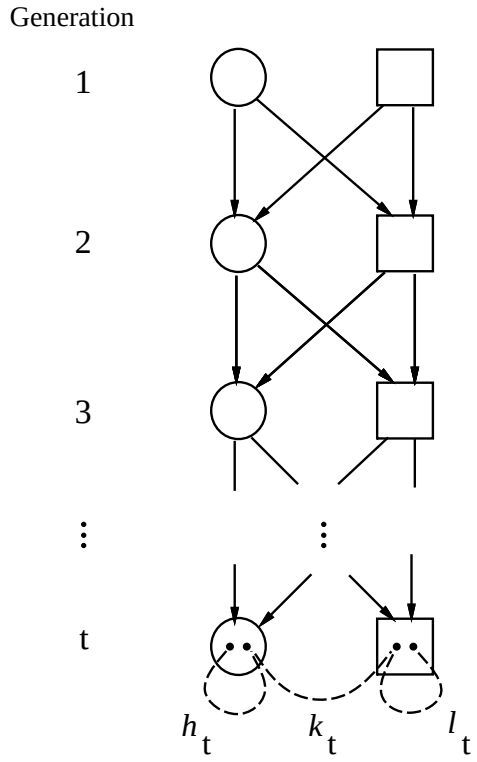


Figure 5.7: The system of repeated full-sib mating, and three probabilities of non-identity used to analyze it.

V.10 Regular Systems of Inbreeding: Full Sib Mating

.

Full-sib mating, which is a slightly more complicated system, will serve to illustrate most of the details of analysis of a regular system of inbreeding. Figure 5.7 shows a diagram of full-sib mating, in which in each generation a pair of full sibs is kept as the source of the next generation. As there are two individuals per generation, there are a total of 3 coefficients of kinship possible within a generation. While the analysis could be carried through in terms of coefficients of kinship, it will prove easier to work with three different quantities. These are illustrated in the lower part of Figure 5.7. We work with probabilities of non-identity-by-descent of pairs of gene copies. The quantity h_t is the probability of non-identity of the two gene copies in a female in generation t. ℓ_t is the corresponding probability of non-identity of the two gene copies in a male. k_t is the probability of non-identity of two genes, one drawn from the female and one from the male.

The first simplification in the analysis is to note that, provided that both the females and males founding each line are drawn from the same base population (which we assume), the symmetry of the situation demands that $h_t = \ell_t$ throughout the process. Hence, from this point on, h_t will be redefined as the probability of non-identity of two genes from the same individual, irrespective of its sex. We are now down to two quantities, h_t and k_t. Because the two gene copies in the same individual definitely came from different parents, and are copies of random gene copies in those parents, clearly

$$h_t \;=\; k_{t-1}. \tag{V-26}$$

However, the two gene copies chosen at random from different parents can be of four different origins. Both may be the maternally-derived copies, or both the paternally-derived copies, or the first maternal and the second paternal, or the first paternal and the second maternal. Each of these possibilities has probability $1/4$. In the first two cases, both came in the preceding generation from the same individual, an event with total probability $1/2$. But only in half of those cases would the two gene copies from the same individual be copies of both gene copies in that individual, so that $(1/4)h_{t-1}$ of the time the two copies in different individuals are non-identical owing to descent from non-identical copies in the same parent. One-fourth of the time they represent copies of the same gene from the same parent, and cannot be non-identical. Half of the time, they came from different parents (and were drawn randomly from those parents). Hence they will be non-identical through descent from non-identical copies in different parents $(1/2)\,k_{t-1}$ of the time. Putting all of these possibilities together, we get

$$k_t \;=\; \frac{1}{4}\,h_{t-1} + \frac{1}{2}\,k_{t-1}. \tag{V-27}$$

The initial values, h_0 and k_0, are clearly 1 if the initial male and female were drawn at random from the base population. We can start with those values and use (V-26) and (V-27) to compute h_t and k_t successively for as many generations as we desire. Figure 5.8 shows h_t and k_t plotted both directly and logarithmically against time (which we measure in generations). It will be apparent from the logarithmic plot that after a few generations, h and k decline geometrically (losing a constant percentage of their current value each generation).

While we may be satisfied to know some arithmetic values of h_t and k_t, we may also be interested in the ultimate rate of decay of h and k, which will be the rate at which f_t approaches 1. There is a short-cut method for finding the rate of decay, which we now briefly examine. We first note that, from (V-26), $h_{t-1} = k_{t-2}$. Substituting k_{t-2} for h_{t-1} in (V-27), we find that

$$k_t \;=\; \frac{1}{2}\,k_{t-1} + \frac{1}{4}\,k_{t-2}, \tag{V-28}$$

so that we can calculate each value of k from the two previous values. Now suppose that k has reached the part of its curve in Figure 5.8 where it is declining geometrically. Then

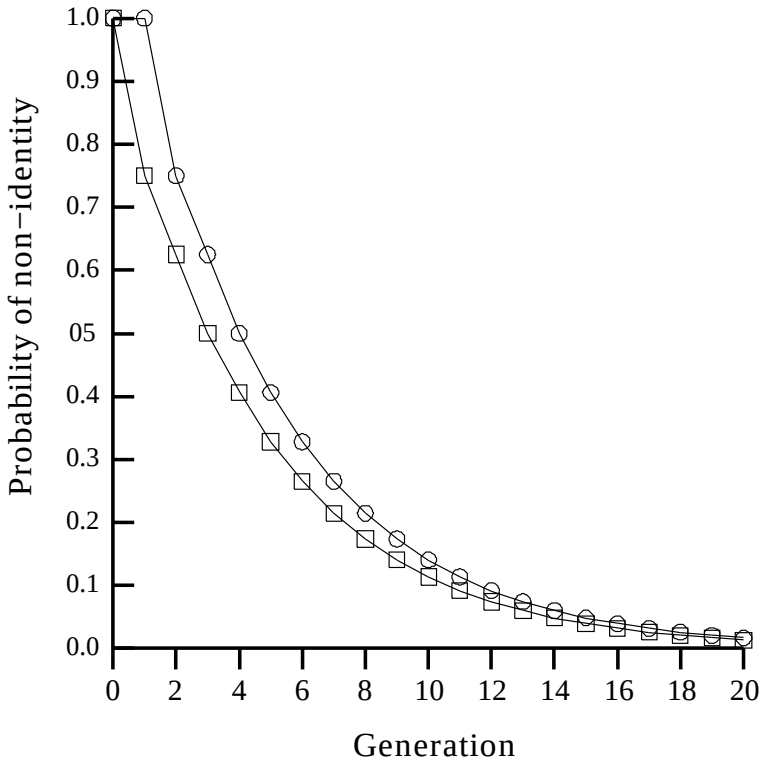


Figure 5.8: The decline of the probability of non-identity under repeated full-sib mating. Circles give the probabilities of non-identity of genes in the same individual, squares the corresponding probability for genes drawn from different individuals.

each k is a constant fraction λ of the previous k. So

$$\begin{aligned} k_t &= \lambda\, k_{t-1} = \lambda^2\, k_{t-2} \\ \text{and} & \\ k_{t-1} &= \lambda\, k_{t-2} \quad . \end{aligned} \tag{V-29}$$

Substituting these into (V-28), we get

$$\lambda^2\, k_{t-2} = \frac{1}{2}\,\lambda\, k_{t-2} + \frac{1}{4}\, k_{t-2}. \tag{V-30}$$

Each term of (V-30) contains k_{t-2}. Since it is never zero, it can be divided out of each term, giving, after moving all terms to the same side of the equation, the following quadratic equation in λ:

$$\lambda^2 - \frac{1}{2}\lambda - \frac{1}{4} = 0. \tag{V-31}$$

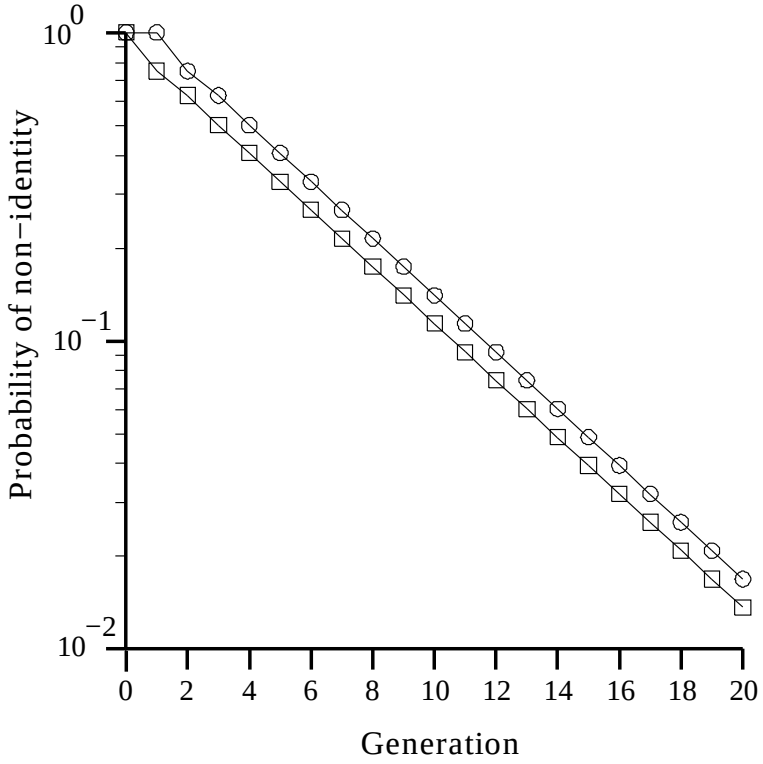


Figure 5.9: Same as the previous figure, except that the vertical axis is scaled logarithmically. Note how quickly the rate of decay of non-identity becomes constant.

The solutions of (V-31) are $(1 \pm \sqrt{5})/4$. Since the λ we are looking for must be positive, the relevant solution is the larger one:

$$\lambda \;=\; \frac{1+\sqrt{5}}{4} \;=\; 0.80901699. \qquad \text{(V-32)}$$

Thus full-sib lines asymptotically lose nearly 20% of heterozygosity each generation. Although we have computed the rate of decline of k, h must decline at the same rate, by equation (V-26). The h in each generation is simply the k in the previous one.

HISTORY. This technique, of eliminating all but one quantity from the equations (V-26) and (V-27), making two one-generation recurrence equations into one two-generation equation (V-28), is of general applicability. It was used (in effect) by Sewall Wright in his 1921 papers, where he analyzed not only full-sib mating, but nine other simple mating systems. The full-sib mating case was first analyzed by Pearl (1913), incorrectly. His

results were corrected by Jennings (1914) and Fish (1914), and retracted by Pearl himself (1914). However, all of these approaches made use of direct genotype enumeration. Jennings (1916) was able to extend this analysis to the case of a parent-offspring mating system. It was only with the elaboration of the inbreeding coefficient by Wright (1921a, b, c) that such computations could be carried out easily for a wide variety of mating systems.

V.11 Regular Systems of Inbreeding: Matrix Methods

.

The equations (V-26) and (V-27) are simultaneous linear homogeneous recurrence equations. The present method for obtaining λ has neither been presented in general form, nor has it found all relevant information. We now examine a technique using the algebra of matrices, which allows complete and systematic solution of such recurrence equations. The reader who is unfamiliar with matrix algebra may wish to familiarize themselves with it before reading this section. Alternatively, no harm will come from skipping this section. Simple numerical computation with equations (V-26) and (V-27) over a number of generations will usually be sufficient to illustrate the properties of full-sib mating, without recourse to more elaborate approaches.

The equations (V-26) and (V-27) can be rewritten in matrix form:

$$\begin{bmatrix} h_t \\ \\ k_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} h_{t-1} \\ \\ k_{t-1} \end{bmatrix} \tag{V-33}$$

which is a matrix equation of the form

$$\mathbf{x}_t \;=\; \mathbf{A}\,\mathbf{x}_{t-1}. \tag{V-34}$$

It follows that

$$\mathbf{x}_t \;=\; \mathbf{A}^t\,\mathbf{x}_0. \tag{V-35}$$

The matrix A has a *characteristic equation* which is obtained by subtracting λ from all of the diagonal elements of $\mathbf{A}$, then taking the determinant of the resulting matrix. Since λ is unknown, the result is an expression in λ, in fact, a polynomial. Computing this polynomial and equating it to zero, we obtain the characteristic equation:

$$(0-\lambda)\left(\frac{1}{2}-\lambda\right)-\frac{1}{4}(1) \;=\; 0. \tag{V-36}$$

This is simply

$$\lambda^2 \; - \; \frac{1}{2}\lambda \; - \; \frac{1}{4} \; = \; 0, \tag{V-37}$$

which is precisely (V-31)! In most cases resulting from inbreeding systems, there are as many distinct roots of the characteristic equation as there are rows (or columns) of **A**. These roots are known as *characteristic values* or *eigenvalues*. Let us assume that in the general case there are n rows (columns) of **A**. If there are not n different roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the characteristic equation, this is usually because we have failed to take advantage of some symmetry in the pedigree, and the net result is that one or more of the quantities in the vector $\mathbf{x}_t$ are superfluous. Suppose that we have pared down the quantities in $\mathbf{x}_t$ to the minimum possible, and have found the n roots of the characteristic equation, which we take to be distinct roots, no two of which are equal.

The standard spectral theory of matrices then tells us that we can find n vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, which satisfy the equations

$$\mathbf{A}\mathbf{y}_i \; = \; \lambda_i \, \mathbf{y}_i, \qquad i \; = \; 1, 2, \ldots, n. \tag{V-38}$$

When we have found these vectors, we can write the initial vector $\mathbf{x}_0$ as a linear combination of these vectors. This means that we can find n constants $c_1, c_2, \ldots, c_n$ such that

$$\mathbf{x}_0 \; = \; c_1 \, \mathbf{y}_1 \; + \; c_2 \, \mathbf{y}_2 \; + \; \ldots \; + \; c_n \, \mathbf{y}_n. \tag{V-39}$$

A property of the vectors $\mathbf{y}_i$ (which are called the *eigenvectors* or *characteristic vectors* of **A**), is that in general we can write

$$\mathbf{x}_t \; = \; c_1 \lambda_1^t \, \mathbf{y}_1 + \cdots + c_n \lambda_n^t \, \mathbf{y}_n. \tag{V-40}$$

This gives us a way to compute any $\mathbf{x}_t$ once we have computed the eigenvalues and eigenvectors of **A**, and the constants $c_1, \ldots, c_n$. In particular, as t increases, the vector $\mathbf{x}_t$ is more and more closely approximated by the term

$$\mathbf{x}_t \; \simeq \; c_1 \, \lambda_1^t \, \mathbf{x_1}, \tag{V-41}$$

where λ_1 is the eigenvalue of **A** which has the largest absolute value. In this case, the elements of **A** are never negative, and each row of **A** never adds up to more than 1, as the elements of the row are probabilities of mutually exclusive events. Two mathematical theorems (Gershgorin's theorem and Frobenius' theorem) assure us that $0 < \lambda_1 < 1$ and that, since all other eigenvalues are closer to zero than λ_1, the convergence in (V-40) will actually take place if $c_1 > 0$. λ_1 is worth notice. It is the asymptotic fraction of heterozygosity which is retained each generation.

There is no point in belaboring this approach further. The reader interested in actual computations will find computer programs available to obtain the eigenvalues and eigenvectors of matrices such as A. The point worth making here is simply that the

manipulations of the recurrence equations in the previous section were not arbitrary or based on trickery: they were a way of obtaining the characteristic equation and its largest root. We can find this root in any case of interest, by finding A and getting its characteristic equation.

OTHER MATRIX APPROACHES. An alternative approach to regular systems of inbreeding which also uses matrix algebra is the one developed by Bartlett and Haldane (1934) for an autotetraploid case. It was stated generally by Fisher (1949 and subsequent editions). It involves setting up a matrix with one column for each of the genotype compositions of a given generation. Thus, in the case of repeated full-sib mating, there are 9 possible compositions in generation t if there are two alleles at the locus. The k-th column of this matrix contains the probabilities that if we are now in the k-th genotype composition, in the next generation we shall be in each of the possible genotype compositions. Fisher shows how to set up these matrices, and to simplify them somewhat. It will always turn out that the largest eigenvalue of this matrix is the same as the value λ_1 obtained by the approach of this section (see, for example, the comments of Wright, 1969, pp. 171-173). In theory these matrices give an even more complete analysis of the inbreeding system than does the smaller matrix **A**. In practice, however, they are quite difficult to work with. They are very large matrices in all but the simplest cases.

V.12 Repeated double first cousin mating.

Figure 5.10 shows the system of repeated double-first-cousins mating, which we take as the third (and final) example of a repeated mating system. The method of analysis is analogous to the case of full-sib mating. We need only three quantities: h_t, k_t, and ℓ_t. If we draw two gene copies from the pedigree,

they can come from three sources. If both come from the same individual, their probability of non-identity is h_t. If they come from individuals who are full sibs, their probability of non-identity is k_t. If they come from different individuals who are not sibs, their probability of non-identity is ℓ_t. By carefully studying the pedigree, you will discover that all possible pairs of individuals from the same generation fall into one of these categories. Since the mating system avoids mating of sibs, the two gene copies in one individual come from different non-sibs in the previous generation. The genes in sibs come half of the time from different parents, who are nonsibs. Half of the time, they come from the same individual, in which case they come from different gene copies half of the time. Finally, the two genes in nonsibs cannot come from the same individual. They will come from sibs half of the time and from nonsibs half of the time. Thus we

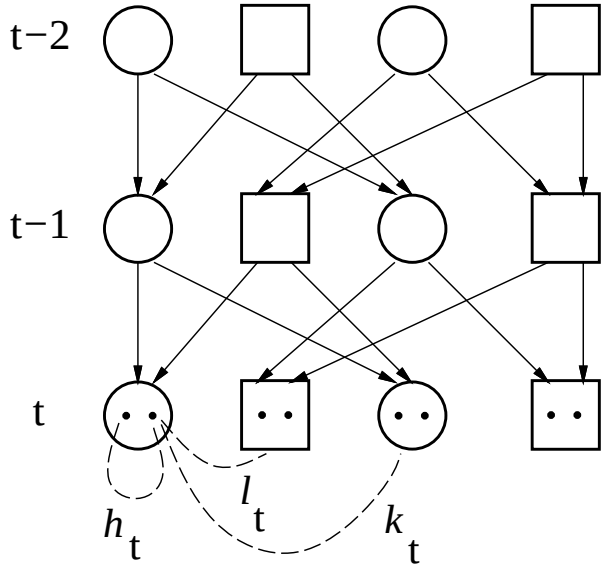


Figure 5.10: The system of repeated double-first-cousins mating, and three probabilities of nonidentity used to analyze it.

obtain:

$$\begin{aligned} h_{t+1} &= \qquad\qquad\quad \ell_t \\ k_{t+1} &= \tfrac{1}{4}h_t \qquad +\tfrac{1}{2}\,\ell_t \\ \ell_{t+1} &= \qquad \tfrac{1}{2}k_t \quad +\tfrac{1}{2}\,\ell_t \end{aligned} \tag{V-42}$$

We can use these equations to compute these quantities in successive generations. The initial values of h, k, and ℓ are 1. After an initial period of a varying rate of decline, all three quantities decline geometrically, with a constant factor λ by which each is multiplied in each generation. We can find λ by eliminating all but one of the three quantities h, k, and ℓ from equations (V-42). Then we use equations (V-29) or their equivalent to obtain an equation in λ which can be solved.

Alternatively, we can set up the matrix described in the previous section, and subtract λ from its diagonal elements. By equating the determinant of that matrix to zero, we obtain the equation for λ:

$$\begin{vmatrix} 0-\lambda & 0 & 1 \\ \frac{1}{4} & 0-\lambda & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2}-\lambda \end{vmatrix} = 0, \tag{V-43}$$

or

$$\lambda^3 - \frac{1}{2}\lambda^2 - \frac{1}{4}\lambda - \frac{1}{8} = 0. \qquad \text{(V-44)}$$

The largest root of (V-44) is $\lambda = 0.919643$.

We can set up systems of size 8, 16, 32, etc., which are analogous to double-first cousin mating in that they mate those individuals at each generation who are least closely related. These "maximum avoidance of inbreeding" systems were analyzed by Wright (1933). All have characteristic equations analogous to (V-31) and (V-44). For example, the equation for octuple third cousin mating is:

$$\lambda^5 - \frac{1}{2}\lambda^4 - \frac{1}{4}\lambda^3 - \frac{1}{8}\lambda^2 - \frac{1}{16}\lambda - \frac{1}{32} = 0 \qquad \text{(V-45)}$$

V.13 Avoiding Inbreeding.

Interestingly enough, the systems of "maximum avoidance of inbreeding" do not have the lowest ultimate rates of approach to homozygosity. With the same number of individuals, it is possible to devise circular half-sib mating systems which, although they inbreed faster initially, have a lower rate of approach to total homozygosity. This was shown by Kimura and Crow (1964), who treated such systems in generality. Robertson (1964) provided a more general framework for this result. He showed that, in general, regular systems of mating with a given number of individuals will have a *lower* rate of approach to homozygosity the more closely related are the individuals who mate!

Of course, the best system of all for avoiding the loss of alleles from a population of fixed size involves the most intense inbreeding of all. If we divide a population of size 20 into 10 full-sib lines, and keep those lines isolated from each other, we stand a good chance of retaining a reasonable fraction of the common alleles present initially. For even though each such line reaches fixation for an allele, different lines may well fix for different alleles. So although each individual becomes homozygous, we can restore a good fraction of the initial heterozygosity of the base population by crossing different lines. By contrast, a repeated mating system which does not break the population into isolated lineages is certain to fix for one allele or another sooner or later, however small its rate of approach to homozygosity.

In spite of the need to conserve the dwindling supply of genetic variability in domestic animals and plants, the system of isolated inbred lines is rarely used. This is because of the disastrous effects of inbreeding on viability and fertility. We now examine these effects briefly.

V.14 The Effects of Inbreeding.

The simplest effects of inbreeding are on the frequencies of phenotypes of recessive alleles. From (V-2), if A is a recessive allele (the size of the letter notwithstanding) the frequency of homozygotes is

$$p^2(1-f) + p\,f, \tag{V-46}$$

which can be rewritten

$$p^2 + f\,p(1-p), \tag{V-47}$$

which shows that the frequency of homozygotes increases with f. How much it increases is a function of the gene frequency p. When $p = 0.1$, the homozygote frequency will rise from 0.01 to 0.1 as f is increased from 0 to 1. This is an increase by a factor of ten. But when $p = 0.001$, the increase will be from 0.000001 to 0.001, a factor of 1000. If a population has a great many rare recessive mutants, inbreeding will make the frequencies of their phenotypes far greater. It is a quite general observation that severely deleterious rare mutants tend to be recessive. There are straightforward functional reasons for this. For loci which code for enzymes, a severely deleterious mutant tends to be one which renders the enzyme inactive. When only half of the enzyme molecules (or polypeptide chains) are inactive, there is still usually enough active enzyme around to give a normal or nearly-normal phenotype. So severely deleterious mutants tend to be recessive more often than dominant.

This is the basis of the effect of inbreeding in reducing fitness. There is much data on the effects of inbreeding, but one set of figures will suffice. The frequency of congenital genetic diseases is about 1% at birth. But if the parents are first cousins ($f = 0.0625$), this figure rises to 2%, implying that we would see a great increase in a mating which produced a completely inbred offspring.

In the case of rare recessive alleles, we have seen that inbreeding produces a great increase in the frequency of homozygotes for that allele. This implies that most of the homozygotes result from identity by descent rather than from randomly-occurring homozygosity. One example should suffice. If a rare allele has $p = 0.001$, and if one mating in a thousand is a first cousin marriage (for which $f = 1/16$ or 0.0625), then the average f is (0.001)(0.0625) = 0.0000625. The frequency of homozygotes is (from (V-47))

$$(0.000001)(0.9999375) + (0.0000625)(0.001) \;=\; 1.062 \times 10^{-6}.$$

The fraction of all homozygotes who result from inbreeding is

$$\frac{6.25 \times 10^{-8}}{1.062 \times 10^{-6}} \;=\; 0.0588.$$

So although only one mating in a thousand involves inbreeding, and those are only first-cousin matings, almost six percent of all homozygotes for the recessive allele come from

those matings! These figures are of interest, in that they are reasonable for many genetic diseases in human populations. One in a thousand is not untypical of rates of cousin marriage in Europe and the U. S.

For an overdominant locus, there can also be a decrease of fitness as a result of inbreeding. Suppose the fitnesses of three genotypes are:

AA	Aa	aa
0.5	1	0.5

Mean fitness of a population is increased above one-half by the presence of heterozygotes. As the level of inbreeding increases, heterozygotes will become rarer, so that mean fitness must decline. Since the frequency of heterozygotes at any stage is their original frequency, multiplied by $1 - f$, this decline must be continual throughout the process.

Both of the preceding cases are part of a more general process. Suppose that the fitnesses of the three genotypes AA, Aa, and aa are respectively w_{AA}, w_{Aa}, and w_{aa}. The mean population fitness is:

$$\begin{aligned}
&[p^2(1-f)+pf]\,w_{AA} \;+\; 2p(1-p)(1-f)\,w_{Aa} \;+\; [(1-p)^2(1-f)+(1-p)f]\,w_{aa} \\
&\quad = \; p^2\,w_{AA} \;+\; 2p(1-p)\,w_{Aa} \;+\; (1-p)^2\,w_{aa} \\
&\qquad + f\;\left[(p-p^2)\,w_{AA} - 2p(1-p)\,w_{Aa} \;+\; ((1-p)-(1-p)^2)\,w_{aa}\right] \\
&\quad = \; \bar{w}_0 \;-\; f\;p(1-p)\;[2w_{Aa} - w_{AA} - w_{aa}],
\end{aligned} \tag{V-48}$$

where $\bar{w}_0$ is (clearly) the average population fitness before the onset of inbreeding. The direction of the effect of inbreeding and its size depend on the sign and magnitude of $2w_{Aa} - w_{AA} - w_{aa}$. This quantity is twice $w_{Aa} - (w_{AA} + w_{aa})/2$, which is the difference between the fitness of Aa and the average of the fitnesses of AA and aa. So we can say that the decline of fitness with inbreeding will be greater the farther above the average of the homozygote fitnesses is the heterozygote fitness. We can thus immediately see that mean fitness will decline under inbreeding if (i) the locus is overdominant, or (ii) the allele with higher fitness is dominant, or (iii) even if there is a slight tendency to partial dominance of the more fit allele. Another prediction of (V-48) is that if we plot mean fitness against f, it will decline (or perhaps increase) *linearly with* f. There are no terms in f^2 in (V-48).

MULTIPLE LOCI. This linearity only holds for a single locus. If fitness were a sum of effects from different loci, each effect's average declining linearly with f, we would expect average fitness of the genotype to be linear with f. But when loci do not interact, it is much more natural to assume that the fitness of a genotype is the *product* of fitnesses

at the different loci. For example, if the probability of surviving a risk of death at locus A is w_{AA}, and the probability of surviving a totally unrelated cause of death at locus B is w_{BB}, the probability of surviving both should be $w_{AA}w_{BB}$. If the loci are also at linkage equilibrium, then it can be shown that the mean fitness of individuals in the population can be computed by taking the products of the mean fitnesses at the separate loci. This means that the mean fitness is a product of terms, each a linear function of f. In taking the products of a series of expressions like (V-48), we introduce higher powers of f, so the decline (increase) of fitness is no longer linear. This calculation is fraught with hidden assumptions. One is that the occurrence of identity by descent at different loci is independent. This assumption may not be met even for unlinked loci if there is any variation from individual to individual in f. It is perhaps best to move on without attempting to untangle this particular difficulty further.

Note that there is nothing in the calculation in (V-48) which is specific to fitness. We could as easily be considering any character which is controlled by a single locus and assumes numerical values. We will return to this point when we consider the effects of inbreeding on quantitative characters.

V.15 Some Comments About Pedigrees

.

Pedigrees, particularly regularly constructed ones, are fascinating structures from a logical and combinatorial point of view. Two of their properties are worth special comment.

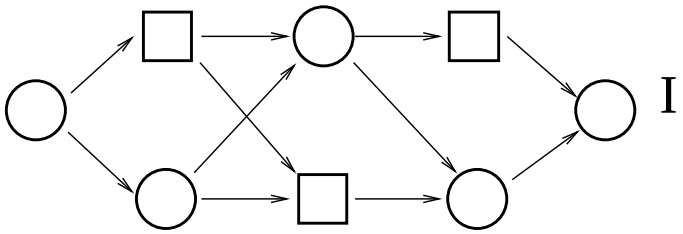
First, it is not possible to sex-label all pedigrees. That is to say, we can draw some pedigrees in which we cannot assign sexes to individuals without finding two individuals of the same sex mating. The reader may wish to stop here and try to invent such a case, then continue reading. The simplest example I know involves three parents and three offspring. The parents are called A, B, and C. Suppose that all three pairwise matings occur, each producing an offspring. So A mates with B, A mates with C, and B mates with C. Since B mates with both A and C, those two individuals must be of the same sex. But A has mated with C, so they must be of different sexes. This is clearly a contradiction. It is resolved by removing our assumption that it is possible to sex-label this pedigree.

Second, a conjecture of Sewall Wright's is worth mentioning. In all the cases we have studied, continuation of the mating system leads to complete inbreeding. Wright noticed that in certain cases this did not occur. These cases turned out to be ones in which the number of ancestors of an individual rise faster than linearly as we move back in time to previous generations. Thus in the case of full sib mating the number of ancestors of an individual are 2, 2, . . . , 2, . . . as we move back in time. But if we had a mating system in

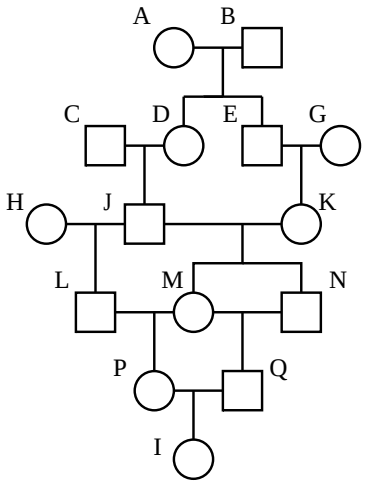
which the number of ancestors was (say) 2, 2, 4, 4, 8, 8, ... we would find that no matter how long the inbreeding has been going on, the level of inbreeding will approach an upper limit short of complete inbreeding. No counterexample is known to Wright's rule. Nevertheless, no rigorous (i.e., correct) proof of it exists, in part because of its generality. I leave it to one of you to prove it.

Exercises

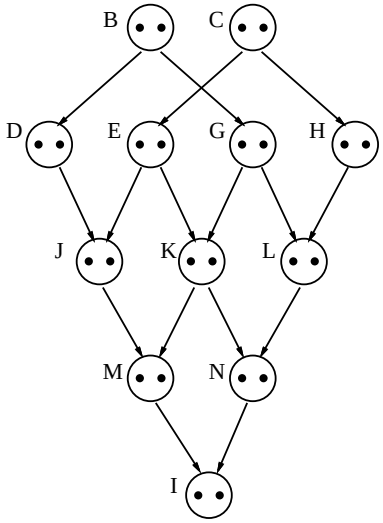
1. If we have a dominant trait whose gene frequency is 0.3, what will be the frequency of the trait when $f = 0$? When $f = 0.2$? When $f = 0.5$? When $f = 1$?
2. James Roosevelt was one son of Eleanor Roosevelt and Franklin Delano Roosevelt, who were fifth cousins (her maiden name was Roosevelt). What was his inbreeding coefficient? (Note – fifth cousins are people who have one parent each that are fourth cousins to each other, and similarly for fourth, third, and second cousins.)
3. Compute f_I in this pedigree: What is f_I when the gene is sex-linked?



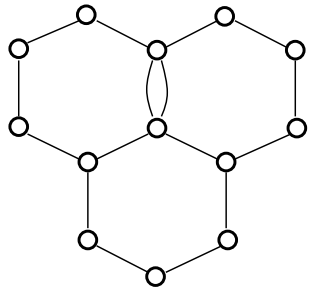
4. What is the inbreeding coefficient of individual I in the following pedigree (written in human genetic form):



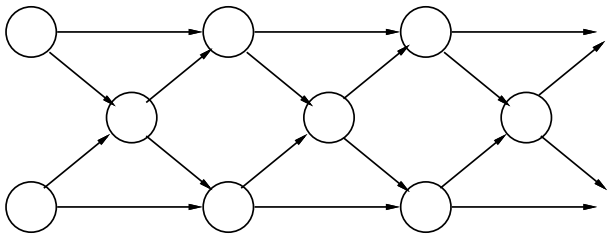
5. Calculate the inbreeding coefficient of individual I in this pedigree. Show the list of loops you find and their contributions.



6. What is the inbreeding coefficient of the bottom individual in this diagram (arrows run downwards and all circles are individuals – note the self-fertilization):



7. For this regular inbreeding system, which runs from left to right, obtain the recurrence relations necessary to analyze it:



8. For a locus with a recessive allele a at a frequency of 0.01, suppose that we divide the population into two parts and then inbreed each subpopulation until the inbreeding coefficient is f. In terms of f, what is

(i) the expected frequency of *aa* in one subpopulation?

(ii) the expected frequency of *aa* in the F_1 cross between the two subpopulations (individuals one of whose parents come from one population and one from the other)?

(iii) the expected frequency of *aa* in the F_2 between the two subpopulations (individuals whose parents are two different F_1 individuals)?

Complements/Problems

1. Suppose that we have a parent-offspring regular system of inbreeding, of the following sort: Individuals #1 and #2 mate to produce individual #3. Then #2 and #3 mate to produce #4, then #3 and #4 to produce #5, and so on. Produce equations for changes in probabilities of non-identity by descent. Obtain the asymptotic rate of decline by finding λ. Compared with the results we obtained for repeated sib-mating, does this system or mating inbreed faster or slower? *(Hint: you will need two coefficients, the probability of non-IBD of an individual, and the coefficient of kinship between the individual and the parent who is about to mate with it. And you will have to figure out how long is one generation.)*

2. What are the recurrence equations for a regular full-sib inbreeding system when we are considering a sex-linked locus?

3. Take the double first-cousins system in Figure 5.10 and turn the picture upside down. What system of inbreeding is it?

4. Suppose that in an infinite random-mating population, a fraction s of (diploid) individuals, chosen at random in each generation, reproduce by self-fertilization, the remainder mating at random. Obtain a recurrence equation for the probability of identity of descent of the two gene copies in an individual chosen at random. How does this f_t change through time?

5. If we take 20 individuals from an infinite, random-mating population and start 10 full-sib lines, continue inbreeding each line for a long time, then cross two of these lines, will the heterozygosity of this hybrid individual be (a) greater, (b) less, or (c) the same as that of a random individual from the original population? *Hint:* (thinking will be more useful than algebraic computation on this one.)

6. Does the result of the previous problem mean that two inbred lines started from a random-mating population will allow us to restore the full variability of the original population by creating a hybrid population by crossing the two lines? Why or why not?

7. If in a population of individuals the average coefficient of inbreeding of newborn individuals is f, and if at a locus there are completely recessive deleterious mutants in the population with gene frequency q and mutation rate to the deleterious alleles u, what is the probability of homozygosity for the mutant allele? If the selection coefficient against that allele is s, what fraction of copies of the deleterious mutant are eliminated in each generation? How large an effect will the inbreeding have on the equilibrium gene frequency of the deleterious mutants?

8. Suppose that we store a large quantity of semen from a prize bull, and carry out the following "parent-offspring" mating system: in each generation a cow is artificially inseminated with some of the semen, and a female offspring is produced. This offspring in turn is artificially inseminated with some of the semen to produce a female offspring in the next generation, and so on. What will be the equations of change in the inbreeding coefficient of these individuals? To what value will f tend through time, and how quickly?

9. Suppose that we start a self-fertilizing line from an *Aa* heterozygous individual. Suppose further that the viabilities of *AA*, *Aa*, and *aa* are in the ratios of $1-s:1:1-s$. In each generation one of the surviving offspring is chosen to continue the self-fertilizing line. Find equations for the change of f_t. To what limit does f_t tend? How does this depend on s? How does mean fitness change with time? How does this compare with the naive result which we would get by applying (V-18) and (V-48)? Explain any discrepancy.

   Suppose that this locus is symmetrically overdominant, so that the fitnesses are

Genotype	*AA*	*Aa*	*aa*
Fitness	$1-s$	1	$1-s$

   How rapidly do the heterozygotes disappear in that case? The model is that a heterozygous individual breeds many offspring (in proportions $\frac{1}{4}:\frac{1}{2}:\frac{1}{4}$), and these then undergo selection with the above fitnesses, deterministically. Then a random survivor is chosen to be the parent. Concentrate on computing the probability that a heterozygote gives a heterozygote as the adult in the next generation. Compute this in terms of s, and explain your logic. *Hint #1: You should follow genotype frequencies among the offspring, not just gene frequencies. Hint #2: Make sure to check your result by seeing whether it does the right thing when $s=0$, which is the no-selection case, and also does the right thing when $s=1$, which will be the case where only heteryozygotes can survive. If it does not do the obvious thing in these two cases then you have made a mistake somewhere. Hint #3: you don't need to do anything involving identity-by-descent to do this one.*

10. Suppose that we have sampled N individuals from an infinite population whose genotype frequencies were in the proportions (V-2). Suppose that we assume that

the sample of size N contains the three genotypes precisely in the proportions given by (V-2). If we compute the chi-square statistic to test departure of the genotype frequencies in the sample from Hardy-Weinberg proportions, what will this quantity be, as a function of N, f and p? How large a sample would be needed, to detect an inbreeding coefficient of 0.05? of 0.01?

11. What is wrong with the preceding calculation as a means of finding the expected value of the chi-square test for departure from Hardy-Weinberg proportions, based on a sample of size N from a population whose genotype frequencies are given by (V-2)? *(Hint:) consider what you would get if $N = 1$, and ask in what ways this differs from the result of the previous problem. Think.)*

12. Suppose that we consider two uniting gametes in a population with given p and f. Let $X = 0$ if the maternal gamete is a, $X = 1$ if it is A. Let $Y = 0$ if the paternal gamete is a, $Y = 1$ if it is A. In terms of p and f: What is the mean of X? The mean of Y? The variance of X? of Y? The covariance of X and Y? The correlation coefficient of X and Y?

13. J.B.S. Haldane considered a system of repeated backcrossing of an individual heterozygous at one locus to a pure line homozygous at that locus. (In repeated backcrossing, in each generation we preserve only the offspring heterozygous at that locus, and breed them to new individuals taken from the pure line). If there is another locus, also heterozygous in the experimental line and also homozygous in the pure line, linked to that locus with recombination fraction r, what are the equations for the probability that the second locus is homozygous in generation t? We repeatedly cross $AB/ab \times ab/ab$, with the offspring that are saved for the next generation being only those that have genotype Aa. Sooner or later, the B allele will be lost, replaced by b. What does this result imply about how large a region around locus A will still be heterozygous after t generations? Is there a simple formula for its expected length?

Chapter VI

FINITE POPULATION SIZE

VI.1 Genetic Drift and Inbreeding: their relationship

You had two parents, they each had two parents, they each had two parents, and so on. If you go back 10 generations, you had 1024 ancestors. At a remove of 20 generations, 1,048,576 ancestors. At a remove of 40 generations (about 1000 years), you must have had over a trillion ancestors! Now, it is quite clear that not that many people have ever lived. How do we resolve the paradox? Quite simply: some of those people were the same individuals, counted many times since they occur many times among your ancestors. Your parents must have been related to each other. Clearly any outbreeding species whose numbers are finite will be subject to the same argument. It therefore becomes of interest to ask how rapidly this sort of inbreeding proceeds. To do this, we make a conceptual model of the process. In this model, we have a population whose size remains constant at N individuals. These are assumed to mate at random, without respect to their relationship to each other. While real populations have a more subtle mating system than this idealization, we will see that, like the Hardy-Weinberg assumptions, this model serves as a useful starting point for discussion of more realistic cases.

Before turning to this task, it will be instructive to consider this finite population of size N from another standpoint. Suppose we took one particular locus, and looked at all the copies of that gene. There are $2N$ of them, and we could label these $1, 2, 3, \ldots, 2N$. Now consider the next generation of individuals (in our idealized model population we have discrete, nonoverlapping generations, unlike actual human populations). Is it possible that each of these $2N$ gene copies is represented exactly once? Yes, but that is unlikely. To happen, each parent would have to be the parent of exactly two offspring, and the two gametes it donates must contain copies of the two different gene copies in that parent (thus an *Aa* heterozygous parent must give an *A* allele to one offspring, and an *a* allele to the other). If there are N parents, each with two offspring, each one has only a 50% chance of giving one copy of each of its two genes to its offspring. So

the probability that each gene copy in the parent generation is represented exactly once among the offspring (even assuming that each parent will have exactly two offspring) is $(1/2)^N$, which becomes vanishingly small very rapidly as we consider larger and larger values of N. For $N = 30$ it is less than one in a billion.

Now suppose that we started out with $2N$ different alleles at a locus, each represented exactly once. The expectation under the Hardy-Weinberg assumptions is that each of these alleles will remain indefinitely in the population with its initial gene frequency, $1/(2N)$. But clearly this is not to be. Some alleles will immediately be lost from the population, by failing to be represented in the next generation. Since there must continue to be $2N$ copies of the gene if population size remains N, the remaining alleles must now be represented more than once each (on the average). The process will then be repeated in the following generation. It will be somewhat harder to lose alleles, since the surviving alleles may be present in more than one copy. Nevertheless some will be lost, and the remainder will be increased in average numbers.

In effect, what is happening as the number of alleles out of the original $2N$ falls, is that fewer and fewer of the gene copies originally present are represented. Ultimately, only one of the original gene copies is represented: the population is fixed for one allele. When it is, two things have happened: the gene frequencies of the alleles have changed, and the copies present are all identical by descent. This makes the point that the random change in gene frequency which is caused by the randomness of Mendelian segregation and by random variations of offspring number is the same process as the increase of identity by descent in a random-mating population. The former process is called *random genetic drift*. Hagedoorn and Hagedoorn (1921) first called attention to its potential evolutionary importance, as did Chetverikov (1926). However, neither of these authors attempted a mathematical treatment. The mathematics of genetic drift was first carefully worked out by Sewall Wright (1929c, 1931). In the following sections more detailed citations will be given.

The reader who has followed the above arguments closely may well be suspicious of one of the steps followed. In the first place, the argument concerning genetic drift seemed special to the case where initially we started with $2N$ different alleles. A moment's thought will show that it is not. If the gene copies in the population are ultimately descended from one of the initial gene copies, then whatever the initial number of alleles, and whatever their frequency one will ultimately fix at the expense of the others.

We have seen that random genetic drift and inbreeding seem to be different aspects of the same process. We now proceed to look into the mathematics of both phenomena and their interconnection. We have not been explicit about the fitnesses of the genotypes. Although genetic drift and inbreeding are always occurring in a finite population, their analysis is far more complex when natural selection, migration, or mutation also occur. For the remainder of this chapter we assume that these forces are absent.

VI.2 Inbreeding due to finite population size

.

Let us consider a population of haploid organisms with constant population size N. In a haploid, there is no Mendelian segregation, so all genetic drift must be due to variations in offspring number. If each of the N individuals had exactly one offspring, then the composition of the population would be exactly reproduced from one generation to the next (barring mutation, and assuming that we are following a single locus). To model the processes of genetic drift and inbreeding, we will have to make some particular assumption about how offspring numbers vary. The assumption we shall make is that each of the N offspring in the next generation is produced by a parent drawn at random independently. This amounts to saying that the pedigree of the group is constructed by drawing in the N individuals of the next generation, then connecting each one to one of the N possible parents, drawn at random and without regard to other links in the pedigree.

This scheme is not chosen simply for its inherent randomness: it corresponds to a life history of the group which has some biological plausibility. Suppose that each parent had a vast number of offspring, but all had the same (vast) number of offspring. Now suppose that this pool of juveniles is subject to mortality by pure random accident, irrespective of genotype and of parentage. The mortality ceases only when the population of juveniles has been whittled down to N. Now consider these N survivors. The first one (we number them arbitrarily) is an offspring of one of the N parents, chosen randomly. The second one is also the offspring of a randomly chosen parent, and what is more important, the fact that the first offspring we examine is descended from parent #17 tells us nothing whatsoever about whether or not the second offspring also comes from parent #17.

That is a result of the vast numbers of juveniles produced by each parent, and the fact that mortality occurs to each juvenile independently. Once it is known that one of the offspring of #17 has survived, the second offspring must be regarded as chosen from among all the other juveniles. But this does not make the next one appreciably less likely also to come from parent #17, as we have ruled out only one of the vast number of juveniles produced by that parent. Our assumption that there is no variation from parent to parent in the (vast) number of juveniles each produces is important, since were it not true, once we knew that parent #17 had provided the first offspring, that information would indicate that this parent was more likely to have produced a large than a small number of juveniles, and thus would also have a higher than average chance of providing the second offspring as well.

Now the model is sufficiently well-specified to permit us to calculate inbreeding coefficients. The reader may be upset at the very notion of inbreeding coefficients in a haploid organism. However, it is meaningful to compute coefficients of kinship between

genes in different individuals. Since we assume that the initial population is also the base population, the initial coefficient of kinship among different individuals is taken to be zero. Letting f_t be the coefficient of kinship of different, randomly chosen individuals in generation t, we will obtain a recurrence relation between f_t and f_{t-1}. We consider two cases. First, the two individuals may be descended from different individuals in the previous generation. Second, they may be descended from the same individual. In the latter case $f = 1$, since they must of necessity contain copies of the same gene. The relative frequencies of these two cases can be obtained by considering that each offspring's parent is drawn at random independently (and "with replacement") from the N parents. Thus if we look at two distinct offspring, once we know from which parent the first of them is descended, the chance that the second one comes from the same parent is simply $1/N$. If the two individuals came from different parents, an event with probability $1 - 1/N$, we may regard them as drawn from two randomly chosen distinct individuals. In that case their coefficient of kinship is simply f_{t-1}. Putting all of this together, we find that

$$f_t \;=\; \frac{1}{N} + \left(1 - \frac{1}{N}\right) f_{t-1}. \tag{VI-1}$$

While this recurrence relation is not difficult to solve, it is made completely transparent by considering the probability of *non-identity* $h_t = 1 - f_t$. We can either substitute $1 - h_t$ and $1 - h_{t-1}$ for f_t and f_{t-1} in (VI-1), or we can reason directly as follows: with probability $1/N$ the parents of the two individuals are the same, in which case $h = 0$. With probability $1 - 1/N$ the parents are distinct and randomly chosen, so that $h = h_{t-1}$. Then

$$h_t \;=\; \left(1 - \frac{1}{N}\right) h_{t-1}. \tag{VI-2}$$

So $1/N$ of the non-identity is lost every generation. Since $h_0 = 1$,

$$h_t \;=\; \left(1 - \frac{1}{N}\right)^t \tag{VI-3}$$

Each generation, the population goes $1/N$ of the remaining distance towards complete inbreeding. It is worth emphasizing that since the gene frequency in a population may wander back and forth, there is no smooth uniform tendency to lose variation in a single population. The approach to complete inbreeding given by (VI-2) and (VI-3) expresses an average of what happens over many populations simultaneously evolving populations.

VI.3 Diploids

If we are discussing inbreeding coefficients it would be convenient to be working with diploid populations. The following is a simple diploid model equivalent to the above

haploid case. Consider a population of hermaphroditic diploids. In each generation the following process takes place: each individual produces a vast (and equal) number of eggs and a vast (and equal) number of sperm. The individuals all spawn into a common gamete pool, like some sessile marine forms. The gametes are thoroughly mixed, so that each of the vast number of juvenile zygotes formed may be regarded as formed by union of a randomly-chosen egg with a randomly-chosen sperm. As in the haploid model, density-dependent mortality acting at random without regard to genotype or parentage, reduces the large number of juveniles to N surviving adults, who will be the parents for the next generation.

This model has the property that if we examine the surviving adults, each may be regarded as if its two parents were drawn from the previous generation independently and at random. Note that it is quite possible for the two parents of an individual to be the same, since the individuals are hermaphrodites, and no prohibition against self-fertilization has been introduced into the model. Also note that when two gene copies are in the same individual, their parents may be regarded as drawn at random, and the same is true when the two gene copies are in different individuals. In effect, each of the $2N$ gene copies in the population comes from a randomly-chosen parent. By the rules of Mendelian segregation, this means that *each of the* $2N$ *gene copies is copied at random from one of the* $2N$ *gene copies of the previous generation*. The resulting inbreeding is straightforward: if we let h_t be the probability of non-identity of two different gene copies (irrespective of whether they are in the same or different individuals),

$$h_t \;=\; \frac{1}{2N}\,(0) + \left(1 - \frac{1}{2N}\right) h_{t-1} \;\;=\;\; \left(1 - \frac{1}{2N}\right) h_{t-1}, \tag{VI-4}$$

so an hermaphroditic diploid organism, mating at random with selfing allowed in a finite population of size N loses $1/(2N)$ of its remaining non-identity, and hence of its heterozygosity, in each generation. In this respect a diploid population of size N is equivalent to a haploid population of size $2N$: when the number of gene copies at locus is the same, so is the rate of inbreeding. Since h_0 is usually taken to be 1, we have that

$$h_t \;=\; \left(1 - \frac{1}{2N}\right)^t, \tag{VI-5}$$

and

$$f_t \;=\; 1 \;-\; \left(1 - \frac{1}{2N}\right)^t. \tag{VI-6}$$

In equations like (VI-3) or (VI-5), we can easily find the number of generations which will be required for half of the heterozygosity to be lost. In the diploid case, setting $h_t = 0.5$ in (VI-5), taking logarithms, and solving for t:

$$t_{0.5} \;=\; \frac{-\ln 2}{\ln\left(1 - \frac{1}{2N}\right)} \tag{VI-7}$$

Table 6.1: Half-life of population heterozygosity (or non-identity) using both exact (VI-7) and approximate (VI-8) formulas, for various population sizes.

	Half-life	
N	Approximate	Exact
1	1.386	1
10	13.863	13.513
100	138.629	138.283
1000	1386.294	1385.948

Now when N is large, since $\ln(1-x) \simeq -x$,

$$t_{0.5} \simeq \frac{-\ln 2}{1/(2N)} = -2N \ln 2 = 1.386N. \qquad \text{(VI-8)}$$

This shows that the time scale for loss of heterozygosity is proportional to population size. This fact, and the accuracy of the approximation (VI-8), are verified by the figures in Table 6.1. Note that the error is never more than four-tenths of a generation.

VI.4 Genetic drift: the Wright-Fisher model

.

HAPLOIDY. We have obtained formulas for the rate of inbreeding due to finite population size. We now want to examine the other side of the finite-populations coin: genetic drift. The model we will introduce was stated by Sewall Wright (1931) and R. A. Fisher (1930) and hence is called the Wright-Fisher model. It is precisely the model(s) of the previous section. The difference is that we are following gene frequency, not inbreeding. For simplicity we will deal only with cases of two alleles, although the Wright-Fisher model is readily extended to multiple alleles.

Consider first the haploid case. We have a fixed population size, N, so that there are only $N+1$ possible gene frequencies: $0/N, 1/N, \ldots, N/N$. Thus, in investigating the gene frequency of the A allele we can just as easily follow the number i of A alleles in the population. Now let us assume that in some population there are presently i copies of A and $N-i$ copies of its allele a. What will happen to the gene frequency of A in the next generation? Recall that we have N offspring, and that each will contain a copy of a randomly chosen gene from the current generation. If $p = i/N$ is the current gene frequency of A, the next generation will represent N tosses of a coin, with probability p that each toss comes up heads (A), for each offspring has the same independent and

random possibility of being a copy of one of the currently-existing A genes. From this we can see that unless $p = 0$ or $p = 1$ in the current generation, the gene frequency in the next generation could assume any of the $N + 1$ possible values $0/N, \ldots, N/N$. For however unlikely it is, it is still possible that all of the N coins will come up heads, or all tails, or any outcome in between. We thus do not have a deterministic single outcome, except when $p = 0$ or $p = 1$.

The best we can hope to do is to characterize the probabilities of the various outcomes. We wish to calculate P_{ij}, the probability that if there are i copies of A this generation, there will be j copies in the next generation. Note that this is a conditional probability, the probability $P(j|i)$ of j given i. By our coin analogy, which is a precise one in the admittedly idealized world of the Wright-Fisher model, we want the probability of obtaining j heads when in N tosses the probability of heads is $p = i/N$. This will be straightforward: it is the binomial probability

$$\begin{aligned} P_{ij} &= \binom{N}{j} p^j (1-p)^{N-j} \\ &= \frac{N!}{j!\,(N-j)!} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j} \end{aligned} \tag{VI-9}$$

(Recall that the notation $\binom{N}{j}$ is the number of combinations of N things taken j at a time). Interestingly enough, formula (VI-9) works not only for the cases where both alleles exist, but also for the cases where $i = 0$ or $i = N$. Then we find $P_{00} = 1$, but all other $P_{0j} = 0$, and $P_{NN} = 1$ but all other $P_{Nj} = 0$. This is what is expected: once A or a becomes fixed in a population, it will remain so forever, as there is no mutation in this model.

DIPLOIDY. Before proceeding to see what we can find out from the P_{ij}, it will be useful to briefly consider the diploid case. Recall that each gene copy in the diploid offspring is independently drawn from a randomly-chosen gene copy in the previous generation. So if we ask only about the gene frequency in the next generation, without regard to how these genes are arranged in genotypes, the result is equally simple. We have $2N$ tosses of a coin whose probability of heads is p, so the probability of getting j copies of A out of $2N$ is simply the binomial probability

$$\begin{aligned} P_{ij} &= \binom{2N}{j} p^j (1-p)^{2N-j} \\ &= \frac{2N!}{j!\,(2N-j)!} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \end{aligned} \tag{VI-10}$$

Thus the diploid case is the same as the haploid case, provided we compare cases with equal numbers of copies of the gene, rather than with equal numbers of individuals.

We could, if we wished, also compute the probability of getting m *AA*'s, n *Aa*'s, and $(N - m - n)$ *aa*'s in the next generation. The reason we will not bother to do so is that the genotype frequencies in this case are an epiphenomenon of the underlying variables, the gene frequencies. If we wanted to find the probability that an individual in the next generation is *AA*, this depends only on the gene frequency in the current generation, being of course p^2. Since gene frequencies are the only variables which affect the status of future generations, we can follow their evolution without ever asking about genotype frequencies.

A MARKOV PROCESS. The P_{ij}, taken together with the initial number of A alleles, completely specifies the process of genetic drift. Because it is random it is called a *stochastic process*. This particular process has the property that its future behavior depends only on its current state (its current gene frequency), not on where it has been in the past. This means that it is a *Markov chain*, named after the Russian probabilist A. A. Markov, who first investigated the behavior of such systems. The states of a Markov process (in this case the different possible numbers of A alleles) can be classified according to how often the process is expected to visit the state. In this case we have two types of state. The states $i = 0$ and $i = 2N$ are *absorbing states* (we are thinking of the diploid case, but the result is entirely analogous in the haploid case). Whenever the process enters either of these states it will stay there forever. The other states are all *transient states*. They are visited only a finite number of times, after which the process never returns to them. In all states for which $0 < p < 1$, there is a probability p^{2N} of going to state $i = 2N$ and a probability $(1 - p)^{2N}$ of going to state $i = 0$ in the next generation. Sooner or later, therefore, the population must fix or lose allele A.

We can use matrix algebra to investigate the behavior of a Markov chain with finitely many states. The reader who is allergic to matrices may wish to skip the rest of this section, as the basic result will be simply that we cannot solve enough of the problem to be of much use.

The basic recurrence equation is as follows. Let $p_k^{(t)}$ be the probability that in generation t the process is in state k. Then

$$p_k^{(t+1)} = \sum_j p_j^{(t)} P_{jk}. \tag{VI-11}$$

The logic of this equation is straightforward: to find the probability that the event is now in state k, we sum over all possible places the process could have been in the previous generation the probability that the process was there, times the probability that it then moved to the state of interest. In matrix terms, if we let $\mathbf{p}^{(t)}$ be the row vector $(p_0^{(t)}, \ldots, p_{2N}^{(t)})$, and if we let $\mathbf{P}$ be the matrix $[P_{ij}]$,

$$\begin{aligned} \mathbf{p}^{(t+1)} &= \mathbf{p}^{(t)}\,\mathbf{P} \\ &= \mathbf{p}^{(0)}\,\mathbf{P}^{t+1}. \end{aligned} \tag{VI-12}$$

The vector $\mathbf{p}^{(0)}$ is the vector specifying the composition of the population in the initial generation. Many of the interesting properties of the Wright-Fisher model are reflected in the sequence $\mathbf{p}^{(t)}$, $t = 0, 1, \ldots$ of state distributions. The vector $\mathbf{p}^{(t)}$ gives the distribution of the different states at time t, over hypothetical replicates of the Wright-Fisher model, all of which are assumed to start in the same state and are assumed to be subject to the same transition probabilities. As t increases, the distribution of states ultimately falls into some equilibrium distribution $(p_0, \ldots, p_{2N})$. Removing the time indices t from (VI-12), this equilibrium distribution must satisfy the matrix equation

$$\mathbf{p} \;=\; \mathbf{p}\,\mathbf{P}. \tag{VI-13}$$

This is the matrix equation $\mathbf{p}(\mathbf{I} - \mathbf{P}) \;=\; \mathbf{0}$, which has an infinite number of solutions, for if a given vector $\mathbf{p}$ satisfies it, so must all multiples of $\mathbf{p}$. To narrow down to the solution we want, we must add the side condition that the elements of the solution $\mathbf{p}$ sum to one. When we do that, we can readily show that the solution $\mathbf{p}$ is a vector with its first and last elements nonzero, and all others zero. In other words, it is of the form $(x, 0, 0, \ldots, 0, 1 - x)$. We will see in the next section that x can be determined without recourse to matrix algebra. So far, all we have shown is that at equilibrium, every population must be fixed for one allele or the other, which is hardly surprising.

EIGENVALUES AND EIGENVECTORS. The general method for working out the vectors $\mathbf{p}^{(t)}$ in cases like (VI-12) is to find all eigenvalues and left eigenvectors of the matrix $\mathbf{P}$. Then if $\mathbf{x}^{(k)}$ is the k-th left eigenvector of $\mathbf{P}$, and λ_k is its associated eigenvalue, if we can represent the initial distribution of states $\mathbf{p}^{(0)}$ by a linear combination of eigenvectors

$$\mathbf{p}^{(0)} \;=\; \sum_{1}^{2N} c_k\,\mathbf{x}^{(k)}. \tag{VI-14}$$

The vector $\mathbf{p}^{(t)}$ can then be computed by multiplying each term by the t-th power of the corresponding eigenvalue:

$$\mathbf{p}^{(t)} \;=\; \sum_{k} c_k\,\lambda_k^t\,\mathbf{x}^{(k)}. \tag{VI-15}$$

If we had expressions for the eigenvalues λ_k, the coefficients c_k (corresponding to a particular initial vector in which we were interested), and the left eigenvectors $\mathbf{x}^{(k)}$, we could write an explicit expression for the elements of $\mathbf{p}^{(t)}$. This would give us much information. The eigenvalues λ_k are in fact known. They are:

$$\lambda_1,\ \ldots,\ \lambda_{2N} \;=\; 1,\ \ 1,\ \ 1 - \frac{1}{2N},\ \ \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right),\ \ \ldots,\ \ \prod_{i=1}^{2N-1}\left(1 - \frac{i}{2N}\right) \tag{VI-16}$$

and were obtained by Feller (1951). (For a simpler derivation of these eigenvalues, see Felsenstein, 1971).

The right eigenvectors of $\mathbf{P}$ may also be obtained (see Karlin, 1966). But there are no known expressions for the left eigenvectors, except for the first two. Those are the eigenvectors corresponding to the absorbing states: $(1, 0, 0, \ldots, 0)$ and $(0, 0, \ldots, 0, 1)$. No one has ever obtained an expression for the third eigenvector, which is the next most interesting one. So the effort to compute $\mathbf{p}^{(t)}$ has so far come to nought, and the same holds for more complex quantities such as the t-step transition probability matrix $\mathbf{P}^t$, or the first-passage times for various states.

However, since we have the transition probabilities P_{ij}, it is possible for small N to compute the quantities $p_k^{(t)}$ numerically, by the simple expedient of repeatedly multiplying the vector $\mathbf{p}^{(0)}$ by the matrix $\mathbf{P}$ in a computer. Figure 6.1 below shows an example of the case $N = 10$ where there are initially 6 copies of A (out of 20 possible). Three different times are shown. Note particularly that when almost all replicates are fixed, the few remaining unfixed are spread out in a nearly uniform distribution over all the unfixed states $1, 2, \ldots, 2N-1$. This is a general pattern no matter what the initial gene frequency (provided it is not 0 or 1).

VI.5 Inbreeding, variances, and fixation probabilities.

FIXATION PROBABILITY. The foregoing section is perhaps excessively gloomy. We cannot find the exact distribution of gene frequencies among replicate populations in the Wright-Fisher model at time t. But we can easily find the mean and variance of the distribution as a function of t, and this gives us much information. Let's start with the mean, where the going is easier. Suppose that x_t is a random variable representing the gene frequency in a population of size N which has been undergoing genetic drift starting at gene frequency p_0. We want to find the mean (expectation) of x_t. Suppose that we knew the expectation of x_{t-1}, and wanted to find the expectation of x_t from this. The expectation of x_{t-1}, which we denote $\mathbb{E}(x_{t-1})$, is its mean over many replicates. In each such replicate, if the current gene frequency is x_{t-1}, the gene frequency in the next generation can be written as $x_t = x_{t-1} + \varepsilon$, where ε is the change of gene frequency in that replicate population. Now the expectation of ε is zero for each given replicate. This is because when we toss a coin $2N$ times with heads probability x_{t-1}, the expectation of the fraction of heads will simply be x_{t-1}. (This property of the binomial distribution can be proven algebraically, but we will not bother to do so here). So taking a given value of the x_{t-1} for a single replicate, and averaging over all possible outcomes in the next generation:

$$\mathbb{E}(x_t) \;=\; x_{t-1}. \tag{VI-17}$$

Now we can further take expectations of that over all replicates, which have different values of x_{t-1}, and we find

$$\mathbb{E}(\mathbb{E}(x_t)) \;=\; \mathbb{E}(x_{t-1}) \tag{VI-18}$$

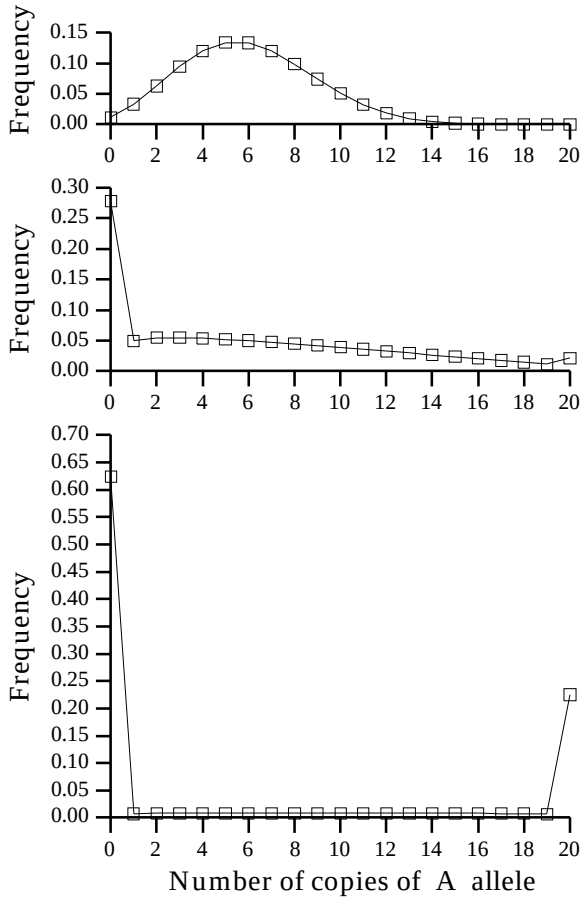


Figure 6.1: Distribution of gene frequencies (given as number of copies of the A allele out of 20) among replicate populations in a diploid Wright-Fisher model with $N = 10$ and initial frequency $p_0 = 0.3$ after 2 (top), 10 (middle), and 40 (bottom) generations. Note that when almost all populations are fixed (T = 40) the remaining populations are distributed nearly uniformly over the unfixed classes.

so that the net result is that the expectation remains the same:

$$\mathbb{E}(x_t) = \mathbb{E}(x_{t-1}) = \ldots = \mathbb{E}(x_1) = p_0. \tag{VI-19}$$

The mean gene frequency over replicates stays at the same value, the initial gene frequency. This has one immediate implication: the probability of fixation of A is the same as its initial frequency. When all populations have become fixed for one allele or the other, the mean gene frequency of A will be the same as the fraction of populations which have fixed A. So this must be the same as the initial gene frequency, by (VI-19).

This accords well with intuition. The process of genetic drift results in all gene copies ultimately being derived from a single copy in the initial generation. The genetic identity of the alleles in those initial copies has no effect on their chances of being the progenitor, since there is assumed to be no natural selection occurring. Therefore the chance that an A allele will be the one chosen to be the progenitor of the population is simply the proportion of initial gene copies which are A. In this sense, although genetic drift makes great changes of gene frequency within any one population, it does not discriminate in favor of one allele as against another, so that over many replicates of the same process it causes no average change in gene frequency.

VARIANCE. While the mean gene frequency remains unchanged, the variance increases through time. Gene frequencies in different replicates are initially the same, and thereafter become more and more different. In generation t the different replicates will vary in gene frequency. If measured in a great many replicates, there is an expected variance of gene frequency among populations, which we can call $\mathrm{Var}\,(p_t)$. We have already seen that there is also an expectation of gene frequency, and that it is the initial gene frequency p_0. The variance of gene frequency is the difference between the expectation of p^2 and the square of the expectation of p, so

$$\mathrm{Var}\,(p_t) \;=\; \mathbb{E}[p_t^2] \;-\; p_0^2. \tag{VI-20}$$

We are going to derive a formula for this variance by using Wahlund's Law, focusing on the average frequency of homozygotes for this allele, averaging over replicates. In the pool of gametes contributing to a replicate population in generation t, the expected probability of an offspring that is a homozygote is p^2, so that the overall frequency of homozygotes is $\mathbb{E}(p_t^2)$, which by the above formula (and by equation (IV-7 for Wahlund's Law, which is equivalent) is

$$P(AA) \;=\; p_0^2 \;+\; \mathrm{Var}\,(p_t) \tag{VI-21}$$

We can also use Sewall Wright's formula for the expected frequency of homozygotes in a population that has been undergoing genetic drift until it has inbreeding coefficient f_t,

$$P(AA) \;=\; p_0^2(1-f_t) \;+\; p_0 f_t \;=\; p_0^2 \;+\; f_t p_0(1-p_0) \tag{VI-22}$$

From equation (VI-6) we already have a formula for f_t,

$$f_t \;=\; 1 \;-\; \left(1-\frac{1}{2N}\right)^t. \tag{VI-23}$$

Substituting it into equation (VI-22) and equating the two expressions, VI-21 and VI-22, and eliminating the common term p_0^2, we get that

$$\mathrm{Var}\,(p_t) \;=\; p_0\,(1-p_0)\left[1 \;-\; \left(1 \;-\; \frac{1}{2N}\right)^t\right] \tag{VI-24}$$

Notice that this formula is zero when $t = 0$ (as it should be), after one generation it is $p_0(1 - p_0)/2N$, the binomial variance in $2N$ trials (as it should be), and ultimately it becomes $p_0(1 - p_0)$. This latter is precisely the variance of a distribution which a fraction p_0 of the time has the value 1, and has value 0 the rest of the time.

GENETIC DRIFT AND MEAN HETEROZYGOSITY. This is also related to the formula for heterozygosity in inbred individuals, which we saw in equation (V-2) is $2p_0(1 - p_0)(1 - f)$. Using the value of f from (VI-23) for the case of genetic drift, we can see that as the variance of gene frequencies among populations increases, the mean heterozygosity decreases. This is, of course, consistent with the Wahlund Effect, as we saw in equation(VI-8).

A COMPUTER SIMULATION. Figure 6.2 shows the outcome of computer simulation, in eight replicates, of random genetic drift in a diploid population of size 10. Both the individual gene frequencies and the means and variances are shown.

It is possible to continue this approach into examination of higher and higher moments of the gene frequency distribution. Although no general expression for the distribution exists, formulas exist for its mean, variance, skewness, and kurtosis. There seems little point in examining the latter two properties here, for as we go beyond the second moment (the variance), the formulas become more and more complicated and the quantities less and less meaningful.

VI.6 Effective population number: avoidance of selfing, two sexes, monogamy.

EFFECTIVE POPULATION NUMBER.

The particular model of the life cycle which was used in the previous section will rarely apply. When it does not, a useful technique will be to compare the rate of inbreeding (or of increase of gene frequency variance) in the more realistic model with that in the idealized model we have discussed. A convenient way to express this is to find that population number in the idealized model which will give the same rate of inbreeding (or the same rate of increase of variance) as we observe in a more realistic model. This is called the effective population number or, equivalently, the effective population size. In what follows we will always compare rates of inbreeding, so that what we will compute will be *inbreeding effective population numbers*, rather than *variance effective population numbers*. These will sometimes differ, but do not in the cases covered in this chapter. The distinction between these two kinds of effective population number was introduced by Crow (1954). The reader will find some further discussion of this point in Crow and Kimura (1970). Ewens (1982) introduced the concept of "eigenvalue effective

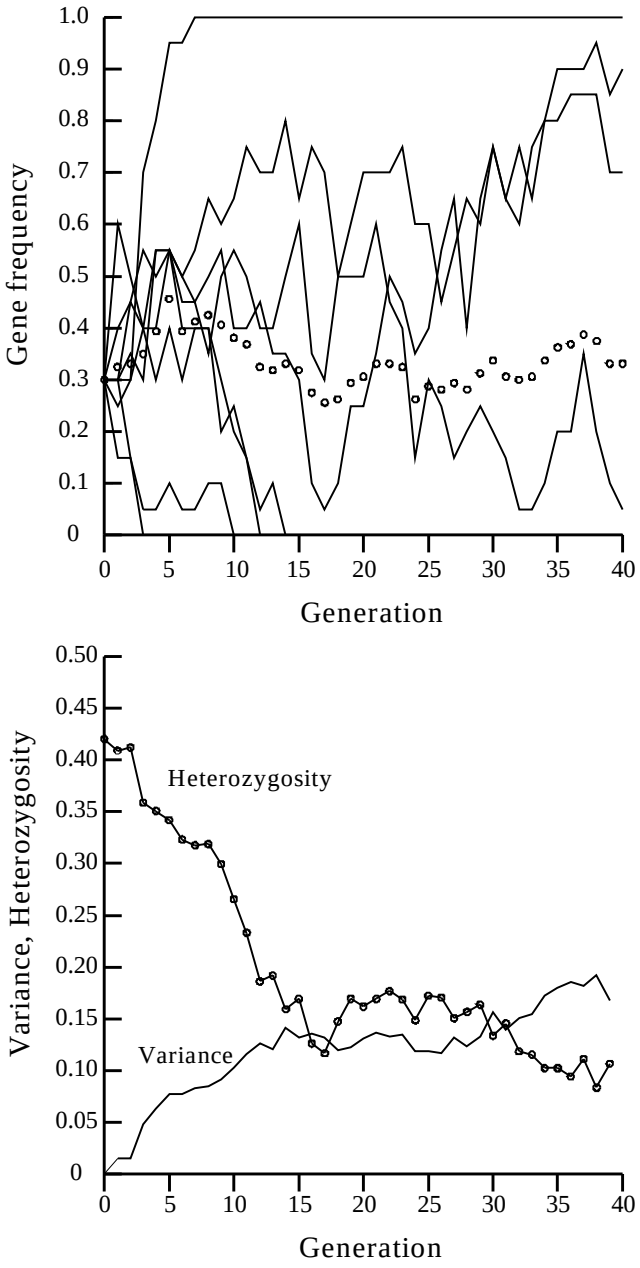


Figure 6.2: Simulated genetic drift in 8 replicates of a diploid Wright-Fisher model with $N = 10$ and $p_0 = 0.3$. The upper graph shows the gene frequencies in the eight replicate populations (lines) as well as the mean gene frequency over those replicates (circles). The lower graph shows for the same simulation the mean heterozygosity within replicates and the variance of gene frequencies among replicates.

population size" as a rigorization and replacement for the variance effective population number.

SELFING NOT ALLOWED. In the idealized model, it was entirely possible that an offspring could be the result of selfing as the result of random collision of an egg and a sperm from the same hermaphroditic individual. If such selfing is impossible, but mating otherwise continues at random, one imagines that the rate of inbreeding will be reduced. By how much is not obvious. Examining this point will make it clear to what extent inbreeding due to finite population size depends on randomly-occurring selfing. Monecious organisms that have both sexes in each individual, but do not allow selfing include earthworms, slugs, snails, and many angiosperms who have both ovules and pollen in each flower. If selfing is prohibited in an hermaphroditic (monecious) population of size N, when we look at two distinct gene copies, their probability of identity by descent may be different depending on whether or not they are from the same individual. After all, gene copies from the same individual cannot be derived from the same individual in the previous generation, whereas copies from different individuals can. Let us examine the change in probabilities of identity through time. Actually, we will use probabilities of non-identity. Let h_t be the probability of non-identity of distinct gene copies from a single randomly-chosen individual. Let k_t be the probability of non-identity of copies chosen at random from two distinct individuals which are themselves chosen at random.

Since distinct copies from a single individual must have come from randomly-chosen distinct individuals in the previous generation.

$$h_{t+1} \;=\; k_t. \tag{VI-25}$$

As for two copies from different individuals, they have probability one-half of having come from the same gene copy in that parent (in which case they cannot be non-identical), and probability one-half of having come from different copies in that parent. Otherwise they must have come from different parents. So

$$k_{t+1} \;=\; \frac{1}{2N}\, h_t + \left(1 - \frac{1}{N}\right) k_t. \tag{VI-26}$$

Substituting (VI-25) into (VI-26) in the usual manner, we get first

$$k_{t+1} \;=\; \left(1 - \frac{1}{N}\right) k_t + \frac{1}{2N}\, k_{t-1} \tag{VI-27}$$

then as the asymptotic multiplier λ of non-identity, the solution of

$$\lambda^2 - \left(1 - \frac{1}{N}\right) \lambda - \frac{1}{2N} \;=\; 0 \tag{VI-28}$$

Table 6.2: Comparison of exact and approximate value of effective population size when selfing is prohibited.

	N_e	
N	Approximate	Exact
2	2.5	2.6180
5	5.5	5.5495
10	10.5	10.5249
100	100.5	100.5025
1000	1000.5	1000.50025

so that

$$\lambda \;=\; \frac{1-\frac{1}{N}+\sqrt{1-\frac{2}{N}+\frac{1}{N^2}+\frac{2}{N}}}{2} \;=\; \frac{1-\frac{1}{N}+\sqrt{1+\frac{1}{N^2}}}{2} \qquad \text{(VI-29)}$$

It is not hard to show that λ is approximately $1-1/(2N+1)$ except for very small N. Since in the idealized model $\lambda \;=\; 1-1/(2N)$, the effective population number is

$$N_e \;\simeq\; N+1/2. \qquad \text{(VI-30)}$$

Table 6.2 compares (VI-30) with the exact value of N_e, computed using

$$N_e \;=\; 1/[2(1-\lambda)] \qquad \text{(VI-31)}$$

and (VI-29). Note the extreme accuracy of (VI-30). It should be clear that an increase of effective population size by half an individual will hardly have any effect on the rate of inbreeding unless N is very small. So randomly-occurring selfing is not a major source of inbreeding with finite population sizes.

SEPARATE SEXES. Now suppose that in addition to preventing selfing, we divide the N individuals into N_f females and N_m males. Each offspring is produced by a randomly chosen female and a randomly chosen male. To analyze the rate of inbreeding, we can use the same two quantities h_t and k_t as before. It might seem that we ought to specify from which sex a gene comes. But the females are simply the first N_f offspring produced, and the males the next N_m (this implicitly assumes that the locus we follow is not sex-linked). Since h_t is the probability of non-identity of the two gene copies from an individual, again we have

$$h_{t+1} \;=\; k_t, \qquad \text{(VI-32)}$$

the fact that the two parent are necessarily of different sexes not affecting the probability of nonidentity of their genes.

When two gene copies are in different individuals, they have chance $1/4$ of both coming from females, in which case they have a $1/N_f$ chance of coming from the same female, and within that a $1/2$ chance of coming from the same gene. There is a similar set of probabilities for males. Half of the time the two gene copies come from parents of different sexes. So

$$k_{t+1} \;=\; \frac{1}{4}\left[\frac{1}{2N_f}h_t + \left(1-\frac{1}{N_f}\right)k_t\right] \;+\; \frac{1}{4}\left[\frac{1}{2N_m}h_t + \left(1-\frac{1}{N_m}\right)k_t\right] + \frac{1}{2}k_t \qquad \text{(VI-33)}$$

This can be rearranged into

$$k_{t+1} \;=\; \left(\frac{1}{8N_f}+\frac{1}{8N_m}\right)h_t + \left(1-\frac{1}{4N_f}-\frac{1}{4N_m}\right)k_t \qquad \text{(VI-34)}$$

Comparing (VI-32) and (VI-34) respectively with (VI-25) and (VI-26), we can see that they will be the same set of equations if we can find a population size N^* such that

$$\frac{1}{N^*} \;=\; \frac{1}{4N_f}+\frac{1}{4N_m} \qquad \text{(VI-35)}$$

so that (VI-34) will become (VI-26), but with N^* instead of N. It has been conventional to define N^* as the effective population size of the population, but this involves a slight inconsistency, for we are then declaring the no-selfing case to be the standard. To continue using the simple Wright-Fisher model as the standard, we can (to good approximation) add $1/2$ to N^*, so that the effective number is (inverting the fraction in VI-35)

$$N_e \;\simeq\; \frac{4N_fN_m}{N_f+N_m}+\frac{1}{2} \qquad \text{(VI-36)}$$

To compute N_e more precisely, one could use (VI-35) to get N^*, then replace N in (VI-29) by that value. Table 6.3 shows how N_e, computed from (VI-36), is affected by different sex ratios when a population of constant size is divided into different numbers of females and males: Note that for extreme sex ratios the effective population number is closer to four times the numbers of the sex in shortest supply, but when both are equally frequent, it gets close to the total number of individuals, being $N+1/2$ when both sexes are equally frequent, just as it was when the individuals were hermaphrodites.

MONOGAMY. Another way in which the Wright-Fisher model departs from reality is the absence of monogamy. Some non-human species form monogamous pairs for life. In the models discussed so far, if one offspring comes from, say, parents #7 and #29, the next offspring might well come from parents #7 and #18. We can instead imagine a population with $N/2$ females and $N/2$ males, which are randomly formed into pairs

Table 6.3: Approximate effective population number for different numbers of females and males.

N_f	N_m	N_e
1	99	4.46
5	95	19.5
10	90	36.5
25	75	75.5
50	50	100.5

(without replacement). Each offspring is produced by choosing one of the $N/2$ pairs at random and having it produce the offspring. This is done $N/2$ times, sampling pairs with replacement. Once again, the only relevant distinction between gene copies will be whether or not they are in the same individual. Without going into great detail, the resulting equations are:

$$h_{t+1} = k_t \tag{VI-37}$$

and

$$\begin{aligned} k_{t+1} &= \tfrac{1}{N/2}\left[\tfrac{1}{2}\cdot\tfrac{1}{2}h_t + \tfrac{1}{2}k_t\right] + \left(1 - \tfrac{1}{N/2}\right) k_t \\ &= \tfrac{1}{2N}\, h_t + \left(1 - \tfrac{1}{N}\right) k_t \end{aligned} \tag{VI-38}$$

Comparison will show that these are precisely the same as equations (VI-25) and (VI-26) so, surprisingly, *enforcing monogamy has no effect on effective population number*.

We have thus found that avoidance of selfing has little effect on effective population number, that enforced monogamy has no effect, but that unequal sex ratios can have a substantial effect, reducing the effective number.

HISTORY. Most of the work reported in this section was first done by Sewall Wright, in his classic 1931 paper. The terminology of effective population number was introduced later by Wright (1938b). The computations regarding monogamy seem to have been done first by Moran and Watterson (1959), as this case was overlooked by earlier workers. Crow and Kimura (1970) discuss the distinction between inbreeding and variance effective numbers more carefully. A careful comparison of the concepts of variance, inbreeding and "eigenvalue" effective population sizes is made by Ewens (1982).

VI.7 Varying population size and offspring number.

VARYING POPULATION SIZE. When population size changes through time, the rate of inbreeding, or of increase of gene frequency variance, will also vary. It is a simple matter to find that population number which would, over the same period of time, lead to

the same amount of inbreeding in a simple Wright-Fisher model. Consider two Wright-Fisher models, one of constant size N_e, the other having a series of sizes in successive generations: $N_1, N_2, N_3, \ldots$ There is no difficulty in defining the latter model: one simply assumes that in generation t, from among the infinite numbers of zygotes produced by random union of gametes of the previous generation, only N_t survive (at random) to adulthood. In the constant size-population, after t generations the probability of non-identity by descent is reduced from 1 to

$$h_t \;=\; \left(1 - \frac{1}{2N_e}\right)^t. \tag{VI-39}$$

In the varying population, this probability is

$$h_t \;=\; \left(1 - \frac{1}{2N_0}\right)\left(1 - \frac{1}{2N_1}\right)\ldots\left(1 - \frac{1}{2N_{t-1}}\right). \tag{VI-40}$$

To find the effective population number, we must equate (VI-39) and (VI-40) and solve for N_e. This we can do by first taking the t-th root of both expressions

$$1 - \frac{1}{2N_e} \;=\; \left[\prod_{i\,=\,0}^{t-1}\left(1 - \frac{1}{2N_i}\right)\right]^{1/t} \tag{VI-41}$$

and then solving:

$$N_e \;=\; \frac{1}{2\left(1 - \left[\prod_i \left(1 - \frac{1}{2N_i}\right)\right]^{1/t}\right)} \tag{VI-42}$$

A useful approximation may be developed as follows: when all of the N_i are large, we can approximate (by a binomial expansion)

$$\left(1 - \frac{1}{2N_i}\right)^{1/t} \;\simeq\; 1 - \frac{1}{2N_i t}. \tag{VI-43}$$

Putting this into (VI-41), after some rearrangement and discarding terms with $1/(N_i N_j)$,

$$N_e \;=\; \frac{1}{\left(\sum_i \frac{1}{N_i}\right)\Big/ t} \tag{VI-44}$$

This formula computes the reciprocal of the average of reciprocals. Simply put, we go onto the scale $1/x$, average the values there, and then take this quantity and come back to the original scale of values of N. This is a well-known quantity: it is known as the *harmonic mean* of the N_i. A well-known property of harmonic means is that they are

Table 6.4: Effective population number after bottlenecks in population number of varying length.

Number of generations at:		Effective population number	
$N_i = 10$	$N_i = 1000$	N_e (approximate)	N_e (exact)
1	99	502.51	496.25
5	95	168.07	164.74
10	90	91.74	89.86
25	75	38.83	38.13
50	50	19.80	19.56
75	25	13.29	13.21
90	10	11.10	11.07
99	1	10.10	10.10

closer to the minimum of the quantities than is the ordinary arithmetic mean. We have seen this property before: in the previous section, the effective population number with two sexes was twice the harmonic mean of the number of the two sexes. In the present case, the reason for using the reciprocals is that we are averaging the increases of the inbreeding coefficient f, and in each generation these are approximately $1/(2N_i)$.

Table 6.4 gives some idea of the effects of varying population number: in it we suppose that population number is 10 for a certain number of generations, and 1000 for the rest of a stretch of 100 generations. The exact effective population number is also given.

Note the relatively strong effect of even a few generations of reduced population size. Note that when half of the generations are at $N_i = 10$ and the other half at $N_i = 1000$, the effective population number is much less than their average, which would be 505. Note that the approximation (VI-44) is quite good. Since its accuracy depends on (VI-43), which is accurate for large N_i, the approximation is better for larger N_i.

VARIATION IN FITNESS. Changes in population size might result from variation in fitness from generation to generation. When fitness varies within a generation, this also affects effective population number. If the variation is genetic, and is due to the genes whose drift or inbreeding we are examining, then we are examining the interaction of genetic drift and selection. We reserve this complex subject until the next chapter. For the present we deal with the case where the fitness variation is not inherited.

Suppose that we have a population following a Wright-Fisher model with selfing allowed, except that at the stage where games are produced, each individual produces an infinitely large but different number of gametes. Suppose that the i-th individual produces a number of gametes proportional to the fitness w_i. It is clear that this can affect the effective population number, since if w_i is very small for all but one individual,

so that this individual does most of the reproducing, the rate of inbreeding will be much increased. We can compute the effective population number as a function of the mean and variance of the w_i within the generation. Since random selfing is allowed, we need only one quantity, h_t, to analyze inbreeding in this case. The probability that a random gamete comes from the i-th individual is $w_i / \sum_i w_i$. The chance that both of two gametes came from individual i is $w_i^2/(\sum_i w_i)^2$. The overall probability that two gametes come from the same parent is $\sum_i w_i^2/(\sum_i w_i)^2$. The equation for change of the probability of nonidentity is:

$$h_{t+1} \;=\; \left(1 - \frac{1}{2}\sum_i w_i^2 \Big/ \left(\sum_i w_i\right)^2\right) h_t. \tag{VI-45}$$

Comparing this to the usual formula

$$h_{t+1} \;=\; \left(1 - \frac{1}{2N_e}\right) h_t, \tag{VI-46}$$

we find that

$$N_e \;=\; \left(\sum_i w_i\right)^2 \Big/ \left(\sum_i w_i^2\right). \tag{VI-47}$$

The variance of fitness is the difference between the expectation of the squares and the square of the expectation, which is $\sum_i w_i^2/N - \bar{w}^2$. The mean fitness is $(\sum_i w_i)/N$, so that $\sum_i w_i \;=\; N\bar{w}$, and using that we can rewrite the expression for N_e as

$$N_e \;=\; N^2\bar{w}^2/(NV_w + N\bar{w}^2). \tag{VI-48}$$

which after dividing the top and bottom of the fraction each by $N\bar{w}^2$ reduces to

$$N_e \;=\; N/(1 + V_w/\bar{w}^2) \;=\; N/(1 + C_w^2). \tag{VI-49}$$

where C_w^2 is the squared coefficient of variation of fitness. Notice from (VI-49) that variation in fitness reduces the effective population number: if the standard deviation of fitness is half its mean, effective population number is reduced by 20%. Careful consideration of this derivation will show that we can only use the expressions for $\bar{w}$ and V_w if we assume that the sum of the w_i is a fixed quantity in each generation.

We may prefer to express the variation in offspring number from individual to individual directly, rather than in terms of V_w. Suppose that in a given generation, n_i is the number of gametes contributed to the surviving offspring (those that reach adulthood) by the i-th parent. The number of ways to choose two distinct gene copies in the offspring is $2N(2N-1)$. Of these, $n_i(n_i-1)$ are choices of copies from the same parent. The total probability of choosing two gene copies from the same parent is

$$P \;=\; \frac{\sum_i n_i(n_i-1)}{2N(2N-1)} \tag{VI-50}$$

This quantity is the equivalent of $1/N_t$, in the discussion of cases where N_t varies between generations. As we have seen a good (approximate) way of computing the effective number when N_t varies is to compute the harmonic mean of the N_t. This is

$$\frac{1}{N_e} \;=\; \mathbb{E}\left[\frac{1}{N_t}\right] \;=\; \mathbb{E}\left[\frac{\sum_i n_i(n_i-1)}{2N(2N-1)}\right] \;=\; \frac{\mathbb{E}\left[\sum_i n_i^2\right] - \mathbb{E}\left[\sum_i n_i\right]}{2N(2N-1)} \;=\; \frac{\sum_i \mathbb{E}[n_i^2] - 2N}{2N(2N-1)}. \tag{VI-51}$$

Note that a sum of N identical terms is N times one of those terms, so that $\sum_i \mathbb{E}[n_i^2] = N\mathbb{E}[n_i^2]$. Now the variance of offspring number of a random individual from a random generation is

$$V_n \;=\; \mathbb{E}[n^2] - \mathbb{E}[n]^2 \;=\; \mathbb{E}[n^2] - 2^2, \tag{VI-52}$$

since the constancy of total population size necessarily implies that $\mathbb{E}[n] = 2$. Solving this for $\mathbb{E}[n^2]$ and substituting, this into (VI-51), we can solve for N_e, getting:

$$N_e \;=\; \frac{2N(2N-1)}{N(4+V_n) - 2N} \;=\; \frac{4N-2}{2+V_n}. \tag{VI-53}$$

This formula needs a bit of interpretation. Note that V_n is the variance of the number of offspring of a randomly chosen individual, in the sense of the number of gametes contributed to the next generation. Thus an offspring produced by random selfing counts as two offspring. Note that as V_n increases, the effective population number declines. In one extreme case, where all individuals contribute exactly two gametes to the next generation, $V_n = 0$, so that $N_e = 2N - 1$. This emphasizes that effective population number need not always be less than census number. When the variance of offspring number is determined by random sampling from a pool of gametes contributed equally by all parents, we have the case of the Wright-Fisher model. There are $2N$ gametes sampled from the pool, and on each draw a given parent has probability $1/N$ of having one of its gametes chosen. The variance of offspring number is simply the binomial variance

$$V_n \;=\; 2N\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right) \;=\; 2-\frac{2}{N}. \tag{VI-54}$$

Substituting this into (VI-53) gives N, as expected. Note one important difference between (VI-48) and (VI-53). In the former case, variation in offspring number can never increase N_e. This is because the random sampling of 2N gametes from the gamete pool creates a certain irreducible variance in offspring number. Formula (VI-49) does not assume this random sampling, so V_n can be smaller than $2 - 2/N$, which is the smallest V_n achievable under the model of (VI-48).

We can regard changes in population size and variation of offspring number as examples of the same phenomenon: variation of the number of gametes contributed to the next generation between random individuals chosen from random generations. When

this variation exceeds the expected from the Wright-Fisher model, there is a reduction in effective population size.

As with most results in this chapter, both (VI-44) and (VI-53) are due to Sewall Wright (respectively 1931 and 1938b). Formula (VI-49) is a variation on a result of Alan Robertson (1961).

VI.8 Other effects on effective population number.

There are many other phenomena which can affect effective population number, and too little space here to discuss them all. I hope to briefly touch on two: overlapping generations and linkage.

OVERLAPPING GENERATIONS. When generations overlap, this may affect the variation of offspring number. One simple model of overlapping generations deserves special mention: the model of P. A. P. Moran (1958). The model has many variants: we will consider the variant which is closest in spirit to the haploid Wright-Fisher model. We have a monecious haploid population of N individuals. In the Moran model, instead of all parent individuals dying simultaneously upon the birth of the offspring generation, one parent dies at a time. Time is divided into units, and at each unit of time, one individual chosen at random dies. Before it dies, a parent chosen at random (with replacement – selfing is allowed) produces the offspring which will replace it. Thus each time unit may see either a small change in the genetic composition of the population or none at all. It is not hard to show that if we choose two different gene copies from the population, and call their probability of non-identity at time t, h_t,

$$h_{t+1} \;=\; \left(1-\frac{2}{N}\right)h_t \;+\; \frac{2}{N}\left(1-\frac{1}{N}\right)h_t \;=\; \left(1-\frac{2}{N^2}\right)h_t. \qquad \text{(VI-55)}$$

The value $2/N^2$ is much smaller than $1/N$, but the two cannot be compared, since one time unit in the Moran model is far less than a generation. Since an individual lives an indeterminate amount of time, we can only compute an average generation time. Since death and reproduction each occur every time unit with probability $1/N$, the average generation length turns out to be the same as the average lifespan: N time units.

Equating the amount of reduction in h_t in a generation with the corresponding value in a haploid Wright model,

$$1-\frac{1}{N_e} \;=\; \left(1-\frac{2}{N^2}\right)^N. \qquad \text{(VI-56)}$$

If N is large, the right-hand side of (VI-56) is nearly $(1-2/N)$, which means that

$$N_e \;\simeq\; N/2. \qquad \text{(VI-57)}$$

The effect of overlapping generations is to cut the effective population size in half. This is not a general rule, although overlapping generations usually result in some reduction of effective population size. There is a substantial literature on this subject (Kimura and Crow, 1963; Nei and Imaizumi, 1966; Felsenstein, 1971; Crow and Kimura, 1972; Hill, 1972c). Hill's paper is worth particular notice because it expresses the effective population number with overlapping generations in terms of the variance of offspring number. This variance is the real reason for the effect of overlapping generations on effective population number. In the Moran model given above, it is not hard to show that the number of offspring of a given individual, over its lifetime, follows a roughly geometric distribution with mean 1. This distribution has a variance of about 2, which is twice as much as the binomial distribution of offspring numbers implied by the Wright-Fisher model. It is precisely this factor of two which causes the reduction of N_e in the Moran model.

More recent work on overlapping generations has generalized my model to diploids (Emigh and Pollak, 1972) and combined it with Hill's approach to more fully take variation of offspring number into account (Waples, 2007; Waples, Do, and Chopelet, 2011). Olsson and Hössjer (2015) have incorporated overlapping generations into the estimation of effective population number by temporal sampling in small populations.

LINKAGE. The effect of linkage on effective population number is more complex than the effect of overlapping generations. A few heuristic examples will have to serve here. The effect is not, strictly speaking, one of linkage alone, but rather of linkage of the locus in question to nearby loci at which selection is occurring. This can greatly exaggerate the effect of background variation in fitness. We have already seen in equation (VI-48) the effect of variation in fitness on effective population number. But this variation was assumed to be non-genetic: if the fitness of an individual was w_i this generation, the fitnesses of its offspring were assumed to be drawn at random. Nei and Murata (1966) have shown that heritability of the fitnesses in the genetic background can exacerbate the effect of fitness variation in increasing inbreeding. But when the background loci are linked to the locus in question, the effect can be much more dramatic. This has been investigated by Hill and Robertson (1966), in the course of computing the effect of linkage on limits to artificial selection. Maynard Smith and Haigh (1974) have called attention to essentially the same phenomenon, calling it the *hitch-hiking effect*. When an advantageous mutant occurs at low frequency in the population, then rises to high frequency, it carries with it alleles which happen to have been present at nearby loci in the original set of chromosomes containing the favorable alleles. If the loci are weakly linked, their random initial linkage disequilibrium rapidly breaks down: the association with the favored allele causes increase of alleles at nearby loci only for as long as the association persists. The more closely linked are the loci, the more dramatic the effect of the selected locus on its hitch-hiking neighbor. The result is large random changes in gene frequency: in effect a great reduction in effective population size. Wagener and Cavalli-Sforza (1975)

have proposed that hitch-hiking can explain much of the variation in gene frequency of genetic diseases (cystic fibrosis and Tays-Sachs syndrome, for example) between human populations. As hitchhiking involves natural selection on linked loci, it will be covered further in chapter VIII.

VI.9 Hierarchical population structure.

In the preceding sections we always used the population from which the initial generation is assumed to be drawn as our base populations. Now we want to explore the consequences of considering different base populations. Suppose we have an infinitely large total population, denoted by T, which consists of infinitely many subpopulations. We let S stand for a randomly selected subpopulation. Now let

h_I = the probability of non-identity of gene copies from a randomly selected individual,

h_S = the probability of non-identity of distinct gene copies sampled from a random subpopulation,

and h_T = the probability of non-identity of two distinct gene copies sampled at random from the total population.

Note that since there are infinitely many subpopulations, gene copies sampled from the total population at random are certain to come from different subpopulations.

Now define three quantities:

$$\begin{aligned} H_{IT} &= h_I/h_T, \\ H_{IS} &= h_I/h_S, \\ H_{ST} &= h_S/h_T. \end{aligned} \tag{VI-58}$$

It is an algebraic necessity that

$$H_{IT} = H_{IS}\, H_{ST}. \tag{VI-59}$$

The interest in this computation lies in the definition of the quantities H_{IT}, H_{IS}, and H_{ST}. Suppose that the total population T was the base population for computation of inbreeding. Then by definition $h_T = 1$, so H_{IT} is simply the probability of non-identity of individual I. What is not perhaps as easy to see is that even when T is not the base population, H_{IT} tells us what the probability of non-identity of I would be *if T were the base population*. We will not try to prove this, since to do so we would have to redefine the H's in terms of correlations rather than probabilities. It is at least reasonable to make this interpretation of H_{IT}, since clearly h_I/h_T is a measure of whether I is more inbred than T, and such a measure is what is required. When S is a population ancestral to

the current population but descended from the original base population, the principle is easily established, as

Prob (The two genes in S are descended from distinct genes in the base population)

= Prob (They were descended from distinct genes in T *and* Those were in turn descended from distinct genes in the base population)

= Prob (They were descended from distinct genes in T)
$\times$ Prob (Randomly chosen distinct genes in T are descended from distinct genes in the base population)

So if H_* is the measure we are looking for measuring the non-identity of S relative to T as the new base population:

$$h_S \;=\; H_*\, h_T \tag{VI-60}$$

so that H_{ST} is the quantity we seek which we have called H_*. There is an implicit assumption here which we must mention: that the two gene copies in S may be regarded as descended from randomly chosen copies in T if they are descended from distinct copies in T. This is what allows us to take the product of probabilities.

We can regard H_{IT}, H_{IS}, and H_{ST} as probabilities of non-identity (at least when they are between 0 and 1). They can be written in terms of the corresponding values of F:

$$(1 - F_{IT}) \;=\; (1 - F_{IS})(1 - F_{ST}). \tag{VI-61}$$

The quantity F_{ST} measures how much more inbred two genes from the same subpopulation are than two genes from different subpopulations. It can be regarded as a measure of the extent to which the subpopulations are differentiated from each other.

POPULATIONS SEPARATED FOR t GENERATIONS. We can get some sense of the meaning of F_{ST} from the following. Suppose that we start with an infinitely large population within which the probability of identity by descent of two random gene copies is f_{w0}. We divide it into an infinite number of subpopulations of size N, and each then undergoes inbreeding (and genetic drift) according to a simple diploid Wright-Fisher model for t generations with no intermigration. Initially the probability of identity of two distinct genes in the same subpopulations is f_{w0}. After t generations it is (from (VI-5))

$$f_{wt} \;=\; 1 - (1 - f_{w0}) \left(1 - \frac{1}{2N}\right)^t. \tag{VI-62}$$

If we choose two gene copies from different populations at time t, their probability of identity is not a function of t, for each was descended from some randomly-chosen copy

in the zero-th generation of its subpopulation, and those had probability f_{w0} of identity. So if f_{bt} is the between-population inbreeding coefficient in generation t,

$$f_{bt} \;=\; f_{b0} \;=\; f_{w0}. \tag{VI-63}$$

Now considering the subpopulations at time t as being the subpopulations S, and the total collection of subpopulations as being T,

$$\begin{aligned} h_I \;&=\; h_S \;&=\; 1 - f_{wt} \\ h_T \;&=\; 1 - f_{bt} \;&=\; 1 - f_{w0}. \end{aligned} \tag{VI-64}$$

So from (VI-62)

$$\begin{aligned} F_{IS} \;&=\; 1 - H_{IS} \;&=\; 1 - h_I/h_S \;&=\; 0, \\ F_{IT} \;&=\; 1 - H_{IT} \;&=\; 1 - h_I/h_T \\ &&=\; 1 - (1 - f_{wt})/(1 - f_{w0}) \;&=\; 1 - \left(1 - \tfrac{1}{2N}\right)^t \\ F_{ST} \;&=\; 1 - h_S/h_T \;&=\; 1 - h_I/h_T \;&=\; 1 - \left(1 - \tfrac{1}{2N}\right)^t. \end{aligned} \tag{VI-65}$$

These tell us the following: since $F_{IT} = F_{ST}$, gene copies from the same individual are no more inbred than are copies from different individuals, so that the inbreeding coefficient of individuals in a Wright-Fisher model relative to their own subpopulation is $F_{IS} = 0$. The quantity F_{ST} is a measure of the accumulated inbreeding, or alternatively the amount of genetic divergence of the populations. This computation makes explicit that all of the inbreeding which accumulates as a result of genetic drift is the result of random changes in gene frequency: the individuals in a subpopulation are not inbred when their own population's current composition is taken as the base population.

Exercises

1. How many generations will it take for a diploid Wright-Fisher model population to lose 90% of its initial heterozygosity?

2. Why isn't the process of genetic drift like that of tossing a coin repeatedly with probability of heads p? In that case we would expect in the long run to get a fraction p of heads, rather than ultimately getting a run of heads (fixation of A) or of tails (a). Where does the analogy break down?

3. In a sample of size $N = 1$ from a random-mating population of size, in which there are found to be one A and one a gene, what will you compute from the gene frequencies to be the expected proportions of the three genotypes? Why does the observed genotype depart from Hardy-Weinberg expectation? Ask and answer the analogous question for $N = 2$. What is the average fraction of Aa individuals in a sample of size $N = 2$ produced by a Wright-Fisher model,

   (i) Given that we find $p_A = 0.5$ in that sample?

   (ii) Given that $p_A = 0.5$ in the population?

   (This can be done by enumerating all possibilities. Be careful to weight them by their probabilities of occurrence.)

4. A new mutant occurs as a single gene copy in a diploid population of size N. What is the probability that it will drift to fixation?

5. As a rare allele happens to drift to fixation, heterozygosity obviously increases at first, then decreases. How can this be reconciled with the notion that f always increases through time?

6. Is there any of the cases we have covered in this chapter which is the same, for a particular value of N, as repeated full-sib inbreeding? Find it and compare the numerical results given in this chapter with those given in chapter V.

7. In an organism like fur seals, perhaps 10% of the bulls do all the breeding in each generation. By how much does this affect effective population number?

8. Wild cheetahs (*Acinonyx jubatus*) have almost no natural genetic variation. They were formerly distributed in Africa, the Near East, and South Asia and had an estimated population size of 100,000. Now they are down to about 10,000 individuals. Someone suggests that this bottleneck of population size explains the loss of genetic variation (as indicated, for example, by their level of heterozygosity). If the reduction of population size to 10,000 happened 100 years ago, and that population size has been maintained since then (with a cheetah generation being about 7 years), by what fraction would the resulting genetic drift (and inbreeding) have reduced their original heterozygosity? What does this imply about this suggested explanation for the loss of cheetah genetic variability?

9. A flock of finches flies to an island and there founds a large population. It is at population size 10 for the first 5 generations and then suddenly grows to a large population size and stays there. What fraction of the heterozygosity at an unselected locus will be expected to have been lost in passing through this bottleneck? Would this mean that most loci would lose all their variability in this event?

10. Suppose that a large diploid population is reduced to a single mating pair, then population sizes doubles every generation thereafter until the original population number is reached. How much inbreeding accumulates during the crisis? How does this compare with the amount of inbreeding which would have accumulated if after the reduction to a single pair, the population had instantly returned to its original size? You can use equation (VI-44) as a starting point. (No, I am not going to specify the original size except to say that it is large).

11. Suppose that in a large population of size N, in each generation one-half of the individuals happen to find good nesting sites and have an average fertility of 3, while the other half of the individuals find inferior nesting sites and have an average fertility of 1. So the gamete pool consists of $3(N/2)G$ gametes from the lucky ones, who all contribute equally, and $(N/2)G$ gametes from the unlucky ones, who all contribue equally. G is very large. Each generation the parents die off, and there is no correlation between the goodness of the nesting site of the parent and that of its offspring. What is the effective population size as a function of N ? *(Hint: you can follow these steps and not use any formula from the book: (1) Try to compute the probability that two gametes drawn at random from the pool come from the same parent, (2) equate that to* $1/N_e$*, and (3) solve for* N_e*. To do step (1) compute the probability that both gametes are ones that came from lucky individuals, and the probability that both came from unlucky ones. Keep in mind that there are* $N/2$ *lucky parents and* $N/2$ *unlucky parents. For each of those cases compute the fraction of those times that they came from the same individual parent. Then multiply and add appropriately).*

12. Among human males in this country, under the traditional naming system, family names behave as if they were Y-linked. What does genetic drift theory tell us about how rapidly diversity of names should disappear if the population stays the same size? if it grows exponentially at a constant rate?

13. If a population starts out at size N and grows by 2% per generation without limit, how much inbreeding will ever accumulate in it? Use an approximation for large N, summing a geometric series.

14. Suppose that in a population with N adults, each parent produces 4 offspring, and 50% of the offspring (taken at random) die before maturity. What will be the effective population size? (Remember that it need not be true that exactly two offspring from each parent survive?).

Complements/Problems

1. If we have a population of diploids who reproduce according to a Wright-Fisher model, but all reproduction is asexual apomixis (clonal reproduction), what are the equations for f_t and g_t be if the population is of size N?

2. See if you can obtain (VI-5) using the probabilities of identity f_t and g_t.

3. How long will it take a diploid population following a simple Wright-Fisher model to lose all but $1/e = 1/2.71828 = 0.36789$ of its initial heterozygosity, if N is large? (Express this as a multiple of N).

4. Suppose that we have a discrete-generations random-mating population of N diploid individuals, except that in each generation there is a constant probability s for each individual that it is produced by selfing. This happens independently for each individual, independently of the others and of whether its own parent was produced by selfing. Consider the probability of non-IBD for two copies in the same individual, and the probability of non-IBD for two copies in different individuals. Work out equations for their change from generation to generation. What is the effective population size as a function of N and s?

5. In a diploid population with random mating, suppose that all matings are monogamous. That is, the individuals are formed randomly into $N/2$ pairs, and offsping are produced by drawing a random pair, having them produce an offspring, and then returning the pair to the pool of pairs. This is done N times. Work out the equations for non-IBD in this population. How does the effective population size compare with the no-selfing-allowed case? *(Hint: you will need to follow two quantities).*

6. Mitochondria are effectively haploid (in that all the mitochondria in an individual are usually copies of one which occurred in the egg). All of them come from the female parent. Suppose that we have a population in which there are N_f females and N_m males in each generation, but which otherwise follows a Wright-Fisher model.

   (i) What is the equation for change of non-IBD for a gene located on the mitochondrion (as these are haploid, we of course mean non-IBD between copies from mitochondria in two random individuals)? What is the effective population number for these genes?

   (ii) Suppose that there is actually a small probability m that an individual has its mitochondria coming from the male parent (and that this event is independent in different individuals). Derive the equations for change of non-IBD

between genes on mitochondria in two different individuals. How many non-IBD quantities do you need to follow? Do you need to have different variables for non-IBD in two males, in two females, or in one male and one female?

(iii) What is the effective population size for a mitochondrial locus as a function of N_f, N_m, and m? Do small values of m make a dramatic difference in the effective population size?

7. An otherwise-idealized infinitely-large population goes through a bottleneck, dropping to 100 individuals, and then increases immediately back up to infinity in the next generation.

   (i) How much inbreeding accumulates in the population?

   (ii) If we consider a period of T generations that includes this bottleneck, what is the effective population size of the population? *(You can use the harmonic mean formula)*

   (iii) Suppose that the population, after this bottleneck, does not go up to size infinity immediately, but instead grows by a fraction r each generation, so that k generations later it is size $100 \times (1+r)^k$. What is the effective population size over a period of T generations (using the harmonic mean formula)? *[You will need to sum a geometric series using* $1 + x + x^2 + x^3 + \cdots + x^n = (1 - x^{n+1})/(1 - x)$ *]*

   (iv) Using this result, if we consider that case and also a case where there is a bottleneck to size 100 for G generations (and then size returns immediately to infinity), is there some relationship between the growth rate r and G that would give the same effective population size (and hence the same amount of inbreeding accumulated)?

8. In the diploid Wright-Fisher model, find the expression for the probability of getting k *AA* and ℓ *Aa* individuals (out of N) in the next generation, given that there are currently i *AA* and j *Aa* individuals out of N. Prove from this, if you can, that formula (VI-10) is correct.

9. In a population of size N which has been produced according to a Wright-Fisher model, if we have k *A* alleles and $2N - k$ *a* alleles, what is the probability that in that same population there are i *AA*, j *Aa*, and $N - i - j$ *aa* genotypes?

10. Show that if, in the notation of section VI.4, if we define the mean number of *A* alleles as

$$\bar{n}_A \;=\; \sum_{k=0}^{2N} k\, p_k^{(t)}$$

that $\bar{n}_A$ does not change with t in the Wright-Fisher model.

11. If we have a 3-allele Wright-Fisher model with the initial frequencies of the 3 alleles being q_1, q_2, and q_3, what are the probabilities of fixation of these three alleles? What is the equation corresponding to (VI-10) in the three allele model?

12. Why didn't we try $N = 1$ in Table 6.2?

13. Suppose that we have a simple Wright-Fisher model with initial frequency of A being p_0. Consider the following three assertions:

    (i) In the first generation (the offspring of the initial generation), the expectation of the genotype frequency of AA is p_0^2.

    (ii) In this first generation, the inbreeding coefficient is $f_1 = 1/(2N)$.

    (iii) In this first generation, the expectation of the genotype frequency of AA is $p_0^2(1 - f_1) + p_0 f_1$.
    Are these consistent? If not, where is the fallacy? Are we conditioning on something in one of these cases?

14. Check (VI-53) by direct computation in the case where in each generation, a randomly chosen individual has N offspring, and all the rest of the individuals in that generation have no offspring.

15. From (VI-53), what can we conclude about how much of the inbreeding in a finite population reproducing according to a Wright-Fisher model comes from random variation in offspring number, and how much comes from the random nature of Mendelian segregation? (Try abolishing one of these effects).

16. In a population in which population size varies randomly from generation to generation, but in which selfing is not allowed, where should we add the $1/2$ to correct for the absence of selfing – to each of the values of N or to the final effective population size?

17. What does it mean when H_{IS}, H_{IT}, or $H_{ST} > 1$?

18. In a case of a population, T, composed of two subpopulations of equal sizes, compute f_T as a function of the within- and between-subpopulation inbreeding coefficients f_w and f_b.

Chapter VII

GENETIC DRIFT AND OTHER EVOLUTIONARY FORCES

VII.1 Introduction

We have already seen the effects of genetic drift when it is the only evolutionary force acting. Its effects are to change gene frequencies in a random and unpredictable manner, resulting in fixation of one allele or another. The other evolutionary forces we have examined (natural selection, migration, and mutation) tend to change gene frequencies in a determinate way, or to push them towards an equilibrium value and hold them there. Genetic drift is the one force which can act as the "thermal noise" in the evolutionary machine. The relative strength of this "noise" compared to the nonrandom forces will determine to what extent the random effects of genetic drift will override other evolutionary forces. The general objective in this chapter will be to try to find simple rules indicating when each evolutionary force will prevail in the face of random genetic drift. (If physicists are listening, it is particularly important to say "random genetic drift" since in their subject "drift" is the name of a nonrandom force).

A subsidiary objective will be to introduce the mathematical technology for treating the interaction of random and systematic processes. This will be done by example, without more than a sketchy treatment of any but the simplest cases. The first evolutionary forces we will treat, mutation and migration, can be investigated in detail by considering only means and variances of gene frequencies (or alternatively, by considering probabilities of identity by descent). When we consider natural selection, this sort of treatment is no longer possible, and we must use the more complicated branching process and diffusion-equation methods.

It is worth reminding the reader that in section VI.4 we saw that there is no general formula for the probability, in a haploid Wright-Fisher model, that a population goes from having i copies of the A allele to having j copies in t generations. Starting with

an initial frequency i/N of the A allele, we cannot predict the distribution of possible t-generation outcomes exactly. This is at the root of the difficulty. Unable to solve exactly for the behavior of a Wright-Fisher model with no mutation, migration, or selection present, we have little hope of achieving exact solutions in the more complicated case when those forces are present. Thus we must rely on partial or approximate solutions. Fortunately, these are available, as they were when only genetic drift was present, and they are quite accurate approximations.

VII.2 Drift versus mutation

THE INFINITE-ALLELES MODEL. The simplest model we can make of the interaction of drift and mutation involves a Wright-Fisher model with N diploid individuals. To simplify things, we allow selfing at random. The model of mutation is different from that used in Chapter III. It is known as the "infinite-alleles" model, or sometimes as the "infinite-isoalleles model". Isoalleles, because all alleles are assumed to be selectively equivalent, there being no fitness differences. Infinite, because *each mutation is to a completely new allele*. There are thus an infinite number of possible alleles. The same allele never recurs twice in different mutations. Thus we need only know whether two different mutational events occurred in the ancestry of two gene copies to know whether they are different alleles. The model is intended as a rough approximation to what would be seen in a stretch of DNA sequence, where most mutations will be in different sites.

This model of mutation makes it particularly easy to work out the consequences of mutation and random genetic drift. We deal in this section only with that information which can be gleaned from means and variances, which fortunately is quite a lot. A more complete set of information can be gained by the diffusion method, as we will discuss later in this chapter. Equivalent to a consideration of means and variances of gene frequencies is a consideration of identity by descent. Suppose that we were to ask what was the probability F_t that two gene copies, randomly chosen without replacement from the same population, are identical by descent. If the occurrence of mutations had no effect on whether we counted genes as identical by descent, then the quantity F_t would follow the same course that it would in a Wright-Fisher model without mutation.

We could simply use (VI-4) to get

$$F_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_t, \tag{VII-1}$$

the usual Wright-Fisher model formula.

But now suppose instead that each time a mutant were to occur, the new allele was no longer counted as being identical by descent to any of the copies of the allele from which it arose. After, all it is our prerogative to define identity by descent any way we choose,

so we may as well try that. The consequences are straightforward. We have to modify (VII-1) as follows. That equation regards two genes as identical by descent provided that they either descended from the same gene copy in the preceding generation, or else came from distinct gene copies which were themselves identical by descent (these possibilities correspond to the two terms of (VII-1)). Now we must add an additional requirement: that neither of the two genes we choose can be a new mutant. We have ruled out the possibility that the new mutants could be identical by descent to any of the other genes in the population, or to each other. The occurrence of mutation is supposed to be random, unconnected with which parental gene is being copied. Each copy of the gene in generation $t+1$ has an equal and independent chance u of being a new mutant. It now follows directly that the right-hand side of (VII-1) must be multiplied by the probability that neither of the two gene copies is a new mutant, so that

$$F_{t+1} \;=\; (1-u)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_t\right]. \tag{VII-2}$$

With the occurrence of mutation, we now find a new behavior of the quantity F_t. It no longer automatically rises towards 1. In fact, it will always reach an intermediate equilibrium value between 0 and 1, provided that N is finite and u is not zero. We can solve for the equilibrium by realizing that at it $F_{t+1} = F_t = F$. Then we can remove the subscripts on F in (VII-2) and rearrange to get

$$F\left[1 - (1-u)^2\left(1 - \frac{1}{2N}\right)\right] \;=\; (1-u)^2 \frac{1}{2N} \tag{VII-3}$$

so that

$$F \;=\; \frac{(1-u)^2}{2N - (1-u)^2(2N-1)} \tag{VII-4}$$

We will be interested mostly in cases where u is very small (10^{-8} being a typical value for a single base, and 10^{-5} for a whole locus). So we can ignore terms in u^2 compared to those in u, so that we can replace $(1-u)^2$ by $1-2u$ to get the approximation (after a little rearrangement)

$$F \;\simeq\; \frac{1-2u}{1-2u+4Nu} \tag{VII-5}$$

The terms $2u$ will then be far smaller than 1, so that we can drop them to obtain the further approximation

$$F \;\simeq\; \frac{1}{1+4Nu} \tag{VII-6}$$

Table 7.1 gives some numerical comparisons of (VII-4) with the approximation (VII-6). Note how good the approximation is. Note also that the predominant feature of (VII-6) is a very good approximation: that the probability of identity by descent is a function of N and u only through their product, Nu. Thus the equilibrium identity by descent is maintained approximately the same by doubling N and halving u (or vice versa).

Table 7.1: Equilibrium probability of identity by descent for two distinct gene copies drawn from the same population in an infinite-alleles model. The exact values are from equation VII-4 and the approximate values from equation VII-6.

	$N =$	100	1000	
u = 0.001	Exact:	0.71398	Exact:	0.19975
	Approx.:	0.71429	Approx.:	0.20
0.0001	Exact:	0.96153	Exact:	0.71426
	Approx.:	0.96154	Approx.:	0.71429

FINITE NUMBERS OF ALLELES. So far our model of mutation has been the infinite-alleles model. It might seem that the fact that an infinite number of alleles are possible allows a much greater amount of variability to exist than if there were only a few possible alleles. This is not so. Suppose that there were only K different alleles, and mutation among these is symmetric. If a mutation carries a gene from one allelic state to one of the others, chosen at random, then we can easily modify the argument to allow this model. In equation (VII-2) the term $(1-u)^2$ is the probability that neither gene is a new mutant, for if so they cannot be identical. But that is only true under the infinite-alleles model. Under the K-alleles model, when one of the alleles mutates it cannot end up being the same allele as the other, but when both mutate this is possible. So the $(1-u)^2$ becomes $(1-u)^2 + u^2/(K-1)$. But an additional term also arises, for now two genes which were different alleles before mutation can be the same after mutation. This yields the recurrence relation

$$\begin{aligned} F_{t+1} &= \left[(1-u)^2 + u^2/(K-1)\right] \left[\tfrac{1}{2N} + (1-\tfrac{1}{2N})F_t\right] \\ &\quad + \left[2u(1-u)/(K-1) + u^2(K-2)/(K-1)^2\right] \left[(1-\tfrac{1}{2N})(1-F_t)\right] \end{aligned} \tag{VII-7}$$

(I leave it to the reader to work out all the terms). If we proceed a bit roughly and drop terms in u^2, and in u/N, this boils down to the approximation

$$F_{t+1} \simeq \frac{1}{2N} + \frac{2u}{K-1} + F_t\left[1 - \frac{1}{2N} - 2u\left(\frac{K}{K-1}\right)\right] \tag{VII-8}$$

which gives the equilibrium solution on setting $F_{t+1} = F_t$ and solving to obtain

$$F = \frac{1 + 4Nu/(K-1)}{1 + 4NuK/(K-1)}. \tag{VII-9}$$

Table 7.2: Equilibrium homozygosities for finite alleles models with different numbers of alleles

			K		
$4Nu$	2	4	8	16	∞
0.1	0.9167	0.9118	0.9103	0.9096	0.9091
0.5	0.75	0.70	0.6818	0.6739	0.6667
1.0	0.667	0.5714	0.5333	0.5161	0.5000
10.0	0.5238	0.3023	0.1954	0.1429	0.0909

Of course, a more accurate but far more complicated formula can be produced directly from (VII-7) by including all terms.

When u is small (VII-8) is quite a good approximation, and we can use it to investigate the effect of the number of possible alleles, K. Note that when K is large the term $1/(K-1)$ in the numerator will be very small, and the term $K/(K-1)$ in the denominator is nearly 1, so that (VII-8) must approach (VII-6) as K becomes large. Note also that as $4Nu$ becomes large the numerator is nearly $1/K$ the denominator. This is as it should be, for in this case the two genes we draw from the population will have quite unrelated mutational histories, with many intervening mutations. The chance that they end up being the same allele is thus $1/K$, which is what it would be if each represents one of the K alleles drawn at random and drawn totally independently of each other.

The qualitative rules for when mutation will maintain variability in the face of genetic drift are hardly affected at all by having only K possible alleles. Here are some values of F from (VII-8): Note that when K is of even moderate size the probability of homozygosity is nearly unaffected by increases in K. A more careful consideration of the Table and of (VII-8) will also show that the conclusion stated above is still valid when K is finite. We can still intuit the behavior of the selectively neutral alleles model in terms of the numbers of new mutants per generation being greater than or less than one.

THE ELECTROPHORETIC LADDER. In the past it was difficult to distinguish alleles. Using protein electrophoresis, alleles that did not differ in the charge of the protein were indistinguishable. Ohta and Kimura (1973) have investigated the effect of this "electrophoretic ladder" on the number of distinguishable alleles. With the availability of population samples of DNA sequences, all alleles can be distinguished. We will discuss further models and methods of analysis for these data in Chapter X.

RATE OF SUBSTITUTION OF ALLELES. So far our discussion of the interaction of mutation and genetic drift has dealt only with equilibrium conditions. With the infinite isoalleles model, the equilibrium of the quantity F does not represent a true equilibrium

of allele frequencies. There is constant turnover of alleles, as new mutations replace pre-existing ones. It is clearly of interest to know how rapidly this turnover occurs. This is a particularly relevant question to protein evolution, since Kimura (1968a) and King and Jukes (1969) have proposed that the bulk of evolutionary amino acid changes in proteins result from the substitution by genetic drift of selectively neutral mutants.

The computation is surprisingly simple. There are two general ways of establishing the same result. The first is a prospective argument. In the current generation we expect there to be $2Nu$ neutral mutants occurring (provided u counts the rate of only the selectively neutral mutations). How many of these are destined to be substituted for the existing alleles? It should be kept in mind that what we are interested in is whether the particular amino acid or nucleotide substitution becomes incorporated into the whole population. Further mutants will occur, so that a given mutant allele may never reach a frequency of 100%. The question we seek to answer is: will the whole population become *descended from* this particular mutant? If so, then the DNA will show that the population has undergone a substitution at that site. Thus we must ignore further mutations when asking whether a given mutant becomes "fixed".

This allows us to take our result directly from the discussion surrounding equation (VI-19) above. The $2N$ copies of the gene in question each has an equal probability of fixation. So each mutant has a probability $1/(2N)$ that it will be the progenitor of future populations. The expected number of mutants arising in the current generation which will substitute throughout the population is therefore $2Nu \times 1/(2N) = u$. So *the rate of substitution of neutral mutations is equal to the neutral mutation rate per haploid genome*.

The reasonableness of this result will be more apparent if we consider the other way of obtaining it: retrospectively. Consider a gene in the current population. Ask how many mutations have occurred since T generations ago. Following the line of ancestry back from a current gene copy, we find that in each generation there was only one gene copy directly ancestral to that gene. Thus there was exactly one opportunity in each past generation for mutation to occur that would affect this particular gene copy. So the average number of mutations affecting this gene will be u in each generation. Thus we can think of the rate of neutral mutant substitution as being u because of an exact cancellation of two effects of population size, one increasing the number of mutants occurring and the other decreasing their chances of fixing, or alternatively we could see the absence of a population size effect as the consequence of there being only one ancestral copy for each gene copy in a population.

A convenient way of seeing what this means in any given case is to notice that the expected number of new mutations each generation is $2Nu$ (as there are $2N$ genes and a fraction u of these are expected to mutate). So the effective number of alleles at equilibrium depends mostly on the number of new mutants arising each generation. This number expresses the balance of forces between genetic drift and mutation, and makes it clear that two populations with different sizes and mutation rates may nevertheless

be expected to have the same amounts of variability, provided that they have the same expected number of new mutants per generation. We may state a qualitative conclusion:

> Substantial genetic variability will be maintained in a population by mutation provided $2Nu \gg 1$, that is, provided there is more than one new mutant at the locus per generation.

It is important to understand what (VII-6) does and does not mean. It gives average homozygosities under genetic drift and mutation, but these are only expectations. If we draw individuals repeatedly from the same population, we will not necessarily obtain the expected proportion of homozygosity. Any one population may go through periods when all but one allele have been lost, and periods when a spate of recent mutations have drifted to high frequencies, leaving it very polymorphic. As we follow the population through time, F will vary above and below expectation, averaging out to the value given in (VII-6). By the same token, if we examined a series of populations simultaneously, where each was isolated from its neighbors and none exchanged migrants, then we would find the gene frequencies and the homozygosity F to vary from population to population, averaging out to its expectation. Thus (VII-6) gives an average over time (once the initial conditions are lost) and also an average over replicate populations.

RESPONSE TO POPULATION SIZE BOTTLENECKS. Another aspect of the time-dependent behavior of a neutral mutation model which may be of interest is how rapidly the level of polymorphism responds to changes in population size. Suppose that we have a population of large size which has been at that size for a very long time, so that $4Nu$ is large. If we reduce population size, how rapidly will variability be lost? Suppose F_t is small, and we have just reduced population size so that N is now small. By equation (VII-2), in the next generation

$$F_{t+1} \simeq (1-u)^2 \frac{1}{2N} \simeq \frac{1}{2N} \tag{VII-10}$$

so that if N is small F_{t+1} may increase substantially in one generation fairly quickly (i.e. if $N = 10$ it will increase by about 0.05 per generation). Once F_t comes to its new equilibrium at a large value, little variability is present. Suppose now that the population size grows back to its old value, so that now $4Nu$ is large again. How rapidly will F_t drop back down to its old value? From (VII-2), if $F_t \simeq 1$, $F_{t+1} \simeq (1-u)^2$ which will usually be very close to 1. Thus F is decreasing by only $2u$ per generation, so that it could take millions of generations to recover the variability. The effect of a bottleneck of population size is thus to rapidly reduce variability, but the rate of recovery after restoration of the population size is slow, as we must wait for new alleles to occur by mutation and to drift to high frequency.

This may seem biologically reasonable, but counterintuitive. The mathematics is, after all, simply that of equation (VII-2). This is a simple linear recursion of F. How can it approach its equilibrium more slowly on one side than on the other? This paradox is resolved if we note that the departure from the equilibrium value of (VII-2), is multiplied each generation by

$$(1-u)^2\left(1-\frac{1}{2N}\right).$$

When N is small, this is dominated by the terms in N, and change is rapid. But when N is large enough that a substantial amount of variability is expected at equilibrium, $2u > 1/(2N)$, and the mutational factor is the dominant one. The reason approach to equilibrium from the two sides differs so much in rate is that N is much different in the two cases. We have compared apples to oranges.

Nei, Maruyama, and Chakraborty (1975) have presented computations for the effects of bottlenecks of population size. These are particularly relevant to human populations, which are now much larger than they were only 100 generations ago.

REFERENCES. The equations for the neutral mutation model were first worked out by Malécot (1948), although apparently without any idea of presenting this process as a model of the maintenance of protein variation in natural populations. Kimura and Crow (1964) obtained similar results independently, as a byproduct of an investigation into the effects of genetic drift on the maintenance of multiple alleles in a population by natural selection. Kimura (1968b) presented the analysis of the effects of assuming a maximum of K alleles. The rate of substitution of neutral mutants was first given by Kimura (1968a) when he famously proposed neutral mutation as the main source of both protein polymorphisms in natural population and the reconstructed rates of amino substitution through evolutionary time. Lewontin and Hubby (1966) had earlier discussed neutral isoalleles as a possible explanation of the electrophoretic variability they observed, in what was the first coherent description of the neutral mutation theory of polymorphism.

We have here been concerned only with a pure Wright-Fisher model. In most cases, more complex models of population reproduction (overlapping generations, varying population size, variation in offspring number, etc.) one can simply replace $4Nu$ by $4N_e u$ throughout the argument without difficulty. Chia and Pollak (1975) present a detailed discussion of varying population size which verifies that one can use the effective population size N_e if the population size does not vary greatly. As we shall see in chapter VIII, linkage is another matter. Linkage of one neutral locus to another is irrelevant, but the presence of a locus with naturally selected variation near a neutral locus can greatly reduce or greatly increase the effects of genetic drift.

At a functional locus which has mutation occurring, there will be a pool of deleterious mutations in the population. If the selection coefficient against these alleles is substantially larger than the mutation rate, this pool will be a small fraction of all copies. How different will these mutant alleles be from each other? If we make two assumptions: that the mutant alleles are all equally deleterious, and that the pool of mutant alleles is roughly constant in size, we can use the neutral mutation model.

We simply pretend that the pool of mutant alleles is a population. New mutant alleles are entering it, approximately $2Nu$ of them per generation. Within the pool the mutant alleles drift, and some are lost. If the pool of mutant copies is a fraction q of all copies at the locus, then there are Nq copies in the pool. The entry of new mutants from outside of the pool looks as if a fraction u/q of the $2Nq$ copies in the pool had mutated. Thus, considering the pool as if it were a population of Nq individuals, fraction of pairs of copies in it that are expected to come from the same mutational event is

$$F = \frac{1}{4Nq(u/q) + 1} = \frac{1}{4Nu + 1} \qquad \text{(VII-11)}$$

which is the predicted probability of homozygosity under neutral mutation at rate u in the full population. Of course, the rate of deleterious mutation u may be different from the rate of neutral mutation.

This result was obtained by Hartl and Campbell (1982). Robertson and Hill (1983) did the corresponding calculation for individually recessive alleles. Slatkin and Rannala (1997) showed that the sampling distributions used for neutral mutants also apply to a pool of equally-deleterious mutant alleles.

As an example, if $N = 10^7$ and $u = 10^{-7}$, we predict that two deleterious mutants will come from the same mutational event with probability $F = 1/5$. In a larger population, there will be even more heterogeneity among mutant alleles. Treating change in the pool of deleterious mutants as if it were neutral mutation is thus an illuminating approximation.

We will see later in this chapter that the assumption of approximate constancy of the size of the pool of mutant alleles is justified if $4Nu$ is substantially greater than 1. For the smaller value of population size in the above examples, this is dubious. So, of course, is the assumption that all deleterious mutants have equal fitnesses.

Box 2: Application: heterogeneity of deleterious mutations

VII.3 Genetic distance

Suppose that we observe two populations, one with gene frequencies $x_1, \ldots, x_k$ of k different alleles, the other with gene frequencies $y_1, \ldots, y_k$. We may wish to estimate F_{ST} for the two populations as a measure of genetic divergence, under the assumption that they have diverged from some initial gene frequencies $z_1, \ldots, z_k$, which we do not know, by a process of genetic drift in isolation. If they have, then we should be able to use the current gene frequencies to estimate the accumulated inbreeding. If we have a good estimate of the population sizes, then this will allow us to calculate the time since divergence of the populations. Measures which estimate F_{ST} or some function of it are called *genetic distances*. Their elaboration was a favorite sport of population geneticists in the early 1970's. The reader will find some discussion of various distance measures in the symposium edited by Crow and Denniston (1974). The literature of genetic distances is very tough sledding because scarcely anyone in it states clearly what problem they are designed to solve.

In fact, there seem to be two classes of genetic distances. One measures whether the heterozygosity between populations is substantially greater than that within populations. The other measures whether homozygosity within populations is substantially greater than that between populations. You might imagine that these are identical questions, but they are not. For example, if the homozygosity within populations is 0.10, and between populations it is 0.05, the homozygosity is twice as great within as between. But the heterozygosities are then respectively 0.90 and 0.95, so the heterozygosity between populations is only $0.95/0.90 = 1.0555$ times as great as within populations!

In fact, it is not obvious that every genetic distance must use heterozygosity or homozygosity – in general these are not "sufficient statistics" that contain all the relevant genetic information. Some of the genetic distances mentioned below will not be functions of the heterozygosity or the homozygosity. Nevertheless, they are useful quantities to examine.

CHANGES OF HETEROZYGOSITY AND HOMOZYGOSITY. If the expected heterozygosity and expected homozygosity within populations are respectively h_w and f_w, and those between populations are respectively h_b and f_b, we can ask about different evolutionary forces, and seek to make a measure of divergence time between two isolated populations. Initially the two population are the same, so that superscripting each with its generation number, $f_w^{(0)} = f_b^{(0)}$ and $h_w^{(0)} = h_b^{(0}$. With both genetic drift and an infinite isoallele model operating, we can find equations for two populations of equal effective size N_e: From equation (VII-2) the expectation of f_w will be:

$$f_w^{(t+1)} = 1 - h_w^{(t+1)} = (1-u)^2 \left[\frac{1}{2N_e} + \left(1 - \frac{1}{2N_e} \right) f_w^{(t)} \right] \qquad \text{(VII-12)}$$

By considering that any two genes randomly sampled from different populations were descended from two genes randomly sampled from those same populations one generation earlier, we see that they are just as likely to be the same allele as they were, provided that neither has mutated:

$$f_b^{(t+1)} \;=\; 1 - h_b^{(t+1)} \;=\; (1-u)^2 f_b^{(t)} \tag{VII-13}$$

Note that the derivation of equations (VII-12) and (VII-13) does not assume that the initial population is in any particular equilibrium state, or that the sizes of the two populations (both N_e) continue to be the same as they were before the populations split. This approach can thus be used for populations that may have recently split into small isolates, where we may be able to ignore mutation.

DIVERGENCE BY GENETIC DRIFT ONLY. There are two cases of particular interest. Suppose that there is no mutation ($u = 0$). Then the equations reduce to

$$h_w^{(t+1)} \;=\; \left(1 - \frac{1}{2N_e}\right) h_w^{(t)} \tag{VII-14}$$

$$h_b^{(t+1)} \;=\; h_b^{(t)} \tag{VII-15}$$

In that case, if the two populations start out identical to each other, so that $h_b^{(0)} = h_w^{(0)}$ we find that since $h_b^{(t)} = h_w^{(0)}$,

$$h_w^{(t)} \;=\; \left(1 - \frac{1}{2N_e}\right)^t h_b^{(t)} \tag{VII-16}$$

A measure of how much greater heterozygosity is between than within is then

$$\left(h_b^{(t)} - h_w^{(t)}\right) / h_b^{(t)} \;=\; 1 - h_w^{(t)} / h_b^{(t)} \;=\; 1 - \left(1 - \frac{1}{2N_e}\right)^t \tag{VII-17}$$

It increases by an amount $t/(2N_e)$ per generation at first. Thus if we could somehow measure $1 - h_w^{(t)}/h_b^{(t)}$ we could solve for $\frac{t}{2N_e}$ to obtain a measure of genetic distance between isolated populations that would for a while increase linearly with time. Note that (VI-65) shows that this measure of genetic distance simply computes F_{ST}.

On might think that this makes finding good measure of genetic distance is easy: simply find the average heterozygosity within populations for the finite set of loci we are observing:

$$H_w \;=\; 1 - \frac{1}{2}\left(\sum_i x_i^2 + \sum_i y_i^2\right) \tag{VII-18}$$

and the average heterozygosity between populations

$$H_b \;=\; 1 - \sum_i x_i y_i, \tag{VII-19}$$

assume that these estimate $h_w^{(t)}$ and $h_b^{(t)}$, and then compute the genetic distance between populations as

$$D \;=\; 1 - H_w / H_b. \tag{VII-20}$$

This seems straightforward, but some questions arise. The quantities h_w and h_b are not the heterozygosities within and between populations, they are the *expected* heterozygosities. Even if we assume that the infinite isoallele model is exactly correct, we need to use many loci, should we average the values of H_w and H_b, *then* compute D, or should we average the values of D for each locus? Also, if we have population samples of different sizes at each locus, how should we weight these? Once these questions are entertained, the door is opened to an endless array of genetic distance measures. We will not attempt to untangle this literature here.

DIVERGENCE BY DRIFT AND MUTATION. The second major approach to genetic distance is due to Nei (1972). It allows for the effects of mutation, but at the cost of having to assume that N_e has been the same for a long time in the progenitor of the two populations as it is in them. It yields a result that is mostly useful when the neutral mutation rate is the same at all loci. If we assume that the base population had a level of homozygosity $f_w^{(0)}$ that is at equilibrium under neutral mutation, then (VII-12) reduces to

$$f_w^{(t+1)} \;=\; f_w^{(t)} \;=\; \ldots \;=\; f_w^{(0)} \tag{VII-21}$$

Combining this with (VII-13) we find that

$$f_b^{(t)} \;=\; (1-u)^2 f_b^{(t-1)} \;=\; (1-u)^{2t} f_b^{(0)} \tag{VII-22}$$

and since $f_w^{(0)} = f_b^{(0)}$ and since $f_w^{(t)}$ does not change through time, we also get that

$$f_b^{(t)} \;=\; (1-u)^{2t} f_w^{(0)} \;=\; (1-u)^{2t} f_w^{(t)}, \tag{VII-23}$$

so that, taking logarithms

$$-\ln\left(\frac{f_b^{(t)}}{f_w^{(t)}}\right) \;=\; -2t\;\ln(1-u) \tag{VII-24}$$

Since u is usually quite small, $-\ln(1-u) \;\simeq\; u$ so that

$$-\ln\left(\frac{f_b}{f_w}\right) \;\simeq\; 2u\,t \tag{VII-25}$$

Thus one approach to genetic distance measures the difference in heterozygosity between populations, the other the difference between homozygosity between them. The first is expected to cope well with differences in size between the base population and the two populations that originate from it. But it does not allow for mutation at the loci. The second approach (Nei's approach) allows for mutation, but assumes that mutation rates are the same at different loci, and that population sizes have not changed from the ancestral population sizes since the populations diverged.

SOME WIDELY-USED MEASURES. A practical genetic distance of the first sort (though not using heterozygosities or homozygosities) is that due to Cavalli-Sforza and Edwards (1967):

$$D_{CSE} \;=\; 4\left[1 - \sum_i \sqrt{x_i y_i}\right]. \qquad \text{(VII-26)}$$

The prescribed method of combining results at different loci is to average the values of D.

Nei's distance measure (Nei, 1972) is in practice:

$$D_N \;=\; -\ln\left[\sum_i x_i y_i \Big/ \left(\left(\sum_i x_i^2\right)^{1/2}\left(\sum_i y_i^2\right)^{1/2}\right)\right] \qquad \text{(VII-27)}$$

The prescribed method of combining results at different loci is to average the values of $\sum x_i y_i$, $\sum x_i^2$, and $\sum y_i^2$.

The behavior of D_{CSE} has been intensively investigated by Heuch (1975). Nei's distance is discussed in his books (1975 and 1987). I have also (Felsenstein, 1985) investigated the behavior of a number of genetic distances under divergence by genetic drift. It is worth pointing out that each of these measures is not expected to perform well when the situation is that appropriate for the other one.

It is well when reading the genetic distance literature to keep the following points in mind:

1. *All genetic distance measures which are derived with the intention of measuring the inbreeding due to genetic drift are roughly proportional to each other when gene frequency differences between populations are small.* In this case it does not matter much which one you use if you want to know whether populations A and B are much more different in gene frequencies than are A and C.

2. *All seem to have various problems when gene frequency differences are large.* I have investigated numerically (Felsenstein, 1985) the way they break down as gene frequencies become large.

3. *Some genetic distance measures are actually measuring other quantities than F_{ST}: for example measures of the probability that two genotypes chosen at random in the populations*

are the same. It seems sufficient that a formula be zero when all the x_i equal the y_i, and positive otherwise, that it be called a "genetic distance". In this sense, you too can have your own genetic distance measure:

$$D_U \;=\; \left[\sum_i (x_i - y_i)^{2H}\right]^{1/(2H)}, \qquad \text{(VII-28)}$$

where H is the number of Hairs in your nose, or if you don't know that, your House number (rounded to the nearest integer), or the last three digits of your Home telephone number.

4. *All genetic distance measures which estimate F_{ST} do so under the implied assumption that all genetic change is due to random genetic drift.* If the gene frequencies in the two populations are diverging due to natural selection, or are being held at constant values by balancing selection, the genetic distance measure ceases to be related to divergence time of the population. It may be taken to be an empirical measure of "genetic distance", but if so, care should be taken to make clear to the reader of the resulting paper why the distance measure is in any way preferable to formula (VII-28) above.

There is much more to say about genetic distances – we have only scratched the surface here. In the Problems/Complements at the end of this chapter you will find some questions about the behavior of these two classes of genetic distance measures when their assumptions are violated. But it is not clear that going deeply into the properties of genetic distance measures is worthwhile, since there are now more powerful methods of inferring population parameters that do not use them, as we will see in chapter X.

VII.4 Drift versus migration.

A ONE-ISLAND MODEL. As there is a balance between mutation and genetic drift, so also is there a balance between migration and genetic drift. We can investigate this most simply in a one-island model. We have an island with N diploid individuals reproducing according to a simple Wright-Fisher model. Nearby lies a continent which has a constant gene frequency $\bar{p}$. Immigrants from the continent affect the island gene frequency, but the continent is too large for the emigrants from the island to alter its gene frequency. We will assume that the immigrants arrive as gametes (this is biologically dubious but mathematically convenient). Since the island follows a Wright-Fisher model, we assume that there is an infinite pool of gametes before the density-dependent death of all but N individuals. A fraction m of the gametes are replaced by immigrants. So if p_t is the frequency of allele A among the gametes before immigration, afterwards it is

$$p_t^* \;=\; (1-m)\; p_t + m\; \bar{p}. \qquad \text{(VII-29)}$$

The gametes combine at random to form diploid individuals, and all but N of these die during the density-dependent mortality on the island before adulthood. The gametes in the pre-immigration pool of the next generation are contributed equally by these surviving adults, so that p_{t+1} will be the same as the gene frequency in the adults of generation t. This will be the result of binomial sampling of $2N$ genes from a pool with gene frequency p_t^*. The expected value of the post-sampling gene frequency is just p_t^* , so that from (VII-29),

$$\mathbb{E}(p_{t+1}) \;=\; (1-m)\,\mathbb{E}(p_t) \;+\; m\,\bar{p} \qquad \text{(VII-30)}$$

(since we can take the constant $(1-m)$ outside of the expectation, and $\bar{p}$ is itself also a constant and therefore has expectation $\bar{p}$). We will concentrate on the "stationary state" in which all of this has gone on for long enough that the effects of the initial frequency p_0 on the island have been lost. In the stationary state we can assume that $\mathbb{E}(p_{t+1}) = \mathbb{E}(p)$. From (VII-30) we then have

$$\mathbb{E}(p) \;=\; \bar{p}. \qquad \text{(VII-31)}$$

The expected gene frequency on the island is thus the same as on the continent. This is a reasonably intuitive result (if you think about it). Most genes on the island came from the continent. If we trace back the ancestry of a gene, at each stage there is probability m that the ancestor was on the continent. Sooner or later every gene on the island turns out to be of continental origin. So it cannot be surprising that the expected gene frequency on the island is the continental gene frequency.

VARIATION OF GENE FREQUENCY. Of course the island will not be exactly at gene frequency $\bar{p}$. This is only the *expected* gene frequency. Genetic drift will continually move the island gene frequency away from its current value. Migration from the mainland will continually pull the island gene frequency back towards $\bar{p}$ by diluting out the island genes with mainland genes. The interesting question is: how far will the island genes be from $\bar{p}$? This can be addressed by asking about the variance of the island gene frequency around its expected value $\bar{p}$. The easiest way to investigate this seems to be to look at the deviation of each population's gene frequency from $\bar{p}$. Let $x = p - \bar{p}$ be the deviation of a population from the expected gene frequency $\bar{p}$. Since we obtain the x's by subtracting a constant ($\bar{p}$) from the p's, the variance of the p's will be the same as the variance of the x's. Note that when we talk of the variance we are, as in the previous section, discussing the variance among independent replicate populations each undergoing the same process, or else the variance in the gene frequency of a single population through time.

From (VII-29) we find that after migration but before genetic drift

$$\bar{p} + x_{t+1}^* \;=\; (1-m)\,(\bar{p} + x_{t+1}) \;+\; m\bar{p} \qquad \text{(VII-32)}$$

so that

$$x_{t+1}^* \;=\; (1-m)\,x_{t+1}. \qquad \text{(VII-33)}$$

and therefore after genetic drift

$$x_{t+1} \;=\; (1-m)\, x_t^* \;+\; \varepsilon_t \tag{VII-34}$$

where ε_t is the change in gene frequency caused by random sampling of $2N$ gametes from a population whose gene frequency is x_t^*. As in we saw when we discussed the expectations, it is easy to demonstrate that ε_t has expectation zero, and also that it is uncorrelated with x_t^*: knowing that x_t^* in a particular population is positive tells us nothing about whether ε_t^* will be positive. All of which is by way of hand-waving our way to the following:

$$\mathbb{E}({x_{t+1}}^2) \;=\; (1-m)^2\, \mathbb{E}({x_t^*}^2) + 2(1-m)\mathbb{E}(x_t^*\, \varepsilon_t) + \mathbb{E}(\varepsilon_t^2) \;=\; (1-m)^2\, \mathbb{E}({x_t^*}^2) \;+\; \mathbb{E}(\varepsilon_t^2), \tag{VII-35}$$

since the cross-product $\mathbb{E}(x_t^*\varepsilon_t)$ will be zero if ε_t is uncorrelated with x_t^*, as we claim. Notice that $\mathbb{E}(\varepsilon_t^2)$ in a population is $p_t^*(1-p_t^*)/2N$, the binomial sampling variance based on a current gene frequency of p_t^*. When the expectation is taken over all replicates (which may have different values of p_t^*) throughout equation (VII-35) we get

$$\mathbb{E}(x_{t+1}^2) \;\;=\;\; (1-m)^2\, \mathbb{E}({x_t}^2) \;+\; \mathbb{E}[(\bar{p}+x_t^*)(1-\bar{p}-x_t^*)/2N]. \tag{VII-36}$$

Note that since x_t^* has expectation zero by (VII-31), $\mathbb{E}({x_t^*}^2)$ is simply the variance of x_t^* and therefore also the variance of p_t^*. This is the the variance (over all replicates) of the gene frequency immediately after migration, and before the adult stage of the life cycle. This will be $(1-m)^2 V_t$. From (VII-36), making use of the fact that x_t^* has expectation zero, we get

$$V_{t+1} \;\;=\;\; (1-m)^2\; V_t \;+\; [\bar{p}(1-\bar{p})/2N - (1-m)^2 V_t/2N]. \tag{VII-37}$$

The rest is straightforward: we are interested in the variance of the adult gene frequencies p_t when a stationary state is reached (i.e. when the initial conditions have become of no importance, when the mean and variance of p_t have reached equilibrium values, although individual population gene frequencies continually vary and do not reach equilibrium). This we get by solving (VII-37) for $V_{t+1} = V_t = V$:

$$V\,(1-(1-m)^2(1-1/(2N)) \;\;=\;\; \bar{p}(1-\bar{p})/2N \tag{VII-38}$$

so that with some rearranging

$$V \;\;=\;\; \frac{\bar{p}(1-\bar{p})}{2N-(2N-1)(1-m)^2}. \tag{VII-39}$$

Note that (VII-39) checks with intuition in those cases where we know the answer. When $m = 0$, so that only genetic drift is operating, $V = \bar{p}(1-\bar{p})$, which is the variance among

a set of populations each of which has probability $\bar{p}$ of being fixed for A. (This is not quite kosher since, when $m = 0$, then $\bar{p}$ is not the expectation of the p_t, but let that pass). When $m = 1$, so that every generation the island consists of only immigrants and then we observe these immediately after genetic drift occurs, $V = \bar{p}(1-\bar{p})/(2N)$, which is simply the binomial variance after one generation of genetic drift starting from the continental gene frequency,

When m is small and N is large, we can simplify (VII-39). When terms in m^2 and m/N are ignored,

$$V \approx \bar{p}(1-\bar{p})/(4Nm+1). \qquad \text{(VII-40)}$$

Note the term $4Nm$, which looks suspiciously like the $4Nu$ which appeared in our discussion of the infinite isoallele model. Now we are ready to answer the question as to when immigration will override the effects of genetic drift. When there is no immigration $4Nm = 0$ and we find that the variance of gene frequencies among different realizations of this process will be $\bar{p}(1-\bar{p})$, indicating that all islands will be fixed, some for A, some having lost A. Only when $4Nm$ is not much smaller than 1 will immigration from the mainland pull the island gene frequency closer to the mainland gene frequency $\bar{p}$. When $4Nm+1$ is large, there will be little variation of the island gene frequency around the continental value. Note that $2Nm$ is the expected number of genes among the adult survivors of population size regulation which are immigrants, so that (although it is gametes which migrate in this particular model), the level of immigration is roughly equivalent to Nm individuals per generation.

We want a crude rule to serve as a rough guide, so that we will regard the immigration rate at which $4Nm = 1$ as being close to the rate at which $Nm = 1$, so that we can state another principle:

> Migration will have a substantial effect in counteracting the effects of local random genetic drift provided that there is one or more immigrant individual into the population each generation.

Like the principle stated in the last section for mutation, this is only a rough guide, but is surprisingly useful in practice.

VII.5 Drift vs. Migration: the Island Model

THE MODEL. may recall the discussion in Chapter IV of different models of migration. In our discussion of the interaction between genetic drift and migration, we have so far used only the simple one-island model. Now we want to look at the n-island model. One limitation of the one-island model was that the genetic drift on the island caused gene frequency changes, but these were not exported back to the mainland. Now suppose that there are n islands, each reproducing according to a simple diploid Wright-Fisher

model with population size N. Let us introduce migration by assuming that, in the gamete stage of the life cycle, a fraction m of the gametes in each population is removed and replaced by gametes randomly and independently sampled from the other $n-1$ populations. Mating follows the migration. We allow mutation to occur before mating, according to the infinite isoalleles model. Thus the model of reproduction is:

$$\underset{(N)}{\text{Adults}} \xrightarrow{\textit{meiosis}} \underset{(\infty)}{\text{Gametes}} \xrightarrow{\textit{migration}} \underset{(\infty)}{\text{Gametes}} \xrightarrow{\textit{mutation}} \underset{(\infty)}{\text{Gametes}} \xrightarrow{\textit{mating}} \underset{(N)}{\text{Adults}}$$

Why are we bothering to look at this case? We have a group of populations, in each of which genetic drift is being counteracted by mutation. If there were no migration, each population would come to contain different alleles, in a balance between mutation and drift. But migration will spread the same alleles into different population. It will increase the number of alleles present in any one population, but at the same time will make the populations more similar. At what point, at what amount of migration will the set of n populations begin to behave like a single large population of Nn individuals? This question has no particular meaning in the one-island model. The n-island model is the simplest in which we can investigate it.

Note that if the number of islands is infinite ($n = \infty$), the calculations of the previous section apply. The average gene frequency of all islands then stays precisely constant at the initial frequency. We can call this $\bar{p}$. Equation (VII-29) and all the other equations of that section apply. Drawing a random gene from the archipelago is the same as drawing it from a continent whose gene frequency never changes.

To analyze the case where n is not infinite, we make use of two quantities, F_W and F_B. These are the probabilities of identity of two genes drawn at random (respectively) from the same population (F_W) and from different populations.

Qian Sophia Zhang has suggested a derivation in terms of the probability that two gene copies that are in the same population are descended from copies that were in the same population in the previous generation. Let's call this P_w. We can also make a similar computation for two copies that were in different populations – the probability P_B that the copies ancestral to them were in the same population. We will compute these shortly. Before we do, note that when two copies are independently drawn from the same population, their probability of having identical alleles, if there has been no mutation yet, is

$$Q = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_w \qquad \text{(VII-41)}$$

This can be seen since there is a chance $1/(2N)$ that the two copies are drawn from the same ancestral copy, in which case they must have the same allele; if they are drawn from different copies, the probability that they have the same allele is F_B.

For two copies whose immediate ancestor copies were in different populations, the probability that they have the same allele is simply F_B. Putting all this together, the probability that we find the same allele in copies from the same population, and in

copies from different populations can be written simply as:

$$\begin{aligned} F_B^* &= P_w\, Q \;+\; (1 - P_w)\; F_W \\ F_W^* &= P_b\, Q \;+\; (1 - P_b)\; F_W \end{aligned} \tag{VII-42}$$

The asterisked quantities F_W^* and F_B^* are the probabilities of two copies having the same allele in this offspring population, given that they have not yet mutated.

Let us use an infinite-alleles model of mutation. In that model, two gene copies in the offspring can have the same allele only if neither has mutated since the previous generation. Thus, after the stage at which mutation occurs, the probabilities of having the same allele are:

$$\begin{aligned} F_B' &= (1-u)^2\,(P_w\, Q \;+\; (1 - P_w)\; F_W) \\ F_W' &= (1-u)^2\,(P_b\, Q \;+\; (1 - P_b)\; F_W) \end{aligned} \tag{VII-43}$$

We still need to obtain the probabilities P_w and P_b. We already have Q.

EQUILIBRIUM SOLUTION. For the moment, we're interested in the equilibrium state in which $F_W' = F_W$ and $F_B' = F_B$. We have two equations in two unknowns, and these can be solved to obtain explicit formulas for F_W and F_B in terms of m, u, n, and N. The results are rather complicated formulas which are not easy to look at. Instead, let us approximate, considering only those cases in which m, u, and $1/N$ are small. This will quite often be biologically realistic. The squares and products of these small quantities, m^2, u^2, mu, m/N, and u/N, can be ignored as compared to m, u, and $1/N$ themselves. When this is done we get the considerably simpler-looking approximate equations (dropping primes since we assume that the quantities F_W and F_B are at equilibrium)

$$\begin{aligned} F_W &= \tfrac{1}{2N} \;+\; (1 - 2u - 2m - \tfrac{1}{2N})F_W \;+\; 2m\, F_B \\ F_B &= [1 - 2u - 2m/(n-1)]\, F_B \;+\; [2m/(n-1)]\, F_W. \end{aligned} \tag{VII-44}$$

The second equation immediately yields (dividing every term by F_W and solving for the ratio F_B/F_W)

$$\frac{F_B}{F_W} \;=\; \frac{m/(n-1)}{u + m/(n-1)}. \tag{VII-45}$$

The quantity F_B/F_W is a measure of how similar genes are between populations, as compared with how similar they are within populations. We call it ρ. It is closely related to Nei's measure of genetic distance, which would be $-\ln\rho$ if enough loci were able to be measured to determine ρ accurately. If genes from different populations are totally unrelated, then $F_B = 0$ so that $\rho = 0$. When the genetic contents of the different islands are so thoroughly similar that we are no more likely to find the same allele simply by virtue of looking in the same population, then $F_W = F_B$ so that $\rho = 1$.

The first result we obtain from the analysis is that this measure of the similarity of population *is independent of N* (to good approximation). It reflects only the balance between the rate at which alleles are exported into another population (governed by m) and the rate at which they become new alleles (governed by u). Note that the quantity $m/(n-1)$ is the immigration rate from one population to one particular other population.

The second result follows easily from the first equation of (VII-44). We notice from (VII-45) that F_B is by definition equal to $F_W\rho$. Making this substitution in the first equation of (VII-44), we get a simple linear equation for F_W so that

$$F_W = \frac{1}{1+4Nu+4Nm(1-\rho)}. \qquad \text{(VII-46)}$$

Now this has a very familiar look to it. It is essentially a modified version of the equation for F in the one-population isoallele model, (VII-6). The relationship between these two equations has a reasonably straightforward verbal interpretation. We simply have to recognize that in the n-island model new alleles are introduced into a population by two routes. One is mutation, an event which occurs with probability u and also occurs in the one-population model. The second route is by migration, which brings in alleles which may or may not be different. Note that ρ is a measure of how much different the alleles in the other populations are. Since $F_B = \rho F_W$, it is as if the immigrant alleles, whose probability of identity with a random resident is F_B, were a mixture of residents and new mutations. If they were such a mixture, with ρ being the proportion of them which are not mutants, then we would have $\rho\, F_W + (1-\rho)\, 0 = \rho\, F_W$ as their probability of identity with the residents. All of which is by way of justifying the assertion that when immigration occurs, it brings in new alleles at a rate which is equivalent to a mutation rate of $m(1-\rho)$. So the presence of migration has the effect of increasing the effective mutation rate from u to $u + m(1-\rho)$. When we make that replacement in equation (VII-6), we get precisely the equation for the present model, (VII-47). Note that ρ itself depends on m and u.

A NUMERICAL EXAMPLE. Equations (VII-45) and (VII-46) provide a good approximation when m, u, and $1/N$ are all small. Table 7.3 shows a series of values of F_W and ρ obtained from the numerical solution of the exact equations (VII-44) for the equilibrium values. The Table is organized with reference to the case where $N = 200$, $n = 10$, $m = 0.001$, and $u = 10^{-4}$. The different parts of the Table show the effects of varying each of these parameters. The numbers in parentheses are the results of the approximate formulae (VII-45) and (VII-46), presented to show that they are reasonably close to the exact solutions. Let us consider the effects of varying each parameter. When we increase N, the population size of an island, we increase the genetic variability maintained in a single population, which is to be expected since we are weakening the effect of genetic

Table 7.3: Effects of varying different parameters in an n-island infinite isoallele model when the parameters are (unless otherwise specified) $N = 200$, $n = 10$, $m = 0.001$, $u = 10^{-4}$. The values in parentheses are those computed from the approximations (VII-45) and (VII-46), presented to show their level of accuracy.

Changes in N:	N	F_W		ρ	
	10	0.9775	(0.9776)	0.5267	(0.5263)
	100	0.8132		0.5267	
	200	0.6852		0.5267	
	500	0.4655		0.5267	
	1000	0.3033		0.5267	
	2000	0.1788	(0.1789)	0.5267	(0.5263)
Changes in n:	n	F_W		ρ	
	2	0.8675	(0.8675)	0.9093	(0.9090)
	5	0.7641		0.7147	
	10	0.6852		0.5267	
	20	0.6231		0.3452	
	50	0.5729		0.1697	
	100	0.5532		0.0919	
	1000	0.5338	(0.5342)	0.0099	(0.0099)
Changes in u:	u	F_W		ρ	
	10^{-6}	0.9921	(0.9921)	0.9911	(0.9911)
	10^{-5}	0.9310		0.9176	
	10^{-4}	0.6852		0.5267	
	10^{-3}	0.3962		0.1002	
	0.01	0.1007		0.0111	
	0.1	0.0104	(0.0122)	0.0012	(0.0011)
Changes in m:	m	F_W		ρ	
	0	0.9259	(0.9259)	0	(0)
	10^{-6}	0.9252		0.0011	
	10^{-4}	0.8680		0.1002	
	10^{-3}	0.6852		0.5267	
	0.01	0.5742		0.9187	
	0.1	0.5572		0.9925	
	0.9	0.5555	(0.5558)	1.0000	(0.9990)

drift. But we have almost no effect on ρ, as predicted from the approximate formula (VII-45) which does not contain N. Thus as the similarity between genes from the same population decreases, the similarity of genes in different populations decreases at the same rate. On the other hand, when we increase n, the number of islands, ρ drops toward zero. This it does because the amount of migration between any two given populations is $m/(n-1)$, which declines to zero as n is increased. Thus as n increases the islands becomes less and less similar. As this happens, new immigrants become more and more equivalent to mutants. The similarity of alleles in a population drops toward a limit, which can be found simply by replacing u by $u+m$ in the one-population isoallele model (VII-6). If we compute $1/[1+4N(u+m)]$ in this case, it is 0.5319, very close to what is observed when n is large.

When u is varied, as expected there is less and less gene identity within a population. The quantity ρ also drops as mutation provides a stronger and stronger force differentiating populations. Finally, the effect of increasing the rate of migration m is to increase the genetic similarity between populations, as one might expect. As migration becomes a stronger force, the set of n populations becomes more and more nearly a single random mating population. When $m = 1 - 1/n$, so that $m = 0.9$, each gene in a population came from one of the other populations chosen at random, so that our archipelago of islands is one random-mating population. In that case the amount of variation in a single population is precisely what would be predicted from an infinite isoalleles model with Nn individuals. The approximation formulae (VII-45) and (VII-46) do a good job for parameters in the ranges we have considered. As expected, they begin to lose accuracy when m, u, or $1/N$ become large.

The reader should be clear about the significance of quantities like F_W and ρ. They are not expected to have these values in any one population or even when averaged over pairs of genes at one locus drawn from the populations of any one archipelago. In the derivation of the equations for these quantities we considered the probability that two genes drawn from a random population *and from a random realization of the process* were identical. It is much the same as in the infinite isoalleles model: the probabilities are only expectations. An individual realization of the process could, as a result of random genetic drift, come to contain only the same one allele in all populations, despite the fact that $F_W < 1$. The values of F_W and F_B are averages over all possible pairs of genes sampled from all possible replicates of the process. As such they indicate only the expected course of events and not the fluctuations we should expect around those averages.

We have dealt only with equilibrium situations in this section. With three evolutionary forces operating, there will be a variety of time scales on which the system will respond to perturbation. If all variability is lost, the rate at which new variability reenters the whole system will obviously be governed by the mutation rate u. If the populations become excessively differentiated, the rate at which migration re-mixes the genetic con-

tents of the populations will be governed by m. If the populations become excessively similar (although containing variability) the rate at which they differentiate by genetic drift will be governed by $1/N$. A complete analysis of rates of return to equilibrium is beyond our scope here: there is insufficient space to discuss it.

RATE OF LOSS OF VARIABILITY WITHOUT MUTATION. However, a body of work is available on the rate of approach to equilibrium in one set of cases of particular importance. These are when there is no mutation ($u = 0$). Then the variability in the populations will be lost as F_W and F_B approach one. The questions which arise immediately are: (1) at what rate will the whole archipelago fix for the same allele, and (2) as the variability of the set of populations is being lost, will the residual variability represent mostly between-population variability, or will all populations have similar genetic composition? We will only sketch the methods used to answer these questions, as they are fairly straightforward. When $u = 0$ equations (VII-44) can be used to show that the equilibrium values of F_W and F_B are both 1. Substituting $H_W = 1 - F_W$ and $H_B = 1 - F_B$ into these equations, we find that they give two linear recurrence equations in two unknowns. These can be solved by matrix methods in the manner discussed above in Chapter V, in section V.11. Both H_W and H_B decline towards zero, and the rate of decline becomes geometric, so that after a while $H'_W \simeq \lambda H_W$ and $H'_B \simeq \lambda H_B$. The value of the largest root of the characteristic equation of the matrix which is found tells us the answer to the first of the two questions posed above. The smaller is this λ, the more rapidly variability (as expressed by either H_W or H_B) is lost. We can write out the quadratic equation with coefficients depending on N and m, and solve it either analytically or numerically.

The result shows an interesting transition, and a behavior somewhat different from the isoallele-mutation case just discussed. When Nm is large, the whole population drifts as if it were a single random-mating population. If it were, then clearly the value of λ would be $1 - 1/(2Nn)$, since this panmictic population would contain Nn individuals. As m is decreased, the rate of loss of variation stays near this value for some time. But there is a rather sudden transition in a particular range of values of m. Below this range, the rate of loss of variation begins to decline. Furthermore, it comes to depend on m, but not on N, whereas before it depended on N but not on m. When

$$\lambda \simeq 1 - \frac{1}{2Nn} \tag{VII-47}$$

and when m is small

$$\lambda \simeq 1 - \frac{2m}{n-1}. \tag{VII-48}$$

Figure 7.1 shows this transition, by graphing the rate of loss of variability, $1 - \lambda$, as a function of m. Clearly the critical range of values of m is near the intersection of the two

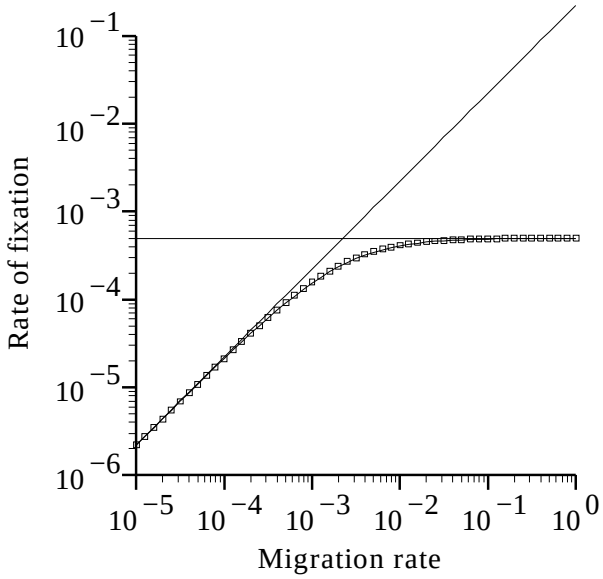


Figure 7.1: Rate of approach to fixation in an island model with $n = 10$ and $N = 100$. The squares are the result of exact calculation of eigenvalues. The curve is the approximation (VII-51).

asymptotes, at the point when (VII-47) equals (VII-48). So it is where

$$\frac{1}{2Nn} \;=\; \frac{2m}{(n-1)} \tag{VII-49}$$

which is where

$$4Nm \;=\; \frac{n-1}{n}. \tag{VII-50}$$

So we can state the principle that the population drifts as if one panmictic population provided Nm is large, but geographic structuring of the population impedes the spread of one or another allele if Nm is small. This is certainly consistent with the behavior of the one-island model. In fact, Nagylaki (1977a) has proven that any reasonable connected geographic structure behaves as if panmictic if the population sizes are large enough, which implies that all the Nm_{ij} are large.

As an aside, it may be worth noting that a good approximation to λ over all ranges of m is

$$\lambda \;\simeq\; 1 - \frac{1}{2Nn + \frac{n-1}{2m}} \tag{VII-51}$$

It is very much as if when m is large the population drifts with an effective population size near Nn, and when m is small it drifts with an effective population size of $(n -$

$1)/(4m)$. An even better approximation for all values of m is the sum of these two effective sizes.

All of the above statements may be verified by consideration of approximations to the full equation for λ, but we shall not do this here. The remaining aspect of the approach to fixation of an n-island model is the extent of geographic differentiation of the islands. This turns out to be completely concordant with the above rules. When the population has effective population size Nn, it shows little geographic differentiation, as judged by the ratio H_W/H_B, which will then be near unity. But when geographic subdivision is effective in impeding genetic drift, then there will be substantial differentiation, with far more heterozygosity if two genes are drawn from different populations than if drawn from the same population.

The populations then become differentiated from each other, and the whole species then has its mean gene frequency drift until it is 0 or 1. The differentiation arises much faster than the whole species drifts. It arises in a number of generations which is a small multiple of the local population size N. A good approximation to the extent of differentiation is that $H_W/H_B \simeq 4Nm/(4Nm+1)$.

REFERENCES. The n-island model was first envisaged by Sewall Wright (1931), who dealt only with the case $n = \infty$ when effectively it will behave like a one-island model in many respects. P. A. P. Moran (1959) first pointed out the transition in the rate of approach to homozygosity as m is varied. Maynard Smith (1970) gave an approximate solution, very much like the one here, to the island model with infinite isoallele mutation. Maruyama (1970) gave a less approximate treatment. The exact solution was given by Nei and Feldman (1972) and also by Latter (1973), who approximated the time dynamics of the model more fully. The intuitive discussion presented here owes much to the treatments by Maynard Smith (1970) and Spieth (1974). Robertson (1964) presented a general rule relating the extent of geographic differentiation to the rate of approaches of the whole species to fixation.

VII.6 Drift vs. Migration: the stepping stone model.

THE MODEL. We have gone into the island model in some detail because all the phenomena seen there also occur in the more complex spatial models. The island model has no geography: each population is in effect equidistant from each other population. There are many models of geographic population subdivision. The best investigated of these are the stepping stone models. We now look briefly at some of the results from this model. Our lengthy discussion of the island model will help us distinguish those effects due to the subdivision of the species into local populations from those which also require a particular geographic arrangement of the populations.

We concentrate first on the simplest case of the stepping stone model. We imagine that there is an infinite chain of equally spaced populations, each of size N. The model of population reproduction is a Wright-Fisher model with immigration, as it was in the island model. The difference is that of the m immigrant gametes, a fraction $m/2$ comes from the neighboring population to the left of the population, and a fraction $m/2$ comes from the neighbors on the right. This is shown diagrammatically in Figure 4.3. The chain of populations is assumed to be infinitely long.

Let us consider the equilibrium state of an infinite isoalleles model in an infinitely long stepping stone model. As before, this will not be an equilibrium of individual allele frequencies, but of the amounts of variability as measured by probabilities of allele identity. The probabilities we use are the F_{ij}, the probability that a gene drawn at random from population i will be the same allele as one drawn at random from population j, the sampling being without replacement if $i = j$.

AN ASIDE: THE GENERAL MIGRATION MATRIX MODEL. For any arbitrary geographic structure there is a general equation which the F_{ij} must satisfy at equilibrium. Using a sequence of lifestages in which the juveniles migrate:

$$\underset{(\infty)}{\text{Juveniles}} \xrightarrow{sampling} \underset{(N)}{\text{Adults}} \xrightarrow{meiosis} \underset{(\infty)}{\text{Gametes}} \xrightarrow{mutation} \underset{(\infty)}{\text{Gametes}} \xrightarrow{mating} \underset{(\infty)}{\text{Juveniles}} \xrightarrow{migration} \underset{(\infty)}{\text{Juveniles}}$$

the equation is

$$F_{ij} \;=\; (1-u)^2 \sum_k \sum_l M_{ik} M_{j\ell} \left[F_{k\ell} + \delta_{k\ell} \left(\frac{1 - F_{k\ell}}{2N} \right) \right]. \qquad \text{(VII-52)}$$

This complex-looking equation is actually rather simple. The factor $(1-u)^2$ comes, of course, from the isoallele mutation model. M_{ik} is the probability that a gene found in population i came in the previous generatian from population k. When $i = k$ it gives the probability that the gene is not a new immigrant. $M_{j\ell}$ is defined similarly. The quantity in brackets is easily interpreted. $\delta_{k\ell}$ is called the Kronecker delta function. It is simply a bookkeeping device. It is zero when $k \neq \ell$ but one when $k = \ell$. This means that when $k \neq \ell$ the quantity in square brackets is simply $F_{k\ell}$ but when $k = \ell$ it is

$$\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{k\ell}.$$

This quantity should be familiar to us by now. Thus the double summation in (VII-53) simply keeps track of all possible places in which the two genes could have been in the previous generation. $M_{ik}M_{j\ell}$ is the probability that genes now respectively in i and j came from k and ℓ, and the quantity in square brackets is their probability of identity if they came from k and ℓ.

THE STEPPING STONE MODEL AGAIN. In this general migration matrix model with an arbitrary number of populations and an arbitrary migration scheme given by the M_{ij}, there is no simple expression giving the solution of (VII-53). But in the one-dimensional linear stepping stone model, things simplify. Each gene can only have come from three possible places, the population where it was found and its two neighbors. So there are nine terms in the double sum:

$$\begin{aligned} F_{ij} \;=\; (1-u)^2\Big[&(m^2/4)F^*_{i-1,j-1} \;+\; (m^2/4)F^*_{i-1,j+1} \;+\; (m^2/4)F^*_{i+1,j-1} \;+\; (m^2/4)F^*_{i+1,j+1} \\ &+\; (m/2)(1-m)F^*_{i+1,j} \;+\; (m/2)(1-m)F^*_{i-1,j} \;+\; (1-m)(m/2)F^*_{i,j-1} \\ &+\; (1-m)(m/2)F^*_{i,j+1} \;+\; (1-m)^2F^*_{i,j}\Big]. \end{aligned} \tag{VII-53}$$

We have assumed that the populations are numbered in order by integers: -3, -2, -1, 0, 1, 2, ... Recall that the three possible migration events of a gene have probabilities $m/2$, $1-m$, and $m/2$ respectively. The asterisks on the F's indicate that each of these is a quantity like the one in square brackets in the preceding equation, being either F or $1/(2N)+[1-1/(2N)]F$ depending on whether or not the two subscripts of F are equal.

Now we can both simplify and approximate. We simplify by assuming that at equilibrium the probability of allelic identity is dependent only on how far apart are the two populations. So $F_{ij} = F_{i+1,j+1} = F_{i+2,j+2} = \ldots = F_{i-1,j-1} = F_{i-2,j-2}$, etc. If we let $F_{ij} = f_{i-j}$ the set of values of the f's will tell us all the values of the F's. So we need only solve for the f_i, the probability of identity between two genes drawn from populations i steps apart. In addition to this simplification we also approximate by assuming that m, u, and $1/N$ are small, and ignoring products and squares of these small quantities. After all of this (VII-53) becomes the set of equations

$$\begin{aligned} f_0 &= \left(1-2m-2u-\tfrac{1}{2N}\right)f_0 + 2mf_1 + \tfrac{1}{2N} \\ f_1 &= mf_0 + (1-2m-2u)f_1 + mf_2 \\ &\vdots \\ f_k &= mf_{k-1} + (1-2m-2u)f_k + mf_{k+1}. \\ &\vdots \end{aligned} \tag{VII-54}$$

We have used in the first of these equations the fact that $f_1 = f_{-1}$, so that we only need to know f_i for positive values of i.

Notice that all but the first equation are identical in pattern. If we consider what will happen to f_k as k becomes large, it is intuitively clear that f_k will decline to zero. The farther apart are two genes, the more certain that mutation will have occurred to one or the other since they last had a common ancestor. We can assume that the decline of f_k is

geometric, so that (for large k) $f_{k+1} \simeq \lambda f_k \simeq \lambda^2 f_{k-1}$. Since from (VII-54)

$$m f_{k+1} - 2(m+u) f_k + m f_{k-1} \;=\; 0, \tag{VII-55}$$

after substituting $\lambda^2 f_{k-1}$ for f_{k+1} and λf_{k-1} for f_k, we obtain the quadratic equation

$$m\lambda^2 - 2(m+u)\lambda + m \;=\; 0 \tag{VII-56}$$

so that

$$\lambda \;=\; \frac{m+u \pm \sqrt{(m+u)^2 - m^2}}{m}. \tag{VII-57}$$

The relevant root is the smaller one. (The full solution of the difference equation (VII-55) is a linear combination of powers of the two roots, but the larger root is greater than 1 and it can be shown that it must have a coefficient of 0, as f_k declines to 0 as k increases). We can approximate the smaller root when $u \ll m$ by

$$\lambda \;\simeq\; 1 - \sqrt{2u/m} \tag{VII-58}$$

so that f_k, which should (for large k) be proportional to λ^k is approximately

$$f_k \;\simeq\; f_0 \;\exp[-k\sqrt{2u/m}], \tag{VII-59}$$

This argument has been a bit fast and loose, making many assumptions and approximations. But we have gone through it because it is one result which can be obtained fairly easily. The expression for f_1, which is in terms of f_0, can also be substituted back into the first equation if we use our approximation to compute f_2 from f_1 and we can then use (VII-54) to obtain an expression for f_0.

FURTHER RESULTS. Many variants of the stepping stone model have been treated in the literature. The methods needed to solve for equilibrium levels of variability or for rates of approach to homozygosity are rather tedious and difficult. Rather than attempt to present them here, we will simply present the main conclusions which have come out of this body of work. References to the work will be found at the end of this section.

As we have seen, in the infinite isoalleles model in a one-dimensional infinite stepping stone model, the identity of alleles in two populations will decay at the same rate as $\exp[-k\sqrt{2u/m}]$ for large distances between the populations. This exponential decline is not quite true for a two dimensional stepping stone model. A two-dimensional stepping stone model is defined similarly to the one dimensional case. It is assumed that a fraction m of gametes in a population are replaced by gametes from the four neighbors. Each neighbor in the grid of populations contributes a proportion $m/4$ of the gametes. If we express the probability of identity as a function of the distance between the two populations, we find that when k is large, f_k is

$$f_k \;=\; c\; f_0\; e^{-k\sqrt{(4u/m)}} \Big/ \sqrt{k} \tag{VII-60}$$

where the constant c is a rather messy expression depending on both u and m, but not on N.

In both cases, more exact formulas are available for computing the f_k, but they yield little in the way of insight. We have dealt only with infinitely long one- and two-dimensional stepping stone models. Models of finite extent (lines, circles, rectangles or tori of populations) have also been analyzed and show similar patterns. Note that although there is no analogue in the n-island model for the rate of decline of allelic identity with distance, there is one direct parallel. Like F_B/F_W, f_k/f_0 depends only on the relation between u and m, and not on N to any extent.

RATE OF LOSS OF VARIABILITY. A clearer picture is obtained by eliminating mutation from the models, and asking at what rate the allelic identity f_k approaches 1 (i.e., how rapidly does the population fix for one allele). Like the island model, after an initial period the rate of decline of variability, both within a single colony and throughout the species range, settles down to a constant rate. The rates of decay of variability show a transition behavior similar to that seen in the island model. Note that we can most meaningfully discuss the approach to fixation in a species of finite extent (a line, circle, rectangle, or torus of populations), for in an infinitely long stepping stone model the whole species can never fix for one allele. When m is large, the whole species drifts as if one panmictic population, with very little geographic differentiation. When m is small, there is local geographic differentiation during the fixation process. The process of fixation then consists primarily of the spread through the species of a patch of nearly-fixed populations. Each population is fixed one way or another, and the patches of fixed populations each spread or shrink until all populations are fixed for one allele.

This transition from the one behavior to the other is qualitatively like that seen in the n-island model. We can define an effective population size for this process of genetic drift, and approximate it for large and small m. As in that case, the effective population size is well approximated by the sum of these. Figure 7.2 shows the transition undergone by the rate of decay of variability (more properly, the rate of approach of the F's to 1 per generation).

In the one-dimensional case, when there are n populations in a line, the effective population size is approximately (when there are a reasonably large number of populations)

$$N_e \simeq \frac{n^2}{\pi^2 m} + Nn \tag{VII-61}$$

while in a two-dimensional rectangle of $n_l \times n_2$ populations we have

$$N_e \simeq \frac{2n_1 n_2}{m} + Nn_1 n_2 \tag{VII-62}$$

Note that when m becomes large the second term in each of these expressions predominates. It is the total number of individuals in the species. When m is small the first

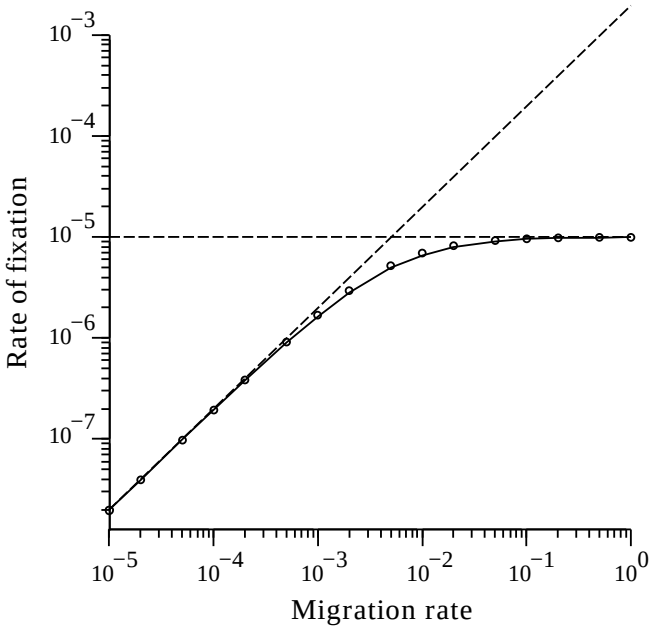


Figure 7.2: Rate of approach to fixation in a one-dimensional stepping-stone model with $n = 50$ and $N = 1000$. The circles show exact values computed using equation (VII-52), the curve uses an approximation based on equation (VII-61), and the dashed lines are the rates based on the two parts of that equation.

terms, which do not depend on the local population size N, dominate. A particularly interesting phenomenon occurs when we look at the point at which the transition to effective panmixia occurs. We can find this by finding the value of m at which the two asymptotes in Figures 7.2 intersect. This is the same as the point at which the two terms of (VII-61) or of (VII-62) are equal. For one dimension the transition occurs when $Nm = n/(2\pi^2)$. For two dimensions we require $Nm = 2$. Note the difference in behavior of the one- and two-dimensional cases. When we have a rectangle of populations, it will drift as one panmictic population whenever there are (substantially) more than two immigrant individuals (four surviving immigrant gametes) expected per generation. This will hold *no matter how many populations there are* in the rectangle. $Nm > 2$ is enough to "thoroughly mix" even a very large species. But in the one-dimensional case, when we increase n, we also increase the amount of migration needed to make the species drift as if one large population. So for a fixed amount of migration, a long enough line of populations will show local genetic differentiation during the fixation process. But a large enough square or rectangle may not show genetic differentiation unless $Nm < 2$.

Note by comparing these results to (VII-51) that the n-island model behaves much more like a two-dimensional than a one dimensional model. This seems to be related to the fact that in both the island and the two-dimensional model there are many routes from one population to another via chains of other populations, while in the one-dimensional model there is only one route.

RELATION TO THE EQUILIBRIUM WITH MUTATION. Now we can relate these results about the rate of genetic drift to the equilibrium variability under an isoallele model. When the mutation rate is so small that we expect far fewer than one mutant per generation *in the whole species* (i.e., $2Nnu << 1$), after each mutation there will be a prolonged period of genetic drift. Thus the species will be quite likely to drift to fixation or loss of the allele before the next mutation occurs at the locus in question. Thus if we ask whether we will see geographic differentiation with u very small, this will be nearly the same as asking whether we see geographic differentiation during the process of approach to fixation of one allele. Nearly, but not quite, the same. When we ask about geographic differentiation with small u, we must also include those cases in which the species has reached fixation. If we use the ratio f_k/f_0 as a measure of geographic differentiation, we will find that since we are by and large examining totally fixed species, $f_k \simeq f_0 \simeq 1$. Thus this measure will show little geographic differentiation. A gene from a distant population is just as likely to be the same allele as one from the same population.

But there is another, equally relevant measure of genetic differentiation, related to the ratio H_W/H_B discussed in the section on the island model. This is $(1-f_0)/(1-f_k)$, which compares the *heterozygosity* within a population to that between populations. Now the instances in which the species is totally fixed contribute no heterozygosity to either $1-f_0$ or $1-f_k$. So in examining this quantity we are, in effect by asking only about those generations during which heterozygosity exists, looking only at those cases not yet fixed. This quantity, which will be substantially less than 1 when there is more heterozygosity between than within populations, behaves just as it does when there is no mutation. There will be found to be a transition between local differentiation and effective panmixia as m is increased, and it will occur at the same value of m as discussed above.

When $4Nnu > 1$ (one dimension) or $4Nn_1n_2u > 1$ (two dimensions) the entire analogy between the equilibrium and the approach to fixation breaks down, and we can no longer look to fixation rates for insight into geographic differentiation of populations.

REFERENCES. We have so far deferred citing the actual literature on the stepping stone model. It was first formulated by Kimura (1953) and independently by Malécot (1950), who obtained the first approximate solutions. Earlier Sewall Wright (1940, 1943, 1946) propounded a model of individuals distributed in a spatial continuum, as did Malécot (1948, 1969). We have not covered the continuum models since they involve a

questionable assumption (Felsenstein, 1975), but they were the first models of genetic drift in a truly geographically subdivided population.

Kimura and Weiss (1964; Weiss and Kimura, 1964) presented a more detailed analysis of the infinite-length stepping stone models. Malécot (1950, 1951) obtained solutions for the equilibrium in some finite-length cases, and Maruyama (1970) first obtained the rate of loss of variability in finite-length stepping stone models. Maruyama has also written many papers providing detailed solutions and approximations to a wide variety of stepping stone models. Readers will find references to these and related papers in his monograph (1977) and also in my own review paper (1976) which contains some erroneous formulas. I have more recently (2015) presented approximations for finite 1- and 2-dimensional stepping-ston models that improve on Maruyama's. The transition in behaviors of the stepping stone model was first discussed by Kimura and Maruyama (1971) and also by Maruyama (1972). Korolev et al. (2010) have found a similar transition in stepping stone model behavior in a continuous approximation using stochastic differential equations. The interpretation of the transition in terms of effective population numbers is my own, and is published here. Nagylaki (1976, 1977a, Nagylaki and Barcilon, 1988) has used continuous-space approximations to investigate the rate of convergence of a stepping-stone model to its equilibrium. This can take quite long.

VII.7 Probability of Fixation of a Mutant

When genetic drift acts in opposition to mutation or to migration, it is possible to gain a clear picture of events by investigating the behavior of means and variances of gene frequencies, or equivalently of probabilities of allelic identity. When natural selection and genetic drift are the forces present, the methods become more complex, leading us ultimately to the diffusion equation methods, the most sophisticated mathematics used in population genetics. We will ease into this morass by considering first the simplest models containing both drift and selection. These treat the probability that a rare advantageous allele introduced into a large population will successfully fix.

PROBABILITY OF FIXATION. Consider an allele introduced as a single copy into a very large population which is reproducing according to a Wright-Fisher model (discrete generations, random mating, no mutation), and which is otherwise totally fixed for the other allele. One might think that if the population were large enough, the course of change in gene frequency would follow the expectation which we get from the deterministic models of Chapter II. But a moment's reflection will suffice to see that this cannot be so. If the gene is very rare (and it is) and has relative fitness $1 + s$ when heterozygous, then the results of section II.6 tell us that we expect the gene frequency to be multiplied by approximately $1 + s$ per generation. This means that if we now have one copy of the

allele, then in the next generation, is $s = 0.2$, we should have 1.2 copies. This is clearly impossible.

The paradox is partially resolved when we realize that the relative fitness $1 + s$ is only an expectation. *On average* the carriers of our allele A will have $1 + s$ times as many offspring as the homozygotes for a. But a single Aa individual may have no offspring, or a few, or a great many. It is only necessary that the mean number of survivors work out to $1 + s$ (relative to aa). If there is any chance at all of having no surviving offspring, then it is possible for the allele A to be lost in the first generation. Even if it survives, it is quite possible for the survivors to themselves all have no surviving offspring. It is clearly quite possible that a new mutant allele will die out as a result of these chance events, even though it is selectively advantageous and occurs in an infinite population. We will try to find the probability that the allele survives long enough, and comes to exist in enough copies, to ultimately complete the process of substituting itself for the existing allele. We return later to the question of how this random process may be reconciled with our notion that when the population size is very large there should be little or no genetic drift.

THE BRANCHING PROCESS. We take as our starting point the probabilities $p_0, p_1, p_2, \ldots$ that an individual A allele in an adult Aa gives rise to 0, 1, 2, ... copies of the A gene among the adults of the next generation. We then return to compute these probabilities after doing the analysis in terms of them. We seek the probability of long-term survival of the A allele. The easiest way to find this is to consider instead the probability λ that the A gene is ultimately lost. These two probabilities (of loss and of survival) must sum to unity. Given a single Aa individual, A will be ultimately lost if it has no offspring in the next generation, an event whose probability we have assumed to be given by p_0. But it may also be lost if it gives rise to only one A in the next generation. That event has probability p_1. Knowing that there is only one A in the next generation, so that it must be in a Aa individual, we can see that since by assumption the relative fitness of Aa remains constant at $1 + s$, the probability that that single Aa offspring contains an A gene which is ultimately lost is λ. So we can put this together to say that the chance that our original A gene dies out through a series of events which start by only one offspring containing A being produced is the product $p_1\lambda$. Note that in trying to compute λ we are in the process finding a formula for it that itself involves λ. This will work out.

It is also possible for there to be two copies of A in the next generation. Since the A allele is extremely rare, it is overwhelmingly likely that these A-bearing adult offspring are both Aa individuals. The probability that there are two A-bearing offspring is p_2. Now to have the A allele be lost, neither of those A genes can give rise to a lineage which ultimately survives. The critical assumption we now use is that, since these two individuals are each diluted out in a very large population, they do not interact in any way with each other, *nor do their descendants*. If this assumption holds, then the event of the dying out of one A line occurs independently of whether the other A line dies out.

Each of these lines is again a line started from a single gene. So the probability that each one dies out is λ, and the probability that both die out is (by independence of these two events) λ^2. Thus the probability that the original A gene has two descendants and that neither of these ultimately has surviving A-bearing offspring is $p_2\lambda^2$.

By simple extension of this argument we can see that the probability that the lineage of A's has 3 offspring, whose descendants all die out is $p_3\lambda^3$, and so on. We have now computed λ in terms of itself as:

$$\lambda \;=\; p_0 + p_1\lambda + p_2\lambda^2 + p_3\lambda^3 + \cdots + p_k\lambda^k + \ldots, \qquad \text{(VII-63)}$$

the summation continuing to the largest number of offspring the original gene could possibly have. This process, in which a single particle (in our case, a gene) can reproduce different numbers of offspring, each of which independently gives rise to a line of offspring according to the same process, is called a *branching process*. The probability of death of a lineage founded by a single gene is found by solving (VII-63) for λ, if that is possible. There will always be a root $\lambda \;=\; 1$ in (VII-63). When the chance of survival is greater than zero, there will also be another root in which $\lambda < 1$.

THE WRIGHT-FISHER MODEL WITH SELECTION. Now let us put some flesh on the bones of equation (VII-63) by specifying values for the p_k. We recall that the original A is in an adult, that the population reproduces according to a Wright-Fisher model, and that the relative fitness of *Aa* is $(1+s)$. To find the probability that there are n *Aa* adults in the next generation, we must specify what we mean by a Wright-Fisher model when there is natural selection. Recall that in a Wright-Fisher model, it is as if an infinite pool of gametes were produced, these combine at random, and then N surviving adults are chosen by the lottery of density-dependent population size regulation. To incorporate fitnesses into the scheme, we need only assume that *in the production of the gamete pool and in subsequent events up to but not including population size regulation, natural selection (or migration, or mutation) is at work*. Our life cycle diagram is:

$$\underset{(\mathrm{N})}{\text{Adults}} \xrightarrow[\text{(fertility differences)}]{\text{selection meiosis,}} \underset{(\infty)}{\text{Gametes}} \xrightarrow{\text{random union}} \underset{(\infty)}{\text{Zygotes}} \xrightarrow[\text{(viability differences)}]{\text{selection}} \underset{(\infty)}{\text{Late preadult zygotes}} \xrightarrow{\text{density regulation}} \underset{(\mathrm{N})}{\text{Adults}}$$

Thus having one copy of A with relative fitness of its bearers being $1+s$ will lead, whether by fertility or viability effects, to a proportion $(1+s)/N$ of the late preadult zygotes being heterozygotes *Aa*, where N is the population size, assumed large. On each of the N "draws" which determine survivors' genotypes, the chance of getting a *Aa* is $(1+s)/N$. The average number of A genes surviving is thus $N(1+s)/N = 1+s$. Thus in a Wright-Fisher model under these conditions, the number of *Aa* survivors is drawn

from a binomial distribution with N trials and the probability $(1+s)/N$ of success on each trial.

But we have assumed that N is very large. These conditions (very small probability of success on each trial) are precisely those in which the number of surviving offspring will nearly follow a Poisson distribution with mean number of offspring $1+s$. The validity of this approximation will increase as we consider cases of larger and larger N. Thus the number of offspring Aa is k with probability

$$p_k \;=\; e^{-(1+s)}(1+s)^k/k!, \tag{VII-64}$$

these being the Poisson probabilities when the mean is $1+s$. To find λ, we substitute these for the p_k in (VII-53). We find that

$$\begin{aligned}\lambda \;&=\; e^{-(1+s)} \;+\; e^{-(1+s)}(1+s)\lambda \;+\; e^{-(1+s)}(1+s)^2\lambda^2/2 \;+\; \ldots \;+\; e^{-(1+s)}(1+s)^k/k! + \ldots \\ &=\; e^{-(1+s)}\left[1+(1+s)\lambda+(1+s)^2\lambda^2/2+\cdots+(1+s)^k\lambda^k/k!+\ldots\right].\end{aligned} \tag{VII-65}$$

The power series $1+x+x^2/2+\cdots+x^k/k!+\ldots$ is simply the Taylor series expansion of e^x, and we can thus write (VII-55b) as

$$\begin{aligned}\lambda \;&=\; e^{-(1+s)}\; e^{\lambda(1+s)} \\ &=\; e^{(\lambda-1)(1+s)}.\end{aligned} \tag{VII-66}$$

The value of λ is found by solving (VII-66) for λ, given the value of s. Unfortunately there is no closed-form expression for λ.

AN APPROXIMATION. An approximate solution can be obtained when s is small by expanding the right side of (VII-66) as a power series in $(\lambda-1)$ and dropping terms beyond the square. This involves the assumption that λ is near 1. We get

$$\lambda \;\simeq\; 1+(\lambda-1)(1+s)+(\lambda-1)^2(1+s)^2/2 \tag{VII-67}$$

or

$$(\lambda-1)\left[1-(1+s)-(\lambda-1)(1+s)^2/2\right] \;\simeq\; 0 \tag{VII-68}$$

which is solved either when $\lambda \;=\; 1$ or when

$$1-\lambda \;\simeq\; \frac{2s}{(1+s)^2} \tag{VII-69}$$

Our analysis shows that when s is small, the probability of survival of a new mutant (for that is what $1-\lambda$ is) is nearly $2s$. When s is negative, the only acceptable solution of (VII-68) is $\lambda \;=\; 1$.

Table 7.4: Comparison of exact probability of survival in a branching process with two approximations.

s	Exact Probability	$2s/(1+s)^2$	$2s$
0	0	0	0
0.01	0.01973	0.01922	0.02
0.02	0.03896	0.03845	0.04
0.05	0.09370	0.09070	0.10
0.10	0.17613	0.16529	0.20
0.20	0.31369	0.27778	0.40
0.50	0.58281	0.44444	1.00
1	0.79681	0.50000	2.00

Clearly a new mutant in a very large population has a nonzero chance of spreading only when $s > 0$. This is certainly consistent, in a qualitative sense, with the deterministic results for selection in an infinite population. Table 7.4 compares exact solutions of (VII-66) with the approximations $2s$ and $2s/(1+s)^2$. The exact probability of survival is between the two approximations, somewhat closer to (VII-69).

When s is small, clearly $2s$ is close enough to the probability of survival to serve as a working rule of thumb. It is worth considering how small a probability of survival this is. When $s = 0.01$, only one new mutant in 50 will succeed in spreading, despite the fact that all are advantageous. Even with s as large as 0.1, large enough to guarantee fairly rapid change in gene frequencies in the deterministic case, only one new mutant in six will establish itself. Obviously, genetic drift is a powerful force when only a few copies of an allele are in existence. Only rarely will an allele, even if advantageous, escape from the risk of loss due to the randomness of births and deaths, and of Mendelian segregation.

But once an advantageous allele reaches a substantial number of copies, its continued survival is better assured. If there are n copies of an allele, it can only be lost by all n lineages of A-bearing individuals going extinct. The probability that this will happen is λ^n, since we still assume that the population is so vast that these different lineages do not interact, and survive or are lost completely independently of one another. The probability that an allele represented initially by n copies is lost is

$$\lambda^n \simeq (1-2s)^n. \tag{VII-70}$$

Once 100 copies of an allele exist, when $s = 0.01$, the probability that it is lost is only 0.14, and once 1000 copies exist, the probability that it will be lost thereafter is less than 3×10^{-9}. This provides us with some insight into the time dynamics of the process of

establishment or loss of a new advantageous mutant.

As $1-2s \simeq e^{-2s}$, The quantity $(1-2s)^n$ can be approximately written as e^{-2ns}. If $2ns$ is substantially greater than 1, this means that the favored allele is very unlikely to be lost. This is true when $n > 1/(2s)$. Thus when the number of copies of the advantageous allele substantially exceeds $1/(2s)$, the allele is virtually guaranteed to fix. This is true even if it is as yet still at a low gene frequency. So most of the loss of advantageous alleles takes place while these alleles are still present in only a few copies (this is not as silly a statement as it sounds). This in turn must be during the first few generations. An allele present in only one or a few copies is constantly at risk of being lost and could not last long in that state. If it survives many generations it must therefore be fortunate enough to have drifted to a larger number of copies.

NUMBER OF COPIES PRESENT. We can now get some insight by asking, not about the probability of survival, but about the average number of copies which we expect to be present after t generations. Starting with one mutant copy of A, we expect to have an average of $1+s$ copies in the next generation. These in turn each will have an average of $1+s$ A-bearing offspring, and so on. After t generations there should be an average of $(1+s)^t$ copies of A. Consider a large value of t, say 1000 generations. If $s = 0.01$, then we expect an average of 20,959 copies of the allele to exist after 1000 generations. But this is an average both over cases in which the allele has been lost and those in which it still survives. The process of loss of the allele will have mostly run its course long before 1000 generations. So, since $1 - 0.01973 = 0.98027$ of the time the allele will be lost, this average of 20,959 represents an average of numbers which are zero 0.98027 of the time! Thus if there are any A alleles around at all (an event which happens only 0.01973 of the time), there must be (on average) 20,959/0.01973 = 1,062,299. Thus the figure of 20,959 will not be typical of any particular population. A few populations will have 50 times that many alleles, but in most cases there will be no A copies left after 1000 generations.

A more careful analysis of the branching process shows that loss does indeed occur within a few generations of the initial occurrence of the allele, if it is going to occur at all. In those cases in which there is survival of the mutant, it drifts upwards in numbers until a substantial number of copies (as we have seen, about $1/(2s)$ of them) are available. Thereafter it increases relatively smoothly by a fraction of s per generation.

This latter is precisely the behavior expected from the deterministic selection equations. We are still left with the problem of how to reconcile the deterministic prediction with the results from the branching process, as both are supposed to apply when N is large. The paradox is resolved by focusing on the number of copies initially present. When there are only a few, the branching process accurately indicates that most likely the favored allele will be lost, and that in any case genetic drift will be a major influence on gene frequencies in early generations. But recall that we have assumed a vast population. Even if the initial gene *frequency* is $p_A = 0.01$, this will represent a large *number* of copies of A. In that case A is expected to increase its frequency smoothly, and have

little chance of being lost. Even if we have a new mutation coming into a population which lacks it, if $2Nu \gg 1$ there will be many copies of the mutant even during the first generation, and these mutants will follow deterministic patterns as a result. Thus there is no contradiction between the results of the branching process and of the deterministic treatment. However very large population sizes (10^9) will often be required to get this consistency, and otherwise the branching process will more accurately reflect what happens when only a few copies of an allele exist.

REFERENCES. The branching process was introduced by Francis Galton and the Reverend Thomas Watson (Watson and Galton, 1874) in the last century to model the extinction of family surnames (the problem had been worked on earlier by I. J. Bienaymé, as pointed out by Heyde and Seneta, 1977). It was first applied to the problem of extinction of a gene by R. A. Fisher (1922), who treated only neutral genes. Haldane (1927) extended the treatment to advantageous genes and obtained the approximation $2s$. The number-of-copies argument of the previous paragraphs is given by John Maynard Smith (1971).

VII.8 The Diffusion Approximation to Fixation Probabilities.

THE WRIGHT-FISHER MODEL WITH SELECTION. The weakness of the branching process approach to fixation probability is that it assumes that all the different copies of the allele reproduce independently of each other. This can be a good approximation only if the allele is at low frequency in a very large population. In this section, we begin to develop the diffusion approximation, which will work for any initial gene frequency of the allele.

Before doing so, it is worth looking at the Wright-Fisher model with selection. Is there any hope of finding the fixation probability exactly? We assume that the life cycle is that given in the last section. We will also restrict consideration to the case where only the viability stage of the life cycle is affected by the genotype. We do not consider fertility differences, which we assume do not exist. Suppose that the viabilities of AA, Aa, and aa are respectively w_{AA}, w_{Aa}, and w_{aa}. These give the probability that a zygote of a certain genotype survives to the life stage where the density-regulation random sampling starts. The sampling itself is simply random choice of N individuals. Now suppose that we know the numbers of AA, Aa, and aa adults in generation t, and wish to find the probabilities of various outcomes in generation $t+1$.

The first thing to notice is that gametes are being produced, the proportion of A genes among them is the same as the gene frequency of A among the adults that produced them. So the future behavior of the population depends, in this case, only on the

current gene frequency and not at all on the current genotype frequencies. This allows us to shift our attention to the number of A genes out of the $2N$ gene copies present in the adults. Suppose that there are i copies of A. Among the gametes, $p = i/(2N)$ are A, since the gametes are produced without fertility differences. The gametes combine at random to form the zygotes (which is of course the same as having each zygote formed by two randomly-chosen adults). So the newly formed zygotes are in Hardy-Weinberg proportions at gene frequency p. Natural selection acts on the infinite population of zygotes. After selection acts, the AA, Aa, and aa survivors are in the frequencies $p^2 w_{AA}/\bar{w}$, $2p(1-p)w_{Aa}/\bar{w}$, and $(1-p)^2 w_{aa}/\bar{w}$ respectively, where $\bar{w}$ is the mean fitness $p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa}$. The N surviving adults are then determined by random sampling from these infinitely many survivors. Thus the probability of finding n_1, n_2, and n_3 survivors of genotypes AA, Aa, and aa is the trinomial sampling probability (the probability of these numbers of outcomes if we toss a three-sided coin)

$$P(n_1, n_2, n_3) \;=\; \frac{N!}{n_1!\; n_2!\; n_3!} \left[p^2 w_{AA}/\bar{w}\right]^{n_1} \left[2p(1-p)w_{Aa}/\bar{w}\right]^{n_2} \left[(1-p)^2 w_{aa}/\bar{w}\right]^{n_3} \tag{VII-71}$$

In practice we are only interested in what the gene frequency among adults in the next generation is. So we are interested in the total probability of all those combinations n_1, n_2, n_3 in which we have a total of j copies of the A allele, i.e., in which $2n_1 + n_2 = j$. So if we have i copies of A in the present generation, the probability that we have j copies in the next generation is

$$P(j|i) \;=\; \sum_{k=0}^{j/2} P(k,\; j-2k,\; N-j+k), \tag{VII-72}$$

the three arguments of P on the right side being determined by the requirements that $n_1 + n_2 + n_3 = N$ and that $2n_1 + n_2 = j$. Recall that $p = i/(2N)$ in (VII-71).

There is a great deal more to say about the Wright-Fisher model with selection, but little space here to say it. Equations (VII-72) and (VII-71) allow us to calculate the transition probabilities $P(j|i)$ of going to j copies of A from i copies. Versions of the model allowing fertility selection as well can also be written down, but we will not do so here. It is worth pointing out that the sampling which takes place in choosing N adults is not the same as sampling $2N$ genes from the zygotes. If the zygotes had not undergone a round of selection, the two processes would be the same, since the two genes in a single individual might as well be sampled independently. But if natural selection gets the population out of Hardy-Weinberg proportions, there is a lack of independence between the two genes in an individual: knowing whether one is A tells us something about whether the other is A also. As an extreme example, if $w_{AA} = w_{aa} = 0$ so that only heterozygotes Aa survive selection, then all N adults must be Aa, so that sampling does not alter the gene frequency from 0.5 at all. Conversely, if no heterozygotes survive, then

each surviving individual sampled contains two A or two a genes, which means that the effect of sampling in changing gene frequencies is twice as great as we might imagine just by looking at the number of genes.

EXACT FIXATION PROBABILITIES. Once we know the transition probabilities $P(j|i)$, we could hope to use them to work out the fixation probabilities. These will be a set of quantities u_i, giving the probability that A becomes fixed given that we start with i copies of A. For each value of i we have the following basic equation:

$$u_i \;=\; \sum_j P(j|i)\; u_j. \tag{VII-73}$$

These equations are analogous to equation (VII-63) in the branching process calculation. They express the fixation probability as an average of the fixation probabilities u_j from all gene frequencies (or gene numbers) to which the population could change from its starting point in one generation, with these being weighted by the probabilities $P(j|i)$ that this particular one-generation change will take place.

Two of the values of the u_i are known in advance. When there are no copies of A present, the gene cannot fix, so that $u_0 = 0$. When all genes in the population are A, the gene has already fixed, so that $u_{2N} = 1$. With the $P(j|i)$ known, and with these two "boundary conditions", equation (VII-73) specifies a set of $2N-1$ equations in $2N-1$ unknowns (two of the u_i have been determined). In any particular case for which fixation probabilities are needed, if N is not too large one can determine the coefficients $P(j|i)$ numerically, and solve for the u_i by numerical methods. Even with a computer it is difficult to deal with cases in which $N > 100$. Unfortunately, there is no known algebraic expression in terms of N, w_{aa}, w_{Aa}, and w_{aa} which solves (VII-73) for the fixation probabilities.

THE DIFFUSION APPROXIMATION. We are therefore faced with the necessity of approximating. The approximation we present here is the diffusion method, which we will also use in the next section to determine equilibrium distributions of gene frequencies. We start by expressing the fixation probability as a function of gene frequency rather than of numbers of copies of the allele. Let $U(p)$ be the fixation probability given that A starts out at gene frequency p. Then $u_i = U(i/(2N))$. We also replace the quantity $P(j|i)$ by the probability of that particular change in gene frequency. Let us call it $P_p(\Delta p)$. We are still on a gene frequency scale that has only $2N+1$ possible values of the gene frequencies, one for each possible number of A alleles possible in the population, so the summation is still discrete and the values of P are still probabilities, not probability densities. Then $P(j|i) = P_{i/(2N)}[(j-i)/(2N)]$, although we will not need to use this relationship. Equation (VII-73) now becomes

$$U(p) \;=\; \sum_{\Delta p} P_p(\Delta p)\; U(p+\Delta p) \tag{VII-74}$$

So far we have simply re-expressed (VII-73) without approximating at all. Note that the summation over Δp is over all possible changes in gene frequency. (These Δp are not infinitesimal quantities).

We now approximate $U(p + \Delta p)$ by replacing it by the first three terms of its Taylor series. Indicating derivatives with respect to p by primes,

$$U(p + \Delta p) \simeq U(p) + \Delta p\, U'(p) + \frac{(\Delta p)^2}{2} U''(p) \tag{VII-75}$$

When we substitute this into (VII-74) we get

$$U(p) \simeq \sum_{\Delta p} P_p(\Delta p)\, U(p) + \sum_{\Delta p} P_p(\Delta p)\Delta p\, U'(p) + \frac{1}{2}\sum_{\Delta p} P_p(\Delta p)(\Delta p)^2\, U''(p). \tag{VII-76}$$

Noting that $U(p)$, $U'(p)$, and $U''(p)$ do not contain Δp, we move them outside the summations:

$$U(p) \simeq U(p)\sum_{\Delta p} P_p(\Delta p) + U'(p) \sum_{\Delta p} P_p(\Delta p)\, \Delta p + \frac{1}{2}U''(p)\sum_{\Delta p} P_p(\Delta p)\, (\Delta p)^2 \tag{VII-77}$$

Now note that $\sum P_p(\Delta p)$ is simply the sum of the probabilities of all conceivable changes in gene frequency. This must be 1. The quantity $\sum P_p(\Delta p)\Delta p$ is the weighted average of all the changes Δp in gene frequency. It is the expected change in gene frequency, $\mathbb{E}(\Delta p)$, which we will call $M(p)$. Finally, the term $\sum P_p(\Delta p)(\Delta p)^2$ is the expectation of the squared change in gene frequency, $\mathbb{E}[(\Delta p)^2]$, which we call $V(p)$. Note that we express M and V as functions of the current gene frequency p because these expectations of Δp and of $(\Delta p)^2$ are different for different gene frequencies. Now (VII-76) becomes

$$U(p) \simeq U(p) + U'(p)\, M(p) + \frac{1}{2}U''(p)\, V(p), \tag{VII-78}$$

or

$$M(p)\, U'(p) + \frac{1}{2}\, V(p)\, U''(p) \simeq 0. \tag{VII-79}$$

This equation is called (in a slightly more general form) the Kolmogorov Backward Equation. (We shall see the Kolmogorov Forward Equation in the next section). Hereafter we drop the $\simeq$ in favor of $=$.

Before we solve it, it is worth inquiring what we have assumed. In dropping terms from the Taylor series for $U(p + \Delta p)$, we in effect assumed that the terms which contributed the bulk of the quantity $U(p)$ in (VII-74) involved small values of Δp, which is to say that $P_p(\Delta p)$ is small except when Δp is small. In other words, p is changing only by small amounts in any one generation. This amounts to the assumption that population sizes are large and selection coefficients small. Of course, our derivation here

is heuristic. Advanced population genetics texts (e.g. Ewens, 2004, section 4.5) may be consulted for a more formal treatment.

Equation (VII-78) is easily solved to give $U(p)$ in terms of $M(p)$ and $V(p)$. It can be rearranged (unless $U'(p) = 0$ or $V(p) = 0$), using the fact that $U''(p)/U'(p)$ is the derivative of $\ln U'(p)$, to be

$$-2M(p)/V(p) \;=\; U''(p)/U'(p) \;=\; \frac{d}{dp}[\ln U'(p)] \tag{VII-80}$$

and we then integrate to get

$$-2\int_c^x M(p)/V(p)\; dp \;=\; \ln U'(x) - \ln U'(c) \tag{VII-81}$$

or

$$U'(x) \;=\; U'(c)\exp\left[-2\int_c^x M(p)/V(p)\; dp\right] \tag{VII-82}$$

where the lower limit of integration c is not specified yet. We can call the right-hand side of (VII-82) $U'(c)G(x)$. Then integrating (VII-82) from 0 to p we get

$$U(p) - U(0) \;=\; U'(c)\int_0^p G(x)\; dx. \tag{VII-83}$$

We know that $U(0) = 0$. The pesky constant $U'(c)$ can be eliminated by noting that since $U(1) = 1$, we can set $p = 1$ in (VII-83) and solve for $U'(c)$. Then finally

$$U(p) \;=\; \int_0^p G(x)\; dx \Big/ \int_0^1 G(x)\; dx \tag{VII-84}$$

where

$$G(x) \;=\; \exp\left[-2\int_c^x M(y)/V(y)\; dy\right]. \tag{VII-85}$$

This can be put into (VII-84) to get the solution we sought. Note that we have changed the variable of integration in (VII-85) to y to avoid having p appear in more than one context. If you are worried by the persistence of c, you may care to take time out to persuade yourself that it will introduce only a multiplicative constant into the expression for $G(x)$, so that as long as we use the same value of c in the $G(x)$ in both numerator and denominator of (VII-84), it will not matter what value of c we use.

A SPECIFIC CASE. All of which is all very well, but we would prefer to know the fixation probabilities in terms of population sizes and selection coefficients, not in terms of the rather mysterious $M(p)$ and $V(p)$. It remains to determine $M(p)$ and $V(p)$ in the particular case we are interested in. The case most easily solved is simple multiplicative selection, where $w_{AA} = (1+s)^2$, $w_{Aa} = 1+s$, and $w_{aa} = 1$. The expectation of Δp

is simply the deterministic change in Δp, since this is the process at work among the zygotes before sampling occurs, and sampling does not alter the gene frequency on average. Then from (II-42) we have

$$M(p) \;=\; \mathbb{E}(\Delta p) \;=\; \frac{sp(1-p)}{1+sp} \tag{VII-86}$$

As for $V(p)$, it is the mean of $(\Delta p)^2$. This latter is the sum of the variance of Δp and the square of its expectation. Now the variance comes entirely from the sampling. As we mentioned, the variance is not the same as what we would get by sampling $2N$ gametes from a pool with frequency p. For one thing we are sampling *after* selection, when the gene frequency has changed a bit. For another, there may be a lack of independence between the two genes sampled in one individual. The first effect is small if $\mathbb{E}(\Delta p)$ is small, which we have to assume to justify the diffusion approach. One can also show rather easily from (VII-71) that in the case of multiplicative fitnesses there is still independence of the presence or absence of A in the two genes of an individual. For other patterns of selection this is not true, but assuming it will give a good approximation as long as selection coefficients are small.

The upshot of this is that

$$V(p) \;\simeq\; p(1-p)/2N \;+\; M^2(p), \tag{VII-87}$$

but since $M^2(p)$ contains an s^2 and we are assuming s is small we will ignore this term. Then

$$V(p) \;\simeq\; p(1-p)/2N. \tag{VII-88}$$

We also approximate (VII-86) by

$$M(p) \;\simeq\; sp(1-p), \tag{VII-89}$$

since s is assumed small. Then for (VII-86) we have

$$2M(y)/V(y) \;\simeq\; 4Ns. \tag{VII-90}$$

Now it is easy to evaluate the fixation probability from (VII-84) and (VII-85). It turns out to be

$$U(p) \;\simeq\; \frac{1-e^{-4Nsp}}{1-e^{-4Ns}} \tag{VII-91}$$

NUMERICAL EXAMPLES. How good is the diffusion approximation in this case? Table 7.5 shows a comparison between exact numerical solution of (VII-73) and the diffusion approximation when $N = 10$ and with two values of s. When $s = 0.01$ the approximation is remarkably good. Considering that N is quite small, this is an amazing performance. When $s = 0.1$ the approximation is not doing quite so well, but is still far better than we

Table 7.5: Comparison of exact fixation probabilities for a Wright-Fisher model with diffusion approximation. Multiplicative selection. $N = 10$.

	$s = 0.01$			$s = 0.1$	
p	exact $U(p)$	approx.	p	exact $U(p)$	approx.
0.05	0.06002	0.06006	0.05	0.17873	0.18465
0.1	0.11885	0.11894	0.1	0.32602	0.33583
0.2	0.23305	0.23321	0.2	0.54756	0.56095
0.3	0.34279	0.34300	0.3	0.69830	0.71184
0.4	0.44825	0.44848	0.4	0.80100	0.81299
0.5	0.54959	0.54983	0.5	0.87107	0.88080
0.7	0.74056	0.74077	0.7	0.95166	0.95671

have any reason to expect. This is a fairly general property of the diffusion approximation. Even though it assumes that the gene frequency changes in myriad small jiggles, it does a remarkably good job of predicting what will happen even when the gene frequency is actually changing by a few large jumps. It is possible to improve the accuracy of the approximation by even more elaborate efforts, but this seems a waste of time since the biological conclusions are in no way altered by that increase in accuracy.

Let us turn to examining the implications of (VII-91). A question which immediately arises is: when will natural selection make any substantial impact on the probability that a new allele fixes? Figure 7.3 shows the fixation probability $U(p)$ plotted as a function of p for various values of $4Ns$. When $4Ns = 0$, the fixation probability is the same as the initial gene frequency. We saw this result in section VI.5 above. It can also be obtained from (VII-91) by taking the limit of $U(p)$ as $s \to 0$. When $s = 0$ both the numerator and denominator of (VII-91) are zero, but we can use L'Hôpital's Theorem to obtain the limit

$$U(p) \;=\; \frac{4Np}{4N} \;=\; p \qquad \text{(VII-92)}$$

which is the same as the exact result obtained from the Wright-Fisher model. When $4Ns > 0$, the fixation probability is increased by natural selection (which is hardly surprising). When $4Ns = 0.01$ there is a rather small effect of selection, but when $4Ns = 100$ it is dramatic. We can tentatively conclude that $4Ns = 1$ is a reasonable value at which to recognize selection as beginning to have a significant impact. Although our experience is limited to one case as yet, we make so bold as to state this as a general principle. It is interesting to examine how many individuals are dying as a result of natural selection when $4Ns = 1$. If the population consisted entirely of the less fit genotype, we note that its fitness is a fraction $1/(1+s)^2 \simeq 1 - 2s$ of the fitness of the most fit genotype. We can say rather hazily that the amount of selection $4Ns = 1$ (so that $s = 1/(4N)$) would be

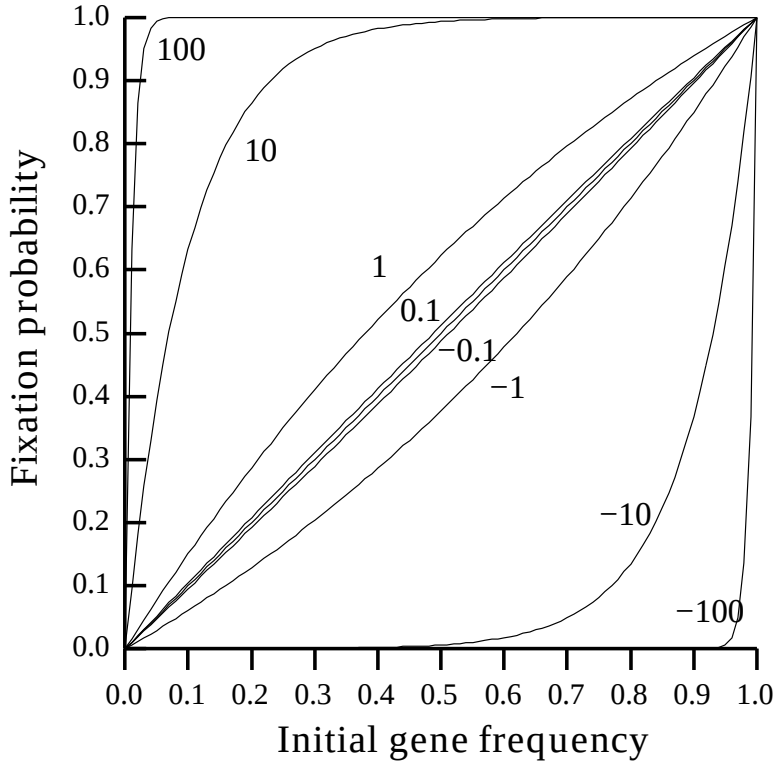


Figure 7.3: Probability of fixation of an allele with multiplicative fitnesses. Results from the diffusion approximation for various values of $4Ns$ and p are shown. The values of $4Ns$ are shown next to the nine curves, except for the diagonal, which has $4Ns = 0$.

equivalent to the death or sterility, from genetic causes, of $2sN = 2(1/(4N))N = 1/2$ of an individual per generation. So we can state our conclusion:

> Natural selection will be effective in the face of genetic drift if at the locus at least one individual every two generations dies or becomes sterile from genetic causes.

This is hardly a precise quantitative rule but certainly can be used to give us a rapid idea of whether selection will be effective. If we knew, for example, that there were 10,000 animals in a population, and that a certain locus has selection coefficients of about 0.01, then simply by observing that $4Ns = 400$ we know that genetic drift will be so weak an effect that natural selection would make a dramatic impact on gene frequencies in the long run. This strength of selection could be thought of as being equivalent to the death

of $(2s)N = (0.02)(10,000) = 200$ individuals per generation if all were of the inferior genotype.

DEPARTURES FROM THE WRIGHT-FISHER MODEL: EFFECTIVE POPULATION NUMBER. We have been using here the population size N as a measure of the strength of genetic drift. It is natural to ask how things are altered if we do not have a pure Wright-Fisher model of population reproduction. Suppose that any or all of the forces discussed in Chapter VII alter the effective population size (for example unequal numbers of the two sexes, prohibition of selfing, monogamy, varying population size, variability of offspring number, or fitness variation at nearby genetic loci). Can we correct for this in some way? The key to coping with these complexities is to notice that in the diffusion approximation, the population size enters in only through its effect in determining the variance of gene frequencies introduced in each generation by genetic drift. The relevant quantity is $V(p)$, the relevant equation is (VII-87). Now recall that when we discussed effective population number, we said that there is usually little distinction between the inbreeding effective number and the variance effective number. The latter is simply that value N_e which gives the correct variance of gene frequencies created by genetic drift in one generation when we predict the variance to be $p(1-p)/(2N_e)$. In most of the cases we treated, the two different definitions of the effective number have the same value, so that we can employ the effective numbers computed from inbreeding considerations to predict variance of gene frequencies. The result of this chain of reasoning is straight-forward: usually we can simply substitute the effective population size N_e for N and obtain as our criterion for the effectiveness of natural selection $4N_es > 1$. Thus in our hypothetical animal population of 10,000 individuals, if departures from a Wright-Fisher model reduced the effective population size to 3,000 individuals, we still have $4N_es = 120$ which implies that selection will be effective, though not quite so effective as would be implied by our computation that it is as if 200 individuals per generation are dying or being made sterile as a result of natural selection.

SELECTION AGAINST A MUTANT. We may also inquire into the effects of selection against an allele on its chances of fixation. Figure 7.3 shows several such cases. It is clear that as $4Ns$ falls below -1 selection starts to have a substantially reduced chance of fixation, reduced far below the initial frequency of the allele. For example an allele with an initial frequency of $p = 0.5$ and $4Ns = -100$ has only a chance of 1.92875×10^{-22} of being fixed! Clearly the rule that $4Ns \gg 1$ implies strong effects of selection works in this context as well: if s is negative we need only focus our attention on the other allele, which will be favored by selection, and ask when natural selection makes that allele likely to fix, which must also tell us when our original allele is likely to be lost.

ACCURACY OF THE BRANCHING PROCESS APPROXIMATION. Of course, even if natural selection is predicted to be effective, genetic drift will still have an effect. This

effect will loom larger the lower is the initial frequency of the allele. In Figure 7.3 all curves $U(p)$ drop towards zero as p approaches zero. It is relevant to ask whether this behavior is essentially that found in the branching process approximation. Does a single new mutant have a probability of fixation near $2s$, or has the finiteness of N dramatically altered the prospects for fixation? For what size populations will we obtain roughly correct results from the branching process approximation? To check this we can look at formula (VII-91) when $p = 1/(2N)$, so that we start with only one copy of the allele in the population. Then

$$U\left(\frac{1}{2N}\right) \;=\; \frac{1-e^{-2s}}{1-e^{-4Ns}} \tag{VII-93}$$

When $4Ns$ is large, the denominator is nearly 1, even for $4Ns$ as small as 10. Then

$$U\left(\frac{1}{2N}\right) \;\simeq\; 1-e^{-2s}. \tag{VII-94}$$

But since $e^{-x} \simeq 1-x$ for small x, we have

$$U\left(\frac{1}{2N}\right) \;\simeq\; 2s. \tag{VII-95}$$

So we can obtain this approximation to the branching process result by assuming $4Ns > 10$ and s small. In fact a comparison of the fixation probabilities obtained from (VII-94) with the exact results of the branching process in Table 7.3 shows that it is an even better approximation to the branching process results than is (VII-95). For example, with $s = 0.01$ the diffusion approximation is 0.01980, compared to 0.01973 for the exact Wright-Fisher model, 0.01922 for the approximation $2s/(1+s)^2$, and 0.02 for the approximation $2s$. So the branching process becomes relevant for large values of N, but only requires $4Ns$ to be 10 or more. Note that the branching process predicts no chance of fixation at all when s is negative. This is certainly close to the diffusion equation prediction if $4Ns < -10$, as Figure 7.3 will testify.

WEAK SELECTION: AN INTUITIVE RESULT. The process of interaction of genetic drift with natural selection must surely seem the hardest process to intuit in population genetics. There is, however, a case where we can get some insight. This is when $4Ns$ is small, so that selection has relatively little effect. If we expand the e^{-4Nsp} in the numerator of (VII-91) and the e^{-4Ns} in the denominator both as power series in $4Ns$, and take terms up to $(4Ns)^2$, dropping the rest by assumed smallness of $4Ns$, we get

$$\begin{aligned} U(p) \;&\simeq\; \frac{1-\left(1-4Nsp+8(Nsp)^2\right)}{1-\left(1-4Ns+8(Ns)^2\right)} \\ &\simeq\; p(1-2Nsp)/(1-2Ns). \end{aligned} \tag{VII-96}$$

We can approximate this when $4Ns$ (and hence $2Ns$) is small by

$$U(p) \simeq p + 2Nsp(1-p). \tag{VII-97}$$

This result can be justified by intuition in the following way. Clearly the first term, p, is simply the fixation probability when there is only genetic drift. The second term is $2N$ times (approximately) the expected change of gene frequency by natural selection in one generation, starting at the initial gene frequency. It is as if genetic drift is acting by first allowing the gene frequency to change deterministically by selection for about $2N$ generations, then suddenly fixing the population in a burst of activity.

Note that to good approximation the expected change of gene frequency by selection is $sp(1-p)$, where p is the current gene frequency in the population. Since the heterozygosity in the population is $2p(1-p)$, this indicates that the change of gene frequency is expected to be proportional to the heterozygosity. Now if natural selection is having little effect on the gene frequencies, we expect heterozygosity to decline by a fraction $1/(2N)$ every generation as a result of genetic drift. Since there should be $(1-1/(2N))$ as much heterozygosity around after one generation of drift, there should be a correspondingly smaller amount of change of gene frequencies, averaged over all possible replicates of the process. The total response to natural selection should then be

$$sp(1-p)\left[1+\left(1-\frac{1}{2N}\right)+\left(1-\frac{1}{2N}\right)^2+\ldots\right].$$

The quantity in brackets is a simple geometric series which is easily summed, and we get $2Nsp(1-p)$. That will be the average over all replicates, of the amount by which selection changes the gene frequency. But recall from Section VI.5 that the average gene frequency over all replicates must be the fixation probability, if the average is taken after the process of genetic drift is complete. So the final result is that the fixation probability is increased by $2Nsp(1-p)$ above the initial gene frequency if drift and selection essentially occur independently of one another. This is precisely the result of equation (VII-97).

So we can think of genetic drift as simply having the effect of eroding the stock of genetic variability on which natural selection acts, with the resulting total amount of natural selection being equivalent to $2N$ generations of selection. Although this principle is valid only for weak selection ($4Ns < 1$) it can be used to obtain useful approximations in many complicated cases in which the diffusion equations cannot be solved (e.g. with multiple alleles or multiple loci).

One interesting subcase is when we have only a single mutant gene initially. When $p = 1/(2N)$, (VII-97) gives

$$U\left(\frac{1}{2N}\right) \simeq s + \frac{1}{2N}, \tag{VII-98}$$

ignoring terms in s/N, on the assumption that these will be small. Looking back at Table 7.5 we can see that the fixation probability is, when $p = 1/(2N)$, being increased by s over the initial frequency by the presence of selection. Once again we get the impression that, with weak selection, genetic drift and selection are working roughly independently of one another. Of course, (VII-98) is valid only when $4Ns$ is small. Otherwise we get a better approximation from the branching process result $U(1/(2N)) \simeq 2s$.

For an exhaustive (and exhausting) discussion of various approximation formulae that can replace (VII-98) with much greater accuracy, and preserving desirable properties of the process, you may wish to consult the paper by McCandlish, Epstein, and Plotkin (2014). Their paper treats the related case of a haploid Moran model.

DOMINANCE, RECESSIVENESS, AND OVERDOMINANCE. The case of multiplicative selection has provided us with the diffusion approximation (VII-91), which we have seen gives us some insight as to the domains of validity of other approximations. When we introduce dominance into the fitness scheme the result is less simple. Suppose that the fitnesses are

$$\begin{array}{ccc} AA & Aa & aa \\ 1+s & 1+hs & 1 \end{array}$$

so that h measures the dominance of A over a. By dropping terms in s^2 on the assumption that these are far smaller than terms in s, we find that

$$M(p) \;\simeq\; sp(1-p)[p(1-2h)+h] \tag{VII-99}$$

and when we use $V(p) \simeq p(1-p)/(2N_e)$, we find that after a bit of algebra we get from (VII-85) using $c = 0$,

$$G(x) \;=\; \exp[-2N_e s(1-2h)x^2 - 4N_e shx] \tag{VII-100}$$

This function has no explicit integral, so we must use numerical integration methods to obtain the integrals in equation (VII-84). Although we have no explicit solution for the function $U(p)$ we can investigate its numerical value in any specific case.

When $h \simeq 1$ so that the mutant allele A is dominant, and when p is small, then particularly when $4N_e s$ is large there is little difference from the multiplicative case. This reflects the importance of the early generations, when all that matters for the process of loss of the allele is the difference hs in fitness between Aa and aa, as the AA genotype is hardly ever present. To good approximation the fixation probability per copy will be twice the effective selection coefficient so that

$$U(p) \;\simeq\; 2hs \times 2Np \;=\; 4Nhsp \tag{VII-101}$$

for small p and large $4N_e hs$. For larger initial values of p the numerical values of $U(p)$ obtained reflect the lessening of the strength of selection as the gene frequency rises. But

since we are primarily interested in small initial values of p, the biological significance of the result can be most succinctly stated by saying that new mutants are effectively "screened" by the population on the basis of their fitness in heterozygotes.

Recessiveness. When $h = 0$ matters are more complicated. A new rare recessive allele will experience very little selection. Only when it rises to a high enough frequency as a result of genetic drift will selection begin to operate. We should expect that such an allele will have a very low chance of fixation. These expectations are borne out by the numerical evaluation of (VII-84). A useful approximation can be developed for $s > 0$ and $4N_e s$ large in the recessive case:

$$u\left(\frac{1}{2N}\right) \simeq \sqrt{\frac{2s}{N\pi}}, \tag{VII-102}$$

provided $N_e \simeq N$.

As N becomes large, this probability declines. For example, when $s = 0.01$ then a dominant allele will have fixation probability near 0.02 for any large value of N. But when the allele is recessive with the same value of s, the fixation probability is approximately 0.0025 when $N = 1000$ but drops to 0.00008 when $N = 10^6$. Clearly a totally recessive allele has very little chance of surviving long enough to drift to a high enough frequency that selection becomes effective in fixing it. If the allele is not completely recessive, the effect of the allele in the heterozygote becomes a far more significant force determining its chance of fixation, and for a single initial mutant formula (VII-101) will be appropriate.

Balancing selection. In both overdominant loci and the simplest forms of frequency-dependent balancing selection selection, diffusion approximations are the same. The function $M(p) = p(1-p)[(s+t) - t\,p]$ in both cases, where s and t are positive. This will have the effect that if $4Ns$ and $4Nt$ are large, the gene frequency will tend to move near the value which is the equilibrium of deterministic selection. There it can stay for a long time – for large values of $4Ns$ and $4Nt$ the process can stay near there for geologically long periods. Although a new mutant can be lost early on, once it reaches the vicinity of the equilibrium, after that the ultimate probability of it fixing does not depend on the initial gene frequency. There is no algebraic formula for the fixation probability, but it can be evaluated by numerical integration of the integrals in equation (VII-84). The fact that overdominance and frequency-dependent balancing selection have the same diffusion approximation means that it will be very hard to tell which of these two forms of balancing selection is acting by observing gene frequency values or gene frequency changes. More direct measurement of fitnesses at different gene frequencies would be needed.

HISTORY AND REFERENCES. The diffusion approach to examining evolution in finite populations was pioneered by Fisher (1922), but his equations contained an error. This was corrected, and the approach put on a sounder footing by Sewall Wright (1931) in a classic paper. Both papers were largely concerned with equilibrium distributions, which we discuss in the next section. Fisher (1930) did, however, concern himself with an equilibrium under a constant flux of mutations. This amounts to consideration of fixation probabilities, since it asks what will be the rate at which new mutations destined to be fixed occur, and this should be near $2NuU(1/(2N))$. Both Fisher (1930) and Wright (1931) obtained $2s/(1-e^{-4Ns})$ for the probability of fixation in the multiplicative case. But Fisher was primarily interested in the gene frequencies to be expected under such a flux of mutations. Wright (1938, 1942) treated cases of irreversible mutation and obtained an approximation to the probability of fixation of an advantageous recessive allele. Haldane (1927) had already obtained by branching process methods a similar approximation differing only by a factor of $\sqrt{2}$. Motoo Kimura obtained the present formula (VII-102) from a diffusion equation (1957). In 1962 he gave a solution to the general Kolmogorov Backward Equation for fixation probabilities (VII-84, VII-85) which is the basis for most contemporary work on fixation probabilities. In particular, Alan Robertson (1960) has used Kimura's results, as well as stating equation (VII-97), in an imaginative application of this approach to finding systems of artificial selection which maximize the probability of fixation of genes favorable to livestock productivity.

We will defer to the next section further discussion of the history of the diffusion methods, since the bulk of early work with these was concerned not with fixation probabilities but with equilibrium distributions of gene frequencies.

VII.9 Approximation to Equilibrium Distributions.

INTRODUCTION. When mutation continually re-introduces alleles into a population, or when migration continually brings them in from a population which itself remains unfixed, then there is no such thing as a probability of fixation. The very concept of fixation then exists as only a temporary state: a population may arrive at a state of fixation for one allele, but sooner or later new mutations or immigration will re-introduce other alleles and thus move the population out of its state of fixation. In these cases we are not dealing with the evolution of gene frequencies as a temporary phase which ends in a state of fixation. Instead gene frequencies continue changing back and forth indefinitely. One naturally wants to know what sort of gene frequencies will be found in a population under such circumstances. Sometimes it will be reasonable to assume that the population has existed at approximately its present size, under more or less the same environment, for a long enough time that we may consider it to be in equilibrium for most alleles. Of course this equilibrium assumption has its risks: the desk on which

this is written would have been under a mile of ice only 500 human generations ago!

Nevertheless the equilibrium distribution of gene frequencies is of great interest and provides much information. We shall first briefly examine the treatment by means of a Wright-Fisher model. As in the case of fixation probabilities, this yields only numerical solutions at best. But it is useful to help clarify the logic of the underlying process. As before, the treatment here will be confined to the case of two alleles in a diploid population.

THE WRIGHT-FISHER MODEL. We have already defined the diploid Wright-Fisher model with selection in the previous section. When mutation and migration are added in, nothing really changes. If mutation occurs in the gamete or in the zygote stage of the life cycle, and if natural selection occurs at the viability stage of the life cycle, and before N adults are randomly sampled to survive density-dependent population size regulation, then we can compute transition probabilities in much the same way as before. The transition probability $P(j|i)$, the probability that there will be j copies of the A allele next generation if there are i now among the $2N$ genes in adults, is obtained by the following process:

1. Compute the gene frequency $p = i/(2N)$ in the adults of the current generation.
2. The proportions of A and a among the infinite number of gametes produced by these adults will then be p and $1 - p$.
3. If a fraction u of the A's mutate to a and a fraction v of the a's mutate to A, after mutation the gamete frequency of A will be $p' = (1 - u)p + v(1 - p)$.
4. If a fraction m of the gametes are replaced by immigrant gametes whose gene frequency of A is fixed at p_I, then after immigration the gamete frequency of A gametes is altered to $p' = (1 - m)p + m\, p_I$.
5. Compute the genotype frequency among newly fertilized zygotes by assuming random combination of gametes, producing Hardy-Weinberg proportions at the new gene frequency p'.
6. Now apply the natural selection. Multiple the frequency of each genotype by its fitness, then divide each by the mean fitness. This gives the proportions of the genotypes among survivors of natural selection. Suppose that these are P, Q, and R respectively for AA, Aa, and aa.
7. The probability that among the N surviving adult zygotes there are k AA, ℓ Aa, and $N - k - \ell$ aa is the trinomial sampling probability

$$\frac{N!}{k!\ \ell!\ (N - k - \ell)!}\ P^k\ Q^\ell\ R^{N-k-\ell}$$

8. To find the probability that there are j A genes among the N adults, sum these probabilities over all combinations of k and ℓ that have $2k + \ell\ =\ j$.

When natural selection is multiplicative it is not hard to show that the process of sampling N adults is exactly the same as sampling $2N$ genes. When selection is not multiplicative, this is only approximately true, but it is a better approximation the weaker is the natural selection (for after all, no selection at all is a case of multiplicative fitnesses). We will make use of this later, as we did in the previous section.

These prescriptions for computing transition probabilities assume one particular kind of life cycle. Similar computations can be made with other life cycles, although in some of these (particularly with fertility differences among genotypes) we will no longer be able to summarize this whole Markov process by simply looking at the gene frequencies $i/(2N)$ and $j/(2N)$ in adults. In these other cases the probability of a given genetic composition in the next generation may depend not only on how many A genes there are but on whether they are concentrated in AA homozygotes or spread among Aa heterozygotes. This makes things more difficult. It is also possible to compute transition probabilities for multiple-allele Wright models, or even for multiple-locus Wright-Fisher models. If overlapping generations are preferred, it is even easier to compute the transition probabilities of a Moran model (see section VI.8), for there are fewer of them as only one individual dies during any time interval.

Once we have the transition probabilities, the equilibrium distribution of gene frequencies (actually, of gene numbers) is given by the solution of

$$f_j \;=\; \sum_{i\,=\,0}^{2N} f_i\, P(j|i), \qquad\qquad j \;=\; 0, 1, \ldots, 2N \qquad\qquad \text{(VII-103)}$$

where f_i is the equilibrium probability that there are i A genes in the population. If we let a population evolve long enough that it has lost all effects of its initial gene frequency, then f_i is the probability that its gene frequency is $i/(2N)$. There are actually only $2N$ equations in $2N + 1$ unknowns (the f_i) in equation (VII-103), as one equation is redundant. But once we add the requirement that the f_i must sum to one, the equations are determined and can be solved. At least, can be solve numerically.

The problem is that no case of any biological interest has an explicit algebraic solution. The set of equations also can be solved numerically in cases with rather small population sizes. When $N = 10$, there are 21 equations in 21 unknowns. When $N = 50$, 101 equations in 101 unknowns. Present-day computers can solve these equations for values of N in the thousands. If natural selection occurs by fertility differences, the cases which can be treated are even smaller, for then there are many more states of the process which must be distinguished for a given N. When $N = 10$, there are 66 different combinations of genotype frequencies possible, so that with fertility selection there will be 66 equations in 66 unknowns. When the states of the process correspond to genotype frequencies rather than gene frequencies, it will be difficult to rapidly solve numerically cases larger than $N = 14$.

These numerical calculations are a useful check on the accuracy of approximations,

but it is approximations which will provide us with insight into the interaction of evolutionary forces.

THE DIFFUSION APPROXIMATION. Fortunately, a good diffusion approximation can be developed for the equilibrium distribution. We will only sketch the derivation here, as it is fairly tedious. It uses much the same methods as in the previous section. We start from a version of the exact equation (VII-103). We replace f_i by a function $f(p)$ of the gene frequency, and replace $P(j|i)$ by the probability $P_p(\Delta p)$ of the change Δp in gene frequency given that the gene frequency before the change is p. So

$$f(p) \;=\; \sum_{\Delta p} f(p-\Delta p) P_{p-\Delta p}(\Delta p) \tag{VII-104}$$

where the sum is over all changes in gene frequency which could have resulted in the current gene frequency p. This is a set of equations like (VII-103). Now we approximate $f(p)$ by a continuous density function $\phi(p)$. Since $\phi(p)$ must be multiplied by the width of the gene frequency intervals to obtain the probability of being in the particular interval around p, let us take the width of the intervals to be a quantity δp. Note that it is not the same as Δp. So we replace $f(p)$ by $\phi(p)\Delta p$. So

$$\phi(p)\,\delta p \;\simeq\; \sum_{\Delta p} \phi(p-\Delta p) P_{p-\Delta p}(\Delta p)\,\delta p \tag{VII-105}$$

Leaving out δp, we can write the density function $\phi(p)$ as follows. We approximate each of the functions $\phi(p-\Delta p)$ and $P_{p-\Delta p}(\Delta p)$ by the first three terms of a Taylor series expansion around p, so that

$$\phi(p) \;\simeq\; \sum_{\Delta p} \left[\phi(p) - \Delta p\;\phi'(p) + \frac{(\Delta p)^2}{2}\,\phi''(p)\right]\left[P_p(\Delta p) - \Delta p\;P_p'(\Delta p) + \frac{(\Delta p)^2}{2}\,P_p''(\Delta p)\right], \tag{VII-106}$$

where primes denote derivatives with respect to p. Collecting terms, ignoring those containing $(\Delta p)^3$ or $(\Delta p)^4$ and making use as before of

$$\begin{aligned} M(p) \;&=\; \mathbb{E}(\Delta p) \;&=\; \sum_{\Delta p} P_p(\Delta p) \\ V(p) \;&=\; \mathbb{E}((\Delta p)^2) \;&=\; \sum_{\Delta p} P_p(\Delta p)(\Delta p)^2 \end{aligned} \tag{VII-107}$$

and

$$\sum_{\Delta p} P_p(\Delta p) \;=\; 1 \tag{VII-108}$$

as well as of

$$\begin{aligned} M'(p) &= dM(p)/dp &&= \sum_{\Delta p} P'_p(\Delta p)\Delta p \\ V'(p) &= dV(p)/dp &&= \sum_{\Delta p} P'_p(\Delta p)(\Delta p)^2 \\ V''(p) &= d^2V(p)/dp^2 &&= \sum_{\Delta p} P''_p(\Delta p)(\Delta p)^2 \end{aligned} \tag{VII-109}$$

we obtain after multiplying out the expressions in equation (VII-106) and using these relationships to replace terms:

$$0 \simeq -\frac{d}{dp}[M(p)\phi(p)] + \frac{1}{2}\frac{d^2}{dp^2}[V(p)\phi(p)] \tag{VII-110}$$

This is known as the Kolmogorov Forward Equation. The term "forward" comes from equation (VII-105) where as one moves from the right to the left-hand side of the equation one moves forward in time.

The solution of this equation to obtain ϕ in terms of M and V is a bit obscure. We can integrate (VII-110) once with respect to p to obtain

$$C \simeq -M(p)\phi(p) + \frac{1}{2}\frac{d}{dp}[V(p)\phi(p)]. \tag{VII-111}$$

The constant C is then found by imposition of a rather mysterious "zero probability flux" condition. This step (for which interested readers may consult Crow and Kimura, 1970, section 8.3 or Ewens, 2004, section 4.5) rules out the possibility that the density of gene frequencies on the (0,1) interval is maintained by a steady creation of mass at 0 and its flow across the interval and ultimate destruction at 1, or the reverse. In effect, it is a "conservation of matter" condition requiring that populations not be created or lost at the endpoints 0 or 1 of the gene frequency scale. C turns out to be 0.

With that done, equation (VII-111) is a simple, first-order linear differential equation. Its solution is fairly easy. If we define $f(p) = \phi(p)V(p)$, equation (VII-111) becomes

$$\frac{2M(p)}{V(p)}f(p) = \frac{df(p)}{dp}$$

and dividing by $f(p)$ it becomes

$$\frac{2M(p)}{V(p)} = \frac{1}{f(p)}\frac{df(p)}{dp} = \frac{dlnf(p)}{dp}.$$

Integrating that, the solution turns out to be

$$\phi(p) = \frac{K}{V(p)}\exp\left[2\int_c^p M(x)/V(x)\,dx\right]. \tag{VII-112}$$

The constant K is simply a scaling parameter that is fixed by the requirement that ϕ, being a density function, has area 1 between $p = 0$ and $p = 1$. The lower limit of integration c can be taken to be anything reasonable, as it is in effect part of the constant K.

The remainder of this section will be concerned with finding $M(p)$ and $V(p)$ in a number of cases of biological interest, using (VII-112) to find $\phi(p)$, and examining the shape of $\phi(p)$ to gain some insight into the simultaneous operation of multiple evolutionary forces. Solution of (VII-112) is relatively easy once $M(p)$ and $V(p)$ are known, as it involves only one integration. Readers who are intimidated by differential equations may be able to resume following the narrative here.

MUTATION AND DRIFT. we have two alleles with probability u of changing by mutation to a a, and probability v that each a will mutate to an it A, then from the consideration in chapter III we find that the deterministic change in gene frequency is

$$p_{t+1} \;=\; (1-u)\; p_t + v\;(1-p_t) \tag{VII-113}$$

so that

$$M(p) \;=\; \mathbb{E}(\Delta p) \;=\; -u\; p + v\;(1-p) \tag{VII-114}$$

and the quantity $V(p) = \mathbb{E}((\Delta p)^2)$ is also determined exactly, but we can approximate it by

$$V(p) \;\simeq\; \frac{p(1-p)}{2N_e} \tag{VII-115}$$

This involves (1) ignoring a term involving $M^2(p)$ on the grounds that squares and products of mutation rates such as u^2, v^2, and uv may safely be ignored in comparison to quantities like u and $1/N_e$, and (2) ignoring the fact that the p in the right side of (VII-115) should in reality be $p + \delta p$, on the grounds that since u is small we may ignore terms like u/N_e.

Now the computation of ϕ goes through straightforwardly:

$$\frac{2M(x)}{V(x)} \;=\; \frac{-2ux + 2v(1-x)}{x(1-x)/2N_e} \;=\; \frac{-4N_e u}{1-x} + \frac{4N_e v}{x}\;, \tag{VII-116}$$

so that

$$\int \frac{2M(x)}{V(x)}\,dx \;=\; 4N_e u\;\ln(1-x) \;+\; 4N_e v\;\ln x \tag{VII-117}$$

and

$$\phi(p) \;=\; \frac{K}{p(1-p)}\;\exp\left[4N_e u\ln(1-p) + 4N_e v\ln p\right] \tag{VII-118}$$

where we have absorbed a number of inconvenient constants into K which we shall leave undetermined. The final equilibrium density is

$$\phi(p) \;=\; Kp^{4N_e v-1}(1-p)^{4N_e u-1} \tag{VII-119}$$

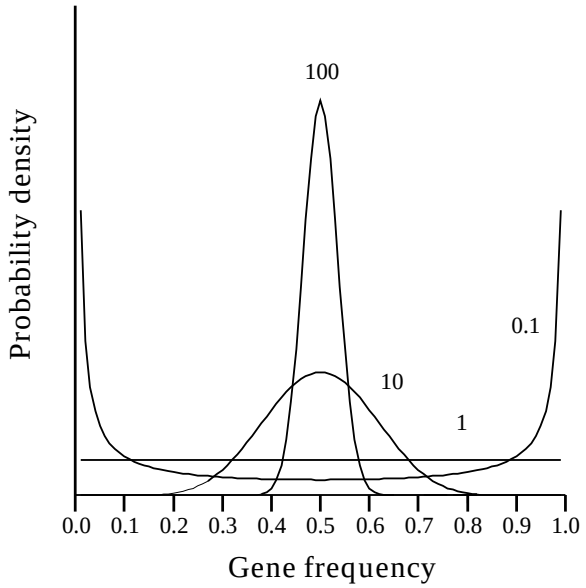


Figure 7.4: Equilibrium distribution of gene frequencies under mutation and genetic drift. In this case $4N_e u = 4N_e v$. The values of $4N_e u$ are shown next to the curves.

A set of curves for various values of $4N_e u$ and $4N_e v$ are given in Figures 7.4 and 7.5. These will allow us to obtain some insight into the behavior of mutation when interacting with genetic drift.

Numerical examples. Examine first Figure 7.4, where $4N_e u = 4N_e v$ (so that mutation is pressing the gene frequency toward an equilibrium at $p = 0.5$). When $4N_e u$ is large, the equilibrium density of gene frequencies is tightly clustered around the deterministic equilibrium. Clearly in these cases genetic drift can hardly ever move a population's gene frequency far from equilibrium before recurrent mutation pushes it back. When $4N_e u$ is small, the equilibrium distribution is U-shaped, with most of the mass concentrated near $p = 0$ or $p = 1$, with occasional movement from one tail of the curve to the other when a new mutant succeeds in spreading through the population. Remember that we are approximating a discrete histogram by a continuous density function, so that although the diffusion approximation never predicts a frequency exactly at 0 or 1, it has a certain fraction of the area under the curve so close to 0 or 1 that this proportion of replicates (or of generations) would be predicted to have gene frequency 0 or 1. Note that when $4N_e u = 4N_e v = 1$, the equilibrium distribution of gene frequencies is a flat rectangle with neither peak nor tails. This lends weight to our assertion that $4N_e u = 1$ is a rough dividing line between cases in which genetic drift overpowers

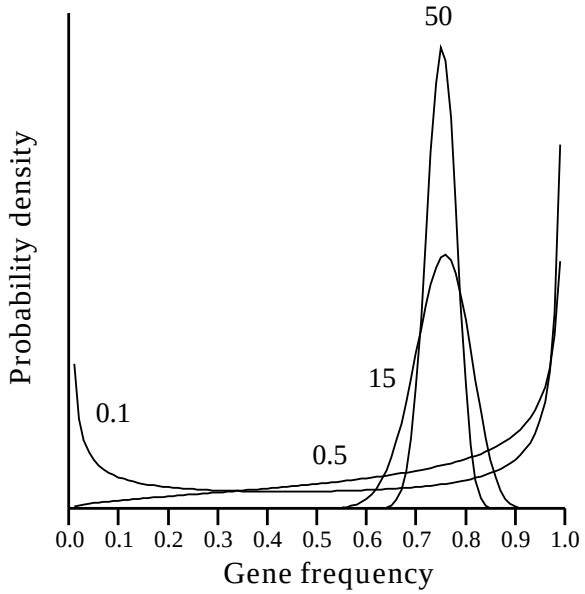


Figure 7.5: Equilibrium distribution of gene frequencies under mutation and genetic drift. In this case $4N_e u = 3(4N_e v)$. The values of $4N_e u$ are shown next to the curves.

mutation and cases in which mutation overpowers genetic drift.

Figure 7.5 shows cases in which $3u = v$, so that the deterministic equilibrium in an infinite population would live at $v/(u+v) = 0.75$. Again when $4N_e u$ and $4N_e v$ are large the gene frequencies lie near their deterministic mutational equilibrium. When $4N_e u$ and $4N_e v$ are both small, we again find a U-shaped distribution, only now with tails of unequal size, so that the average gene frequency over generations (or over replicates) will not be 0.5 (in fact it will be 0.75). We find that $4N_e u = 1$ and $4N_e v = 1$ are the points at which the tails of the distribution disappear. But now these are not both true at the same time, as witness the case where $4N_e u = 0.5$ and $4N_e v = 1.5$.

In fact, equation (VII-119) describes a well-known statistical distribution, the Beta distribution, whose properties were worked out long ago. The expectation of the distribution, which will also be its mean $\bar{p}$ over many independent replicates, is precisely

$$\mathbb{E}(p) \;=\; \frac{4N_e v}{4N_e u + 4N_e v} \;=\; \frac{v}{u+v} \qquad \text{(VII-120)}$$

which is precisely the mutational equilibrium gene frequency. This established that even when genetic drift carries the gene frequency far from its equilibrium value, the average gene frequency over generations or over replicates will still be the same. The variance of

the equilibrium gene frequency distribution is also well-known. It turns out to be

$$\text{Var}\,(p) \;=\; \frac{(4N_e u)(4N_e v)}{(4N_e u + 4N_e v)^2(4N_e u + 4N_e v + 1)} \tag{VII-121}$$

which can also be written

$$\text{Var}\,(p) \;=\; \frac{\bar{p}(1-\bar{p})}{1 + 4N_e u + 4N_e v} \tag{VII-122}$$

where $\bar{p}$ is the mean $\mathbb{E}(p)$. Since $\bar{p}(1-\bar{p})$ would be the variance of a set of populations which are all fixed, a fraction $\bar{p}$ of them for allele A, this equation for the variance is consistent with out picture that populations will be at or near fixation when both $4N_e u$ and $4N_e v$ are small and near $\bar{p}$ when both are large.

It can be seen that the results are very consistent with those that we obtained in section VII.2 by looking at the probabilities of identity of two alleles. This is far more than a coincidence: the two-allele model which gives equation (VII-8) is essentially identical to the present model. There is a close connection between probabilities of identity of alleles and variances of gene frequencies, so that we expect (and find) a good consistency between the variance of the gene frequency distribution and the probability of identity.

MIGRATION AND DRIFT. Migration could serve to maintain an equilibrium if migrants came from a mainland into an island, the gene frequencies on the mainland being (say) Q and $1-Q$ and never changing. In this case the change in gene frequency is easily computed:

$$M(p) \;=\; \mathbb{E}(\Delta p) \;=\; m\,Q + (1-m)\,p \;-\; p \;=\; m\,(Q-p). \tag{VII-123}$$

We also approximate the function $V(p)$ in this case, which will be valid if m is small and N large:

$$V(p) \;\simeq\; \frac{p(1-p)}{2N_e}. \tag{VII-124}$$

While it would be easy to go on to compute the equilibrium distribution from this, a shortcut is available. Like (VII-114), the formula for $M(p)$ is a simple linear function of p. In fact, by putting it into the form

$$M(p) \;=\; m\,Q \;-\; [m\,Q + m\,(1-Q)]\,p \;=\; -m\,(1-Q)\,p \;+\; m\,Q\,(1-p) \tag{VII-125}$$

we can see that it is really the same as (VII-114), provided that we substitute mQ for v and $m(1-Q)$ or u. The $V(p)$ function is also exactly the same as in that case. So we immediately know that the equilibrium distribution of gene frequencies is exactly the same as in the case of mutation vs. drift, save only that we make the substitution of $4N_e mQ$ for $4N_e v$ and $4N_e m(1-Q)$ for $4N_e u$. Thus there is a fairly exact analogy of

migration with mutation, at least when a single-locus island model is employed. The result is

$$\phi(p) \;=\; K\, p^{4N_e mQ-1}\,(1-p)^{4N_e m(1-Q)-1}, \tag{VII-126}$$

where K is, as usual, the constant that enables the area under the curve to be 1. Drawing directly on the results of the previous case, we find the mean and the variance of the equilibrium gene frequency to be:

$$\mathbb{E}(p) \;=\; \frac{4N_e mQ}{4N_e mQ + 4N_e m(1-Q)} \;=\; Q \tag{VII-127}$$

and

$$\mathrm{Var}\,(p) \;=\; \frac{\bar{p}(1-\bar{p})}{1+4N_e m} \tag{VII-128}$$

where $\bar{p} \;=\; \mathbb{E}(p) \;=\; Q$ as before. Note that we have perfect agreement of our results for the mean and variance of gene frequency in the one-island model: (VII-127) is the same as (VII-31) and (VII-128) is the same as (VII-40) in section VII.4 above.

In this case we can see the whole equilibrium distribution, not just the mean and variance. Its general properties are of course the same as in the mutation case: when $4N_e m$ is large, the island gene frequency is near Q, and when when $4N_e m$ is small, the island gene frequency is near 0 or 1. Figure 7.6 shows equilibrium distributions for a variety of values of $4N_e m$ when $Q = 0.4$.

SELECTION VERSUS DRIFT: A GENERAL FORMULA. When we have natural selection acting with two alleles and constant fitnesses, a relatively simple formula for the equilibrium distribution can be derived. Of course, if only genetic drift and natural selection are acting, there would be no equilibrium distribution, except in the extreme case where both homozygote genotypes are lethal. Barring that, drift would sooner or later result in fixation of one or the other allele, and the fixation probability calculations of the previous section would be more relevant than the equilibrium distribution. But in reality mutation is never absent, and so it is of interest to add mutation and see what equilibrium distribution is obtained in the presence of all three forces.

In section II.8 we saw that there was a simple expression for the change of gene frequencies under selection with two alleles, namely

$$\Delta p \;=\; \frac{p(1-p)}{2\bar{w}}\,\frac{d\bar{w}}{dp}. \tag{VII-129}$$

To compute the expectation of Δp, we must also take into account mutation. Alone, it would have

$$\Delta p \;=\; v - (u+v)p. \tag{VII-130}$$

As we saw in Chapter III, we get a good approximation to the net result of selection and mutation by simply summing these two formulas. This amounts to ignoring the fact that

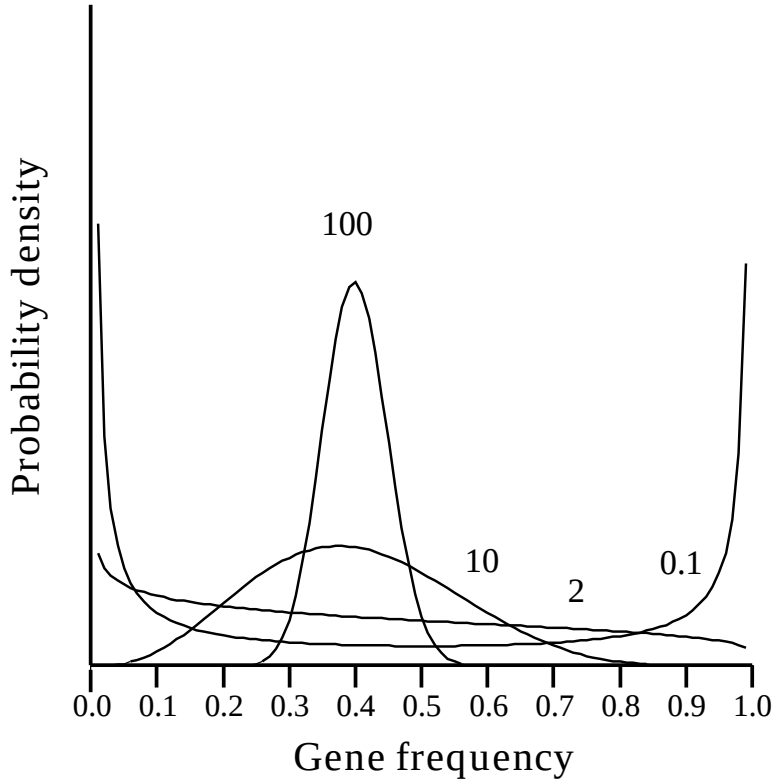


Figure 7.6: Equilibrium gene frequency distributions in a balance between migration and genetic drift, when there is a one-island model with gene frequency 0.4 on the continent. The values of $4N_em$ are shown next to the curves.

(if mutation precedes selection) the change due to selection has to be calculated based on the gene frequencies *after mutation*. It is a good approximation to ignore this if both selection coefficients and mutation rates are sufficiently small that we can ignore their product. So we make use of

$$M(p) \;=\; \frac{p(1-p)}{2\bar{w}}\frac{d\bar{w}}{dp} + v(1-p) - up. \qquad \text{(VII-131)}$$

As in the previous cases we discussed, we also ignore the slight effects of selection and mutation on $\mathbb{E}((\Delta p)^2)$ and use

$$V(p) \;=\; \frac{p(1-p)}{2N_e}. \qquad \text{(VII-132)}$$

Now we get easily

$$2\frac{M(x)}{V(x)} \;=\; \frac{2N_e}{\bar{w}}\frac{d\bar{w}}{dx} \;+\; \frac{4N_ev}{x} \;-\; \frac{4N_eu}{1-x}. \qquad \text{(VII-133)}$$

The derivative $(1/\bar{w})\, d\bar{w}/dx$ is the derivative of $\ln \bar{w}$, so that upon integrating

$$2 \int^{p} \frac{M(x)}{V(x)}\, dx \;=\; 2N_e \ln \bar{w} \;+\; 4N_e v \ln p \;+\; 4N_e u \ln(1-p) \tag{VII-134}$$

and we can plug this into equation (VII-112) and get the equilibrium distribution

$$\phi(p) \;=\; K\, p^{4N_e v - 1}\, (1-p)^{4N_e u - 1}\, \bar{w}^{2N_e}, \tag{VII-135}$$

(with K as usual the constant that enables the area under the curve to be 1). Keep in mind that $\bar{w}$ is itself a function of p. The effect of raising $\bar{w}$ to the $2N$-th power is to greatly exaggerate its peaks and valleys, the more so the larger is N_e. Thus the effect of a large population is to greatly increase the height of the equilibrium distribution's density function $\phi(p)$ in the neighborhood of the highest values of $\bar{w}$ and greatly decrease it elsewhere. Of course the mutation terms (the factors of p and $(1-p)$ have much the same effect, except that they attract the equilibrium distribution to the region of the mutational equilibrium. Which of these effects is more important will depend on the relative sizes of mutation rates and selection coefficients.

We saw in the diffusion approach to fixation probabilities that the strength of natural selection in the face of genetic drift depended essentially on one parameter, $4N_e s$. Formula (VII-135) does not at first sight seem to show this, as the selection coefficients are part of the formula for $\bar{w}$, while N_e is in its exponent. However the behavior is nearly the same if selection is not very strong. Take for example the case of an advantageous recessive allele.

$$\bar{w} \;=\; 1 + sp^2 \tag{VII-136}$$

so that

$$\bar{w}^{2N_e} \;\simeq\; (1+sp^2)^{2N_e} \tag{VII-137}$$

but if s is small

$$1 + sp^2 \;\simeq\; e^{sp^2} \tag{VII-138}$$

so that

$$\bar{w}^{2N_e} \;\simeq\; e^{2N_e\, sp^2} \tag{VII-139}$$

a formula that depends only on the product of the selection coefficient and the population size. Other selection schemes will show similar behavior when selection is not very strong. This is the case in the derivation of these equilibria, since in effect we take selection coefficients and mutation rates very small while taking N very large, in such a way that products like Ns and Nu remain constant.

SOME NUMERICAL EXAMPLES OF SELECTION, MUTATION AND DRIFT. With this general formula in hand it is easy to generate equilibrium distributions. In fact, it is fairly easy to intuit the shape of the distributions without any calculation! One of the reasons of looking at such distributions is to hone one's intuition, and I hope that readers will make some attempt to treat the examples in this section in that way.

Mutation versus selection. The first set of cases we will examine involves the equilibrium between mutation and selection, where an allele is straightforwardly deleterious. We have mutation at equal rates between the two alleles ($u = v$), and the selection scheme

$$\begin{array}{ccc} AA & Aa & aa \\ (1+s)^2 & 1+s & 1 \end{array}$$

Figure 7.7 shows the equilibrium densities for 16 different combinations of $4Ns$ and $4Nu$. Computations were done with $N = 1000$. When $4Ns = 0.1$, which is the bottom row of the Figure, we see hardly any effect of selection. The distributions are nearly symmetrical around $p = 0.5$. As $4Nu$ increases the gene frequencies huddle more and more closely around their mutational equilibrium value. However there are some signs of asymmetry, particularly in the relative heights of the two tails of the distribution when $4Nu = 0.1$. When $4Ns = 1$ we start to see more signs of the effectiveness of selection. When $4Nu$ is small, the tails of the distribution are definitely asymmetric: the gene frequency spends more time near zero than near one, and this must be due to natural selection resisting genetic drift towards fixation but assisting genetic drift toward loss. As $4Nu$ increases, this asymmetry is still evident when $4Nu = 1$, but disappears as $4Nu$ becomes large since then mutation is a stronger force than selection, even from a deterministic point of view, since $u \gg s$. When $4Ns$ is large (the top two rows of the Figure), selection effects are evident. When $4Nu$ is small, selection creates a marked asymmetry in the sizes of the two tails of the distribution, to the point where one tail becomes so small that it cannot be seen in our Figures. When $4Nu$ is larger, there are few populations near fixation or loss. Most tend to cluster around the equilibrium gene frequency, which in a deterministic analysis of this case is u/s.

The effects of population size can most easily be seen by moving from the lower left to the upper right along the diagonal of this Figure. These are a series of cases in which $u/s = 0.1$, so that all have the same deterministic equilibrium gene frequency. The distribution goes from a U-shaped one which is nearly symmetric to a peak near the mutation-selection equilibrium. In the process it goes from one influenced mainly by the mutation-drift balance to one influenced mainly by the balance between selection and mutation. It is interesting to note that as N is made small, mutation becomes a more important force than selection in influencing the relative heights of the two tails of the distribution. This is paradoxical: if $4Ns$ is becoming small, so is $4Nu$, so that there seems no reason to expect mutation to become more important than selection. One resolution of the paradox is to note that while selection is strongest when gene frequencies are intermediate, mutation is most active when gene frequencies are extreme. Now as N is made small, genetic drift is becoming the dominant evolutionary force. As it does, gene frequencies spend more time near 0 or 1 and less at intermediate values, and it is this which makes mutation a more important determinant of gene frequency than selection.

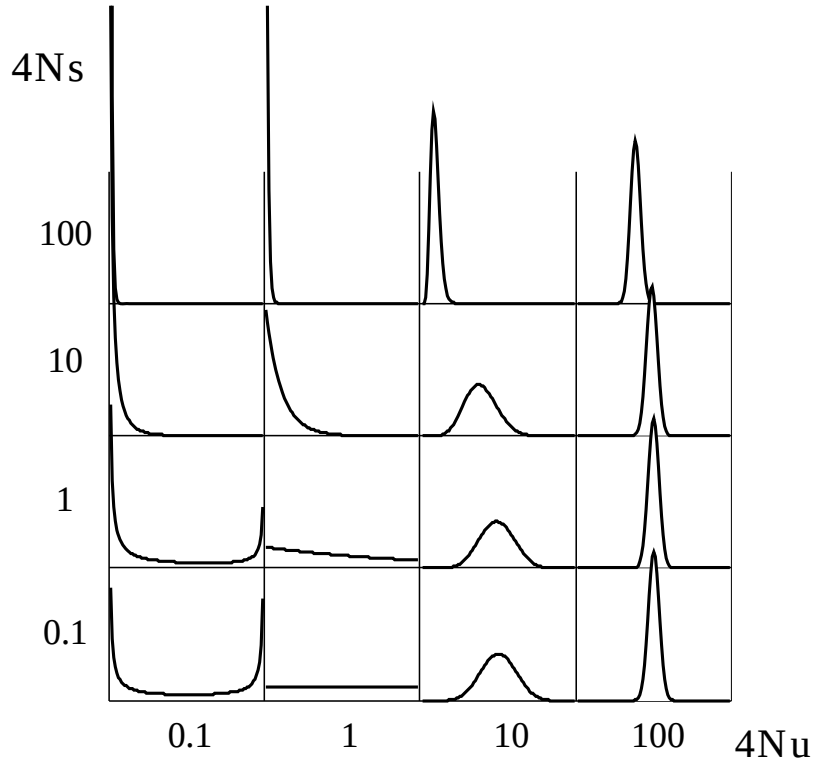


Figure 7.7: Equilibrium distribution of gene frequencies in a case of mutation with one allele deleterious, in a finite population, for different values of $4Nu$ and $4Ns$. Explanation is in text.

Balancing selection. The second set of examples is shown in Figure 7.8. These again involve $N = 1000$, and $u = v$, but now the selection is symmetric overdominant selection with fitnesses

$$\begin{array}{ccc} AA & Aa & aa \\ 1-s & 1 & 1-s. \end{array}$$

Again 16 combinations of $4Nu$ and $4Ns$ are shown. Now the curves are always symmetric about 0.5, since this is the equilibrium value for both the selection and mutation processes. Once again, there is little sign that selection is effective when $4Ns$ is small. Now, however, when $4Ns$ is large it creates a peak near $p = 0.5$, for the effect of selection is to pull gene frequencies toward that value. When $4Ns = 10$ and $4Nu = 0.1$, there is a peak in the center in addition to the tails near 0 and 1. As $4Ns$ is increased with $4Nu = 0.1$, the peak in the center increases in size and the tails shrink in size (although they may not be visible in the Figure, there still are tails on the distribution). This results

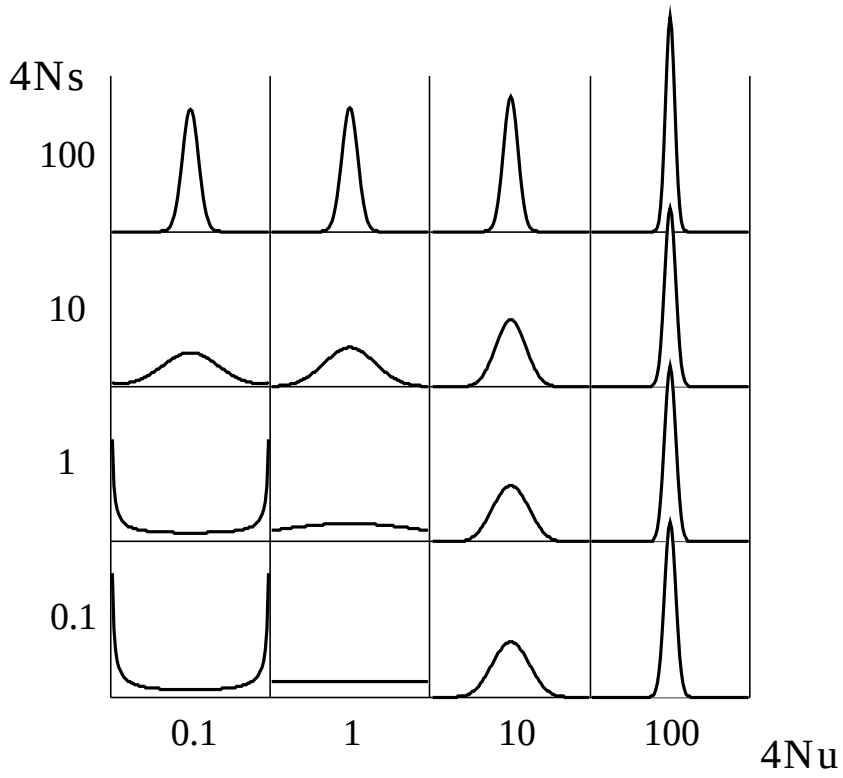


Figure 7.8: Equilibrium distributions of gene frequencies at an overdominant locus in a finite population with mutation present, for different values of $4Nu$ and $4Ns$. Explanation is in text.

from the relatively small fraction of generations which a population is expected to spend in the tails. Once or the other allele reaches fixation, then (if $4Nu = 0.1$) we expect one new mutant to occur every twenty generations. These mutants will be at a strong selective advantage when rare. It will not be long before one of them becomes established in the the population, and the gene frequency returns to 0.5. But once there, the large value of $4Ns$ means that it will stay in that vicinity far longer than it stayed near zero.

For larger values of $4Nu$ and $4Ns$, the two forces, selection and mutation, combine to create a peak at $p = 0.5$, and the peak is higher the larger are these two quantities.

HISTORY AND REFERENCES. The diffusion approach to finding equilibrium distributions of gene frequencies has a complex history. As already mentioned, R. A. Fisher (1922) was the first to attempt it. He made a transformation of scale, $y = \sin^{-1}\sqrt{p}$, and argued that the distribution of the quantity y would follow a particular differential equation, a variant of the heat equation of physics. He treated a number of cases in this way,

coming to this conclusion that selection must nearly always overwhelm the effects of genetic drift, a conclusion valid for the parameter values he assumed. Fisher had, however, an error in his logic. When Sewall Wright (1929c, 1931) treated the same phenomena by entirely different methods, it was discovered that there was a discrepancy in these results, and the error was discovered. Wright's methods differed from both Fisher's and from the ones we have used here. Wright was able to achieve an approximate solution of an integral equation derived from (VII-103). He gave both formulas and figures for a large number of cases, and an extensive discussion of their biological significance (extended and repeated many times elsewhere, e.g. Wright (1932)). While Fisher's priority may be argued, Wright's presentation seems to have had the predominant influence on subsequent papers, arriving at our formula (VII-135) as well as a multiple-allele generalization of it (1937).

In the meantime a formally different approach was under development by the great Russian probabilist A. N. Kolmogorov. He had published (1931) the first comprehensive treatment of diffusion processes, and arrived at the forward and backward equations cited above. These he applied to population genetics almost immediately (1935, 1938). It was subsequent to this work that Wright (1945) realized that his equilibrium distributions could be obtained from Kolmogorov's Forward Equation.

Later work. There is a large volume of work springing from these pioneering efforts, far too much to cover in these pages, and too extensive to cite adequately. Much attention has been focused on the transient behavior of diffusion approximations. The largest contributor has been Motoo Kimura. He has found the rate of approach to fixation or loss as well as the distribution of unfixed classes in two-allele cases with multiplicative selection (1955a), and also given the general time-dependent solution for the distribution of gene frequencies at two alleles when there is no mutation or selection (1955c). Kimura has also presented (1956b) an exact solution for a three-allele genetic drift diffusion process, and large-time asymptotic results for multiple alleles (1955b). Together with Tomoko Ohta he has also obtained diffusion equations for time to fixation of a new mutant (1969a, b). Ewens (1963a) and Watterson (1962) had previously presented results for the mean time to fixation. Nei (1968) has obtained information about the frequency of lethal chromosomes from a diffusion approach. Recently (Steinrücken, Wang, and Song, 2013) there has been more progress in obtaining transition probabilities for finite times for diffusion approximations of multi-allele diploid selection.

Other major work on the time-dependent behavior of diffusion approximations which are on their way to fixation has included Warren Ewens's (1963b) derivation of the "sojourn time" of a population at different gene frequencies, and Alan Robertson's (1962) detailed examination of the effect of overdominance in delaying fixation. Surprisingly, overdominance which has an extreme equilibrium gene frequency in infinite populations often accelerates rather than retards fixation in finite populations.

Infinite isoalleles model. Another line of work using diffusion approximations has involved predicting the numbers and frequencies of alleles present in an infinite isoalleles model, and the use of data from populations to test the adequacy of such a model. Ewens (1964) and Kimura and Crow (1964b) treated the distribution of gene frequencies of neutral isoalleles, and Ewens (1972) was able to construct a statistical test of the neutrality hypothesis by an ingenious conditioning procedure. In particular, he showed that the parameter $4Nu$ is best estimated from the number of alleles present in a sample drawn from the population, with the relative frequencies of the alleles in the sample adding nothing to the estimation! Unfortunately we rarely see either a truly isolated population or an infinite-isoalleles type of mutation in real cases, so the test has only occasionally been applied. Takeo Maruyama (1973, 1974) has found a different invariance in a two-allele model with geographic structure: under a neutral model when the total heterozygosity over all local populations is plotted against the average gene frequency over all local populations, the result is a rectangular distribution, independent of the population sizes or migration rates. This intriguing result cannot be extended to multiple alleles or to the infinite-isoalleles model.

Diffusion equations can also be used to treat random variation of selection coefficients, although care must be taken. The literature on this subject involves a certain amount of controversy (Kimura, 1954; Jensen and Pollak, 1969; Ohta, 1972; Gillespie, 1973; Jensen, 1973; Karlin and Levikson, 1974; Karlin and Lieberman, 1974).

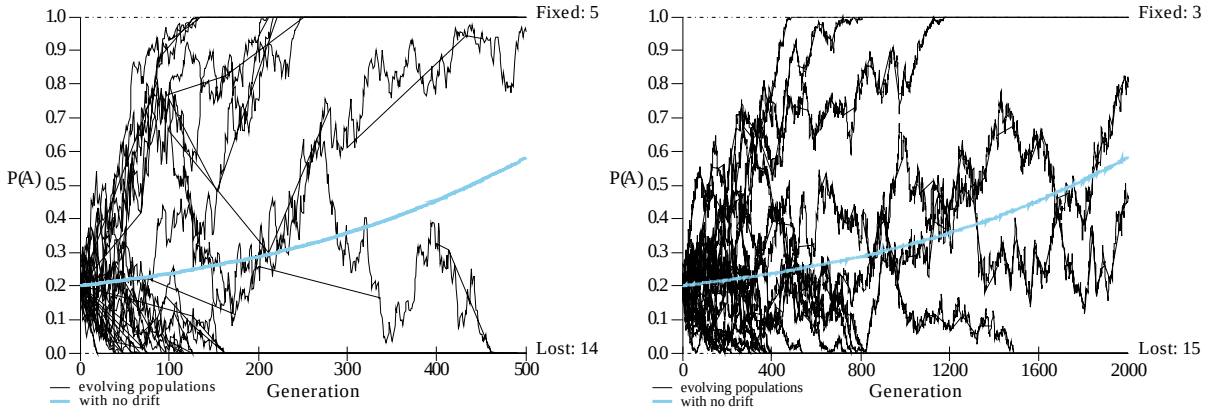
Rigorization. In the midst of all of these various lines of work applying diffusion methods, there have been a series of papers attempting to formalize the logical basis of the use of the diffusion approximations and obtain estimates of the error involved. Feller seems to have been drawn into his classical mathematical work on boundary conditions in diffusion approximations (1952, 1954) by his experience in formalizing the diffusion approximation. Feller (1951) and Karlin and McGregor (1964a) have taken the approach of showing that if we take a sequence of Wright-Fisher models (or Moran models) with increasing values of N but with the same $4Nu$ and $4Ns$, then an appropriately transformed time scale the process of gene frequency change becomes a diffusion process (in which gene frequency changes infinitely often in infinitely small jumps). Watterson (1962) has shown by this approach that the presence of diploidy and of two sexes do not cause serious trouble for our ability to approximate gene frequency change by a one-variable diffusion process involving only the overall gene frequency. Norman (1975) has plugged a gap in Watterson's proof. The limitation of all of these papers has been that the reliance on taking limits as $N \to \infty$ has introduced uncertainty as to whether the approximation is good for any particular population size N. Reassurance is available in a paper by Ethier and Norman (1977), who give bounds on the accuracy of the diffusion approximation in cases of a balance between mutation and genetic drift.

Many further references on diffusion approximations in population genetics may be found in Ewens's (2004) book, in Maruyama's (1977) monograph, or in W.-H. Li's (1977a)

In diffusion approximations, fixation probabilities starting from a given gene frequency and equilibrium distributions depend on the mutation rates, migration rates, and selection coefficients only through the products $4Nu$, $4Nm$, and $4Ns$. These distributions and probabilities are generally extremely accurate. Similar rules apply when there are multiple loci, and they also apply to rates of recombination, with $4Nr$ playing the same role.

Suppose that we want to know the equilibrium distribution for a multi-locus case where there is no formula available. We could carry out a computer simulation of the case, but what if we want to investigate a case where N is very large, for example if $N = 10^8$? This is where the diffusion scaling rules help. If we have $N = 10^8$, $4Nu = 2.5$, and (say) $4Ns = 10$, then $u = 6.25 \times 10^{-7}$ and $s = 2.5 \times 10^{-8}$. Both deterministic and random forces are expected to be very weak, and change to be slow. Simulation would be difficult. But we know that another case which has $4Nu = 2.5$ and $4Ns = 10$ should have nearly the same distribution. If we instead simulate a population that has $N = 1000$, $u = 0.000625$, and $s = 0.0025$, the higher values of u and s mean that the deterministic forces are stronger, and the smaller value of N means that simulations will be much faster. This simulation then stands in for a whole series of cases with the same values of $4Nu$ and $4Ns$, and we only need to simulate one of them.

These different cases will have similar probabilities of fixation and similar equilibrium distributions. Their time scales of gene frequency change will be different. A population of size 1000 will do in 1000 generations what a population of size 1,000,000 does in 1,000,000 generations, if they have the same values of $4Nu$, $4Nm$, $4Ns$ and $4Nr$. The figures below show this. Note the different horizontal time scales.



Simulations of 20 lines starting from gene frequency 0.2 of a recessive advantageous allele in populations with $4Ns = 4$. The left case has $N = 100$, the right case $N = 400$. The smooth curve shows deterministic change.

Box 3: The diffusion scaling rules

reprinting of many classic papers on diffusion approximations in population genetics.

Multiple alleles and multiple loci Only modest progress has been made in extending diffusion methods to loci with more than two alleles, or to multiple loci. For a particular "parent-independent" model of multiallele mutation, Li (1977b) has shown that the equilibrium distribution of the multiple alleles is a close analogue of equation (VII-135) above. Fernhead (2006) has argued that with multiple unlinked loci an analogous result can be obtained. Taylor (2008) has treated an island model with large numbers of islands exchanging migrants, and temporally-varying selection on those islands. If the number of islands is made large, this gives an overall gene frequency which is described by a single-population diffusion equation. Beyond these interesting cases, there has not yet been progress.

VII.10 The Relative Strength of Evolutionary Forces

The rules presented in this chapter for deciding which evolutionary forces will prevail in determining means and variances of gene frequencies, fixation probabilities, and equilibrium distributions, are simple and fairly general. They involve the comparison of the quantities $4N_e u$, $4N_e m$, and $4N_e s$ with each other and with 1. Each of the three evolutionary forces (mutation, migration, and selection) will be important in the face of genetic drift when the corresponding quantity exceeds 1.

The understanding gained from there quantities can be widened if it is realized that we can express them simply in terms of the relative rates at which the various processes change gene frequency. Mutation changes gene frequencies by about u per generation. Migration changes them by about m per generation. Selection changes gene frequencies by approximately s per generation. Correspondingly, we may think of natural "time scales" on which these processes affect gene frequencies. A substantial change in gene frequency will take $1/u$ generations if accomplished by mutation, $1/m$ if by migration, and $1/s$ if by natural selection. This picture of evolutionary forces is consistent with what we have covered in preceding chapters. The equilibrium gene frequency maintained by mutation in the face of natural selection is about u/s. This may be thought of as the amount of gene frequency that will accumulate by mutation during the time it takes for selection to substantially reduce gene frequency. Similarly, our consideration of patches and clines in Chapter IV persistently invoked the ratio m/s, which has a similar interpretation.

The rules obtained in this chapter involve a new force, random genetic drift. Its time scale is less self-evident. Gene frequency changes in any one generation are about $\sqrt{p(1-p)/(2N)}$ in magnitude. But to some extent they cancel each other, so we cannot take take $\sqrt{1/(8N)}$ as a measure of their size. (The 8 comes from the fact that $p(1-p)$ is near $1/4$ for a wide range of values of p). In fact, the time scale for genetic drift is

about $4N_e$ generations. A new mutant takes an average of about that many generations to complete fixation if it is destined to be fixed, and in $4N_e$ generations over 85% of a population's initial heterozygosity is expected to be lost by genetic drift. Thus, if we ask how far genetic drift will be able to change gene frequency during a process of (say) mutation, it will be substantial only if $1/(4N_e) > u$, that is to say, if $4N_e u < 1$. For migration we must compare $1/(4N_e)$ to m, getting $4N_e m < 1$, and for natural selection we compare $1/(4N_e)$ to s, getting as our condition $4N_e s < 1$, Thus the rules involving genetic drift are consistent with those involving only the deterministic forces, provided that we take the time scale of genetic drift to be about $4N_e$ generations.

We have encountered only a few exceptions to this picture. When an allele is rare and completely recessive, the effective amount of selection acting on it is very small, as homozygotes are rarely formed. So the amount of selection is far less than simple consideration of the selection coefficient s would suggest. This considerably alters the simple picture of evolutionary forces which we have been presenting. It is difficult to find a simple rationalization for equation (VII-102) in terms of these time scales, for example.

Another exception was seen when we discussed Figure 7.8, involving three forces, mutation, selection, and genetic drift. The increase of gene frequency by mutation is nearly u when the gene is rare, but the decrease of the gene frequency by selection is about sp in that case. In those cases where $4N_e u < 1$ so that the population would usually have p at or near 0 or 1, we saw a relative weakening of the effect of selection relative to mutation, since, unlike mutation, selection is most effective at intermediate gene frequencies. Thus when N is small, mutation can be more important than selection even though $s \gg u$. An interesting perspective on this behavior is given by Rouzine et al. (2001), who prefer to think of there being three regions of different behavior rather than the two I have discussed.

Aside from these caveats, we will not go far wrong by concluding that even though the mathematics of the processes are complicated, these different forces of evolution do not show subtle or complex interactions. They can be teased apart by a little work and a modicum of intuition.

Exercises

1. Suppose that in a population one allele has frequency $1/2$ and n others have frequency $1/(2n)$. What is the effective number of alleles present, expressed in terms of n?

2. Whooping cranes (*Grus americana*) have a population size of about 350 at present. If they maintain an effective population size this large for a great length of time, how many new mutants will occur per generation if $u = 10^{-7}$? What will be the

long-term effective number of alleles if these are all neutral isoalleles? How many generations will they take to approach this equilibrium level of variability?

3. Under an infinite isoalleles model, by how many allele substitutions will a given protein differ between two species if they have been separate populations for 100 million years, if a generation one year, mutation rate is 10^{-7} per year and all mutations can be detected? By how much will this quantity vary? Why? (Assume that we have sampled one sequence from each of the two species).

4. Suppose that an island on which reproduction follows a Wright-Fisher model receives immigrant gametes from two neighboring islands. The first of these provides 0.01 of the gametes and remains fixed for A. The second provides 0.02 of the gametes and remains fixed for a. The island we are concerned with maintains a population size of $N = 1000$ organisms. Compute the mean and standard deviation of the gene frequency when an equilibrium state has been reached.

5. Suppose two populations start each with two alleles, and they both happen to drift to fixation for the same allele. What will be Cavalli-Sforza & Edwards' genetic distance between the two resulting populations? What will be Nei's genetic distance? What if they had happened to drift to fixation for different alleles?

6. Suppose that two islands, each of size N, exchange migrants, the migration rate being m. If there is infinite isoallele mutation going on, we should be able to use (VII-45). What value of m corresponds to complete random union of gametes across the pair of islands? Using this value, do we find an expression for F_W that is the same as in a one-population infinite isoalleles model with $2N$ individuals? What about the comparison with one-population models when $m = 0$?

7. If we take $m = 1$ in a one-dimensional stepping stone model, does this correspond to complete random mating along the whole chain of populations (i.e., to the absence of geographical structure)?

8. Kimura and Maruyama (1971) pointed out that an infinite isoalleles model of neutral mutation can give a pattern of gene frequency looking very much like a migration-selection smooth cline of gene frequency. For what values of u, m, and N would a smooth cline of gene frequency across the whole species range be a likely observation in a one-dimensional habitat? In a two-dimensional habitat?

9. Use Kimura's formula for the fixation probability of an allele that is deleterious and has multiplicative fitnesses, so that the fitnesses for AA, Aa, and aa are $(1-s)^2 : 1-s : 1$. If the deleterious allele has initial frequency 0.9 and if $s = 0.1$, in a population of size 10,000 what is the probability that the deleterious allele fixes in the population (reaches frequency 1)?

How small does the deleterious effect have to be to allow the deleterious allele A to have a probability 0.5 of being fixed? Does this result surprise you – is natural selection being more or less effective than you expected?

10. Suppose that we have an overdominant locus with the following fitnesses:

    Based on the branching process approximation, what is the probability of fixation of a single A mutant? Based on the Wright-Fisher model, what is it (think of gene and genotype frequencies at various points in the life cycle)? Why the discrepancy? What is the branching process result actually computing if not the fixation probability?

11. Suppose that a locus has relative fitnesses:

AA	Aa	aa
4	2	1

    Compute the fixation probability from the diffusion equation formula (VII-91) of A when it has initial gene frequency 0.1 in a population of 10 individuals. Then consider the fitnesses

AA	Aa	aa
$1/4$	$1/2$	1

    and compute the fixation probability of allele A when it has initial gene frequency 0.9 in a population of size 10. What would be the relationship between these two fixation probabilities? Is it satisfied? Why or why not?

12. For the underdominant locus

BB	Bb	bb
1	$1-s$	1

    sketch what you feel the equilibrium distribution of gene frequencies will look like for each of the following parameter combinations when $u = v$:

    (i) $N = 1000$, $s = 0.01$, $u = 0.0001$

    (ii) $N = 1000$, $s = 0.001$, $u = 0.01$

    (iii) $N = 1000$, $s = 0.001$, $u = 0.0001$

Complements/Problems

1. In the infinite isoalleles model with a single diploid population of size N, we note that the number of new mutants occurring per generation is not the constant number $2Nu$ but a random variable with this number as its mean. Is this an extra source of random variation, or is it already accounted for in the computations?

2. In the one-island model of the equilibrium between drift and migration, have we already taken into account the random variability in the number of immigrants per generation?

3. Derive a good approximation for the average total number of mutational events which have occurred in the lines leading from a gene in one individual in a population, and from another randomly chosen individual back to their common ancestor. (In other words, if each mutant occurs at a different nucleotide position in the DNA so that we can see a record of all mutants, by how many nucleotides will two randomly chosen gene copies from the same population differ?) Use the infinite isoalleles model for a single population.

4. Reconcile equations (VII-8) and (VII-122) with each other when $u = v$ in a two-allele case. Do this by using (VII-8) to compute the probability that a randomly chosen individual is homozygous, and noting that this can also be written $\mathbb{E}[p^2 + (1-p)^2]$ and expressed in terms of $\bar{p}$ and V_p. Are the two predictions the same?

5. How is Nei's genetic distance measure expected to behave if the two subpopulations diverge by genetic drift with no mutation or migration? Is it expected to be dependent on the initial gene frequencies? Assume that the population has been at equilibrium for given values of N_e and u and that upon the two populations separating, the effective population size changes to a new, smaller, value $N*$ and mutation is absent from that point on.

6. How is Nei's genetic distance measure expected to behave if there are two classes of loci with different mutation rates, with half of all loci in one class and half in the other? Will it rise linearly with time?

7. How will the genetic distance measures that measure $1 - h_w/h_b$ be affected if there is mutation according to an infinite isoallele model? Develop equations for the change of h_w and h_b and use these.

8. There is a very complete analogy between the island model with $n = \infty$ and the one-island model with immigration from a continent. Why? How complete is the analogy? Where does it break down?

9. Compute the temporal correlation between the gene frequency of allele A in a one-island model. Use (VII-31), (VII-33), and (VII-40) to compute $\mathbb{E}[p_{t+1}p_t]$ and to obtain the covariance of gene frequency in successive generations. How great an interval must there be between the two samples from a single population, in order that their gene frequencies effectively be independent of one another? Could the temporal correlation of variability be used to estimate m, or is it too dependent

on m only through the quantity Nm, so that we could not know m unless we also knew N?

10. How do we have to alter the formulas for F_B and F_W in the n-island model if the immigrants are drawn at random from all n populations instead of only from the other $n-1$ populations?

11. Using the approximate solutions to the n-island model, obtain an expression for $(F_W - F_B)/(1 - F_W)$, which is a measure of the genetic distance related to H_B/H_W, the relative heterozygosity between and within populations. Compare it to Nei's distance measure $-\ln(F_B/F_W)$. Which one is more sensitive to the population size? Find cases where the two give discordant answers as to how much genetic distance there is between islands.

12. Consider a locus undergoing infinite-isoallele neutral mutation in a population, but which is near a strongly-selected overdominant locus. The overdominant locus has two alleles at equal frequencies, and these frequencies do not change. The recombination fraction between the loci is r. Suppose that we consider the two groups of chromosomes which are defined by the two alleles at the selective locus. These can be treated as if they are separate populations. They are each of constant size, and they exchange genes at the unselected locus by the process of recombination. Compute the effective number of alleles maintained in this pair of populations, as a function of N, u, and r. By making the proper comparison, discover whether having an overdominant locus nearby affects the amount of variation maintained at a locus.

13. In a two-allele model with geographic structure (say an island model), consider a random variable which has the value 1 if a randomly chosen gene is A, 0 if it is a. If we pick a gene at random from another population in the same generation, what will be the correlation between the values? Express it in terms of F_W and F_B. (You need not find the values of F_W and F_B to do this, just give an expression in terms of them). Does this result look familiar?

14. Suppose that in a diploid population on an island, multiplicative selection favors allele a, with the selection coefficient "in favor of" allele A being $s < 0$. Migrants from a continent continually bring in A genes, for on the continent all individuals are AA. Suppose gametes immigrate and $2Nm \ll 1$, so that the fate of each immigrant gene is decided long before another arrives. What does the branching process approximation tell us about the ultimate fate of the island's genetic composition? Is this consistent with intuition? Why not?

15. In the case described in the preceding problem, use equation (VII-91) to find the probability that a given immigrant A gene succeeds in driving out the locally fa-

vored a allele. Obtain from this and from $2Nm$ the time until the successful allele finally arrives, given initially no A alleles on the island. What does this say biologically about the time that a patch of local adaptation can persist in the face of immigration?

16. We are often interested in cases where population sizes are rather large, so that a gene frequency will not move far from its equilibrium. Near the equilibrium many evolutionary forces can be well approximated by saying that in deterministic situations they are expected to multiply the deviation of gene frequency from its equilibrium by a factor c:

$$p' - p_e \simeq c\,(p - p_e).$$

The effect of genetic drift can be approximated by saying that it causes a variance of gene frequency equal to

$$\mathrm{Var}\,(\Delta p) = \frac{p_e(1-p_e)}{2N}.$$

Use these to obtain approximations for $M(p)$ and $V(p)$, and solve for the equilibrium distribution of gene frequency. How does it compare to a Normal (Gaussian) distribution with mean μ and standard deviation σ?

Calculate c and p_e for the case of an overdominant locus. What mean and variance of gene frequencies does this predict? For what parameter values does this approximation break down? How can we compare it to a true equilibrium distribution since we haven't assumed any mutation?

17. In the case of geometric fitnesses, set up the matrices for the exact equations (VII-73) for the fixation probability for the case of $N = 1$. Solve them. How does the solution compare numerically in this case of an outrageously small population to the diffusion approximation (VII-91) evaluated when $p_0 = \frac{1}{2}$?

18. A more precise approximation than equation (VII-89) is

$$M(p) = \frac{sp(1-p)}{1+sp}$$

and

$$V(p) = M(p)^2 + \frac{(p + M(p))(1 - p - M(p))}{2N}.$$

Do what you can to evaluate this integral exactly (by partial fractions?). Does it lead to a better approximation?

19. If we compare large populations with small populations, which will do a better job of eliminating deleterious mutations? (Compare the mean of the frequency of

a deleterious mutation when holding u and s constant, but changing N. You may have to use numerical integration).

20. Suppose that the fitnesses at a locus for AA, Aa, and aa are respectively $1-2s$: $1-s$: 1. If $s = 0.1$ and there is mutation from a to A at rate 10^{-3}, and back mutation from A to a at rate 10^{-4}, about what would be expect the equilibrium gene frequency to be for the deleterious allele A in an infinite population? You don't need to get the full solution, just a good approximation.

    In a finite population, how much will the frequency of the deleterious allele vary around this? You should use `PopG` and run 100 or 200 lines, each with this pattern of mutation and selection. Start with different initial gene frequencies and run until you get similar results. As the program does not show you the average gene frequencies of the lines, you can roughly tell how variable the outcome is by the fraction of lines that have lost A.

    Try different population sizes. How large a population size is necessary to have the gene frequencies stay close to the equilibrium gene frequency, as judged by their not having lost A? What are the values of $4Nu$ and $4Ns$ needed to have the gene frequency be close to the equilibrium value?

    How do these results compare to the results of using diffusion equations to compute the equilibrium distribution of gene frequencies? Does the rate of back mutation make much difference? Why or why not?

21. When advantageous mutations (of selective advantage s in heterozygotes) occur at rate u per locus a population of size N, will there be more advantageous mutations fixed per generation if N is larger? Is all of this effect due to there being more advantageous mutations occurring in a larger population?

Chapter VIII

MULTIPLE LINKED LOCI

VIII.1 Introduction

Until about 1960 most evolutionary genetic theory was single-locus. There had been study of the decay of linkage disequilibrium by Jennings (1917), Robbins (1918) and Geiringer (1944, 1945) in the absence of any selection or genetic drift, and Wright's 1935b examination of intermediate optima was a multiple-locus selection model, but no one before the 1950s made a serious attempt to work out the theory of multiple linked loci under natural selection; Wright's adaptive surface results implicitly assumed that all loci were continually in linkage equilibrium.

This changed with the papers by Kimura (1956b) and Lewontin and Kojima (1960). Kimura's treatment of a two-locus system under selection assumed overlapping generations, with the approximation that Hardy-Weinberg proportions were always maintained. Lewontin and Kojima's paper was a more exact treatment of the discrete-generations case, and as such has been the basis for the literature that followed. The fact that the subject was taken up after a delay of almost 40 years after Robbins's paper is probably due to computers becoming available at that time.

Most of the burst of work in the following 15 years was inspired by the hope of finding some simple reparameterization that would greatly facilitate generalizations about the outcome of selection on linked loci. One of the goals was to find out what function of fitnesses and recombination rates was being maximized by natural selection. If natural selection was not maximizing mean fitness, we might at least find out from that function how the details of the genetic system altered the outcome. As we shall see these hopes have not been realized. Nevertheless evolutionary genetic theory has gained considerable insight into the interaction of selection and linkage, and into the effects of genetic drift in systems of multiple linked loci.

We will start the story with the simplest case, that of two loci each with two alleles in a haploid, and consider different measures of linkage disequilibrium.

VIII.2 A Haploid 2-locus Model

If we have two alleles at each of two loci in a haploid organism, there are of course 4 possible haplotypes: AB, Ab, aB, and ab. For simplicity let is designate their haplotype frequencies x_1, x_2, x_3, and x_4 respectively. The gene frequencies of A and B are each the sums of two haplotype frequencies:

$$\begin{aligned} p_A &= x_1 + x_2 \\ p_B &= x_1 + x_3 \end{aligned} \tag{VIII-1}$$

and the usual measure of linkage disequilibrium is

$$D = x_1 - p_A\, p_B \tag{VIII-2}$$

which we have seen in Chapter I (equation I-50) can also be written as

$$D = x_1 x_4 - x_2 x_3 \tag{VIII-3}$$

In the absence of selection, mutation, migration, or genetic drift, and the presence of random mating, the linkage disequilibrium measure D is expected to decline by a fraction r each generation.

SELECTION WITH NO RECOMBINATION. If we have no recombination, there is a different measure of disequilibrium that behaves nicely in the presence of simple types of selection. This is the crossproduct ratio, which we will call R,

$$R = \frac{x_1 x_4}{x_2 x_3} \tag{VIII-4}$$

If there is linkage equilibrium ($D = 0$) then $x_1 x_4$ equals $x_2 x_3$, so that the crossproduct ratio $R = 1$. If there is positive linkage disequilibrium $R > 1$, and if there is negative linkage disequilibrium $R < 1$. However, beyond that, there is no straightforward relationship between these two measures. They are sensitive to different aspects of departure from linkage equilibrium.

To see why R is useful, note that in the absence of recombination each haplotype is in effect a separate clone, whose proportion in the population is affected by its own fitness as compared to the mean population fitness:

$$x_i' = x_i w_i / \bar{w} \tag{VIII-5}$$

where of course

$$\bar{w} = x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4. \tag{VIII-6}$$

Taking the crossproduct ratio for the x_i', the mean fitnesses cancel and we get a startlingly simple result:

$$\frac{x_1' x_4'}{x_2' x_3'} = \left(\frac{w_1 w_4}{w_2 w_3}\right) \frac{x_1 x_4}{x_2 x_3} \tag{VIII-7}$$

R is multiplied every generation by the crossproduct ratio of the fitnesses $w_1 w_4/(w_2 w_3)$. Its value in any future generation can be predicted with ease. Obviously this crossproduct ratio of fitnesses depends only on the relative fitnesses of haplotypes, not their absolute fitnesses, as the ratio causes any multiplier that affects all the four haplotype fitnesses to cancel.

What this means for the effects of different patterns of gene interaction is most easily seen by expressing the w_i differently. Suppose that we take w_4 to be 1, and express the fitnesses as:

$$\begin{array}{ll} AB & (1+s)(1+t)(1+E) \\ Ab & 1+s \\ aB & 1+t \\ ab & 1 \end{array} \tag{VIII-8}$$

You will find readily that with these fitnesses

$$\frac{w_1 w_4}{w_2 w_3} = 1 + E \tag{VIII-9}$$

So if E is zero, there will be no change in the crossproduct ratio R from selection. This means that if $R = 1$, so that $D = 0$, there will be no linkage disequilibrium created by natural selection. If we start in linkage equilibrium we will stay there forever.

Note that if only one of the loci is under selection (say A) we must have $t = 0$ and $E = 0$, since the presence of a B on the AB chromosome cannot cause its fitness to be different. So whenever only one locus is under selection, R will not change as a result of that selection. It is only when the fitnesses interact that selection creates or intensifies linkage disequilibrium.

Additive fitness. These equations point strongly toward a definition of interaction that measures it as departure from multiplication of fitnesses. Many people prefer to think of interaction as departure from additivity. If the fitnesses were perfectly additive, so that the fitness of AB in (VIII-8) were $1 + s + t$, then $E = -st$, so that R will decrease under selection and negative linkage disequilibrium would be generated. If we put the fitnesses on a log scale, it is on that scale that their additivity will correspond to having $E = 0$. Figure 8.1 shows numerical results for R and D for the case where we start with the initial frequencies of A and B both being 0.1, no initial linkage disequilibrium, the values of s and t being respectively 0.2 and 0.3, and either $E = 0.1$ or $E = -0.06$. Note

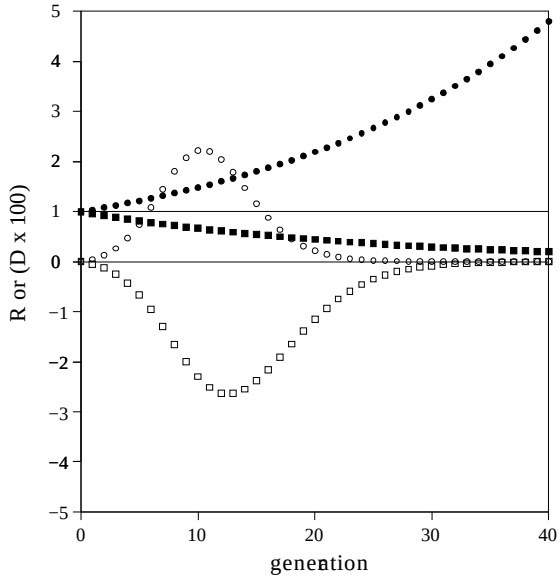


Figure 8.1: Results of iterating equations VIII-5 and observing the values of R and D. In all cases the initial frequencies of A and B were 0.1 without linkage disequilibrium between them. s was 0.2 and t was 0.3. The curves with circles are R and $100 \times D$ for the case where $E = 0.04$. The curves with squares are for $E = -0.04$, which is additive fitnesses. The solid lines are for $E = 0$. The black symbols are for R, and the open symbols are for D.

that R changes continually in the predicted direction, but that D moves away from zero and then back towards it as the alleles A and B move toward fixation.

This illustrates the different properties of the two measures of linkage disequilibrium. To see why they will be expected to differ, consider the case where A and B both have very low frequencies. If all the A's and all the B's are in AB haplotypes, then the value of D cannot exceed the frequency of that haplotype, which will be small. But the value of the crossproduct ratio R will be infinite.

We ought to add that in the case of no recombination the four haplotypes are in effect acting like four haploid alleles. We can work out their "allele" frequencies in all future generations. The ratio of any two of them (say x_1/x_2) gets multiplied by w_1/w_2 each generation (as in equation (II-108)). So after t generations the ratio will have been multiplied by $(w_1/w_2)^t$. It follows easily from this that the frequency of x_i in generation t, $x_i(t)$, is:

$$x_i(t) \;=\; x_i(0)\, w_i^t \;/\big(x_1(0)\, w_1^t \;+\; x_2(0)\, w_2^t \;+\; x_3(0)\, w_3^t \;+\; x_4(0)\, w_4^t\big)\,. \qquad \text{(VIII-10)}$$

The exact analogy to multiple alleles allows us to use equation (II-113) to show that mean

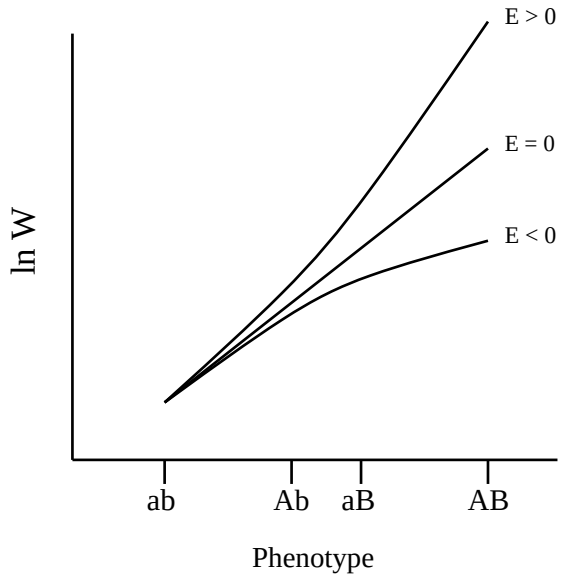


Figure 8.2: Three cases in which the logarithm of fitness is a function of an underlying phenotype, itself having additive effects of the alleles at the two loci. The top curve will have $E > 0$, the bottom $E < 0$, and the line in between will have $E = 0$.

fitness can never decrease during these changes.

EPISTASIS. The quantity E measures the strength of gene interaction, which is often called *epistasis*. Originally the term meant a particular type of interaction, in which one locus is said to be epistatic to another if it has a genotype that masks the effect of the other locus (as would be the case here if $E = -t/(1+t)$ or $E = -s/(1+s)$). But within evolutionary genetics, the term has since been used more broadly.

The meaning of the quantity E will be seen more clearly in a particular case. Suppose that there is a quantitative character controlled by the two loci A and B with no epistasis. Call the value of this character X. Now suppose that the logarithm of fitness ($\ln W$) is a nonlinear transformation of this character, as shown in Figure 8.2. There you will see three cases. In the top curve, which is curving upwards, $E > 0$. In the bottom one, which is curving downwards, $E < 0$. When the logs of the fitnesses are additive, the fitnesses themselves are multiplicative. This is the case with the straight line, which is when $E = 0$.

By an elementary application of the Mean Value Theorem from calculus, it can shown (Felsenstein, 1965) that in this case the sign of E will always be the same as the sign of the curvature of the curve relating $\ln W$ to X, so that E is positive when the curve is

curving upwards, negative when it curves downwards, and zero when it is a straight line. Thus in cases where epistasis results from such a transformation, we can often easily see whether selection will result in positive or negative disequilibrium.

SELECTION AND RECOMBINATION. If we now add recombination to the model, we must specify at what stage of the life cycle we observe the haplotype frequencies. If we have the life cycle:

$$\text{Newborns} \xrightarrow{\textit{selection}} \text{Survivors} \xrightarrow{\textit{random mating}} \text{Diploids} \xrightarrow{\textit{meiosis}} \text{Newborns}$$

then we can rather easily work out the equations for the x_i in successive generations using successively equations (VIII-5) (for the selection phase) and (I-47) for the decay of linkage disequilibrium following random mating. If we define

$$D^* = \left(\frac{x_1 w_1}{\bar{w}}\right)\left(\frac{x_4 w_4}{\bar{w}}\right) - \left(\frac{x_2 w_2}{\bar{w}}\right)\left(\frac{x_3 w_3}{\bar{w}}\right) \tag{VIII-11}$$

the linkage disequilibrium after selection has acted, we get

$$x_i' = \frac{x_i w_i}{\bar{w}} - k_i r D^*, \quad i = 1,2,3,4 \tag{VIII-12}$$

where k_i is a bookkeeping device to simplify the expressions:

$$\begin{aligned} k_1 &= k_4 = 1 \\ k_2 &= k_3 = -1. \end{aligned} \tag{VIII-13}$$

As an exercise, you should try to work out what these equations will be if we observe the population immediately after selection instead of immediately after meiosis.

It would be nice to go one from this point, as one can in the case of no recombination, and work out the future frequencies of all four haplotypes. In fact, we cannot. Certain special cases can be studied (such as $E = 0$ when there is initial linkage disequilibrium), but in general there are no analytical solutions for haplotype frequency dynamics. Figure 8.3 shows the changes in R in the case of Figure 8.2 in which $E = 0.06$, for three different levels of recombination. It is evident that the less recombination, the more linkage disequilibrium. It is clear from the figure that disequilibrium is greater when there is tight linkage, and less when there is loose linkage.

My paper (Felsenstein, 1965) and the papers of Maynard Smith (1968) and Eshel and Feldman (1970) should be consulted for further discussion of the effects of linkage on the rate of change of gene frequencies under directional selection.

INTERACTION AND LINKAGE – AN EXAMPLE. A particularly interesting case arises when there are negative values of s and t, but a strongly positive value of E. For

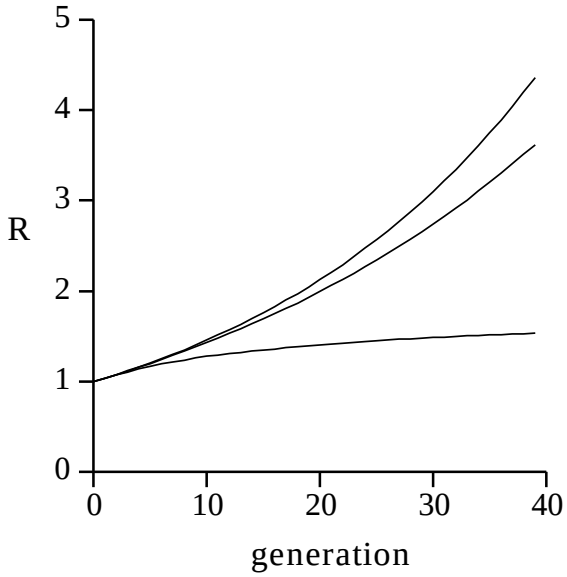


Figure 8.3: Course of change of the crossproduct ratio measure of linkage disequilibrium R for the case where initially there is linkage equilibrium, where initially p_A and p_B = 0.01, and where $s = 0.2$, $t = 0.3$, and $E = 0.06$. The three curves are for $r = 0$ (top), $r = 0.01$ (middle), and $r = 0.1$ (bottom).

simplicity suppose that $s = t < 0$ but E is positive enough that $(1+s)^2(1+E) > 1$. Then A and B are individually deleterious (compared to ab) but when they are combined, they are advantageous. This is a simplest genetic case expressing the dilemma of adaptations that must occur together to be advantageous. The question is, whether the genetic system will have the sense to evolve them.

If we start with a population of ab haplotypes, and introduce the two mutants in initial linkage equilibrium as rare alleles, then even though we cannot solve equations for the haplotype frequencies in an arbitrary generation, we can do so as long as they are rare. If we let $x_4 = 1 - x_1 - x_2 - x_3$ and write the equations (VIII-12) in terms of x_1, x_2 and x_3, after some algebra we find that if the fitness of AB is called $1 + u$, the equations can be written as:

$$\begin{aligned} x_1' &= x_1(1+u)(1-r) + O(x^2) \\ x_2' &= x_2(1+s) + x_1 r(1+u) + O(x^2) \end{aligned} \qquad \text{(VIII-14)}$$

where the term $O(x^2)$ stands for terms which are a square of one of the x_i or a product of two of them. These terms we will ignore when x_1, x_2, and x_3 are all small. We omit the equation for x_3 as it is the same as for x_2 in this case. Without the x^2 terms these

equations can be written as linear equations in matrix form:

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} (1+u)(1-r) & 0 \\ r(1+u) & 1+s \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \text{(VIII-15)}$$

x_1 and x_2 will increase when rare if and only if the leading eigenvalue of the matrix on the right-hand side of (VIII-15) is greater than 1. The matrix is triangular, which means its characteristic equation is easily found to be the product of the diagonal elements, after λ has been selected from each of them:

$$((1+u)(1-r)-\lambda)\,(1+s-\lambda) \;=\; 0. \qquad \text{(VIII-16)}$$

This has two roots, which are instantly recognizable as $(1+u)(1-r)$ and $1+s$, the diagonal elements of the matrix. Since s is negative, the latter one is always less than 1. The former is greater than 1 only when $r < u/(1+u)$. So, for example, when $u = 0.1$, the haplotypes AB and Ab (and therefore also aB will increase when rare if $r < 1/11$. The reader may wonder why Ab and aB are increasing, if selection is acting to decrease them. This increase happens because AB is increasing in frequency, and each generation generates some Ab and aB haplotypes by recombination with the ab haplotypes that make up most of the population.

The condition for the increase of AB is easily interpretable: AB haplotypes have their frequency multiplied by $1+u$ by selection and then a fraction r of their offspring turn out, due to recombination, not to be AB. So the condition $(1+u)(1-r) > 1$ simply is the condition for more than one AB offspring to result per AB parent. The analysis here assumes that we can ignore the product of the frequencies of Ab and aB compared to the frequency of AB. If the condition for the increase of AB is on the borderline (that is, if $(1+u)(1-r) = 1$ then these neglected second-order terms will become important and the analysis must be redone.

We can see that if Ab and aB are of low fitness and AB of high fitness, this double adaptation is able to increase when rare if the linkage between the genes involved is sufficiently tight. Once it reaches a substantial frequency, the loss of AB offspring due to recombination becomes less serious, as there are fewer ab haplotypes around to mate with, and the AB haplotype continues to increase toward fixation.

In a sense this provides us with a clearer picture of where the boundary is between having two separate adaptations (A and B) and having one dual adaptation (the haplotype AB). The condition of r acts like a filter, allowing establishment of only those interacting adaptations whose loci are closely enough linked. This will act as a mechanism for organizing the genome to place loci that interact closer to each other – simply because pairs of adaptations that interact but are not closely linked do not get established. The existence of evolutionary-genetic arguments like this one has long led geneticists to suspect that there must be some recognizable clustering of loci in genetic maps according to function. The empirical evidence for such a clustering is maddeningly poor, however.

That linkage is accomplishing something that will not happen otherwise is made clearer by considering what happens if there is no linkage disequilibrium. In that case each locus will show a gene frequency change whose direction can be predicted by considering the slope of mean fitness in the direction of increase in that gene frequency (for diploids we would consider equation II-116). The mean fitness of the population will be

$$\bar{w} \;=\; p_A p_B(1+u) \;+\; p_A(1-p_B)(1+s) \;+\; (1-p_A)p_B(1+s) \;+\; (1-p_A)(1-p_B). \tag{VIII-17}$$

Figure 8.4 shows a contour map of the fitnesses in the case where u = 0.1 and s = -0.1, plotted against the gene frequencies. The population is starting near the lower-left corner, with low initial frequencies of A and B. If it climbs the adaptive surface, it will not be able to increase the frequency of either the A or the B allele. Only if the allele frequencies start at high enough frequencies to be on the peak in the upper-right corner will they increase. If recombination is frequent, the population will be near (but not exactly at) linkage equilibrium and the picture in the Figure will predict the outcome with fair accuracy. But when linkage is tight the fitness of A and of B alleles is no longer well-predicted by the assumption that they are randomly associated into haplotypes. The presence of a A will then become associated with the presence of B. The fitness of A will be increased by the coupling linkage disequilibrium which makes A and B appear largely in AB; they are thus of higher fitness than a and b. This departure from the adaptive-surface picture will, as we will see, turn out to be a general phenomenon for closely-linked loci.

The conditions in the present case were worked out by Crow and Kimura (1965) in a paper on the evolution of recombination.

VIII.3 Linkage and Selection in Diploids

The original papers on linkage and selection by Kimura (1956) and Lewontin and Kojima (1960) used diploid models. We will develop here equations that are equivalent to Lewontin and Kojima's (1960) discrete-generations model. The result will be similar to equation (VIII-12) but a bit more complicated. One source of the complication is that the natural selection occurs at the same life stage where the recombination also occurs. This makes the form of the equations slightly different.

One again we have four gametes, AB, Ab, aB and ab, and number these 1, 2, 3, and 4. x_i is the frequency of gamete i in the pool of gametes that will, by random mating, be assembled into the diploids of the current generation. The frequency of a diploid composed of gametes i and j will be $x_i x_j$ among newborns. If natural selection occurs thereafter with fitness w_{ij}, the frequency of genotype ij after selection will be $x_i x_j w_{ij}/\bar{w}$,

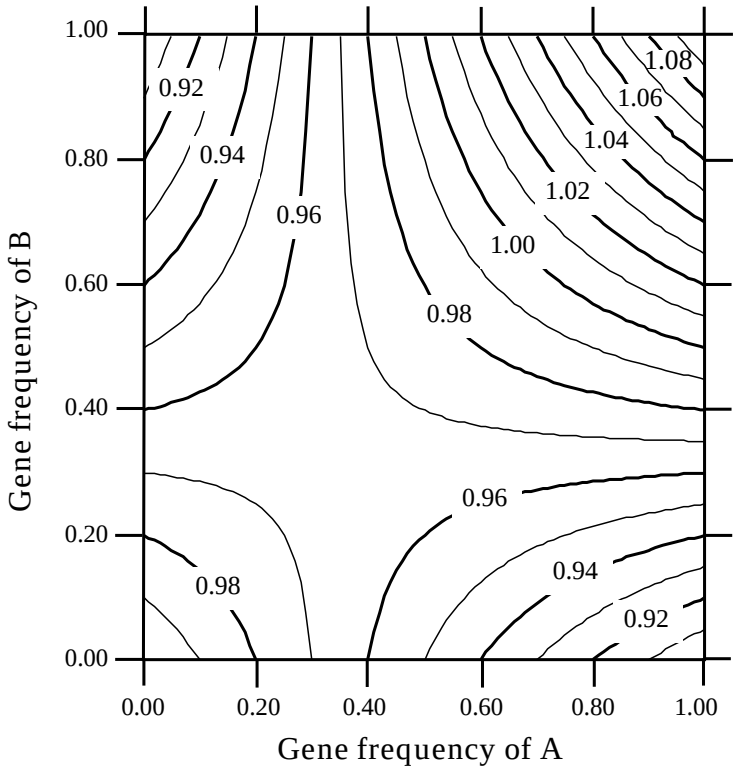


Figure 8.4: Contours of mean population fitness as a function of gene frequency in the absence of linkage disequilibrium, for the case in which the fitnesses of *AB*, *Ab*. *aB* and *ab* are $1+u \ : \ 1-s \ : \ 1-s \ : 1$. If linkage equilibrium is maintained, which will nearly be true if the loci are unlinked, the population will always move uphill and thus will climb the peak on which it is located. There are two peaks, one at the lower-left and one at the upper-right. The case shown is for $u = 0.1$ and $s = 0.1$.

where as usual $\bar{w}$ is the average of the w_{ij}, weighted by the genotype frequencies:

$$\bar{w} \ = \ \sum_i \sum_j x_i x_j w_{ij}. \qquad \text{(VIII-18)}$$

Recombination occurs among these survivors. It has no effect except on four of the 16 diploid genotypes: *AB/ab*, *Ab/aB*, *aB/Ab*, and *ab/AB*. These are genotypes 14, 23, 32, and 41 respectively. Table 8.1 shows boxes for all of the 16 pairs of haplotypes, and in each shows what fraction of *AB* gametes that genotype will produce, when there is a recombination fraction of r. We can get the frequency of x_1 in the next generation by

Table 8.1: All 16 possible diploid genotypes, each showing the fraction of gametes of haplotype AB that it will produce. Blank cells produce none.

	AB	Ab	aB	ab
AB	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}(1-r)$
Ab	$\frac{1}{2}$		$\frac{1}{2}r$	
aB	$\frac{1}{2}$	$\frac{1}{2}r$		
ab	$\frac{1}{2}(1-r)$			

summing these gamete contributions, each multiplied by the frequency of that genotype:

$$\begin{aligned} x_1' \;=\; & x_1x_1w_{11}/\bar{w} \;+\; \tfrac{1}{2}x_1x_2w_{12}/\bar{w} \;+\; \tfrac{1}{2}x_1x_3w_{13}/\bar{w} \;+\; \tfrac{1}{2}(1-r)\,x_1x_4w_{14}/\bar{w} \\ & +\; \tfrac{1}{2}x_2x_1w_{21}/\bar{w} \;+\; \tfrac{1}{2}r\,x_2x_3w_{23}/\bar{w} \\ & +\; \tfrac{1}{2}x_3x_1w_{31}/\bar{w} \;+\; \tfrac{1}{2}r\,x_3x_2w_{32}/\bar{w} \\ & +\; \tfrac{1}{2}(1-r)\,x_4x_1w_{41}/\bar{w} \end{aligned} \tag{VIII-19}$$

This expression can be simplified by noting that there is a $1/\bar{w}$ in every term on the right-hand side. We can also make a simple assumption that will allow collapsing more terms. This is that $w_{ij} = w_{ji}$ for all pairs of values i and j. This amounts to assuming that the fitness of a genotype cannot depend on whether haplotype i came from the mother or the father. This is reasonable in most cases but rules out certain types of maternal effects.

If we assume that, and collect terms, we get

$$x_1' \;=\; \left(x_1x_1w_{11} + x_1x_2w_{12} + x_1x_3w_{13} + (1-r)\,x_1x_4w_{14} + r\,x_2x_3w_{23}\right)/\bar{w}. \tag{VIII-20}$$

Collecting terms in r, factoring x_1 out of the terms that do not contain r, we can simplify this, especially if we assume further that $w_{14} = w_{23}$. This latter assumption amounts to specifying that the fitness of a double heterozygote does not depend on its phase, so that the fitness of *AB/ab* equals that of *Ab/aB*. This is again reasonable in most cases but rules out cis-trans effects. The resulting expression is:

$$x_1' \;=\; \left(x_1\left(x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14}\right) - r\left(x_1x_4 - x_2x_3\right)w_{14}\right)/\bar{w}. \tag{VIII-21}$$

In this equation, two terms are notable. We have already seen in equation (VIII-3) that $(x_1x_4 - x_2x_3)$ is the linkage disequilibrium measure D. The expression $x_1w_{11} + x_2w_{12} +$

$x_3w_{13} + x_4w_{14}$ is also worth note. It is the average fitness of all genotypes that receive an AB gamete from one parent, averaged over all the gametes they might receive from the other parent. Thus we can define an average fitness of a haplotype (just as in Chapter II we defined the average fitness of an allele in (II-124). In general, we can similarly define the mean fitnesses of all four haplotypes:

$$\bar{w}_i \;=\; \sum_{j=1}^{4} x_j w_{ij}. \tag{VIII-22}$$

We can follow the same argument that lead to (VIII-20) for the other three haplotypes, and if we do we will find, using (VIII-3) and (VIII-22), that the full set of equations is

$$\begin{aligned} x_1' &= (x_1\,\bar{w}_1 \;-\; r\,D\,w_{14})\,/\bar{w} \\ x_2' &= (x_2\,\bar{w}_2 \;+\; r\,D\,w_{14})\,/\bar{w} \\ x_3' &= (x_3\,\bar{w}_3 \;+\; r\,D\,w_{14})\,/\bar{w} \\ x_4' &= (x_4\,\bar{w}_4 \;-\; r\,D\,w_{14})\,/\bar{w}. \end{aligned} \tag{VIII-23}$$

We can use the bookkeeping device k_i in (VIII-13) to simplify this to

$$x_i' \;=\; (x_i\,\bar{w}_i \;-\; k_i\,r\,Dw_{14})\,/\bar{w}, \quad i \;=\; 1,2,3,4. \tag{VIII-24}$$

Equations (VIII-23) were derived by Lewontin and Kojima (1960). The natural next step would be to derive formulas for the future frequencies of all four gamete types in terms of the current frequencies x_i the fitnesses w_{ij}, and the recombination fraction r. Alas, no one has been able to construct general expressions for this (and this should not be surprising since they cannot in many much simpler single-locus cases that are subcases of this). The formulas are easily iterated numerically, and over the years substantial insight has been gained empirically by considering numerical examples. The other way that insight has been gained is by the exact solution of particular cases, to which we now turn.

Table 8.2 will help to show how the w_{ij}'s relate to the genotypes, keeping in mind that $w_{ij} = w_{ji}$:

VIII.4 Linked polymorphisms

Much effort has gone into examining linked overdominant loci. In particular, a number of simple fitness schemes have been investigated that can be at least partly analyzed analytically.

Table 8.2: The relationship between genotypes and fitnesses of haplotype pairs

	BB	Bb	bb
AA	w_{11}	w_{12}	w_{22}
Aa	w_{13}	$w_{14} = w_{23}$	w_{24}
aa	w_{33}	w_{34}	w_{44}

LEWONTIN AND KOJIMA'S SYMMETRIC MODEL. Lewontin and Kojima (1960) examined the behavior of the symmetrical overdominant fitness model. The symmetries ensured that the fitness table would remain the same when we exchange either $A \leftrightarrow a$ or $B \leftrightarrow b$ or both. The fitness table is given in Table 8.3:

Table 8.3: Lewontin and Kojima's symmetric fitness model

	BB	Bb	bb
AA	a	b	a
Aa	c	d	c
aa	a	b	a

It seems evident that if there are any nontrivial equilibria in the system, they will involve equal gene frequencies of alleles A and a, and equal frequencies of B and b too. In that case, it will be true that

$$x_1 + x_2 \;=\; x_3 + x_4 \;=\; x_1 + x_3 \;=\; x_2 + x_4 \;=\; \frac{1}{2}, \tag{VIII-25}$$

so that we can write x_2, x_3, and x_4 all as functions of x_1. In fact, the linkage disequilibrium D is also a function of x_1, being

$$D \;=\; x_1 - \frac{1}{4}. \tag{VIII-26}$$

so that we can also write:

$$x_1 \;=\; x_4 \;=\; \frac{1}{4} + D \tag{VIII-27}$$

and

$$x_2 \;=\; x_3 \;=\; \frac{1}{4} - D \tag{VIII-28}$$

The changes of the x_i are now all replaced by changes of D, provided that we have started at, and thus remain at, gene frequencies of $\frac{1}{2}$ at both loci. We can rewrite $\bar{w}$ using Table 8.3, (VIII-18), (VIII-27) and (VIII-28) as

$$\begin{aligned}\bar{w} &= a\left(2\left(\tfrac{1}{4}+D\right)^2+2\left(\tfrac{1}{4}-D\right)^2\right)+b\left(4\left(\tfrac{1}{4}+D\right)\left(\tfrac{1}{4}-D\right)\right)\\ &\quad +c\left(4\left(\tfrac{1}{4}+D\right)\left(\tfrac{1}{4}-D\right)\right)+d\left(2\left(\tfrac{1}{4}+D\right)^2+2\left(\tfrac{1}{4}-D\right)^2\right)\end{aligned} \tag{VIII-29}$$

and this can be simplified readily to

$$\bar{w} = \frac{1}{4}(a+b+c+d)+4D^2(a-b-c+d). \tag{VIII-30}$$

The haplotype mean fitnesses $\bar{w}_i$ given in (VIII-22) can also be simplified to

$$\begin{aligned}\bar{w}_1 = \bar{w}_4 &= \left(\tfrac{1}{4}+D\right)a+\left(\tfrac{1}{4}-D\right)b+\left(\tfrac{1}{4}-D\right)c+\left(\tfrac{1}{4}+D\right)d\\ \bar{w}_2 = \bar{w}_3 &= \left(\tfrac{1}{4}+D\right)b+\left(\tfrac{1}{4}-D\right)a+\left(\tfrac{1}{4}-D\right)d+\left(\tfrac{1}{4}+D\right)c\end{aligned} \tag{VIII-31}$$

which simplifies to

$$\begin{aligned}\bar{w}_1 = \bar{w}_4 &= (a+b+c+d)\tfrac{1}{4}+(a+d-b-c)D\\ \bar{w}_2 = \bar{w}_3 &= (a+b+c+d)\tfrac{1}{4}-(a+d-b-c)D.\end{aligned} \tag{VIII-32}$$

Substituting (VIII-32) and (VIII-30) into the first equation of (VIII-23) we get

$$\frac{1}{4}+D' = \frac{\left(\tfrac{1}{4}+D\right)\left((a+b+c+d)\tfrac{1}{4}+(a-b-c+d)D\right)-rdD}{\tfrac{1}{4}(a+b+c+d)+4D^2(a-b-c+d)} \tag{VIII-33}$$

which simplifies to

$$D' = D\,\frac{\tfrac{1}{2}(a+d)-rd}{\tfrac{1}{4}(a+b+c+d)+4D^2(a-b-c+d)}. \tag{VIII-34}$$

We are looking for equilibria of the system, which will be roots of the equation obtained by setting $D' = D$. One such root is clearly $D = 0$. The other is the value of D that makes the fraction on the right-hand side of (VIII-34) be 1. This yields the equation

$$4D^2(a-b-c+d) = \frac{1}{4}(a-b-c+d)-rd \tag{VIII-35}$$

which has two roots,

$$D = \pm\frac{1}{4}\sqrt{1 - \frac{rd}{4(a-b-c+d)}}. \qquad \text{(VIII-36)}$$

We now have up to three equilibrium values of D. It will be helpful to plot an example. Suppose that a, b, and c are all 0.9, and $d = 1.0$. That is a case in which making one locus homozygous reduces fitness by 0.1, but making two homozygous is not as bad as expected: it does not reduce fitness any further. Figure 8.5 shows the equilibrium values of D plotted against the possible values of r.

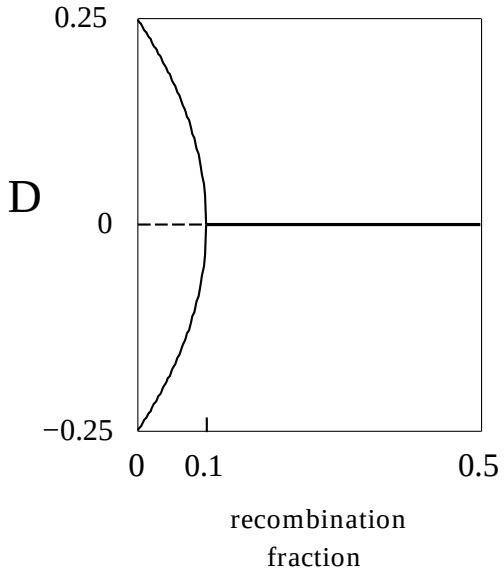


Figure 8.5: Equilibrium values of D for the Lewontin-Kojima model when all genotypes other than the double heterozygote *AaBb* have a fitness of 0.9 relative to that genotype. The solid curves show the stable equilibria, plotted as functions of the recombination fraction r. The dashed line is the unstable equilibrium.

There is always an equilibrium with $D = 0$. Below $r = (a - b - c + d)/d$ there is a change, with three equilibria, $D = 0$ and the two values from (VIII-36). Above $r = (a - b - c + d)/d$ those two equilibria do not exist, as the quantity inside the square root is negative, so that the solutions would be imaginary.

Are all of these solutions relevant? Not all need be stable equilibria. We can investigate the stability of the equilibrium at $D = 0$ simply using equation (VIII-34). When

$D \simeq 0$ we can drop the term in D^2 and approximate,

$$D' \simeq D \frac{\frac{1}{2}(a+d) - rd}{\frac{1}{4}(a+b+c+d)}. \tag{VIII-37}$$

A necessary condition for the stability of the equilibrium at $D = 0$ is that if D is perturbed to a small nonzero value, it return to 0. This requires that the multiplier of D on the right-hand side of (VIII-37) be less than 1. Then we require

$$\frac{\frac{1}{2}(a+d) - rd}{\frac{1}{4}(a+b+c+d)} < 1. \tag{VIII-38}$$

Since $a+b+c+d > 0$, we can multiply through by it and solve for r, getting

$$rd > \frac{1}{4}(a-b-c+d) \tag{VIII-39}$$

or if $d > 0$,

$$r > \frac{1}{4} \frac{a-b-c+d}{d}. \tag{VIII-40}$$

Thus the condition for the existence of the paired equilibria (VIII-36) are the same as the condition for instability of $D = 0$. When the paired equilibria exist above and below it, $D = 0$ is not stable. Above the critical value of r, they disappear (having collided with each other and with $D = 0$ at the critical recombination value and disappeared into the complex plane), and there only $D = 0$ exists, and it is a stable equilibrium. We have not made a full analysis of the stability of $D = 0$, since have not allowed for deviations of the gene frequencies from $\frac{1}{2}$, but when we do we find that they do not affect the conditions for stability of D.

The paired equilibria in (VIII-36) are stable when they exist. This can be verified by using (VIII-34) to ask whether small departures from the equilibria will grow. We will not do this here, but simply note that the equilibria are stable.

A useful reparameterization of the fitness scheme in Table 8.3 is given in Table 8.4. On substituting in (VIII-36) 1 for d, $1-s$ for b, $1-t$ for c, and $1-s-t+e$ for a, we get

$$D = \pm\frac{1}{4}\sqrt{1 - \frac{4r}{e}}. \tag{VIII-41}$$

the other equilibrium being of course $D = 0$. The condition for stability of the equilibrium $D = 0$ in (VIII-41) also simplifies, becoming by substitution into (VIII-40)

$$r > \frac{1}{4}\, e. \tag{VIII-42}$$

The fitness scheme in Table 8.4 clarifies which fitness effects are affecting D. The parameter e measures the gene interactions by measuring the departure of the effect of

Table 8.4: A reparameterization of Lewontin and Kojima's symmetric fitness model

	BB	Bb	bb
AA	$1-s-t+e$	$1-s$	$1-s-t+e$
Aa	$1-t$	1	$1-t$
aa	$1-s-t+e$	$1-s$	$1-s-t+e$

homozygosing the two loci from what would be expected from making only one of them homozygous. Note that the departure is a departure from an additive prediction, not a multiplicative prediction. In this, e is a different measure from E, the epistasis measure in (VIII-8), which measured departure from multiplicative interaction. The fact that one measure appears naturally in that context and the other here has been the cause of much disagreement over how to best measure epistasis.

In the present case, the paired equilibria (VIII-41) appear whenever $e > 0$ and r is sufficiently small. One particularly common pattern when the two loci do not have any particular biochemical or developmental interaction will be for the fitnesses to be multiplicative. That case we expect the fitness of the double homozygotes to be $(1-s)(1-t) = 1-s-t+st$, so that $e = st$. In that case, whenever $r < st/4$ there will be a pair of stable equilibria, one with coupling and one with repulsion disequilibrium. Thus there is every reason to expect such disequilibrium to occur between many pairs of closely linked loci, because it can occur even when the loci that are closely linked affect totally unrelated aspects of fitness.

The pattern of having a coupling-repulsion pair of stable equilibria when r is small, and otherwise linkage equilibrium, is special to this particular model. But a similar pattern, that of having a not-quite-symmetrical pair of equilibria when r is small, and otherwise an equilibrium with almost no linkage disequilibrium, will occur quite frequently in a variety of models.

Note that I have not explained all of the behaviors of the Lewontin-Kojima model. It is also possible for there to be boundary equilibria in which both loci become fixed, and in those cases the paired equilibria that are found when r is small are actually unstable, if we allow the gene frequencies to depart from $1/2$.

FITNESS AND DISEQUILIBRIUM: MORAN'S COUNTEREXAMPLE. If we take equation (VIII-30) and the reparameterized fitnesses in Table 8.4, we can express the mean fitness of the population as

$$\bar{w} = \left(1 - \frac{s}{2} - \frac{t}{2} + \frac{e}{4}\right) + 4D^2 e. \qquad \text{(VIII-43)}$$

Table 8.5: An example showing Moran's phenomenon

	BB	Bb	bb
AA	1	0.5	1
Aa	0.5	1.1	0.5
aa	1	0.5	1

This implies that if e is positive, which is the case that can lead to paired equilibria, any increase in the absolute value of D away from 0 will increase the mean fitness of the population. However, in any given case, r might not be small enough to allow disequilibrium to persist.

P. A. P. Moran (1964) used cases like this to make a dramatic point about the ability of multi-locus selection to optimize mean fitness. He actually used a somewhat different parameterization of fitnesses, but the point can be made using Lewontin and Kojima's model. If, for example, we take $s = t = 0.5$, and $e = 0.5$, so that the fitness table is given in Table 8.5: Then if $r > 0.25$ the equilibrium $D = 0$ is stable. With tighter linkage the two paired equilibria in equations (VIII-44) exist. Notice what happens if $r > 0.25$ and we start with any nonzero value of D beyond the equilibrium: D will decrease continuously towards the equilibrium, *and as it does the mean fitness of the population continuously decreases*, as shown in Table 8.6. This depressing behavior will occur at all values of r beyond 0.25. If we start with either positive or negative D, its value will continually subside toward 0, and from (VIII-30) we can immediately see that this will result in a continual decline in mean fitness. The phenomenon will occur in the Lewontin-Kojima model for all positive s and t and for all positive values of e. However for some of them the equilibrium gene frequencies of $1/2$ will be unstable and ultimately one allele will be lost at each locus, and the population mean fitness can increase as that happens. In the cases where the gene frequencies approach $1/2$ at both loci, as they change towards them, in the process mean fitness will tend to increase. This can counterbalance the decrease of mean fitness by reduction of D.

COADAPTED GENE COMPLEXES AND RECOMBINATION. Moran's phenomenon shows that two-locus systems do not automatically maximize mean fitness. The reason seems basically to be the breakdown of linkage disequilibria by recombination. If there is no recombination ($r = 0$) then the mean fitness will be maximized. This can be seen because then the four haplotypes are inherited as if they were four alleles at a single locus. If we set $r = 0$ in the equations of change (VIII-23) they become exactly the same as the multiple-allele equations (II-122), for the corresponding assignment of fitnesses

Table 8.6: Change of D and $\bar{w}$ in the numerical example

Generation	D	$\bar{w}$
0	0.2000000	0.9510
1	0.1577287	0.8845
2	0.1337493	0.8537
3	0.1175011	0.8357
4	0.1054454	0.8239
5	0.0959848	0.8155
6	0.0882713	0.8093
7	0.0818050	0.8044
8	0.0762684	0.8006
9	0.0714486	0.7975
10	0.0671963	0.7949
11	0.0634033	0.7927
12	0.0599889	0.7908
13	0.0568914	0.7892
14	0.0540627	0.7876
15	0.0514648	0.7867
20	0.0410332	0.7824
25	0.0334432	0.7799
30	0.0276221	0.7784
35	0.0230095	0.7773
40	0.0192763	0.7766
45	0.0162114	0.7762
50	0.0136704	0.7758
∞	0	0.775

to genotypes. It follows from this that the mean fitness can never decrease from one generation to the next.

Recombination reduces the mean fitness by breaking up "coadapted gene complexes" by breaking down linkage disequilibrium. Consider the fitnesses shown above in Table 8.5. If we have no linkage disequilibrium, and are at equilibrium gene frequencies of $\frac{1}{2}$, the average fitness of *Aa* heterozygotes will be 0.8, and the average fitness of *aa* and *AA* homozygotes will be 0.75, yielding a mean fitness which is 0.775. But if there is complete linkage, we can have a linkage disequilibrium which, for example, associates *b* completely with *A*, so that there are only two haplotypes in the population, *AB* and *ab*. These will when homozygous have fitness 1, and when heterozygous fitness 1.1. It is clear that this state of affairs has higher mean fitness. The breakdown of the association

between A and b (which is often called "coadaptation"), results in lowered mean fitness. Selection "wants" to eliminate the repulsion haplotypes Ab and aB, but recombination between the desirable haplotypes AB and ab keeps reintroducing them. (It is worth noting that this is one of two equilibria. There is a symmetrical equilibrium in which A is associated with b and a with b and it is the two coupling haplotypes AB and ab that selection tries to eliminate.)

By examining (VIII-43) and (VIII-41) it is easy to show that not only can recombination lead to a continual decrease of mean fitness during the course of evolution, but it leads to an equilibrium state which has lowered mean fitness. This is quite frequently found in multiple-locus models. If all alleles at all loci are present at a stable equilibrium, so that all haplotypes are present (as will be the case when there is also not complete linkage between any pair of loci) we can use the single-locus multiple-alleles result make a simple argument to this effect. If all haplotypes are present at equilibrium with or without recombination present, unless the equilibrium haplotype frequencies with recombination are the same as with no recombination, they must have lower mean fitness, since the equilibrium haplotype frequencies without recombination are at a local maximum of the mean fitness, and as mean fitness is a quadratic function of the haplotype frequencies it cannot have any other local maximum with all haplotype frequencies present.

This seemingly dysfunctional property of recombination raises the question of why recombination is present at all. It may be a byproduct of other cellular phenomena, such as DNA repair, but we will see, later in this Chapter, that there are cases in which recombination is advantageous, and they must also be taken into account in making an evolutionary explanation of the presence of recombination.

The notion of an adaptive topography is thus compromised by recombination. There is no rule that mean fitness always increases, or even sometimes increases, nor is there any rule that the final mean fitness is above the initial mean fitness. Nevertheless in "real" cases it is often found that the net effect of selection in the presence of recombination is to increase the mean fitness of the population, comparing final to initial values. It is just that it does not do so in all cases. The genetic system is not perfectly designed to increase mean fitness (probably because evolution has had only a limited opportunity to try alternative genetic systems). But if one cannot be Panglossian and believe that mean fitness changes are always for the best, your very presence reading these pages is indirect evidence that there have, on the whole, been more increases than decreases of mean fitness in the course of evolution.

THE GENERAL SYMMETRIC MODEL. Another, less limited model that can be exactly analyzed is the General Symmetric model, which was first introduced by Bodmer (Bodmer and Parsons, 1964; Bodmer and Felsenstein, 1967) and was first rigorously analyzed by Karlin and Feldman (1969, 1970). It is shown in Table 8.7.

Table 8.7: The General Symmetric Model

	BB	Bb	bb
AA	$1-\delta$	$1-\beta$	$1-\alpha$
Aa	$1-\gamma$	1	$1-\gamma$
aa	$1-\alpha$	$1-\beta$	$1-\delta$

Notice the sense in which this fitness scheme is symmetric. In the Lewontin-Kojima symmetric model, if we relabeled the fitness table by exchanging the A and a allele symbols, the table would be unchanged, and this would also be true if we exchanged the B and b allele symbols. In the General Symmetric model neither of these symmetries exists, because of the difference between α and δ. But in the General Symmetric model if we simultaneously exchange $A \leftrightarrow a$ and $B \leftrightarrow b$ the fitness table is unchanged. Lewontin and Kojima's model is a subcase of this one, the case in which $\alpha = \delta$.

We will change some of the parameters to make the expressions more meaningful in our presentation of this model. Suppose that we let

$$\begin{aligned} \beta &= s \\ \gamma &= t \\ \alpha &= s + t + e_2 \\ \delta &= s + t - e_1 \end{aligned} \tag{VIII-44}$$

The table of fitnesses then becomes the one shown in Table 8.8. Nevertheless there are symmetric equilibria as well. These are fairly readily derived. If we assume that the gene frequencies at both loci and 0.50:0.50, then we can write the gamete frequencies of the four haplotypes as a result of (VIII-1) and (VIII-2) as

$$\begin{aligned} x_1 &= \tfrac{1}{4} + D \\ x_2 &= \tfrac{1}{4} - D \\ x_3 &= \tfrac{1}{4} - D \\ x_4 &= \tfrac{1}{4} + D. \end{aligned} \tag{VIII-45}$$

Using these in place of the x_i we can write the basic diploid two-locus equations (VIII-23)

Table 8.8: The General Symmetric Model, reparameterized

	BB	Bb	bb
AA	$1-s-t+e_1$	$1-s$	$1-s-t-e_2$
Aa	$1-t$	1	$1-t$
aa	$1-s-t-e_2$	$1-s$	$1-s-t+e_1$

in terms of D. We have using (VIII-22)

$$\begin{aligned}
\bar{w}_1 &= \left(\tfrac{1}{4}+D\right)(1-s-t+e_1) + \left(\tfrac{1}{4}-D\right)(1-s) + \left(\tfrac{1}{4}-D\right)(1-t) + \left(\tfrac{1}{4}+D\right) \\
\bar{w}_2 &= \left(\tfrac{1}{4}+D\right)(1-s) + \left(\tfrac{1}{4}-D\right)(1-s-t-e_2) + \left(\tfrac{1}{4}-D\right) + \left(\tfrac{1}{4}+D\right)(1-t) \\
\bar{w}_3 &= \left(\tfrac{1}{4}+D\right)(1-t) + \left(\tfrac{1}{4}-D\right) + \left(\tfrac{1}{4}-D\right)(1-s-t-e_2) + \left(\tfrac{1}{4}+D\right)(1-t) \\
\bar{w}_4 &= \left(\tfrac{1}{4}+D\right) + \left(\tfrac{1}{4}-D\right)(1-t) + \left(\tfrac{1}{4}-D\right)(1-s) + \left(\tfrac{1}{4}+D\right)(1-s-t+e_1)
\end{aligned} \tag{VIII-46}$$

which simplifies to

$$\begin{aligned}
\bar{w}_1 &= \bar{w}_4 = 1-\tfrac{1}{2}s-\tfrac{1}{2}t+\left(\tfrac{1}{4}+D\right)e_1 \\
\bar{w}_2 &= \bar{w}_3 = 1-\tfrac{1}{2}s-\tfrac{1}{2}t-\left(\tfrac{1}{4}-D\right)e_2
\end{aligned} \tag{VIII-47}$$

The mean fitness is (as is the case with multiple alleles) the weighted mean of these haplotype mean fitnesses:

$$\bar{w} = 1-\frac{1}{2}s-\frac{1}{2}t+2\left(\frac{1}{4}+D\right)^2 e_1 - 2\left(\frac{1}{4}-D\right)^2 e_2. \tag{VIII-48}$$

We can then write the first equation of (VIII-23), assuming $x_i' = x_i$, as

$$0 = \left(\frac{1}{4}+D\right)^2\left(1-2\left(\frac{1}{4}+D\right)\right)e_1 + 2\left(\frac{1}{4}+D\right)\left(\frac{1}{4}-D\right)^2 e_2 - rD. \tag{VIII-49}$$

This is a cubic equation in D, which will in general have three roots, though not always ones that are feasible. The fact that s and t have cancelled out of the equation suggests

that the parameterization in Table 8.1 is a natural one. Although the solutions of this cubic can be written down explicitly, the result is not particularly illuminating.

When $r = 0$ the equation has a particularly simple form (it can also be derived by simply requiring that $\bar{w}_1 = \bar{w}_2$). It factors into

$$\left(\frac{1}{4} + D\right)\left(\frac{1}{4} - D\right)\left(\left(\frac{1}{4} + D\right)e_1 + \left(\frac{1}{4} - D\right)e_2\right) \;=\; 0, \qquad \text{(VIII-50)}$$

which yields the solutions $D = \frac{1}{4}$, $D = -\frac{1}{4}$, and

$$D \;=\; \frac{e_1 + e_2}{4(e_2 - e_1)}. \qquad \text{(VIII-51)}$$

This is generically similar to the results with Lewontin and Kojima's model. There are paired equilibria at complete linkage equilibrium when $r = 0$, as well as another equilibrium with a modest amount of linkage disequilibrium. When $e_1 = -e_2$, the results reduce to those of Lewontin and Kojima, as they must. The pattern as r is changed is also similar to Lewontin and Kojima's model. The paired equilibria approach each other and collide, and at that point the "central" equilibrium becomes stable, with the amount of linkage disequilibrium it contributes declining rapidly with increasing r.

For example, when $e_1 = 0.1$ and $e_2 = -0.15$, the central equilibrium when $r = 0$ has $D = 0.05$. Table 8.9 shows the three equilibria in this case as they change with different values of r. For small values of r, there are paired strong disequilbria, which in fact are stable equilibria, plus an unstable equilibrium with a positive value of D. As r increases the unstable equilibrium value of D rises, and the positive stable value falls, and they collide at just above $r = 0.014062$. These two equilibria annihilate each other by having complex values of D which are impossible in the real world. That leaves the negative value of D as the only stable one, and it gradually approaches zero, roughly inversely with r. Even with free recombination ($r = 0.5$) the equilibrium value of D is not quite zero.

Other patterns of symmetric equilibria are also possible in the General Symmetric model. For example, when $e_1 = -0.1$ and $e_2 = -0.15$, there is only one stable equilibrium, which has a small positive value of D, and that drops gradually toward zero as r increases. Lewontin and Feldman (1988) have shown that the presence of an equilibrium with large amounts of recombination with D near zero will characterize two-locus, two-allele models more generally, given that they have any equilibrium.

It was long assumed that all the two-locus equilibria of the General Symmetric model must have gene frequencies 0.50 : 0.50 at all loci. Karlin and Feldman (1970) made the startling discovery that there are "unsymmetric" equilibria, which have unequal gene frequencies of A and a (though they have the same pair of gene frequencies for alleles B and b).

Table 8.9: Three equilibria of the linkage disequilibrium value D for a general symmetric model in which $e_1 = 0.1$ and $e_2 = -0.15$. The three equilibrium values of D are shown, except when they are not achievable values, in this case, when they are complex numbers.

r			
0	-0.25	0.05	0.25
0.01	-0.215109	0.0773891	0.18772
0.011	-0.211441	0.082675	0.178766
0.012	-0.20774	0.089315	0.168425
0.013	-0.204005	0.0985179	0.155487
0.014	-0.200237	0.118186	0.132051
0.01406	-0.200009	0.123618	0.126391
0.015	-0.196435	–	–
0.02	-0.176953	–	–
0.03	-0.13637	–	–
0.04	-0.0977967	–	–
0.05	-0.0685037	–	–
0.10	-0.0224614	–	–
0.20	-0.0092443	–	–
0.30	-0.0058105	–	–
0.40	-0.0042360	–	–
0.50	-0.0033327	–	–

In addition to the General Symmetric Model, Puniyani and Feldman (2006) have discovered an interesting semi-symmetric model, symmetric and one locus but not the other, which lends itself to investigation of the number of equilibria and the levels of recombination needed for them to be stable or unstable.

MULTIPLICATIVE OVERDOMINANT LOCI. Most schemes of interaction among loci are arbitrary. One is not – multiplicative fitnesses. We have noted earlier in this chapter that fitnesses tend to be multiplicative across loci when the loci do not interact. There has been some work on linked overdominant loci whose fitnesses are multiplicative. Bodmer and Felsenstein (1967) showed that there is an equilibrium in such case where the loci are each at their equilibrium gene frequencies, and they are at linkage equilibrium with each other. However, they also showed that if the fitnesses were $1 - s_1 \ : \ 1 \ : \ 1 - t_1$ and $1 - s_2 \ : \ 1 \ : \ 1 - t_2$, then if the recombination fraction

$$r \ < \ \left(\frac{s_1 t_1}{s_1 + t_1} \right) \left(\frac{s_2 t_2}{s_2 + t_2} \right) \qquad \text{(VIII-52)}$$

the two loci go into linkage equilibrium with each other and there are two paired equilibria. When $r = 0$ the disequilibrium is complete. The haplotype frequencies are easy to compute in that case: the fitnesses of the three possible genotypes are then either $(1-s_1)(1-s_2) : 1 : (1-t_1)(1-t_2)$ or $(1-s_1)(1-t_2) : 1 : (1-t_1)(1-s_2)$.

If such disequilibria form, the haplotypes will be more strongly overdominant than were the indvidual loci. Franklin and Lewontin (1970) showed by exact iteration of haplotype frequencies that if there are many overdominant loci sufficiently near each other, the genome could "congeal" into a small number of haplotypes, each strongly overdominant and at intermediate haplotype frequencies. Interest in such a phenomenon has waned, since it does not seem to be commonly found in nature.

SOME PERSPECTIVE ON INTERACTING POLYMORPHISMS. The behaviors in interacting two-locus polymorphisms are complicated and interesting. It is hard to make many generalizations. Much work was done on them in the late 1960s and early 1970s. The unstated hope of that work was that some general rules could be found – perhaps even a function that was maximized by evolution. We knew from Moran's result that it would not be mean fitness, but whatever it was, it would give us insight into how the details of the genetic systems compromised optimization of fitness.

Alas, we were to be disappointed – the maximand was never found. The advances that were made tended to be disproofs of generalities rather than proofs of them.

Some interesting cases that were discovered were ones that showed, not an equilibrium, but a limit cycle. This was suggested by Akin (1979) using approximate methods. Hastings (1981) found sets of fitnesses that showed sets of fitnesses that showed these stable limit cycles in exact numerical iterations.

VIII.5 Intermediate optimum models

A common pattern of natural selection in nature must be for a higher fitness to be associated with an intermediate value of a character, close to an optimum. Sewall Wright (1935b) investigated such cases, in the era before the exact effects of linkage disequilibrium could be known. Lewontin (1964b) and Singh and Lewontin (1966) have used exact computer iteration of haplotype frequencies to discover what equilibrium states result from this optimum selection. The models assume that the loci interact additively to determine a phenotype, with fitness a function of the departure of the phenotype from an intermediate optimum value. It might seem obvious in such a case that with, say, 10 loci, that optimum selection toward an intermediate phenotype will be a force to maintain variability at the individual loci, keeping them segregating in the population. However, this is not the case. Lewontin and Singh and Lewontin found that the population moved rapidly into a state where all loci had intermediate gene frequencies, and the mean phenotype was almost precisely at the optimum value. But this state then slowly changed.

Individual loci gradually fixed.

If we think of the alleles at each loci as + or − alleles, based on the direction of their effect on the phenotype, Lewontin and Singh found that some alleles fixed for the + allele, and some for the − allele. Ultimately all loci would be fixed, or all but one would be fixed. The state of the population would be a mixed fixation that brought the phenotype close to the optimum value, with at most one locus still segregating. For example, if we have 10 loci, (*A* through *J*), each with two alleles, and if the phenotype is simply the number of capital letters in the genotype, then the phenotypes can range from 0 to 20. If the optimum phenotype is 10, the population might end up fixed for the genotype *AABBccDDeeffGGHHiijj*. All individuals would then have the optimum phenotype (we have neglected to allow for environmental variance of the phenotype). By contrast, if each locus segregated for both alleles in equal frequencies, the average phenotype would be 10, but only a small fraction of individuals would have this phenotypic value. If the optimum was instead 11, no mixed fixation could achieve this, but if five loci were fixed for the + allele and four for the − allele, the remaining one locus would show overdominance. Thus by achieving a mixed fixation, the population has moved toward the highest possible mean fitness. Wright had already (1935, 1952) argued that the population would proceed toward these mixed fixations. There are numbers of them that will be nearly or precisely tied in mean fitness. He suggested that movement among these equilibria would play a major role in his Shifting Balance Theory of evolution. Closely related is the model by Lande (1976b) in which there is a line of genetic equilibria along which genetic drift can move the population, while still keeping the population near the optimum phenotype.

VIII.6 Selection on modifiers

It has long been known that aspects of the genetic system can be modified by genetic variants. This includes sizes of genetic effects, dominance, degree of linkage, and rates of mutation. To make a simple population-genetic model we need to consider a new allele at a modifier locus which has no direct effect on fitness, but acts by modifying something else such as the fitnesses at another locus, or the rate of recombination or the rate of mutation. A simple deterministic haploid model will point out some of the properties of such a case. Imagine a locus with two alleles, A and a, where allele a is deleterious and maintained by mutation, with the rate of mutation to a being u. As we saw in Chapter III , the deleterious allele will be maintained at an equilibrium frequency of u/s. Now suppose that at another locus, allele B arises which has no effect except that it makes the effect of allele a on fitness smaller. The fitness table might be similar to the fitness scheme VIII-8:

$$\begin{array}{cc} AB & 1 \\ Ab & 1 \\ aB & 1-s+E \\ ab & 1-s \end{array}$$

Given the recombination fraction between the loci, r, we can ask whether allele B will increase when rare. Its only effect is to increase the fitness of one genotype, so the answer is yes, it will increase. It is never selected against. But it will increase very slowly. If the two loci are unlinked, they will be near linkage equilibrium. If both alleles are rare, allele B will be favored by an amount E, but only u/s of the time. So it will increase at a rate that corresponds to a selection coefficient of only $(u/s)E$. If the increase of fitness E is a fraction f of the original selection coefficient s, this will be only $(u/s)fs$ which is uf. Thus the selection coefficient favoring B is tiny. This leads us to wonder whether selection will have much effect in modifying the fitness of allele a. With tighter recombination, we can analyze this case, though it does not yield to linearizing the equations of the frequencies of haplotypes, as we did with equation (VIII-15) the terms that result in increase of the frequency of haplotype aB are quadratic. They are also very small, for reasonable values of s and u. Tight linkage does not greatly speed up the increase of allele B. It is hard to avoid the conclusion that modifiers have little influence on the effects of rare deleterious alleles.

But only as long as they are rare. There is a difficulty with this model. The direct effect of the modifier B is small, but only if allele a is rare. If a is at some intermediate gene frequency, then it is even hard to know which locus is the modifier. The fitness effect E causes fitness differences between the alleles of locus A and also between the alleles of locus B. There seems to be no model that has one locus act only to modify the fitness effects of the other, and not have any direct effect on the fitnesses at its own locus.

MODIFICATION OF DOMINANCE. There has been a considerable amount of work on modifiers. R. A. Fisher (1928) proposed that deleterious mutants were often recessive because modifiers had been selected to increase the fitness of the heterozygotes in which the mutant alleles were usually found. Sewall Wright (1929a, b) and J. B. S. Haldane (1930a) were skeptical of Fishers argument, as they were more familiar with the biochemistry of gene action, and could see reasons for dominance that would be much stronger than the weak effects of selection for modifiers of the fitness of the rare heterozygote. The controversy between Fisher and Wright became heated and was the occasion for the final breakdown of communications between them. Brian Charlesworth (1979) has reviewed the evidence, coming down on Wright and Haldane's side of the argument. His paper should be consulted for further references.

MODIFICATION OF RECOMBINATION. Natural selection can also modify recombination rates. A simple model of this would have two loci whose fitnesses interact, and

a third locus nearby whose only effect was to change the rate of recombination between those two loci. In a deterministic model with constant fitnesses, the same result is always found: if there is epistasis between the original two loci, the modifier typically changes so as to reduce the recombination between them (Nei, 1967; Feldman, 1972; for more recent work see the review by Feldman, Otto, and Christiansen, 1996). The effect of this reduction is to make it less likely that recombination will break up favorable gene combinations. As we will see below when we discuss linkage disequilibrium created by genetic drift, this can create opportunities for selection of modifiers that increase levels of recombination. Feldman, Otto, and Christiansen, as well as Otto and Feldman (1997), note some other deterministic scenarios that can also select modifiers that increase recombination.

MODIFICATION OF MUTATION RATE. A more puzzling problem is selection on modifiers of mutation rates. One would like to think that present-day mutation rates are "tuned" by selection of modifiers to achieve the best compromise between having too many deleterious mutations and too few advantageous mutations. And this is exactly what is found (Holsinger and Feldman, 1983). But their model is for a completely self-fertilizing species. A similar result was obtained by Leigh (1970) for asexual organisms. But as Leigh found for outcrossing sexual organisms, there is no selection for an optimum mutation rate. Selection should act only to decrease the mutation rate. Liberman and Feldman (1986) have a detailed analysis of a two-locus system in which one locus modifies the mutation rate of the other, and found the same outcome.

In asexuals or completely selfing species, the mutator allele bears the full responsibility for all mutations it causes, because they continue to reside in the same descendants where it resides. In an outcrossing sexual species, the mutator can be irresponsible. It is expected to stay with the mutations that it causes for only a few generations, segregating away from them quickly. Thus most deleterious effects they have, and most advantageous effects they have, do not decrease or increase the frequency of the mutator allele. Thus it is hard to see how a mutation rate optimal for the species would result.

The exception will be when the modifier of mutation acts locally. If mutations reside nearby, they will hang around longer and the modifier might have its gene frequencies more nearly reflect the net effect of the mutation rate that it causes. From this one might predict some conflict between local mutation rate modifiers and general mutation rate modifiers. The matter has not been examined either theoretically or empirically.

GENERAL REDUCTION PRINCIPLE. Marc Feldman and his coworkers have generalized the theory of modifier genes, demonstrating that a large class of modifiers has the property that their effect is "viability analogous". In deterministic models, the quantity that the modifier controls can be treated as if it were the inverse of a viability, and natural selection on the modifier then acts to decrease the quantity. Feldman and Liberman (1986) showed this for rates of mutation, recombination, and migration. Altenberg

and Feldman (1987) gave a more general proof for many kinds of modifiable evolutionary parameters. Altenberg (2011, 2012) has since given even more general proofs for modification of mutation and migration rates. Zhivotovsky, Feldman, and Christiansen (1996) gave a multiple-locus proof for invasion of new alleles modifying recombination. Altenberg, Liberman, and Feldman (2017) have generalized the reduction further. Of course, the resulting "reduction principle" cannot be the whole story, otherwise there would be almost no mutation, no recombination, and no migration.

VIII.7 Genetic drift and linkage

Genetic drift also produces linkage disequilibrium. When there are multiple haplotypes, genetic drift will change the frequencies of all of them, and this is very unlikely to leave the population precisely in a state of linkage equilibrium. Of course, it has no tendency to preferentially associate particular alleles. For any pair of alleles, one at each of two loci, it can lead to either positive or negative association. Analytical results for the extent of disequilibrium produced by genetic drift are difficult, and we do not have the extensive space here that they require.

However, we can at least get some rough idea of the amount of linkage disequilibrium produced by genetic drift by considering a simple case with two alleles at each of two loci, and a population that starts in linkage equilibrium and undergoes one generation of change. If recombination precedes genetic drift in the life cycle, there is no change of haplotype frequencies by recombination in this generation. Genetic drift then changes the haplotype frequency x_1 of AB, as well as the gene frequencies of alleles A and B (which we will call p and q). If the changes of the four haplotype frequencies are, respectively, e_1, e_2, e_3, and e_4 then the linkage disequilibrium after the bout of genetic drift will be

$$\begin{aligned} D' &= x_1 + e_1 - (p + e_1 + e_2)(q + e_1 + e_3) \\ &= D + e_1(1 - p - q) - e_2\, q - e_3\, p + (e_1 + e_2)(e_1 + e_3) \end{aligned} \tag{VIII-53}$$

The e_i are the changes that result from (in the diploid case) multinomial sampling of $2N$ of the haplotypes to be the genotypes of the adults. Each of the e_i has expectation zero. The variance of e_i is the same as the variance of its x_i, the usual multinomial variance $x_i(1 - x_i)/(2N)$. They also covary: the covariance of e_i and e_j is $-x_i\, x_j/(2N)$. Putting these together, it can be shown that $\mathbb{E}[D'] = 0$. To compute the $\mathbb{E}[(D')^2]$, and thus consider the expectations of all the products of terms in equation (VIII-53). Some of these have three of the e_i, terms such as $\mathbb{E}[e_1^2 e_2]$. All of those have expectations with coefficients involving $1/N^2$ or even higher powers. We will ignore those. After much tedious algebra (and remember, this is the easiest case) we get the variance we need to

compute:

$$\text{Var}\,[D'] \;\approx\; \frac{p(1-p)q(1-q)}{2N} \tag{VIII-54}$$

That holds at (or near) $D = 0$, and, as we have omitted terms in higher powers of $1/N$, for large N. What does it mean about the standing variability in D? Naively, we can model D as being multiplied each generation by $(1-r)$, and then having a random amount ε added whose variance is approximately $p(1-p)q(1-q)/(2N)$. So if we imagine this process continuing until an equilibrium variance of D is reached,

$$\text{Var}\,[D] \;=\; \text{Var}\,[(1-r)D \,+\, \varepsilon] \;=\; (1-r)^2\text{Var}\,[D] \,+\, p(1-p)q(1-q)/(2N), \tag{VIII-55}$$

since ε does not covery with D. Solving this for $\text{Var}\,[D]$ we get

$$\text{Var}\,[D] \;=\; \frac{p(1-p)q(1-q)}{2N(1-(1-r)^2)} \tag{VIII-56}$$

or, to good approximation when r is small,

$$\text{Var}\,[D] \;=\; \frac{p(1-p)q(1-q)}{4Nr} \tag{VIII-57}$$

These are approximations: for small $4Nr$ they give a variance higher than can actually be achieved, since D cannot exceed $\pm\frac{1}{4}$. But they give us an idea when to expect substantial disequilibrium to be maintained by genetic drift, namely, when $4Nr$ is not large. An approximation similar to this was introduced by Sved (1971).

A NUMERICAL EXAMPLE. As an example, suppose that humans have an historical effective population size of $N_e \approx 10,000$, and note that there is approximately one recombination per 10^8 nucleotides in the human genome. We can take the scale over which we expect noticeable linkage disequilibrium to be the distance along the genome at which $4Nr = 1$. If we take the recombination fraction between points B bases part as $r = B \times 10^{-8}$ then $4Nr = 1$ when $40,000 \times B \times 10^{-8} = 1$ which is $B = 2,500$. This calculation was first made by Hill and Robertson (1983). They assumed $N_e = 1,000$ which gives a shorter distance, $B = 250$.

This is a small expected length of tracts of linkage disequilibrium. Disequilibrium in the human genome extends much further, often tens of kilobases. One likely explanation for this is that our assumption of a uniform rate of recombination is oversimplified. If recombination actually has "hotspots", with low-recombination regions in between, then we expect longer stretches of disequilibrium when we are in between hotspots, and we expect disequilibrium to have difficulty extending across a hotspot.

In chapter X, when we discuss coalescent genealogical trees of genes with recombination, we will return to this calculation and see that it is directly relevant there. As we will see there, treelike genealogies of haplotypes and regions with strong linkage disequilibrium are really the same thing.

WHY THIS IS NOT QUITE RIGHT. The argument leading to equation (VIII-57) sounds reasonable, but when examined more carefully, it falls apart. We have casually assumed that the gene frequencies remain at p and q, when there is no force holding them there. As the gene frequencies drift and the haplotype frequencies drift, there will be a distribution of D around zero. In the long run the individual loci start to reach fixation. As soon as one locus loses its variation, the value of D is necessarily zero. If we consider the r^2 measure of relative linkage disequilibrium, which is standardized by $\sqrt{p(1-p)q(1-q)}$ (Hill and Robertson, 1968) or the D' measure, which is standardized by the maximum value that it could have given its sign and the gene frequencies (Lewontin, 1964a), these become $0/0$ as soon as one locus becomes fixed for one allele.

Exact equations can be derived for the expectations of D^2 in a population undergoing pure genetic drift without mutation (Hill and Robertson, 1968). These involve the expectations of three quantities, D^2, $D(1-2p)(1-2q)$, and $p(1-p)q(1-q)$. Matrix equations can be set up iterating the expectations from generation to generation, but there is no analytical expression giving their expectations t generations in the future. The expectation of D^2 rises and then falls away to zero. Computer simulations show that $r^2 = D^2/p(1-p)q(1-q)$ gradually approaches a stationary distribution as drift continues. Progress has only recently begun to be made on computing the expectation of r^2 in a future generation (Song and Song, 2007), but Hill and Robertson's iteration equations can be used to approximate it from the ratio of expectations of numerator and denominator.

Ewens (2004, section 6.6) discusses work by Ohta and Kimura (1979a, 1979b) which uses diffusion equation methods to obtain analytical expressions for the expectations of the three quantities, although it requires solving a cubic equation. Hill and Weir (1988) give a fairly general treatment of four loci, obtaining equations for the expectations of variances and covariances among pairwise D. They allow for mutation, genetic drift, and recombination. As in the two-locus case, explicit formulas for these quantities t generations from now are not available.

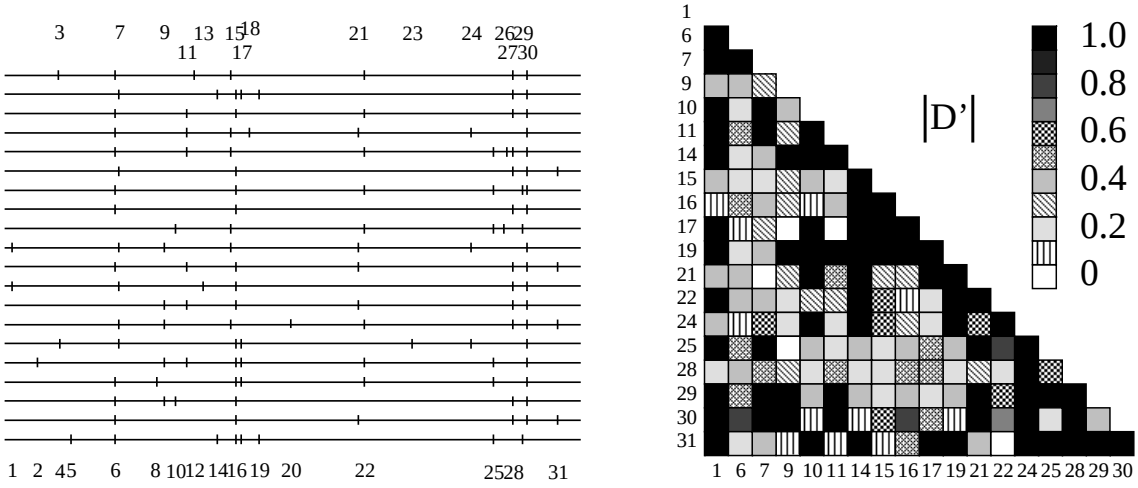
An alternative approach to analytical treatment of the expected degree of disequilibrium involves computing the joint probabilities of identity by descent at two linked loci. This too involves iterating three probabilities of dual identity, ones in which the two loci are spread over 2, 3, or 4 haplotypes, and here too analytical formulae are hard to come by (Weir and Cockerham, 1969). These different approaches are all, in some sense, equivalent.

VIII.8 Drift, linkage disequilibrium, and selection

Once linkage, genetic drift, and selection all interact, it should not surprise anyone that the outcome is hard to model and hard to discuss. There are, however, some important

The patterns of linkage disequilibrium produced by genetic drift are often summarized as the creation of "haplotype blocks" in which groups of adjacent sites are in strong linkage disequilibrium, while between these groups linkage disequilibrium is said to be near 0. The International HapMap Consortium (2005) project has worked to identify "tag SNPs" that would be highly correlated with the common haplotypes in these blocks.

It is not entirely clear that we should expect a block structure of haplotypes when linkage disequilibrium is produced by genetic drift and eroded by recombination. Below you will see haplotypes that resulted from genetic drift on a simulated 10,000-site genome in which we have sampled 20 genomes, in which 31 SNPs were seen. (I used Richard Hudson's seminal `ms` program with $4Nu = 0.001$ and per-site recombination rate of $4Nr = 0.01$). The left diagram shows the haplotype patterns, the right shows the extent of disequilibrium between all pairs of the 19 informative sites, as indicated by the absolute values of D'. Sites that have only one copy of the rarer SNP allele have been omitted from the second figure, as they necessarily show $D' = \pm 1$ with all SNPs.



Simulated haplotypes in a recombining region of chromosome with the parameter values described in the text, and the resulting patterns of disequilibrium.

A pure block structure would show strong triangles along the diagonal, with little disequilibrium elsewhere. This is not quite what is seen in the simulation. However this test is a bit unfair: the recombination rate chosen is higher than in most places in the human genome. Furthermore the presence of recombination "hot spots" with low-recombination regions in between will favor the occurrence of haplotype blocks for which tag SNPs should be useful.

Box 4: Haplotype blocks?

cases that can be discussed straightforwardly.

HITCHHIKING, SELECTIVE SWEEPS, AND PERIODIC SELECTION.

If genetic drift produces random linkage disequilibria, sometimes of one sign, sometimes another, does this affect the course of natural selection at a typical locus? The answer is yes (as you should have suspected from the very fact that I raised it here). The easiest place to see this is in the phenomenon of hitchhiking. It first was noticed in experimental evolution in bacteria in chemostat populations. A mysterious phenomenon called *periodic selection* was encountered. A population of bacteria, one which had little or no recombination, had a locus with two alleles whose frequencies were being monitored through time. The frequency remained relatively constant, because the locus was near enough to being neutral. Then, suddenly, one allele began to rise rapidly in frequency. The phenomenon was first discussed by Atwood et. al. (1951) who found some cases in which these perturbations succeeded one another in periodic fashion, hence the name "periodic selection".

The simplest explanation is that a favorable mutant has arisen at another locus and rapidly increased in the population. Whichever haplotype it has arisen in rapidly takes over the population. Thus one of the alleles at the marker locus suddenly increases. When there is no recombination, that allele at the marker locus will go to fixation, limited only by the rate of mutation away from that allele at the marker locus, or by mutation at the selected locus.

A subpopulation of haplotypes largely taking over the whole population is referred to as a *selective sweep*. The lucky allele at a neutral marker locus increases as a result of *hitchhiking*, as it is merely a passenger in a fast-moving selective sweep.

When there is recombination in the population, the hitchhiking effect is less complete. This has been investigated by Maynard Smith and Haigh (1974) and by Thomson (1977). As the favored haplotype increases, recombination gradually redistributes the neutral hitchhiking allele from it to the nonfavored haplotype, and the favored haplotype also comes to have more and more of the nonhitchhiking allele. In the end, the increase of the hitchhiking allele is limited – it does not reach fixation. The favored allele at the selected locus sweeps through the population, but the nearby neutral allele stops short of fixation.

Figure 8.6 shows the final frequency of a neutral allele that occurs only in the favored haplotype, which sweeps upward from an initial frequency of 0.001, initially present in all copies of that haplotype. The recombination fraction between the locus under selection and the neutral locus is shown, on a logarithmic scale. What is noticeable is that more selection means that the region around the selected locus in which linked neutral alleles are made substantilly more frequent by the selective sweep is proportionately larger.

How much the hitchiking allele ends up increasing in the population depends on the selection coefficient s of the favored haplotype, and the recombination fraction r between

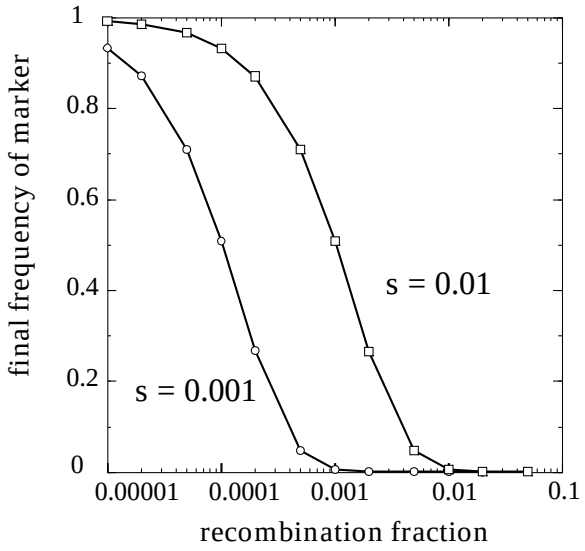


Figure 8.6: Final frequency reached by a neutral marker allele, when that allele is present in all copies of the initial haplotype in which a favored allele occurs at a nearby locus, and only there. The initial frequency of the haplotype is 0.01, and the selection coefficient favoring it (s) is either 0.001 or 0.01. The final frequency of the nearby neutral allele is plotted against the recombination fraction between the two loci. The curves decline with more recombination until they approach the initial frequency of the haplotype, 0.01.

the hitchiking locus and the selected site on the favored haplotype. In fact, it depends almost entirely on the ratio s/r. If s is doubled, the favored haplotype moves through the population roughly twice as fast, as we have seen in Chapter II. The amount of gene flow at the neutral locus into the favored haplotype is less the faster that haplotype increases, as then there is simply less time for recombination to move gene copies at the neutral locus into the favored haplotype from the less favored haplotype. To have the same amount of net gene flow (at the neutral locus) into the favored haplotype, we need to have twice as much recombination. So the ratio s/r is important to the net effect on the hitchiking neutral locus.

If s/r is large, there will be a large hitchhiking effect and the marker alleles in the original favored haplotype will be swept nearly to fixation. If s/r is small, the population composition at the neutral marker locus will be little changed by the selective sweep. The length of chromosome that is swept through the population is enough that the recombination fraction between the loci is proportional to s. Note in this numerical example that nearby markers with recombination fraction as large as $s/10$ are swept up

to a final gene frequency of 1/2.

Another consideration is the initial frequency of the favored haplotype. If it is very low, it will take a longer time until it approaches fixation, with correspondingly more gene flow between favored and unfavored haplotypes and thus less net effect on the population frequency of the marker locus. So the amount of recombination necessary to prevent a selective sweep from having much effect is not only a function of s/r. We must also take into account effects of random genetic drift in the period when the favorable allele is rare. If it survives this period, it will rise more rapidly toward fixation than is implied by the deterministic theory used here. Durrett and Schweinsberg (2004) have pointed out that the deterministic approximation can be a bad one in this case, and they have suggested an alternative approximation.

Repeated effects of hitchhiking by selection at nearby loci are also called *genetic draft*, in analogy to genetic drift. They could cause large stochastic changes in gene frequency even when the population size is large. John Gillespie (2000, 2001) has named this phenomenon and made the most serious investigations of it.

There is a similar phenomenon when geographic differences in fitness, a locus with different alleles adapted to local populations induces a reduction in effective migration rate among populations. This happens by selection against migrants. In effect it is a selective sweep frozen in place by the fitness differences. Petry (1983) has showed that this induces a reduction of the effective migration rate among local populations, simply by reduced fitness of the immigrant alleles at one locus, and that this affects the distribution of neutral genetic variation at nearby loci that are closely enough linked.

A related phenomenon, "gene surfing", is the spread of alleles, and the stripelike patterns of their distribution, when one population invades nearby smaller populations. Edmonds, Lillie, and Cavalli-Sforza (2004) emphasized the importance of gene surfing as an explanation of these patterns. Hallatschek and Nelson (2008) developed predictions for these patterns. Note that, unlike hitchhiking, gene surfing results from the spread of a population, not from linkage disequilibrium of a neutral locus with an advantageous one.

THE HILL-ROBERTSON EFFECT. What about the effect of two selected loci on each other? This was investigated by Hill and Robertson (1966). They discovered a phenomenon which has come to be called the Hill-Robertson Effect. Two closely linked selected loci each interfere with the effectiveness of selection at the other locus. This effect is strongest when the recombination fraction is small enough that $4Nr$ is small, and the effect of selection on the other locus is substantial when $4Ns$ is greater than 1.

I will give one verbal argument and one simple derivation for a special case to persuade you of this. The verbal argument considers the effect of a selected locus on the effective population size which is relevant to the other locus. We have already seen in Chapter VI that variation in fitness from individual to individual reduces effective population size (see, for example, equations VI-49 and VI-53). Note that the variation

could be environmental, or it could be the result of genetic variation at other loci. If the background loci are far away, their genotypes in the descendants change greatly over a rather short time scale. An haplotype in this generation may carry allele B – in a few generations a typical descendant may instead carry allele b at that locus. But if the background loci are closely linked, B hangs around for many generations in the descendants. This magnifies its effect. If B raises the fitness, it continues to do so for some time, and similarly if allele b lowers the fitness, descendants have their fitness lowered as well.

Thus linked background loci can cause large swings up and down of haplotype frequency. It is as though effective population size at a locus were small. Each locus does this to its neighbors on the chromosome.

A simple case. Hill and Robertson gave approximations, verbal arguments, and computer simulations. We can demonstrate the phenomenon more precisely in one simple case. Suppose that we have a haploid population with two loci that are completely linked. Each has a favored allele that increases the fitness by a fraction s. Both loci initially have precisely one copy of the favored allele (there is no additional mutation beyond the occurrence of these favored alleles).

If we have one copy each, and there is no initial association between the favored alleles, there are two cases. $1 - 1/N$ of the time the two copies are on different haploid genomes. $1/N$ of the time they are on the same haploid genome. Now use the haploid version of Kimura's formula for fixation probability. This is equivalent to equation (VII-91) but with the population size N halved. Thus

$$U(p,s) \;=\; \frac{1 \;-\; \exp(-2Nsp)}{1 \;-\; \exp(-2Ns)} \tag{VIII-58}$$

Considering a favored mutant at one of our two loci, in the case where the mutants are on different haploid genomes, it is as if they were copies of the same allele. The probability that one or the other of them fixes is $U(2/N,\, s)$, and if it does, the probability that it is the copy at the particular locus that we are watching is $1/2$. If the two copies are in the same haploid genome, the chance that they fix is $U(1/N, 2s)$, and then they do the particular favored mutant fixes. So the net probability that it fixes is the weighted average of these fixation probabilities:

$$\left(\frac{1}{N}\right)\; U(1/N, 2s) \;+\; \left(1 \;-\; \frac{1}{N}\right)\; \frac{1}{2}\, U(2/N, s) \tag{VIII-59}$$

We can compare this to the fixation probability of a single copy of the favorable allele at one locus, with no other selection happening nearby. This is simply $U(1/N, s)$.

Figure 8.7 shows the ratio of the fixation probabilities of the two-locus and one-locus cases, when $N = 100,000$ and various values of the selection coefficient s. For small s, the one- and two-locus cases have almost the same fixation probability. As s passes 0.1,

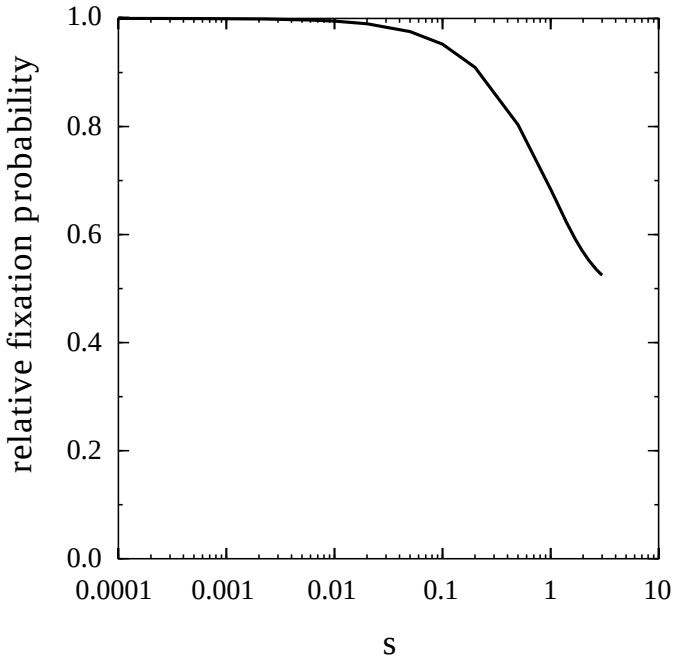


Figure 8.7: The fixation probability of a single copy of a favored allele when a favored allele in a single copy with the same selection coefficient (s) is tightly linked but not on average associated with it. What is plotted is the relative fixation probability, compared to the case where there is no other selected locus nearby. The haploid model discussed in the text is used, with a population size of $100,000$. Computed using equations (VIII-59) and (VIII-58).

the selection at a nearby locus starts to have a noticeable effect in reducing the fixation probability.

What is happening is that genetic drift creates random linkage disequilibrium between the two loci. Sometimes it is coupling disequilibrium, with the two favored alleles on the same haplotype. Sometimes it is repulsion disequilibrium, with the two favored alleles on different haplotypes. In the first case the two favored alleles help each other – the fitness difference between A and a is twice as great as it would be without the disequilibrium. Selection at each locus helps the favored allele at the other. In the second case, the repulsion disequilibrium means that selection at each locus slows the response at the other one.

You might suspect that this will all cancel out in the end, that the net result will be that the change at each locus is not affected by the selection at the other. The results of Hill and Robertson, and the simple calculation presented here, show that this is not so –

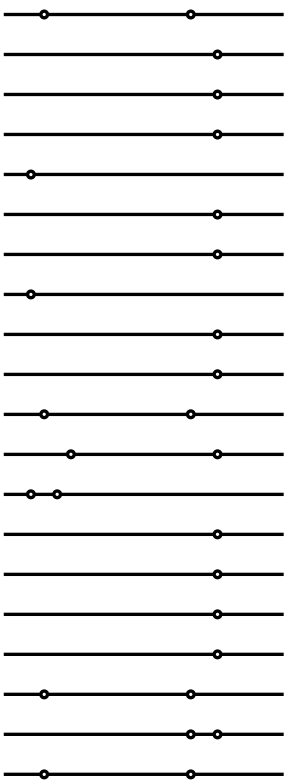


Figure 8.8: Haplotypes from an asexual haploid population of size 20, from a simulation of natural selection against deleterious mutants in which the population has just reached a state in which Muller's Ratchet advances, with each haplotype containing at least one deleterious mutation. The circles are the deleterious mutants.

the negative effects of repulsion disequilibrium are stronger than the positive effects of coupling disequilibrium. The more closely the two loci are linked, the longer the random disequilibrium will persist, and the larger its net effect will be.

It is important to realize that, although this is an effect of linkage disequilibrium, it is quite distinct from the deterministic effects of linkage disequilibrium caused by interaction of the fitnesses at the two loci. The Hill-Robertson effect occurs even when the loci have multiplicative fitnesses (so that their fitnesses in effect do not interact). When there are both deterministic and random linkage disequilibria, it will often be the case that the random linkage disequilibrium will cause a Hill-Robertson effect that will be far more noticeable than the effects of the deterministic linkage disequilibrium.

IMPLICATIONS OF THE HILL-ROBERTSON EFFECT. The Hill-Robertson effect oc-

curs in any case where linked loci are undergoing natural selection and genetic drift. As such, it shows up in many guises. Some are:

Muller's Ratchet H. J. Muller (1958, 1964) noticed that when deleterious mutations accumulate in a nonrecombining genome, it is possible to get into a state in which all haplotypes have at least one deleterious mutant. Natural selection alone cannot then eliminate all deleterious mutations. He noted (1964, p. 8) that

> If we disregard advantageous mutations ... we find that an asexual population incorporates a kind of ratchet mechanism, such that it can never get to contain, in any of its lines, a load of mutations smaller than that already existing in its at present least-loaded lines.

This has come to be called "Muller's Ratchet". Figure 8.8 shows a population of asexual haploids at the moment when the mutant-free chromosomes have been lost. Every chromosome has at least one deleterious mutant allele, though no locus is yet fixed for the deleterious allele. The ratchet has advanced; from now on every haploid genotype will contain at least one deleterious mutant allele. As further deleterious mutations occur, the ratchet will operate repeatedly. The species continually loses ground. The ratchet can be unsprung by recombination, which can reintroduce haplotypes that have no recombination. In the absence of recombination only back mutation can undo the ratchet, and that will be a weak force unless most loci have accumulated deleterious mutations.

The Fisher-Muller explanation for the evolution of recombination Earlier, R. A. Fisher (1930) and H. J. Muller (1932, 1958) had described another related phenomenon. They both, apparently independently, realized that if there were selective sweeps occurring at multiple tightly-linked loci, that the selection for favorable alleles at one locus could interfere with the response at another. If there is no recombination, and two favorable alleles arise, they will mostly arise in different haplotypes, and then both cannot reach fixation – one or the other would ultimately be lost. Of course, there is a small chance that the second of them to arise would arise in the descendants of the first, in which case they would help each other reach fixation. They pointed out that this showed a strong advantage of recombination – favorable mutants arising in different individuals can become combined into one genome. This was the first valid explanation for the evolution of recombination based on its genetic effects. (Previous arguments that recombination "creates variation" were based on misunderstandings of the population genetics of multiple loci, and were not correct – unfortunately those arguments are still often found in textbooks).

The Fisher-Muller phenomenon was, like Muller's ratchet, a special case of the Hill-Robertson effect. In a haploid case, in the Fisher-Muller case the favored alle-

les start at frequency $1/N$ each. In the Muller's ratchet case they start at $1 - 1/N$ instead. Otherwise the advantage of recombination has the same source. Crow and Kimura (1965) gave approximate formulae for the degree of advantage of recombination. Felsenstein (1974) modified these and showed the results of computer simulation verifying the reality of the phenomenon. Felsenstein and Yokoyama (1976) showed that it could also lead to selection for a modifier of the amount of recombination.

Background selection Deleterious mutations occur throughout the genome, and are generally held to low frequency by countervailing natural selection. It is to be expected from the Hill-Robertson effect that in regions of the genome in which recombination is restricted, random linkage disequilibrium between the deleterious alleles will on average lead selection against them to be less effective. This has come to be called "background selection". Charlesworth (1994) argued that, for loci that are tightly linked, it is approximately equivalent to a reduction of the effective population size to the fraction of mutation-free haplotypes.

The degeneration of Y chromosomes A dramatic application of the Hill-Robertson effect was made by Charlesworth (1978b). Although attempts had been made to model the degeneration of Y chromosomes in X/Y sex determination systems, there had been no convincing explanation for the fact that functional genes tend to largely disappear from the Y chromosomes, but to remain on X chromosomes. Charlesworth realized that Muller's ratchet provided an explanation. Y chromosomes are under strong selection to have no recombination with the corresponding X chromosomes. Thus Y chromosomes become clonally reproducing, and they never have opportunities to recombine with each other. X chromosomes do have opportunities to recombine with each other in females. The Y chromosomes are then subject to Muller's Ratchet, which makes the functional gene copies disappear, leaving behind a genetic desert, with only sex-determining genes maintained. As the functional genes disappear, there will be stronger and stronger selection for "dosage compensation" for them, in which single copies on the X are made to function as well as two copies do in the female. Charlesworth's paper is a landmark in the understanding of the evolution of Y chromosomes.

The same phenomena also argue for a tendency of areas of the genome with very low recombination rate to have genes in that region become inactivated. This has been used by Charlesworth, Sniegowski, and Stephan (1994) to explain the tendency of transposons and repeated DNAs to accumulate near centromeres and near telomeres of chromosomes.

VIII.9 Migration and linkage disequilibrium

Selection and genetic drift can both create linkage disequilibrium. So can migration. I have already discussed this in Chapter IV. It was discussed by Cavalli-Sforza and Bodmer (1971, p. 69) and by Prout in Mitton and Koehn (1973). We have seen in equations (IV-10) and (IV-11) that, even if there is no linkage disequilibrium in any population, a mixture of populations can be in linkage disequilibrium if the gene frequencies of alleles at the two loci covary across populations. Thus, if we have one population that is all $AABB$, and another that is all $aabb$, then each is in linkage equilibrium. But if we make a mixture of the two (in any nontrivial proportion), there will be strong linkage disequilibrium in the mixture. If gene flow continues into the admixed population, this will continually reinforce the linkage disequilibrium even as it dies away owing to random mating within the populations. As the gene frequencies in the source populations become more similar, the amount of linkage disequilibrium created by migration becomes smaller and smaller, so ultimately all of the disequilibrium disappears.

One situation in which linkage disequilibrium will be maintained is when natural selection keeps the gene frequencies in the populations different. The linkage disequilibrium within each population is then maintained at an equilibrium, where it is both dying away by recombination and also constantly replenished by immigration. This has been discussed by Li and Nei (1974) and by Feldman and Christiansen (1974), and most completely for the case of selective clines by Slatkin (1975).

Exercises

1. Suppose that we have an infinite haploid population, and two loci that are completely linked (with no recombination between them), and that there are two alleles at each locus (A and a, B and b). If the initial haplotype frequencies of AB, Ab, aB, and ab are 0.01, 0.03, 0.03, and 0.93, and the fitnesses of the haploid genotypes are

AB	Ab	aB	ab
2	1	2	1

   if we wait 5 generations, what will be
   (a) The gene frequency of A ?
   (b) The gene frequency of B ?
   (c) The crossproduct ratio measure of linkage disequilibrium ?
   (d) The usual measure of linkage disequilibrium D ?

2. Suppose that in an infinite haploid population, with two loci, each of which has two alleles, that the fitnesses of the four possible haplotypes are

AB	Ab	aB	ab
1.1	1	1	1.1

If we start with equal frequencies of each of these four haplotypes, and no recombination between the loci what happens to the haplotype frequencies? What would happen (qualitatively) if the population were instead large but finite?

3. In the cases of the previous question, what would happen if there were a small amount (say 0.01) of recombination between the loci?
4. In an infinite diploid population, there are two loci, each with two alleles, and at each locus the fitnesses are $1/2 : 1 : 1/2$. The two loci do not interact, so that the fitnesses are multiplicative across loci (the fitness of $AA\,BB$, for example, is $1/4$). Use the Lewontin-Kojima results to predict the linkage disequilibrium between these loci, and how this depends on the recombination fraction r.
5. A locus is a "balanced lethal system" if there are two alleles, and all homozygotes die. If two balanced lethal loci are linked to each other with recombination fraction r, in an infinite population. Is there a value of the recombination fraction below which there will be expected not to be linkage equilibrium? Assume, of course, that only double heterozygotes survive. Owing to having only two possible genotypes of the surviving adults, you should be able to work this out exactly.

Complements/Problems

1. For the case in Exercise 2 above, where the haplotype frequencies have fitnesses $1.1 : 1 : 1 : 1.1$, and the recombination fraction is r, calculate what the equilibrium haplotype frequencies will be as a function of r for the interior equilibrium at which the gene frequencies at both loci are $1/2$. Is this situation stable?
2. When I calculated the conditions for increase of a rare advantageous haplotype containing two otherwise-deleterious alleles in equation (VIII-15), I casually dropped the equation for haplotype aB. Check whether this is justified by writing a similar set of equations with the three rare haplotypes AB, Ab, and aB, dropping all quadratic terms in the equations. Show that the expressions for the eigenvalues of this 3×3 matrix yield the same conditions as before.
3. Consider a neutral locus with two alleles, B and b, which is tightly linked to an overdominant locus with two alleles, tightly enough that there is no recombination. Describe some of the possible outcomes if we start with all haplotypes present, in a finite population. Include a description of the final equilibrium states, as well as states that persist a long time, but not indefinitely.
4. Set up the equations for one generation of change in the frequencies of the haplotypes at two linked loci, each with two alleles, where one of them is under (haploid) natural selection with fitnesses $1 + s : 1$ and the other one is neutral. Can you write a computer program to iterate these equations for t generations? If we start with a population that is mostly ab with a low frequency (such as 0.001) of AB, can

you verify some of the final gene frequencies of the neutral allele after the selective sweep is over, as shown in Figure 8.6?

Chapter IX

QUANTITATIVE CHARACTERS

IX.1 What is a Quantitative Character?

We have seen the complexity of treating the effects of natural selection with multiple loci. Hardly any generalities are available in such a case. Yet animal and plant breeders need to be able to predict the outcome of artificial selection on traits of economic interest, traits which are undoubtedly affected by many loci and alleles. Evolutionists need to be able to interpret natural variation in measurable characters and make statements about the strength of the evolutionary forces involved in the maintenance of the variation. Human geneticists are also frequently confronted with traits which are *polygenic*, that is, affected significantly by many loci. They must be able to compute the probabilities of various outcomes in an unborn child, or in an individual thought to be at risk of developing a condition later in life.

If so little could be said about outcomes when we knew the fitnesses of all genotypes and their frequencies, we at least had recourse to strongarm methods for computing gamete frequencies in successive generations. This is inelegant and computationally difficult, but at least it is possible when all fitnesses and genotype frequencies are known. Unfortunately, in the particular applications just mentioned, this is not the case. In each case we are usually dealing not with discrete phenotypes, but with a continuously measurable trait, such as height, length of jaw, or blood glucose concentration. Sometimes we are instead measuring a count of discrete entities. These cases are respectively that of a *quantitative character* and of a *meristic* character. Even when we believe ourselves to be dealing with truly discrete phenotypes, we are often actually measuring a quantitative character. Recall that Mendel's original characters included the height of the pea plant. The key to his ability to make a single-gene analysis of this character was the fact that the distribution of available phenotypes resolved itself into two distinct peaks, which could be assigned the names "tall" and "short". There were so few plants of medium height that Mendel was able to discretize this particular quantitative character.

Figure 9.1 shows another, hypothetical example. Here we have a single locus with two alleles, B and b. Each of the three genotypes has a particular distribution of phenotypes. There is no general rule as to what the distribution of phenotypes will be which we get from a given genotype. However we should be able to characterize such a distribution by, among other things, its mean. The means of the distributions in our example are:

BB	Bb	bb
5	4	2

In the different parts of the Figure we see cases with different amounts of variability in each of the three distributions. Of course, when we observe a population with a certain gene frequency (in the Figure it is 0.4) we cannot tell which individuals come from which distributions if the distributions overlap. As the variability of the distributions increases, we find first that there comes to be overlap, with ambiguity as to the genotypes of many individuals whose phenotypes we observe. But at this point we can still tell that there are three component peaks in the overall distribution of phenotypes. However, when variability of the individual genotypes becomes greater still, the peaks disappear and all we see is one broad peak. The same thing happens if there is more than one locus. If the genotypic means are:

	BB	Bb	bb
AA	6.5	6	4
Aa	5.5	5	3.5
aa	4.5	3	2

and we have linkage equilibrium with $p_A = 0.4$ and $p_B = 0.5$, then Figure 9.2 shows the distributions of phenotypes which we might see with different amounts of variability in the distribution of phenotypes produced by a genotype. Once again, we see distinct nonoverlapping phenotypes when the variability is small, but as the variability increases our ability to identify genotypes declines, until ultimately there is only a single smooth distribution.

If this is the situation when we observe a quantitative character, how can we have any hope of predicting effects of selection or phenotypes of relatives if we are ignorant of the exact genetic basis of a trait, and therefore also of genotype frequencies? In the general case, there is nothing that can be done. But if we are willing to make a certain kind of oversimplified model of the way genes and environment act to determine the phenotype, we find that general rules do exist, providing us with useful guidelines for plant and animal breeding and medical genetics. It is with this approximate approach that we concern ourselves for the rest of this chapter.

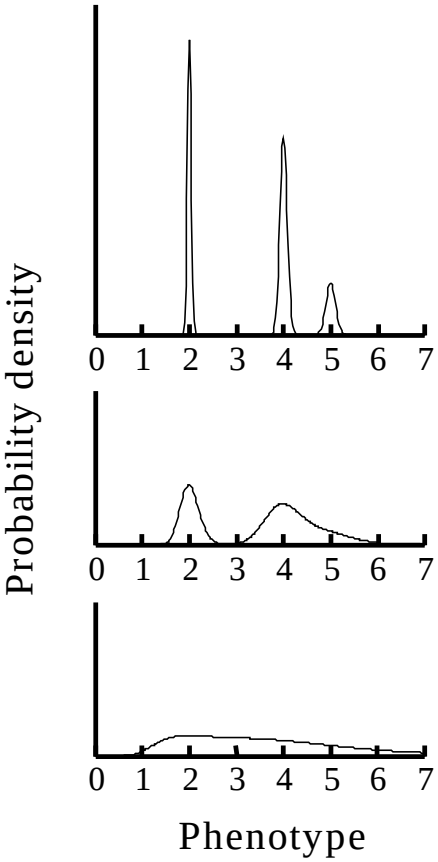


Figure 9.1: Phenotype distributions from a one-locus genetic model with various amounts of variability in the distributions produced by the genotypes. There are three genotypes *BB*, *Bb*, and *bb* whose mean phenotypes are 5, 4, and 2. The gene frequency of *B* is 0.6. The component distributions used here are lognormal with equal variances on the log scale. Standard deviations on the log scale are, for the top, middle, and bottom graphs, 0.02, 0.10, and 0.35.

IX.2 The Model

Our model the quantitative character will be determined by some number of loci, plus some environmental influences. We make a series of rather restrictive assumptions:
Assumption No. 1: The phenotype is the sum of effects contributed by each of the n loci, plus an environmental effect, so that we may write

$$P = \mu + g_1 + g_2 + \cdots + g_n + e \qquad \text{(IX-1)}$$

where μ is an arbitrary starting point (not necessarily the population mean). This is really

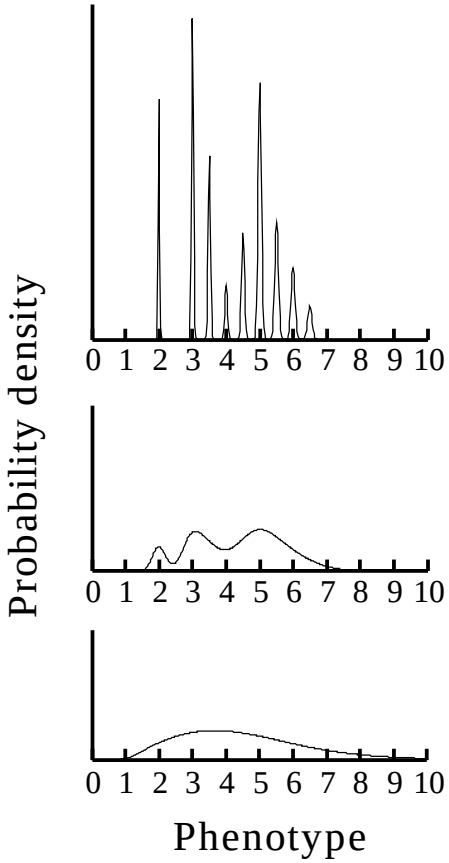


Figure 9.2: Phenotype distribution produced by a two-locus model. See text for details of the genetics. (In this example lognormal distributions with equal variances in the log scale are used for the distributions of the phenotypes). Standard deviations on the log scale are, for the top, middle, and bottom graphs, 0.01, 0.10, and 0.3.

a very special assumption. It places strong constraints on the kinds of gene interaction which may be present. In real life genes concerned with a trait may interact in wondrous ways. But in our model, the effect of changing the genotype at one locus is always to add or subtract the same increment to the phenotype. For example, with four loci the following scheme is one that satisfies our assumptions:

$$P = 7 + \begin{Bmatrix} 2 & \text{if } AA \\ 0 & \text{if } Aa \\ \text{-1} & \text{if } aa \end{Bmatrix} + \begin{Bmatrix} 1 & \text{if } BB \\ 0 & \text{if } Bb \\ \text{-1} & \text{if } bb \end{Bmatrix} + \begin{Bmatrix} \text{-2} & \text{if } CC \\ \text{-2} & \text{if } Cc \\ 0 & \text{if } cc \end{Bmatrix} + \begin{Bmatrix} 0.3 & \text{if } DD \\ 2 & \text{if } Dd \\ 0.3 & \text{if } dd \end{Bmatrix} + \begin{matrix} \text{environmental} \\ \text{effect} \end{matrix} \qquad \text{(IX-2)}$$

This scheme predicts that the phenotype of an *AA Bb cc DD* individual will be 7 + 2 + 0 + 0 + 0.3, plus an environmental effect. Similarly we can determine from this scheme the genotypic contribution to the phenotypes for each of the other 80 possible genotypes. Note that the 81 genotype contributions are here specified by 12 quantities, so we immediately see that not all phenotypic schemes can be specified in this fashion. For instance, try as you may you will not be able to find two sets of 3 contributions each which will result in the phenotypes in the two-locus example of the previous section, the example which gave rise to Figure 9.2. To prove that it can't be done, you should (for example) compare the effect of substituting *Bb* for *BB* in a genotype which has *AA* at the other locus, with the substitution of *Bb* for *BB* in an *aa* individual. Under our model both of these substitutions must have the same effect, which they do not have in this two-locus example.

Assumption No. 2: The genotypes at the n loci are independent of each other.

This amounts to the statement that the population is in linkage equilibrium at all combinations of loci. It will be violated by any force which tends to produce linkage disequilibrium, such as selection of many types, random genetic drift, and migration. In animal and plant breeding, and also in human genetics, artificial crossing or human migration is quite likely to result in a population which is initially in linkage disequilibrium and has not had enough time to come back to linkage equilibrium.

Assumption No. 3: The environmental contribution to the phenotype is drawn from a distribution independently of the genotype and independently of the environmental contributions in other individuals.

This is the assumption most frequently violated in a serious fashion when quantitative genetics theory is used to analyze data. It can be seen that the general tendency of these assumptions is to erect a model of a phenotype determined by additive independent causes, with only the genetic factors being shared among relatives. As we shall see, the correlations among relatives can then be used to make some statements about the genetics of the trait. But when environmental factors act which are common to relatives, then unless this is known it may cause us to mistake the resulting correlations of phenotypes for evidence of genetic factors, if we are mistakenly making Assumption 3.

In addition to excluding environmentally-based correlations of relatives, these assumptions exclude interactions between loci by requiring locus effects to be additive, and they also exclude correlation of environmental effects with genotypic effects; this may be violated if the presence of a particular genotype makes more likely the presence of a particular environment. Such a correlation leads to confounding of the effects of these factors, and consequent inability to distinguish them.

Note, however, that dominance (an interaction between the two alleles at a single locus) is not excluded. In the four locus scheme in (IX-2) *C* is completely dominant over *c* with respect to their contribution to phenotype *P*. *D* and *d* are overdominant, *a* is partially dominant over *A*, and there is complete absence of dominance at the *B* locus.

This ability to include dominance leads to many interesting complications.

SCALE TRANSFORMATIONS. Everything said so far assumes that that the genes act additively on the scale which we happen to be measuring. But it is by no means obvious that the scale we measure is the scale on on which additivity occurs. Suppose that we measured the weight of an animal. The weight will be closely related to body volume in most cases. But why do we assume that genes add increments to the volume? Could they not as easily act additively on the linear dimensions of the organism, with the volume (and hence the weight) being simply the cube of a linear size measurement? Is it not even possible that the genes are additive on the cross-sectional area, with volume being proportional to the $3/2$ power of this quantity, and length to its square root? Faced with this diversity of possibilities, what are we to do?

If we knew something about the way genes acted in contributing to the character, we could gain some insight into what is the proper scale. But this presumes the very thing we are most likely not to know, once we have been forced to the unpleasant expedient of using quantitative genetic theory. Alternatively, we can make use of a family of scale transformations such as the power transformation

$$y = \frac{(x^p - 1)}{p} + 1 \qquad \text{(IX-3)}$$

which gives us a wide variety of scale changes by varying one parameter, p. When $p = 1$ it is no change of scale at all. When $p = 1/2$ or $1/3$ it is close to the square root or cube root, and so on, and when $p = 0$ it can be shown (by L'Hôpital's Rule) that (IX-3) becomes $y = 1 + \ln x$. There is nothing in (IX-3) that prevents p from being greater than 1 or less than zero, and these may be useful regions to consider. One would in theory make analyses of one's data for various values of p, pick the value which resulted in the best fit, and then attempt to correct for the fact that p is estimated from the same data by reducing the degrees of freedom in the analysis by one.

The value $p = 0$ is special because the logarithm will often be a highly reasonable scale on which to assume additivity. In particular, if the various genotypic and environmental factors act multiplicatively on the original scale, they will be found to act additively on the logarithm of the measurement, since $\ln(xy) = \ln x + \ln y$. Multiplicative action of factors is easy to envision if we can persuade ourself that a given factor (whether genetic or environmental) acts by making a percentage increment in the trait rather than an absolute increment. This will frequently be quite reasonable. It is certainly easier to envision a genotypic change as acting to increase (say) weight by 10% of its previous value, rather than by an absolute amount of 10 grams. If so, then the change multiplies weight by 1.1, and by taking logarithms we see that it adds the amount $\ln(1.1) = 0.0953$ to the logarithm of the weight.

Multiplicative gene action is certainly more reasonable than additive gene action when we are dealing with traits which have a natural zero point. If one subtracts fixed

amounts from a quantity it may become negative, but if one multiplies by fixed positive quantities it can at worst approach zero. Thus if a particular change is supposed to have the effect of decreasing the weight by 10g, we are faced with the problem that this may be unreasonable if the weight starts out at 9g. But if it decreases weight by 20%, this may be 10g if the organism starts at 50g but is only 1.8g if it starts at 9g. We then never predict a negative weight.

Some quantitative genetic theory (that concerned with response to selection) and much data analysis requires that we assume that the character follows a normal distribution. If our character has a natural zero point, it cannot exactly follow a normal distribution, for that distribution has tails which spread out to $+\infty$ and $-\infty$. The lower tail would predict the existence of individuals with negative phenotypes. It may still be a good enough approximation to use in practice, but there is at least then some pressure (if only from slight embarrassment) to consider taking logarithms. For if the logarithm of the phenotype were the quantity which was normally distributed, then as it went to $-\infty$ the original phenotype would only approach zero. So it is entirely conceivable that a trait with a natural zero point has its logarithm normally distributed. I would go so far as to state a conclusion:

> For a trait with a natural zero point, first take the logarithm of the phenotypes and base analysis on it. Do not return to the original scale unless you can come up with positive reasons why the genetic or environmental factors are likely to act additively on that scale.

This must be taken with a grain of salt. It is certainly better than always staying on the original scale. But it would be better yet to estimate the appropriate scale from some data set, as with the family of transformations (IX-3).

IX.3 Means

A ONE-LOCUS ANALYSIS. One of the properties of an additive model of the phenotype is that it greatly simplifies formulas for the mean and variance of the character. In this section we concentrate on the mean. It is always true (even if the individual terms are not independent) that the expectation of a sum is the sum of the individual expectations, so that from (IX-1)

$$\mathbb{E}(P) \;=\; \mathbb{E}(g_1) + \mathbb{E}(g_2) + \cdots + \mathbb{E}(g_n) + \mathbb{E}(e). \tag{IX-4}$$

We are going to work with the individual terms $\mathbb{E}(g_i)$, with a view to establishing relationships which hold for a single locus. However (IX-4) will enable us to assert that these relationships also hold for the whole phenotype, so that we are accomplishing

more than we seem to be. However we must first do something about the term $\mathbb{E}(e)$. This will be the mean of the environmental effects on all members of the population. It is often conveniently disposed of by assuming that it is zero. This does not involve any extra assumption. Consider expression (IX-2) to see this point: we might assume that the effects at (say) the first locus were 9, 7, and 6 for *AA*, *Aa*, and *aa* while the mean environmental effect was (say) 2.3. But how would this differ from assuming that the effects of genotypes at the first locus were 11.3, 9.3, and 8.3 with a mean environmental effect of zero? In fact, both schemes make the same prediction of the phenotype of any particular genotype. We have simply removed 2.3 from the postulated environmental effect in each individual, only to add it to the postulated contribution of the first locus.

We could have as easily moved any amount we wanted to from one locus to another. In assuming that $\mathbb{E}(e) = 0$ we make no restrictive assumption. The reader who is troubled by the sleight-of-hand involved here may take comfort in the realization that we actually need not go through the procedure at all. All of the relationships we investigate involve the effects on the mean phenotype of a given genetic change. So we are actually investigating differences between the means under two situations. The implicit assumption is that we have changed only the genotypes and not the environments. That in turn requires that the environmental effects, or at least their mean, be independent of the genotypic effects. Then writing (IX-1) as $P = G + E$ we have

$$\mathbb{E}(P^*) - \mathbb{E}(P) = \mathbb{E}(G^*) - \mathbb{E}(G) + \mathbb{E}(E) - \mathbb{E}(E) = \mathbb{E}(G^*) - \mathbb{E}(G), \tag{IX-5}$$

where the asterisk denotes the population after some genetic change. Thus we can ignore the environmental effects in computing changes in the mean, provided our assumptions are satisfied.

INBREEDING EFFECTS. We have seen that we can gain insight into a multi-locus trait by considering one locus at a time. If we have a single locus with two alleles, and the contributions of *AA*, *Aa*, and *aa* to the phenotype are respectively a_{11}, a_{12}, and a_{22}, then the contribution of this locus to the mean is

$$\mathbb{E}(g) = P\,a_{11} + Q\,a_{12} + R\,a_{22} \tag{IX-6}$$

where P, Q, and R are the genotype frequencies. We are interested in the effects of inbreeding on the population mean phenotype. When the population has gene frequency p of A and inbreeding coefficient f, then using our standard genotype frequency formulas given in (V-2), and discovered by Sewall Wright (1921a):

$$\begin{aligned}\mathbb{E}(g) &= [p^2(1-f) + pf]\,a_{11} + 2p(1-p)(1-f)a_{12} + [(1-p)^2(1-f) + (1-p)f]\,a_{22} \\ &= p^2a_{11} + 2p(1-p)a_{12} + (1-p)^2a_{22} + fp(1-p)\,[a_{11} + a_{22} - 2a_{12}].\end{aligned} \tag{IX-7}$$

The first three terms, which lack f, are simply the mean contribution in an outbred population. Note that f enters into (IX-7) linearly, with no terms in f^2. This means that each locus has a mean contribution of the form $A + Bf$, where the gene frequencies enter into A and B. Then the overall phenotypic mean is of the form $\sum A + f(\sum B)$, which will also be linear in f. Thus we have the general result that *the mean phenotype is expected to be linearly related to the inbreeding coefficient* f.

But in which direction should the phenotype be expected to change? What will be the sign of $\sum B$? It is much less easy to come up with a general rule, but quite often $\sum B$ is negative. Note that the quantity in brackets in the last term of (IX-7) is

$$[a_{11} + a_{22} - 2a_{12}] \;=\; 2\left[\frac{a_{11} + a_{22}}{2} - a_{12}\right]. \qquad \text{(IX-8)}$$

If a_{12} is equal to the average of a_{11} and a_{22}, in other words if the heterozygote contribution is the average of the two homozygote contributions, then this quantity is zero. When a_{12} exceeds this average, then the mean declines with inbreeding (note the plus sign preceding this term in (IX-7)). When a_{12} is less than the average, inbreeding raises the population mean. This refers to a single locus only: the general picture will be that inbreeding will increase the mean contribution at some loci and decrease it at others. The net effect of inbreeding will depend on the overall sign of $\sum B$, there being no reason to expect B always to have the same sign. However there does appear to be a vague generalization available: If the trait is positively correlated with vigor, size, or fitness, inbreeding tends to reduce it. It seems that the heterozygote tends to be closer to the higher homozygote, so that B tends on average to be negative.

This observation was made in the early years of the century, by the pioneering corn geneticists E. M. East and D. F. Jones (1919), and G. H. Shull (1908). East and Jones proposed alternative theories of the occurrence of inbreeding depression. East (1936) favored the view that the individual loci connected with yield in corn were overdominant. Jones (1917) proposed an alternative hypothesis that the individual loci tended to have the allele with the higher homozygote be dominant. There is no way of distinguishing these hypotheses simply from examination of the mean: both predict inbreeding depression, and values of the a_{ij} can be chosen under either hypothesis to predict any degree of inbreeding. The controversy over these two views continued for many decades, taking a biochemical form in the arguments of Muller (1950) and Fincham (1972). It is a curious fact that East and Jones were colleagues and close collaborators at Harvard for many years. Their work and Shull's laid the basis for the spectacular success of hybrid corn in the American midwest in the 1930's. I suspect that the fascination with genetic effects of inbreeding in the early 1900s owed something to the fact that Mendelian genetics had explanations for it, while in pre-Mendelian theories it had no explanation.

The qualification which must be made to this picture is that we assumed that there is no change of gene frequency during inbreeding, which is to say that there is no selection.

Of course formulas (V-2) do not assume that gene frequencies remain at p in any one inbred line: they only assume that the gene frequency is on average p over all inbred lines, if there are a great many of them. This amounts to a no-selection assumption. But we have invoked the correlation of the trait with fitness to obtain an expected direction for inbreeding changes! So there is quite likely to be natural selection on the trait itself during inbreeding, resulting in a higher mean phenotype than predicted here, though probably still depressed somewhat. In addition we have seen in the previous chapter that selection at nearby loci (associative overdominance) will tend to retard genetic drift in finite populations. This phenomenon will also occur in most lines during the course of inbreeding, and tend to have a further effect in slowing the course of inbreeding depression.

MEANS OF CROSSES AND BACKCROSSES. The second derivation we shall do concerning means involves the related question of crosses between pure lines. As in the case of inbreeding, it looks at the mean contribution to a character from a single locus. The relationships that are found turn out to apply for any number of loci, as long as the genetic contribution to the character is determined by the sum of effects of the individual loci.

Suppose that we have two completely inbred lines. If one is fixed for A and the other for a, the mean contributions of this locus are $P_1 = a_{11}$ and $P_2 = a_{22}$ respectively. If we now cross these lines to get an F1 strain, the mean contribution in that strain is $F_1 = a_{12}$, since it is necessarily composed entirely of heterozygotes. Therefore there is no prediction that we can make of the F1 mean from the mean of the two parental populations, since knowing a_{11} and a_{22} does not allow us to predict a_{12} as long as there are no general rules concerning dominance. One might think that there would be no further generalizations which could be made. But consider the F2 and the two possible backcrosses. In the former case the genotypic composition is 1/4 AA, 1/2 Aa, and 1/4 aa, so that

$$\begin{aligned} F_2 &= \tfrac{1}{4}a_{11} + \tfrac{1}{2}a_{12} + \tfrac{1}{4}a_{22} \\ &= \tfrac{1}{2}\left(\tfrac{1}{2}a_{11} + \tfrac{1}{2}a_{22}\right) + \tfrac{1}{2}a_{12} \\ &= \tfrac{1}{2}\left(\tfrac{1}{2}P_1 + \tfrac{1}{2}P_2\right) + \tfrac{1}{2}F_1. \end{aligned} \qquad \text{(IX-9)}$$

This linear relationship holds for the contributions at each locus separately, and therefore will also hold for the overall phenotypic means as well. Now we have a prediction of the F2 phenotypic mean from those of the P1, P2, and F1 strains. In effect, what it tells us is that if we consider the F1 mean as well as the "midparent" (the average of the two parental strains), that although the F1 may differ from the midparent, the F2 will have moved halfway back toward the midparent. This is commonly found in hybrid corn: where the F1 is far superior to the original parental lines, the F2 falls far back down

towards the parent lines' performance, primarily as a result of the formation of inferior homozygotes.

Noting that the F2 is in Hardy-Weinberg proportions, the F3 will be the same as the F2, provided that it is formed by cross-fertilization of F2 individuals. If it is formed by selfing, the decline toward the midparental value will continue. The backcrosses BC1 = F1 $\times$ P1 and BC2 = F1 $\times$ P2 follow similar rules, e.g.

$$BC_1 \;=\; \frac{1}{2}a_{11} + \frac{1}{2}a_{12} \;=\; \frac{1}{2}P_1 + \frac{1}{2}F_1. \tag{IX-10}$$

so that each backcross should have a mean equal to the average of its parents' means, unlike the F1 and the F2.

Our entire derivation has assumed that A is fixed in one parent and a in the other, but the resulting rules are far more general. If the same allele is fixed in both populations, the same rules are easily seen to apply (since for that locus the contributions to P_1, P_2, F_1, etc. are all equal). But the results also generalize to the case where both P1 and P2 are not fixed, but are segregating an Hardy-Weinberg proportions. If the frequency of A in P1 is p_1 and if it is p_2 in P2, then

$$\begin{aligned} P_1 &= p_1^2\, a_{11} + 2p_1(1-p_1)\, a_{12} + (1-p_1)^2\, a_{22} \\ P_2 &= p_2^2\, a_{11} + 2p_2(1-p_2)\, a_{12} + (1-p_2)^2\, a_{22} \\ F_1 &= p_1 p_2\, a_{11} + [p_1(1-p_2) + p_2(1-p_1)]\, a_{12} + (1-p_1)(1-p_2)\, a_{22} \end{aligned} \tag{IX-11}$$

and since the F2 will be in Hardy-Weinberg proportions with gene frequency $p_3 = (p_1 + p_2)/2$ the F2 mean will be

$$F_2 \;=\; \left[\frac{1}{2}(p_1+p_2)\right]^2 a_{11} + 2\left[\frac{1}{2}(p_1+p_2)\right]\left[1 - \frac{1}{2}(p_1+p_2)\right] a_{12} + \left[1 - \frac{1}{2}(p_1+p_2)\right]^2 a_{22} \tag{IX-12}$$

Some algebra will then convince the reader that

$$F_2 \;=\; \frac{1}{2}\left(\frac{1}{2}P_1 + \frac{1}{2}P_2\right) + \frac{1}{2}F_1. \tag{IX-13}$$

Similarly the backcross and F3 relationships can be established in this case as well.

We expect these relationships among crosses of lines to hold very generally, as long as the initial populations are in Hardy-Weinberg proportions (although at different gene frequencies), and as long as the trait is determined additively by the genes. In fact, the fit of the F2 and backcross means to these predictions may be used to check on which scale the genes are most nearly additive, for that will be the scale on which the fit is best (all else being equal, as it never is).

That there is a close relationship between these results on the crosses of strains and the previous results on inbreeding depression may be seen by considering self-fertilization starting with F2 individuals, who constitute a population in Hardy-Weinberg proportions. We can predict the means of the self-fertilized F3, F4, etc. either from the above approach or by computing an f. It will be found that the self-fertilized population has a mean which moves halfway towards its ultimate limit each generation, a result wholly compatible with (IX-7).

The rules concerning means appear to have been established by Serebrovsky (1936a, 1936b) The linearity of the inbreeding effects when expressed as a function of f is probably due to Sewall Wright. We have already commented on the history of the controversy over the causes of inbreeding depression. When two pure (i.e., inbred) lines are crossed, it is frequently found that the hybrid is superior to either. This fact, well-known in the 19th century, is in effect an observation of inbreeding depression in reverse, and the same two classes of hypothesis (overdominance and dominance) have long been applied to explain it. A nomenclatural point is in order here. Contrary to much contemporary usage, *heterosis*, a term defined originally by Shull (1914) is the phenomenon of hybrid vigor, irrespective of whether it is caused by dominance or overdominance. *Overdominance* (Hull, 1945) is the superiority of the heterozygote at a single locus, and is only one possible explanation for heterosis.

IX.4 Additive and Dominance Variance

VARIANCES AND COVARIANCES. The additivity of our model of the phenotype has enabled us to analyze means one locus at a time. A similar simplification is possible with regard to variances, but requires one more assumption. Means are additive over loci whether or not the effects of the loci are independent of each other, but the variance is only additive over loci if the individual loci effects are uncorrelated. This will hold if the loci are all jointly at linkage equilibrium with respect to each other, for then knowing the contribution which one locus makes to the phenotype tells us nothing about the genotype (hence about the contribution) at any other locus. Then the locus effects are independent, and hence must also be uncorrelated. So under the assumption of linkage equilibrium we can write

$$\text{Var}\,(P) \;=\; \text{Var}\,(g_1) + \text{Var}\,(g_2) + \cdots + \text{Var}\,(g_n) + \text{Var}\,(e). \qquad \text{(IX-14)}$$

With respect to the environmental contribution we have implicitly assumed that it is uncorrelated with any of the genetic effects. Thus we have now made use of Assumptions No. 2 and No. 3.

A similar additivity holds with regard to the covariances between relatives, but before we comment further on that it may be useful to remind the reader of the meaning of

covariance, and its connection to correlation. Recall that the variance is defined as the expectation of the squared deviation from the mean:

$$\text{Var}\,(X) \;=\; \mathbb{E}[(X-\mu_X)^2] \tag{IX-15}$$

μ_X being the expectation of X. The covariance corresponds to this in a sense: it is the expectation of the product of deviations of two variables, each from its own mean:

$$\text{Cov}\,(X,Y) \;=\; \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]. \tag{IX-16}$$

If X and Y are positively related, this means that when X is above its mean Y will also tend to be above its mean, and when X is below its mean Y will also tend to be below its mean. So in this case the product $(X-\mu_X)(Y-\mu_Y)$ will usually be a product of two positive or of two negative quantities, so that its expectation will tend to be positive. This can be seen from examination of Figure 9.6 (below), which shows a scattergram plot of a large sample from a distribution (in fact a bivariate normal distribution) in which X and Y have a positive covariance. In this case the axes are the means, so that we can easily see whether X or Y exceed their means. There are far fewer points in the upper left and lower right quadrants than in the other two, so that only will the product $(X-\mu_X)(Y-\mu_Y)$ be negative.

We will make use of several properties of the covariance:

1. The covariance of X with itself is its variance. To see this, set $Y = X$ in (IX-16) and compare the result to (IX-14).

2. If Y is completely independent of X, then their covariance is zero. A detailed proof of this will be found in the better statistics texts, but we can make this intuitively plausible by pointing out that for each possible value of X, that is, for each possible value of $X-\mu_X$, all possible values of $Y-\mu_Y$ are possible, their relative probabilities being unchanged by our knowledge of X. So, the average contribution to (IX-16) from this category of outcomes will be $(X-\mu_X)$ times the expectation of $(Y-\mu_Y)$. But the latter must be zero, since it is $\mathbb{E}(Y)-\mathbb{E}(\mu_Y) = \mathbb{E}(Y)-\mu_Y = \mu_Y-\mu_Y$. So each possible value of X makes on average a zero contribution to (IX-15), and it follows from this that the covariance is zero.

3. If $X = a+b$ and $Y = c+d$, then $\text{Cov}\,(X,Y) = \text{Cov}\,(a,c)+\text{Cov}\,(a,d)+\text{Cov}\,(b,c)+\text{Cov}\,(b,d)$. In short, the general result is that the covariance of two sums is the sum of all possible covariances between a term from one sum and a term from the other. This can be shown using (IX-15), but we will not do so here. A particular case which we will use often is when a is independent of d and b is independent of c. In other words, the case where only quantities having the same position in the sum are not independent. In this case, we of course have

$$\text{Cov}\,(X,Y) \;=\; \text{Cov}\,(a,c)+\text{Cov}\,(b,d). \tag{IX-17}$$

The context in which this will arise is when the sum represents the sum of effects from different loci, where the loci are in linkage equilibrium, and where X and Y are the phenotypes in two relatives. Then a and b will be independent, so that a and d, which are effects at different loci in different individuals, are also going to be independent. Likewise since c and d are independent, so also will be c and b. The only nonzero terms will be the covariances between effects at the same locus in different individuals.

Finally, we should recall the definition of the correlation coefficient between two variables, X and Y. It is the ratio of their covariance to the product of their standard deviations:

$$\rho_{XY} = \mathbb{E}\left[\frac{(X-\mu_X)}{\sigma_X}\frac{(Y-\mu_Y)}{\sigma_Y}\right] = \text{Cov}\left(\frac{(X-\mu_X)}{\sigma_X}, \frac{(Y-\mu_Y)}{\sigma_Y}\right) \tag{IX-18}$$

We can therefore think of the correlation coefficient as being the covariance of the two variables after they have been standardized, that is, after their means have been scaled to zero (by subtraction of μ) and their standard deviations to 1 (by division by σ). It is an interesting fact that the correlation coefficient can never be greater than 1 or less than -1. This can be shown using the Cauchy-Schwartz Inequality. It is relatively easy to see that the correlation of a variable X with itself will be 1, since it will be the covariance of the standardized variable with itself, which is in turn the variance of the standardized variable, which is one.

PHENOTYPIC VARIANCE. One can write a straightforward expression for the variance of a phenotype P as a function of the individual genotype effects. Since the loci and the environmental effect are all assumed to be independent, the variance will be a sum of individual locus variances plus the variance of the environmental effect. Looking at the effect of a single locus with two alleles,, with the contributions of AA, Aa, and aa to the phenotype being a_{11}, a_{12}, and a_{22}, we find that its variance is straightforwardly

$$\begin{aligned} \text{Var}\,(g) &= \mathbb{E}(g^2) - [\mathbb{E}(g)]^2 \\ &= p^2a_{11}^2 + 2p(1-p)a_{12}^2 + (1-p)^2a_{22}^2 \\ &\quad - [p^2a_{11} + 2p(1-p)a_{12} + (1-p)^2a_{22}]^2. \end{aligned} \tag{IX-19}$$

This is a straightforward but somewhat dull formula which seems to offer few insights. One could go forward in a similar plodding fashion to compute complicated formulas for covariances among relatives. This would not only be very difficult, but would not necessarily be useful. After all, the resulting formula for the covariances between the overall phenotypes are going to depend on all of the genotype effects a_{jk} plus all of

the gene frequencies, plus an unknown environmental variance. This is a large number of quantities. If the covariances among relatives are complicated functions of many unknown parameters, there is little hope that we could predict anything about them, or use any of them to predict other quantities such as the response to artificial selection. Since we would not know the underlying parameters, the formulas involving covariances, variances, and response to selection would be at best of academic interest.

VARIANCE COMPONENTS. It turns out that the situation is not so gloomy as this. All of these observable quantities (variances, covariances, and selection responses) will depend on the unknown parameters, but *only through three intermediate quantities*. These are written V_A, V_D, and V_E, and are known as the additive, dominance, and environmental variances. They are *variance components*, because the variance of the phenotype is their sum $V_A + V_D + V_E$.

The various variances and covariances will depend on the multitude of unknown parameters only insofar as these affect V_A, V_D, and V_E. The same is (approximately) true of the selection response, which we discuss later. The implications of these facts are striking. Suppose we had (say) 6 covariances between different kinds of relatives, say between full sibs, half-sibs, aunt-nephew, mother-offspring, grandparent-offspring, and first cousins. If all of these depend only on V_A, V_D, and V_E, we should be able to use three of the covariances to estimate these three quantities, then use those in turn to predict the other covariances. Which is to say that there are relationships among various covariances, since all depend on only three quantities under our admittedly idealized model. We therefore have some hope of using some covariances to predict others, or even to predict the response of a character to artificial selection.

This has been the great strength of quantitative genetic theory in animal and plant breeding: the ability to estimate these three variance components from covariances among relatives, then use them to predict the response to different selection schemes with at least modest reliability. The flip side of this happy picture is that once we have estimated the three variance components, further observations of different covariances will only refine those estimates and will not lead us to knowledge of the individual genotype effects or gene frequencies. For any given set of values of V_A, V_D, and V_E there are vast numbers of combinations of gene frequencies and genotype contributions that will yield these same three values. So we can make little progress in working out the genetics of the trait by observing variances and covariances. The very robustness of the predictions of quantitative genetic theory means that the quantities being predicted will provide no insight into the underlying causes of the variation. Perhaps this is the version of the Uncertainty Principle appropriate for quantitative genetics.

ADDITIVE EFFECTS. Having revealed our goal, we must now show that these three quantities can in fact be obtained. We will start by considering the three genotype effects a_{11}, a_{12}, and a_{22} at a two-allele locus. The first step will be to express each of them as a

sum of three parts. The quantity a_{ij} will be replaced by $\mu + \alpha_i + \alpha_j + \delta_{ij}$. Our objective will be to find values of μ, α_1, α_2, and of δ_{11}, δ_{12} and δ_{22} such that three conditions hold:

1. The α_i account for as much as possible of the variance in the quantity a_{ij} in the particular population we consider,

2. The contributions of the four terms to the quantity a_{ij} have zero covariance, so that they are uncorrelated (that is, if we could somehow pick random individuals and look at this locus, and record for each individual the four quantities μ, α_i, α_j, and δ_{ij}, we will find zero covariance between these quantities over the whole population, and

3. Each of the last three terms α_i, α_j, and δ_{ij} has mean zero.

The third condition immediately allows us to determine the value of μ. It is the same for all three genotypes at the locus, and must therefore be the population mean of the a_{ij}:

$$\mu = p^2 a_{11} + 2p(1-p)a_{12} + (1-p)^2 a_{22}. \tag{IX-20}$$

The more difficult task is the determination of α_1 and α_2. Once they are determined, we can get the last term δ_{ij} by subtraction by the simple requirements that the four terms add up to a_{ij} in each genotype. Thus the bulk of our derivation goes into getting α_1 and α_2. We will do this by a rather indirect regression technique. The reader who gets a bit overwhelmed by all the covariances in this derivation may wish to skip to the next subsection, although a careful study of this derivation will pay off in terms of an understanding of what the variance components V_A, V_D and V_E do and do not mean.

The following three equations obviously hold, by the definitions of our quantities:

$$\begin{aligned} a_{11} &= \mu + \alpha_1 + \alpha_1 + \delta_{11} \\ a_{12} &= \mu + \alpha_1 + \alpha_2 + \delta_{12} \\ a_{22} &= \mu + \alpha_2 + \alpha_2 + \delta_{22}. \end{aligned} \tag{IX-21}$$

Now note that these equations can be rewritten in the shorthand form

$$a_{ij} = \mu + 2\alpha_2 + x(\alpha_1 - \alpha_2) + \delta_{ij}, \tag{IX-22}$$

provided that x is a quantity which is 2 when the genotype is A_1A_1, is 1 when it is A_1A_2, and is 0 when the genotype is A_2A_2. In short, x simply tells us how many A_1 alleles there are in the genotype at this locus. Equation (IX-22) may be seen as telling us what is the dependence of genotype contribution a on the allele count x, the α_i as determining the coefficients of this regression equation, and the δ_{ij} as the deviations of the a_{ij} from the regression prediction $\mu + 2\,\alpha_2 + x(\alpha_1 - \alpha_2)$. What we are going to do is to determine the

α_{ij} by carrying out a least squares fit of the a_{ij} to this regression line. There is such a regression. Imagine sampling individuals from our population and recording for each the x and the quantity a_{ij}. We would find that when our "sample" was the whole population,

$$\begin{array}{rll} p^2 & \text{of the time they were} & (2, a_{11}), \\ 2p(1-p) & \text{of the time they were} & (1, a_{12}), \\ \text{and} \quad (1-p)^2 & \text{of the time they were} & (0, a_{22}). \end{array}$$

So we do indeed have two random variables x and a.

Relation to regression. It may seem rather arbitrary to choose the least squares criterion of fit. After all, we are trying to satisfy the three criteria stated above. What do these have to do with a least-squares regression? In fact, everything. For knowing the prediction terms $\mu + 2\alpha_2 + x(\alpha_1 - \alpha_2)$ are chosen by a least squares fit to the a_{ij} guarantees us that all three conditions will be met if the α_i are determined in this way. For fitting regression lines by least squares is well-known to guarantee the following properties:

1. The regression prediction accounts for as much of the variance in the dependent variable as possible,
2. The deviation of the points from the regression line (the residual) is uncorrelated with the regression prediction, and
3. The regression line passes through the point $(\bar{x}, \bar{y})$ where both variables have their mean values, and the mean of the deviations from the regression line is zero.

The first condition is clearly the same as the first of our previous three requirements. The second guarantees us that the δ_{ij} will be uncorrelated (that is, have zero covariance) with the quantity $\mu + \alpha_i + \alpha_j$ which is our regression prediction. We will still need to establish that the two α's in an individual are uncorrelated but this will be easy. The third condition is actually our previous third requirement in disguise. For the expectation (the population mean) of x is simply

$$\begin{aligned} \mathbb{E}(x) &= 2 \times p^2 + 1 \times 2p(1-p) + 0 \times (1-p)^2 \\ &= 2p. \end{aligned} \tag{IX-23}$$

The regression line at $x = 2p$ will have height

$$\mu + 2\alpha_2 + 2p(\alpha_1 - \alpha_2) \;=\; \mu + 2p\alpha_1 + 2(1-p)\alpha_2. \tag{IX-24}$$

Since this must be equal to the population mean of the a_{ij}, which we already know to be equal to the constant μ, we must have, as a result of determining the α_i by the least squares regression, that (dropping the 2)

$$p\,\alpha_1 + (1-p)\alpha_2 \;=\; 0. \tag{IX-25}$$

But consider the first term α_i in the sum $\alpha_i + \alpha_j$. It is dependent only on the identity of the maternally-derived allele A_i. A fraction p of the time that allele is A_1 and $1-p$ of the time it is A_2. So (IX-23) tells us that the mean value of α_i, which will be $p\alpha_1 + (1-p)\alpha_2$ is zero. By an exactly similar argument the mean value of α_j will also be zero. And since the deviation of a_{ij} from the regression prediction is δ_{ij}, this quantity too has expectation zero over the whole population. So we have now nearly satisfied our original three conditions by choosing the α's by least squares. All that remains is to show that the two terms α_i and α_j are not correlated. This is easily seen to be true, as a consequence of Hardy-Weinberg proportions. The size of α_i (that is, whether it is α_1 or α_2) depends only on the allele identity of A_i (which is either A_1 or A_2). Similarly α_j depends only on the identity of A_j. But by the Hardy-Weinberg law these are independent, since the population results from random mating. So α_i and α_j are independent and hence uncorrelated.

Now that we have established that the least squares procedure will fulfill our requirements, all that remains is to carry it out. Figure 9.3 shows the regression in diagrammatic form. It is well known in statistics that the least squares solution to the slope of the regression line is given by

$$b \;=\; \mathrm{Cov}\,(y,x)/\mathrm{Var}\,(x) \tag{IX-26}$$

where y is the dependent variable. Thus to determine the slope, which will be $\alpha_1 - \alpha_2$ by (IX-22), we must evaluate the covariance of a and x and the variance of x. Our "sample" is the whole population, so that we compute these quantities as expectations over the population, the various outcomes being in their expected proportions. The gene dosage x is the number of copies of the A allele. It has values 2, 1, and 0. The variance of x is easily determined:

$$\begin{aligned}
\mathrm{Var}\,(x) &= \mathbb{E}(x^2) - [\mathbb{E}(x)]^2 \\
&= p^2 \times 4 + 2p(1-p) \times 1 - (2p)^2 \\
&= 2p(1-p).
\end{aligned} \tag{IX-27}$$

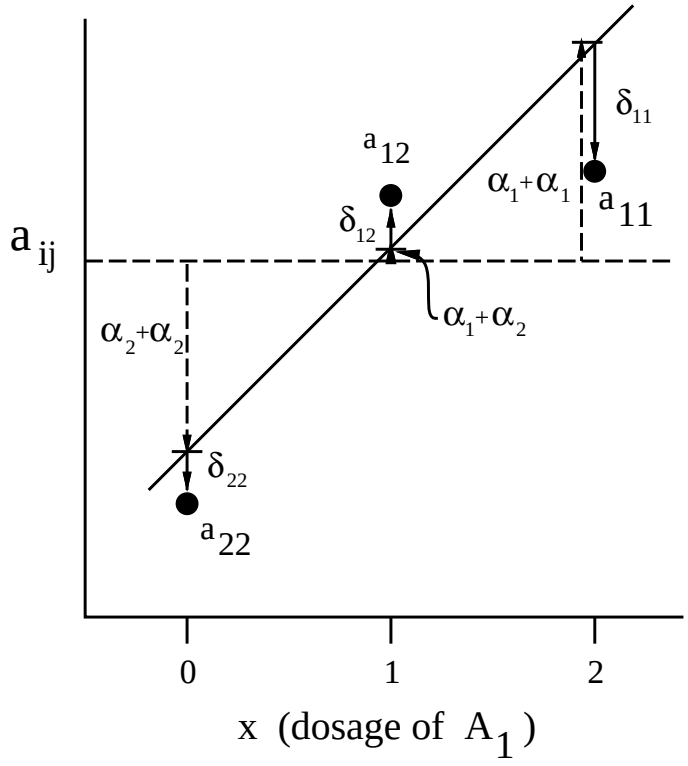


Figure 9.3: The regression of a on x. The line is determined by a least squares fit, weighting each point by its population frequency. The dotted arrows show the fitted additive values $\mu + \alpha_2 + \alpha_2$, $\mu + \alpha_1 + \alpha_2$, and $\mu + \alpha_1 + \alpha_1$. The solid arrows show the residuals.

The covariance of a and x requires a bit more computation:

$$\begin{aligned} \text{Cov}\,(a,x) &= \mathbb{E}(ax) - \mathbb{E}(a)\,\mathbb{E}(x) \\ &= 2 \times a_{11}(p^2) \;+\; 1 \times a_{12}[2p(1-p)] \;+\; 0 \times a_{22}(1-p)^2 \;-\; \mu(2p) \qquad \text{(IX-28)} \\ &= 2p^2 a_{11} + 2p(1-p)a_{12} - 2p\mu. \end{aligned}$$

Now since we are requiring that $\alpha_1 - \alpha_2 = \text{Cov}\,(a,x)/\text{Var}\,(x)$ it follows that $(\alpha_1 - \alpha_2)\text{Var}\,(x) = \text{Cov}\,(a,x)$ so that our equation for the slope of a least squares fit is

$$2p(1-p)(\alpha_1 - \alpha_2) \;=\; 2p^2 a_{11} \;+\; 2p(1-p)a_{12} \;-\; 2p\mu, \qquad \text{(IX-29)}$$

or

$$(1-p)\alpha_1 - (1-p)\alpha_2 \;=\; pa_{11} + (1-p)a_{12} - \mu. \qquad \text{(IX-30)}$$

The slope is one of the two results of doing a least squares regression. The height of the line is the other, and it can be found from the requirement that the line pass through the point $(\bar{x}, \bar{y})$. We have already seen that this gives us equation (IX-25). Adding that equation to (IX-30) we get

$$\alpha_1 \;=\; pa_{11} + (1-p)a_{12} - \mu. \tag{IX-31}$$

and using (IX-28) together with (IX-19) we find that

$$\alpha_2 \;=\; pa_{12} + (1-p)a_{22} - \mu. \tag{IX-32}$$

Now we have found the slope and height of our regression of a on x, and determined α_1 and α_2 from them. The equations for the δ_{ij} follow by substituting (IX-20), (IX-31) and (IX-32) into (IX-22). It turns out that

$$\delta_{11} \;=\; (1-p)^2\,(a_{11} - 2a_{12} + a_{22}). \tag{IX-33}$$

The expressions for δ_{12} and δ_{22} are the same, but with the $(1-p)^2$ replaced, respectively, by $-p(1-p)$ and p^2.

One interpretation of (IX-31) and (IX-32) is worth noting. Since p of the A_1 alleles occur in A_1A_1 homozygotes, and the rest in A_1A_2 heterozygotes, (IX-31) represents the difference between the average contribution at the A locus in an A_1-bearing individual, and the population mean. So it can be thought of as the *average excess* of A_1-bearing individuals over the population mean. Similarly (IX-31) shows that α_2 is the average excess of those individuals in which a randomly-chosen A_2 allele is to be found. This is strongly reminiscent of our results for natural selection in chapter II, where the mean relative fitness $\bar{w}_A$ entered in. In fact, equation (II-35) showed that

$$\Delta p \;=\; p\,\bar{w}_A/\bar{w} \;-\; p \;\;=\;\; p\,(\bar{w}_A \;-\; \bar{w})/\bar{w} \tag{IX-34}$$

which shows the quantity $\bar{w}_A - \bar{w}$ playing an important role. It is precisely the average excess of allele A if the phenotype is the relative fitness w.

Now we have found the quantities μ, α_i, α_j, and δ_{ij} for each genotype. But we never made clear why breaking the genotype contribution a_{ij} into these four uncorrelated parts was worth doing. It now remains for us to show that these can be used to provide definitions of the quantities V_A, V_D, and V_E and insight into their properties.

We have done the above derivation for a two-allele case, but the results for multiple alleles are entirely analogous. The least squares fit is obtained by minimizing the weighted mean square of the residual:

$$Q \;=\; \sum_i \sum_j p_i p_j (a_{ij} - \mu - \alpha_i - \alpha_j)^2 \tag{IX-35}$$

and resulting formulas for the α_i,

$$\alpha_i \;=\; \sum_j p_j\, a_{ij} - \mu \tag{IX-36}$$

have exactly the same interpretation as average excesses.

ADDITIVE AND DOMINANCE VARIANCES. We now know that each individual's phenotype could, if we knew the precise genotype, be partitioned into a series of additive components. Suppose that we write the genotype as $A_i A_j B_k B_l C_m C_n$ and the corresponding breakdown of the phenotype as

$$P \;=\; \mu + \alpha_i + \alpha_j + \delta_{ij} + \mu' + \alpha'_k + \alpha'_l + \delta_{kl} + \mu'' + \alpha''_m + \alpha''_n + \delta_{mn} + e \tag{IX-37}$$

where the primes distinguished between the different loci, and e is the environmental contribution to the phenotype. We can now rearrange this into four groups of terms:

$$P \;=\; (\mu + \mu' + \mu'') \;+\; (\alpha_i + \alpha_j + \alpha'_k + \alpha'_l + \alpha''_m + \alpha''_n) \;+\; (\delta_{ij} + \delta_{kl} + \delta_{mn}) \;+\; e \tag{IX-38}$$

The first group of terms are constants which do not depend on the genotype. These will not contribute to the variance of the phenotype. This first group is the estimate of the phenotype which we would make if we did not know the genotype: it is simply the population mean. The second group refines this estimate by incorporating the average effects. These first two groups of terms, taken together, make an estimate of the phenotype based only on genes taken one at a time. This estimate is called the *breeding value*. The variance of the breeding value will be entirely due to the α's, of course. This variance is called the *additive variance*, and written V_A. The next group of terms adds to the estimate the predicted interactions between the two gene copies at each locus. Each of the individual terms δ_{ij} is called a *dominance deviation*. There is no conventional term for the sum of the δ's. The sum of the dominance deviations has a variance called the *dominance variance*, V_D. The environmental effect e has a variance known as the *environmental variance*, V_E.

Notice that we have set up all the terms in (IX-35) to be uncorrelated with each other. This means that the four groups of terms must also be uncorrelated, so that in effect phenotype has been divided into for parts which are uncorrelated:

$$P \;=\; \mu + A + D + E \tag{IX-39}$$

which means that the variance of this sum is the sum of their variances:

$$\mathrm{Var}\,(P) \;=\; V_A + V_D + V_E \tag{IX-40}$$

since the mean μ does not vary from individual to individual. We have now succeeded in showing that the variance in phenotype can be divided into three parts. If by some

feat of genomics the genotype of each individual became known, we could compute all the α_i and δ_{ij} and obtain these three variance components straightforwardly. In practice this is unlikely, but as we shall see in the next section, the covariances (and hence the correlations) between relatives can be computed from these three variance components. Hence we can reverse the process and estimate the variance components from the covariances of relatives.

The dominance variance. It is nevertheless instructive to find the formulas computing the variance components V_A and V_D from the genotype effects a_{ij}, for then we can get a picture of how various patterns of dominance affect these components. The additive variance V_A is the variance of a sum of terms, two for each locus and all uncorrelated. The dominance variance V_D is also the variance of a sum of uncorrelated terms, one per locus. Thus each of these variances is itself a sum of individual locus additive (or dominance) variances. Let us compute from the a_{ij} and the gene frequency the additive and dominance variances at a single locus for the two-allele case.

Each of the terms α_i has is an α_1 with probability p and an α_2 with probability $1-p$. The variance of one term is thus

$$\text{Var}\,(\alpha_i) \;=\; p\,\alpha_1^2 + (1-p)\,\alpha_2^2 \tag{IX-41}$$

since the mean of α_i is known to be zero by (IX-25), which was in turn a byproduct of our way of defining the α's. We can now substitute from equations (IX-31) and double the result (since there are two α's for this locus, each with variance given by (IX-41)). After some algebra we find that

$$\text{Var}\,(\alpha_i + \alpha_j) \;=\; 2p(1-p)\,[(pa_{11} + (1-p)a_{12}) \;-\; (pa_{12} + (1-p)a_{22})]^2, \tag{IX-42}$$

after using (IX-20) to eliminate μ.

The dominance deviations also have a mean of zero, so that

$$\begin{aligned}\text{Var}\,(\delta_{ij}) \;&=\; p^2\,\delta_{11}^2 + 2p(1-p)\,\delta_{12}^2 + (1-p)^2\,\delta_{22}^2 \\ &=\; p^2(a_{11} - \mu - \alpha_1 - \alpha_1)^2 \;+\; 2p(1-p)(a_{12} - \mu - \alpha_1 - \alpha_2)^2 \\ &\quad +(1-p)^2(a_{22} - \mu - \alpha_2 - \alpha_2)^2\end{aligned} \tag{IX-43}$$

which after a similar, but larger amount of algebra turns out to be

$$\text{Var}\,(\delta_{ij}) \;=\; p^2(1-p)^2(a_{11} - 2a_{12} + a_{22})^2. \tag{IX-44}$$

This last formula has an interesting property. Suppose that the genotypic effect at this locus were actually additive, that is, that the heterozygote effect a_{12} was the arithmetic mean of the two homozygote effects. Then $a_{12} = (a_{11} + a_{22})/2$ and it requires only

Table 9.1: Fraction of all genetic variance at a locus due to dominance variance, where p is the frequency of a completely dominant allele.

p	fraction
0.1	0.053
0.2	0.111
0.3	0.176
0.4	0.25
0.5	0.333
0.6	0.429
0.7	0.538
0.8	0.667
0.9	0.818

a few simple steps to see that (IX-44) is zero. So when there is additive gene action within a locus there is no dominance variance. Since the genotypic contribution to the total phenotypic variance is $V_A + V_D$, we find that in this case all of the genotypic variance will be additive variance. So dominance variance disappears when there is no dominance, exactly as its name implies. Now suppose that instead the locus showed complete dominance of the A_1 allele. Then in formula (IX-42) we have $a_{11} = a_{12}$, so that it reduces to

$$\text{Var}\,(\alpha_i + \alpha_j) \;=\; 2\,p(1-p)^3\,(a_{12} - a_{22})^2. \qquad \text{(IX-45)}$$

A similar calculation is easily done with the dominance variance. Now note that (IX-44) does *not* disappear when there is complete dominance. In fact, if we compute the fraction of the genotypic variance which is due to dominance variance, we find the numbers given in Table 9.1. Interestingly, when the dominant allele is rare there may be mostly additive variance. This corresponds to the observation that we will be hard-pressed to tell whether a rare dominant allele is in fact dominant, since it appears so rarely in homozygotes. So in this respect as in others the behavior of the rare allele in heterozygotes is all that we need know.

Note also that the average dominance over loci is not meaningful here. If some loci show complete dominance, others complete recessivity, such that in some average sense there is no dominance, there will still be dominance variance in existence. For the dominance variance V_D is the sum of terms from individual loci, and for V_D to be zero (IX-42) shows us that there can be no dominance *at any locus*.

We thus have found from these formulae that the amount of dominance variance is a poor indicator of the type of gene action at the individual loci. There can be mostly additive variance even when all loci show complete dominance. It would be a brave quantitative geneticist indeed who would make strong statements about gene action

based on the relative sizes of V_A and V_D. While they may be useless for this purpose they will be of great help in analyzing covariances, to which we now turn.

IX.5 Covariances Between Relatives

Now we are ready to compute covariances between relatives. We imagine ourselves to be dealing with a series of pairs of individuals. In each pair the two individuals are relatives of a specific sort, such as an individual and its grandmother. If we measure a given phenotype on all individuals, we can compute the covariance between the phenotypes of relatives. Each phenotype is a sum of individual locus effects and an environmental effect, so that calling the phenotypes in the two relatives X and Y we have

$$\begin{aligned} X &= g_1 + g_2 + \cdots + g_n + e \\ Y &= g'_1 + g'_2 + \cdots + g'_n + e'. \end{aligned} \tag{IX-46}$$

Now we have already seen that the covariance of two sums is the sum of all possible pairwise covariances involving one term from one sum and one term from the other. Consider some of these terms. Recall that we are assuming that the environmental effect is independent of the genotype, so that terms like $\mathrm{Cov}\,(g_i, e')$ or $\mathrm{Cov}\,(g'_i, e)$ must be zero, since if e is independent of the genotypic effect g_i of a locus in the same individual, it must all the more certainly be independent of the genotypic effect g'_i of that locus in a relative. Likewise the environments were assumed independent in the two individuals, so that $\mathrm{Cov}\,(e, e')$ is also zero. This assumption is not very realistic in many cases, but we make it here for heuristic reasons.

This leaves us with terms involving one g and one g'. Now recall that we also assumed linkage equilibrium. This means that the effects of two different loci in the same individual, g_i and g_j, are independent. This implies that all terms of the form $\mathrm{Cov}\,(g_i, g'_j)$, which involve both different loci and different individuals, should be zero. If g_i and g_j are independent, then surely g_i and g'_j are too. We are left only with those terms which involve the same locus in the two individuals:

$$\mathrm{Cov}\,(X, Y) = \mathrm{Cov}\,(g_1, g'_1) + \mathrm{Cov}\,(g_2, g'_2) + \cdots + \mathrm{Cov}\,(g_n, g'_n) \tag{IX-47}$$

We are now in a position to compute the covariance locus by locus.

In any particular pair of individuals, if we know whether they are relatives, we can compute two coefficients:

- f_1, the probability that a random copy of a gene at a locus is identical by descent to one of the two copies at that locus in the other individual.
- f_2, the probability that both of the two copies at a locus are identical by descent, each to one of the two copies at that locus in the other individual.

If we knew that both of these coefficients were zero, then there would be no identity by descent between the two individuals. The contribution at a locus in one individual would be $\alpha_i + \alpha_j + \delta_i j$ and in the other individual $\alpha_k + \alpha_\ell + \delta_{k\ell}$. The indices i, j, k, and ℓ would be the result of independent random sampling of four genes from the population, so that there would be no identity by descent between the two individuals at this locus. Some of the i, j, k, and ℓ could be the same, by accident, but not owing to identity by descent.

On the other hand, if we knew that $f_2 = 1$, then the contribution from the locus in both individuals would be $\alpha_i + \alpha_j + \delta_{ij}$. In this case $\mathrm{Cov}\,(g_i, g_i')$ would involve the covariance of a quantity with itself. This is precisely the variance of the quantity, as we have already noted. But that in turn is the contribution to the variance $V_A + V_D$ at this locus. (Note an assumption that we have implicitly made: by assuming that each individual is the product of random mating, we have assumed that the two genes in the same individual are not identical by descent. That in turn means that if each of the two genes in one individual are identical by descent to a gene in the relative, they must be identical by descent to two different genes in that relative. Thus if the first individual is genotype A_iA_j, the second must be either A_iA_j or A_jA_i. We are not allowing both the A_i and the A_j to be identical by descent to the same gene.)

If $f_1 = 1$ but $f_2 = 0$, the genotypes must be A_iA_j and A_iA_k, or A_iA_j and A_kA_j, or some such. Then the genotype contributions at the locus are of the form $\alpha_i + \alpha_j + \delta_{ij}$ and $\alpha_i + \alpha_k + \delta_{ik}$. In the covariance these sums, the only terms which could be nonzero are $\mathrm{Cov}\,(\alpha_i, \alpha_i)$ and $\mathrm{Cov}\,(\delta_{ij}, \delta_{ik})$. The latter term is in fact zero, for it turns out that the dominance deviations of two genotypes are uncorrelated, even if they share one allele in common (the other being chosen at random). This can be proven using (IX-33). In fact terms of the form

$$\delta_{11}\ [\ p\delta_{11} + (1-p)\delta_{12}\ ] \tag{IX-48}$$

are always zero, since this equals

$$\begin{aligned} &\delta_{11}\ [p(a_{11} - \mu - 2\alpha_1)\ +\ (1-p)(a_{12} - \mu - \alpha_1 - \alpha_2)] \\ &=\ \delta_{11}\ [pa_{11} + (1-p)a_{12} - \mu - \alpha_1 - p\alpha_1 - (1-p)\alpha_2] \end{aligned} \tag{IX-49}$$

which is easily shown to be zero using (IX-25) and (IX-31). Continuing in this fashion one can show that the covariance of δ_{ij} with δ_{ik} is zero. So the covariance at the locus is simply $\mathrm{Cov}\,(\alpha_i, \alpha_i)\ =\ \mathrm{Var}\,(\alpha_i)$. But this is half the contribution of this locus to the additive variance V_A.

In all but the simplest cases some loci will have two, some one, and some no genes in the first individual which are identical by descent to genes in the second individual. We now make use of the following property of covariance: if variable x has probability p_1 of being the random quantity x_1, and probability p_2 of being the random quantity x_2, then its covariance with anything else is simply $\mathrm{Cov}\,(p_1x_1 + p_2x_2,\ y)$ which can be

shown to be equal to $p_1\text{Cov}(x_1, y) + (1 - p_1)\text{Cov}(x_2, y)$. In other words x has the same covariances as the weighted average of the covariance of x_1 with y and the covariance of x_2 with y, the weights being p_1 and $1 - p_1$.

Applying this to our situation where there are three possibilities, we can write the covariance as

$$\text{Cov}(g_i, g_i') \;=\; f_1\, V_A^{(i)} + f_2\, V_D^{(i)}. \tag{IX-50}$$

where a quantity like $V_A^{(i)}$ is the contribution of locus i to the additive variance. Now we need only use (IX-47) to add these over all loci to obtain

$$\text{Cov}(X, Y) \;=\; f_1\, V_A + f_2\, V_D \tag{IX-51}$$

This is the result we have been striving for. It tells us the covariances between relatives as a function of V_A, V_D, and two probabilities of identity by descent. Those in turn are easy to compute for simple relationships.

CORRELATIONS. The formula for the covariance of relatives can easily be used to find the correlation coefficient between the relatives. Recall that the correlation is simply the covariance divided by the product of the standard deviations. In our model, each relative is drawn from the population at random (although the two members of each pair are not drawn independently). For example, if we are examining grandparent-offspring pairs, it does not matter whether we choose the grandparents at random and then find a grandchild of each one, or whether we choose the grandchildren at random and then get a grandparent of each one: in either case if we were to consider the set of grandparents by itself, it would in effect be randomly sampled from the population. We are assuming that there are no evolutionary forces acting to change gene frequencies or create disequilibria, so that even though the grandparents and grandchildren are sampled from different generations, each group considered alone might just as well have been sampled from the same generation. Which leads us to the conclusion that the standard deviations of the trait in the two relatives will be the same, so that their product will simply be the variance of the trait. Thus the correlation between relatives will simply be the ratio of their covariance to the variance of the trait:

$$\rho_{XY} \;=\; \text{Cov}(X, Y) \big/ \sqrt{\text{Var}(X)\text{Var}(Y)} \;=\; \text{Cov}(X, Y) \big/ \text{Var}(X) \tag{IX-52}$$

PARENTS AND OFFSPRING: HERITABILITY. Now we can use these formulas to examine a specific relationship: that of parents and offspring. We have to compute the coefficients in (IX-47) first. The coefficient of kinship between parent and offspring is 0.25 so the coefficient of V_A is 1/2. Another way of seeing this result is to consider the probability that a gene drawn from the offspring came from that particular parent, which is clearly 1/2. The coefficient P_2 is the probability that both genes in the offspring are identical by descent to genes in the parent. This is impossible, since we are in effect

assuming that there are no other pedigree paths connecting these two individuals. So the second coefficient is zero. Then

$$\begin{aligned} \text{Cov}_{PO} &= \tfrac{1}{2}V_A \\ \rho_{PO} &= \tfrac{1}{2}\frac{V_A}{V_A+V_D+V_E} \end{aligned} \tag{IX-53}$$

The quantity $V_A/(V_A+V_D+V_E)$ will appear repeatedly in this chapter: it is simply the fraction of the variance in the trait which is additive variance. It is usually written h^2 and is called the *heritability*. The name may be misleading, for h^2 is not the fraction of all variation which is due to genetic causes. That would be $(V_A+V_D)/(V_A+V_D+V_E)$ and is sometimes called *heritability in the broad sense*.

FULL SIBS AND HALF-SIBS. If we have pairs of individuals which are half-sibs, sharing (say) a common mother but different fathers, then the covariance is easy to calculate. Once again, there is no chance that both genes in the half-sibs are identical by descent. The chance that a given randomly-chosen gene in one half-sib is identical by descent to a gene in the other is 1/4, for that is the probability that we have both chosen the maternally-derived gene, and that that particular gene in the mother was passed to the other half-sib. So

$$\begin{aligned} \text{Cov}_{HS} &= \tfrac{1}{4}V_A \\ \rho_{HS} &= \frac{1}{4}\,\frac{V_A}{V_A+V_D+V_E} \\ &= \tfrac{1}{4}h^2. \end{aligned} \tag{IX-54}$$

The covariance of full sibs is a bit more complicated. Choosing a gene from one sib, we know that it came from one of the parents. The chance that that particular parent also passed the same gene to the other sib is 1/2. So we have computed the first coefficient. The second (P_2) is the chance that *both* genes in one sib will be found in the other. A moment's consideration will show that the two events are independent (see Figure 9.4) so that the second coefficient must be 1/4. So we can write

$$\begin{aligned} \text{Cov}_{FS} &= \tfrac{1}{2}V_A + \tfrac{1}{4}V_D \\ \rho_{FS} &= \frac{1}{2}\frac{V_A}{V_A+V_D+V_E} + \frac{1}{4}\frac{V_D}{V_A+V_D+V_E} \end{aligned} \tag{IX-55}$$

Thus we see that the correlation between full sibs is greater than between parents and offspring. This is only true to the extent that there is dominance variance in the trait. Parents and offspring share half their genetic material. So do full sibs, but in addition, full sibs have the possibility of getting (with probability 1/4) precisely the same diploid

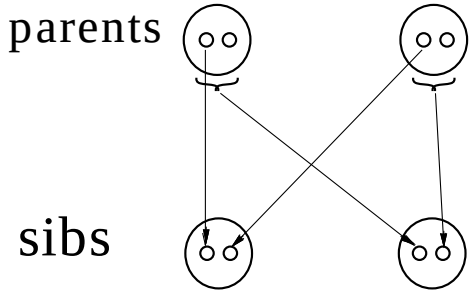


Figure 9.4: Diagram of the relationship of two full sibs.

Table 9.2: Covariances and correlations for different degrees of relationships.

	Covariance	*Correlation*
Parent-offspring	$\frac{1}{2}V_A$	$\frac{1}{2}h^2$
Half sibs	$\frac{1}{4}V_A$	$\frac{1}{4}h^2$
Full sibs	$\frac{1}{2}V_A + \frac{1}{4}V_D$	$\frac{1}{2}h^2 + \frac{1}{4}V_D/(V_A + V_D + V_E)$
Grandparent-offspring	$\frac{1}{4}V_A$	$\frac{1}{4}h^2$
Aunt/Uncle - Niece/Nephew	$\frac{1}{4}V_A$	$\frac{1}{4}h^2$
Dizygotic twins	$\frac{1}{2}V_A + \frac{1}{4}V_D$	$\frac{1}{2}h^2 + \frac{1}{4}V_D/(V_A + V_D + V_E)$
Monozygotic twins	$V_A + V_D$	$h^2 + V_D/(V_A + V_D + V_E)$
Full first cousins	$\frac{1}{8}V_A$	$\frac{1}{8}h^2$
Unrelated individuals	0	0

genotype at a locus. Thus if there is an effect on the phenotype due not only to the genes taken singly, but to the particular combination of two genes at a locus, this effect will be shared by both sibs 1/4 of the time.

OTHER RELATIONSHIPS. The same logic enables us to arrive at formulas for covariances and correlations for other degrees of relationship. Here is a table. It will be a useful exercise for the reader to see if they can reproduce the results in the table. These covariances only hold true under the rather limited assumptions of our model. We shall discuss the pitfalls involved when we discuss the estimation of heritability.

IX.6 Regression of Offspring on Parents

In our discussion of the effects of artificial selection on a phenotype, we will be particularly interested in the relationship of offspring and parents. If a parent is chosen which is above the population mean, what distribution of phenotypes do we expect among its offspring? One way of addressing concerns such as this is to examine the regression of offspring on parents. If we were to make a scattergram in which we chose individual parent-offspring pairs at random from the population, then plotted each as a point on an (X, Y) plane where the horizontal coordinate was the parent's phenotype, and the vertical coordinate the offspring's phenotype, we would see a cloud of points looking very much like Figure 9.5. Suppose that we tried to fit a straight line through the cloud of points by least squares, to predict the offspring phenotypes from the phenotype of the single parent whose phenotype we know. It is a well-known statistical fact that the slope of the least-squares regression is given by the ratio of covariance of the two variables to the variance of the variable that is on the horizontal axis:

$$\beta_{OP} = \frac{\text{Cov}_{PO}}{\text{Var}_O} = \frac{1}{2}\frac{V_A}{V_A + V_D + V_E} = \frac{1}{2}h^2 \qquad \text{(IX-56)}$$

Thus the slope should always be less than 1 (though it may not be if our assumptions are violated). It is more instructive to look at the regression after shifting our axes so that each runs through the corresponding population mean (of parents or of offspring). This shift should not affect the slope. Figure 9.6 shows this plot. Since the least-squares regression line always passes through the sample means, we expect it to pass through the origin in the shifted graph, as shown in the Figure.

The fact that the regression coefficient is less than $1/2$ tells us that we predict the offspring of an individual to be, on average, less than half as far from the population mean as it is. This would seem to be a paradox, for it seems to indicate that the offspring generation will be closer to the mean than the parent generation. If such a process were to continue indefinitely, there would soon be no variability left at all! Yet this cannot be so, since we know that this population is in Hardy-Weinberg proportions and linkage equilibrium, and its genotypic composition will not change over time. In fact, the same phenomenon is seen if we go backwards in time: computing the regression coefficient of parents on offspring, we find that

$$\beta_{PO} = \frac{\text{Cov}_{PO}}{\text{Var}_O} = \frac{1}{2}\frac{V_A}{V_A + V_D + V_E} = \frac{1}{2}h^2 \qquad \text{(IX-57)}$$

Thus looking back in time we also seem to see variance decreasing. This seems to say that the variance of the phenotype just happens to be at a maximum at the moment we look at it. This is such a ridiculous notion that we know that something must be wrong with our argument.

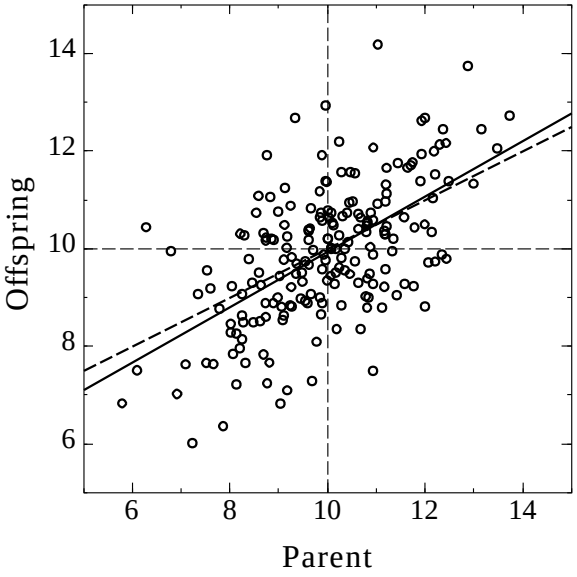


Figure 9.5: Scattergram of a simulated sample of parent-offspring pairs, and the regression line through these points. The expectations of both parents and offspring are 10, and the expected slope is 0.5, and this is shown by the dashed lines. The empirical regression line of the 200 points which passes through the empirical means at (9.969, 9.917) has slope 0.569, and is shown by the solid line.

The flaw lies in the fact that we are only predicting the average phenotype of the offspring of a given individual. Different offspring will have phenotypes which vary around their expectation. This is an additional source of variance in the next generation. Although the means from different parents will be more tightly clustered around the overall population mean than were the phenotypes of the parents, the individual values will vary more than their predictions, and the variance will be just as great in the next generation as it is at present. A small regression coefficient reflects an inability to predict where on the scale the offspring will be, not a prediction that it will be near the population mean. The easiest way to see this is to consider what happens if $h^2 = 0$. Then we cannot make any prediction of offspring's phenotype from parent's, but since (say) V_E may be substantial, we are surely wrong in using our measurement of h^2 to make a positive prediction that all offspring will lie at the population mean.

Interesting enough, it is precisely the case of parent-offspring regression which caused Francis Galton (1889) to coin the term "regression coefficient". Galton noticed that when the offspring of parents far from the mean were looked at, these offspring had (on average) "regressed" towards the mean. The regression coefficient was intended to measure

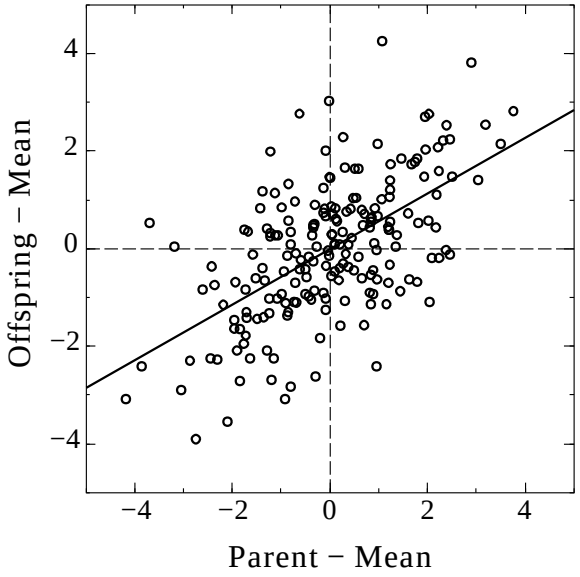


Figure 9.6: Scattergram of the same simulated sample of parent-offspring pairs, plotted on scales showing the departure from the empirical mean, and showing the least squares regression line through these points.

by how much.

REGRESSION ON THE MIDPARENT. It is a natural extension of the preceding discussion to ask what happens if we try to predict the offspring's phenotype from the phenotypes of both parents. Since the two parents contribute equally to the offspring (at least, in our model they do), the natural quantity to consider is the average phenotype in the two parents. This is called the *midparent*, and it seems reasonable to ask how well the offspring phenotype can be predicted from it. The midparent is

$$X_{MP} \;=\; \frac{1}{2}X_1 + \frac{1}{2}X_2, \tag{IX-58}$$

where X_1 and X_2 are the phenotypes of the two parents. The covariance will be

$$\mathrm{Cov}\,(X_{MP}, Y) \;=\; \mathrm{Cov}\left(\frac{1}{2}X_1 + \frac{1}{2}X_2, Y\right) \;=\; \frac{1}{2}\mathrm{Cov}\,(X_1, Y) + \frac{1}{2}\mathrm{Cov}\,(X_2, Y). \tag{IX-59}$$

Although we have not mentioned it previously, one can easily show from the definition of covariance that constants like the $1/2$ can be removed: $\mathrm{Cov}\,(cX, Y) \;=\; c\;\mathrm{Cov}\,(X, Y)$. Now we know the two covariances in (IX-59) to each be $\frac{1}{2}V_A$, so that

$$\mathrm{Cov}\,(X_{MP}, Y) \;=\; \frac{1}{2}\left(\frac{1}{2}V_A + \frac{1}{2}V_A\right) \;=\; \frac{1}{2}V_A. \tag{IX-60}$$

The regression of offspring on midparent is

$$\beta_{O.MP} = \frac{\text{Cov}(X_M, Y)}{\text{Var}(X_M)} = \frac{\frac{1}{2}V_A}{\frac{1}{2}(V_A + V_D + V_E)} = h^2 \qquad \text{(IX-61)}$$

The denominator is determined by the fact that the variance of an average of two independent phenotypes is half as great as the variance of one of them. The result (IX-61) gives us some insight into the meaning of the term “heritability". It measures the fraction of the variation in the phenotype which can be predicted from the phenotypes of the two parents. As we have seen, this is a very different thing from the fraction of variance which can be assigned to genetic causes.

IX.7 Estimating variance components and heritability.

Since the covariances between relatives can be written in terms of the three variance components, it follows that we can estimate the variance components from the covariances. In fact we shall see in later sections that to predict the response to artificial selection we need only two quantities: the total variance of the phenotype and the additive variance V_A. So we can make estimates of these using only two quantities: the observed variance and one covariance. There are three widely-used procedures:

1. Parent offspring regression. If we collect a series of parent-offspring pairs we can estimate both the phenotypic variance and the parent-offspring covariance (the latter depending only on V_A). By doubling the regression of offspring on parent we get an estimate of h^2.

2. Half-sib covariances. Alternatively we could have data on groups of half-sibs. The members of each group all have the same father, but different mothers. Different groups have different fathers. From these numbers one can obtain, via an analysis of variance (ANOVA), estimates of the overall phenotypic variance and of the component of variance due to membership in the half-sib groups. This latter component should be equal to the half-sib covariance which is $\frac{1}{4}V_A$. These numbers again allow us to make estimates of $V_A + V_D + V_E$ and of V_A, and thereby of the heritability.

3. Maximum likelihood. The entire set of data is taken, with the assumption of multivariate normality of the observations, and with the quantitative genetic model supplying values for the variances and covariances in terms of the parameters (say μ, V_A, V_D, and V_E). The likelihood will be the value of the multivariate normal density function above the point which is the observations. The parameters are changed until this is maximized. This approach (Hill and Nicholas, 1974; Shaw,

1987) has the great advantage of using all of the data in an efficient, if computationally tiresome, manner.

PITFALLS AND LIMITATIONS. It is important not to wander into doing these analyses without first acquiring an understanding of some of the difficulties of interpretation and a healthy respect for them. The most serious single problem encountered concerns environmental correlations between individuals. We have been casually assuming that the environmental contribution to the phenotype is independent in different individuals. When one is collecting data this assumption is frequently not met. Relatives not only share common genetic material between them, but frequently also live in similar environments. This leads to an extra term in their covariance, due to the fraction of their environmental variance which is due to factors both share in common. Thus the covariance formulas should actually look something like

$$\text{Cov}\,(X, Y) \;=\; a\,V_A + d\,V_D + e\,V_E, \tag{IX-62}$$

where a and d are the coefficients we have been discussing, which involve the fraction of additive effects and dominance deviations shared between the relatives. These, as we have seen, can be computed from our knowledge of population genetics. But e is another matter. It is the fraction of environmental effects on the phenotype which are due to causes shared by both relatives. Knowing e for a given relationship (say aunt-niece) requires us to have a model of how the environment acts on the trait. We are customarily quite ignorant of this. Unless we knew e it would be quite impossible to estimate V_A, V_D, and V_E. If we allow all possible models of environmental effects to be entertained, these could predict (at least in principle) all possible patterns of covariances purely on the basis of V_E and an arbitrary set of e's, without needing to invoke genetic effects at all. For instance, a 25% correlation among half-sibs might indicate that the trait's variation was entirely additive, so that the 25% reflects genetic relationships, 25% of the additive genetic effects on the character being common to two half-sibs. But it could also be expected to result if the trait had no genetic variation, with only environmental effects, about 25% of which are due to factors common to both half-sibs. How are we ever to untangle this confounding of genetic and environmental factors?

Randomization of environmental effects. In animal and plant breeding we can hope to reduce this confounding by exercising our control over the environments of the organisms. This does not mean that we need try to eliminate environmental variation: all that is necessary is to make environmental effects uncorrelated in relatives. Thus we may want to allocate cattle randomly to pastures, or seeds randomly to plots, so as to prevent relatives from experiencing a common environment more frequently than would be expected at random. Even so, there is a certain amount of common environmental effect which is irreducible. Mammalian offspring in the same litter will have common

prenatal environment and pre-weaning care and nutrition. Seeds from the same mother plant will start with a nutritional package of endosperm which tends to be similar. Even if the offspring are thereafter placed in totally unrelated environments, this common environment will bring about some correlation, which may be mistaken for evidence of genetic variation. It is in an attempt to avoid this that designs for estimating heritability usually avoid computing the covariance of individuals with the same mother or covariances of mother with offspring. For example, in the parent-offspring regression design, it is customary to compute the covariance of father with offspring. In the half-sib design, the half-sibs are paternal half-sibs, which have different mothers. The assumption behind this is that the father contributes only a sperm or a pollen grain. When the father also determines part of the environment of the offspring, even these designs will lead to overestimation of heritability. With animals or plants from natural populations, randomization of the environment is next to impossible, unless one can bring the individuals into the laboratory for cultivation, and do the randomization there. This is a particular problem with humans, who are not amenable to laboratory culture.

When one cannot randomize. With a population experiencing environments in an uncontrolled manner, the best one can hope to do is to measure the environment of each individual, in hopes of removing environmental effects which are common to relatives, leaving behind only a residue which is genetic. The problem with this approach is that we must know which aspects of the environment are the ones relevant to the trait, and we must be able to measure those. This requires us to have a comprehensive understanding of the way in which the environmental factors affect the trait. But that is usually the very thing we are most interested in, so we must know the answer before we can obtain it! In human population genetics, particularly with behavioral measurements such as I.Q. test scores, this problem becomes quite serious. In assuming that certain factors (say income) are sufficient measures of a person's environment, the researcher builds their own social and political assumptions into the conclusion. If the result is then used to bolster these views, there is then a logical circularity. It seems that the best that one can do with this problem is to admit its existence and try to make one's own social and political assumptions explicit, so that those viewing the conclusions can evaluate the results more readily. There is much more that can be said on this subject, but no space here to say it. There are many sources of error other than common environment, and it is worth mentioning a few:

1. *Genotype-environment interactions:* we have in our additive model allowed dominance, which is an interaction effect of two genes. But we have assumed that the environmental effect does not depend on the genotypic effect, and vice versa. If they are interdependent, then this interaction is a source of variation which cannot easily be attributed either to genotype or environment. There is no guarantee that a genotype which raises the phenotype in one environment will not lower it in

another. To the extent that this is important, it means that we cannot use a model in which the phenotype is simply the sum of genotypic and environmental effects. Feldman and Lewontin (1975) have strongly criticized the use of heritability on these grounds, arguing that in the presence of genotype-environment interactions it is a meaningless and potentially misleading quantity.

2. *Genotype-environment covariances.* This sounds like the same thing as genotype-environment interaction, but it is not. If there is an association between the distribution of genotypic effects and the environmental effects, this can cause confusion as to which is acting to cause the phenotypes. For example, in livestock breeding it is not uncommon for the most productive genotypes to be found more commonly in the herds of the best-financed breeders, which are likely also to have the most favorable environments (such as the best food). Under those circumstances the assumption of randomization of environmental effects will be violated, and the environmental effects may be mistaken for genetic effects.

3. *Genetic interactions.* The different loci may also interact, giving rise to extra interaction terms. These genetic interactions affect the covariances between relatives, and may result in misestimation of heritability.

4. *Age and sex effects.* Sometimes the trait in individuals of different ages or sexes is influenced by different genes or different environments. For instance, the weight of an adult may be a somewhat different character than the weight of its offspring. If we wrongly assume in our statistical analysis that adults and juveniles (or males and females) have the same means and variances, we can go considerably wrong in the analysis. Thus the half-sib design has advantages over the parent-offspring design, which may involve measuring both adults and juveniles without knowing whether the trait measured is comparable in both. Even when organisms are measured at the same age, environmental changes from one generation to the next may have altered the statistical and genetic properties of the trait.

5. *Maternal effects.* In mammals, with their large contribution of the parents to the environment of the offspring, and even in nonmammalian species with effects of the egg on the offpsring's environment, maternal effects will be common and can bias the results obtained with an oversimplified model of independent environmental effects. Falconer (1965) has given a simple linear model of maternal effects on a single trait, which can be used to calculate the size of these biases.

6. *Sample size.* A common mistake is to take a small sample of relatives, and compute from it the covariances and heritabilities without noticing that they have large statistical errors. This can lead to exact predictions being made from a totally inadequate set of measurements. In a sample of 20 pairs of relatives, for example,

an observed correlation of 0.4 is not significantly different from zero correlation! It would be of questionable validity to predict from a parent-offspring correlation of 0.4 that the grandparent-offspring correlation will be 0.2, if the sample size were this small.

7. *Epigenetic effects.* In recent years it has been discovered that modifications of the DNA, known as epigenetic effects, can have phenotypic effects and be transmitted to subsequent generations. These do not change the actual DNA sequence. In general these effects are lost within 3-4 generations and are replaced by new modifications. As far as is known the modifications have random effects, so that they do not form a "Lamarckian" system of inheritance, in the sense that their phenotypic effects are uncorrelated in direction with what the organisms needs to cope with that particular environment. Epigenetic effects that are correlated in near relatives will contribute to environmental covariances between relatives, and may be mistaken for genetic effects, leading to an overestimate of long-term selection response. They cannot be the basis for long-term evolutionary change, as they continually revert to their original state on the timescale of a few generations. Slatkin (2009) and Tal et al. (2010) have considered the effects of epigenetic modifications on genetic covariances and heritabilities, pointing out the difficulty of accounting for covariances of all but close relatives by epigenetic effects.

8. *Niche construction.* Organisms affect their own environment, in ways that in turn effect their phenotype (and their fitness). Although this might simply be considered an indirect genetic effect, and therefore already taken into account when heritabilities are estimated, there is one aspect of these "niche construction" effects that needs further consideration. That is that the effects on the environment can persist across generations, so that organisms are thereby affecting their offspring's phenotypes as well. Laland et al. (1999) and Laland and Brown (2006) have made pioneering models of these effects. The model of Falconer (1965) shows some of the selection response to be accounted for by the maternal effect, and thus is in some ways touching on the same issues.

IX.8 History and References

Before the rediscovery of Mendel's work, the dominant theory of heredity involved "blending" inheritance, according to which the offspring would always be phenotypically intermediate between its parents. This was clearly an inadequate theory, since it predicted that all siblings should be identical. Jenkin (1867) pointed out that under this scheme half of the variability (more properly, half of the variance) should disappear each generation. In the last decade of the 1800's Francis Galton (1889, 1897) set forth an

alternative theory. This was developed further by Karl Pearson (1898), and this school of genetical theory has come to be called the Biometricians. In its most sophisticated statements, their theory held that the phenotype value represented a part which was genetic, and that this could be written as

$$P = G + E \tag{IX-63}$$

where

$$G = \frac{1}{4}(P_1 + P_2) + \frac{1}{8}(P_{11} + P_{12} + P_{21} + P_{22}) + \frac{1}{16}(P_{111} + P_{112} + \dots) + \frac{1}{32}(etc. \dots) \tag{IX-64}$$

where P_1 and P_2 are the *phenotypes* of the two parents, P_{11}, P_{12}, P_{21}, and P_{22}, the phenotypes of the four grandparents, and so on through all more remote generations. Pearson called this the Law of Ancestral Heredity. It has some ambiguous and troublesome aspects, but it does make predictions of covariances among relatives. It seems not to predict that there will be a higher covariance among sibs than between parent and offspring. In quantitative genetics, predictions like this have coefficients that depend on heritabilities and thus differ from character to character. Galton and Pearson searched instead for a single equation that would fit all characters, which we now know is impossible.

When Mendel's laws were rediscovered, a controversy ensued between Mendelians and Biometricians. The reader will find good accounts of this in Schwartz (2008) and in Provine's (1971) book. For a careful reexamination of the what modern quantitative genetics and statistics have to say about Galton's work, the monograph by Bulmer (2003) is essential. The covariances between relatives were only part of the rejection of the Biometricians' work; it was also undermined by the ability of the Mendelians to explain the effects of inbreeding. Pearson (1904, 1909a, 1909b) did compute parent-offspring correlations under Mendelian inheritance, under the supposition of complete dominance and equal gene frequencies of the two alleles. Pearson argued that the Mendelian predictions were lower than observed correlations. But the statistician G. Udny Yule (1906) argued that this lack of fit was due to the restrictive assumptions Pearson made.

There the matter lay until it was settled at one blow by R. A. Fisher (1918) in a massive paper which was only beginning to be completely understood half a century later. Fisher defined the three variance components and derived the correlations and covariances of relatives. His paper also contained a detailed treatment of the effects of assortative mating, and it is this aspect of his paper which has only been well-understood in recent years. Sewall Wright (1921a), working independently of Fisher, obtained correlations and covariances among relatives which ignored the dominance variance V_D, which his path coefficient methods could not treat. Most American quantitative genetics theory in the next two decades is based on Wright's work, and most English work on Fisher's. It is an interesting fact that much of modern statistics springs from this controversy and from the work of Galton, Pearson, and Fisher. This includes regression, correlation,

chi-square testing, and the analysis of variance. Although Wright spent more time on biology than on statistics, his method of path coefficients has recently spread to sociology and economics. The most popular methods for deriving the covariance among relatives today were developed from G. Malécot's (1939) reformulation, which made the derivation accessible to many quantitative geneticists and thus stimulated much further work.

IX.9 Response to artificial selection

We have been able to establish a connection between various variances, covariances, and correlations thus far. All can be expressed in terms of three variance components. This in itself is interesting, but of limited practical use. What has made quantitative genetics useful in animal and plant breeding is that it also provides us with a prediction of the response to artificial selection. This in turn can be used to design selection programs. However before we can derive the response to artificial selection, we need to establish yet one more assumption. Until now, we have not specified how many loci were involved in our quantitative character. It is true that if there were only one locus, with large genotypic effects and no overlap of the genotype, then we could use traditional genetic analysis, and we would not be interested in variance components, as we would be able to carry out a far more penetrating analysis. But even in that case, the variance components could be obtained and the covariances among relatives calculated. For there has been nothing in our analysis which restricted it from applying to such a case. Once the assumptions are satisfied, the variance components, covariances and correlations exist and can be computed. The results apply to any number of loci, any number of alleles, and any distribution of environmental effects. For an analysis of response to selection, we must abandon this generality and restrict things somewhat.

NORMALLY DISTRIBUTED PHENOTYPES. We shall assume that each locus contributes only a small fraction of the variance in the character. The character is assumed to be *polygenic*, affected by genetic variation at many loci. We also assume that the environmental effect is drawn from a normal distribution. These assumptions enable us to specify the distribution of the character. Recall that the phenotype is the sum

$$P = g_1 + g_2 + g_3 + \cdots + g_n + e. \tag{IX-65}$$

The g_i are independent of each other, and we have assumed that each contributes only a small fraction of the variance. In mathematical statistics, the Central Limit Theorem tells us that the distribution of a sum such as $g_1 + \cdots + g_n$ will approach a normal distribution as we consider cases with more and more loci (provided each locus contributes less and less of the variance as we progress from case to case). We therefore make the

approximation of saying that the sum of the genotypic effects follows a normal distribution. Now our phenotype P is the sum of two parts, each of which is drawn from a normal distribution. Furthermore, these two parts are independent of each other. The identity of the genotype has been assumed not to affect the environmental effect. The sum of two independent normal variables is again normally distributed, so that we conclude that the phenotype will follow a normal distribution. Cases of actual polygenic characters which are not normally distributed are quite frequent. Presumably they do not follow the rather special assumptions of our model. In some cases a scale transformation will restore normality. More often, there is no particular reason to believe that the assumptions of our model are actually true, and the use of quantitative genetic theory amounts of an act of faith, an approximation made out of lack of any alternative. Figure 9.7 shows the approach to normality as we consider cases with larger and larger numbers of loci. The character is the sum of effects at n loci, with complete dominance at each locus and at each locus a gene frequency of 0.4 for the dominant allele. The locus effects are scaled so that all loci have equal effect, and so that the environmental effects contribute 20% of the variance. The approach to normality can be seen to be quite rapid: it should be a good approximation even when the character is controlled by only a modest number of loci.

Multivariate normality. There are two properties of normal distributions which we will need. The first concerns pairs of relatives. It can be shown that if there are enough loci to have normality of the distribution of phenotypes, then if we draw pairs of relatives from the population and look at their phenotypes, these pairs of numbers will follow a bivariate normal distribution. Such a distribution was used to produce the numbers used to plot Figure 9.5. Furthermore any set of k relatives (say a mother, father, and four offspring) will, if drawn repeatedly from a population, yield a k-tuple of phenotypes which are drawn from a k-variate normal distribution. This fact is highly useful in constructing statistical tests of hypotheses, but we shall make use of it for only one purpose here. That purpose is the establishment of another consequence of normality. This is the linearity of conditional means. What we mean by this is that if we choose a female whose phenotype is (say) 37.5 cm, and have her mate with a randomly chosen male, then the average phenotype of the offspring she, and all other females with the same phenotype, produces will be exactly the prediction we would make from the regression line expressing the regression of offspring on their parents, evaluated at a parental phenotype of 37.5 cm. This may sound completely tautologous, but it is not. Any other joint distribution other than a bivariate normal one would not give this precise linearity of the dependence of offspring mean on parental phenotype. One set of examples which do not show precisely this linearity is all the n-locus cases. It is only as n becomes large that the linearity becomes precise. A similar linearity of offspring means holds when we specify the phenotypes of both parents. From both of them we can predict the offspring phenotype, and the regression line will turn out

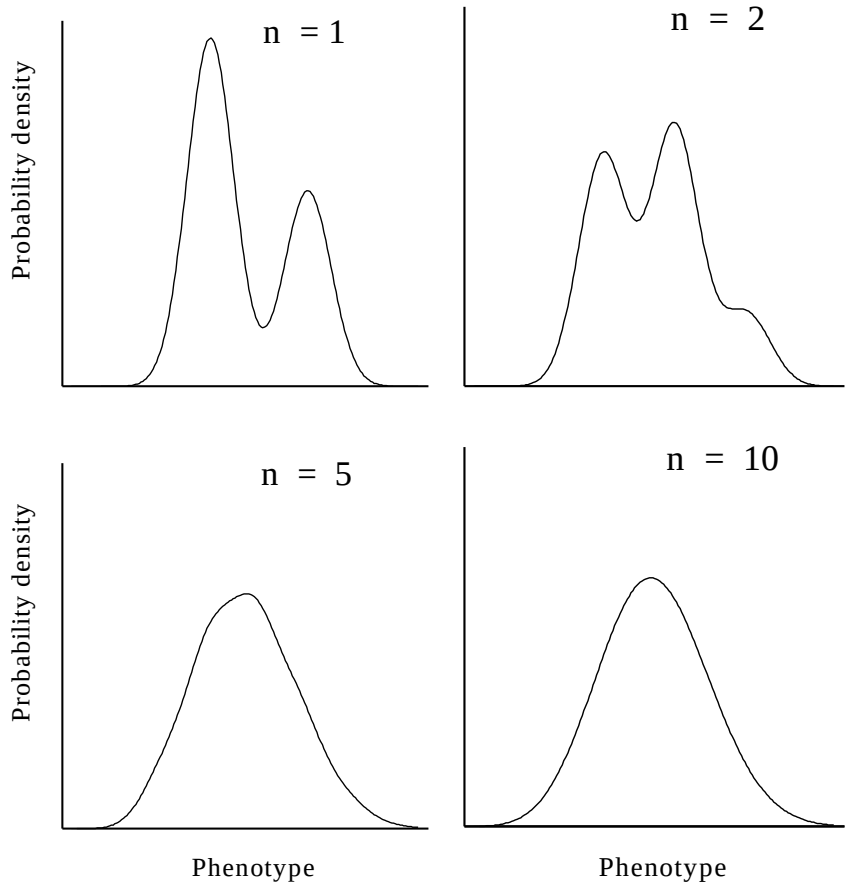


Figure 9.7: Distribution of phenotypes in phenotypes controlled by different numbers of loci. Each locus is completely dominant, with gene frequency of the dominant allele being 0.4. All loci have equal effects. The cases are scaled to have equal variance, with environmental effects contributing 20% of that variance.

also to give the mean phenotype of offspring produced by all pairs of parents having the same two parental phenotypes as these. The predictions we make from the parents are easily written in terms of the regression of offspring on parent. For the case where we choose one parent (say the mother) and based on its phenotype try to predict its offspring's phenotype, then the prediction is obtained from the regression line

$$\mathbb{E}(Y) \;=\; \mu + \beta_{OP}\,(X - \mu) \tag{IX-66}$$

where X and Y are the phenotypes of mother and offspring. In the case where we know the phenotypes of both parents, it can be shown that the prediction is made by a

least-squares regression on two variables, and this turns out to be simply

$$\mathbb{E}(Y) \;=\; \mu + \beta_{OP}\,(X_1 - \mu) + \beta_{OP}\,(X_2 - \mu), \tag{IX-67}$$

X_1 and X_2 being the phenotypes of the two parents. Schraiber and Landis (2015) have investigated further conditions under which normality will not hold – they find reasons to be concerned particularly when there are few loci determining the character, and when there is a "heavy-tailed" distribution of mutational effects, which includes both mutations of small effect, and much rarer mutations of large effect.

RESPONSE TO SELECTION. Now we are in a position to use these formulas to obtain the selection response. Suppose that we were to select from a population a set of individuals for use as parents, and mate them at random, then predict the phenotypes from the offspring of each mating. The average of the predictions will tell us how far the offspring mean will be above the original population mean. Note that in (IX-62) β_{OP} will be $\frac{1}{2}h^2$. Averaging over all the X_1's and X_2's which we choose, we find that

$$\mathbb{E}(\bar{Y}) \;=\; \mu + \frac{1}{2}h^2(\bar{X}_1 - \mu) + \frac{1}{2}h^2(\bar{X}_2 - \mu), \tag{IX-68}$$

where $\bar{Y}$ is the offspring mean and $\bar{X}_1$ and $\bar{X}_2$ are the mean phenotypes of the female and male parents which we have selected. We can immediately see that the response to selection depends only on the mean phenotypes of the selected female and male parents, and thus not at all on the way in which they are mated. However if we wish to use the formula again in the next generation of a selection program, we must know that our assumptions hold in that generation as well. This will scarcely ever be precisely true, but if our mating of the selected parents is random it may nearly be true. If we select equal numbers of males and females, and let $\bar{X}$ be the average phenotype of the selected parents, then (IX-68) can be rewritten as the Breeder's Equation

$$\mathbb{E}(\bar{Y} - \mu) \;=\; h^2(\bar{X} - \mu). \tag{IX-69}$$

In a sense, then, h^2 is the fraction of the selection applied to the parents which has an effect on their offspring. If $h^2 \;=\; 0.3$, then by choosing parents who average 10 kg above the population mean, we obtain offspring who are expected to average 3 kg above the original population mean. Note that formulas (IX-68) and (IX-69) are applicable when we select only one sex and choose the members of the other sex at random. If, say, we mate selected females with randomly-chosen males, then $\bar{X}_2$ will be equal to μ, so that the last term of (IX-68) vanishes. This is exactly what we would have found by using (IX-66) instead of (IX-67). Note also that when there are unequal numbers of the two sexes, (IX-68) shows that we should base our predictions on the simple average of $\bar{X}_1$ and $\bar{X}_2$, not on the overall average parental phenotype.

RESPONSE TO TRUNCATION SELECTION. We have still not said how we selected the parents. An animal or plant breeder trying to make as large a change as possible in one generation will usually try to use as parents the individuals with the largest values of the phenotype which they want to increase. Suppose that the breeder knows that they want to breed from (say) the upper 20% of the herd. How much response can be expected? As we can see from (IX-69) and its predecessors, this depends entirely on how far above the population mean the selected individuals will be. When the character follows a normal distribution, this is easily computed. We can look at a standard normal distribution (one with mean zero and variance 1) and ask what is the mean of the individuals who make up the top 20% of the distribution. Suppose that this is designated as i. We are actually interested in the mean phenotype of the top 20% of a normal distribution with mean μ and standard deviation σ. If we express this as a deviation I from its mean, we are saying that the mean phenotype is $\mu + I$. I is widely used and is called *the selection differential*.

If we have observed I and want to make a prediction from this, we need only multiply I by the heritability of the character. If we have not yet observed I, we need to predict what it will be. The fact that we are considering normal distributions makes this easy. We know that in the standard normal distribution the top 20% is i standard deviations ($i \times 1$) above its mean (0). Then in any other normal distribution the top 20% is also i standard deviations above its mean. So the mean phenotype of selected individuals is

$$\mu + I \;=\; \mu + i\,\sigma \tag{IX-70}$$

or

$$I \;=\; i\,\sigma. \tag{IX-71}$$

This implies that we need only a single table, which tells us for each fraction of individuals retained what the value of i will be. From this along with the standard deviation we can use (IX-71) to find I, the expected difference between the phenotypes of the selected parents and the population mean. Equation (IX-64) tells us that we expect a fraction h^2 of this selection differential to show up in the *selection response* of the mean of the offspring. An example may be useful. Suppose that we have a population of mice weighing an average of 20g, with standard deviation of 2g. We have taken covariances of relatives and obtained an estimate of $h^2 = 0.3$, and wish to predict the result of breeding from the top 10% of the population. For the top 10% one can show that i, the *standardized selection differential*, is 1.76. Then we expect the top 10% to be $2 \times 1.76 = 3.52$ g above the population mean, so that the mean of selected parents should be 23.52g. The mean of the resulting offspring will be $(0.3)(3.52) = 1.056$ g above the original population mean. So the offspring mean will be expected to be 21.056 g. All that is missing is the table of i as a function of the fraction selected. We need to compute, for a standardized normal distribution, the mean of the top S of the curve, where S is the fraction selected.

A bit of algebra gives this, as follows. The selected parents will show a distribution of phenotypes which is the tail of a normal distribution. It will have density function

$$f(x) = \begin{cases} \frac{1}{S}\frac{1}{\sqrt{2\pi}}\exp[-x^2/2], & x \geq c \\ \\ 0 & x < c, \end{cases} \tag{IX-72}$$

where c is the point at which we must truncate a normal distribution so that a fraction S of the area is above the cutoff point. The average value of x in this distribution is what we seek. It is:

$$\begin{aligned} i = \mathbb{E}(x) &= \int_{-\infty}^{\infty} x f(x)\, dx \\ &= \int_{c}^{\infty} \frac{x}{S\sqrt{2\pi}} \exp[-x^2/2]\, dx \\ &= \frac{1}{S\sqrt{2\pi}} \int_{c}^{\infty} x\, e^{-x^2/2}\, dx \end{aligned} \tag{IX-73}$$

and noticing that the quantity under the integral sign is the derivative of $-\exp(-x^2/2)$, we get

$$i = \frac{1}{S\sqrt{2\pi}} e^{-c^2/2}, \tag{IX-74}$$

so that the standardized selection differential is simply the height of a standard normal curve at the cutoff point c, divided by the fraction of area (S) above the cutoff point. The cutoff point c and the height of the curve at that point can be determined using standard normal tables. Table 9.2 shows values of i: Note that even when we select the top 0.001 of the parents, we get a selection differential of only 3.4 standard deviations: there effectively aren't any individuals available who are more than four standard deviations above their population mean.

SELECTION EFFECTS AT A SINGLE LOCUS. To what is the selection response due? If the response has been the result of creating a departure from Hardy-Weinberg proportions, or creating linkage disequilibrium, then we would not expect it to be retained in future generations: the population will fall back to its previous mean. But if the selection response is mostly a result of changing gene frequencies, then we may expect the gain to persist, and perhaps we may even expect a further gain if the selection procedure is repeated each generation. We now show that the gain is in fact due to changes of gene frequency. To do this we need the amount of change due to selection caused at one locus during the process of truncation selection. Suppose that we have a two-allele locus. Let us ask how likely it is that a copy of the A_1 allele survived selection. When we computed the average excess α_1 for this allele, this was the mean phenotype of all bearers of A_1

Table 9.3: Standardized selection differentials, i, corresponding to various fractions of parents selected (S), as well as cutoff points c.

S	i	c
0.90	0.195	-1.2816
0.80	0.350	-0.8416
0.70	0.497	-0.5244
0.60	0.644	-0.2533
0.50	0.798	0
0.40	0.966	0.2533
0.30	1.159	0.5244
0.20	1.400	0.8416
0.10	1.755	1.2816
0.05	2.063	1.6449
0.01	2.665	2.326
0.001	3.367	3.090

alleles, weighted by the the number of A_1 alleles they carry and expressed as an excess over the population mean. Furthermore, knowing that an individual carries A_1 tells us nothing about the rest of its genome or the environmental variance. Since the bulk of the variance comes from those sources, and only a little of it has been eliminated by fixing one allele at A_1, we will approximate well by saying that A_1-bearing individuals have a mean of $\mu + \alpha_1$ and a variance of σ^2. When we save all individuals whose phenotypes are beyond a value of c, what fraction of A_1 genes survive selection? Let $\phi(x; \mu, \sigma^2)$ be the density of the normal distribution with mean μ and variance σ^2. The fraction of surviving A_1-bearing individuals is

$$\begin{aligned} w_1 &= \int_c^{\infty} \phi(x; \mu + \alpha_1, \sigma^2)\, dx \\ &= \int_c^{c+\alpha_1} \phi(x; \mu + \alpha_1, \sigma^2)\, dx \; + \; \int_{c+\alpha_1}^{\infty} \phi(x; \mu + \alpha_1, \sigma^2)\, dx \end{aligned} \tag{IX-75}$$

The rightmost term is the area beyond $c + \alpha_1$ in a normal distribution centered at $\mu + \alpha_1$. This will be the same as the area beyond c in a distribution centered at μ. This we have already specified to be S, the fraction saved. So

$$\begin{aligned} w_1 &= \int_c^{c+\alpha_1} \phi(x; \mu + \alpha_1, \sigma)\, dx \; + \; S \\ &\simeq \alpha_1 \phi(c; \mu, \sigma) \; + \; S \end{aligned} \tag{IX-76}$$

The second line of this expression is obtained by assuming that α_1 is small. The integral of a curve over a short interval of width α_1 is nearly its height times α_1. Now we notice that the height of the normal curve at the cutoff point c will simply be iS/σ, where i is the standardized selection differential. So

$$\frac{w_1}{S} \simeq 1 + \frac{\alpha_1 i}{\sigma} \tag{IX-77}$$

Thus the A_1 alleles are expected to increase in frequency by a fraction $i\alpha_1/\sigma$ of their previous value. There is not the space for it here, but one can show using the new gene frequency $p' \simeq p(1 + i\alpha_1/\sigma)$ and making use of (IX-22) and (IX-31), that the contribution of this locus towards the mean of the offspring is increased by

$$(i/\sigma)\ [\ 2p\,\alpha_1^2 + 2(1-p)\,\alpha_2^2\ ]. \tag{IX-78}$$

Thus, adding over loci, the changes in gene frequency bring about an increase of the mean by

$$\Delta G = iV_A/\sigma = i\sigma V_A/(V_A + V_D + V_E) = ih^2\sigma = Ih^2 \tag{IX-79}$$

But this is a fraction h^2 of the original selection differential. So the changes in gene frequency account for *all* of the response to selection. The gain due to selection should thus be permanent and enduring.

EFFECTS OF REPEATED SELECTION. The first time selection is applied, the gain is entirely due to changes of gene frequency at the individual loci influencing the trait. If the assumptions which were used in establishing this result continue to hold, we should expect to see continued gain from repeated generations of selection. Chief among the assumptions is the independence of genotypes at different loci, in other words, linkage equilibrium, and if a single generation of truncation selection generates no linkage disequilibrium, it should remain in that state. Unfortunately for our analysis, truncation selection does generate some linkage disequilibrium. In fact, it tends to put alleles which increase the phenotype in repulsion, so that the net effect is to decrease the variance of the trait. The usual method of analysis in animal and plant breeding is to ignore this disequilibrium, and assume that the loci are loosely linked, so that little disequilibrium will remain in the next generation. If this assumption is a reasonable one, then we should again expect to see a selection response of Ih^2 in the next generation.

Bulmer (1971) has presented an approximation which corrects for the linkage disequilibrium induced between unlinked loci. Even if we ignore disequilibrium we must still know what selection does to h^2 and σ. We shall assume that the environmental effects remain the same from generation to generation, and that the loci continue to contribute additively to the phenotype. Other than linkage disequilibrium, the only quantities which could change from one generation to the next are the gene frequencies. If there are many loci contributing substantially to the character, then the average excess

α_1 of one allele will be small compared to the standard deviation of the character. Formula (IX-77) then provides us with some assurance that selection will make only small changes in gene frequency at any one locus. Thus the variance components V_A and V_D will change little in any one generation. Therefore h^2 and σ will also change little. The usual procedure in predicting response to selection is to assume that h^2 and σ remain unchanged from their initial values, so that if a constant fraction S of the population is saved each generation, the response in each generation will be $ih^2\sigma$.

COMPLICATIONS AND LIMITATIONS. The calculation of the response to artificial selection is quite assumption-bound, so that there are a great many places where the argument could go wrong. One of the most vulnerable assumptions is the constancy of the range of environmental effects. In actual animal and plant breeding, the conditions of husbandry are changing continuously, just as the genetic characteristics of the strains are. Under modern industrialized agriculture, there is frequently a tendency for increased mechanization of rearing to lower costs but also to result in more unfavorable conditions for the animal or plant. So the environment is continually deteriorating, and this creates a serious problem for anyone who wants to know whether selection is in fact bringing about the predicted improvement. Selection which is in fact counteracting the deteriorating environment may appear to be having no effect. Alternatively, selection may receive credit for increases of yield which are actually the result of improved agricultural techniques. Clearly in any breeding program it is worth paying a great deal of attention to long-term environmental changes, and attempting to measure them and correct for them.

Natural selection. A second source of difficulties is natural selection. We have assumed that it is absent in our model of a quantitative character, but clearly this is unjustified in general. Characters which have not been under strong artificial selection may owe their present values to natural selection. This natural selection is far more likely to be stabilizing selection than directional selection. In applying artificial selection we are reshaping the organism to our own requirements, not those of natural selection. It is quite likely that among our selected parents, fertility is lowest in those which appear best to us for our own requirements. Likewise among their offspring, those may survive least well which most closely fit our requirements. Natural selection and artificial selection will then be antagonists. As the phenotype departs farther from the original population mean, the intensity of natural selection may increase. Ultimately we will reach a point in the selection program where natural selection prevents further progress from artificial selection. If we see the plateau of selection response, we may mistakenly conclude that it is a result of fixation of the favorable alleles. Imagine our surprise when we cease artificial selection, and observe the phenotype gradually receding towards its original value, owing to the unopposed operation of natural selection! A more careful approach would involve checking at the plateau to see whether heritability had reached zero. If it

has not, then we have the paradoxical situation of being unable to get further response to artificial selection in spite of the presence of additive genetic variance. There is an increasing suspicion among animal and plant breeders that many of these paradoxical plateaus in selection response are the result of natural selection.

Gene interaction. A third source of complications is gene interaction. We have assumed that on some scale the different loci have effects that add up. Sometimes the proper choice of a scale transformation will bring one nearer to this idealization, but in general genes interact by a variety of complex mechanisms, so that there is no reason to expect perfect additivity. Some considerable effort has gone into incorporating genetic interaction into quantitative genetic theory. Cockerham (1954) and Kempthorne (1954) independently arrived at a method for breaking down the phenotypic variance into multiple components when interaction is present. They were able to find reasonably simple formulas for the covariances between relatives in terms of these variance components. Griffing (1960) made substantial progress towards predicting selection response when interaction between loci is present. The difficulty with these papers is that the approach they use is not very useful in practice. Even if we only allow for two locus interaction, the variance is broken down into six components:

$$V_T = V_A + V_D + V_{AA} + V_{AD} + V_{DD} + V_E. \tag{IX-80}$$

Finding enough different covariances of relatives, and large enough sample sizes, to estimate these six quantities is essentially impossible in practice. The variance components approach to interaction has had little impact in practice. The best we can do in practice is to ensure that we remove interaction effects as much as possible by scale transformation. Beyond that we can only look for an understanding of the effects of individual loci, which is easier wished for than achieved.

Genetic drift. Another serious complication is the presence of genetic drift. This can cause changes in the variance components. A careful study of this problem has been made by Robertson (1960). He found that the fixation probability formulas of Kimura could be applied, to predict the limiting phenotype under selection. The average result of genetic drift is to reduce the heritability. If only N adults are preserved each generation, the additive genetic variance will ultimately disappear, and no further progress will be made thereafter. This is actually simply another way of stating that in a finite population, there is a nonzero chance of losing advantageous alleles. Thus we may be less interested in rates of change of phenotype than in the ultimate selection limit. A particularly interesting result concerns the case in which n individuals are screened with only a fraction retained. In that case if n is fixed, the more strongly we select, the fewer adults we will choose. There is then a tradeoff between the immediate response, which is greater the more strongly we select, and the selection limit, which is less the

fewer adults we save. On the other hand, if we select weakly we will make little progress before our additive genetic variance is lost. Robertson showed that the optimum procedure, from the point of view of the selection limit, is to save the upper half of the population. This can be proven using our equation (IX-74) to compute the value of $4Ns$, and finding the level of selection which maximized that quantity. Hill and Robertson (1966) subsequently investigated the effects of linkage on selection limits. As we saw in section VIII.8, they found an interesting phenomenon (the "Hill-Robertson effect") in that selection at neighboring loci in a small population tended to interfere with selection response at all of these loci, even if the loci neither interact in their fitnesses nor are in initial linkage disequilibrium. This phenomenon seems to be quite general. Hill (1971, 1972a, 1972b) has made a more exhaustive study of the effects which finite population size will have on the variability of selection response.

Two lesser concerns are with the finiteness of the number of loci and with the fact that we never quite achieve the selection intensities we expect. Latter (1965) has looked into the effects of having a few loci with large effects on the character. He finds that this will rarely cause serious difficulties. Hill (1976) and Rawlings (1976) have investigated the possibility that the top (say) 20% of the group of individuals may not be near the top 20% of a normal distribution, owing to the fact that there are groups of relatives among them. This seems to be a secondary problem compared to the others we have mentioned.

IX.10 History and References

While selecting the top of a herd or crop is an ancient practice, it was only in the 1930's that quantitative genetics attempted to predict the response to selection. Haldane (1930c) was the first to compute the selection intensity by reference to tails of normal curves. Using the work of Fisher, Wright, and Haldane, Jay L. Lush fought for the introduction of quantitative genetics into animal breeding. His book, *Animal Breeding Plans* (1937) was a landmark in introducing these techniques to wider audiences. Of particular note during the early years of quantitative genetics were the papers of Fairfield Smith (1936) and Hazel (1943) on "index selection" (selection based on a combination of traits), of Hazel and Lush (1943) and Lush (1947a,b) on selection based on the performance of near relatives, of Comstock, Robinson, and Harvey (1949) and Dickerson (1952) on selection of two lines based on the performance of the cross between them, and of Robertson and Lerner (1949) and Dempster and Lerner (1950) on all-or-none traits. We have already cited a number of more recent papers which extend or check selection theory. The reader interested in further enlightenment will find it in the excellent books by Falconer and MacKay (1996) and Lynch and Walsh (1998).

Exercises

1. Suppose that we have a trait controlled by two alleles at each of two unlinked loci, and that (i) all double homozygotes have phenotype 1, (ii) all single heterozygotes have phenotype 2, and (iii) all double heterozygotes have phenotype 4. If we cross two strains, one *AABB* and the other *aabb*, what are the mean phenotypes of these parent strains? of the F1? of the F2? of the two backcrosses F1 $\times$ P1 and F1 $\times$ P2? Are the rules concerning means of crosses obeyed? Why or why not?

2. In the above case, would the rules concerning means of crosses be obeyed if instead of measuring the phenotype we measured its logarithm? its square root?

3. Suppose that a trait is the sum of effects at two loci. At one locus the contributions of *AA*, *Aa*, and *aa* are 1, 3, and 2. At the other locus the contributions of *BB*, *Bb*, and *bb* are 2, 1, and 6. If both *A* and *B* have gene frequencies of 0.6, compute the mean phenotype when the inbreeding coefficient is f. Does this depend on the probability F_{11} that both loci are simultaneously inbred? Why or why not?

4. Compute the total phenotypic variance in a one-locus trait where there are no environmental effects and the phenotypic means of *AA*, *Aa*, and *aa* are 1, 1, and 0. Obtain this variance as a function of gene frequency, p. By direct enumeration of all possibilities and their relative frequencies obtain the covariance of parents and offspring in this case. What do these two expressions tell you about the way that the heritability depends on p? Does this depend on the a_{ij}?

5. Suppose that there is a quantitative character in which there is only one gene affecting it, and the rest of the variation is environmental effects. If the average phenotypes of the three genotypes are for *AA*, *Aa* and *aa* the quantities 11.0, 11.0, and 8.0, and there are Hardy-Weinberg proportions of the genotypes with gene frequency p of allele *A*,

   (a) What will be the mean phenotype in the population, as a function of p ?

   (b) What will be the variance of the phenotype as a function of p ? (Note that the variance is the expectation (theoretical mean) of the square of the character, x^2, minus the square of the expectation of the character).

   (c) If the frequency of allele *A* is 0.4, and if the environmental effect on the phenotype of an individual is drawn at random from a distribution that has variance 10, and simply added to the genetic effect, what fraction of the variance of the character is genetic?

   (d) Is this the heritability of the character? Why or why not? (You are not asked to calculate that number).

6. In terms of the three variance components V_A, V_D, and V_E, and/or the heritability, what are the covariance and the correlation between your maternal half-sib and your paternal uncle?

7. Use (IX-51) and (IX-52) to compute the covariance and correlation of an individual with itself. What is wrong with this?

8. Suppose that we find no inbreeding depression in a trait which is the sum of effects at loci each with two alleles. Does this mean that the trait has no dominance variance if the trait depends on only one locus? two loci?

9. Consider a trait whose numerical value is controlled by two loci, without any environmental variance. If the values specified by the genotypes at these two two-allele loci are:

	BB	Bb	bb
AA	6	7	8
Aa	7	8	9
aa	8	9	10

   and both genes have 50:50 gene frequencies (so that $p_A = 0.5$ and $p_B = 0.5$)

   (i) What will the mean value of this phenotype be?

   (ii) What will its variance be? Its standard deviation?

   (iii) Are these two loci individually additive in the effects of their alleles? Do they show any interaction (epistasis)?

   (iv) In view of (iii), what will the heritability of the trait be?

   (v) If we carry out artificial selection, saving all those individuals whose phenotypes are 8 or greater, what will the response to selection be in the first offspring generation?

10. Suppose that we have a trait that shows a mean of 105.2, a variance of 30.6, and a father-offspring correlation of 0.45. What is the heritability? If we choose fathers that measure exactly 110 on the scale, what will be the mean trait value of their offspring if they each mate with a randomly chosen female?

11. Suppose that we find a parent-offspring correlation of 0.1 and a grandparent-offspring correlation of 0.03. Can we determine heritability from this? Why or why not?

12. Suppose that we have two genotypes, one with mean phenotype 10 and the other with mean phenotype 11. The environmental effect is normally distributed with

mean zero and variance 1.2. Use a table of areas of the normal distribution to compute the fitnesses of these genotypes under a regime of truncation selection where all individuals above 11.2 are saved.

13. Suppose that we have an overdominant locus with the mean phenotypes 3, 4, and 2 for genotypes AA, Aa, and aa, a gene frequency of $2/3$ for A, and an environmental contribution which is normally distributed with variance 1. Compute the heritability of this trait. Based on this, what will be the response to one generation of truncation selection in which all individuals above 5 are saved?

14. In the preceding exercise, use tables of the normal distribution to compute the fitnesses of the three genotypes when we carry out truncation selection saving individuals with phenotypes above 5. Do we expect to see any response to one generation of this selection?

15. Is there a discrepancy between the answers to exercises 13 and 14? Explain why we do or do not expect one to exist.

Complements/Problems

1. If the two loci in Exercise 1 of this chapter instead had recombination fraction r between them, how would the mean phenotype of the F2 generation depend on r?

2. If two traits are each determined multiplicatively by many loci, and we are interested in their ratio, does taking logarithms make this ratio additively determined by the loci in the sense that the log (of the ratio) is now additively determined by the loci? Is this result altered if some of the same loci contribute two the two traits?

3. If we have a trait (say, numbers of bristles) which has for each genotype a Poisson distribution, then it is known that its environmental variance can approximately be made constant over all genotypes by working not with the number of bristles but with its square root. To be able to use this transformation and our model of determination of a phenotype, what would be have to assume about the way effects at different loci combined?

4. Do the relationships among means of inbred lines and their crosses and backcrosses hold when we have multiple alleles? When we have a sex-linked trait with two alleles and with hemizygotes which always resemble the corresponding homozygote?

5. Does the linearity of the effect of inbreeding on the population mean continue to hold in the case of multiple alleles?

6. Use (IX-35) to confirm the multiple-allele formula (IX-36) by differentiating Q with respect to α_i and equating the result to zero.

7. What is the correlation between the additive effect A and the total phenotype $\mu + A + D + E$? Can you see from this why heritability is written h^2 ?

8. If we have locus at which recessive deleterious mutants occur at rate u, with the fitness of the homozygote between these alleles being $1 - s$, we know that we should have an equilibrium frequency $\sqrt{u/s}$ of the deleterious mutant allele.

   (i) What is the additive genetic variance of fitness at this equilibrium?

   (ii) What is the total variance of fitness in this case?

   (iii) What is the heritability of fitness in this case?

   (iv) Suppose that there are n such loci with all of them being in linkage equilibrium, so that genotypes at one locus are independent of genoptypes at any other. The natural logarithm of the fitness predicted for an individual by its genotype is the sum of the logs of the predictions for each locus (because the predicted fitness is the products of the predicted fitnesses at each locus). What is the heritability of log predicted fitness? How does it depend on n? Why didn't we ask about the predicted fitness itself in this case, instead of its logarithm?

9. If one predicted the phenotype of an individual from the mean of two of its siblings (by regression), will this be a better or worse prediction of its phenotype than the midparent? How does the presence of environmental correlations affect this result?

10. Which is expected to be the better predictor of an individual's phenotype, if no environmental correlations are present, the midparent or the mean of the four grandparents? Are the two predictions going to be the same?

11. In terms of h^2, what is the (genetic) correlation between offspring and midparent? What is the regression coefficient of midparent on offspring (i.e. with X being the phenotype of the offspring and Y the midparent)?

12. In a one locus two-allele case with genotypic means 1, 3, and 4 of genotypes AA, Aa, and aa and no environmental variance, is the regression of offspring on parent perfectly linear?

13. Suppose that one genotype has mean μ_1 and environmental variance σ_1^2, and that another has mean μ_2 and environmental variance σ_2^2. What is the rule as to which genotype has higher fitness under truncation selection if $\mu_1 \neq \mu_2$ and $\sigma_1 = \sigma_2$? If $\mu_1 = \mu_2$ and $\sigma_1 \neq \sigma_2$? What does the latter say about one of the possible unpleasant side effects of truncation selection? What is the general rule for arbitrary μ_1, μ_2, σ_1, and σ_2?

14. Can we use the ultimate limit reached under truncation selection to get an idea of how many loci are affecting variation in a trait? What are the limitations of this approach?

15. *(Harder)* Suppose that we take n males and n females from a random-mating population, and construct a *diallel cross* by making all n^2 possible matings and measuring one offspring from each. If we arrange the resulting numbers in a square and do an analysis of variance, what are the expected variance components for rows, columns, and interaction in terms of V_A, V_D, and V_E?

16. *(Harder)* How are the results altered in the preceding problem by using instead males from n inbred lines and females from n other inbred lines, each line being totally inbred starting independently from the same population?

Chapter X

MOLECULAR POPULATION GENETICS

X.1 Introduction

The flood of molecular sequences in molecular evolution has reached inside populations. It is no longer true that each species is necessarily represented by only one sequence of each gene ("the mouse sequence", "the human sequence"). Now it has become more common for some attempt to be made to assess population-level variability by collecting population samples of sequences at a single locus. We are now seeing more and more studies that sequence multiple loci in the same individuals and populations, adding an additional dimension to the data. In addition, genomics is no longer content with sequencing one genome of one species, but has expanded, not only to multiple species, but to characterization of genetic variability within and between populations.

These developments of the 1980's and 1990's, together with the previous expansion of restriction-sites data, has created a new field of evolutionary genetics, molecular evolutionary genetics. It is now joined by evolutionary genomics.

X.2 Mutation models

While the genetic drift, natural selection, and migration of sequences is normal, the way they mutate is distinct. When a sequence with many sites in it undergoes mutation, one does not merely see a series of distinct alleles, but to some extent one can reconstruct the history of the mutations by examining the sequences in detail. With n sites, there are in a nucleotide sequence 4^n possible sequences.

THE JUKES-CANTOR MODEL. The simplest possible model of mutation simply assumes that the same mutational process, with the same mutation rate, occurs at all n

sites independently. The simplest model for mutation at a single site is to assume that each nucleotide has the same probability of mutation, and when it mutates it changes to one of the three other possible bases with equal probability. This is the model of Jukes and Cantor (1969).

If the mutation rate is μ per unit generation per site, and if P_{ij} is the probability that base i changes to base j in one generation, then the mutation matrix looks like this:

$$\mathbf{P} = \begin{bmatrix} 1-\mu & \mu/3 & \mu/3 & \mu/3 \\ \mu/3 & 1-\mu & \mu/3 & \mu/3 \\ \mu/3 & \mu/3 & 1-\mu & \mu/3 \\ \mu/3 & \mu/3 & \mu/3 & 1-\mu \end{bmatrix}. \tag{X-1}$$

For the whole sequence, the probability that one mutates, in one generation, from one given sequence to another that differs from it in m out of the n sites, is easily computed as the probability of getting particular changes at those m sites and no change at the others:

$$(\mu/3)^m (1-\mu)^{n-m}. \tag{X-2}$$

We will have use for the transition probability at a site between base i and base j in t generations. This is of course the ij element of the t-th power of the mutation matrix $\mathbf{P}$. While that can be directly computed, we can get it more easily by a slightly indirect argument, which we will also need for some further models below.

Note that if t is large and μ is small, we may approximate the process by having time (measured in units of generations) be a continuous scale, with a constant risk of mutation. Now notice also that if we had a type of mutation that changed a base to one of the four bases chosen at random, with equal probability of the four outcomes, this would look almost like the Jukes-Cantor model, except that $1/4$ of the time it would make no change at all. Now imagine that we have this altered mutation model but we increase the rate to $\frac{4}{3}\mu$. In that model the mutations that change the bases occur at rate μ, continuously in time.

But in that altered model (the one with increased mutation rate) it is quite easy to compute the probability that we end up with base j after having started with base i. If there has been any mutation at all during the time time span t, then the probability of ending up at base j is $1/4$. Now mutation in this continuous-time version of the model has a dynamics like waiting for the first atomic decay of a radioactive substance. Thus the probability that there is no decay at all in t units of time is the exponential $\exp(-\frac{4}{3}\mu t)$. Putting all of this together, we easily show that the transition probabilities (they are called that for mathematical reasons, not having anything to do with transitions

and transversions) are

$$\begin{aligned} P_{ij} &= \tfrac{1}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right) \qquad (i \neq j) \\ P_{ij} &= e^{-\frac{4}{3}\mu t} + \tfrac{1}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right) \quad (i = j) \end{aligned} \tag{X-3}$$

We can also compute, from (X-3), the expected fraction of sites in which two sequences will differ, as a function of mutation rate and time. This is useful in some kinds of phylogenetic inferences. Adding up the three values of P_{ij} for the changes from base i to all three different bases, the expected fraction D of sites differing becomes

$$D = \frac{3}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right). \tag{X-4}$$

This equation can easily be inverted to solve for the "branch length" μt as a function of D by solving for the exponential and then taking logarithms:

$$\mu t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right). \tag{X-5}$$

KIMURA'S 2-PARAMETER MODEL. The Jukes-Cantor model is easy to analyze, but lacks some of the structure of more realistic models of base change. The most notable absence is the inequality of transitions and transversions. This was corrected by Kimura (1980), whose "2-parameter model" (sometimes called the K2P model) we now describe. This model has the simplest form of transition-transversion inequality possible. All four bases are equally frequent. They all mutate with equal rates μ. When they do, the probability of transition is R times as great as the total probabilities of both possible transversions. The result is the mutation matrix for bases ordered A, G, C, T:

$$\mathbf{P} = \begin{bmatrix} 1-\mu & \frac{R}{R+1}\mu & \frac{\mu}{2R+2} & \frac{\mu}{2R+2} \\ \frac{R}{R+1}\mu & 1-\mu & \frac{\mu}{2R+2} & \frac{\mu}{2R+2} \\ \frac{\mu}{2R+2} & \frac{\mu}{2R+2} & 1-\mu & \frac{R}{R+1}\mu \\ \frac{\mu}{2R+2} & \frac{\mu}{2R+2} & \frac{R}{R+1}\mu & 1-\mu \end{bmatrix} \tag{X-6}$$

I have expressed this matrix, not in Kimura's original parameterization, but in one of my own that uses the more immediately meaningful parameters μ (the total rate of potentially observable mutations) and R (the transition/transversion ratio).

For many purposes what we need for each of these models is the transition probabilities, where the "transitions" are all changes of state. In molecular biology changes among purines or among pyrimidines are referred to as "transitions" and all other changes as "transversions". I will try to make it clear when the word transition is used in its stochastic processes sense and when it is used in its molecular biology sense. In the latter case the word transversion is usually nearby.

The simplest method of computing the transition probabilities for Kimura's 2-parameter model is to use a method similar to the one used above for the Jukes-Cantor model. Let's assume that time is continuous, as before. We can reframe the K2P model as involving two kinds of events. One (type II) occurs at rate $\beta = 2\mu/(R+1)$. When it occurs, a random one of the four bases is chosen to replace the present base. This gives us the correct rate of transversions such as $A \to C$ and $A \to T$. However it does not give us enough transitions such as $A \to G$. To get those to come out, we must add another kind of event (type I) with rate $\alpha = (R - \frac{1}{2})\mu/(R+1)$ that can make either no change or a transition, and cannot make a transversion. If this event happens, and the base was a purine, one of the two purines is chosen at random to replace the base. If it was a pyrimidine, one of the two pyrimidines is chosen at random to replace the base.

This keeps the model the same, but defining these imaginary events makes the calculations easy. If a lineage has even a single event of type II, its final state is random among all 4 bases. If it has no event of type II but at least one of type I, it is random among both purines, or among both pyrimidines, depending on which type of base was present. Note that sometimes either event will result in no change, so not all these events are real, but imagining them does result in the right probabilities.

The upshot is that the probability of a particular transversion change (say getting a T given that one started with an A) is

$$\text{Prob}(T \mid A, t) \;=\; \left(1 \;-\; e^{-\beta\,\mu\,t}\right)\;\frac{1}{4} \tag{X-7}$$

and the probability of the transition change is

$$\text{Prob}(G \mid A, t) \;=\; e^{-\beta\,\mu\,t}\;\left(1 \;-\; e^{-\alpha\,\mu\,t}\right)\;\frac{1}{2} \;+\; \left(1 - e^{-\beta\,\mu\,t}\right)\;\frac{1}{4} \tag{X-8}$$

THE HKY MODEL. The Kimura K2P model allows for inequality of transitions and transversions, but is still unrealistic in always leading to equal equilibrium frequencies of the four bases. Hasegawa, Kishino, and Yano (1985) introduced a model that allows for inequality of base frequencies. Its rates of change for the bases ordered A, G, C, T

can be written as

$$\begin{bmatrix} -- & (\alpha+\beta)\ \mu\ \pi_G & \beta\ \mu\ \pi_C & \beta\ \mu\ \pi_T \\ (\alpha+\beta)\ \mu\ \pi_A & -- & \beta\ \mu\ \pi_C & \beta\ \mu\ \pi_T \\ \beta\ \mu\ \pi_A & \beta\ \mu\ \pi_G & -- & (\alpha+\beta)\ \mu\ \pi_T \\ \beta\ \mu\ \pi_A & \beta\ \mu\ \pi_G & (\alpha+\beta)\ \mu\ \pi_C & -- \end{bmatrix} \tag{X-9}$$

The transition (and transversion) probabilities for this model can be worked out similarly to Kimura's model (see Felsenstein, 2004, chapter 13). The equilibrium frequencies of this model are the quantities $(\pi_A, \pi_G, \pi_C, \pi_T)$ which are, in effect, parameters of the model. To get the overall transition/transversion ratio to be R and the overall rate of base change to be μ it is necessary to set

$$\alpha \ = \ \frac{R}{R+1}\ \frac{1}{F} \ - \ \frac{1}{R+1}\ \frac{1}{1-F} \tag{X-10}$$

$$\beta \ = \ 1 \ - \ \alpha\ F \tag{X-11}$$

where

$$F \ = \ 2\ \pi_A\ \pi_G \ + \ 2\ \pi_C\ \pi_T. \tag{X-12}$$

Note that for some small values of R some rates in the matrix become negative – not all values of R can be achieved.

There is a similar but slightly different model introduced by me, the F84 model, and a more general model due to Tamura and Nei (1993) that includes both as special cases. For all of these the net probabilities of change can be calculated (see Felsenstein, 2004, chapter 13).

THE GENERAL TIME-REVERSIBLE MODEL. A 9-parameter model can be defined that has the often-desired property of reversibility. This ensures that the fraction of all changes that are from state i to state j is expected to be equal to the fraction that are from state j to state i. (Note that this is *not* the same as assuming that the transition probability matrix is symmetric).

We will not make use of the property of reversibility here, but it is still worth stating the model, as the only widely-used 9-parameter model. It has rate matrix

To: / From:	A	G	C	T
A	$--$	$\pi_G\,\alpha$	$\pi_C\,\beta$	$\pi_T\,\gamma$
G	$\pi_A\,\alpha$	$--$	$\pi_C\,\delta$	$\pi_T\,\epsilon$
C	$\pi_A\,\beta$	$\pi_G\,\delta$	$--$	$\pi_T\,\zeta$
T	$\pi_A\,\gamma$	$\pi_G\,\epsilon$	$\pi_C\,\zeta$	$--$

(X-13)

This has four base frequencies and six other rates; there are a total of nine parameters if we standardize the total rate of change to 1 by insisting that

$$2\pi_A \pi_G \alpha + 2\pi_A \pi_C \beta + 2\pi_A \pi_T \gamma + 2\pi_G \pi_C \delta + 2\pi_G \pi_T \epsilon + 2\pi_C \pi_T \zeta = 1 \qquad \text{(X-14)}$$

The GTR model was introduced by Lanave et al. (1984). There are no simple formulas for transition probabilities, at least not ones any simpler than using the general solution to the cubic equation. It is best to compute transition probabilities by numerical calculation of the eigenvalues and eigenvectors. This can be done by expressing the rate matrix as a product of a diagonal matrix and a symmetric matrix (see Felsenstein, 2004, chapter 13).

X.3 Approximate mutation models

While these mutation models attempt to approach realism, calculations using them may suffer from intractability. Two models, one of which we have already seen, sacrifice some of the realism for greater tractability.

THE INFINITE-ALLELES MODEL. We have already seen (in Chapter VII) the infinite-alleles model of Kimura and Crow (1964). It assumes that all alleles mutate at equal rates, and when they do, all give rise only to new alleles that have never been seen before. When it is used for nucleotide sequences, this model in effect treats all sequences as differing, without inquiring at which sites they differ or by how many sites. Any statistical treatment based on this method will thus necessarily discard all information about which alleles are historically close to each other.

THE INFINITE-SITES MODEL. To model this historical information an "infinite-sites" model has been proposed (Kimura, 1969; Ewens, 1974; Watterson, 1975). In this model each mutational event occurs at a new site, introducing an alternative allele at that site. The original haplotype is known, and so is the presence or absence of the alternative allele at each other site. The model has been used only in the case where there is no recombination between sites; the exact order of sites is thus unimportant.

The infinite-sites model does retain rather powerful historical information: it is possible to reconstruct for each haplotype which haplotype was its parent. Figure 10.1 shows an example of haplotypes produced by mutation in the infinite-sites model. The sites that have mutated are shown, with the new mutant always designated as 1 and the original state as 0. The historical information is accessible. For instance, the haplotype 0100 is intermediate between haplotypes 0000 and 0110.

Although it is possible to have a version of the infinite-sites model that has recombination between the sites, this does not lead to tractable mathematics, so work with the infinite-sites model generally assumes that there is no recombination. Given that, the left-right order of sites is arbitrary. Thus, the haplotypes in Figure 10.1 could just as well

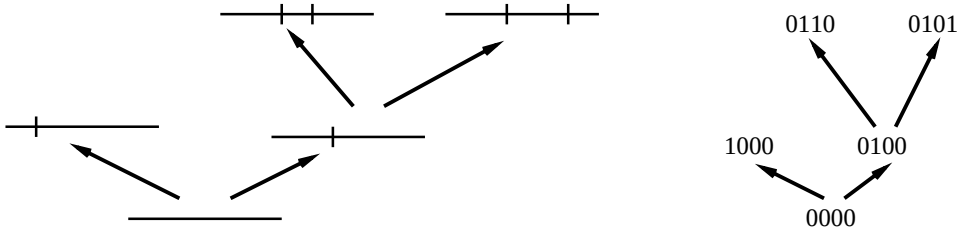


Figure 10.1: An example of haplotypes produced by mutation in the infinite-sites model. On the left are the alleles, with the arrows indicating the events and the locations of mutation indicated by vertical marks. On the right is the same case with the states at each site indicated by 0 and 1, and only the sites that actually mutate shown.

have been been (in the order in which the haplotypes are shown across the Figure) 0010, 0000, 1000, 1001, and 1010.

One of the consequences of the infinite-sites model is that each mutational event gives rise to at most one additional haplotype. "At most", because some of the haplotypes can be lost by genetic drift. For any given pair of sites, there can be only three combinations of states. Thus, in the Figure, sites 1 and 4 show three different combinations: 10, 00, and 01. Combination 11 could arise for these two sites only by recombination, as recurrent mutation is not allowed in the model. Since recombination is assumed to not be present within a locus, no more than three combinations can be present at two sites. It can be shown that if all pairs of sites pass the three-state test, the set of haplotypes can have arisen by mutation in an infinite-sites model.

We can test whether a set of sequences could possibly result from an mutation in an infinite-sites model by asking, for each pair of sites, how many states are present. If some DNA haplotypes that have no more than two different bases present at each site have, at sites 12 and 18, states CA, CG, AG, and GG, then they fail the test. This three-state test was introduced by Hudson (2001); it is a version of a test introduced earlier in systematics by E. O. Wilson (1965). Strobeck and Morgan (1978) have shown that in a model with multiple sites, intragenic recombination may have a substantial effect in generating new sequences in addition to the effect of mutation.

Before showing how models such as these may be used in molecular population genetics, we need to introduce the standard model for genealogies of gene copies, the coalescent.

X.4 The Coalescent

The presence of possibilities of inferring the history of mutations means that we can to some extent see the ancestry of the sequences in a population sample. If we have a segment of DNA that does not undergo recombination very often (and we will see how often this is), then the ancestry of the sequences in a sample forms a tree. It is not a phylogenetic tree, because the sequences at the tips of its branches are not from different species. We have seen, in chapter VI, that a pair of gene copies in an isolated random-mating population will be identical by descent an average of $2N_e$ generations ago. This classic result of Sewall Wright's has more recently been extended in an important result by J. F. C. Kingman (1982a, 1982b, 1982c). It characterizes the tree of ancestry of gene copies in such an isolated population.

Kingman's result is an approximation, but a very accurate one in most cases. Consider the lineages of ancestry from n copies of a gene, in a population of size N, where $n \ll N$. Each copy will have such a lineage, extending back to a parent, a grandparent, a greatgrandparent, and beyond. It is important to realize that we are talking about the lineage of copies which are traced back to previous copies, not individuals traced back to previous individuals. Figure 10.2 shows three such lineages of gene copies in a population of 12 individuals. Each individual has two immediate ancestors, but each gene copy has only one (for example your maternal copy of the Hemoglobin β locus may come from your mother's father, and specifically from his paternal copy, which comes in turn from his father's maternal copy, and so on.

The figure shows a correct simulation of a population that reproduces according to the Wright-Fisher model, and the ancestry of three copies of the gene. These lineages combine as one goes backwards, until there is only a single lineage. This process of the merging of lineages is called a *coalescent*, a term introduced by Kingman.

In the Wright-Fisher model each lineage in effect "chooses" its immediate ancestor, both in terms of choosing the parent individual and choosing the gene copy within that individual. There are $2N$ copies of the gene to choose from. We may ask what is the probability that, in a given generation, two copies "choose" the same parent copies. If there are n copies and the first two of them happen to be the ones that choose the same parent copy, the probability of this happening is

$$\frac{1}{2N}\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)\ldots\left(1-\frac{n-2}{2N}\right) \qquad \text{(X-15)}$$

because the first copy chooses some parent copy or other (with probability 1), the second independently happens to choose the same copy (with probability $1/(2N)$), the third chooses a parent copy different from that one (with probability $1-1/(2N)$), the fourth chooses a parent copy different from those two copies, and so on. This is only one of the $n(n-1)/2$ possible pairs of copies that could have the same parent. These events,

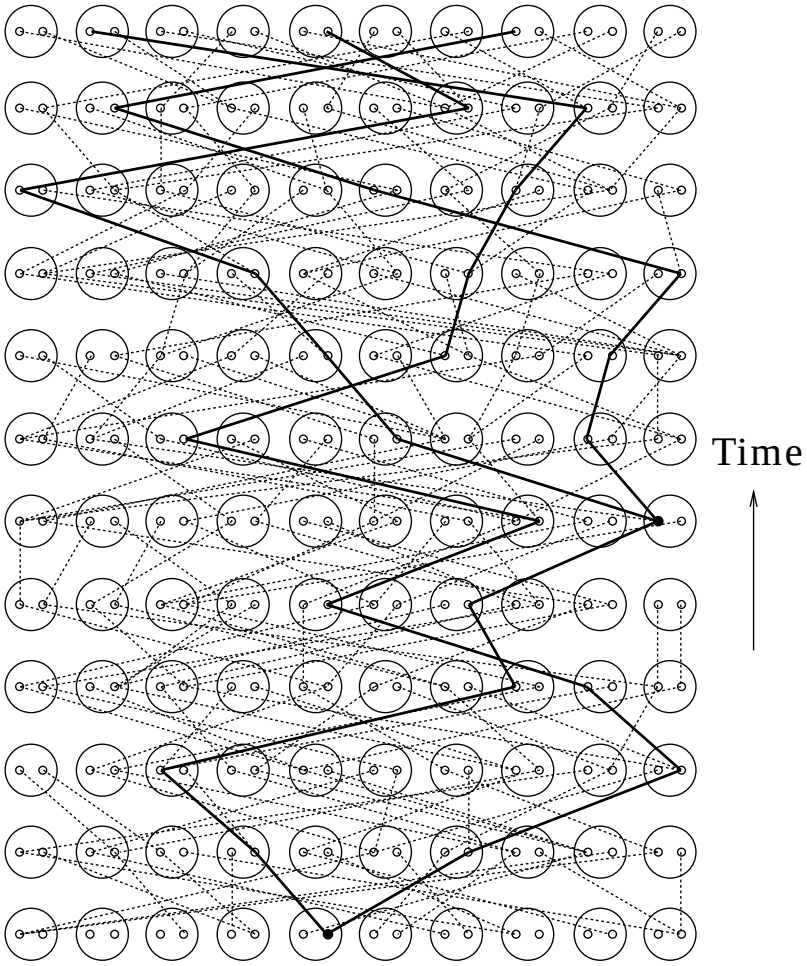


Figure 10.2: The coalescence of lineages in population reproducing according to a Wright-Fisher model. The dark lines show the genealogy at the gene level of a sample of three copies of the gene.

which each have exactly two individuals with the same parent, are mutually exclusive, so we can add their probabilities. Collecting together terms in $1/N$ and $1/N^2$ and so on, we get for the probability that exactly two lineages coalesce in this generation

$$\frac{n(n-1)}{2}\,\frac{1}{2N} \;+\; \text{terms in } \frac{1}{N^2} \tag{X-16}$$

A similar argument yields the probability that three lineages happen to coalesce in the same generation as

$$\frac{n(n-1)(n-2)}{6}\,\frac{1}{4N^2} \;+\; \text{terms in } \frac{1}{N^3} \tag{X-17}$$

This will be considerably smaller than the probability (X-16) of pairwise coalescence if $(n-2)/(6N)$ is considerably less than 1. Probabilities of coalescence of more than three lineages in the same generation are even smaller. Hence when $n \ll N$ we need only consider pairwise coalescences.

A good approximation to the process is that in each generation there is a small probability $n(n-1)/(4N)$ of coalescence. If there is a coalescence, it is of two lineages. Which two? The answer should be obvious – a random pair of lineages. We have a process that goes back in time, generation by generation, having a constant small probability of a coalescence. If we ask what the distribution of the time (going backwards) until coalescence is, it is the distribution of the time to the first "heads" in a series of coin tosses with a small probability of "heads". The average number of tosses will be the reciprocal of the heads probability, $4N/(n(n-1))$.

Technically the distribution is a geometric distribution, but it is excellently approximated by an exponential distribution with the same mean. An exponential distribution is the distribution of the time until a random event that can occur at any point in a continuous time, such as the time until the next radioactive decay detected by a Geiger counter. Making that approximation, we get a process which goes back an exponentially distributed number of generations, with mean $4N/(n(n-1))$, then coalesces two random lineages. At that point there are now $n-1$ lineages. It is obvious that the process now continues, but with $n-1$ replacing n throughout the mathematics. So the next coalescent event, proceeding backwards, occurs after a time that is exponentially distributed with mean $4N/((n-1)(n-2))$, and involves two random lineages. This continues until there are only two lineages left, and these coalesce after a time that is exponentially distributed with mean time $2N$ generations, which is precisely the time predicted in Sewall Wright's original work.

Figure 10.3 shows such a coalescent. The time u_k is exponentially distributed with mean $4N/(k(k-1))$.

How long will it take for the sample of n genes to coalesce to one copy? We can get an idea by adding up the mean times. Noting that $1/(k(k-1)) = 1/(k-1) - 1/k$, we find that

$$\begin{aligned}
&\frac{4N}{n(n-1)} + \frac{4N}{(n-1)(n-2)} + \frac{4N}{(n-2)(n-3)} + \cdots + \frac{4N}{2} \\
&= 4N\left(\frac{1}{n-1} - \frac{1}{n} + \frac{1}{n-2} - \frac{1}{n-1} + \frac{1}{n-3} - \frac{1}{n-2} + \cdots + \frac{1}{1} - \frac{1}{2}\right) \qquad \text{(X-18)} \\
&= 4N\left(1 - \frac{1}{n}\right).
\end{aligned}$$

The expected time for a whole population to coalesce is thus nearly $4N$ generations.

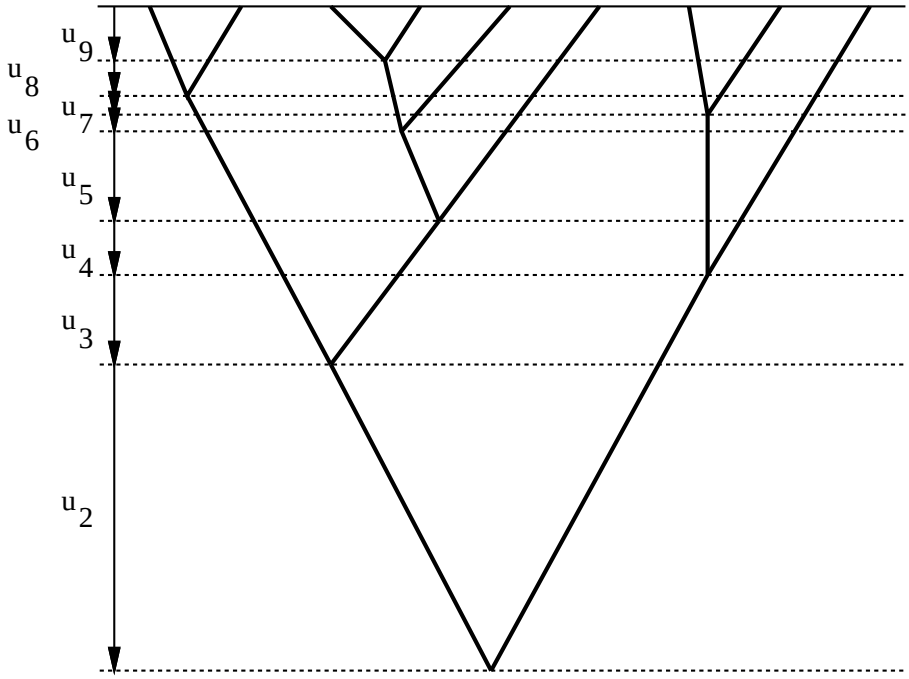


Figure 10.3: Kingman's n-coalescent process, an approximation to the genealogy of gene copies in an isolated random-mating population

A similar process applies to mitochondrial genes, which are effectively haploid and are only descended from the females in the population. One simply replaces $4N$ by $2N_f$, so that the time to coalescence is about 4 times less.

THE APPROXIMATION. The accuracy of the coalescent as an approximation to the genealogical tree of gene copies in a Wright-Fisher model has been investigated by Fu (2006). The approximation, which is quite accurate, becomes more so the smaller n is compared to N. Technically it is a diffusion approximation. If we take $N \to \infty$, the time for n lineages to coalesce to fewer than n does not approach a limit, because it has mean $4N/(n(n-1))$, which becomes infinite. But if we make the same time scale change that a diffusion approximation does (as we saw in Chapter VII), there is a limit: if time is measured in units of N generations, then the distribution approaches an exponential distribution with mean $4/(n(n-1))$ of these units. The distribution then converges. Note that this same limit also guarantees that with probability 1, the coalescence involves two lineages and not more.

This change of time scales is often ignored. It is often good enough for practical purposes to say that the coalescence time is exponentially distributed with mean $4N/(n(n-1))$, as I have done above. Many other complications in the breeding sys-

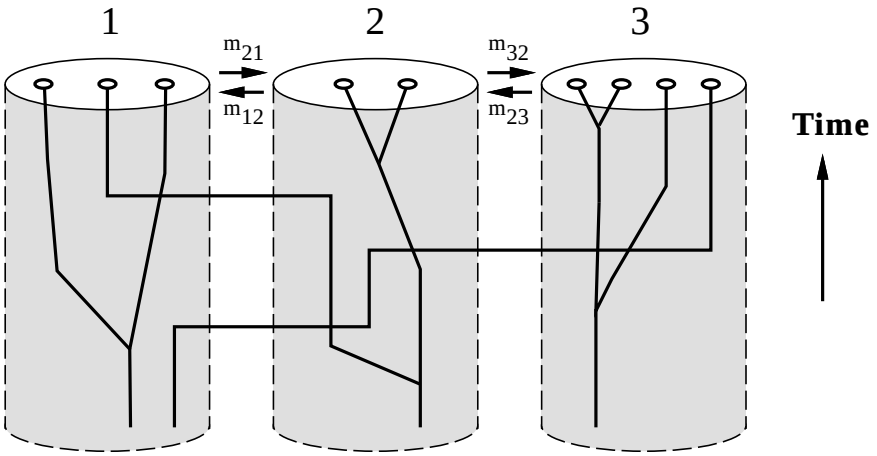


Figure 10.4: Migration and coalescence in three populations. Looking backward, the events are, in order, coalescence in population 3, coalescence in population 2, migration to population 1 from population 2, migration to population 3 from population 2, coalescence in population 3, migration to population 2 from population 1, coalescence in population 1, and coalescence in population 2.

tem can be handled by simply replacing the actual population size N with the effective population size N_e.

X.5 Coalescents with migration

If there are two or more populations, with constant rates of migration m_{ij} from population j into population i, the coalescent distribution of genealogies is easy to obtain. As we go back in time, with n_i lineages in population i, whose size is N_i, there is a constant rate $n_1(n_1-1)/(4N_1)$ of coalescence in population 1, $n_2(n_2-1)/(4N_2)$ in population 2, and so on for all the populations. There is also a constant rate m_{ij}, for each lineage in population i, of events in which it proves to be newly arrived from population j by a migration event at that time. Figure 10.4 shows a series of events in three populations in which there is migration and coalescence.

We can use this figure to show how to draw a genealogy in a coalescent with three populations. We start at the top. There are 3 lineages in population #1, 2 in population #2, and 4 in population #3. In population #1, the 3 lineages have a rate $3 \times 2/(4N)$ of coalescence. Each of these 3 also has a rate m_{12} of migration events (we are looking backwards). Likewise, in population #2, the rate of coalescence is $2/(4N)$ and the total

rate of migration events is $2m_{21} + 2m_{23}$. In population #3, the rate of coalescence is $4 \times 3/(4N)$ and the rate of migrations is $4m_{32}$.

We don't draw genealogies separately for the populations. Instead we take the total rate of occurrence of events:

$$\frac{6}{4N} + \frac{2}{4N} + \frac{12}{4N} + 3\,m_{12} + 2\,m_{21} + 2\,m_{23} + 4\,m_{32} \qquad \text{(X-19)}$$

consider the time back to the next (i.e., previous) event to be drawn from an exponential distribution with this rate of events. When that time is drawn, we then must decide which event happened. It is like sitting waiting for a telephone call, when the rate of events is 0.02 business calls per minute and 0.01 personal calls per minute. Their total is 0.03, so we will wait an exponentially distributed length of time with mean 1/0.03 = 33.333 minutes. When a call occurs, it has probability 0.02/0.03 = 2/3 of being a business call, and 0.01/0.03 = 1/3 of being a personal call.

So if the total rate of events is $5/N + 3\,m_{12} + 2\,m_{21} + 2\,m_{23} + 4\,m_{32}$, the probability that the event is a coalescence in population #1 is $6/(4N)$ divided by this, and so on. We choose among the events in proportion to their rates of occurrence. Having chosen an event (in the case of the genealogy shown in the Figure it was a coalescence in population #3), we change the genealogy by carrying out the coalescence. If it had been a migration we would move the lineage into the appropriate population. Now there are 3, 2, and 3 lineages in the three populations. We recalculate the rates of events accordingly, and also recalculate their total rate. Then once again we draw the time back to the next (previous) event, and again draw what kind of event it is. This continues until the last remaining lineages coalesce and there is only one lineage.

Coalescents with migration are easy to sample in this way. Their properties are less easy to derive mathematically. Takahata and Slatkin (1990) could derive the mean time to coalescence for two lineages, one from each of two populations, but they found no simple form for the density function of the time to coalescence. It would have to be a mixture of sums of different numbers of exponential densities, depending on how many migration events occurred on the way back to the coalescence. Wakeley and Lessard (2006) have found that if there are low rates of migration, the genealogy of samples from multiple populations involve two limiting processes, coalescence within populations and a longer-term coalescence between lineages ancestral to each population.

X.6 Coalescents with population growth

If there is only a single population, but it is changing size, the coalescent is also complicated. The rate of coalescence is no longer constant as we go back in time. Instead, if the population size is $N(t)$ when we have gone back t units of time, the rate of coalescence

at that moment, for k lineages, is $k(k-1)/(4N(t))$. If we pass through a population bottleneck where $N(t)$ is small, there will be a higher rate of coalescence in that period.

The simplest way of drawing a genealogy is to imagine that, when population size is small, it is as if the time clock is running faster, so that there is more opportunity for coalescence. Following that line of argument, the total opportunity for coalescence, back to actual time t, can be calculated and equated to a fictional time τ that would give the same opportunity for coalescence in a population that had constant size:

$$\frac{k(k-1)\,\tau}{4N(0)} \;=\; \int_0^t \frac{k(k-1)\,u}{4N(u)}\,du \tag{X-20}$$

where the variable of integration, u is time back (not mutation rate). What this does is to add up the total rate of coalescence back to time t, and allow us to calculate τ, a number of generations in a population of constant size that would give the same amount of coalescence. Of course the $k(k-1)$ and the 4 can be eliminated.

For a population that is growing exponentially at rate g as we go forward in time,

$$N(t) \;=\; N(0)\;\exp(-g\,t) \tag{X-21}$$

in which case we can use (X-20) to calculate that

$$\tau \;=\; \frac{1}{g}\left(e^{gt}-1\right) \tag{X-22}$$

To draw a coalescent genealogy, we use the constant population size $N(0)$ and draw a genealogy. On going back to a coalescence at time T, we consider this as the fictional time τ and solve for what the real time would have been, using equation X-22 and solving for t in terms of τ. A somewhat more extended version of this derivation is given in my book on phylogenies (Felsenstein, 2004, pp. 460-461).

This false-time-scale argument was developed by Kingman (1982c) and is also described by Slatkin and Hudson (1991). They pointed out that with exponential population growth in which $4N(0)g$ is substantial, the shape of the coalescent tree becomes closer and closer to a "star" tree which has all of its splitting near the base.

RECOMBINATION. In the arguments so far, there has been no recombination within a locus. Each gene copy was descended from a single gene copy in the previous generation. Suppose that there was a recombination between sites 221 and 222 in a locus, but no others occurred all the way back to coalescence. The front end of the gene, the first 221 sites, have an ordinary coalescent. For the rest, we follow the same genealogy, except for the lineage leading back to the recombination event. When we get back to that event, the rear end of the gene, say sites 222 to 1000, come from a different copy in the immediate parent. Following their ancestry back one has a different coalescent lineage. It goes back and ultimately coalesces with one of the other lineages.

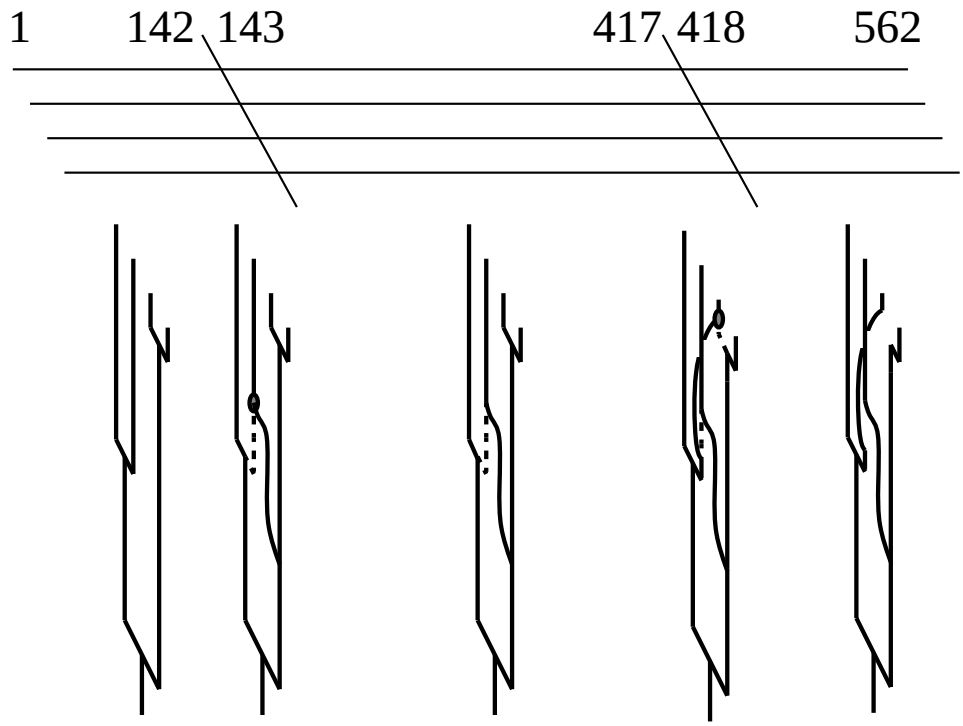


Figure 10.5: Three coalescent trees generated by two recombination events along a chromosome during the ancestry of four haplotypes. The recombination events are the gray disks. Note that one lineage becomes a "ghost" lineage after the first recombination, but its bottom part and a coalescent event are restored to relevance after a new lineage coalesces with it following the second recombination.

Thus there is an ordinary coalescent for sites 1 to 221, and another one for sites 222-1000, which differs by having one lineage unhooked and then allowed to coalesce elsewhere. Figure 10.5 shows the result of two recombination events. As one moves along the chromosome, one has one coalescent, then another, and then a third. Note that some coalescent events that were in the first coalescent and not in the second show up again in the third.

As one moves long the chromosome and passes points where, in the coalescent ancestry, there was a recombination, the genealogical tree gradually changes. Moving far enough along the chromosome, the tree becomes very different.

TREES AND D'S. How far along the chromosome is enough for this? We can imagine two sites far enough apart that they have recombination fraction r, in a population of size N. At each site there is a straightforward coalescent – the question is whether these

are the same. Following a single lineage down to the root of the coalescent is about $4N$ generations. The expected number of recombination events between these sites in $4N$ generations is $4Nr$. When this number exceeds 1, we expect the two sites to become separated on most lineages before either coalesces. It turns out that $4Nr > 1$ defines the amount of recombination at which sites have substantially different coalescent trees.

Recall from Chapter VIII that this is also the condition for two sites to have substantial linkage disequilibrium generated by genetic drift. This is not an accident. When sites are in linkage disequilibrium, it is because they share genetic drift events in their ancestry, because they trace back to the same ancestors. Shared coalescent trees and noticeable D's indicate the same associations.

But how far is this? Recall that in the example in Chapter VIII, if we have an organism with one recombination every 100 million bases in each generation, and an effective population size of 100,000, the distance along the chromosome at which $4Nr = 1$ is when $r = 1/(4N)$, which is a mere 250 bases. However, if recombination has "hot spots", the regions between those hot spots will be longer, and the recombinations will be clustered in the hot spots, where the tree will change rapidly.

The upshot of all this is that no one locus shows us the tree of ancestry of the species. Instead each region of the genome has a different tree, going back to many different coalescent events. There may be a mitochondrial Eve, but there is a Y-chromosome Adam (who did not know Eve and lived at a different time). There is also a Cytochrome Sam and a Hemoglobin Helen, and many others. If trees change, say, every 10,000 bases, your ancestry involves about 320,000 different trees. If you happen to discover the tree for one region of the genome, you ought to think twice (or perhaps 320,000 times), before claiming that it shows "the ancestry" of the species.

THE ANCESTRAL RECOMBINATION GRAPH. Taking all the trees for a set of haplotypes, we can superimpose them and make a graph showing recombination events. Each recombination event is a fork splitting downwards. Below each event, we indicate which sites have ancestry along each lineage. Figure 10.6 shows this for the trees from Figure 10.5. It is possible to work back through time, drawing the ancestral recombination graph. The process is very similar to that used for coalescents with recombination. You need to know, at each time in the past, what are the possible events that can occur and their rates (the probability of occurrence per unit time). For example, in the graph in Figure 10.6, as we work backward in time, down the graph, the events are

1. a recombination separating sites 417 and 418,
2. a coalescence of lineages,
3. a recombination separating sites 142 and 143,
4. a coalescence of lineages,
5. a coalescence of lineages, and
6. a coalescence of the remaining two lineages.

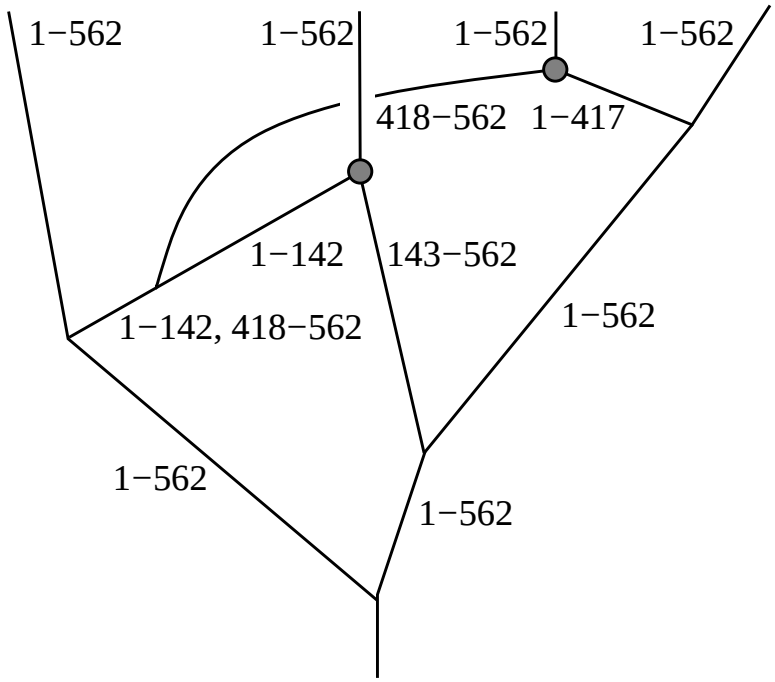


Figure 10.6: The ancestral recombination graph for the case shown in Figure 10.5. Each lineage lists the sites that have their ancestry along that lineage.

If we have generated this graph down as far as event number 4, at that point we have 4 lineages. The first has all sites 1-562, the second has sites 1-142 and 418-562, the third has sites 143-417, and the fourth has all sites 1-562. There is then a total rate $4 \times 3/(4N)$ of coalescence of lineages. There is also a rate of recombination. In the first lineage there are 561 intervals between bases at which recombination can happen. The second may look as if it can have recombination in any of (142-1)+(562-418) places. But not so: a recombination between sites 142 and 418 does have an observable consequence, so we have to count them too. There are actually 561 places where there could be a recombination in that lineage. The third lineage has 417-143 = 274 places, and the fourth has 561 places. Thus the total rate of recombination is the rate r of recombination per base multiplied by 561+561+274+561 = 1,957.

To generate the next event down the recombination graph, we need only draw an interval of time from an exponential distribution whose mean is $1/(3/N + 1957r)$. Then we need further random numbers to determine whether the event is a coalescence (as it was on the graph in the Figure) or a recombination. For each of those we need to choose which lineages coalesce, or which interval of the 1957 suffered the recombination.

Ancestral recombination graphs were first discussed by Hudson (1983). He also pro-

duced a computer program that has been the basis for most subsequent programs simulating these graphs. A more detailed mathematical treatment will be found in the paper by Griffiths and Marjoram (1997).

FURTHER READING ON COALESCENTS. My book on phylogenies (Felsenstein, 2004, Chapters 26-28) may be consulted for a more detailed description of the coalescent. Hein, Schierup, and Wiuf (2004) have written the first book devoted entirely to the coalescent. It contains many clear and illuminating descriptions of the population genetic theory involved. A more recent book by Wakeley (2008) is a strong competitor, with particularly clear explanations and examples.

X.7 Some summary statistics

In the first years after DNA sequences became available, the most widely used methods of estimating population parameters such as $4N$ and the neutral mutation rate μ were to compute summary statistics. We can use what we know about the coalescent to simplify these arguments. Generally these estimators are examples of the Method of Moments. This approach computes the expectation of the statistic in terms of our parameters, equate it to the observed value, and solve for the parameters. Although tempting because it is often easy to do, this approach is usually lacking in statistical power.

NUCLEOTIDE DIVERSITY. Kimura (1968b) introduced the nucleotide diversity π, defined as the average number of differences per site between pairs of sequences drawn from a sample. The expectation of this quantity is easy to compute under neutral models of substitution. As it is the average of all pairs of sequences, its expectation is the expectation of any pair, say the first two sequences. If the two sequences coalesce t generations ago, the probability density for t will be the usual coalescent density

$$f(t) \;=\; \frac{1}{2N}\, e^{-\frac{t}{2N}} \tag{X-23}$$

For a Jukes-Cantor model, we can simply average the fraction of sites different in a total time of $2t$ generations (down t and back up another t), weighting by this density function:

$$\mathbb{E}[\pi] \;=\; \int_0^\infty \frac{1}{2N}\, e^{-\frac{t}{2N}}\, \frac{3}{4}\left(1 \;-\; e^{-\frac{4}{3}\mu(2t)}\right) dt \tag{X-24}$$

Collecting terms this becomes

$$\mathbb{E}[\pi] \;=\; \int_0^\infty e^{\frac{3}{8N} - t}\left(\frac{1}{2N}\right) dt \;-\; \int_0^\infty e^{\frac{3}{8N} - t}\left(\frac{1}{2N} \;+\; \frac{8}{3}\mu\right) dt \tag{X-25}$$

which is easily evaluated and turns out to be

$$\mathbb{E}[\pi] \;=\; \frac{3}{4} \;-\; \frac{3}{8N}\;\frac{1}{\frac{1}{2N}+\frac{8}{3}\mu}, \tag{X-26}$$

and this in turn works out to be

$$\mathbb{E}[\pi] \;=\; \frac{\Theta}{1 \;+\; \frac{4}{3}\Theta} \tag{X-27}$$

where $\Theta = 4N\mu$. Thus for small values of Θ we can simply estimate it as π, while for slightly larger values we can use:

$$\hat{\Theta} \;=\; \frac{\pi}{1 \;-\; \frac{4}{3}\pi} \tag{X-28}$$

Note that what we can estimate is not either μ or N, but instead their product $N\mu$, here used in the more natural form of $4N\mu$. This can be seen by thinking about the coalescent genealogy. If we double N, the coalescent gets twice as deep. If at the same time we halve μ, we expect half as many mutations per site per unit time, and so we expect the same total number of mutations to be visible. Since what we see in a contemporary sample is the pattern of differences caused by mutation, these two cases will be indistinguishable. The product $N\mu$ is thus what we can actually infer.

NUMBER OF SEGREGATING SITES. Watterson (1975) introduced this statistic, using the infinite-sites mutation model which he introduced in that paper. One takes the sample of n sequences and counts at how many sites there is variation. Consider the sequences to be mutating according to the infinite-sites model, so that each mutation is a different site. What is the expectation of the quantity S, the number of segregating sites? All mutations can be seen, as none ever obscure each other or reverse each other. So the expected number of segregating sites is the expected number of mutations on the coalescent tree before all lineages coalesce into one.

Suppose that the total mutation rate along the sequence is U. We can compute the expected numbers of mutations in each of the $n-1$ coalescence intervals. In the one which has k lineages and coalesces to $k-1$, the expected time is $4N/(k(k-1))$ and there are k lineages having that length, so the expected total tree length in this interval is $4N/(k-1)$. The expected number of mutations in that interval is U times this, or $4NU/(k-1)$. Adding up over all intervals, from n lineages down to the bottommost one that has 2 lineages:

$$\mathbb{E}[S] \;=\; 4NU\left(\frac{1}{n-1}+\frac{1}{n-2}+\cdots+\frac{1}{3}+\frac{1}{2}+\frac{1}{1}\right). \tag{X-29}$$

To make a method-of-moments estimator, we simply divide S by the quantity in parentheses to get a quantity that has expectation $\theta = 4NU$:

$$\widehat{\theta} \;=\; S \Big/ \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n-1}\right), \tag{X-30}$$

It has become conventional in population genetics to have θ be computed as the product of $4N$ and the total mutation rate of the sequence. Note that to make it into $4N\mu$, where μ is the mutation rate per site, it is necessary to divide by the number of sites.

As an example, consider a set of sequences that evolved in a population where the per-site value $\Theta = 0.003$. In a simulated sample of 10 sequences of 500 sites evolving according to a Jukes-Cantor model, I found 6 sites varying, with a mean pairwise difference of 0.008533. Using (X-28) the estimate of Θ is 0.00863. Using Watterson's estimator we would get $\theta = 6/(1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6 + 1/7 + 1/8 + 1/9) = 2.1209$. However is this per-locus, so we would get an estimate of $\Theta = 0.004242$ once we divide by 500. Both of these are higher than the true value, with the Watterson estimator being closer. They are not identical, showing that they respond to somewhat different aspects of the data. As Watterson's estimator is derived from a model that does not allow multiple "hits" at one site, we would expect it to be, on average, a bit low. However, that is unlikely to be an important effect for values of Θ this small.

Interestingly, it is possible to argue in a similar fashion that the estimator of Θ should be the same, even if the sequences are known to undergo recombination. At each base, the probability that the site is segregating is as given above, and since S is the sum over sites, its expectation is not affected if the sequence has recombination causing different sites to have different coalescent genealogies.

TAJIMA'S TEST. Tajima (1989) uses these two estimates of Θ to test the neutral mutation theory. Taking the difference between the nucleotide diversity and Watterson's estimator, he uses formulas for variances of these estimators, and derives one for their covariance. He is then able to compute a standard deviation for their difference, and divide by this. If we have deleterious mutants at low frequencies in the locus, these would be expected to increase the number of segregating sites without having much impact on the mean number of differences between sequences. This will cause his statistic to become negative. He argues that balancing selection on some sites at this locus would make the statistic tend to be positive.

Tajima's test is fairly widely used. It is fairly robust and simple, though it does require that we have only a single population.

THE SITE-FREQUENCY SPECTRUM. Neutral mutation and genetic drift predict the distribution of gene frequencies in an isolated random-mating population, and this can be used to predict the distribution of outcomes in a sample of n copies from the population. Usually this is done using the infinite-alleles model and the elegant theorems

by Ewens (1972), one of the high-water marks of theoretical population genetics. This distribution of sample outcomes can then be compared to SNPs found in samples.

For simplicity, and to show that the result holds more broadly, I will derive it for a single site that evolves according to a Jukes-Cantor model of mutation. One of the convenient properties of that model is that we can easily derive the distribution of population gene frequency of one of the four bases, say C. In a Jukes-Cantor model with four bases, we can simply lump the other three bases, A, G, and T and consider then to be a single allele, not-C. For those two alleles, we then see that C has probability μ of mutating to not-C, and not-C has probability $\mu/3$ of mutating to C. The resulting distribution is then available from equation VII-119. In fact it is exactly the distributions shown in Figure 7.5, where the probability reverse mutation is $1/3$ of the probability of forward mutation. The density function will be

$$\phi(p) \;=\; K\, p^{4N\mu-1}(1-p)^{\frac{4}{3}N\mu-1} \tag{X-31}$$

where p is the frequency of the C allele (and using N for the effective population size).

When we take a sample from a population that has gene frequency p, the probability of finding C a total of i times in a sample of n copies will of course simply be the binomial sampling probability

$$\text{Prob}\;(i \mid n, p) \;=\; \frac{n!}{i!(n-i)!}\, p^i (1-p)^{n-i} \tag{X-32}$$

To get the overall probability of i copies, averaged over the distribution of p, we have to multiply these two and integrate over all possible values of p. Moving the constants outside the integral, this is

$$\text{Prob}\;(i \mid n) \;=\; K\,\frac{n!}{i!(n-i)!}\int_0^1 p^{4N\mu-1}(1-p)^{\frac{4}{3}N\mu-1}\, p^i (1-p)^{n-i}\, dp \tag{X-33}$$

The integral can be rewritten by collecting terms in p and in $1-p$:

$$\text{Prob}\;(i \mid n) \;=\; K\,\frac{n!}{i!(n-i)!}\int_0^1 p^{i+4N\mu-1}\,(1-p)^{n-i+\frac{4}{3}N\mu-1}\, dp. \tag{X-34}$$

The integral is well-known, from the theory of Beta distributions, to be

$$\frac{\Gamma(i+4N\mu)\,\Gamma(n-i+\frac{4}{3}N\mu)}{\Gamma(n+\frac{16}{3}N\mu)}. \tag{X-35}$$

The result is

$$\text{Prob}\;(i \mid n) \;=\; K\,\frac{n!}{i!(n-i)!}\,\frac{\Gamma(i+4N\mu)\,\Gamma(n-i+\frac{4}{3}N\mu)}{\Gamma(n+\frac{16}{3}N\mu)}. \tag{X-36}$$

We are interested only in the relative values for different numbers of copies of allele C, and we will rescale these to add to 1. So we can aggregate all the terms that lack i into a constant K, and then we have

$$\text{Prob } (i \mid n) \;=\; K \frac{\Gamma(i + 4N\mu)\,\Gamma(n - i + \frac{4}{3}N\mu)}{i!(n-i)!} \tag{X-37}$$

That is fairly simple, but we can do better if we allow μ to be small. In eukaryotes the mutation rate per base is in fact about 10^{-8}, so this should be reasonable. As μ approaches 0, the numerator terms become $\Gamma(i)\Gamma(n-i)$. For integers, it is a well-known property of the Gamma function that $\Gamma(k) = (k-1)!$ The factorials then cancel appropriately so that we finally get

$$\text{Prob } (i \mid n) \;=\; \frac{K}{i(n-i)}. \tag{X-38}$$

which is the result found by Ewens (1972; see also his 2004 book, section 3.6.2). This simple formula allows us to predict this distribution, the *site frequency spectrum* or SFS. Computing it is simple: after the quantities $1/(i(n-i))$ are computed, we simply sum them and divide them all by that sum so that they add to 1. We can infer that a similar result will hold for other, more realistic models of mutation; although this has not been formally proven, you won't get me to bet against it.

As SNP chips have made it possible to screen populations for large numbers of SNP loci in ever-larger samples, the SFS has become more widely used. In most cases there are only two alleles at each SNP locus, and in most of thoses cases we cannot tell which allele arose most recently. It is usually necessary to use the frequency of the rarer allele, so that the observed allele frequencies are less than or equal to 0.5. If our data are summarized in this way, we need to fold the expected distribution around 0.5, so that there are classes only for values of i that do not exceed $n/2$. This doubles the height of every class in our histogram, except for the 0.5 class if the sample size is even.

As beautiful as this theory is, its application has some limits. Technically it is for a single population which has reached a stationary state under mutation and genetic drift. Thus it does not predict the SFS (or the joint SFS) for two or more populations that are connected by migration. The assumption of a stationary state also makes the SFS incorrect for populations that have been changing size, as is certainly the case for most human populations. Evans, Shvets, and Slatkin (2007) have given differential equations for the change of the gene frequency spectrum toward its equilibrium. Another interesting attempt to use a combination of simulation methods and diffusion-equations analysis to cope with these complications has been made by Gutenkunst et al. (2009). Progress has also been made on developing a theory of joint SFS in two recently-diverged populations (Chen et al., 2007). But for one population, Myers, Fefferman, and Patterson (2008) have shown that information about historical population sizes is absent from the site frequency spectrum.

The major limitation of the SFS is that it ignores linkage disequilibrium between sites, particularly nearby sites. There is information in the joint patterns of occurrence of SNP alleles at nearby loci, and all of that information is lost when we examine loci one at a time. To retain all information we need full likelihood or Bayesian methods, which we now will discuss.

X.8 Likelihood calculations

Summary statistics are simple and robust, but are not necessarily efficient. To make an efficient estimation of Θ we need to ask whether it is possible to compute the likelihood for a sample of sequences. For most of the 1980s no one seems to have even posed this question (but see Strobeck, 1984). It is perhaps not surprising that little progress was made, that summary statistics methods were in use, but is astonishing that no one even pointed out that likelihood-based estimators were of potential interest.

As we shall see, likelihood (and Bayesian) methods are difficult, but since the work of Griffiths (1989), Griffiths and Tavaré (1994), and Kuhner, Yamato, and Felsenstein (1995) they have become practical. They involve sampling a large number of possible genealogies from the huge set of all possible genealogies that could connect the sequences in the sample. Although they have been slow to be adopted, these methods are the future of data analysis for sequences sampled within species.

THE LIKELIHOOD. Computing the likelihood under the neutral mutation model is almost impossible to consider unless we can think of the genealogy underlying the sample. If we knew the genealogy G^* that connects the members of the sample, we could use it to estimate the effective population size N_e (Felsenstein, 1992a). However, this genealogy is not available to us: it is only hinted at by the sequences, especially if the number of segregating sites is small, as it often is. For the case of an infinite-sites model, Griffiths (1989) proposed to compute the likelihood by a recursion which sums over all possible genealogies. This is possible because, in the infinite-sites model, each site defines a partitioning of the sequences which reflects a feature of the true coalescent tree. Griffiths' recursion method, while an intellectual breakthrough, was not practical for more samples of more than a few sequences.

In more realistic models of sequence evolution, parallel and reverse mutations allow any possible sequence sample to have some nonzero probability of occurrence, no matter what the true coalescent. This rules out use of exact recursions. As we will see, the number of possible genealogies is so great that, unless there is a remarkable breakthrough yielding a formula that sums over genealogies, only sampling methods have any chance to yield usable likelihood or Bayesian methods.

SUMMING OVER TREES. To give a simple idea of the logic involved, let's consider first the simple case of two sequences. As above, we are interested in the case of estimation

of N and μ in a single random-mating population which has maintained its current size for a long time. With nonrecombining sequences, the coalescent tree is simply two sequences coalescing, with the coalescence time drawn from an exponential distribution with mean $2N$.

The likelihood is the probability of the two sequences, summed (integrated, in this case) over all possible coalescence times, with each term weighted by the probability of that coalescence time:

$$L \;=\; \text{Prob}(D \mid N, \mu) \;=\; \int_0^\infty \text{Prob}(t \mid N)\ \text{Prob}(D \mid t, \mu)\ dt \qquad \text{(X-39)}$$

The term $\text{Prob}(t \mid N)$ is the density function of the exponential distribution with mean $2N$. The other term, $\text{Prob}(D \mid t, \mu)$ is familiar in phylogenetic inference – it is known there as the likelihood for this tree. Details of its calculation may be found in my recent book (Felsenstein, 2004). On larger trees, it can be efficiently calculated by a "pruning" algorithm that calculates conditional likelihoods recursively down the tree.

For our purposes, we only need to note that since branch lengths in transition probability formulas for DNA models occur only as products like $\mu\, t$ with the mutation rates,

$$\text{Prob}(D \mid t, \mu) \;=\; \text{Prob}(D \mid \mu t, 1), \qquad \text{(X-40)}$$

which simply means that, if the mutation rate is μ, the probability of the outcome in t units of time is the same as it would be if the mutation rate were 1, but only μt units of time had elapsed. So if the mutation rate is 10^{-9} per site, the probability when 10^8 generations has elapsed is the same as if the mutation rate were 1 per site and 0.1 generations had elapsed.

Putting this into equation X-39 together with the density function of the two-sequence coalescent, the likelihood becomes

$$L \;=\; \text{Prob}(D \mid N, \mu) \;=\; \int_0^\infty \frac{1}{2N}\, e^{-\frac{t}{2N}}\ \text{Prob}(D \mid \mu\, t, 1)\ dt. \qquad \text{(X-41)}$$

Changing variables to a new time scale u which is measured in expected numbers of changes per site, so that $\mu\, t = u$ we replace t by u/μ and dt by $(1/\mu)\, du$. This leaves us with

$$L \;=\; \text{Prob}(D \mid N, \mu) \;=\; \int_0^\infty \frac{1}{2N\mu}\, e^{-\frac{u}{2N\mu}}\ \text{Prob}(D \mid u, 1)\ du. \qquad \text{(X-42)}$$

The essential point about this is that the likelihood turns out to be a function, not of the quantities N and μ separately, but only of their product $N\mu$.

Figure 10.7 shows the Kingman distributions of divergence time μt for three different values of Θ and the probability of getting the two sequences under a Jukes-Cantor model, when they are 1000 bases long and differ at 5 sites. This is one of the few cases simple enough to integrate the product of the Kingman densities and the probability of the

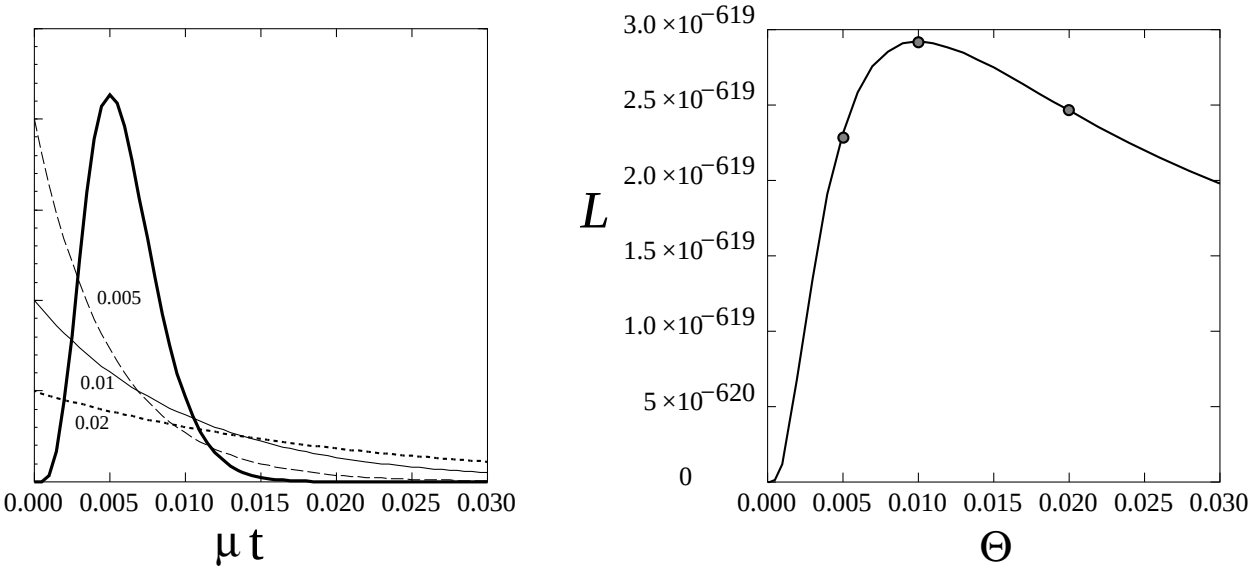


Figure 10.7: Inference from two sequences of 1000 bases that are 0.5% different. Left, the Kingman distributions of time of divergence for three different values of Θ, and also the probability of the sequences for different times of divergence (dark curve). Right, likelihoods computed by integrating different priors against that curve. The values for those three values of Θ are shown as points.

sequences. Integrating these for Kingman densities for different values of Θ, we get the likelihood curve shown on the right-hand side of the figure, with the three values 0.005, 0.01, and 0.02 indicated by circular points. Note that $\Theta = 0.005$ has a lower likelihood, because the smaller values of t have too low a probability of giving rise to the observed data. $\Theta = 0.01$ is the maximum likelihood value; $\Theta = 0.02$ leads us to expect too many values of t that would lead to too large a divergence of the sequences, so it has lower likelihood.

In a case this simple, a single numerical integration can allow us to average over our uncertainty about the coalescent tree. In cases with more sequences, the space of trees grows much larger. The relevant entities are trees which have their interior nodes in a particular time order. These are called "labeled histories", and Edwards (1970) has shown how to count them. A labeled history is essentially a sequence of coalescences. When there are n sequences, the number of different possible pairwise coalescences is simply $\binom{n}{2}$, which is $n(n-1)/2$. Taking the product of these for $n, n-1, n-2, \ldots, 2$ we find that there are $n!(n-1)!/2^{n-1}$ possible labeled histories. This can be a very

large number: for $n = 10$ there are fully 2.571×10^9 of them. Each such tree has $n - 1$ interior nodes, each of which can slide up and down between the next highest and lowest of them. As with the two-sequence case, the times of the nodes matter. Thus the summation over all trees for 10 sequences is actually a set of 2.571×10^9 integrals, each of them 9-dimensional!

Formally, we can write the likelihood in all these cases as

$$L = \sum_{G^*} \text{Prob}(G^*|N) \text{ Prob}(D|G^*, \mu) \tag{X-43}$$

provided that we understand that the summation is over all labeled histories and is also an integration over all node times within each of these. In general, we can change the time scale, as we could in the two-species case, scaling time in units of expected mutations per site. As it did in that case, it removes μ as a separate parameter and causes it to enter only as the product with N, so that the parameter is again $\Theta = 4N\mu$. Letting G be the tree with branch lengths in units of expected mutations per site rather than generations, the equation can then be written

$$L = \sum_{G} \text{Prob}(G \,|\, \Theta) \text{ Prob}(D \,|\, G) \tag{X-44}$$

These equations were first given by me (Felsenstein, 1988; see also 1992b).

If the summation could be done analytically, and result in a closed-form formula, we could go forward with this approach straightforwardly. So far, no one has discovered a way to do this.

MONTE CARLO INTEGRATION. Given that a numerical approach is then needed, we are faced with a vast number of high-dimensional integrals. Doing even one of them is extraordinarily difficult by conventional numerical integration. The normal method is to lay a grid over the space and evaluate heights of the curve at each grid point. For a one-dimensional numerical integration, one can usually achieve good coverage of the relevant areas with, say, 1000 points. For a 9-dimensional integration, a lattice with that many points would allow us only 2 points in each dimension. And we have not one, but billions of integrations to do.

To deal with such apparently hopeless cases, applied mathematicians have developed Monte Carlo integration. The general idea is that instead of defining a grid, we sample points at random from the domain, and evaluate the height of the function above each. If enough points are taken, this gives us a good estimate of the average height of the function, and thus of the value of the integral. You can see that if a relatively smooth function is evaluated over a two-dimensional space such as the floor of a room, a sample of, say, 1000 points at random from the floor would give a good result. The name "Monte Carlo" refers to the famous gambling casino, as the method relies heavily on the randomness.

However, it is less easy to see what happens with higher dimensionality. One is tempted, in the present case, to draw trees at random from the Kingman prior, and evaluate $\text{Prob}(D \mid G)$ for each. Equation X-44 is of this form: the probability of the data is averaged over the Kingman prior. With a large enough sample this will work, in theory. In practice, it does not. Results are wildly variable from one run to another, and are clearly not getting a consistent answer.

The reason appears to be that most of the trees drawn conflict with the data strongly enough that they make little contribution to the integral. Only a tiny fraction of all trees group similar sequences together, and these account for most of the area under the integral. If we evaluated a function over the floor of a room, and the function consisted of a high peak over one floor tile, but was nearly zero everywhere else, you can see that a Monte Carlo approach would be likely to fail. Most of the points would be drawn from places where the function was nearly zero. The estimate of the integral would depend heavily on exactly how many points were drawn over the one floor tile, and this would vary greatly from one run to another.

IMPORTANCE SAMPLING. To cope with this problem, *importance sampling* was developed. If there were some way of concentrating the sampling in the relevant region, the integral could be reliably evaluated. You might wonder if this was so. After all, if many of the points are then concentrated in the part of the domain where the function if highest, won't we get a misleadingly large estimate of the integral? This can be avoided by correcting these samples for their greater concentration in that region. If a certain area has twice as many points as another, we need to take each of those points as representative of only half as much area.

Doing this importance sampling correction, the noisiness of the integral is greatly reduced. We define an appropriate density function $g(x)$ and sample from it. We weight each of the samples inversely by how dense the samples will be in that region. This is seen in a simple manipulation of an integral. If the function we integrate is $f(x)$, the integral of this function can be rewritten as

$$\int f(x)\, dx \;=\; \int \frac{f(x)}{g(x)}\, g(x)\, dx \;=\; \mathbb{E}_g \left[\frac{f(x)}{g(x)} \right], \tag{X-45}$$

as the expectation of a quantity $h(x)$ under distribution $g(x)$ is the integral of the product $h(x)\, g(x)$. So our original integral of $f(x)$ is simply the expectation of $f(x)/g(x)$ evaluated at points drawn from the density function $g(x)$. The expectation is approximated by averaging the values of f/g for a large sample of points. If the function $g(x)$ is chosen carefully enough, it can greatly reduce the uncertainty in the integral. In the most optimistic case, if $g(x)$ is proportional to $f(x)$, each sample computes the constant of proportionality, which happens also to be the value of the integral! Only a single point would be needed. Of course, we are never in a situation this good.

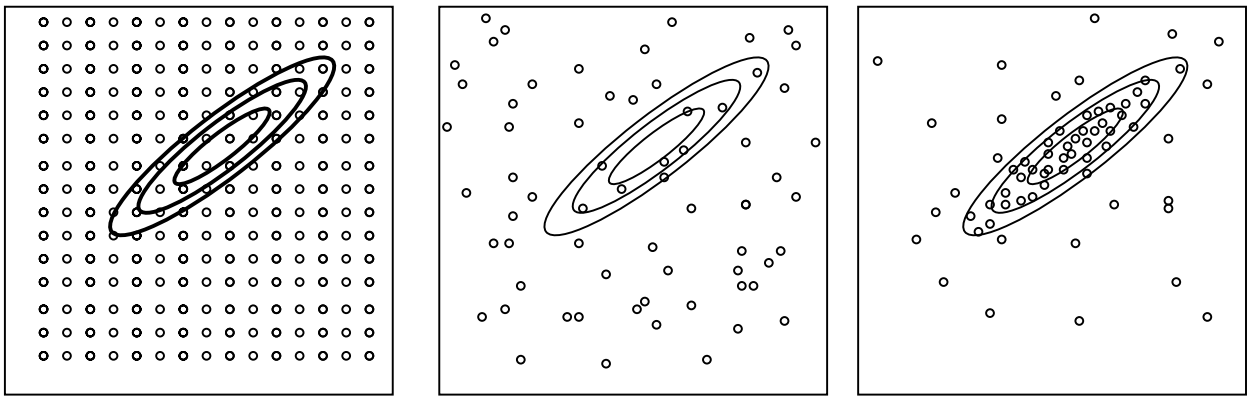


Figure 10.8: Three contour plots of the same function (in a nonbiological case) showing (left) a conventional grid for numerical integration, (center) a random sampling of points over the rectangle, and (right) points randomly sampled from a distribution that concentrates in the region where the function is large. The right-hand random sampling makes a much better estimate of the integral than the other two, if the samples are each weighted inversely according to how densely points are sampled in that region

Figure 10.8 shows importance sampling in a two-dimensional domain. The leftmost of the three contour plots shows a function and a rectangular grid of points. The center plot shows the same function with randomly sampled points. The rightmost plot shows points concentrated in the region in which the function is high. It should be apparent that it makes a much more relevant sample than the other two.

COMPUTING LIKELIHOODS. A number of different importance sampling methods have been developed for likelihoods with coalescents. For many of them, one draws from an importance density $g(G \mid \Theta)$ a series of coalescent trees $G_1, G_2, \ldots, G_n$. for some particular value $\Theta = \Theta_0$. To estimate the likelihood for other values of Θ one wants to use this sample to estimate the integral in equation X-44. This is done by computing an average:

$$\mathbb{E}_g \left[\frac{\text{Prob}(D \mid G)\ \text{Prob}(G \mid \Theta)}{g(G \mid \Theta_0)} \right] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Prob}(D \mid G_i)\ \text{Prob}(G_i \mid \Theta)}{g(G_i \mid \Theta_0)} \qquad \text{(X-46)}$$

Usually a sample of hundreds of thousands of trees G is needed to attain any accuracy.

There are two major variants of this approximation of the likelihood curve. In some cases we know the function $g(G \mid \Theta)$. In others we know it only up to a constant. This is possible because the Metropolis-Hastings sampling uses only the ratios of different g's for different trees G' and G.

A major issue in the likelihood approach is that the sampling is much more accurate when the "driving value" Θ_0 is close to the values of Θ for which we need the likelihoods.

BAYESIAN SAMPLERS. An alternative method that is coming into wide use is to take a Bayesian approach, where one has a prior distribution Prob (Θ) on Θ. We want to know what the posterior distribution of Θ is. The easiest way to do this is to consider the joint distribution of Θ and G, and sample from it. The posterior distribution of Θ is then approximated by simply taking the pairs (Θ_i, G_i) and ignoring the G's. By Bayes' Theorem,

$$\begin{aligned} \text{Prob}(\Theta, G \mid D) &= \frac{\text{Prob}(\Theta, G)\ \ \text{Prob}(D \mid \Theta, G)}{\text{Prob}(D)} \\ &= \frac{\text{Prob}(\Theta)\ \ \text{Prob}(G \mid \Theta)\ \ \text{Prob}(D \mid \mathit{Theta}, G)}{\text{Prob}(D)} \end{aligned} \tag{X-47}$$

The last term in the numerator can be simplified to $\text{Prob}(D \mid G)$ if G has its time scale as average mutations per site, but this is not essential. The denominator is intractable, but it need not be computed, as we shall note below. One can sample from the posterior by using Hastings-Metropolis sampling and then the denominator cancels out, and we are using the ratio of numerators to do the acceptance and rejection. For the Bayesian approach, there are no arbitrarily-chosen parameters that control the sampling: it involves trying both new G's and new Θ's.

GRIFFITHS-TAVARÉ INDEPENDENT SAMPLING. The pioneering importance sampling method for coalescents was Griffiths and Tavaré's (1994) independent sampling method. It was developed from the exact recursion calculation of Griffiths (1989). They thought of their method as approximating it by sampling paths through the recursion. It is not immediately obvious that this can also be seen as importance sampling of genealogies. The sequence of events in their recursion correspond to mutations to particular bases at particular sites and coalescences of particular lineages in the past. The history of a set of sequences is described by these events. In choosing a sample path through the recursion, they are specifying the past history of events.

Their sampler chooses particular sites to have had a mutation, or particular pairs of identical sequences to coalesce. The original method assumed an infinite sites model. The use of a DNA sequence model instead was difficult, because they had no bias in their sampling toward having two different sequences become more similar as they were followed back into the past. Their sampler did have the proper correction for the probabilities of events, but when used on sequence models it would sometimes have a very low chance of coming up with a sequence of events that accounted for a reasonable fraction of the total probability.

Nevertheless, their sampling method was not only pioneering, it had some advantages. The importance sampling function g is known, and there is no undetermined constant of proportionality. Each step in the sampling is quite rapid. Most significantly, each sample path, each reconstruction of the past history of events, is independent of the others. Thus their method cannot get stuck in one region of tree (or history) space. This category of methods are sometimes called IS (Independent Sampling) methods.

These advantages are counterbalanced by the frequency with which an improbable sequence of events is reconstructed, which can make a very large number of samples necessary. Griffiths and Tavaré's paper allowed not only for a constant population size, but also could estimate population growth rate in an exponentially growing population. Griffiths and Marjoram (1997) extended the method to deal with recombining coalescents, using ancestral recombination graphs. Bahlo and Griffiths (2000) extended the method to multiple populations with migration.

Stephens and Donnelly (2000) developed a biased reconstruction of mutations which went far towards making reasonable reconstructions. It tended to reconstruct more often mutations that carried a DNA sequence toward the others. The bias of their sampling was correctly compensated for in the importance sampling weighting. The result was a tenfold speedup of the method.

MARKOV CHAIN MONTE CARLO SAMPLING. Another approach was proposed by Kuhner, Yamato, and Felsenstein (1995). We used Metropolis-Hastings sampling to draw points from the distribution of genealogies. The Metropolis algorithm involves proposing changes from a current genealogy G to a new one G'. If we are trying to sample from an importance sampling density $g(x)$, we evaluate the density at the new point and at the old one. If $g(G')/g(G)$ is greater than 1, we accept the new point (and thus move to the new genealogy G). If $g(G')/g(G)$ is less than or equal to 1, we draw a random fraction R, accepting the new point when $R < g(G')/g(G)$. This has the effect of accepting a fraction $g(G')/g(G)$ of the time. In effect, it is a Markov process which achieves the desired equilibrium distribution. It is of the increasingly popular class of Markov chain Monte Carlo (MCMC) methods.

It can be shown that if the proposal distribution is able, in principle, to move anywhere in the space, the resulting distribution of points will be the desired distribution defined by the function $g(G)$. The samples are autocorrelated, so that a large number of samples may be needed to explore the space. The twin dangers are moving too far and altering the tree so much that the new tree is highly likely to be rejected, and moving too little so that one gets stuck in the initial area and does not adequately explore the space.

The importance sampling density used is simply proportional to the product $\text{Prob}(G \mid \Theta_0) \times \text{Prob}\ (D \mid G)$ for a driving value. The unknown constant of proportionality turns out to be the likelihood at the driving value, $L(\Theta_0)$. This in turn means that, although the method does not infer the likelihood $L(\Theta)$, it infers the likelihood ratio $L(\Theta)/L(\Theta_0)$.

The proposed moves in this case were erasures of a portion of the tree and its reformation with possible local rearrangements of branches and changes of times of coalescence. Later papers extended the method, altering the rearrangement scheme somewhat, to deal with exponentially growing populations (Kuhner, Yamato, and Felsenstein, 1998), with migration among two or more populations (Beerli and Felsenstein, 1999, 2001), and recombining sequences (Kuhner, Yamato, and Felsenstein, 2000).

Each move in these methods is more work than with the Griffiths-Tavaré independent sampling, as the probabilities of the data sum over all possible past histories of mutation, using the standard "pruning" algorithms for recursive computation of likelihoods on a tree. The trees are also necessary autocorrelated, and the possibility exists of failing to explore the space well enough. On the other hand, the importance sampling density is closer to the desired form, and most samples will not be wasted.

For further developments (including work on ascertainment correction with SNPs, haplotype inference, and some remarkable progress on coalescents with natural selection), the reader may want to consult the review in my book (Felsenstein, 2004, Chapters 26-28). The paper by Felsenstein et al. (1999) goes into some detail on why the Griffiths-Tavaré sampler is best regarded as carrying out importance sampling. Some of the material on this particular subject in my book (on page 481 of the book) is incorrect. A new method of importance sampling (Slatkin, 2002) promises further progress, but this has not yet been carefully investigated.

needs evaluation

APPROXIMATE BAYESIAN COMPUTATION. A more popular, if less powerful approach avoids using the full density function Prob $(G \mid \Theta)$Prob $(D \mid G)$. Instead it chooses some easily-computed sample statistics (such as the average heterozygosity). For any proposed value of the parameter Θ, we do a simple computer simulation to draw a set of DNA sequences and from that to obtain a value of the statistic S. If the S is sufficiently close to the observed value, that Θ is accepted as a sample from the posterior distribuion of Θ. This is called *Approximate Bayesian computation* or *ABC*. It was first used in population genetics by Fu and Li (1997) for a single parameter, and extended to multiple summary statistics by Weiss and von Haeseler (1998). A useful review is the one by Beaumont, Zhang, and Balding (2002).

The advantage of ABC is that it is fast and relatively easy. The disadvantage is that the choice of summary statistic needs to be a wise one, or there will be a loss of statistical power. Nevertheless, the samples from the posterior distribution using ABC are valid ones, provided the threshold for closeness to the observed summary statistic S is small. If these are not required to be small, the posterior distribution will be inaccurately assessed. If they are required to be small, the number of acceptances will be low and the simulation will need to run longer.

AN OBJECTIVE. When they can be done, sampling methods (both IS and MCMC)

are the state of the art in statistical inference from population samples of molecular sequences. When they are not practical, one must fall back on methods such as ABC. The sampling methods give hope of a "black box" which will accommodate many of the possible complications of evolutionary models (multiple loci, diploid genotypes, recombination, population size changes, migration, even simple kinds of natural selection). The user will specify what evolutionary scenarios to allow and what kinds of data have been supplied. The user will need to understand the evolutionary models employed, but may be relatively insulated from having to master the details of the sampling. The program will then run the sampler and provide a likelihood surface, or a Bayesian posterior distribution, for the genetic or population parameters. We are not there yet, though many of the pieces have been tested. The great unknown is how much sampling will be necessary in complicated models, and how long it will take for long stretched of genome.

Exercises

1. If two sequences differ at 65% of their positions and have evolved by a Jukes-Cantor model, what is the best estimate of the branch length between them?

2. Use the Jukes-Cantor model for the following computation: if two sequences differ at 10% of their sites, what is the branch length between them? If the second sequence then evolves into a third one by changing a completely different 10% of its sequence, what was the branch length between these two (that should be easy)? Compare the total branch length to the branch length you get when taking the first and third sequence and considering that they differ at 20% of their sites. Why the discrepancy?

3. A sample of 5 DNA sequences of 100 sites length has five segregating sites, each having a single copy of its variant nucleotide. Compute the estimate of θ from Watterson's number-of-segregating sites estimator. Compute $4N\mu$ from the nucleotide diversity. Taking the number of total sites into account, compare these [careful! you have to alter one of these numbers]. Are they supposed to be the same?

4. There are two populations (of diploid organisms), #1 having 10,000 individuals, #2 having 1,000 individuals. These population sizes have remained the same for a long time. We have sampled 6 copies of a gene from population #1 and 4 copies from population #2.

   (a) Using the coalescent approximation, if we have no migration between the populations, and we go back in time until we finally find that there are two of the 6 copies in population #1 come from the same copy, on average how many generations will that be? If we go back in time from the present until all 6

copies come from the same copy, on average how many generations will that be?

(b) If we do the same thing in population #2, on average how many generations will that be? If we go back in time from the present until all 4 copies come from the same copy, on average how many generations will that be?

(c) What is the probability that, going back from the present, that the first of these populations (going backwards) to have a coalescence will be population #1? *[Hint: calculate for each population a probability of coalescence per unit time, where time is scaled in generations. Roughly, this should be the reciprocal of the expected number of generations until that event. The probability that the first one (going backwards in time) coalesces in population #1 will be proportional to the probability of coalescence per unit time in that population, compared to the similar quantity in the other population.]*

(d) Now suppose that there is a 0.0001 migration rate in each direction between the two populations, so that in each generation there is a probability 0.0001 that a lineage is newly arrived from the other population. This probability applies separately to each lineage. What are the relative probabilities, going back one generation from the present, that (i) There is a coalescence in population #1, (ii) there is a coalescence in population #2, (iii) there is a lineage in #1 that is newly arrived from population #2, and (iv) there is a lineage in #2 that is newly arrived from population #1.

5. For four samples of a chromosome, each with two loci, A and B, that we are following, draw a coalescent tree that includes a single recombination event, such that the common ancestor at locus A is at a different time than the common ancestor at locus B. Is that always the case on any two-locus coalescent? Indicate on each branch of the coalescent which locus or loci that end up in the sampled chromosome copies are inherited along that branch. Note that you do not need to specify which alleles at the loci are present.

   Also write out the two coalescent trees for the individual loci. *Hint – look at the section on coalescents with recombination.*

Problems/Complements

1. Calculate the expectation of the nucleotide diversity between a pair of sequences under the Kimura 2-parameter model. How does it depend on the transition/transversion ratio R?

2. Suppose that a fraction f of the time a mitochondrion comes from the male parent

instead of the female parent.

(a) What is the probability that two gene copies in different individuals come from the same copy in the previous generation?

(b) Does this depend on whether the two individuals are both females, both males, or one of each? Why or why not?

(c) What is the distribution of the number of generations back to coalescence?

(d) If we have a population of 100,000 individuals with a 1% chance that each one has mitochondria derived from the male parent, what is the mean time to coalescence of a mitochondrial gene? How much larger is this than it would be if all mitochondria were derived from the female parent?

3. Consider a Moran model (described in Chapter VI) in which, at each instant in continuous time, one individual in a haploid population is killed and replaced by a copy of one of the others. What is the exact distribution of time to coalescence of two copies? What is the exact process that corresponds to the coalescent? How does it compare to the coalescent that has the same effective population size?

4. What is the exact distribution of trees from a sample of three gene copies from a diploid Wright-Fisher model with $N = 4$?

5. Draw an ancestral recombination graph with one recombination event, in which two different loci have different times of their Most Recent Common Ancestor (MRCA).

Chapter XI

POLYGENIC CHARACTERS IN NATURAL POPULATIONS

XI.1 Phenotypic Evolution Models

The theory of quantitative genetics is a short-term theory, projecting response to selection for a few generations starting with a population in linkage equilibrium. It is far harder to predict long term distributions of quantitative characters or to take linkage disequilibrium into account. We have already seen in Chapter IX that optimizing selection will generate linkage disequilibrium. It is in part the realization of that fact, along with the short-term nature of the theory, which has prevented the application of quantitative genetics theory to natural selection. In the last decade this has begun to change. A class of models known as *phenotypic evolution models* has been developed, starting with the work of Kimura (1965), Slatkin (1970) and Bulmer (1971). These models attempt to take linkage disequilibrium effects into account, while remaining within a quantitative genetics framework so that gene and gamete frequencies need not be followed explicitly. These models approximate the genetics of the trait, some so severely that the genetics disappears from view, and the model then speaks only of the evolution of phenotypes. While the proponents of these approaches sometimes regard them as more general than explicitly genetic models, they seem to involve rather restrictive genetic assumptions in the ways in which the phenotypes of parents are allowed to affect those of their offspring.

EFFECT OF OPTIMIZING SELECTION. We will have space here only to sketch a simple case: the balance between optimizing selection and mutation. Optimizing selection will continually reduce the variance of the selected phenotype. Mutation will increase it. A balance will be reached between these two forces, and we are trying to find what it will be. Let us start by assuming that the phenotype in which we are interested follows a normal distribution with mean zero and variance σ^2, so that its distribution has density

function:

$$f(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\exp[-x^2/2\sigma^2]. \tag{XI-1}$$

The kind of natural selection to which we expose these individuals is called *optimizing* selection. The fitness function is shaped like a normal distribution:

$$w(x) \;=\; \exp[-x^2/(2S)]. \tag{XI-2}$$

Thus the highest fitness is at the phenotype which also happens to be the population mean, namely zero. We make that simplification purely to avoid having to worry about the mean. In general, terms for the mean do appear in phenotypic evolution models. If we now look at the distribution of phenotypes among survivors of selection (assuming the fitness to be expressed through viability, or else weighting individuals by their fertilities), this is easily shown to have a density function proportional to $f(x)w(x)$. This turns out to be a normal distribution with mean zero and variance

$$\sigma_{as}^2 \;=\; 1/(1/\sigma^2 + 1/S). \tag{XI-3}$$

Thus, the effect of selection has been to reduce the variance of the phenotype by an amount depending on the parameter S, which reflects the strength of selection. A small S indicates strong selection, for then, by (XI-2), fitness will fall off rapidly as the phenotype departs from zero. This part of the argument is easy: the difficulties, as well as the differences between the various models, arise when we ask what this reduction in phenotypic variance implies for the offspring distribution.

XI.2 Kimura's model

Kimura (1965) made the pioneering model of the mutation-selection balance in quantitative characters. He assumed that the gametes would have a normal distribution of genetic effects. If we take the reduction of the phenotypic variance that is implied by equation (XI-3), half of it will come by creating a negative correlation between the effects in the two gametes, and half by reducing the variances of the gamete effects, by eliminating more extreme haplotypes. So the change in genetic variance from selection is (if there is no environmental variance)

$$\frac{1}{2}\left(G - \frac{GS}{G+S}\right) \;=\; -\frac{1}{2}\frac{G^2}{G+S} \tag{XI-4}$$

If we then assume that mutation adds U to the genetic variance, the net change in the genetic variance of an individual is

$$-\frac{1}{2}\frac{G^2}{G+S} + U. \tag{XI-5}$$

At equilibrium we can equate this to zero, and obtain as the equilibrium genetic variance Kimura's result:

$$G = U + \sqrt{U^2 + 2US} \qquad \text{(XI-6)}$$

Note that this derivation assumes that the gametes remain normally distributed. As we will see, this is not uncontroversial.

XI.3 Lande's model

The most sophisticated development of the normally-distributed phenotypic evolution models was by Lande (1976b), who made them model changes due to linkage disequilibrium. If the reduction of phenotypic variance is accomplished mostly by changing the gene frequencies, then we should expect the variance to continue at its new value in the next generation. On the other hand, if it reflects primarily the creation of linkage disequilibria, then we expect that as part of that disequilibrium breaks down the variance should return part of the way towards its previous value. It is here that the genetic assumptions become critical. We will describe Lande's scheme briefly, then present the equations for a restricted version of it.

Lande starts by assuming that the phenotype is the sum of individual allele effects at n loci. There are no dominance effects allowed in his model. The allele effects are then assumed to follow a $2n$-variate normal distribution. This is a strong simplifying assumption. It can be regarded as an approximation to the situation we would have if there were two alleles per locus. It could not then be exact because, among other things, the effects at one locus are then not normally distributed. Alternatively, one could imagine that there were an infinite number of possible alleles at each locus, and that the allele effects follow a normal distribution. A multivariate normal distribution is completely characterized by its means, variances and covariances. Thus, as long as we can approximate the joint distribution of allele effects as being $2n$-variate normal, we can obtain a complete description of the distribution if we know the mean and variances at each of the $2n$ sites, and the pairwise covariances between them.

In this model it is usual to assume that there is random mating. This ensures that the covariances between sites on different gametes are zero at the beginning of each generation. The means and variances reflect the gene frequencies at each locus, and the covariances between sites on the same gamete are the equivalent of linkage disequilibria. In Lande's model we expect linkage disequilibrium to arise when optimizing selection acts. Suppose that capital letters represent alleles which increase the phenotype. An individual copy of A is more likely to survive if it is in an individual which has a b than a B at the next locus, so that after selection there will be a lack of independence between loci. This extends to genes on different gametes as well: a A is more likely to survive if the gamete opposite it (the one which came from the other parent) has a b than if

it has a B. This is true as well for two genes at the same locus: an Aa is more likely to survive selection than an AA, so that the two genes at one locus are not independent after selection. Lande is able to compute the means, variances, and covariances after selection in terms of the means, variances, and covariances before selection. This involves matrix algebra, and is too complex a derivation to give here.

A SYMMETRIZED VERSION. Instead, let us impose some further restrictions. Let us assume that the loci are completely exchangeable: all have the same means, all the same variances, and all pairwise covariances are equal. This state of complete symmetry can only be maintained if all pairs of loci have the same recombination fraction. That in turn will only be true if all pairs of loci are completely linked or all completely unlinked. Suppose that the n loci are all unlinked. We now have a model in which the variance at each of the $2n$ genes is v, and the covariance (before selection) between pairs of genes from the same parent is c, and the mean effect at each gene is m. Under this symmetry all of Lande's matrix expressions become much simpler, though we shall still not give them here. A simple result emerges (readers interested in its derivation can consult Appendix 2 of the paper by Felsenstein, 1979b). Let us focus on the changes in the variances and covariances. The optimizing selection will reduce the variances v, make the covariances c more negative, and create a negative covariance between genes which are on opposite gametes. The generalization which emerges for the case of exchangeable loci is that there is an equal reduction x in each of these terms. Thus, after selection $v' = v - x$ and $c' = c - x$. Also, a negative covariance of $-x$ is created between each pair of genes on opposite gametes (these covariances being zero before selection, due to random mating). Recombination will not affect the variances v, but it will have an easily calculable effect on the covariance. Two genes at different loci in the gametes produced by a survivor of selection are equally likely to have come from the same or from opposite gametes. So

$$c' \;=\; \frac{1}{2}(c - x) + \frac{1}{2}(-x) \;=\; \frac{1}{2}c - x \qquad \text{(XI-7)}$$

At the beginning of the next generation, after these gametes have combined at random, the phenotypic variance will consist of two parts. One part is due to the variance terms. There are $2n$ of these, each being v'. The other part of the phenotypic variance is due to the covariances c'. There are $2n(n-1)$ of these. Let us call these two parts of the variance respectively V and C. It should be clear from all of this that V will be changed by $-2nx$ by selection, and C by $-2n(n-1)x$. It remains to determine x. Recall that the total change of phenotypic variance is divided equally among the $4n^2$ possible terms. The total change of variance can be computed from (XI-3) to be $-\sigma^4/(\sigma^2 + S)$. Since $\sigma^2 \;=\; V + C$, we have

$$4n^2x \;=\; (V + C)^2/(S + V + C) \qquad \text{(XI-8)}$$

so that we finally obtain for the changes in V and C under selection and recombination:

$$\begin{aligned} V' &= V - \frac{1}{2n}\frac{(V+C)^2}{S+V+C} \\ C' &= \frac{1}{2}C - \frac{n-1}{2n}\frac{(V+C)^2}{S+V+C} \end{aligned} \tag{XI-9}$$

EFFECT OF MUTATION. Now we can easily add the change caused by mutation. Suppose that we regard mutation as adding a random amount to the effect of each gene. If e and e' are two such random increments, then since $\text{Cov}\,(x+e, x+e') = \text{Cov}\,(x, x')$, the mutation effects do not alter the covariances. Since $\text{Var}\,(x+e) = \text{Var}\,(x) + \text{Var}\,(e)$, they do increase the variances. Thus we model mutation by saying that it adds a quantity with mean zero and variance u to each gene effect. The net effect is to add $U = 2nu$ to the total of the variances, so that if mutation follows selection and recombination in the life cycle we can write simply

$$V'' = V' + U. \tag{XI-10}$$

When the whole system reaches equilibrium (which we assume it will), we must have $V'' = V$ and $C' = C$. Using (XI-9) and (XI-10) this gives

$$V'' - V = 0 = U - \frac{1}{2n}\frac{(V+C)^2}{S+V+C} \tag{XI-11}$$

and

$$\frac{1}{2}C = -\frac{n-1}{2n}\frac{(V+C)^2}{S+V+C} \tag{XI-12}$$

These can easily be solved for V and C in terms of U, S, and n. The result is

$$C = -2(n-1)U \tag{XI-13}$$

and

$$V = (3n-2)U + \sqrt{n^2U^2 + 2nUS}, \tag{XI-14}$$

predicting a total genetic variance at equilibrium of

$$V + C = nU + \sqrt{n^2U^2 + 2nUS}. \tag{XI-15}$$

Thus we are able to make an approximate calculation under this simplified version of Lande's model of the amount of variance and covariance maintained at an equilibrium between mutation and optimizing selection. An interesting feature of this equilibrium is that the amount of variance maintained depends on the number of loci n. It is not altogether obvious that this would be so, for we have already taken U to be the total mutational increment of the genetic variance, summed over all loci. But the model maintains a substantial amount of genetic variation at equilibrium. In effect, the term in n comes from the interference caused by variation at each locus in the selection at the others.

ENVIRONMENTAL VARIANCE. So far the model has assumed that the character has no environmental variance. If we add to the model an environmental variance E, then we must distinguish between the breeding value and the phenotype. If A is the breeding value and $A + e$ the phenotype, then we can calculate the fitness as a function of the breeding value in the following fashion:

$$w(A) \;=\; \sum_{e} \text{Prob}(e)\;\; \exp[-(A+e-P)^2/(2S)]. \tag{XI-16}$$

When $\text{Prob}(e)$ is taken to be a normal distribution with mean zero and variance E, the summation is an integration. We finally find that

$$w(A) \;=\; K \exp[-(A-P)^2/2(S+E)], \tag{XI-17}$$

where K is a constant which need not concern us. Thus the effect of the environmental variance is to weaken the selection by replacing S by $S + E$ throughout the derivation of this section. The equilibrium genetic variances and covariances V, C, and $V + C$ can be obtained in this fashion. Of course, it must be kept in mind that the equilibrium phenotypic variance will be $V + C + E$, not just $V + C$. In effect, the environmental variance weakens the selection by causing some of it to be expended uselessly in eliminating extreme individuals who owe their phenotypes to the environment. The presence of environmental variance means that the phenotype is no longer a reliable guide to the breeding value, and this lessens the effect of selection on the phenotype.

THE MEAN. As an aside, we may add that under this model, the population mean follows the equation:

$$M' \;=\; \frac{M(S+E) + P(V+C)}{S+E+V+C} \tag{XI-18}$$

If the mean were to start at a different value than the optimum phenotype P, this equation simply predicts that at equilibrium $M' = M = P$. This can hardly be a great surprise.

STRENGTHS AND LIMITATIONS. Lande's model allows us to obtain an equilibrium solution for the amount of variance maintained by mutation, and also to describe the effect of linkage disequilibrium (via the covariance terms) without having to follow $2n$ different quantities. This is not achieved without cost. Although the population is assumed to have all gene effects multivariate normally distributed, this cannot be strictly true. Even within the confines of Lande's model, multivariate normality is violated. The problem is with the recombination process. Although the optimizing selection leaves the survivors in a multivariate normal distribution, the recombination process will give a gamete pool which is a mixture of gametes which have undergone different kinds of recombination. This mixture cannot be multivariate normal. The conditions for maintenance of exact multivariate normality under Lande's model have been investigated

(Felsenstein, 1977), and it is found to be essentially impossible as long as recombination exists. Thus Lande's model is an approximation. There has as yet been no detailed investigation of the validity of the approximation, but it seems likely that it is a good one if selection is weak and recombination moderate to strong. The advantage to using Lande's model or one of the other phenotypic evolution models is that they allow us to explicitly allow for the changes in genetic variance as a result of selection, mutation, and recombination. For some further insights into the behavior of Lande's model see the paper by Chevalet (1994).

XI.4 Bulmer's model

A different prediction of the equilibrium between mutation and selection was made by Bulmer (1974) and it is instructive to compare it with Lande's. His derivation is a bit complex; we will simplify it by crude but painless approximations. Bulmer worked out approximations for a character controlled by n two-allele loci, and in effect showed that selection at the loci does not interact. We simply take this non-interaction as an assumption.

Imagine a single locus in which there is a mutation-selection balance. At equilibrium the mean fitness at the locus is reduced by $2u$. The mean fitness at that locus is essentially $1-2u \simeq \exp(-2u)$. With n loci the mean fitness is then approximately $\exp(-2nu)$. What level of genetic variability will lead to such a reduction in fitness? If the quantitative character has no environmental variation, so that its variance $\sigma^2 = G$ is entirely genetic, then we can compute the mean fitness by integrating the product of (XI-1) and (XI-2):

$$\bar{w} = \int \frac{1}{\sqrt{G}\sqrt{2\pi}} \exp[-x^2/(2G)] \exp[-x^2/(2S)]\ dx \tag{XI-19}$$

which can be evaluated by noting that it is

$$\begin{aligned} &\tfrac{1}{\sqrt{G}\sqrt{2\pi}} \int \exp[-x^2/(2G)] \exp[-x^2/(2S)]\ dx \\ &= \sqrt{\tfrac{S}{G+S}} \left(\tfrac{1}{\sqrt{2\pi}\sqrt{\frac{1}{\frac{1}{G}+\frac{1}{S}}}} \int \exp\left[-\tfrac{x^2}{2}\left(\tfrac{1}{G}+\tfrac{1}{S}\right)\right]\ dx \right) \end{aligned} \tag{XI-20}$$

We recognize the integral of a normal distribution in the expression in large parentheses; it is 1, leaving us with

$$\bar{w} = \sqrt{\frac{S}{G+S}} = e^{-2nu}. \tag{XI-21}$$

Solving for G, we get

$$G = S\left(e^{4nu}-1\right) \tag{XI-22}$$

or, to good approximation if $4nu$ is small,

$$G \simeq 4nuS \tag{XI-23}$$

It will also help to recall that (from the previous section), the effect of environmental variance E is to increase S to $S + E$. That immediately gives us the result for genetic variance in that more general case:

$$G = 4nu(S + E) \tag{XI-24}$$

Note that in expression (XI-24) the size of the mutational effects is completely absent! This is the analogue of the effect in ordinary mutational load arguments, where the selection coefficient s does not appear in the expression for the mutational load. Doubling the size of the mutational effects would lead to a lower frequency of the mutants, for the same net genetic variance maintained by the mutation-selection balance.

COMPARISON OF THE TWO MODELS. In Lande's model, the size of mutational effects affects U and the result. It is thus immediately apparent that Lande's and Bulmer's approximations must differ. A numerical example will help. Suppose that a quantitative character has its variation due to 30 loci, each with mutation rates 10^{-5}. The mutants change the character by 0.1 phenotypic units each, on average. Thus the variance due to mutation from each locus in each generation is $2 \times 10^{-5} \times 10^{-2} = 2 \times 10^{-7}$. So Lande's U is 0.000006. Now suppose that $S = 10$, so that in the 0.1 units on the phenotypic scale that a typical mutation moves the phenotype, the fitness drops by $\exp(-0.01/20) \simeq 0.0005$.

Substituting into Lande's formula (XI-12) we get

$$G = V + C = 0.00018 + \sqrt{3.6 \times 10^{-11} + 60 \times 6 \times 10^{-6} \times 10} = 0.06018. \tag{XI-25}$$

while in Bulmer's formula we get instead

$$G = 4 \times 30 \times 0.00001 \times 10 = 0.0120. \tag{XI-26}$$

so that Lande's predicts five times as much variance as Bulmer's. The difference is not quite as great when measured in standard deviations, Lande's predicting 0.24532 while Bulmer predicts 0.109545.

The difference between the two predictions is greater when mutation rates are smaller, when number of loci is larger, when mutation effects are larger, or when selection is stronger (S is smaller).

Why the difference between the two predictions? Lande's argument assumes that the distribution of genotypes is normal, which is most nearly achieved when selection is weak. Bulmer's argument ignores (as given here) or approximates away (in its original

form) the linkage disequilibria that arise between loci. It may be doubted that, for many cases of interest, the Lande result is more accurate. In the above numerical example, the equilibrium genetic standard deviation is predicted to be no more than the size of two mutation effects. This suggests that in few cases will a mutant re-mutate before selection eliminates it. So the linkage disequilibria may have little effect, and normality may be hard to assume.

For a more detailed examination of these questions, see the comprehensive study by of these models by Turelli (1984).

XI.5 Other models

The other phenotypic evolution models can mostly be obtained as special cases of our symmetrized version of Lande's model. Kimura's model (1965) is essentially the case in which $n = 1$, so that C is always zero. Bulmer's earlier model of 1971 is the case where $n = \infty$, which implies that V remains unchanged at its initial value if there is no mutation. Cavalli-Sforza and Feldman (1976) give a system of equations reminiscent of the one-locus case, but they do not take into account the negative covariance between the effects of the two copies of a gene after selection, and consequently their results differ from Lande's. Slatkin's (1970) system imposes an external constraint in the form of an assumed constancy of within-sibship variance: this constancy does not obtain in Lande's model and amounts to an arbitrary assumption.

A PHILOSOPHICAL DIFFERENCE. From a genetic point of view, the other phenotypic models are justified to the extent that they can be derived from a model such as Lande's, which attempts to take genetic factors into account. Many authors prefer the position that their models are arbitrary assumptions about the evolution of phenotypes, without specifically genetic assumptions. The difficulty with this position is that if it is then asserted that these phenotypic models are more general than the genetic models, one has to account for the fact that all of the other models either arise as special cases of Lande's model or are incompatible with it.

SOME FURTHER REFERENCES. The phenotypic evolution models have found application in a number of contexts, particularly in ecology. Of particular interest are the papers of Roughgarden (1972, 1974a, 1974b) Slatkin and Lande (1976), and Slatkin (1979) on the evolution of niche overlap, the papers of Lande (1976a, 1977) on long-term evolutionary effects, and the work of Feldman and Cavalli-Sforza (Cavalli-Sforza and Feldman, 1976, 1978; Feldman and Cavalli-Sforza 1977) on models incorporating cultural transmission.

Complements/Problems

1. In the phenotypic evolution models of Lande and Bulmer given above, how is the equilibrium variance of a trait under optimizing selection vs. mutation affected by doubling the number of loci and halving the contribution to the mutational variance U in each locus?

2. Evolutionary quantitative geneticists often assume, for simplicity, that there is an equilibrium between new variation arising by mutation and old variation disappearing by selection and by genetic drift. For Bulmer's model, with 10 loci affecting the trait and having equal strengths of selection and equal mutation rates, see if you can use the diffusion equations for an equilibrium between mutation and selection. Work out (numerically) the variance in the amount of genetic variance contributed by that locus. For Bulmer's model the individual loci contribute additively to the overall genetic variance. With 10 loci, we can add these variances. What will be the coefficient of variation, from generation to generation, of the genetic variance maintained by the balance of mutation, genetic drift, and selection?

REFERENCES

Abbot, P. and 136 other authors. 2011. Inclusive fitness theory and eusociality. *Nature* **471:** E1–4. doi: 10.1038/nature09831.

Akin, E. 1979. *The Geometry of Population Genetics.* Lecture Notes in Biomathematics No. 31, Springer-Verlag, Berlin.

Alexander, R. D. 1974. The evolution of social behavior. *Annual Review of Ecology and Systematics* **5:** 325–383.

Altenberg, L., and M. W. Feldman. 1987. Selection, generalized transmission and the evolution of modifier genes. I. The reduction principle. *Genetics* **117:** 559–572.

Altenberg, L. 2011. An evolutionary reduction principle for mutation rates at multiple loci. *Bulletin of Mathematical Biology* **73:** 1227–1270.

Altenberg, L. 2012. Resolvent positive linear operators exhibit the reduction phenomenon. *Proceedings of the National Academy of Sciences, USA* **109 (10):** 3705–3710.

Altenberg, L., U. Liberman, and M. W. Feldman. 2017. Unified reduction principle for the evolution of mutation, migration, and recombination. *Proceedings of the National Academy of Sciences, USA* **114 (12):** E2392–E2400. doi:10.1073/pnas.1619655114

Anderson, W. W. 1971. Genetic equilibrium and population growth under density-regulated selection. *American Naturalist* **105:** 489–498.

Asmussen, M. A. and M. W. Feldman. 1979. Density dependent selection 1: A stable feasible equilibrium may not be attainable. *Journal of Theoretical Biology* **64:** 603–618.

Asmussen, M. A. 1979. Density-dependent selection II. The Allee effect. *American Naturalist* **114:** 796–809.

Atkinson, F. V., G. A. Watterson, and P. A. P. Moran. 1960. A matrix inequality. *Quarterly Journal of Mathematics. Oxford Series* **11:** 137–140.

Atwood, K. C., L. K. Schneider and F. J. Ryan. 1951. Periodic selection in Escherichia coli. *Proc. Natl. Acad. Sci.* **37:** 146–155.

Bahlo, M. and R. C. Griffiths. 2000. Inference from gene trees in a subdivided population. *Theoretical Population Biology* **57:** 79–95.

Bartlett, M. S., and Journal of B. S. Haldane. 1934. The theory of inbreeding in autotetraploids. *Journal of Genetics* **29:** 175–180.

Beaumont, M. A., W. Zhang, and D. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162 (4):** 2025–2035.

Beerli, P. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152:** 763–773.

Beerli, P. and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci.* **98:** 4563–4568.

Bernstein, F. 1931. The geographical distribution of blood groups and its anthropological significance (German). pp. 227–243 in Comitato Italiano per lo Studio dei Problemi della Populazione. Istituto Poligrafico dello Stato, Rome.

Bodmer, W. F. and P. A. Parsons. 1962. Linkage and recombination in evolution. *Advances in Genetics* **11:** 1–100.

Bodmer, W. F. 1965. Differential fertility in population genetics models. *Genetics* **51:** 411–424.

Bodmer, W. F. and J. Felsenstein. 1967. Linkage and selection: theoretical analysis of the deterministic two locus random mating model. *Genetics* **57:** 237–265.

Bodmer, W. F. and L. L. Cavalli-Sforza. 1968. A migration matrix model for the study of random genetic drift. *Genetics* **59:** 565–592.

Bulmer, M. G. 1971. The effect of selection on genetic variability. *American Naturalist* **105:** 201–211.

Bulmer, M. G. 1974. Linkage disequilibrium and genetic variability *Genetical Research* **23:** 281–289.

Bulmer, M. G. 2003. *Francis Galton: Pioneer of Heredity and Biometry.* Johns Hopkins University Press, Baltimore.

Cannings, C. 1971. Natural selection at a multiallelic autosomal locus with multiple niches. *Journal of Genetics* **60:** 255–259.

Cannings, C. 1973. An increasing fitness function for a population with many niches. *Nature* **241:** 47.

Castle, W. E. 1903. The laws of Galton and Mendel and some laws governing race improvement by selection. *Proceedings of the American Academy of Arts and Sciences* **39:** 233–242.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19:** 233–257 (also published in: Evolution **21:** 550–570 1967).

Cavalli-Sforza, L. L. and G. Zei. 1967. Experiments with an artificial population. pp. 473–478 in *Proceedings of the Third International Congress of Human Genetics*, ed. J. F. Crow and J. V. Neel. The Johns Hopkins Press, Baltimore.

Cavalli-Sforza, L. L., and W. F. Bodmer. 1971. *The Genetics of Human Populations*. W. H. Freeman, San Francisco.

Cavalli-Sforza, L. L., and M. W. Feldman. 1976. Evolution of continuous variation: direct

approach through joint distribution of genotypes and phenotypes. *Proc. Natl. Acad. Sci.* **73:** 1689–1692.

Cavalli-Sforza, L. L., and M. W. Feldman. 1978. The evolution of continuous variation. III. Joint transmission of genotype, phenotype, and environment. *Genetics* **90:** 391–425.

Cavalli-Sforza, L. L. and M. W. Feldman. 1978. Darwinian selection and "altruism". *Theoretical Population Biology* **14:** 268–280.

Ceppelini, R., M. Siniscalco, and C. A. B. Smith. 1955. The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics* **20:** 97–115.

Charlesworth, B. 1971. Selection in density-regulated populations. *Ecology* **52:** 469–474.

Charlesworth, B. 1978a. Some models of the evolution of altruistic behavior between siblings. *Journal of Theoretical Biology* **72:** 297–319.

Charlesworth, B. 1978b. Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl. Acad. Sci.* **75(11):** 5618–5622.

Charlesworth, B. 1979. Evidence against Fisher's theory of dominance. *Nature* **278** 848–849.

Charlesworth, B. 1980. *Evolution in Age-Structured Populations.* Cambridge Studies in Mathematical Biology 1. Cambridge University Press, Cambridge.

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* **63(3):** 213–227.

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215–220.

Chen, H., R. E. Green, S. Pääbo, and M. Slatkin. 2007. The joint allele-frequency spectrum in closely related species. *Genetics* **177:** 387–398.

Chetverikov, S. S. 1926. On certain aspects of the evolutionary process from the standpoint of modern genetics (Russian). *Zhurnal Eksperimental'noi Biologii* **2:** 3–54 (translation in Proceedings of the American Philosophical Society **105:** 167–195 1961).

Chevalet, C. 1994. An approximate theory of selection assuming a finite number of quantitative trait loci. *Genetique Selection Evolution* **26:** 379–400.

Chia, A. B., and E. Pollak. 1974. The inbreeding effective number and the effective number of alleles in a population that varies in size. *Theoretical Population Biology* **6:** 149–172.

Clarke, B. C. and P. O'Donald. 1964. Frequency-dependent selection. *Heredity* **19:** 201–206.

Clarke, B. C. 1972. Density-dependent selection. American Naturalist **106:** 1–13.

Cockerham, C. C. 1954. An extension of the concept of partitioning of hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39:** 859–882.

Cockerham, C. C. and P. M. Burrows. 1971. Populations of interacting autogenous components. *American Naturalist* **105:** 13–29.

Cockerham, C. C. 1971. Higher-order probability functions of identity of alleles by descent. *Genetics* **69:** 235–246.

Cockerham, C. C., P. M. Burrows, S. S. Y. Young, and T. Prout. 1972. Frequency-dependent selection in randomly mating populations. *American Naturalist* **106:** 493–515.

Comstock, R. E., H. F. Robinson, and P. H. Harvey. 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. *Agronomy Journal* **41:** 360–367.

Conley, C. 1975. An application of Wazewski's method to a non-linear boundary value problem which arises in population genetics. *Journal of Mathematical Biology* **2:** 241–249.

Cooper, A. J., B. S. Blumberg, P. L. Workman, and J. R. McDonough. 1963. Biochemical polymorphic traits in a U. S. White and Negro population. *American Journal of Human Genetics* **15:** 420–428.

Crow, J. F. 1954. Breeding structure of populations. II. Effective population number. pp. 543–556 in Statistics and Mathematics in Biology, edited by O. Kempthorne, T. A. Bancroft, J. W. Gowen and J. L. Lush. Iowa State College Press, Ames, Iowa.

Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory.* Harper and Row, New York (Burgess, Minneapolis 1978. There may be a reprinting by Princeton University Press soon).

Crow, J. F., and M. Kimura. 1965. Evolution in sexual and asexual populations. *American Naturalist* **99:** 439–450.

Crow, J. F., and M. Kimura. 1972. The effective number of a population with overlapping generations: a correction and further discussion. *American Journal of Human Genetics* **24:** 1–10.

Crow, J. F., and C. Denniston. 1974. *Genetic Distance.* Plenum Press, New York and London.

Crow, J. F., and K. Aoki. 1982. Group selection for a polygenic behavioral trait: a differential proliferation model. *Proc. Natl. Acad. Sci.* **79:** 2628–2631.

Crow, J. F. 1988. Eighty years ago: the beginnings of population genetics. *Genetics* **119:** 473-âĂŞ476.

Crow, J. F. 1999. Hardy, Weinberg and language impediments. *Genetics* **152:** 821–825.

Cruden, D. 1949. The computation of inbreeding coefficients in closed populations. *Journal of Heredity* **40:** 248–251.

Cvijović, I., B. H. Good, E. R. Jerison, and M. M. Desai. 2015. Fate of a mutation in a fluctuating environment. *Proceedings of the National Academy of Sciences, USA* **112 (36):** E5021–E5028.

Danforth, C. H. 1921. The frequency of mutation and the incidence of hereditary traits in man. Proceedings of the. Second International Congress of Eugenics, Vol. 1, pp. 120–128.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B* **39:** 1–38.

Dempster, E. R., and I. M. Lerner. 1950. Heritability of threshold characters. *Genetics* **35:** 212–236.

Dempster, E. R. 1955. Maintenance of genetic heterogeneity. *Cold Spring Harbor Symposia on Quantitative Biology* **20:** 25–32.

Denniston, C. 1974. An extension of the probability approach to genetic relationships: one

locus. *Theoretical Population Biology* **6:** 58–75.

Denniston, C. 1975. Probability and genetic relationship: two loci. *Annals of Human Genetics* **39:** 89–104.

Dickerson, G. E. 1952. Inbred lines for heterosis tests? pp. 350–351 in Heterosis, ed. J. W. Gowen. Iowa State College Press, Ames, Iowa.

Durrett, R., and J. Schweinsberg. 2004. Approximating selective sweeps. *Theoretical Population Biology* **66:** 129–138. doi:10.1016/j.tpb.2004.04.002

Düsing, C. 1884. The Regulation of the Sex Ratio in the Multiplication of People, Animals, and Plants (German). Verlag von Gustav Fischer, Jena.

Düsing, K. 1883. Die Factoren welche die Sexualität entscheiden *Jenaische Zeitschrift für Naturwissenschaft* **16:** 428–464.

East, E. M., and D. F. Jones. 1919. *Inbreeding and Outbreeding: Their Genetic and Sociological Significance.* J. B. Lippincott, Philadelphia.

East, E. M. 1936. Heterosis. Genetics **21:** 375–397.

Edmonds, C. A., A. S. Lillie, L. L. Cavalli-Sforza. 2004. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences, USA* **101:** 975–979.

Edwards, A. W. F. 1967. Fundamental theorem of natural selection. *Nature* **215:** 537–538.

Edwards, A. W. F. 1968. Simulation studies of genealogies. Heredity **23:** 628.

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B* **32:** 155–174.

Edwards, A. W. F. 1990. Fisher, $\bar{W}$, and the Fundamental Theorem. *Theoretical Population Biology* **38:** 276–284. doi:10.1016/0040-5809(90)90014-M

Edwards, A. W. F. 1994. The fundamental theorem of natural selection. *Biological Reviews* **69:** 443–474.

Edwards, A. W. F. 2008. G. H. Hardy (1908) and Hardy-Weinberg Equilibrium *Genetics* **179:** 1143–1150.

Edwards, A. W. F. 2014. R.A. Fisher's gene-centred view of evolution and the Fundamental Theorem of Natural Selection. *Biological Reviews* **89:** 135–147. doi:10.1111/brv.12047

Eigen, M. and P. Schuster. 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58 (10):** 465–523.

Eigen, M., J. McCaskill, and P. Schuster. 1988. Molecular quasi-species. *Journal of Physical Chemistry* **92:** 6881–6891.

Emigh, T. H. and E. Pollak. 1979. Fixation probabilities and effective population numbers in diploid populations with overlapping generations. *Theoretical Population Biology* **15:** 86–107.

Emik, L. O., and C. E. Terrill. 1949. Systematic procedures for calculating inbreeding coefficients. *Journal of Heredity* **40:** 51–55.

Encode Project Consortium (442 authors). 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489 (Sep. 6, no. 7414):** 57–74.

Engels, W. R. 1983. Evolution of altruistic behavior by kin selection: an alternative approach. *Proc. Natl. Acad. Sci.* **80:** 515–518.

Eshel, I., and M. W. Feldman. 1970. On the evolutionary effect of recombination *Theoretical Population Biology* **1:** 88–100.

Ethier, S. N., and M. F. Norman. 1977. Error estimates for the diffusion approximation of the Wright-Fisher model. *Proc. Natl. Acad. Sci.* **74:** 5096–5098.

Evans, S. N., Y. Shvets, and M. Slatkin. 2007. Non-equilibrium theory of the allele frequency spectrum *Theoretical Population Biology* **71:** 109-âĂŞ119. doi:10.1016/j.tpb.2006.06.005

Ewens, W. J. 1963. The mean time for absorption in a process of genetic type. *Journal of the Australian Mathematical Society* **3:** 375–383.

Ewens, W. J. 1963. The diffusion equation and a pseudo-distribution in genetics. *Journal of the Royal Statistical Society B* **25:** 405–412.

Ewens, W. J. 1964. The maintenance of alleles by mutation. *Genetics* **50:** 891–898.

Ewens, W. J. 1969. *Population Genetics*. Methuen, London.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3:** 87–112, 240, 376.

Ewens, W. J. 1974. A note on the sampling theory for infinite sites and infinite states models. *Theoretical Population Biology* **6:** 143–148.

Ewens, W. J. 1979. *Mathematical Population Genetics*. Biomathematics, vol. 9. Springer-Verlag, Berlin.

Ewens, W. J. 1982. On the concept of the effective population size. *Theoretical Population Biology* **21:** 373–378.

Ewens, W. J. 1989. An interpretation and proof of the fundamental theorem of natural selection. *Theoretical Population Biology* **36:** 167–180. doi:10.1016/0040-5809(89)90028-2

Ewens, W. J. 2004. *Mathematical Population Genetics. I. Theoretical Introduction.* 2nd edition. Springer-Verlag, New York.

Eyland, E. A. 1971. Moran's island migration model. *Genetics* **69:** 399–403.

Falconer, D. S. 1965. Maternal effects and selection response. pp. 763–774 in *Genetics Today, Proc. XI International Congress of Genetics*, Vol. 3, ed. S. J. Geerts. Pergamon Press, Oxford.

Falconer, D. S. and T. F. C. MacKay. 1996. *Introduction to Quantitative Genetics, 4th ed.* Longman Scientific and Technical, Burnt Mill, Harlow, Essex.

Fearnhead, P. 2006. The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theoretical Population Biology* **70:** 376-âĂŞ386. doi:10.1016/j.tpb.2006.02.001

Feldman, M. W., and F. B. Christiansen. 1974. The effect of population subdivision on two loci without selection. *Genetical Research* **24:** 151–162.

Feldman, M. W., 1972. Selection for linkage modification: I. Random mating populations. *Theoretical Population Biology* **3:** 324–346.

Feldman, M. W., and R. C. Lewontin. 1975. The heritability hang-up. *Science* **190:** 1163–1168.

Feldman, M. W., and L. L. Cavalli-Sforza. 1976. Cultural and biological evolutionary processes, selection for a trait under complex transmission. *Theoretical Population Biology* **9:** 238–259.

Feldman, M. W., and L. L. Cavalli-Sforza. 1977. The evolution of continuous variation. II. Complex transmission and assortative mating. *Theoretical Population Biology* **11:** 161–181.

Feldman, M. W. and U. Liberman. 1986. An evolutionary reduction principle for genetic modifiers. *Proc. Natl. Acad. Sci.* **83:** 4824–4827.

Feldman, M. W., S. P. Otto, and F. B. Christiansen. 1996. Population genetic perspectives on the evolution of recombination. *Annual Review of Genetics* **30:** 261–295.

Feller, W. 1951. Diffusion processes in genetics. pp. 227–246 in *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. University of California Press, Berkeley and Los Angeles.

Feller, W. 1952. The parabolic differential equations and the associated semi-groups of transformations. *Annals of Mathematics* **55:** 468–519.

Feller, W. 1954. Diffusion processes in one dimension. *Transactions of the American Mathematical Society* **77:** 1–31.

Felsenstein, J. 1965. The effect of linkage on directional selection. *Genetics* **52:** 349–363.

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68:** 581–597.

Felsenstein, J. 1971. The rate of loss of multiple alleles in finite haploid populations. *Theoretical Population Biology* **2:** 391–403.

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics* **78:** 737–756.

Felsenstein, J. 1975. A pain in the torus: some difficulties with models of isolation by distance. *American Naturalist* **109:** 359–368.

Felsenstein, J. 1976. The theoretical population genetics of variable selection and migration. *Annual Review of Genetics* **10:** 253–280.

Felsenstein, J. and S. Yokoyama. 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. *Genetics* **83:** 845–859. (Corrigendum **85:** 372)

Felsenstein, J. 1977. Multivariate normal genetic models with a finite number of loci. pp. 227–246 in *Proceedings of the International Conference on Quantitative Genetics, August 16–21, 1976*, ed. E. Pollak, O. Kempthorne, and T. B. Bailey, Jr. Iowa State University Press, Ames, Iowa.

Felsenstein, J. 1979. r- and K-selection in a completely chaotic population model. *American Naturalist* **113:** 499–510.

Felsenstein, J. 1979. Excursions along the interface between disruptive and stabilizing selection. *Genetics* **93:** 773–795.

Felsenstein, J. 1985. Phylogenies from gene frequencies: a statistical problem. *Systematic Zoology* **34:** 300–311.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22:** 521–565.

Felsenstein, J. 1992a. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59:** 139–147.

Felsenstein, J. 1992b. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo approach. *Genetical Research* **60:** 209–220.

Felsenstein, J., M. K. Kuhner, J. Yamato, and P. Beerli. 1999. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. pp. 163–185 in *Statistics in Molecular Biology and Genetics*, ed. F. Seillier-Moiseiwitsch. IMS Lecture Notes-Monograph Series, volume 33. Institute of Mathematical Statistics and American Mathematical Society, Hayward, California.

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

Felsenstein, J. 2015. Covariation of gene frequencies in a stepping-stone lattice of populations *Theoretical Population Biology* **100:** 88-âĂŞ97. doi:10.1016/j.tpb.2014.12.004

Fincham, J. R. S. 1972. Heterozygous advantage as a likely general basis for enzyme polymorphism. *Heredity* **28:** 387–391.

Fish, H. D. 1914. On the progressive increase of homozygosis in brother-sister matings. *American Naturalist* **48:** 759–761.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52:** 399–433.

Fisher, R. A. 1922. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42:** 321–341.

Fisher, R. A. 1928. The possible modification of the response of the wild type to recurrent mutation. *American Naturalist* **62:** 115–126.

Fisher, R. A. 1929. The evolution of dominance: reply to Professor Sewall Wright. *American Naturalist* **63:** 553–556.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford (Second Edition, Dover, New York 1958).

Fisher, R. A. 1937. The wave of advance of advantageous genes. *Annals of Eugenics* **7:** 355–369.

Fisher, R. A. 1949. *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

Fisher, R. A. 1950. Gene frequencies in a cline determined by selection and diffusion. *Biometrics* **6:** 353–361.

Fleming, W. H. 1975. A selection-migration model in population biology. *Journal of Mathematical Biology* **2:** 219–233.

Franklin, I., and R. C. Lewontin. 1970. Is the gene the unit of selection? *Genetics* **65:** 707–734.

Fu Y.-X., and W.-H. Li. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution* **14:** 195–199.

Fu, Y.-X. 2006. Exact coalescent for the WrightâĂŞFisher model. *Theoretical Population Biology* **69:** 385-âĂŞ394. doi:10.1016/j.tpb.2005.11.005

Galton, F. 1889. *Natural Inheritance*. Macmillan, London.

Galton, F. 1897. The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society* **61:** 401–413.

Geiringer, H. 1944. On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics* **15:** 25–27.

Geiringer, H. 1948. Further remarks on linkage theory in Mendelian heredity. *Annals of Mathematical Statistics* **16:** 390–393.

Geppert, H., and S. Koller. 1938. *Inheritance Mathematics* (German). Leipzig.

Gillespie, J. H. 1973. Natural selection with varying selection coefficients: a haploid model. *Genetical Research* **21:** 115–120.

Gillespie, J. H. 1973. Polymorphism in random environments. *Theoretical Population Biology* **4:** 193–195.

Gillespie, J. H. and C. H. Langley. 1974. A general model to account for enzyme variation in natural populations. *Genetics* **76:** 837–848.

Gillespie, J. H. 1978. A general model to account for enzyme variation in natural populations. V. The SAS-CFF model. *Theoretical Population Biology* **14:** 1–45.

Gillespie, J. H. 1979. Molecular evolution and polymorphism in a random environment. *Genetics* **93:** 737–754.

Gillespie, J. H. 1982. A randomized SAS-CFF model of natural selection in a random environment. *Theoretical Population Biology* **21:** 219–237.

Gillespie, J. H. 1991. *The Causes of Molecular Evolution. Oxford Series in Ecology and Evolution, no. 2.* Oxford University Press, Oxford, U.K.

Gillespie, J. H. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155 (2):** 909-919.

Gillespie, J. H. 2001. Is the population size of a species relevant to its evolution? *Evolution* **55 :** 2161âĂŞ2169.

Gillois, M. 1964. The relation of identity in genetics (French). Dr. Natural Sciences Thesis, Université de Paris, Paris.

Gillois, M. 1965. The relation of identity in genetics. I. Mendelian axioms and postulates (French). *Annales de l'Institut Henri Poincare, Sect. B* **2:** 1–38.

Glass, B. and C. C. Li. 1953. The dynamics of racial intermixture: an analysis based on the American Negro. *American Journal of Human Genetics* **5:** 1–20.

Gliddon, C., and C. Strobeck. 1975. Necessary and sufficient conditions for multiple-niche polymorphism in haploids. *American Naturalist* **109:** 233–235.

Gregorius, H.-R. 1982. Selection in diplo-haplonts. *Theoretical Population Biology* **21:** 289–300.

Griffing, B. 1960. Theoretical consequences of truncation selection based on the individual phenotype. *Australian Journal of Biological Sciences* **13:** 307–343.

Griffiths, R. C. 1989. Genealogical-tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology* **27:** 667–680.

Griffiths, R. C., and S. Tavaré. 1994. Sampling theory for neutral alleles in a varying

environment. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **344:** 403–410.

Griffiths, R. C. and P. Marjoram. 1997. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3:** 479–502.

Gutenkunst, R., R. D. Hernandez, S. H. Williamson, and C. D. Bustamente. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5 (10):** e1000695.

Hadeler, K. P. 1976. Travelling population fronts. pp. 585–592 in Population Genetics and Ecology, ed. S. Karlin and E. Nevo. Academic Press, New York.

Hagedoorn, A. L., and A. C. Hagedoorn. 1921. *The Relative Value of the Processes Causing Evolution*. M. Nijhoff, The Hague.

Haldane, J. B. S. 1924. A mathematical theory of natural and artificial selection. Part I. *Transactions of the Cambridge Philosophical Society* **23:** 19–41.

Haldane, J. B. S. 1926. A mathematical theory of natural and artificial selection. Part III. *Proceedings of the Cambridge Philosophical Society* **23:** 363–372.

Haldane, J. B. S. 1926. A mathematical theory of natural and artificial selection. Part IV. *Proceedings of the Cambridge Philosophical Society* **23:** 607–615.

Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection. Part V: Selection and mutation. *Proceedings of the Cambridge Philosophical Society* **23:** 838–844.

Haldane, J. B. S. 1930a. A note on Fisher's theory of the origin of dominance, and on a correlation between dominance and linkage. *American Naturalist* **64:** 87–90.

Haldane, J. B. S. 1930b. A mathematical theory of natural and artificial selection (Part VI, Isolation). *Proceedings of the Cambridge Philosophical Society* **26:** 220–230.

Haldane, J. B. S. 1930c. A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate. *Proceedings of the Cambridge Philosophical Society* **27:** 131–136.

Haldane, J. B. S. 1932. *The Causes of Evolution*. Longmans, Green, and Co., London (Cornell University Press, Ithaca, New York 1966).

Haldane, J. B. S. 1933. The part played by recurrent mutation in evolution. *American Naturalist* **67:** 5–19.

Haldane, J. B. S. 1935. The rate of spontaneous mutation of a human gene. *Journal of Genetics* **31:** 317–326.

Haldane, J. B. S. 1937. The effect of variation on fitness. *American Naturalist* **71:** 337–349.

Haldane, J. B. S. 1939. The equilibrium between mutation and random extinction. *Annals of Eugenics* **9:** 400–405.

Haldane, J. B. S. 1940. The conflict between selection and mutation of harmful recessive genes. *Annals of Eugenics* **10:** 417–421.

Haldane, J. B. S. 1948. The theory of a cline. *Journal of Genetics* **48:** 277–284.

Haldane, J. B. S. and S. D. Jayakar. 1963. Polymorphism due to selection of varying direction. *J. Genetics* **58:** 237–242.

Hallatschek, O., and D, R. Nelson. 2008. Gene surfing in expanding populations *Theoretical Population Biology* **73:** 158-âĂŞ170. doi:10.1016/j.tpb.2007.08.008

Hamilton, W. D. 1963. The evolution of altruistic behavior. *American Naturalist* **97:** 354–356.

Hamilton, W. D. 1964a. The genetical evolution of social behavior. I. *Journal of Theoretical Biology* **7:** 1–16.

Hamilton, W. D. 1964b. The genetical evolution of social behavior. II. *Journal of Theoretical Biology* **7:** 17–52.

Hamilton, W. D. 1971. Selection of selfish and altruistic behavior in some extreme models. pp. 57–91 in *Man and Beasts. Comparative Social Behavior*, ed. J. F. Eisenberg and W. S. Dillon. Smithsonian Press, Washington, D.C.

Hansen, T. F. 1992. Evolution of stability parameters in single-species population models: stability or chaos? *Theoretical Population Biology* **42:** 199–217. doi:10.1016/j.tpb.2009.12.003

Hanson, W. D. 1966. Effects of partial isolation (distance), migration, and different fitness requirements among environmental pockets upon steady state gene frequencies. *Biometrics* **22:** 453–468.

Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science* **28:** 49–50.

Hartl, D. L. and R. B. Campbell. 1982. Allele multiplicity in simple Mendelian disorders. *American Journal of Human Genetics* **34:** 866–873.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22:** 160–174.

Hastings, A. 1981. Stable cycling in discrete-time genetic models. *Proc. Natl. Acad. Sci.* **78:** 7224–7225.

Hazel, L. N., and J. L. Lush. 1942. The efficiency of three methods of selection. *Journal of Heredity* **33:** 393–399.

Hazel, L. N. 1943. The genetic basis for constructing selection indexes. *Genetics* **28:** 476–490.

Hedrick, P. W., M. Ginevan, and E. P. Ewing. 1976. Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics* **7:** 1–32.

Hedrick, P. W. 1986. Genetic polymorphism in heterogeneous environments: a decade later. *Annual Review of Ecology and Systematics* **17:** 535–566.

Hein, J., M. Schierup, and C. Wiuf. 2004. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory.* Oxford University Press, Oxford.

Heuch, I. 1975. The relationship between separation time and genetic distance based on angular transformations of gene frequencies. *Biometrics* **31:** 685–700.

Heyde, C. C., and E. Seneta. 1977. *I. J. Bienaymé, Statistical Theory Anticipated.* Springer-Verlag, New York.

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* **8:** 269–294.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38:** 226–231.

Hill, W. G. 1971. Design and effectiveness of selection experiments for estimating genetic parameters. *Biometrics* **27:** 293–311.

Hill, W. G. 1972. Estimation of realised heritabilities from selection experiments. I. Divergent selection. *Biometrics* **28:** 747–765.

Hill, W. G. 1972. Estimation of realised heritabilities from selection experiments. II. Selection in one direction. *Biometrics* **28:** 767–780.

Hill, W. G. 1972. Effective size of populations with overlapping generations. *Theoretical Population Biology* **3:** 278–289. Hill, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33:** 229–239.

Hill, W. G. and F. W. Nicholas. 1974. Estimation of heritability by both regression of offspring on parent and intra-class correlation of sibs in one experiment. *Biometrics* **30:** 447–468.

Hill, W. G. 1976. Order statistics of correlated variables and implications in genetic selection programmes. *Biometrics* **32:** 889–902.

Hill, W. G., and B. S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33:** 54–78.

Hoekstra, R. F. 1975. A deterministic model of cyclical selection. *Genetical Research* **25:** 1–15.

Holsinger, K. E. and M. W. Feldman. 1983. Modifiers of mutation rate: Evolutionary optimum with complete selfing. *Proc. Natl. Acad. Sci.* **80:** 6732–6734.

Hopf, F. A., R. E. Michod, and M. J. Sanderson. 1988. The effect of the reproductive system on mutation load. *Theoretical Population Biology* **33:** 243–265.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23:** 183–201.

Hudson, R. R. 2001. Two-locus sampling distributions and their application. *Genetics 2001* **159:** 1805–1817.

Hull, F. H. 1945. Recurrent selection for specific combining ability in corn. *Journal of the American Society of Agronomy* **37:** 134–145.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437 (7063):** 1299–1320.

Jacquard, A. 1974. *The Genetic Structure of Populations*. Springer-Verlag, Berlin, Heidelberg, and New York.

Jenkin, F. 1867. Origin of Species. *North British Review* **46:** 277–318.

Jennings, H. S. 1914. Formulae for the results of inbreeding. *American Naturalist* **48:** 693–696.

Jennings, H. S. 1916. Numerical results of diverse systems of mating. *Genetics* **1:** 53–89.

Jennings, H. S. 1917. The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. *Genetics* **2:** 97–154.

Jensen, L., and E. Pollak. 1969. Random selective advantages of a gene in a finite population. *Journal of Applied Probability* **6:** 19–37.

Jensen, L. 1973. Random selective advantages of genes and their probabilities of fixation. *Genetical Research* **21:** 215–219.

Jones, D. F. 1917. Dominance of linked factors as a means of accounting for heterosis. *Genetics* **2:** 466–479.

Jukes, T. H. and C. Cantor. 1969. Evolution of protein molecules. pp. 21–132 in *Mammalian Protein Metabolism,* ed. M. N. Munro. Academic Press, New York.

Karigl, G. 1982. A mathematical approach to multiple genetic relationships. *Theoretical Population Biology* **21:** 379–393.

Karlin, S., and J. McGregor. 1964. On some stochastic models in genetics. pp. 245–279 in *Stochastic Models in Medicine and Biology*, ed. J. Gurland. University of Wisconsin Press, Madison.

Karlin, S. 1966. *A First Course in Stochastic Processes.* Academic Press, New York.

Karlin, S. and M. W. Feldman. 1969. Linkage and selection: new equilibrium properties of the two-locus symmetric viability model. *Proc. Natl. Acad. Sci.* **62:** 70–74.

Karlin, S., and M. W. Feldman. 1970. Linkage and selection: two locus symmetric viability model. *Theoretical Population Biology* **1:** 39–71.

Karlin, S. and J. McGregor. 1972a. Application of method of small parameters to multi-niche population genetic models *Theoretical Population Biology* **3:** 186–209.

Karlin, S. and J. McGregor. 1972b. Polymorphisms for genetic and ecological systems with weak coupling. *Theoretical Population Biology* **3:** 210–238.

Karlin, S., and U. Lieberman. 1974. Random temporal variation in selection intensities: case of large population size. *Theoretical Population Biology* **6:** 355–382.

Karlin, S., and B. Levikson. 1974. Temporal fluctuations in selection intensities: case of small population size. *Theoretical Population Biology* **6:** 383–412.

Karlin, S. 1976. Population subdivision and selection migration interactions, pp. 617–657 in *Population Genetics and Ecology*, ed. S. Karlin and E. Nevo, Academic Press. New York.

Karlin, S. 1977. Gene frequency patterns in the Levene subdivided population model. *Theoretical Population Biology* **11:** 356–385.

Karlin, S. 1982. Classifications of selection-migration structures and conditions for a protected polymorphism. pp. 61–204 in *Evolutionary Biology, Vol. 14*, ed. M. K. Hecht, B. Wallace, and G. T. Prance. Plenum Press, New York.

Kempthorne, O. 1954. The correlation between relatives in a random mating population. *Proceedings of the Royal Society B* **143:** 103–113.

Kempthorne, O. 1957. *An Introduction to Genetic Statistics*. John Wiley, New York.

Kerr, B., P. Godfrey-Smith, and M. W. Feldman. 2004. What is altruism? Trends in Ecology and Evolution **19:** 135–140.

Kimura, M. 1953. "Stepping-stone" model of population. *Annual Report of the National Institute of Genetics, Japan* **3:** 62–63.

Kimura, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics* **39:** 280–295.

Kimura, M. 1955a. Stochastic processes and distribution of gene frequencies under natural selections. *Cold Spring Harbor Symposia on Quantitative Biology* **20:** 33–53.

Kimura, M. 1955b. Random drift in a multi-allelic locus. *Evolution* **9:** 419–435.

Kimura, M. 1955c. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci.* **41:** 144–150.

Kimura, M. 1956a. Random genetic drift in a tri-allelic locus: exact solution with a continuous model. *Biometrics* **12:** 57–66.

Kimura, M. 1956b. A model of a genetic system which leads to closer linkage by natural selection. *Evolution* **10:** 278–287.

Kimura, M. 1956c. Rules for testing stability of a selective polymorphism. *Proc. Natl. Acad. Sci.* **42:** 336–340.

Kimura, M. 1957. Some problems of stochastic processes in genetics. *Annals of Mathematical Statistics* **28:** 882–901.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* **47:** 713–719.

Kimura, M., and J. F. Crow. 1963. The measurement of effective population number. *Evolution* **17:** 279–288.

Kimura, M., and J. F. Crow. 1964. On the maximum avoidance of inbreeding. *Genetical Research* **4:** 399–415.

Kimura, M., and G. H. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49:** 561–576.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* **49:** 725–738.

Kimura, M. 1965. A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proc. Natl. Acad. Sci.* **54:** 731–736.

Kimura, M. 1968a. Evolutionary rate at the molecular level. *Nature* **217:** 624–626.

Kimura, M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11:** 247–269.

Kimura, M. 1969a. Natural selection and gene substitution. *Genetical Research* **13:** 127–141.

Kimura, M., and T. Ohta. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61:** 763–771.

Kimura, M. 1969b. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics* **61:** 893–903.

Kimura, M., and T. Ohta. 1969. The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* **63:** 701–709.

Kimura, M., and T. Maruyama. 1971. Pattern of neutral polymorphism in a geographically structured population. *Genetical Research* **18:** 125–131.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution **16:** 111–120.

King, J. L. 1967. Continuously distributed factors affecting fitness. *Genetics* **55:** 483–492.

King, J. L., and T. Jukes. 1969. Non-Darwinian evolution: random fixation of selectively neutral mutations. *Science* **164:** 788–798.

Kingman, J. F. C. 1961. A mathematical problem in population genetics. *Proc. Cambridge Philosophical Society* **57:** 574–582.

Kingman, J. F. C. 1961. On an inequality in partial averages. *Quarterly Journal of Mathematics. Oxford Series* **12:** 78–80. doi: 10.1093/qmath/12.1.78

Kingman, J. F. C. 1982a. The coalescent. *Stochastic Processes and their Applications* **13:** 235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability* **19A:** 27–43.

Kingman, J. F. C. 1982c. Exchangeability and the evolution of large populations. pp. 97–112 in *Exchangeability in Probability and Statistics. Proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th-9th April, 1981, in honour of Professor Bruno de Finetti.* ed. G. Koch and F. Spizzichino. North-Holland Elsevier, Amsterdam.

Kolmogorov, A. N. 1931. On analytic methods in probability calculation (German). *Mathematische Annalen* **104:** 415–458.

Kolmogorov, A. N. 1935. Deviations from Hardy's formula in partial isolation. *Comptes Rendus (Doklady) de l'Academie des Sciences de l'U. R. S. S. Nouvelle Serie* **3:** 129–132.

Kolmogorov, A. N., I. G. Petrovskiy, and N. S. Piskunov. 1937. The study of the diffusion equation with increase of the quantity of matter and its application to a biological problem (Russian, French). *Byulleten' Moskovskogo Gosudarsvennogo Universiteta (A)* **1:** 1–6, *Bulletin de l'Université d'Etat a Moscou (Serie Internationale) A* **1:** 1–25.

Kolmogorov, A. 1938. On the solution of a biological problem (German, Russian). *Izvestiya Naukno-Issledelvatel'skogo Instituta Matematiki i Mekhaniki pri Tomskom Gosudarstvennom Universitete* **2:** 1–6 (German), 7–12 (Russian).

Korolev, K. S, M. Avlund, O. Hallatschek, and D. R. Nelson. 2010. Genetic demixing and evolution in linear stepping stone models. *Reviews of Modern Physics* **82(2):** 1691–1718.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140:** 1421–1430.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149:** 429–434.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156:** 1393–1401.

Laland, K. N., J. Odling-Smee, and M. W. Feldman. 1999. Evolutionary consequences of niche construction and their implications for ecology. *Proc. Natl. Acad. Sci.* **96:** 10242–10247.

Laland, K. N. and G. R. Brown. 2006. Niche construction, human behavior, and the adaptive lag hypothesis. *Evolutionary Anthropology* **15:** 95–104.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* **20:** 86–93.

Lande, R. 1976a. Natural selection and genetic drift in phenotypic evolution. *Evolution* **30:** 314–324.

Lande, R. 1976b. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetical Research* **26:** 221–235.

Lande, R. 1977. The influence of the mating system on the maintenance of genetic variability in polygenic characters. *Genetics* **86:** 485–498.

Latter, B. D. H. 1965. The response to artificial selection due to genes of large effect. I. Changes in gene frequency at an additive locus. *Australian Journal of Biological Sciences* **18:** 585–598.

Latter, B. D. H. 1973. The island model of population differentiation: a general solution. *Genetics* **73:** 147–157.

Leigh, E. G., Jr. 1970. Natural selection and mutability. *American Naturalist* **104:** 301–305.

León, J. A. and B. Charlesworth. 1978. Ecological versions of Fisher's Fundamental Theorem of Natural Selection. *Ecology* **59:** 457–464.

Lessard, S. 1997. Fisher's fundamental theorem of natural selection revisited. *Theoretical Population Biology* **52:** 119-âĂŞ136. doi:10.1006/tpbi.2002.1570

Lessard, S. and S. Karlin, 1982. A criterion for stability-instability at fixation states involving an eigenvalue one with applications in population genetics. *Theoretical Population Biology* **22:** 108–126.

Levene, H. 1949. On a matching problem arising in genetics. *Annals of Mathematical Statistics* **20 (1):** 91-94.

Levene, H. 1953. Genetic equilibrium when more than one ecological niche is available. *American Naturalist* **87:** 331–333.

Levitt, P. R. 1975. General kin selection models for genetic evolution of sib altruism in diploid and haplodiploid species. *Theoretical Population Biology* **4:** 85–128.

Lewontin, R. C. 1958. A general method for investigating the equilibrium of gene frequency in a population. *Genetics* **43:** 419–433.

Lewontin, R. C. and K. Kojima. 1960. The evolutionary dynamics of complex polymorphisms. *Evolution* **14:** 458–472.

Lewontin, R. C. 1964a. The interaction of linkage and selection. I. General considerations; heterotic models. *Genetics* **49:** 49–67.

Lewontin, R. C. 1964b. The interaction of linkage and selection. II. Optimum models. *Genetics* **50:** 757–782.

Lewontin, R. C. and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54:** 595–609.

Lewontin, R. C. 1974. *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York and London.

Lewontin, R. C. and M. W. Feldman. 1988. A general asymptotic property of two-locus selection models. *Theoretical Population Biology* **34:** 177–193.

Li, C. C., and L. Sacks. 1954. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10:** 347–360.

Li, C. C. 1955. The stability of an equilibrium and the average fitness of a population. *American Naturalist* **89:** 281–295.

Li, W.-H. and M. Nei. 1974. Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology* **6:** 173–183.

Li, W.-H. (editor). 1977a. *Stochastic Models in Population Genetics*. Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania.

Li, W.-H. 1977b. Maintenance of genetic variability under mutation and selection pressures in a finite population. *Proceeding of the National Academy of Sciences, USA* **74:** 2509–2513.

Liberman, U. and M. W. Feldman. 1986. Modifiers of mutation rate: a general reduction principle. *Theoretical Population Biology* **30:** 125–142.

Lin, N. and C. D. Michener. 1972. Evolution of sociality in insects. *Quarterly Review of Biology* **47:** 313–159.

Lush, J. L. 1937. *Animal Breeding Plans*. Iowa State College Press, Ames, Iowa (Third Edition, 1945).

Lush, J. L. 1947. Family merit and individual merit as bases for selection. Part I. *American Naturalist* **81:** 241–261.

Lush, J. L. 1947. Family merit and individual merit as bases for selection. Part II. *American Naturalist* **81:** 362–379.

Lynch, M. and J. B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, Massachusetts.

MacArthur, R. H. 1962. Some generalized theorems of natural selection. *Proc. Natl. Acad. Sci.* **48:** 1893–1897.

Malécot, G. 1939. *Mathematical Theories of Generalized Mendelian Heredity* (French). Guilhot, Paris.

Malécot, G. 1948. *The Mathematics of Heredity* (French). Masson, Paris.

Malécot, G. 1950. Some probability schemes for the variability of natural populations (French). *Annales de l'Université de Lyon, Sciences, Section A* **13:** 37–60.

Malécot, G. 1951. A stochastic treatment of linear problems (mutation, linkage, migration) in population genetics (French). *Annales de l'Université de Lyon, Sciences, Section A* **14:** 79–117.

Malécot, G. 1969. *The Mathematics of Heredity*. W. H. Freeman, San Francisco.

Mange, A. P. 1964. FORTRAN programs for computing Wright's coefficient of inbreeding in human and nonhuman pedigrees. *American Journal of Human Genetics* **16:** 484.

Mange, A. P. 1969. Wright's coefficient of inbreeding, F, for human pedigrees. pp. 72–81 in *Computer Applications in Genetics*, ed. N. E. Morton. University of Hawaii Press, Honolulu.

Maruyama, T. 1970. On the rate of decrease of heterozygosity in circular stepping stone models. *Theoretical Population Biology* **1:** 101–119.

Maruyama, T. 1970. Effective number of alleles in a subdivided population. *Theoretical Population Biology* **1:** 273–306.

Maruyama, T. 1972. Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics* **70:** 639–651.

Maruyama, T. 1973. A diffusion model for geographically structured populations. pp. 40–44 in *Genetic Structure of Populations*, ed. N. E. Morton. University of Hawaii Press, Honolulu.

Maruyama, T. 1974. A simple proof that certain quantities are independent of the geographic structure of a population. *Theoretical Population Biology* **5:** 148–154.

Maruyama, T. 1977. *Stochastic Problems in Population Genetics.* Lecture Notes in Biomathematics No. 17. Springer-Verlag, Berlin, Heidelberg, and New York.

Matessi, C. and S. D. Jayakar. 1976. Conditions for the evolution of altruism under Darwinian selection. *Theoretical Population Biology* **9:** 360–387.

Mather, K. 1969. Selection through competition. I. *Heredity* **24:** 529–540.

May, R. M. 1972. On the relationship among various types of population models. *American Naturalist* **107:** 46–57.

May, R. M. 1974. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* **186:** 645–647.

May, R. M., J. Endler, and R. E. McMurtrie. 1975. Gene frequency clines in the presence of selection opposed by gene flow. *American Naturalist* **109:** 659–676.

May, R. M. 1976. Simple mathematical models with very complicated dynamics. *Nature* **261:** 459–476.

May, R. M. and G. F. Oster. 1976. Bifurcations and dynamic complexity in simple ecological models. *American Naturalist* **110:** 573–599.

Maynard Smith, J. 1962. Disruptive selection, polymorphism and sympatric speciation. *Nature* **195:** 60–62.

Maynard Smith, J. 1968. Evolution in sexual and asexual populations. *American Naturalist* **102:** 469–473.

Maynard Smith, J. 1970. Population size, polymorphism, and the rate of non-Darwinian evolution. *American Naturalist* **104:** 231–237.

Maynard Smith, J. 1970. Genetic polymorphism in a varied environment. American Naturalist **104:** 487–490.

Maynard Smith, J. 1971. What use is sex? *Journal of Theoretical Biology* **30:** 319–335.

Maynard Smith, J. and G. R. Price. 1973. The logic of animal conflict. *Nature* **246: 15–18**. doi:10.1038/246015a0

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* **23:** 23–35.

Maynard Smith, J. 1974. The theory of games and the evolution of animal conflict. *Journal of Theoretical Biology* **47:** 209-âĂŞ221.

Maynard Smith, J. 1981. Will a sexual population evolve to an ESS? *American Naturalist* **117:** 1015–1018.

Maynard Smith, J. 1982. *Evolution and the Theory of Games.* Cambridge University Press, Cambridge, U.K.

McCandlish, D. M., C. L. Epstein, and J. B. Plotkin. 2014. Formal properties of the probability of fixation: Identities, inequalities and approximations. *Theoretical Population Biology* 99: 98-113. dx.doi.org/10.1016/j.tpb.2014.11.004

Milkman, R. D. 1967. Heterosis as a major cause of heterozygosity in nature. *Genetics* **55:** 493–495.

Mitton, J. B., and R. K. Koehn. 1973. Population genetics of marine pelecypods. III. Epistasis between functionally related isoenzymes of *Mytilus edulis* (with an appendix by T. Prout). *Genetics* **73:** 487–496.

Moran, P. A. P. 1958. Random processes in genetics. *Proc. Cambridge Philosophical Society* **54:** 60–71.

Moran, P. A. P., and G. A. Watterson. 1959. The genetic effects of family structure in natural populations. *Australian Journal of Biological Sciences* **12:** 1–15.

Moran, P. A. P. 1959. The theory of some genetical effects of population subdivision. *Australian Journal of Biological Sciences* **12:** 109–116.

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

Moran, P. A. P. 1964. On the nonexistence of adaptive topographies. *Annals of Human Genetics* **27:** 383–393.

Morton, N. E., J. F. Crow, and H. J. Muller. 1956. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl. Acad. Sci.* **42:** 855–863.

Muller, H. J. 1918. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3:** 422–499.

Muller, H. J. 1932. Some genetic aspects of sex. *American Naturalist* **66:** 118–138.

Muller, H. J. 1950. Our load of mutations. *American Journal of Human Genetics* **2:** 111–176.

Muller, H. J. 1950. Evidence of the precision of genetic adaptation. pp. 165–229 in *Harvey Lectures, Series 18*. Charles C. Thomas, Springfield, Illinois.

Muller, H. J. 1958. Evolution by mutation. *Bulletin of the American Mathematical Society* **64:** 137–160.

Muller, H. J. 1964. The relation of recombination to mutational advance. *Mutation Research* **1:** 2–9.

Myers, S., C. Fefferman, and N. Patterson. 2008. Can one learn history from the allelic spectrum? *Theoretical Population Biology* **73:** 342-âĂŞ348. doi:10.1016/j.tpb.2008.01.001

Nagylaki, T. and J. F. Crow. 1974. Continuous selective models. *Theoretical Population Biology* **5:** 257–283.

Nagylaki, T. 1975. Conditions for the existence of clines. *Genetics* **80:** 595–615.

Nagylaki, T. 1975. A continuous selective model for an X-linked locus. *Heredity* **34:** 273–278.

Nagylaki, T. 1976. The decay of genetic variability in geographically structured populations. *Theoretical Population Biology* **10:** 70–82.

Nagylaki, T. 1977a. Decay of genetic variability in geographically structured populations. *Proceedings of the National Academy of Sciences, USA* **74(6):** 2523–2325.

Nagylaki, T. 1977b. *Selection in One- and Two-Locus Systems*. Lecture Notes in Biomathematics No. 15. Springer-Verlag, Berlin.

Nagylaki, T. and V. Barcilon, 1988. The influence of spatial inhomogeneities on neutral models of geographic variation. II. The semi-infinite linear habitat. *Theoretical Population Biology* **33:** 311–343.

Nei, M. and M. Murata. 1966. Effective population size when fertility is inherited. *Genetical Research* **8:** 257–260.

Nei, M., and Y. Imaizumi. 1966. Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. *Heredity* **21:** 183–190, 344.

Nei, M. 1968. The frequency distribution of lethal chromosomes in finite populations. *Proc. Natl. Acad. Sci.* **60:** 517–524.

Nei, M. 1971. Fertility excess necessary for gene substitution in regulated populations. *Genetics* **68:** 169–184.

Nei, M. 1972. Genetic distance between populations. *American Naturalist* **106:** 283–292.

Nei, M., and M. W. Feldman. 1972. Identity of genes by descent within and between populations under mutation and migration pressures. *Theoretical Population Biology* **3:** 460–465.

Nei, M. 1975. *Molecular Population Genetics and Evolution*. American Elsevier, New York. North Holland, Amsterdam.

Nei, M., T. Maruyama, and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. *Evolution* **29:** 1–10.

Nei, M. 1978. *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Norman, M. F. 1975. Diffusion approximation of non-Markovian processes. *Annals of Probability* **3:** 358–364.

Norman, M. F. 1975. An ergodic theorem for evolution in a random environment. *Journal of Applied Probability* **12:** 661–672.

Norton, H. T. J. 1915. Appendix to *Mimicry in Butterflies* by R. C. Punnett. Cambridge University Press, Cambridge.

Norton, H. T. J. 1928. Natural selection and Mendelian variation. *Proc. London Mathematical Society (Series 2)* **28:** 1–45.

Nowak, M. 1990. An evolutionarily stable strategy may be inaccessible. *Journal of Theoretical Biology* **142:** 237–241.

Nowak, M. A., C. E. Tarnita, and E. O. Wilson. 2010. The evolution of eusociality. *Nature* **466:** 1057–1062.

Nowak, M. A. C. E. Tarnita, and E. O. Wilson. 2011. Reply (to Abbot et al. paper and 4 other papers). *Nature* **471:** E9–E4, author reply E9–E10. doi: 10.1038/nature09836.

Ohta, T. 1972. Fixation probability of a mutant influenced by random fluctuation of selection intensity. *Genetical Research* **19:** 33–38.

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

Ohta, T., and M. Kimura. 1969a. Linkage disequilibrium due to random genetic drift. *Genetical Research* **13:** 47–55.

Ohta, T. and M. Kimura. 1969b. Linkage disequilibrium at a steady state determined by random genetic drift and recurrent mutation. *Genetics* **63:** 229–238.

Olsson, F., and O. Hössjer. 2015. Estimation of the variance effective population size in age structured populations. *Theoretical Population Biology* **101:** 9–23. doi:10.1016/j.tpb.2015.02.003

Otto, S. P., and M. W. Feldman. 1997. Deleterious mutations, variable Orlove, M. J. 1979. A reconciliation of inclusive fitness and personal fitness approaches: a proposed correcting term for the inclusive fitness formula. *Journal of Theoretical Biology* **81:** 577–586.

epistatic interactions, and the evolution of recombination. *Theoretical Population Biology* **51:** 134–147. doi:10.1006/tpbi.1997.1301

Pearl, R. 1913. A contribution towards an analysis of the problem of inbreeding. *American Naturalist* **47:** 577–614.

Pearl, R. 1914. On the results of inbreeding a Mendelian population: a correction and extension of previous conclusions. *American Naturalist* **48:** 57–62.

Pearson, K. 1898. Mathematical contributions to the theory of evolution. On the law of ancestral heredity. *Proceedings of the Royal Society* **62:** 386–412.

Pearson, K. 1904. Mathematical contributions to the theory of evolution. XII. On a generalized theory of alternative inheritance, with special reference to Mendel's laws. *Philosophical Transactions of the Royal Society A* **203:** 53–86.

Pearson, K. 1909. The theory of ancestral contributions in heredity. *Proceedings of the Royal Society B* **81:** 219–224.

Pearson, K. 1909. On the ancestral gametic correlations of a Mendelian population mating at random. *Proceedings of the Royal Society B* **81:** 225–229.

Petry, D. 1983. The effect on neutral gene flow of selection at a linked locus. *Theoretical Population Biology* **23:** 300–313.

Price, G. R. 1970. Selection and covariance. *Nature* **227:** 520–521.

Price, G. R. 1972a. Extension of covariance selection mathematics. *Annals of Human Genetics* **36:** 129–140.

Price, G. R. 1972b. Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* **36:** 129-âĂŞ140.

Prout, T. 1968. Sufficient conditions for multiple niche polymorphism. *American Naturalist* **102:** 493–496.

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics.* University of Chicago Press, Chicago.

Puniyani, A. and M. W. Feldman. 2006. A semi-symmetric two-locus model. *Theoretical Population Biology* **69:** 211-âĂŞ215. doi:10.1016/j.tpb.2005.07.004

Punnett, R. C. 1917. Eliminating feeblemindedness. *Journal of Heredity* **8:** 464–465.

Rawlings, J. O. 1976. Order statistics for a special class of unequally correlated multinomial variables. *Biometrics* **32:** 875–887.

Reiersøl, O. 1962. Genetic algebras studied recursively and by means of differential operators. *Mathematica Scandinavica* **10:** 25–44.

Robbins, R. B. 1918. Some applications of mathematics to breeding problems. III. *Genetics* **3:** 375–389.

Robbins, R. B. 1922. Selection through choice of seed from dominant plants of an allogamous population. *Genetics* **7:** 508–512.

Robertson, A., and I. M. Lerner. 1949. The heritability of all-or-none traits: viability of poultry. *Genetics* **34:** 395–411.

Robertson, A. 1960. A theory of limits in artificial selection. *Proceedings of the Royal Society B* **153:** 234–249.

Robertson, A. 1961. Inbreeding in artificial selection programmes. *Genetical Research* **2:** 189–194.

Robertson, A. 1962. Selection for heterozygotes in small populations. *Genetics* **47:** 1291–1300.

Robertson, A. 1964. The effect of non-random mating within inbred lines on the rate of inbreeding. *Genetical Research* **5:** 164–167.

Robertson, A. and W. G. Hill. 1983. Identity for different mutations for deleterious genes. *Nature* **301:** 176.

Roughgarden, J. 1971. Density-dependent natural selection. *Ecology* **52:** 453–468.

Roughgarden, J. 1972. Evolution of niche width. American Naturalist **106:** 683–718.

Roughgarden, J. 1974. The fundamental and realized niche of a solitary population. *American Naturalist* **108:** 232–235.

Roughgarden, J. 1974. Niche width: biogeographic patterns among Anolis lizards. *American Naturalist* **108:** 429–442.

Roughgarden, J. 1979. *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. Macmillan, New York.

Rouzine, I. M., A. Rodrigo, and J. M. Coffin. 2001. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiology and Molecular Biology Reviews* **65(1):** 151–85.

Sacks, J. M. 1967. A stable equilibrium with minimum average fitness. *Genetics* **56:** 705–708.

Scheuer, P. A. G. and S. P. H. Mandel. 1959. An inequality in population genetics. *Heredity* **13:** 519–524.

Schraiber, J. G., and M. J. Landis. 2015. Sensitivity of quantitative traits to mutational effects and number of loci. *Theoretical Population Biology* **102:** 85-âĂŞ93. doi:10.1016/j.tpb.2015.03.005

Schwartz, J. 2008. *In Pursuit of the Gene: From Darwin to DNA*. Harvard University Press, Cambridge, Massachusetts.

Serebrovsky, A. S. 1936. An attempt at a new method of genetical analysis of quantitative characters. *Comptes Rendus (Doklady) de l'Academie des Sciences de l'U.R.S.S., Nouvelle Serie* **2:** 49–51.

Serebrovsky, A. S. 1936. Second variant of the M-S2 triangle method. *Comptes Rendus (Doklady) de l'Academie des Sciences de l'U.R.S.S., Nouvelle Serie* **4:** 21–24.

Shaw, R. G. 1987. Maximum-likelihood approaches applied to quantitative genetics of natural populations. *Evolution* **4:** 812–826.

Shull, G. H. 1908. The composition of a field of maize. *American Breeder's Association Report* **4:** 296–301.

Shull, G. H. 1914. Duplicate genes for capsule form in *Bursa bursa-pastoris*. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **12:** 97–149.

Singh, M. and R. C. Lewontin. 1966. Stable equilibria under optimizing selection. *Proc. Natl. Acad. Sci.* **56:** 1345–1348.

Slatkin, M. 1970. Selection and polygenic characters. *Proc. Natl. Acad. Sci.* **66:** 87–93.

Slatkin, M. 1973. Gene flow and selection in a cline. *Genetics* **75:** 733–756.

Slatkin, M. 1975. Gene flow and selection in a two-locus system. *Genetics* **81:** 787–802.

Slatkin, M., and R. Lande. 1976. Niche width in a fluctuating environment – density independent model. *American Naturalist* **110:** 31–55.

Slatkin, M. and D. Charlesworth. 1978. The spatial distribution of transient alleles in a subdivided population: a simulation study. *Genetics* **89:** 793–810.

Slatkin, M. 1979. Frequency and density-dependent selection on a quantitative character. *Genetics* **93:** 755–771.

Slatkin, M. and J. Maynard Smith. 1979. Models of coevolution. *Quarterly Review of Biology* **54:** 223–263.

Slatkin, M. and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA in stable and exponentially growing populations. *Genetics* **129:** 555–562.

Slatkin, M. and B. Rannala. 1997. The sampling distribution of disease-associated alleles. *Genetics* **147:** 1855–1861.

Slatkin, M. 2002. A vectorized method of importance sampling with applications to models of mutation and migration. *Theoretical Population Biology* **62:** 339–348. doi:10.1016/S0040-5809(02)00007-2

Slatkin, M. 2009. Epigenetic inheritance and the missing heritability problem. *Genetics* **182(3):** 845–850.

Smith, C. A. B. 1957. Counting methods in genetical statistics. *Annals of Human Genetics* **21:** 254–276.

Smith, H. F. 1936. A discriminant function for plant selection. *Annals of Eugenics* **7:** 240–250.

Sokal, R. R., and F. J. Rohlf. 1969. *Biometry*. W. H. Freeman, San Francisco.

Song, Y. S. and Song, J. S. 2007. Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. *Theoretical Population Biology* **71:** 49–60.

Spieth, P. T. 1974. Gene flow and genetic differentiation. *Genetics* **78:** 961–965.

Steinrücken, M., Y. X. R. Wang, and Y. S. Song. 2013. An explicit transition density expansion for a multi-allelic WrightâĂŞFisher diffusion with general diploid selection. *Theoretical Population Biology* **83:** 1-âĂŞ14. doi:10.1016/j.tpb.2012.10.006

Strobeck, C. 1975. Selection in a fine-grained environment. *American Naturalist* **109:** 419–425.

Strobeck, C., and K. Morgan. 1978. The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88:** 829–844.

Strobeck, C. 1984. Estimation of the neutral mutation rate in a finite population from DNA sequence data. *Theoretical Population Biology* **24:** 160–172.

Sved, J. A., T. E. Reed, and W. F. Bodmer. 1967. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics* **55:** 469–481.

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2:** 125–141.

Tajima, F. 1989. Statistical method For testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

Takahata, N. and M. Slatkin. 1990. Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology* **38:** 331–350.

Tal, O., E. Kisdi, and E. Jablonka. 2010. Epigenetic contribution to covariance between relatives. *Genetics* **184(4):** 1037–1050.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10:** 512–526.

Taylor, J. E. 2008. Environmental variation, fluctuating selection and genetic drift in subdivided populations. *Theoretical Population Biology* **74:** 233-âĂŞ250. doi:10.1016/j.tpb.2008.07.005

Thompson, E. A. 1974. Gene identities and multiple relationships. *Biometrics* **30:** 667–680.

Thomson, G. 1977. The effect of a selected locus on linked neutral loci. *Genetics* **85:** 753–788.

Trivers, R. L. 1972. Parental investment and sexual selection. pp. 136–179 in *Sexual Selection and the Descent of Man*, ed. B. G. Campbell. Aldine, Chicago.

Trivers, R. L. 1974. Parent-offspring conflict. *American Zoologist* **14:** 249–264.

Turelli, M. and D. Petry. 1980. Density-dependent selection in a random environment: an evolutionary process that can maintain stable population dynamics. *Proceedings of the National Academy of Sciences, USA* **77:** 7501–7505.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology* **25:** 138–93.

Turner, J. R. G. 1967. The effect of mutation on fitness in a system of two co-adapted loci. *Annals of Human Genetics* **30:** 329–334.

Wade, M. J. 1978. Kin selection: a classical approach and a general solution. *Proc. Natl. Acad. Sci.* **75:** 6154–6158.

Wagener, D. K., and L. L. Cavalli-Sforza. 1975. Ethnic variation in genetic disease: possible roles of hitchhiking and epistasis. *American Journal of Human Genetics* **27:** 384–364.

Wahlund, S. 1928. The combination of populations and the appearance of correlation examined from the standpoint of the study of heredity (German). *Hereditas* **11:** 65–106.

Wakeley, J., and S. Lessard. 2006. Corridors for migration between large subdivided populations, and the structured coalescent. *Theoretical Population Biology* **70:** 412-âĂŞ420. doi:10.1016/j.tpb.2006.06.001

Wakeley, J. 2008. *Coalescent Theory.* Roberts and Company, Greenwood Village, Colorado.

Wallace, B. 1968. *Topics in Population Genetics*. W. W. Norton, New York.

Wallace, B. 1970. *Genetic Load: Its Biological and Conceptual Aspects*. Prentice-Hall, Englewood Cliffs, New Jersey.

Walsh, B. 1983. Conditions for protection of an allele in linear homogeneous stepping stone models. *Theoretical Population Biology* **24:** 78–106.

Waples, R. S. and M. Yokota. 2007. Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175:** 219-âĂŞ233.

Waples, R. S., C. Do, and J. Chopelet. 2011. Calculating N_e and N_e/N in age-structured populations: a hybrid Felsenstein-Hill approach. *Ecology* **92:** 1513–1522. dxdoi.org/10.1890/10-1796.1

Warren, H. C. 1917. Numerical effects of natural selection acting on Mendelian characters. *Genetics* **2:** 305–312.

Watson, H. W., and F. Galton. 1874. On the probability of extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland* **4:** 138–144.

Watterson, G. A. 1962. Some theoretical aspects of diffusion theory in population genetics. *Annals of Mathematical Statistics* **33:** 939–957 (Erratum **34:** 352 1963).

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7:** 256–276.

Weinberg, W. 1908. On the detection of heredity in man (German). *Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg* **64:** 368–382.

Weinberg, W. 1909. On the laws of heredity in man (German). *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **2:** 276–330.

Weir, B. S. and C. C. Cockerham. 1969. Group inbreeding with two linked loci. *Genetics* **63:** 711–742.

Weir, B. S. 1979. Inferences about linkage disequilibrium. *Biometrics* **35:** 235–254.

Weiss, G. H., and M. Kimura. 1965. A mathematical model of the stepping stone model of genetic correlation. *Journal of Applied Probability* **2:** 129–149.

Weiss, G., and A. von Haeseler. 1998. Inference of population history using a likelihood approach. *Genetics* **149:** 1539–1546.

Wentworth, E. N. and B. L. Remick. 1916. Some breeding properties of the generalized Mendelian population. *Genetics* **1:** 608.

West-Eberhard, M. J. 1975. The evolution of social behavior by kin selection. *Quarterly Review of Biology* **50:** 1–33.

Wilke, C. 2005. Quasispecies theory in the context of population genetics. *BMCEvolutionary Biology* **5:**: 44 (doi:10.1186/1471-2148-5-44)

Wilkes, M. V. 1975. How Babbage's dream came true. *Nature* **257:** 541–544.

Williamson, M. H. 1974. The analysis of discrete time cycles. Pp. 17–33 in *Ecological Stability*, ed. M. B. Usher and M. H. Williamson. Chapman and Hall, London.

Wilson, D. S., and E. O. Wilson. 2007. Rethinking the theoretical foundation of sociobiology. *Quarterly Revew of Biology* **82:** 327–348.

Wilson, E. O. 1965. A consistency test for phylogenies based on contemporaneous species. *Systematic Zoology* **14:** 214–220.

Wilson, E. O. 1975. *Sociobiology: the New Synthesis*. Harvard University Press, Cambridge, Massachusetts.

Workman, P. L., B. S. Blumberg, and A. J. Cooper. 1963. Selection, gene migration and polymorphic stability in a U. S. White and Negro population. *American Journal of Human Genetics* **15:** 429–437.

Wright, S. 1921a. Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6:** 111–123.

Wright, S. 1921b. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics* **6:** 124–143.

Wright, S. 1921c. Systems of mating. V. General considerations. *Genetics* **6:** 167–128.

Wright, S. 1922. Coefficients of inbreeding and relationship. *American Naturalist* **56:** 330–338.

Wright, S., and H. C. McPhee. 1925. An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. *Journal of Agricultural Research* **31:** 377–383.

Wright, S. 1929a. Fisher's theory of dominance. *American Naturalist* **63:** 274–279.

Wright, S. 1929b. The evolution of dominance. *American Naturalist* **63:** 556–561.

Wright, S. 1929c. Evolution in a Mendelian population. *Anatomical Record* **44:** 287.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16:** 97–159.

Wright, S. 1932. The roles of mutation, crossbreeding, and selection in evolution. *Proc. 6th International Congress of Genetics* **1:** 356–366.

Wright, S. 1933. Inbreeding and homozygosis. *Proc. Natl. Acad. Sci.* **19:** 411–420.

Wright, S. 1935a. The analysis of variance and the correlation between relatives with respect to deviations from an optimum. *Journal of Genetics* **30:** 243–256.

Wright, S. 1935b. Evolution in populations in approximate equilibrium. *Journal of Genetics* **30:** 257–266.

Wright, S. 1937. The distribution of gene frequencies in populations. *Proc. Natl. Acad. Sci.* **23:** 307–320.

Wright, S. 1938. The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci.* **24:** 372–377.

Wright, S. 1938. Size of population and breeding structure in relation to evolution. *Science* **87:** 430–431.

Wright, S. 1940. Breeding structure of populations in relation to speciation. *American Naturalist* **74:** 232–248.

Wright, S. 1942. Statistical genetics and evolution. *Bulletin of the American Mathematical Society* **48:** 223–246.

Wright, S. 1943. Isolation by distance. *Genetics* **28:** 114–138.

Wright, S. 1945. The differential equation of the distribution of gene frequencies. *Proc. Natl. Acad. Sci.* **31:** 382–389.

Wright, S. 1946. Isolation by distance under diverse systems of mating. *Genetics* **31:** 39–59.

Wright, S. and Th. Dobzhansky. 1946. Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. *Genetics* **31:** 125–156.

Wright, S. 1952. The genetics of quantitative variability. pp. 5–41 in *Quantitative Inheritance*, ed. E. C. R. Reeve and C. H. Waddington. Her Majesty's Stationery Office, London.

Wright, S. 1969. *Evolution and the Genetics of Populations. Vol. 2. The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Yokoyama, S. and J. Felsenstein. 1978. A model of kin selection for an altruistic trait considered as a quantitative character. *Proc. Natl. Acad. Sci.* **75:** 420–422.

Yule, G. U. 1906. On the theory of inheritance of quantitative compound characters on the basis of Mendel's law: a preliminary note. *Report of the Third International Conference on Genetics* 140–142.

Zhivotovsky, L. A., M. W. Feldman, F. B. Christiansen. 1994. Evolution of recombination among multiple selected loci: a generalized reduction principle. *Proc. Natl. Acad. Sci.* **91:** 1079–1083.