Mathematical Population Genetics

Introduction to the Stochastic Theory

Lecture Notes

Guanajuato, March 2009

Warren J Ewens

Preface

These notes should, ideally, be read before the Guanajuato meeting starts. They are intended to give background material in mathematical population genetics and also, in part, to form the background for some of the material given by other lecturers. At the very least, the first approximately 30 pages should be read before the meeting.

The notes are abstracted from Ewens (2004) *Mathematical Population Genetics*, from which further details not covered in these notes may be found.

Some standard genetical terms will be used and it is assumed that the reader is familiar with the meanings of these. These terms include gene, genotype, allele, (gene) locus, haploid, diploid, homozygote, heterozygote, heterozygosity, monoecious, dioecious, linkage, polymorphism and recombination.

Even if one is primarily a mathematician, one should not do mathematical population genetics in isolation. It should be considered as part of science, particularly (of course) of genetics and more recently genomics, and the relevance of mathematical population genetics to evolution, medicine, and other scientific activities should always be kept in mind. For this reason these notes start with a brief historical sketch and remarks about the beginnings of population genetics theory.

The historical background

Darwin and after

Although these notes describe stochastic processes in evolutionary genetics, it is appropriate to start with a brief summary of the historical background, and then the describe briefly the non-stochastic, or deterministic, theory of introductory population genetics.

The *Origin of Species* was published in 1859. Apart from the controversies it brought about on a nonscientific level, it set biologists at odds as to various aspects of the theory. That evolution had occurred was not, on the whole, questioned. What was more controversial was the claim that the agency bringing about evolution was natural selection, and, among selectionists, there was disagreement about the the nature of a selectively induced evolutionary changes.

These difficulties arose mainly because the nature of the hereditary mechanism was not generally, known, since Mendel's work, and hence the mechanism of heredity, was in effect unknown before 1900.

In so far as a common view of heredity existed at the time, it would have been that the characteristics of an individual are, or tend to be, a blending of the corresponding characteristics of his parents. The blending hypothesis brought perhaps the most substantial scientific objection to Darwin's theory. It is easy to see that with random mating, the variance in a population for any characteristic will, under the blending theory, decrease by a factor of one-half in each generation. Thus uniformity of characteristics would essentially be obtained after a few generations, so that eventually no variation would exist upon which natural selection could act. Since, of course, such uniformity is not observed, this argument is incomplete. But since variation of the degree observed could only occur by postulating further factors of strong effect which cause the characteristics of offspring to deviate from those of their parents, it cannot then be reasonably argued that selectively favored parents produce offspring who closely resemble them and who are thus themselves selectively favored. This argument was recognized by Darwin as a major obstacle to his theory of evolution through natural selection, and it is interesting to note that later versions of the *Origin* were, unfortunately, somewhat influenced by this argument.

The year 1900 saw the rediscovery of Mendelism. The particulate nature of this theory was of course appealing to the saltationists. Rather soon many biologists believed in a non-Darwinian process of evolution through mutational jumps – the view that "Mendelism had destroyed Darwinism" was not uncommon. On the other hand, the biometricians continued to believe in the Darwinian theory of gradualist evolution through natural selection and were thus, in the main, disinclined to believe in the Mendelian mechanism, or at least that this mechanism was of fundamental importance in evolution.

Given the above, it is therefore paradoxical that in actual fact, not only are Darwinism and Mendelism compatible, the Darwinian theory relies crucially on the Mendelian mechanism. Further, it would be difficult to conceive of a Mendelian system without some form of natural selection associated with it. To see why this should be so, it is now necessary to turn to the beginnings of the mathematical theory of population genetics.

The Hardy–Weinberg law

We consider a random-mating monoecious population (that is, a population with no concept of two separate sexes) which is so large that genotype frequency changes may be treated as deterministic, and focus attention on a given gene locus at which two alleles may occur, namely A_1 and A_2. Suppose that in any generation the proportions of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are u, $2v$, and w, respectively. Since random mating obtains, the frequency of matings of the type $A_1A_1 \times A_1A_1$ is u^2, that of $A_1A_1 \times A_1A_2$ is $4uv$, and so on. If there is no mutation and no fitness differentials between genotypes, elementary Mendelian rules indicate that the outcome of an $A_1A_1 \times A_1A_1$ mating must be A_1A_1 and that in an indefinitely large population, half the $A_1A_1 \times A_1A_2$ matings will produce A_1A_1 offspring, and the other half will produce A_1A_2 offspring, with similar results for the remaining matings.

It follows that since A_1A_1 offspring can be obtained only from $A_1A_1 \times A_1A_1$ matings (with overall frequency 1 for such matings), from $A_1A_1 \times A_1A_2$ matings (with overall frequency $\frac{1}{2}$ for such matings), and from $A_1A_2 \times A_1A_2$ matings (with frequency $\frac{1}{4}$ for such matings), and since the frequencies of these matings are u^2, $4uv$, $4v^2$, the frequency u' of A_1A_1 in the following generation is

$$u' = u^2 + \tfrac{1}{2}(4uv) + \tfrac{1}{4}(4v^2) = (u+v)^2. \tag{1}$$

Similar considerations give the frequencies $2v'$ of A_1A_2 and w' of A_2A_2 as

$$2v' = \tfrac{1}{2}(4uv) + \tfrac{1}{2}(4v^2) + 2uw + \tfrac{1}{2}(4vw) = 2(u+v)(v+w), \tag{2}$$

$$w' = \tfrac{1}{4}(4v^2) + \tfrac{1}{2}(4vw) + w^2 = (v+w)^2. \tag{3}$$

The frequencies u'', $2v''$ and w'' for the next generation are found by replacing u', $2v'$ and w', by u'', $2v''$ and w'' and u, $2v$ and w by u', $2v'$ and w' in (1)–(3). Thus, for example, using (1) and (2),

$$\begin{aligned} u'' &= (u'+v')^2 \\ &= (u+v)^2 \\ &= u', \end{aligned}$$

and similarly it is found that $v'' = v'$, $w'' = w'$. Thus, the genotype frequencies established by the second generation are maintained in

the third generation and consequently in all subsequent generations. Frequencies having this property can be characterized as those satisfying the relation

$$(v')^2 = u'w'. \tag{4}$$

Clearly if this relation holds in the first generation, so that

$$v^2 = uw, \tag{5}$$

then not only would there be no change in genotypic frequencies between the second and subsequent generations, but also these frequencies would be the same as those in the first generation. Populations for which (5) is true are said to have genotypic frequencies in Hardy–Weinberg form.

Since $u + 2v + w = 1$, only two of the frequencies u, $2v$ and w are independent. If, further, (5) holds, only one frequency is independent. Examination of the recurrence relations (1)-(3) shows that the most convenient quantity for independent consideration is the frequency $x = u + v$ of the allele A_1. These conclusions may be summarized in the form of a theorem:

Theorem (Hardy–Weinberg). Under the assumptions stated, a population having genotypic frequencies u (of A_1A_1), $2v$ (of A_1A_2) and w (of A_2A_2) achieves, after one generation of random mating, stable genotypic frequencies x^2, $2x(1-x)$, $(1-x)^2$ where $x = u + v$ and $1 - x = v + w$. If the initial frequencies u, $2v$, w are already of the form x^2, $2x(1-x)$, $(1-x)^2$, then these frequencies are stable for all generations.

The important consequence of this theorem lies in the stability behavior. If no external forces act, there is no intrinsic tendency for any variation present in the population, that is, variation caused by the existence of the three different genotypes, to disappear. This shows immediately that the major earlier criticism of Darwinism, namely the fact that variation decreases rapidly under the blending theory, does not apply with Mendelian inheritance. It is clear directly from the Hardy–Weinberg Law that under a Mendelian system of inheritance, variation tends to be maintained.

The deterministic theory of natural selection

The twin cornerstones of the Darwinian theory of evolution are variation and natural selection. Since the different genotypes in a population will often have different fitnesses, that is will differ in viability, mating success, and fertility, natural selection will occur. We now outline the work done during the 1920s and 1930s in this direction. This quantification amounts to a scientific description of the Darwinian theory in Mendelian terms.

It is necessary, at least as a first step, to make a number of assumptions and approximations about the evolutionary process. Thus although mutation is essential for evolution, mutation rates are normally so small that for certain specific problems we may ignore mutational events. Further, although the fitness of an individual is determined in a complex way by his entire genetic make-up, and even then will often differ from one environment to another, we start by assuming as a first approximation that this fitness depends on his genotype at a single locus, or at least can be found by "summing" single locus contributions to fitness. It is also difficult to cope with that component of fitness which relates to fertility, and almost always special assumptions are made about this. If fitness relates solely to viability then much of the complexity is removed, and for convenience we make this assumption in these notes.

Suppose then that the fitnesses and the frequencies of the three genotypes A_1A_1, A_1A_2, and A_2A_2 at a certain locus "A" are as given below:

	A_1A_1	A_1A_2	A_2A_2
fitness	w_{11}	w_{12}	w_{22}
frequency	x^2	$2x(1-x)$	$(1-x)^2$

(6)

We have written the frequencies of these genotypes in the Hardy–Weinberg form appropriate to random mating. Now Hardy–Weinberg frequencies apply only at the moment conception, since from that time on differential viabilities alter genotype frequencies from the Hardy–Weinberg form. For this reason we count frequencies in the population at the moment of conception of each generation.

Clearly the most interesting question to ask is: What is the behavior of the frequency x of the allele A_1 under natural selection?

Since we take the fundamental units of the microevolutionary process to be the replacement in a population of an "inferior" allele by a "superior" allele, the answer to this question is essential to an understanding of the microevolutionary process as directed by natural selection.

The first step is to find the frequency x' of A_1 in the following generation. By considering the fitnesses of each individual and all possible matings, we find that

$$x' = \frac{w_{11}x^2 + w_{12}x(1-x)}{w_{11}x^2 + 2w_{12}x(1-x) + w_{22}(1-x)^2}\,. \tag{7}$$

Clearly continued iteration of this recurrence relation yields the successive values taken by the frequency of A_1. Unfortunately simple explicit expressions for these frequencies are not always available, and resort must be made to approximation.

Before discussing these approximations, we observe that x' depends on the ratios of the fitnesses w_{ij} rather than the absolute values, so that x' is unchanged if we multiply each w_{ij} by any convenient scaling constant. It is therefore possible to scale the w_{ij} in any way convenient to the analysis at hand. Different scalings are more convenient for different purposes. We indicate below (in (8), (9) and (10)) three alternative representations of the fitness values; on different occasions different representations might prove to be the most useful. It should be emphasized that nothing is involved here other than convenience of notation.

Fitness Values			
A_1A_1	A_1A_2	A_2A_2	
w_{11}	w_{12}	w_{22}	(8)
$1+s$	$1+sh$	1	(9)
$1-s_1$	1	$1-s_2$	(10)

We normally assume that except in extreme cases, perhaps involving lethality, the fitness differentials s, sh, s_1 and s_2 are small, perhaps of the order of 1%. In this case we ignore small-order terms in these parameters.

In the form (9), that parameter h is called a "dominance" parameter, and the value $h = 1/2$ corresponds to the case of no dominance, where the fitness of A_1A_2 is half-way between that of A_1A_1 and A_2A_2.

Using the fitness scheme (9), the recurrence relation (7) leads, to a sufficiently close approximation, to

$$x' - x = sx(1-x)\{x + h(1-2x)\}. \tag{11}$$

If we measure time in units of one generation, this equation may be approximated, in turn, by

$$dx/dt = sx(1-x)\{x + h(1-2x)\}. \tag{12}$$

If the time required for the frequency of A_1 to move from some value x_1 to some other value x_2 is denoted by $t(x_1, x_2)$, then clearly

$$t(x_1, x_2) = \int_{x_1}^{x_2} \big(sx(1-x)\{x + h(1-2x)\}\big)^{-1}\, dx. \tag{13}$$

Naturally this equation applies only in cases where, starting from x_1, the frequency of A_1 will eventually reach x_2.

While an explicit expression for $t(x_1, x_2)$ is possible, it is usually more convenient to use the expression (13) directly. Suppose first that $s > sh > 0$. Then it is clear from (12) that the frequency of A_1 steadily increases towards unity. However, as this frequency approaches unity, the time required for even small changes in it will be large, due to the small term $1 - x$ in the denominator of the integrand in (13). This behavior is even more marked in the case $h = 1$ (A_1 dominant to A_2 in fitness), for then the denominator in the integrand in (13) contains a multiplicative term $(1-x)^2$. This very slow rate of increase is due to the fact that, once x is close to unity, the frequency of A_2A_2, the genotype against which selection is operating, is extremely low. In the important particular case $h = \frac{1}{2}$, that is no dominance in fitness, (13) assumes the simple form

$$t(x_1, x_2) = \int_{x_1}^{x_2} \left\{\tfrac{1}{2}sx(1-x)\right\}^{-1}\, dx. \tag{14}$$

It is possible to evaluate the times required for any nominated changes in the frequency of A_1 from (13) and (14). This procedure quantifies, at least approximately, important aspects of the unit microevolutionary process of the replacement of an "inferior" allele by a "superior" allele.

The case where $s < sh < 0$ is a mirror-image to the case just considered, and needs no further discussion.

In both cases considered above the fitness of the heterozygote is intermediate between that of the two homozygotes. When the heterozygote A_1A_2 is at a selective advantage over both homozygotes it is convenient to use the fitness notation (10), where $s_1 > 0, s_2 > 0$. It is easy to see that the population evolves to a stable equilibrium where the frequency of A_1 is $s_2/(s_1 + s_2)$. Thus both A_1 and A_2 remain forever in the population.

We do not pursue the deterministic theory further in these notes, since the material above is sufficient for the various stochastic theory calculations that we shall make later.

The beginnings of population genetics theory

It was thus beginning to become clear, following the derivation of the Hardy-Weinberg law and the elements of the deterministic theory of natural selection, that a reconciliation between Darwinism and Mendelism was not only possible but indeed inevitable. The marriage of these two fields can be said to have produced the subject of population genetics.

Thee is one important deficiency in the analyses given above: the population considered is assumed to be infinite in size, so that that random, or stochastic, changes in gene frequencies are not allowed. However, all population sizes are, or course, finite, and thus the stochastic aspect of evolutionary population genetics must be investigated. This was of course recognized from the earliest times, and thus the stochastic theory of population genetics is one of the oldest examples of stochastic processes applied in science. **From now on, these notes focus entirely on stochastic processes in evolutionary genetics.** This implies that large areas of the modern theory of population genetics, especially "population genomics", where the theory is deterministic (because of the complexities of the entire genome), are not covered.

A remark about notation

Standard notation in probability and statistics is to denote random variables by upper case notation. Since we consider the stochastic

theory in these notes, we attempt so far as possible to use this notational convention. However, it is not possible to use this convention all the time, and sometimes upper case notation is used for quantities that are not random variables. Here are some examples:-

(i) Population sizes are denoted in upper case (for example N), whereas sample sizes are denoted in lower case (for example n).

(ii) While the *number* of genes in a stochastic model is denoted in upper case, the corresponding *frequencies*, or *proportions*, are often denoted in the corresponding lower case.

(iii) To conform with the Markov chain convention of using the symbol p_{ij} to denote a transition probability from "i" to "j", the lower case notations "i" and "j" are often used to denote the values of discrete random variables in Markov chain models.

Parameters, in particular θ, are usually denoted in Greek, in accordance with standard statistical notation. However, mutation rates are denoted in lower case Roman (usually u and v).

The stochastic theory

Finite Markov chains

It is assumed that you are familiar with the basic notions of finite Markov chains. However, all the stochastic theory considered below is in terms of finite Markov chains, so a very brief introduction to the theory of these chains is given in this section.

Consider a discrete random variable X which at time points 0, 1, 2, 3, ... takes one or other of the values 0, 1, 2, ..., M. We shall say that X, or the system, is in state E_i if X takes the value i. Suppose that, at some time t, the random variable X is in state E_i. Then if the probability p_{ij} that, at time $t+1$, the random variable is in state E_j is independent of t and also of the states occupied by X at times $t-1$, $t-2$, ..., the variable X is said to be Markovian, and its probability laws follow those of a finite Markov chain. If the initial probability (at $t=0$) that X is in E_i is a_i then the probability that X is in the state E_i, E_j, E_k, E_ℓ, E_m ... at times 0, 1, 2, 3, 4 ... is

$ap_ip_{ij}p_{jk}p_{k\ell}p_{\ell m}\ldots$.

Complications to Markov chain theory arise if periodicities occur, for example, if X can return to E_i only at the time points t_1, $2t_1$, $3t_1, \ldots$ for $t_1 > 1$. Further minor complications arise if the states E_0, $E_1, \ldots, E_M$ can be broken down into non-communicating subsets. To avoid unnecessary complications, which never in any event arise in genetical applications, we suppose that no periodicities exist and that, apart from the possibility of a small number of absorbing states, (E_i is absorbing if $p_{ii} = 1$), no breakdown into non-communicating subsets occur.

It is convenient to collect the p_{ij} into a matrix $P = \{p_{ij}\}$, so that

$$P = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0M} \\ p_{01} & p_{11} & \cdots & p_{1M} \\ \vdots & & & \\ p_{M0} & p_{M1} & \cdots & p_{MM} \end{pmatrix}. \tag{15}$$

The probability $p_{ij}^{(2)}$ that X is in E_j at time $t+2$, given it is in E_i at time t, is evidently

$$p_{ij}^{(2)} = \sum_k p_{ik}p_{kj}.$$

Since the right-hand side is the (i,j)th element in the matrix P^2, and if we write $P^{(2)} = \{p_{ij}^{(2)}\}$, then

$$P^{(t)} = P^t \tag{16}$$

for $t = 2$. More generally (16) is true for any positive integer t. In all cases we consider, P^t can be written in the "spectral expansion" form

$$P^t = \lambda_0^t r_0 \ell_0' + \lambda_1^t r_1 \ell_1' + \cdots + \lambda_M^t r_M \ell_M' \tag{17}$$

where λ_0, $\lambda_1, \ldots, \lambda_M$ ($|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_M|$) are the eigenvalues of P and $(\ell_0, \ldots, \ell_M)$ and $(r_0, \ldots, r_M)$, normalized so that

$$\ell_i' \mathbf{r}_i = \sum_{j=0}^{M} \ell_{ij} r_{ij} = 1, \tag{18}$$

are the corresponding left and right eigenvectors, respectively.

Suppose E_0 and E_M are absorbing states and that no other states are absorbing. Then $\lambda_0 = \lambda_1 = 1$ and if $|\lambda_2| > |\lambda_3|$ and $i, j = 1, 2, \ldots, M-1$,

$$p_{ij}^{(t)} = r_{2i}\ell_{2j}\lambda_2^t + o(\lambda_2^t) \tag{19}$$

for large t. Thus the leading non-unit eigenvalue λ_2 plays an important role in determining the rate at which absorption into either E_0 and E_M occurs.

Let π_j be the probability that eventually E_M (rather than E_0) is entered, given initially that X is in E_i. By considering values of X at consecutive time points it is seen that the π_i satisfy

$$\pi_i = \sum_{j=0}^{M} p_{ij}\pi_j, \quad \pi_0 = 0, \quad \pi_M = 1. \tag{20}$$

The mean times $\bar{t}_i$ until absorption into E_0 or E_M occurs, given that X is in E_i, similarly satisfy

$$\bar{t}_i = \sum_{j=0}^{M} p_{ij}\bar{t}_j + 1, \quad \bar{t}_0 = \bar{t}_M = 0. \tag{21}$$

Starting with X in E_i the members of the set of mean times $\{\bar{t}_{ij}\}$ that X is in E_j before absorption into either E_0 or E_M satisfy the equations

$$\bar{t}_{ij} = \sum_{k=0}^{M} p_{ik}\bar{t}_{kj} + \delta_{ij}, \quad \bar{t}_{0j} = \bar{t}_{Mj} = 0, \tag{22}$$

where $\delta_{ij} = 1$ and $i = j$ and $\delta_{ij} = 0$ otherwise. Further,

$$\bar{t}_{ij} = \sum_{n=0}^{\infty} p_{ij}^{(n)}, \quad \bar{t}_i = \sum_{j=1}^{M-1} \bar{t}_{ij}. \tag{23}$$

Consider now only those cases for which E_M is the absorbing state eventually entered. Writing X_t for the value of X at time t, we get

$$\begin{aligned} p_{ij}^* &= \text{Prob}\{X_{t+1} \text{ in } E_j \mid X_t \text{ in } E_i, E_M \text{ eventually entered}\} \\ &= \text{Prob}\{X_{t+1} \text{ in } E_j \text{ and } E_M \text{ eventually entered} \mid X_t \text{ in } E_i\} \\ &\quad \div \text{Prob}\{E_M \text{ eventually entered} \mid X_t \text{ in } E_i\} \\ &= p_{ij}\pi_j/\pi_i, \qquad (i, j = 1, 2, \ldots, M), \end{aligned} \tag{24}$$

using conditional probability arguments and the Markovian nature of X. Let $\widetilde{P}$ be the matrix derived from P by omitting the first row and first column and let

$$V = \begin{pmatrix} \pi_1 & & & \\ & \pi_2 & & \\ & & \ddots & 0 \\ 0 & & & \pi_M \end{pmatrix}. \tag{25}$$

Then if $P^* = \{p^*_{ij}\}$, Eq. (24) shows that

$$P^* = V^{-1}\widetilde{P}V. \tag{26}$$

Standard theory shows that the eigenvalues of P^* are identical to those of P (with one unit eigenvalue omitted) and that if $\ell'(\mathbf{r})$ is any left (right) eigenvector of $\widetilde{P}$, then the corresponding left and right eigenvector of P^* are $\ell' V$ and $V^{-1}\mathbf{r}$. Further, if $P^{*(n)}$ is the matrix of conditional n step transition probabilities,

$$P^{*(n)} = (P^*)^n = V^{-1}\widetilde{P}^n V$$

so that

$$p^{*(n)}_{ij} = p^{(n)}_{ij}\pi_j/\pi_i, \tag{27}$$

a conclusion that can be reached directly as with (24). If $\bar{t}^*_{ij}$ is the conditional mean time spent in E_j, given initially X in E_i, then

$$\begin{aligned} \bar{t}^*_{ij} &= \sum_{n=0}^{\infty} p^{*(n)}_{ij} \\ &= (\pi_j/\pi_i)\sum_{n=0}^{\infty} p^{(n)}_{ij} \\ &= \bar{t}_{ij}\pi_j/\pi_i. \end{aligned} \tag{28}$$

If there is only one absorbing state interest centers solely on properties of the time until the state is entered. Taking E_0 as the only absorbing state and E_i as the initial state, the mean time t_i until absorption satisfies (21) with the single boundary condition $\bar{t}_0 = 0$, and the mean number of visits to E_j satisfies (22) with the single condition $\bar{t}_{0j} = 0$.

If there are no absorbing states P will have a single eigenvalue and all other eigenvalues will be strictly less than unity in absolute value. Equation (17) then shows that

$$\lim_{t\to\infty} P^t = \mathbf{r}_0 \ell_0' \tag{29}$$

and since r_0 is of the form $(1, 1, 1, \ldots, 1)'$,

$$\lim_{t\to\infty} p_{ij}^{(t)} = \ell_{0j} \quad \text{for all } i. \tag{30}$$

Using a slightly different notation we may summarize this by saying

$$\lim_{t\to\infty} p_{ij}^{(t)} = \phi_j, \tag{31}$$

where $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_M)$ is the unique solution of the two equations

$$\boldsymbol{\phi}' = \boldsymbol{\phi}' P, \quad \sum_{j=0}^{M} \phi_j = 1. \tag{32}$$

The vector $\boldsymbol{\phi}$ is called the stationary distribution of the process and in genetical applications exists only if fixation of any allele is impossible (e.g. if all alleles mutate at positive rates).

If the matrix P is a continuant (so that $p_{ij} = 0$ if $|i - j| > 1$) explicit formulae can be found for most of these quantities. We do not give these formulae here, but will give examples of them when considering genetical models involving continuant Markov chains.

We conclude our discussion of finite Markov chains by introducing the concept of time reversibility. Consider a Markov chain admitting a stationary distribution $\{\phi_0, \phi_1, \ldots, \phi_M\}$. Then we define the process to be reversible if, at stationarity,

$$\text{Prob}\{X_t, X_{t+1}, \ldots, X_{t+n}\} = \text{Prob}\{X_t, X_{t-1}, \ldots, X_{t-n}\} \tag{33}$$

for every t and n. A necessary and sufficient condition for this is that the stationary state has been reached and that the equation

$$\phi_i p_{ij} = \phi_j p_{ji} \tag{34}$$

hold for all i, j. Certain classes of Markov chains are always reversible. For example, if the transition matrix is a continuant, the Markov chain at stationarity is reversible. Certain other chains, in particular several having genetical relevance, are also reversible: we consider these later when discussing the uses to which the concept of reversibility can be put.

The "simple" Wright-Fisher model

It is necessary, in order to arrive at a theoretical estimate of the importance of the stochastic factor, to set up stochastic models which reasonably describe the behavior of a finite population. Perhaps more than in any other part of population genetics theory, the choice of a model is arbitrary, and no-one pretends that Nature necessarily follows at all closely the models we construct. Although they did not use the terminology of Markov chain theory, the methods used by Fisher and Wright, who initiated the theory in the 1920's, are in fact those of this theory and its close relative, diffusion theory. Here we present some of the conclusions of Fisher and Wright, but unlike them we present these in the terminology of Markov chains.

We consider, as the simplest possible case, a diploid population of fixed size N. Suppose that the individuals in this population are monoecious, (that is, there is no concept of two separate sexes), that no selective difference exist between the two alleles A_1 and A_2 possible at a certain locus "A," and that there is no mutation. There are $2N$ genes in the population in any generation, and it is sufficient to center our attention on the number X of A_1 genes. Clearly in any generation X takes one or other of the values $0, 1, \ldots, 2N$, and we denote the value assumed by X in generation t by $X(t)$.

We must now assume some specific model which describes the way in which the genes in generation $t+1$ are derived from the genes in generation t. Clearly many reasonable models are possible and, for different purposes, different models might be preferable. Naturally, biological reality should be the main criterion in our choice of model, but it is inevitable that we consider mathematical convenience in this choice. The Markov chain model discussed below, although it was not written down explicitly by Fisher and Wright, was clearly in effect known to them both, since they both gave several formulas deriving from it.

The model assumes that the genes in generation $t+1$ are derived by sampling with replacement from the genes of generation t. This means that the number $X(t+1)$ of A_1 genes in generation $t+1$ is a binomial random variable with index $2N$ and parameter $X(t)/2N$. More explicitly, the model assumes that, given that $X(t) = i$, the

probability p_{ij} that $X(t+1) = j$ is given by

$$p_{ij} = \binom{2N}{j}(i/2N)^j\{1-(i/2N)\}^{2N-j}, \quad i,j = 0,1,2,\ldots,2N. \quad (35)$$

More precisely, we refer to the model (35) as the "simple" Wright–Fisher model, since it does not incorporate selection, mutation, population subdivision, two sexes or any other complicating feature. The purpose of introducing it is to allow an initial examination of the effects of stochastic variation in gene frequencies, without any further complicating features being involved. More complicated models that introduce factors such as selection and mutation, and which allow more than two alleles, but which share the binomial sampling characteristic of (35), will all be referred to generically as "Wright–Fisher" models. We emphasize that all of these models are no more than crude approximations to biological reality.

In the form of (35), it is clear that $X(\cdot)$ is a Markovian random variable with transition matrix $P = \{p_{ij}\}$, so that in principle the entire probability behavior of $X(\cdot)$ can be arrived at through knowledge of P and the initial value $X(0)$ of X. In practice, unfortunately, the matrix P does not lend itself readily to simple explicit answers to many of the questions we would like to ask, and we shall be forced, later, to consider alternative approaches to these questions.

On the other hand, (35) does enable us to make some comments more of less immediately. Perhaps the most important is that whatever the value $X(0)$, eventually $X(\cdot)$ will take either the value 0 or $2N$, and once this happens there will be no further change in the value of $X(\cdot)$. We refer to this as fixation (of A_2 and A_1 respectively). Genetically this corresponds, of course, to the fact that since the model (35) does not allow mutation, once the population is purely A_2A_2 or purely A_1A_1, no variation exists, and no further evolution is possible at this locus. It was therefore natural for both Fisher and Wright to find, assuming the model (35), the probability of eventual fixation of A_1 rather than A_2, and perhaps more important, to attempt to find how much time might be expected to pass before fixation of one or other allele occurs.

The answer to the first question is $X(0)/2N$. This conclusion may be arrived at by a variety of methods, the one most appropriate to Markov chain theory being that if the probability of fixation of A_1,

given that currently $X(\cdot) = i$, is denoted π_i, then the choice

$$\pi_i = i/(2N) \tag{36}$$

satisfies the standard Markov chain fixation probability difference equations (20), together with the appropriate boundary conditions (in this model) $\pi_1 = 0, \pi_{2N} = 1$. Setting $i = X(0)$ leads to the required solution. A second way of arriving at the value $X(0)/2N$ is to note that $X(\cdot)/2N$ is a martingale, that is satisfies the "invariant expectation" formula

$$E\{X(t+1)/2N \mid X(t)\} = X(t)/2N, \tag{37}$$

and then use either martingale theory or informal arguments to arrive at the desired value. A third approach, more informal and yet from a genetical point of view perhaps more useful, is to observe that eventually every gene in the population is descended from one unique gene in generation zero. The probability that such a gene is A_1 is simply the initial fraction of A_1 genes, namely $X(0)/2N$, and this must also be the fixation probability of A_1.

It is far more difficult to assess the properties of the (random) time until either loss or fixation of the allele A_1 occurs. The most obvious quantity to evaluate is the mean time $\bar{t}\{X(0)\}$, measured with unit time corresponding to one generation, until $X(\cdot)$ reaches 0 or $2N$, starting from $X(0)$. No simple explicit formula for this mean time is known, and we now discuss indirect arguments concerning this mean and also some approximations to this mean.

The indirect arguments center around the eigenvalues of the matrix $P = \{p_{ij}\}$ defined by (35). Standard Markov chain theory shows that one eigenvalue of this matrix is automatically 1. Denoting this eigenvalue by λ_0, the remaining eigenvalues will later in these notes be shown to be

$$\lambda_j = [(2N)(2N-1)\dots(2N-j+1)]/[(2N)^j], \quad j = 1, 2, \dots, 2N. \tag{38}$$

Thus $\lambda_1 = 1$ and λ_2, largest non-unit eigenvalue, is $1 - (2N)^{-1}$. From (19), the probability that $X(t) = j$, for $j = 1, 2, \dots, 2N-1$, given that $X(0) = i$, is of the form

$$\text{const}\{1 - (2N)^{-1}\}^t + o\{1 - (2N)^{-1}\}^t \tag{39}$$

for t large. Thus implies that the probability that the number of A_1 genes reaches 0 or $2N$ by time t increases very slowly as a function of t.

It also follows from (35) that if we put $x(t) = X(t)/2N$,

$$\mathrm{E}\big(x(t+1)\{1-x(t+1)\} \mid x(t)\big) = \{1-(2N)^{-1}\}x(t)\{1-x(t)\}, \quad (40)$$

so that the expected value of the so-called "heterozygosity measure" $2x(\cdot)\{1-x(\cdot)\}$ decreases by a factor of $1-(2N)^{-1}$ each generation. This confirms the conclusion drawn from equation (39). We conclude that although genetic variation, that is the existence of both A_1 and A_2 in the population, must ultimately be lost under the model (35), the loss of genetic variation is usually very slow. We might even suspect that the mean time for loss of variation is of order N generations, and will in fact soon confirm that, apart from cases where the initial number of A_1 genes is very low or very high, this is indeed the case. This slow rate of loss may be thought of as a stochastic analogue of the "variation-preserving" property of infinite genetic populations.

On the other hand, we will see later that indirect arguments concerning the mean time until loss of genetic variation found by using the eigenvalue λ_2, and to a lesser extent to the complete set (38) of eigenvalues, are sometimes misleading, and we turn now to approximations for this mean time.

Taylor series approximations

We consider the Markov chain model (35), and for convenience put $i/2N = x$, $j/M = x + \delta x$, and write $\bar{t}(x)$ as the mean time for loss or fixation of A_1, given a current frequency x. We assume that $\bar{t}(x)$ is a twice differentiable function of a continuous variable x. Then from standard theory,

$$\begin{aligned} \bar{t}(x) &= \sum_{\delta x} \mathrm{Prob}\{x \to x+\delta x\}\bar{t}(x+\delta x) + 1 && (41)\\ &= \mathrm{E}\{\bar{t}(x+\delta x)\} + 1 && (42)\\ &\approx \bar{t}(x) + \mathrm{E}(\delta x)\frac{d\bar{t}(x)}{dx} + \tfrac{1}{2}\mathrm{E}(\delta x)^2\frac{d^2\bar{t}(x)}{dx^2} + 1, && (43) \end{aligned}$$

where all expectations are conditional on x and in (43) only the first three terms in an infinite Taylor series have been retained. This leads to the equation

$$\mathrm{E}(\delta x)\frac{d\bar{t}(x)}{dx} + \tfrac{1}{2}\mathrm{E}(\delta x)^2\frac{d^2\bar{t}(x)}{dx^2} \approx -1 \quad (44)$$

Since from (35)

$$\mathrm{E}(\delta x) = 0, \quad \mathrm{E}(\delta x)^2 = \frac{x(1-x)}{2N}, \tag{45}$$

the approximation (44) gives

$$\frac{x(1-x)}{4N}\,\frac{d^2\bar{t}(x)}{dx^2} \approx -1. \tag{46}$$

The solution of this equation, subject to the obvious boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$, is

$$\bar{t}(p) = -4N\{p\log p + (1-p)\log(1-p)\}, \tag{47}$$

where $p = i/2N$ is the initial frequency of A_1. Since the above calculations involve approximations, it is more correct to say that

$$\bar{t}(p) \approx -4N\{p\log p + (1-p)\log(1-p)\}. \tag{48}$$

We shall see later that this Taylor series approximation is also the so-called diffusion approximation to the required mean time, although we have here not made any reference to diffusion processes. It is known that this is an extremely accurate approximation.

In the case $i = 1$, so that $p = (2N)^{-1}$, the value appropriate if A_1 is a unique new mutation in an otherwise pure A_2A_2 population, equation (48) leads to

$$\bar{t}\{(2N)^{-1}\} \approx 2 + 2\log 2N \text{ generations}, \tag{49}$$

while when $p = \frac{1}{2}$,

$$\bar{t}\{\tfrac{1}{2}\} \approx 2.8N \text{ generations}. \tag{50}$$

This very long mean time, for equal initial frequencies, is of course intimately connected with the fact that the leading non-unit eigenvalue of (35) is very close to unity, differing from unity by a term of order N^{-1}.

A further question, taken up as long ago as the 1930's by Fisher and Wright, is the following. Suppose that, in an otherwise pure A_2A_2 population, a single new mutant A_1 gene arises. No further mutation occurs, so from this point on the model (35) applies. The mean fixation time $\bar{t}_1$ may be written in the form

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j}, \tag{51}$$

where $\bar{t}_{1,j}$ is the mean number of generations for which the number of A_1 genes takes the value j before reaching either 0 or $2N$. Both Fisher and Wright found that

$$\bar{t}_{1,j} \approx 2j^{-1}, \quad j = 1, 2, \ldots, 2N-1, \tag{52}$$

so that, using (51),

$$\bar{t}_1 \approx 2\big(\log(2N-1) + \gamma\big), \tag{53}$$

where γ is Euler's constant 0.5772157 ... This approximation is very close to that given in (49). (The best approximation known is $\bar{t}_1 \approx 2\log(2N) + 1.355076$.)

There is an ergodic equivalent to the expressions in (51) and (52) which is perhaps of more interest than (51) and (52) themselves, and which is indeed the route by which Fisher arrived at these formulae. Consider a sequence of independent loci, each initially pure "A_2A_2", and at which a unique mutation A_1 occurs in generation k in the kth member of the sequence. We may then ask how many such loci will be segregating for A_1 and A_2 after a long time has passed, and at how many of these loci will there be exactly j "A_1" genes. It is clear that the mean values of these quantities are $\bar{t}_1$ and $\bar{t}_{1,j}$, respectively, and this gives us some idea, at least insofar as the model (35) is realistic, of how much genetic variation we may expect to see in any population at a given time. The question of the amount, and the nature, of the genetic variation that can be expected in a population at any given time will be discussed later, and at much greater length. In that discussion ergodic arguments, moving from a "time" calculation to a result concerning the make-up of a contemporary population, will again be used.

Conditional processes

We return to the model (35) and consider now only those cases for which the number of A_1 genes eventually takes the value $2N$. We first find the transition matrix of the conditional process when the condition is made that eventually this fixation event occurs.

If the typical term in this conditional process transition matrix

is denoted p^*_{ij}, we get, from (24) and (36)

$$\begin{aligned} p^*_{ij} &= \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(\frac{2N-i}{2N}\right)^{2N-j}\frac{j}{i} \\ &= \binom{2N-1}{j-1}\left(\frac{i}{2N}\right)^{j-1}\left(\frac{2N-i}{2N}\right)^{2N-j}. \end{aligned} \tag{54}$$

An intuitive explanation for the form of p^*_{ij} is that, under the condition that A_1 fixes, at least one A_1 gene must be produced in each generation. Then p^*_{ij} is the probability that the remaining $2N-1$ gene transmissions produce exactly $j-1$ A_1 genes.

Conditional Markov chain theory (see the discussion following (26)) shows that the eigenvalues of P^* are identical to those of P (with one unit eigenvalue omitted). However, as the results just given above and the diffusion approximation to the conditional process (given below) both show, the properties of the conditional process are quite different from those of the original (unconditional) process (35). This is one reason why the use of eigenvalues in assessing properties of the time until loss of genetic variation in the unconditional process, as discussed above, can lead to misleading conclusions.

We now consider approximate calculations for the conditional process parallel to those leading to those leading to (46) for the unconditional process. To use equations and approximations parallel to those in (41) – (43), we have to find an expression for $E^*(\delta x)$ and $E^*(\delta x)^2$, the means and of δx and $(\delta x)^2$ respectively, in the conditional process whose transition matrix is given in (54). The intuitive comments following (54) show that, given the number i of A_1 genes in any generation, the mean number of A_1 genes in the next generation is $1+(2N-1)i/(2N)$. The difference between this value and i is $1-i/(2N)=1-x$, where $x=i/(2N)$. Thus the mean increase in the proportion of A_1 genes, that is $E^*(\delta x)$, is $(1-x)/(2N)$. Similar arguments show that, if terms of order n^{-2} are ignored, $E^*(\delta x)^2$ is $x(1-x)/(2N)$.

If the conditional mean time to fixation is denoted by $\bar{t}^*(x)$, given a current frequency x of A_1, these arguments lead to the approximating equation

$$\frac{(1-x)}{2N}\frac{d\bar{t}^*(x)}{dx}+\frac{x(1-x)}{4N}\frac{d^2\bar{t}^*(x)}{dx^2}=-1. \tag{55}$$

The solution of this equation, subject to $\bar{t}^*(1) = 0$ and the requirement

$$\lim_{x \to 0} \bar{t}^*(x) \text{ is finite,} \tag{56}$$

and assuming initially $x = p$, is

$$\bar{t}^*(p) = -4Np^{-1}(1-p)\log(1-p), \tag{57}$$

so that the Taylor series approximation to the conditional mean time is

$$\bar{t}^*(p) \approx -4Np^{-1}(1-p)\log(1-p). \tag{58}$$

(As with the unconditional process, we see later that this is also the diffusion approximation to this mean time.) We observe from (57) that

$$\bar{t}^*\{(2N)^{-1}\} \approx 4N - 2 \text{ generations,} \tag{59}$$

$$\bar{t}^*\left\{\tfrac{1}{2}\right\} \approx 2.8N \text{ generations,} \tag{60}$$

$$\bar{t}^*\{1 - (2N)^{-1}\} \approx 2\log 2N \text{ generations.} \tag{61}$$

The approximation (60) is to be expected from the approximation (50), since by symmetry, when the initial frequency of A_1 is $\frac{1}{2}$, the conditioning should have no effect on the mean fixation time. On the other hand, the approximations (59) and (61) provide new information. The approximation (59) shows that, while when the initial frequency of A_1 is $(2N)^{-1}$ it is very unlikely that fixation of A_1 will occur, in the small fraction of cases when fixation of A_1 does occur, an extremely long fixation time may be expected. The approximation (61) shows that, while when the initial frequency of A_1 is $(2N)^{-1}$, so that loss of A_1 is likely to occur, a quite small loss time may be expected.

Analogous arguments show that if the condition is made that eventually A_1 is lost from the population, the Taylor series approximation $\bar{t}^{**}(p)$ for this conditional mean time for this to occur is given by

$$\bar{t}^{**}(p) = \frac{-4Np\log p}{1-p}. \tag{62}$$

Since the probability that A_1 is eventually fixed in the population is p and the probability that it is eventually lost is $1-p$, the equation $\bar{t} = p\bar{t}^* + (1-p)\bar{t}^{**}$ must hold, and a comparison of (47), (57) and (62) shows that this is the case.

Further conclusions will be given later when we consider the full diffusion approximation to the Wright–Fisher model (35).

Some remarks about the Wright-Fisher model

Perhaps unfortunately, the simple Wright–Fisher model has assumed a "gold standard" status, and serves as a reference distribution for several calculations in population genetics theory. This has arisen largely for historical reasons, and the fact that this is only one model among many, and is far less general and plausible than the Cannings model, to be discussed later, is seldom mentioned. We mention two examples where the fact that the Wright–Fisher model is no more than a reference model has been often overlooked, with unfortunate consequences.

First, the concept of the "effective population size", discussed in more detail later, is defined with reference to the simple Wright–Fisher model (35). A certain model has effective population size N_e if some characteristic of the model has the same value as the corresponding characteristic for the simple Wright–Fisher model (35) whose actual size is N_e. Further, the comparison of several characteristics are possible, and this leads to different concepts, or varieties, of the effective population size. Except in simple cases, the concept is not directly related to the actual size of a population. For example, a population might have an actual size of 200 but, because of a distorted sex ratio, have an effective population size of only 25. This implies that some characteristic of the model describing this population, for example a leading eigenvalue, has the same numerical value as that of a Wright–Fisher model with a population size of 25. It would be more useful if the adjective "effective" were replaced by "in some given respect Wright–Fisher model equivalent". Misinterpretations of effective population size calculations frequently follow from a misunderstanding of this fact.

Second, the fundamental genetic parameter θ will be introduced later in further discussion of the Wright–Fisher model and in other models. For the Wright-Fisher model θ assumes the value $4Nu$, and the identification of θ and $4Nu$ is very common in the literature. However, for models other than Wright–Fisher models a different definition of θ is needed. This is particularly true of the Cannings model as well as of the Moran model, both of which are discussed

later. The identification of θ with $4Nu$ arises in effect from an inappropriate assumption that the simple Wright–Fisher model (35) is the stochastic evolutionary model relevant to the situation at hand. The rather more general definition of θ as $4N_eu$ partly overcomes this problem, but not entirely, since (as mentioned above) there are several distinct concepts of the effective population size N_e.

Mutation

One-way mutation

Suppose now that A_1 mutates to A_2 at rate u but that there is no mutation from A_2 to A_1. It is then reasonable to replace the model (35) by

$$p_{ij} = \binom{2N}{j} (\psi_i)^j (1-\psi_i)^{2N-j} \tag{63}$$

where $\psi_i = i(1-u)/2N$. Here interest centers on properties of the time until A_1 is lost, either using an eigenvalue approach or mean time properties. For the moment we consider only mean time properties and note that an argument parallel to that leading to (46) shows that, to a first approximation, the mean time $\bar{t}(x)$, given a current frequency x, satisfies the approximating equation

$$-ux\frac{d\bar{t}(x)}{dx} + \frac{x(1-x)}{4N}\frac{d^2\bar{t}(x)}{dx^2} = -1. \tag{64}$$

If initially $x = p$, the solution of this equation, subject to the requirements $\bar{t}(0) = 0$,

$$\lim_{x\to 1} \bar{t}(x) \text{ is finite,}$$

is

$$\bar{t}(p) = \int_0^1 t(x,p)\,dx \tag{65}$$

where for $\theta(= 4Nu) \neq 1$,

$$t(x;p) = 4Nx^{-1}(1-\theta)^{-1}\{(1-x)^{\theta-1} - 1\}, \quad 0 < x \leq p, \tag{66}$$

$$t(x;p) = 4NKx^{-1}(1-\theta)^{-1}(1-x)^{\theta-1}, \quad p \leq x \leq 1, \tag{67}$$

where $K = 1 - (1-p)^{1-\theta}$.

The corresponding formula for the case $\theta = 1$ is found from (66) and (67) by standard limiting processes. It may be shown that with the definition of $t(x,p)$ in (66) and (67), $\bar{t}(p)$ may be written as

$$\bar{t}(p) = \sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}\Big(1-(1-p)^j\Big). \tag{68}$$

This is of course only an approximate formula, so more accurately we should write

$$\bar{t}(p) \approx \sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}\Big(1-(1-p)^j\Big). \tag{69}$$

The function $t(x,p)$ defined in (66) and (67) is more informative than it initially appears since, as we see later, $t(x,p)\delta x$ provides an excellent approximation to the mean number of generations for which the frequency of A_1 takes a value in $(x, x+\delta x)$ before reaching zero.

For the application of the above theory to the "infinitely many alleles" model and to molecular population genetics, two particular values of p have to be considered. The first of these is the values $p = 1/(2N)$. For this value of p the quantity K in (67) is approximately $(1-\theta)/(2N)$, and equations (66) and (67) give, approximately,

$$t(x;\frac{1}{2N}) = 4N, \quad 0 < x \le \frac{1}{2N}, \tag{70}$$

$$t(x;\frac{1}{2N}) = 2Nx^{-1}(1-x)^{\theta-1}, \quad \frac{1}{2N} \le x \le 1, \tag{71}$$

This leads to the approximation

$$\bar{t}(\frac{1}{2N}) \approx 2\big(1 + \int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{\theta-1}\,dx\big). \tag{72}$$

The second important value of p is when $p = 1$. For this value the approximation (69) gives, immediately,

$$\bar{t}(1) \approx \sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}. \tag{73}$$

The value $\theta = 2$ has no particular biological significance. It does however lead to one remarkable calculation. For this value of θ, (69) gives

$$\bar{t}(p) \approx \frac{-4Np \log p}{1-p} \tag{74}$$

when $p \neq 1$, and by a straightforward limiting process,

$$\bar{t}(p) \approx 4N \tag{75}$$

when $p = 1$. This value also follows directly from (73) when $\theta = 2$.

Note that the formulae (74) and (62) are identical. This is unexpected, since one formula applies for a conditional process without mutation, whereas the other applies for an unconditional process with mutation. It can be shown (see the discussion in the paragraph following equation (298) below) that the entire properties of the two processes (and not simply these two mean times) are identical.

Two-way mutation

Suppose next that A_2 also mutates to A_1 at rate v. It is now reasonable to define ψ_i in (63) by

$$\psi_i = \{i(1-u) + (2N-i)v\}/2N. \tag{76}$$

There now exists a stationary distribution $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_{2N})$ for the number of A_1 genes, where ϕ_i is the stationary probability that the number of A_1 genes takes the value i. The exact form of this distribution is complex and is not known, and we consider later the diffusion approximation to it. On the other hand, certain properties of this distribution can be extracted from (63) and (76). The stationary distribution satisfies the equation $\boldsymbol{\phi}' = \boldsymbol{\phi}' P$, where P is defined by (63) and (76), so that if ξ is a vector with ith element i $(i = 0, 1, 2, \ldots, 2N)$ and μ is the mean of the stationary distribution,

$$\mu = \boldsymbol{\phi}'\xi = \boldsymbol{\phi}' P \xi.$$

The ith $(i = 0, 1, 2, \ldots, 2N)$ component of $P\xi$ is

$$\sum j \binom{2N}{j} \psi_i^j (1-\psi_i)^{2N-j}$$

and from the standard formula for the mean of the binomial distribution, this is $2N\psi_i$ or

$$i(1-u) + (2N-i)v.$$

Thus,

$$\begin{aligned}\boldsymbol{\phi}' P\xi &= \sum\{i(1-u)+(2N-i)v\}\alpha_i \\ &= \mu(1-u)+v(2N-\mu).\end{aligned}$$

It follows that

$$\mu = (1-u)\mu + v(2N-\mu)$$

or

$$\mu = 2Nv/(u+v). \tag{77}$$

Similar arguments show that the variance σ^2 of the stationary distribution is

$$\sigma^2 = 4N^2uv/\{(u+v)^2(4Nu+4Nv+1)\} + \text{small order terms.} \tag{78}$$

Further moments can also be found, but we do not pursue the details. The above values are sufficient to answer a question of some interest in population genetics, namely "what is the probability of two genes drawn together at random are of the same allelic type?" If the frequency of A_1 is x and terms of order N^{-1} are ignored, this probability is $x^2+(1-x)^2$. The required value is the expected value of this taken over the stationary distribution, namely

$$\mathrm{E}\{x^2+(1-x)^2\} = 1 - 2\mathrm{E}(x) + 2\mathrm{E}(x^2).$$

If $u = v$, $4Nu = \theta$, equation (77) and equation (78) together show that this is

$$\text{Prob (two genes of same allelic type)} \approx (1+\theta)/(1+2\theta). \tag{79}$$

This expression is more revealing than might originally be thought. Since $u = v$, there is a complete symmetry between the properties of A_1 and A_2 genes. One might then have thought that the probability in (79) should be 1/2. However, this is not the case. Indeed, if θ is small this probability is close to 1. The reason for this is that small values of θ correspond to very low mutation rates. For such low rates, the most likely situation, at any time, is that the number of A_1 genes in the population is likely to be close to 0 or close to $2N$. In both cases the probability that two genes drawn at random from the population are of the same allelic type is close to 1. We confirm this observation later when discussion the stationary distribution of the number of A_1 genes.

The probability in (79) can be arrived at in another way, which we now consider since it is useful for purposes of generalization. Let the required probability be F and note that this is the same in two consecutive stationary generations. Two genes drawn at random in any generation will have a common parent gene with probability $(2N)^{-1}$, or different parent genes with probability $1-(2N)^{-1}$, which will be of the same allelic type with probability F. The probability that neither of the genes drawn is a mutant, or that both are, is $u^2+(1-u)^2$, while the probability that precisely one is a mutant is $2u(1-u)$. It follows that

$$\begin{aligned} F &= \{u^2+(1-u)^2\}\{\frac{1}{2N}+F(1-\frac{1}{2N})\} \\ &\quad + 2u(1-u)(1-F)(1-\frac{1}{2N}). \end{aligned}$$

Thus exactly

$$F = \frac{1+2u(1-u)(2N-2)}{1+4u(1-u)(2N-1)},$$

and approximately

$$F \approx (1+\theta)/(1+2\theta), \tag{80}$$

in agreement with (79). We later consider a third approach which yields the same answer.

The eigenvalues of the model defined jointly by (63) and (76) are known. They are best found by a very general approach due to Cannings, which finds eigenvalues in a wider range of models than that defined by (63) and (76). We now turn to these models.

The Cannings (exchangeable) model

An important generalization of the Wright–Fisher form of model is due to Cannings (1974). We consider a "population" of genes of fixed size $2N$, reproducing at time points $t = 0, 1, 2, 3, \ldots$. The stochastic rule determining the population structure at time $t+1$ is quite general, provided that any subset of genes at time t has the same joint probability distribution of "offspring" genes at time $t+1$ as any other subset of the same size. Thus, if the ith gene leaves Y_i offspring genes, we require only that $Y_1 + \ldots + Y_{2N} = 2N$ and that the joint distribution of $Y_i, Y_j, \ldots, Y_k$ be independent of the

choice of $i, j, \ldots, k$. In particular all genes must have the same offspring probability distribution. This distribution must have mean 1, by symmetry, and we denote the variance of this distribution by σ^2. **This interpretation of σ^2 is used throughout these notes when Cannings models are considered.** In some Cannings models a gene present at time t can also be present at time $t+1$, and is then counted as one of its own offspring. An example of this is discussed later.

The Wright–Fisher model (35) is clearly a particular case of the Cannings model since in the model (35) $Y_1, Y_2, \ldots, Y_{2N}$ have a symmetric multinomial distribution. Thus the Cannings model is more general than the Wright–Fisher model, and by choosing σ^2 appropriately it can be made much more realistic than the Wright–Fisher model.

Our first calculation concerning the Cannings model relates to eigenvalues. Let the genes be divided into two allelic classes, A_1 and A_2, as for the Wright–Fisher model (35), and denoting as below the number of A_1 genes at time t by $X(t)$, Then we have

Theorem (Cannings, (1974)). If

$$p_{ij} = \text{Prob}\{X(t+1) = j \mid X(t) = i\}, \quad i, j = 0, 1, 2, \ldots, 2N,$$

then the eigenvalues of the matrix $\{p_{ij}\}$ are

$$\lambda_0 = 1, \quad \lambda_j = \text{E}(Y_1 Y_2 \cdots Y_j), \quad j = 1, 2, \ldots, 2N. \tag{81}$$

Further, $\lambda_0 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_{2N}$.

As noted above, in the Wright–Fisher model (35), any set Y_1, Y_2, …, Y_j has a multinomial distribution with index $2N$ and common parameter $(2N)^{-1}$. This implies that if we write

$$\frac{(2N)!}{y_1! y_2! \ldots y_j! (2N - y_1 - \ldots - y_j)!} = \binom{n}{\mathbf{y}},$$

the eigenvalue $\lambda_j, j = 1, 2, \ldots, 2N$ is given by

$$\begin{aligned} \lambda_j &= \sum \cdots \sum y_1 y_2 \ldots y_j \binom{n}{\mathbf{y}} \left(\frac{1}{2N}\right)^{\sum y_i} \left(1 - \frac{j}{2N}\right)^{2N - \sum y_i} \\ &= (2N)(2N-1)\ldots(2N-j+1)/(2N)^j. \end{aligned} \tag{82}$$

This confirms the values given in (38), which were found originally (Feller, (1951)) by other methods.

The theorem shows that, for the Cannings model, the leading non-unit eigenvalue is $\lambda_2 = \mathrm{E}(Y_1 Y_2)$. Now $\sum Y_j \equiv 2N$, so that the variance of $\sum Y_j$ is 0. Then by symmetry,

$$2N \ \mathrm{var}(Y_i) + 2N(2N-1) \ \mathrm{covar}(Y_i, Y_j) = 0.$$

This implies that

$$\mathrm{covar}(Y_i, Y_j) = -\sigma^2/(2N-1), \tag{83}$$

where $\sigma^2 = \mathrm{var}(Y_i)$. From this,

$$\begin{aligned} \lambda_2 &= \mathrm{E}(Y_1 Y_2) \\ &= \mathrm{Covar}(Y_1, Y_2) + \mathrm{E}(Y_1)\mathrm{E}(Y_2) \\ &= 1 - \sigma^2/(2N-1). \end{aligned} \tag{84}$$

To confirm this formula we observe that, in the Wright–Fisher model, Y_i has a binomial distribution with index $2N$ and parameter $(2N)^{-1}$. Thus $\sigma^2 = (2N-1)/(2N)$ and

$$\lambda_2 = 1 - \{2N-1\}/\{2N(2N-1)\} = 1 - (2N)^{-1},$$

agreeing with the "$j = 2$" case in the expression in equation (82).

Other properties of the Cannings model follow easily. For example, by symmetry, the probability of eventual fixation of any allele in such a model must be its initial frequency. Further, suppose that there are $X(t)$ A_1 genes in the Cannings model at time t, and write $X(t) = i$ for convenience. If we relabel genes so that the first i genes are of allelic type A_1 and the remaining genes of allelic type A_2,

$$\begin{aligned} \mathrm{Var}\{X(t+1) \mid X(t)\} &= \mathrm{Var}(Y_1 + \ldots + Y_i) \\ &= i\sigma^2 + i(i-1) \ \mathrm{Covar}(Y_1, Y_2) \\ &= i(2N-i)\sigma^2/(2N-1), \end{aligned} \tag{85}$$

from equation (83). If $x(t) = X(t)/2N$, it follows that

$$\mathrm{var}\{x(t+1) \mid x(t)\} = x(t)\{1-x(t)\}\sigma^2/(2N-1). \tag{86}$$

This equation shows that the rate of loss of heterozygosity in the population is directly proportional to σ^2. As an extreme case, suppose that one individual is chosen at random from the population to produce *all* the offspring in the next generation. It follows

that $x(t+1)$ is either 1 (with probability $x(t)$) or 0 (with probability $1-x(t)$). The variance of $x(t+1)$ is then easily shown to be $x(t)\{1-x(t)\}$. We now check that this result is given by (86). The number of offspring of a randomly chosen individual is, in the model considered, either 0 (with probability $1-1/(2N)$) or $2N$ (with probability $1/(2N)$). The variance σ^2 of this offspring distribution is easily seen to be $2N-1$, and inserting this value into (86) we do indeed obtain the result $x(t)\{1-x(t)\}$.

We now introduce mutation, and assume that A_1 genes mutate to A_2 at rate u, with reverse mutation at rate v. We also assume that if mutation does not exist the conditions for Theorem 1 above hold. Cannings was able to find an expression for the eigenvalues in this "mutation" model. The details are complex and here we only give the results for particular cases. For the Wright–Fisher model, the eigenvalues of the matrices defined by (63) and (76) are $\lambda_0 = 1$ and, for $j = 1, 2, \ldots, 2N$,

$$\lambda_j = (1-u-v)^j [2N(2N-1)\cdots(2N-j+1]/(2N)^j. \qquad (87)$$

The leading non-unit eigenvalue λ_1 is $1-u-v$ and is thus independent of N. This is extremely close to unity and suggests a very slow rate of approach to stationarity in this model. These eigenvalues apply also in any Cannings one-way mutation model, for which we simply put $v = 0$ in (87).

The Cannings model has an interesting relationship with another class of models frequently used in population genetics, namely conditional branching process model. (These are in fact also a particular case of the Cannings model.) In the conditional branching process model it is supposed that each gene produces k offspring with probability f_k ($k = 0, 1, 2, 3, \ldots$), with the numbers of offspring from different parents being assumed independent. If $f(s) = \sum f_i s^i$, the generating function of the distribution of the total number of offspring genes is $[f(s)]^{2N}$. We now make the condition that the total number of such offspring is $2N$. If at time t there were i A_1 genes, the probability p_{ij} that at time $t+1$ there will be j A_1 genes is, for these models,

$$p_{ij} = \frac{\text{coeff } t^j s^{2N} \text{ in } [f(ts)]^i [f(s)]^{2N-i}}{\text{coeff } s^{2N} \text{ in } [f(s)]^{2N}}. \qquad (88)$$

Transition probabilities of this form were introduced by Moran and Watterson (1958) who used them to find explicit expressions for the

leading non-unit eigenvalue in dioecious populations with various family structures.

Karlin and McGregor (1965) analyzed the conditional branching process model in detail. They show in particular that the eigenvalues of the matrix $\{p_{ij}\}$ are

$$\lambda_0 = \lambda_1 = 1,\ \lambda_k = \frac{\text{coeff } s^{2N-k} \text{ in } [f(s)]^{2N-k}[f'(s)]^k}{\text{coeff } s^{2N} \text{ in } [f(s)]^{2N}},\ k = 2, 3, \ldots, 2N. \tag{89}$$

These must agree with the values found in equation (81), since a conditional branching process is a Cannings model. We check that this agreement holds for the eigenvalue λ_2. It is clear from (88) that

$$\sum_j p_{ij} t^j = \frac{\text{coeff } s^{2N} \text{ in } [f(ts)]^i [f(s)]^{2N-i}}{\text{coeff } s^{2N} \text{ in } [f(s)]^{2N}}\,.$$

Differentiating twice with respect to t and putting $t = 1$,

$$\sum_j j(j-1)p_{ij} = \lambda_2 i(i-1) + \eta_2 i, \tag{90}$$

where λ_2 is defined by equation (89) and η_2 is some constant independent of i and j. Now $\sum j p_{ij} = i$ by symmetry, and $\sum j(j-1)p_{1j} = \sigma^2$, where σ^2 is defined after (83). Thus putting $i = 1$ in (90) we get $\eta_2 = \sigma^2$ and then putting $i = 2$,

$$\sum_j j(j-1)p_{2j} = 2\lambda_2 + 2\sigma^2$$

so that

$$\sum_j j^2 p_{2j} = 2\lambda_2 + 2\sigma^2 + 2. \tag{91}$$

But the left-hand side in (91) is $E(Y_1+Y_2)^2$, where Y_i is the (random) number of offspring genes left by parental gene i. It follows that

$$2 + 2\sigma^2 + 2E(y_1 y_2) = 2\lambda_2 + 2\sigma^2 + 2$$

or

$$\lambda_2 = E(Y_1 Y_2),$$

as required. Parallel calculations can be made for the remaining eigenvalues, but we do not pursue the details here.

The Moran model

Introduction

The conclusions reached so far apply only for the Wright–Fisher model and, more generally, to the Cannings model. Different conclusions are reached for models other than these, and we consider now a rather different model, due to Moran (1958). Moran's model has the additional advantage of allowing explicit expressions for many quantities of evolutionary interest, although, strictly, it applies only for haploid populations.

Consider then a haploid population in which, at time points $t = 1$, 2, 3, ..., an individual is chosen at random to reproduce. After reproduction has occurred, an individual is chosen to die (possibly the reproducing individual but not the new offspring individual).

As is discussed later, the model can be generalized by allowing mutation. We consider first the simplest case where there is no mutation. Suppose the population consists of $2N$ haploid individuals (we use this notation to allow direct comparison with the diploid case), each of whom is either A_1 or A_2. Suppose also that, at time t, the number of A_1 individuals is i. Then at time $t+1$ there will be $i-1$ A_1 individuals if an A_2 is chosen to give birth and an A_1 individual is chosen to die. The probability of this, under our assumptions, is

$$p_{i,i-1} = i(2N - i)/(2N)^2. \tag{92}$$

Similar reasoning shows that

$$p_{i,i+1} = i(2N - i)/(2N)^2, \tag{93}$$

$$p_{i,i} = \{i^2 + (2N - i)^2\}/(2N)^2. \tag{94}$$

The matrix defined by these transition probabilities is a continuant, so that the standard theory of continuant Markov chain transition matrices can be applied to it. In the standard notation of continuant matrices, $p_{i,i-1} = \mu_i, p_{i,i+1} = \lambda_i$, and

$$\rho_0 = 1, \rho_k = \frac{\mu_1\mu_2\cdots\rho_k}{\lambda_1\lambda_2\cdots\lambda_k}.$$

Thus for the Moran model above,

$$\lambda_i = \mu_i = i(2N - i)/(2N)^2, \qquad \rho_i = 1, \quad i = 0, 1, 2, \ldots, 2N. \tag{95}$$

It follows either from continuant Markov chain theory, or from the kind of argument used above for the Wright–Fisher model, that the probability π_i of eventual fixation of A_1, given currently i A_1 individuals, is

$$\pi_i = i/2N. \tag{96}$$

Further, continuant Markov chain theory shows, using notation developed above, that if initially there are iA_1 genes, the mean number $\bar{t}_{ij}$ of birth-death events at which there are jA_1 genes is given by

$$\begin{aligned} \bar{t}_{ij} &= 2N(2N-i)/(2N-j), \quad j = 1, 2, \ldots, i, \\ \bar{t}_{ij} &= 2Ni/j, \quad j = i+1, \ldots, 2N-1. \end{aligned} \tag{97}$$

Thus, immediately, the mean number $\bar{t}_i$ of birth-death events until loss or fixation of A_1, given initially iA_1 genes, is given by

$$\bar{t}_i = 2N(2N-i)\sum_{j=1}^{i}(2N-j)^{-1} + 2Ni\sum_{j=i+1}^{2N-1} j^{-1}. \tag{98}$$

Further, given that A_1 is eventually fixed,

$$\begin{aligned} \bar{t}^*_{ij} &= 2N(2N-i)j/\{i(2N-j)\}, \quad j = 1, 2, \ldots, i, \\ \bar{t}^*_{ij} &= 2N, \quad j = i+1, \ldots, 2N-1, \end{aligned} \tag{99}$$

$$\bar{t}^*_i = 2N(2N-i)i^{-1}\sum_{j=1}^{i} j(2N-j)^{-1} + 2N(2N-i-1) \tag{100}$$

An interesting example of these formulae is the case $i = 1$, corresponding to a unique A_1 mutant in an otherwise purely A_2 population. Here $\bar{t}^*_{1j} = 2N$ for all j so that, given that the mutant is eventually fixed, the number of A_1 genes takes, on average, each of the values $1, 2, \ldots, 2N-1$ a total of $2N$ times. The conditional mean fixation time is given by

$$\bar{t}^*_1 = 2N(2N-1) \tag{101}$$

birth-death events. The variance of the conditional absorption time can also be written down but we do not do so here.

The exact results found above for the Moran model are unwieldy, so we now give simple approximate expression for the most important conclusion derived from them. It is evident from the expression (98) that

$$\bar{t}(p) \approx -(2N)^2\{p\log p + (1-p)\log(1-p)\}, \tag{102}$$

where $p = i/2N$. The similarity between this formula and (47) is interesting. A factor of $2N$ may be allowed in comparing the two to convert from birth-death events to generations. There remains a further factor of 2 to explain, and we show later why this factor exists.

The eigenvalues of the Moran model transition matrix can be found by using Cannings' Theorem. Take any collection of j genes and note that the probability that one of these is chosen to reproduce is $j/2N$, with the same probability that one is chosen to die. For this model a gene can be (and indeed usually is) one of its own "offspring". Using the notation of Cannings' Theorem, the product $Y_1 Y_2 \ldots Y_j$ can take only three values:

0 if one of these genes is chosen to die and the gene so chosen is not chosen to reproduce,

2 if one of the genes is chosen to reproduce and none is chosen to die,

1 otherwise.

Thus $\lambda_0 = 1$ and

$$\begin{aligned} \lambda_j &= \mathrm{E}(Y_1 Y_2 \ \ldots \ Y_j) \\ &= 2j(2N-j)/(2N)^2 + 1 - j(4N-j-1)/(2N)^2 \\ &= 1 - j(j-1)/(2N)^2, \quad j = 1, 2, \ldots, 2N. \end{aligned} \tag{103}$$

The largest non-unit eigenvalue is

$$\lambda_2 = 1 - 2/(2N)^2. \tag{104}$$

The fact that λ_2 is very close to unity agrees with the very large mean absorption times (98) for intermediate values of i.

It is possible to find the right eigenvector $\mathbf{r}$ and left eigenvector ℓ' corresponding to this eigenvalue. They are given by

$$\begin{aligned} \mathbf{r} &= (0, 1(2N-1), 2(2N-2), \ldots, i(2N-i), \ldots, 1(2N-1), 0)' \\ \ell' &= \left(-\tfrac{1}{2}(2N-1), 1, 1, 1, \ldots, 1, -\tfrac{1}{2}(2N-1)\right). \end{aligned}$$

One-way mutation in the Moran model

If mutation from A_1 to A_2 is allowed (at rate u), with no reverse mutation, A_1 must eventually become lost, and interest centers on

properties of the time for this to occur. The model defined in (92) – (94) is now amended to

$$\begin{aligned} p_{i,i-1} &= \{i(2N-i)+ui^2\}/(2N)^2 = \mu_i \\ p_{i,i+1} &= i(2N-i)(1-u)/(2N)^2 = \lambda_i \\ p_{i,i} &= 1-p_{i,i-1}-p_{i,i+1}. \end{aligned} \tag{105}$$

Continuant Markov chain theory can now be used to find explicit exact $\bar{t}_{ij}$ and thus $\bar{t}_i$. We do not present these here since it will be more useful (see (109) - (113) below) to proceed via approximations.

Two-way mutation in the Moran model

If mutation from A_2 to A_1 (at rate v) is also allowed, the model becomes

$$\begin{aligned} p_{i,i-1} &= \{i(2N-i)(1-v)+ui^2\}/(2N)^2 = \mu_i \\ p_{i,i+1} &= \{i(2N-i)(1-u)+v(2N-i)^2\}/(2N)^2 = \lambda_i \\ p_{i,i} &= 1-p_{i,i-1}-p_{i,i+1}. \end{aligned} \tag{106}$$

The typical value ϕ_j in the stationary distribution $\boldsymbol{\phi}$ for the number of A_1 genes is found to be

$$\phi_j = \phi_0 \frac{(2N)!\Gamma\{j+A\}\Gamma\{B-j\}}{j!(2N-j)!\Gamma\{A\}\Gamma\{B\}} \tag{107}$$

where $\Gamma\{\cdot\}$ is the well-known gamma function, $A = 2Nv/(1-u-v)$, $B = 2N(1-v)/(1-u-v)$, $C = 2Nu/(1-u-v)$, $D = 2N/(1-u-v)$ and $\alpha_0 = \Gamma\{B\}\Gamma\{A+C\}/[\Gamma\{D\}\Gamma\{C\}]$. Although these expressions are exact they are rather unwieldy, and we consider in a moment a simple approximation to ϕ_j.

The Markov chain defined by (106), having a stationary distribution and a continuant transition matrix, is automatically reversible. This is not necessarily true for other genetical models: for example it can be shown that the Wright–Fisher Markov chain defined jointly by (63) and (76) is not reversible. What does reversibility mean in genetical terms? All the theory we have considered so far is *prospective*, that is, given the current state of a Markov chain, probability statements are made about its future behavior. Recent developments in population genetics theory often concern the *retrospective* behavior: the present state is observed, and questions

are asked about the evolution leading to this state. For reversible processes these two aspects have many properties in common, and information about the prospective behavior normally yields almost immediately useful information about the retrospective behavior. We shall see later how the identity of prospective and retrospective probabilities can be used to advantage in discussing various evolutionary questions.

The eigenvalues of the transition matrix defined by (106) can be found by applying the Cannings theory for cases involving mutation. It is found after some calculation for these eigenvalues that $\lambda_0 = 1$ and

$$\lambda_j = 1 - \frac{j(u+v)}{(2N)} - \frac{j(j-1)(1-u-v)}{(2N)^2}, \quad j = 1, \ldots, 2N. \quad (108)$$

These eigenvalues apply also in the case $v = 0$. The leading non-unit eigenvalue is $1-(u+v)/(2N)$, and since $2N$ time units in the process we consider may be thought to correspond to one generation in the Wright–Fisher model, this agrees closely with the value $1 - u - v$ found in (87) in that model.

We now consider approximations for several of the above quantities. First, we recall the result given in (102) for the mean number of birth-death events before loss of genetic variation when there is no mutation. In the case of the one-way mutation model (105), an approximating expression for the mean time for loss of A_1 in this model, given initially that there are i A_1 genes, (that is, $\bar{t}_i$), for the case $\theta \neq 1$ is

$$\begin{aligned} \bar{t}_i \;\approx\; & (2N)^2(1-\theta)^{-1}\Big(\int_0^p x^{-1}\{(1-x)^{\theta-1}-1\}dx \\ & + \int_p^1 x^{-1}(1-x)^{\theta-1}\{1-(1-p)^{1-\theta}\}\,dx\Big) \end{aligned} \quad (109)$$

birth-death events, where $p = i/(2N)$, $x = j/(2N)$ and θ is defined (for this formula) for as $2Nu$. (The more appropriate and exact expression $2Nu/(1-u)$ for θ in the Moran model given in (152) below. Since u is very small, the value $2Nu$ is very close to this expression, so that when considering approximation formulae, we use this approximating value.) Clearly (109) does not apply for the

case $\theta = 1$, and the appropriate formula for this value of θ is given below.

The particular case $p = (2N)^{-1}$ will be of interest to us later in the context of the infinitely many alleles model. For this value of p, the expression (109) reduces, to a close approximation, to

$$\bar{t}_i \approx 2N\big(1 + \int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{\theta-1}\,dx\big) \tag{110}$$

birth-death events.

The two cases $\theta = 1, \theta = 2$ have no particular biological relevance. However, it is interesting to consider the expression for $\bar{t}_i$ in these two cases, since the expressions for $\bar{t}_i$ simplify for them. The expression given in (109) does not apply for the case $\theta = 1$, and a limiting ($\theta \to 1$) calculation is needed. This gives

$$\begin{aligned} \bar{t}_i &\approx -(2N)^2 \int_0^p x^{-1}\log(1-x) + (2N)^2 \log p \log(1-p), \quad p \neq 1, \\ \bar{t}_i &\approx 2\pi^2 N^2/3, \quad p = 1. \end{aligned} \tag{111}$$

For $\theta = 2$, (109) gives

$$\bar{t}_i \approx -(2N)^2 \frac{p\log p}{(1-p)}, \quad p \neq 1, \tag{112}$$

$$\bar{t}_i \approx (2N)^2, \quad p = 1. \tag{113}$$

It is interesting to compare the values in (111) and (113) when $p = 1$. The value in (113) is about 60% of the value in (111), and this shows the effect of doubling the mutation rate in speeding up the loss of the allele A_1.

We consider finally an approximation to the stationary distribution in the two-way mutation model, given exactly in (107). We put $x = j/(2N)$, $u = \alpha/(2N)$, $v = \beta/(2N)$ and let j and $2N$ increase indefinitely with x, α and β fixed. Using the Stirling approximation $\Gamma\{y + a\}/\Gamma\{y\} \sim y^a$ for large y, moderate a, the stationary probability ϕ_j in (107) becomes, approximately,

$$\phi_j \sim (2N)^{-1} \frac{\Gamma\{\alpha+\beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}} x^{\beta-1}(1-x)^{\alpha-1}, \tag{114}$$

at least for values of x not extremely close to 0 or 1. Clearly this approximation expression is far simpler than the exact value (107), and that is why this approximation is often used. It also confirms a conclusion reached above, that if mutation rates are very low, so the both α and β are very small, the stationary distribution is U-shaped, so that it is most likely that the number of A_1 genes in any given generation is either extremely small or extremely close to $2N$.

M-allele models

The models considered so far can easily be extended to allow M different alleles at the locus in question, where M is an arbitrary positive integer. (For the ABO blood group system, for example, there are three possible alleles, A, B and O at the gene locus for this blood group, so that for this case $M = 3$.) For an M-allele model the population configuration at any time can be described by a vector $(X_1, X_2, \ldots, X_M)$, where X_i is the number of genes of allelic type A_i at that time. We assume, as for the "two-allele" model considered above, that the population size is fixed at the value N in all generations, so that $X_1 + X_2 + \ldots + X_M = 2N$. Thus only $M-1$ elements in the above vector are independent, but for reasons of symmetry we retain all elements in this vector.

The first case to consider is that where there is no mutation, and our first task is to consider the M-allele generalizations of some of the results found above for the Wright-Fisher, the Cannings and the Moran models. In doing this, it is sometimes convenient to consider some specific allele, say the allele A_i, on its own, all other alleles being classed as non-A_i, and if this is done some of the above theory can be applied. Some examples of this strategy are given below.

We consider first the M-allele generalization of the Wright–Fisher model (35), for which we assume that

$$\begin{aligned} &\Pr\{X_i(t+1) \text{ genes of allele } A_i \text{ at time } t+1 \mid X_i(t) \text{ genes of allele} \\ &\qquad\qquad i \text{ at time } t, \quad i = 1, 2, \ldots, M\} \\ &= \frac{(2N)!}{X_1(t+1)! \ldots X_M(t+1)!} \psi_1^{X_1(t+1)} \ldots \psi_M^{X_M(t+1)} \end{aligned} \tag{115}$$

where $\psi_i = X_i(t)/(2N)$.

The eigenvalues of the Markov chain transition matrix defined implicitly by (115) are the same as the values given in the expres-

sion (38), but for the M-allele model λ_j has multiplicity $(M+j-2)!/\{(M-2)!/j!\}$, $(j=2,3,\ldots,2N)$. The eigenvalue $\lambda_0=1$ has total multiplicity M. These eigenvalues have the interesting interpretation (Littler, (1975)) that

$$\Pr\{\text{at least } j \text{ allelic types remain present at time } t\} \sim \text{const } \lambda_j^t. \tag{116}$$

M-allele processes raise questions that are more complex than two-allele processes. In the M-allele process without selection or mutation, (that is the process discussed above), one of the M alleles will eventually become fixed in the population. It is easy to see, by considering any one allele (say A_i), and grouping all other alleles as "non-A_i", that the probability that any specified allele is eventually fixed is its initial frequency. On the other hand, some interesting "mean time" questions are not so simply found. For example, it is much harder to find the probability that a given allele is the first one lost from the population. It is also quite difficult to find the mean number of generations until some (unspecified) allele is fixed in the population, or (more difficult) until exactly two alleles exist in the population. We later use diffusion theory to approach some of these questions.

When mutation exists between all alleles there will exist a multidimensional stationary distribution of allelic numbers. The means, variances and covariances in this distribution can be found by procedures analogous to those leading to the expressions in (77) and (78). We consider in detail only the case where mutation is symmetric: here the probability that any gene mutates is assumed to be u, and given that a gene of allelic type A_i has mutated, the probability that the new mutant is of type A_j is $(M-1)^{-1}$, $(j\neq i)$. By symmetry, the mean number of genes of allelic type A_i alleles in the stationary distribution must be $2N/M$. However, it sometimes occurs that this is not a likely value for the actual number of genes of any allelic type to arise, and we see this best by finding the probability F that two genes taken at random from the population are of the same allelic type. Generalizing the argument that led to (80) we find, ignoring terms of order u^2, that

$$F=\big((2N)^{-1}+\{1-(2N)^{-1}\}F\big)(1-2u)+\big(1-(2N)^{-1}\big)(1-F)\big(2u/(M-1)\big).$$

If we write $\theta=4Nu$, this gives

$$F\approx(M-1+\theta)/(M-1+M\theta). \tag{117}$$

This expression agrees with that in (80) for the case $M = 2$. For large M,

$$F \approx (1+\theta)^{-1}, \tag{118}$$

an expression we return to later.

These formulas demonstrate the theme discussed below (79). In (79) and more generally in (118), if θ is small, then $F \approx 1$. This implies that it is very likely that one or other allele appears with high frequency with the remaining alleles having negligible frequency, despite the fact that all alleles are selectively equivalent. The imbalance arises because of stochastic effects, and is quite different from that predicted by considering the mean allele frequencies only.

The eigenvalues of the matrix defined by the symmetric mutation model are found from the values (82) if λ_j is multiplied by

$$\{1 - \frac{uM}{(M-1)}\}^j.$$

The multiplicity of λ_j is, as in the "no mutation" case, $(j + M - 2)!/\{j!(M-1)!\}$.

In view of the comments concerning the Cannings model made above, it is plausible that the approximations (117) and (118) hold for that model, with however θ now defined by $\theta = 4Nu/\sigma^2$. There is also an M-allele Moran model which allows various exact formulae, but we do not consider this here.

Infinitely many alleles models

Introduction

Infinitely many alleles models were inspired by the knowledge of the gene as a sequence of nucleotides. There are four possible nucleotides at each site in this sequence, a, g, c and t, and an "allele" is simply one specific sequence, such as *tccgagtgcat...tc*. In a typical gene, consisting of a sequence of 3000 nucleotides, there are 4^{3000} possible sequences, that is 4^{3000} possible alleles. For essentially all practical purposes we may take this number as infinity, thus leading to the infinitely many alleles model. Thus this model is one of *molecular population genetics*, since it is inspired by knowledge of the molecular nature of the gene. Some comments about molecular population genetics theory are given in the next subsection.

Another model inspired by the knowledge of the gene as a sequence of nucleotides is the "infinitely many sites" model, described in detail in these lectures by Dr Joyce. However, some aspects of this model are discussed in these notes also.

In this section we consider both population and sample properties of the "infinitely many alleles" versions of the Wright–Fisher, the Cannings and the Moran models. The discussion of the Wright–Fisher model is more extensive than that for the Cannings model. This arises for two reasons. The first is that calculations for the Wright–Fisher model are comparatively straightforward, and the second is that results for this model can be taken over almost directly for the Cannings model, with an appropriate change in the definition of the parameter θ arising in all the formulas found, as hinted at the end of the previous section.

Results for the Wright–Fisher and the Cannings infinitely many alleles models are usually approximations. By contrast, the infinitely many alleles Moran model allows many exact calculations.

Mutation is intrinsic to all infinitely many alleles models, but the nature of the new mutants is different from anything assumed so far, the key difference being that all mutant genes are assumed to be of a new allelic type, not currently or previously seen in the population. This has several important implications that are discussed in detail below.

Some remarks about molecular population genetics

Before discussing infinitely many alleles models, we make some comments about molecular population genetics, since the statistical theory associated with molecular population genetics is carried out in terms of the infinitely many alleles, and also the infinitely many sites, models.

First, the theory considered so far in these notes concerns alleles given labels such as "A_1", "A_2", etc. These are simply arbitrary notations. However, at the molecular level, the actual genetic material is known, so that the nucleotide symbols a, g, c and t refer to actual rather than arbitrary entities. The fact that the molecular theory thus concerns ultimate and real entities is of great importance, and allows evolutionary inferences not possible with classical population genetics theory.

These inferences are based on samples, and thus properties of samples of genetic material are considered often in the following notes. These inferences relate to *retrospective* rather than *prospective* evolutionary questions. The theory given so far in these notes is largely prospective – given numerical values for various genetic parameters, the theory shows what the evolution (or in a stochastic model the likely evolution) of a population will be. By contrast, the aim of the retrospective theory is to describe the course that evolution has taken, and thus to gain empirical insight into evolutionary questions. This change of viewpoint leads to the use of statistical methods, analyzing current genetical data. The current emphasis on statistical inference procedures is the most important new direction in population genetics theory. Knowledge of the actual genetical material is essential for these inferences, and the entire retrospective analysis must therefore be carried out in the framework of molecular population genetics.

To help with the discussion of inferential properties through the use of samples, the sample size is denoted in these notes by n (genes) and the population size by N. Since a diploid population is assumed the number of genes in the population is $2N$. Because we often compare sample and population properties, we sometimes write suffices "n" and "$2N$" when appropriate (for example K_n and K_{2N} to distinguish between the number of different alleles in a sample and in the population).

The Wright–Fisher infinitely many alleles model

Population properties

The Wright–Fisher infinitely many alleles model follows the generic binomial sampling characteristic of all Wright–Fisher models. The nature of the mutation mechanism, as stated above, implies that if the mutation rate (always to new allelic types) is u, and if in generation t there are $X_i(t)$ genes of allelic type A_i ($i = 1, 2, 3, \ldots$), then the probability $\text{Prob}\{X_0(t+1), X_1(t+1), X_2(t+1), \ldots \mid X_1(t)X_2(t), \ldots\}$ that in generation $t+1$ there will be $X_i(t+1)$ genes of allelic type A_i, together with $X_0(t+1)$ new mutant genes, all by assumption of different and novel allelic types, is

$$\frac{(2N)!}{\Pi X_i(t+1)!}\,\Pi\pi_i^{X_i(t+1)}, \tag{119}$$

where $\pi_0 = u$ and $\pi_i = X_i(t)(1-u)/(2N)$, $i = 1, 2, 3, \ldots$.

This model differs fundamentally from previous mutation models (which allow reverse mutation) in that since each allele will sooner or later be lost from the population, there can exist no nontrivial stationary distribution for the frequency of any allele. Nevertheless we are interested in stationary behavior, and it is thus important to consider what concepts of stationarity exist for this model. To do this we consider delabeled configurations of the form $\{a, b, c, \ldots\}$, where such a configuration implies that there exist a genes of one allelic type, b genes of another allelic type, and so on. The specific allelic types involved are not of interest. The possible configurations can be written down as $\{2N\}$, $\{2N-1, 1\}$, $\{2N-2, 2\}$, $\{2N-2, 1, 1\}, \ldots, \{1, 1, 1, \ldots 1\}$ in dictionary order. Here, for example, the configuration $\{2N\}$ is that for which all $2N$ genes in the population are of the same (unspecified) allelic type, the configuration $\{2N-1, 1\}$ is that for which $2N-1$ genes in the population are of the same (unspecified) allelic type and the remaining gene is of some other type, and the configuration $\{1, 1, 1, \ldots 1\}$ is that for which all $2N$ genes in the population are of different allelic types. The number of such configurations is $p(2N)$, the number of partitions of $2N$ into positive integers. For small values of N values of $p(2N)$ are given by Abramowitz and Stegun (1965, Table 24.5), who also provide asymptotic values for large N. It is clear that (119) implies certain transition probabilities from one configuration to another. Although these probabilities are extremely complex and the Markov chain of configurations has an extremely large number of states, nevertheless standard theory shows that there exists a stationary distribution of configurations. Unfortunately, no explicit expression is known for this stationary distribution, and we now discuss some partial and approximate results.

We consider first the probability that two genes drawn at random from the population at stationarity are of the same allelic type. For this to occur neither gene can be a mutant and, further, both must be descended from the same parent gene (probability $(2N)^{-1}$) or different parent genes which were of the same allelic type. Writing $F_2^{(t)}$ for the desired probability in generation t, we get

$$F_2^{(t+1)} = (1-u)^2\big((2N)^{-1} + \{1-(2N)^{-1}\}F_2^{(t)}\big). \tag{120}$$

At stationarity $F_2^{(t+1)} = F_2^{(t)} = F_2$ and thus

$$F_2 = \{1 - 2N + 2N(1-u)^{-2}\}^{-1} \approx (1+\theta)^{-1}, \qquad (121)$$

where, as as is standard for Wright–Fisher models, θ is defined as $4Nu$. This is identical to the limiting ($K \to \infty$) value in (118). In view of the fact that there is no concept of the stationary distribution for the frequency of any allele in the infinitely many alleles case, this fact is perhaps surprising.

We consider further calculations of this kind when discussing sample properties.

We now turn to eigenvalue calculations. Equation (120) can be rewritten in the form

$$F_2^{(t+1)} - F_2^{(\infty)} = (1-u)^2\{1-(2N)^{-1}\}\{F_2^{(t)} - F_2^{(\infty)}\}, \qquad (122)$$

and this implies that $(1-u)^2\{1-(2N)^{-1}\}$ is an eigenvalue of the Markov chain configuration process discussed above. A similar argument using (135) shows that a second eigenvalue of the configuration process is $(1-u)^3\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}$. Equations (137) and (141) suggest that $(1-u)^4\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}\{1-3(2N)^{-1}\}$ is an eigenvalue of multiplicity 2. It is found more generally that

$$\lambda_i = (1-u)^i\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}\cdots\{1-(i-1)(2N)^{-1}\} \qquad (123)$$

is an eigenvalue of the configuration process matrix and that its multiplicity is $p(i) - p(i-1)$, where $p(i)$ is the partition number given above. This provides a complete listing of all the eigenvalues.

We consider next an approximation for the mean number of alleles existing in the population at any time. Any specific allele A_m will be introduced into the population with frequency $(2N)^{-1}$, and after a random number of generations will leave it, never to return. The frequency of A_m is a Markovian random variable with transition matrix given in (63), with ψ_i defined immediately below (63). There will exist a mean time $\mathrm{E}(T)$, measured in generations, that A_m remains in the population. The mean number of new alleles to be formed each generation is $2Nu$, and the mean number to be lost each generation through mutation and random drift is $\mathrm{E}(K_{2N})/\mathrm{E}(T)$, where $\mathrm{E}(K_{2N})$ is the mean number of alleles existing in each generation in the entire population. It follows, by balancing the mean number of alleles gained each generation with the mean

number lost, that at stationarity,

$$\mathrm{E}(K_{2N}) = 2Nu\mathrm{E}(T). \tag{124}$$

An approximation to $\mathrm{E}(T)$ is given in (72), and together with (124) this gives

$$\mathrm{E}(K_{2N}) \approx \theta + \int_{(2N)^{-1}}^{1} \theta x^{-1}(1-x)^{\theta-1}\, dx. \tag{125}$$

A more detailed approximation, again using an ergodic argument, is possible. If $\mathrm{E}(K_{2N}(x_1, x_2))$ is the mean number of alleles present in the population with frequency in any interval (x_1, x_2) ($(2N)^{-1} \leq x_1 < x_2 \leq 1$), then from (71)

$$\mathrm{E}(K_{2N}(x_1, x_2)) \approx \int_{x_1}^{x_2} \theta x^{-1}(1-x)^{\theta-1}\, dx. \tag{126}$$

The function

$$\phi(x) = \theta x^{-1}(1-x)^{\theta-1}, \tag{127}$$

which is the integrand in (126), is called the "frequency spectrum" of the process considered, and several important conclusions can be found conveniently from it, as shown below. Ignoring small-order terms, it has the (equivalent) interpretations that, to a close approximation, the mean number of alleles in the population whose frequency is in $(x, x+\delta x)$, and also the probability that there exists an allele in the population whose frequency is in this range, is, for small δx, equal to $\theta x^{-1}(1-x)^{\theta-1}\delta x$.

The form of the frequency spectrum also shows that when θ is small, the most likely situation to arise at any time is that where one allele has a high frequency and the remaining alleles are all at a low frequency. This occurs for two reasons. The first of these is historical: Different alleles enter the population an different times, and an "older" allele has had more time to reach a high frequency than a "younger" allele. Second, imbalances in allelic frequencies arise through stochastic fluctuations, as in the M-allele model as discussed above. This imbalance agrees qualitatively with that found surrounding (118) for the M-allele model.

A final result obtained from the frequency spectrum is the following. Practical population geneticists have long been interested in

the degree of genetic variation present in a population. In practice there will almost always be some variation, so that in practice what is meant is "non-trivial variation", or "non-trivial polymorphism." The classic definition of such a polymorphism, given by Harris (1980, p. 331), is that a locus is polymorphic if the population frequency of the most frequent allele in the population of interest is no more than 0.99. Thus in this sense a population is not polymorphic if the frequency of any allele exceeds 0.99. From the frequency spectrum, the probability of polymorphism is

$$\begin{aligned} & 1 - \theta \int_{0.99}^{1} x^{-1}(1-x)^{\theta-1}\,dx \\ \approx\ & 1 - \theta \int_{0.99}^{1} (1-x)^{\theta-1}\,dx \\ =\ & 1 - (0.01)^{\theta}. \end{aligned} \tag{128}$$

For $\theta = 0.1$, for example, this probability is only about 0.37. However, for larger values of θ, for example for $\theta > 1$, this probability exceeds 0.99.

If the Harris value 0.01 in this definition is replaced by the general value δ, for some small δ, then (128) is replaced by the more flexible value

$$\text{Probability of population polymorphism } = \ 1 - \delta^{\theta}. \tag{129}$$

Several results for the infinitely many alleles model can be obtained directly from two-allele theory. For example, we may wish to find the mean number of generations until all alleles currently existing in the population have been replaced by new alleles, not currently existing in the population. This may be found from two-allele theory by identifying all currently existing alleles with the allele A_1, initially having current frequency $p = 1$ in the population, and seeking the mean number of generations until loss of this allele. This expression is given in (73). A slightly more accurate approximation is

$$4N \sum_{j=1}^{2N} \{j(j+\theta-1)\}^{-1} \text{ generations.} \tag{130}$$

The individual terms in (130) have an important interpretation regarding the past history of the population. This will also be discussed by Dr Joyce.

As discussed previously, the case $\theta = 2$ is of some interest. For this value of θ the expression in (130) reduces to

$$4N - 2 \tag{131}$$

generations. We return to this value later (see the discussion following (295)).

Returning to the case of a single allele A_1 with initial frequency 1, a calculation generalizing that leading to (130) can be made for selective models. It has been claimed that most gene fixation processes in evolution concern very slightly deleterious alleles. Consider then an infinitely many alleles model in which a given allele A_1 has initial frequency 1. We suppose that A_1A_1 individuals have fitness 1, that all A_1A_j heterozygotes have fitness $1 - s$, and that all other genotypes have fitness $1 - 2s$. The mean time until one or other deleterious allele fixes must exceed the mean time until loss of A_1, and the latter mean time may be found immediately from two-allele theory using a generalization of (66). If $\alpha = |2Ns|$ this mean time is, in generations,

$$T(1) = 2N \int_0^1 t(x)\, dx, \tag{132}$$

where

$$t(x) = x^{-1}(1-x)^{\theta-1} \exp(2\alpha x) \int_0^x (1-y)^{-\theta} \exp(-2\alpha y)\, dy. \tag{133}$$

This mean time is extremely large even for moderate values of α, increasing (for $\theta = 1$) from $40N$ generations for $\alpha = 2.5$ to $5 \times 10^6 N$ generations for $\alpha = 10$. We conclude that the evolutionary role of these recurrent deleterious mutants is negligible if α is 5 or more.

Although the theory is by no means clear, it is plausible that to a first approximation, all the results given in this section continue to apply in more complicated Wright–Fisher models, involving perhaps two sexes or geographical structure, if the parameter θ is defined as

$$\theta = 4N_e u, \tag{134}$$

where N_e is one or other version of the effective population size (a concept that is discussed later).

Sample properties

We can think of the result in (121) as a property of a sample of two genes. It is possible to extend the arguments leading to this result to consider samples of size three, four, and so on. Consider then the probability $F_3^{(t+1)}$ that three genes drawn at random in generation $t+1$ are of the same allelic type. These three genes will all be descendants of the same gene in generation t, (probability $(2N)^{-2}$), of two genes (probability $3(2N-1)((2N)^{-2})$) or of three different genes (probability $(2N-1)(2N-2)((2N)^{-2})$). Further, none of the genes can be a mutant, and it follows that

$$F_3^{(t+1)} = (1-u)^3(2N)^{-2}\big(1+3(2N-1)F_2^{(t)}+(2n-1)(2N-2)F_3^{(t)}\big). \tag{135}$$

At equilibrium $F_3^{(t+1)} = F_3^{(t)} = F_3$, and rearrangement in (135) yields

$$F_3 \approx 2(2+\theta)^{-1}F_2 \approx 2!/[(1+\theta)(2+\theta)]. \tag{136}$$

Continuing in this way we find

$$\begin{aligned} F_n^{(t+1)} &= (1-u)^n[(2N-1)(2N-2)\cdots(2N-n+1)(2N)^{1-n}F_n^{(t)} \\ &\quad + \text{terms in } F_{n-1}^{(t)},\ldots,F_2^{(t)}] \end{aligned} \tag{137}$$

and from this, that for small values of n,

$$F_n \approx \frac{(n-1)!}{(\theta+1)(\theta+2)\cdots(\theta+n-1)}. \tag{138}$$

We can also interpret F_n as the probability that a sample of n genes contains only one allelic type, or, in other words, that the sample configuration is $\{n\}$.

This conclusion may be used to find the probability of the sample configuration $\{n-1,1\}$. The probability that in a sample of n genes, the first $n-1$ genes are of one allelic type while the last gene is of a new allele type is $F_{n-1}-F_n$. The probability we require is, for $n \geq 3$, just n times this, or

$$\text{Prob}\{n-1,1\} = n\{F_{n-1}-F_n\} \approx n(n-2)!\theta/[(1+\theta)(2+\theta)\cdots(n-1+\theta)]. \tag{139}$$

For $n = 2$ the required probability is

$$\text{Prob}\{1,1\} \approx \theta/(1+\theta). \tag{140}$$

The probabilities of other configurations can built up in a similar way. We illustrate this by considering the probability $F_{2,2}^{(t+1)}$ that, of four genes drawn at random in generation $t+1$, two are of one allelic type and two of another. Clearly none of the genes can be a mutant, and furthermore they will be descended from four different parent genes of configuration $\{2,2\}$, from three different parent genes of configuration $\{2,1\}$, the singleton being transmitted twice, or from two different parent genes, both transmitted twice. Considering the probabilities of the various events, we find

$$\begin{aligned} F_{2,2}^{(t+1)} &= (1-u)^4(2N)^{-3}\big((2N-1)(2N-2)(2N-3)F_{2,2}^{(t)} \\ &\quad + 2(2N-1)(2N-2)F_{2,1}^{(t)} + 3(2N-1)F_{1,1}^{(t)}\big). \end{aligned} \tag{141}$$

Retaining only higher-order terms and letting $t \to \infty$, we obtain

$$F_{2,2} \approx (3+\theta)^{-1}F_{2,1} = 3\theta/\big((1+\theta)(2+\theta)(3+\theta)\big). \tag{142}$$

Continuing in this way we find an approximating partition probability formula for a sample of n of genes, where is is assumed that $n << N$. This probability can be presented in various ways. Perhaps the most useful formula arises if we introduce the random variables $A_1, A_2, \ldots, A_n$, where A_j is the number of alleles that arise exactly j times in the sample. It is necessary that $\sum_j jA_j = n$, and we make this restriction from now on. If we introduce the vector $\mathbf{A} = (A_1, A_2, \ldots, A_n)$, we find that

$$\text{Prob}(\mathbf{A} = \mathbf{a}) = \frac{n!\,\theta^{\sum a_j}}{1^{a_1}2^{a_2}\cdots n^{a_n}\,a_1!\,a_2!\cdots a_n!\,S_n(\theta)}\,. \tag{143}$$

In this expression, $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, $\sum_j ja_j = n$ and $S_n(\theta)$ is defined by

$$S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1). \tag{144}$$

The expression (143) was derived by Ewens (1972) and Karlin and McGregor (1972).

We denote $\sum A_j$, the (random) number of different allelic types seen in the sample, by K_n, and $\sum_j a_j$, the corresponding number of

alleles observed in a given sample, by k_n. By suitable summation in (143) the probability distribution of the random variable K_n may be found as

$$\text{Prob}\,(K_n = k_n) = |S_n^k|\theta^k/S_n(\theta), \tag{145}$$

where $|S_n^k|$ is the coefficient of θ^k in $S_n(\theta)$. Thus $|S_n^k|$ is the absolute value of a Stirling number of the first kind. From (145), the mean of K_n is

$$\text{E}(K_n) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1}\,, \tag{146}$$

the variance of K_n is

$$\text{var}(K_n) = \theta \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}, \tag{147}$$

and the probability that $K_n = 1$ is

$$\frac{(n-1)!}{(\theta+1)(\theta+2)\cdots(\theta+n-1)}. \tag{148}$$

This agrees, as it must, with the expression in (138).

A formula equivalent to (143) is the following. Suppose that in the sample we observe k_n different allelic types. We label these in some arbitrary order as types $1, 2, \ldots, k_n$. Then the probability that $K_n = k_n$ and also that with the types labeling in the manner chosen, there are n_1, n_2, …, n_{k_n} genes respectively observed in the sample of these various types, is

$$\frac{n!\theta^{k_n}}{k_n! n_1 n_2 \cdots n_{k_n} S_n(\theta)}\,. \tag{149}$$

The frequency spectrum (127) can be used to confirm various sampling theory results. For example, an allele whose population frequency is x is observed in a sample of size n with probability $1-(1-x)^n$. From this and (126) it follows that the mean number of different alleles observed in a sample of size n is approximately

$$\int_0^1 \{1-(1-x)^n\}\theta x^{-1}(1-x)^{\theta-1}\,dx, \tag{150}$$

and the value of this expression is equal to that given in (146).

As a second example, the probability that only one allele is observed in a sample of n genes can be found from the frequency spectrum as

$$\begin{aligned}&\theta \int_0^1 x^n \{x^{-1}(1-x)^{\theta-1}\}\, dx \\ &= (n-1)!/\big((1+\theta)(2+\theta)\cdots(n-1+\theta)\big)\end{aligned}$$

and this agrees with the expression in (138). More complex formulas such as (143) can be re-derived using multivariate frequency spectra, but we do not pursue the details.

The Cannings infinitely many alleles model

The reproductive mechanism in the non-overlapping generations Cannings infinitely many alleles model follows that of the general principles of the Cannings two-allele model. That is, the model allows any reproductive scheme consistent with the exchangeability and symmetry properties of the two-allele model. The mean number of offspring genes from any "parental" gene is 1, and the variance of the number of offspring genes is σ^2, necessarily the same for each parental gene. The mutation assumptions are as described above, in particular that all mutant offspring genes are assumed to be of novel allelic types.

Many of the results of the Wright–Fisher infinitely many alleles model apply for the Cannings model, at least to a close approximation, **provided that the parameter θ, arising in many formulas associated with the Wright-Fisher model, is replaced throughout by θ/σ^2.** Therefore we do not explore the Cannings model further, and instead use Wright–Fisher formulae, with this change of definition of θ, when considering the Cannings model.

The Moran infinitely many alleles model

The Moran infinitely many alleles model is the natural extension to the infinitely many alleles case of the Moran two alleles model. Haploid individuals, which we may identify with genes, are created and lost through a birth and death process, as in the two-alleles

case, with the standard infinitely many alleles model assumptions that an offspring gene is a mutant with probability u and that any new mutant is of an entirely novel allelic type.

The stochastic behavior of the frequency of any allelic type in the population is then governed by (105), implying (as for the Wright-Fisher and Cannings models) that there can be no concept of stationarity of the frequency of any nominated allelic type. On the other hand, as for those models, there will exist a concept of the stationary distribution of allelic configurations. The possible configurations of the process are the same as those for the Wright-Fisher and Cannings models, but for the Moran model an exact population probability can be given for each configuration, both in the population as a whole and also in a sample of any size taken from the population.

Population properties

We first consider the population as a whole. Suppose that B_j ($j = 1, 2, \ldots, 2N$) is the number of allelic types with exactly j representative genes in the population, so that $\sum_{j=1}^{2N} jB_j = 2N$. The quantity B_j is the population analogue of the sample number A_j in (143). We introduce the vectors $\mathbf{B} = (B_1, B_2, \ldots, B_{2N})$ and $\mathbf{b} = (b_1, b_2, \ldots, b_{2N})$, where $\sum_{j=1}^{2N} jb_j = 2N$. The exact stationary distribution of the population configuration process is

$$\Pr(\mathbf{B} = \mathbf{b}) = \frac{(2N)!\ \theta^{\sum b_j}}{1^{b_1} 2^{b_2}\ \cdots\ (2N)^{b_{2N}}\, b_1! b_2! \cdots b_{2N}! S_{2N}(\theta)}\,. \qquad (151)$$

Here $S_{2N}(\theta)$ is defined by replacing n by $2N$ in (144), and θ is defined for this model by

$$\theta = 2Nu/(1-u). \qquad (152)$$

This is a different definition of θ than that applying for the Wright–Fisher model, and is always to be used as the definition of θ when referring to the Moran model.

The expression on the right-hand side of (151) is of exactly the same form as (143), with n replaced by $2N$ and A_j by B_j. Since equation (151) is an exact one for the Moran model, several of the calculations arising from (143) are exact for the Moran model population. For example, the distribution of the number K_{2N} of allelic

types in the population is given exactly by (145), with n replaced by $2N$. Also, from (148), the probability that $K_{2N} = 1$ is, exactly,

$$\frac{(2N-1)!}{(1+\theta)(2+\theta)\cdots(2N-1+\theta)}\,. \tag{153}$$

Further, the mean of K_{2N} is given exactly by (146), with n replaced by $2N$ and the variance K_{2N} is

$$\mathrm{var}(K_{2N}) \;=\; \theta \sum_{j=1}^{2N-1} \frac{j}{(\theta+j)^2}. \tag{154}$$

In all of these expressions, it must be kept in mind that the definition of θ for the Moran model (namely $2Nu/(1-u)$) differs from that of the Wright-Fisher model (namely $4Nu$), so that the formal identity of expressions in the two models is perhaps misleading.

An exact expression is available for the Moran model (discrete) frequency spectrum, for which (127) gives the (continuous) approximate Wright–Fisher model formula. To find this we consider first the "two-allele" model (105). In the infinitely many alleles case we think of A_1 as a new arisen allele formed by mutation and A_2 as all other alleles. Standard theory can be used to find the mean number $\mu(T)$ of birth and death events before the certain loss of A_1 from the population. This is

$$\mu(T) = (2N+\theta) \sum_{j=1}^{2N} j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \tag{155}$$

The form of ergodic argument that led to (125) shows that at stationarity, the mean of the number K_{2N} of different allelic types represented in the population is $u\mu(T)$, which is

$$\theta \sum_{j=1}^{2N} j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \tag{156}$$

where here and throughout we use the standard gamma function definition

$$\binom{M}{m} = \frac{\Gamma(M+1)}{m!\Gamma(M-m+1)} \;=\; \frac{M(M-1)\cdots(M-m+1)}{m!}$$

for non-integer M. The expression (156) provides the further information that the typical jth term gives the stationary mean number of alleles arising with j representing genes in the population at any time. In other words, the exact frequency spectrum for the Moran model is

$$\theta\, j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \quad j = 1, 2, \ldots, 2N. \qquad (157)$$

A standard asymptotic formula for the gamma function for large N shows the parallel between this exact expression with the continuous Wright-Fisher frequency spectrum (127).

Various special cases of (157) are of interest. For example, when $\theta = 1$, (157) simplifies to j^{-1}, in which form the parallel with the Wright-Fisher approximation (127) is obvious. However, the different formulae for θ for the two models should be kept in mind when this comparison is made.

Two of the above expressions are of independent interest. First, the expression (155) has the further interpretation that its typical term is the mean number of birth-and-death events for which there are exactly j copies of the allele in question before its loss from the population. It is interesting to evaluate the expression in (155) for specific values of θ. When $\theta = 2$, it is about $2N \log(2N)$ birth and death events, or about $\log(2N)$ "generations". The corresponding approximation for the Wright–Fisher model, found from (66), is also $\log(2N)$ generations, but again this formal equality is misleading because of the different definitions of θ in the two cases.

Second, the expression (156) simplifies to

$$\frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+2N-1}.$$

This is identical to the expression given in (146), with n replaced by $2N$, as we know it must be.

Many further exact results for the Moran model are available. Here are several.

First, if at any time there is only one allele in the population, we say that that allele is "quasi-fixed" in the population. (We do not use the expression "fixed", since in an infinitely many alleles model this allele will eventually be lost from the population.) The probability that a new mutant eventually becomes quasi-fixed can

be found as follows. Call the allelic type of the new mutant A_1 and group together all other genes as "A_2" genes. Then standard continuant Markov chain theory shows that the probability that a new mutant allele eventually becomes quasi-fixed in the population is C^{-1}, where

$$C = \sum_{j=0}^{2N-1} \binom{2N+\theta-1}{j} \Big(\binom{2N-1}{j} \Big)^{-1}. \tag{158}$$

This is a different probability than the probability that, at any specified time, the population is "quasi-fixed" for some unspecified allele. This latter probability is given by the $j = 2N$ term in the exact Moran frequency spectrum (157), namely

$$\frac{\theta}{2N} \Big(\binom{2N+\theta-1}{2N} \Big)^{-1}, \tag{159}$$

or, more simply,

$$\frac{(2N-1)!}{(1+\theta)(2+\theta)\cdots(2N-1+\theta)}. \tag{160}$$

To illustrate the difference between the probability defined by (158) and the probability defined by (160), when $\theta = 1$ the probability of quasi-fixation of a specified new mutant allele is, from the expression (158),

$$\frac{1}{2N} \big(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2N}\big)^{-1} \tag{161}$$

while the quasi-fixation probability of an unspecified allele, given by the expression (160), is $1/(2N)$. These two expressions differ by a multiplicative factor of

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2N}, \tag{162}$$

and it is not a coincidence that this is the mean number of alleles in the population at any time.

Second, it is immediate that the probability that a gene drawn at random from the population is of an allelic type represented j times in the population is found by multiplying the expression in (157) by $j/(2N)$. This gives

$$\frac{\theta}{2N} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right) \tag{163}$$

for this probability. We check that the sum of this expression over $j = 1, 2, \ldots, 2N$ is 1.

Third, (156) allows an exact calculation of the probability of population polymorphism, as defined by Harris. Any allele having a frequency exceeding 0.99 must be the most frequent allele in the population, and at most one allele can have such a frequency. Thus the probability that the most frequent allele in the population has frequency exceeding 0.99 is the mean number of alleles with frequency exceeding 0.99. Taking $0.99(2N)$ as an integer M, (156) shows that the Harris probability of polymorphism is

$$1 - \theta \sum_{j=M+1}^{2N} j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right). \tag{164}$$

This is close to $1 - (0.01)^\theta$, the approximate value found above for the Wright–Fisher model using the frequency spectrum (see the expression in (128)). As with other such comparisons, this apparent similarity is misleading because of the different definitions of θ in the two models.

The final result concerns the mean number of birth and death events until all alleles present in the population at any time are lost. This is the Moran model analogue (once an adjustment is made between generations and birth-death events) for the mean number of generations until allele current alleles in a population are lost in the Wright-Fisher model, an approximation for which is given in (73). In the case of the Moran model an exact calculation is available, namely that the required mean number of birth and death events is

$$2N(2N+\theta)(\theta-1)^{-1} \sum_{j=1}^{2N} j^{-1} \left(1 - \binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right). \tag{165}$$

A formula different from (165), found by applying l'Hôpital's rule, applies for the case $\theta = 1$. An excellent approximation to this value is given in (111).

In the case $\theta = 2$, the expression (165) gives, exactly, $8N^2(N + 1)/(2N + 1)$. This is very close to the approximation $4N^2$ birth and death events given in (113). This can be thought of as corresponding to $2N$ "generations", which can be compared to the approximate value $4N$ for the Wright–Fisher process given in (75). Note the

"factor of two" difference between these values: we return to this factor later.

There are several further comments to make about (165). First, The typical (jth) term in (165) is the mean number of birth and death events for which there are exactly j genes present of the various original alleles in the population before the eventual loss of all these alleles. Thus the expression (165) gives more information than might otherwise be thought.

Second, although the identity is not immediately obvious, the expression in (165) is identical to the expression

$$2N(2N+\theta)\sum_{j=1}^{2N}\frac{1}{j(j+\theta-1)}. \tag{166}$$

The expression in (166) may be written equivalently as

$$\sum_{j=1}^{2N}\frac{1}{v_j+w_j}, \tag{167}$$

where

$$v_j=\frac{ju}{2N},\quad w_j=\frac{j(j-1)(1-u)}{(2N)^2}. \tag{168}$$

Coalescent theory explains why the mean age of the oldest allele can be expressed in the form defined by (167) and (168). Again, this will be discussed by Dr Joyce.

We next consider the mean number of birth-death events until all alleles present in the population at any time are lost. The value given in (130) for the Wright–Fisher model for this mean is a diffusion approximation. In the case of the Moran model an exact calculation can be made by using two-allele theory, and regarding all the alleles existing in the population at a given time as A_1, and that initially there are $2NA_1$ genes in the population. Watterson (1976) found that the required mean number of birth-death events is

$$2N(2N+\theta)(\theta-1)^{-1}\sum_{j=1}^{2N}j^{-1}\Bigl(1-\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\Bigr), \tag{169}$$

(A formula different from (169), found by applying l'Hôpital's rule, applies for the case $\theta=1$.) In the case $\theta=2$ (169) gives, exactly,

$8N^2(N+1)/(2N+1)$, or about $4N^2$, birth-death events. This can be thought of as corresponding to $2N$ "generations". This differs from the value in (131) for the Wright–Fisher model, and we note again the "factor of two" difference between the two models.

We make several further comments about (169). First, as with the corresponding result for the Wright–Fisher model, we may think of (169) as providing, in this case exactly, the mean age of the oldest allele in the population. Second, the typical (jth) term in (169) is the mean number of birth-death events for which there are exactly j genes present of the various original alleles alleles in the population before the eventual loss of all these alleles. Thus the expression (169) gives more information than might otherwise be thought.

The exact frequency spectrum (156) provides two results almost immediately of interest in this section. The first uses the concept of size-biased sampling, discussed in more detail later. In the Moran model the probability that an individual drawn at random is of an allelic type having exactly j copies in the population is found by multiplying the jth term in (156) by $j/(2N)$. This gives a value of

$$\theta(2N)^{-1}\left(\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\right). \tag{170}$$

for this probability. This calculation is of use when "age" properties of the alleles in the population are considered.

Second, (156) allows an exact calculation of the "Harris" probability of population polymorphism, as defined in (128). Any allele having a frequency exceeding 0.99 must be the most frequent allele in the population, and at most one allele can have such a frequency. Thus the probability that the most frequent allele in the population has frequency exceeding 0.99 is the mean number of alleles with frequency exceeding 0.99. Taking $0.99(2N)$ as an integer J, (156) shows that the probability of polymorphism is

$$1-\theta\sum_{j=J+1}^{2N} j^{-1}\left(\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\right). \tag{171}$$

This is close to $1-(0.01)^\theta$, the approximate value found above for the Wright–Fisher model using a diffusion approximation. As with other such calculations, this apparent similarity is misleading because of the different definition of θ in the two models.

Many further exact and elegant results can be found for the Moran model, but since our main interest is in samples of genes from a population rather than the entire population itself, we do not consider these further.

Sample properties

We turn now to sample properties in the Moran model. Suppose now a sample of n genes is taken, without replacement, from a population obeying the infinitely many alleles Moran model. Perhaps surprisingly, all the sampling results given above as approximations for the Wright-Fisher model are exactly correct for this sample, provided of course that θ is interpreted as $2Nu/(1-u)$. In particular, the basic formula (143) applies exactly for the infinitely many alleles Moran model. This implies that the formulas (145), (146), (147) and (148), all of which derive from (143), are exactly correct for that model.

A most important concept concerning a sample of genes is that of a partition structure (Kingman, (1978)). There should be no particular significance attached to the sample size n, and we can regard a sample of size n genes as one arising from a sample of size $n+1$, one of which was accidently lost. We reasonably requires a consistency of formulae for the two sample sizes. To formalize this we denote the left-hand side in the partition formula (143) as as $\mathrm{P}_n(a_1, a_2, \ldots)$ The method of arriving at a sample of n genes as just described then implies that this must be equal to

$$\frac{a_1+1}{n+1}\mathrm{P}_{n+1}(a_1{+}1, a_2, \ldots){+}\sum_{j=2}^{n+1}\frac{j(a_j+1)}{n+1}\mathrm{P}_{n+1}(a_1, .., a_{j-1}{-}1, a_j{+}1, \ldots). \tag{172}$$

The right-hand side in (143) does satisfy this requirement, but Kingman raised the much more general question: how may one characterize probability structures satisfying (172)? He called structures having this property "partition structures", and showed that for all such structures of interest in genetics, $\mathrm{P}_n(a_1, a_2, \ldots)$ could be represented in the form

$$\mathrm{P}_n(a_1, a_2, \ldots) = \int \mathrm{P}_n(a_1, a_2, \ldots |\mathbf{x})\mu(d\mathbf{x}), \tag{173}$$

where $P_n(a_1, a_2, \ldots | \mathbf{x})$ is a complicated sum of multinomial probabilities whose exact form we do not write down. Kingman called μ the "representing measure" of $P_n(a_1, a_2, \ldots)$ and found that for the partition formula (143) this representing measure is the so called "Poisson–Dirichlet" distribution - see Kingman (1975) for the properties of this distribution.

The consistency requirement (172) is a natural one for a sample of genes. We shall, however, find a perhaps more important interpretation for this requirement when considering the past history of the population from which the sample was taken.

Kingman also took up the question of "non-interference", defined by the requirement that if a gene is taken at random from the sample, and all r genes of its allelic type then removed from the sample, the partition probability structure of the remaining $n - r$ genes should be the same as that of an original sample of $n - r$ genes. Non-interference implies that $P_n(a_1, \ldots, a_r, \ldots)$ must satisfy the requirement

$$\frac{ra_r}{n} P_n(a_1, \ldots, a_r, \ldots) = c(n, r) P_{n-r}(a_1, \ldots, a_{r-1}, \ldots), \qquad (174)$$

where $c(n, r)$ does not depend on $a_1, a_2, \ldots$ Kingman then showed that of all partition structures of interest in genetics, the only one also satisfying the requirement (174) is (143).

Further exact Moran model results relating to "time" and "age" properties, both for the entire population and for a sample of genes from the population, will be discussed by Dr. Joyce.

Complications and the effective population size

Introduction

All the theory described above (and also that described later) makes a large number of assumptions, genetical, modeling and demographic. The main genetical assumption is that there is no selection involved between the alleles that we consider. Clearly, and especially in light of the Darwinian paradigm, this means that a very large proportion of population genetics theory, that relating to selection, is not considered. Another important aspect of reality that is ignored is the existence, for the great majority of species of interest to us, of

two sexes and the diploid nature of the individuals in those populations. From the modeling point of view, the three models considered (Wright-Fisher, Cannings and Moran) cannot be expected to describe accurately any real-life population, even though they do provide some insights into the evolutionary genetic behavior of real populations. Finally, many demographic features, such as the geographical dispersion of a population and changes over time in the size of the population of interest, have been ignored.

The concept of the effective population size is meant to address some of these deficiencies, and in this section we define this concept and examine some of its properties.

Three concepts of the effective population size

Even though the "simple" Wright-Fisher model is far less plausible than several other available models as a description of biological reality, it has, perhaps for historical reasons, assumed a central place in population genetics theory. This model has three properties that relate to the population size:

(i) its maximum non-unit eigenvalue $\lambda_{\max}$ is $1 - (2N)^{-1}$,

(ii) the probability π_2 that two genes taken at random are descendants of the same parent gene is $(2N)^{-1}$,

(iii) $\mathrm{var}\{x(t+1) \mid x(t)\} = x(t)\{1 - x(t)\}/(2N)$, where $x(t)$ is the fraction of A_1 genes in generation t.

Thus the population size N in the simple Wright-Fisher model obeys the equations $N = \frac{1}{2}(1 - \lambda_{\max})^{-1}$, $N = (2\pi_2)^{-1}$ and

$$N = \frac{x(t)\{1 - x(t)\}}{2\,\mathrm{var}\{x(t+1) \mid x(t)\}}.$$

In view of these properties it is perhaps natural, if the Wright–Fisher model (35) is to be used as a standard, to define the various effective population sizes in diploid models that are more complicated and realistic then (35) in the following way:

$$\text{eigenvalue effective population size} = \tfrac{1}{2}(1 - \lambda^*_{\max})^{-1}, \quad (175)$$

$$\text{inbreeding effective population size} = (2\pi^*_2)^{-1}, \quad (176)$$

$$\text{variance effective population size} = \frac{x(t)\{1 - x(t)\}}{2\,\mathrm{var}^*\{x(t+1) \mid x(t)\}}\,. \quad (177)$$

We write these for convenience as $N_e^{(e)}, N_e^{(i)}$ and $N_e^{(v)}$ respectively. Here $\lambda^*_{\max}$ is now defined as the largest nonunit eigenvalue of the transition matrix of the complicated model considered and π_2^* is now defined as the probability, in this model, that two genes taken at random in any generation are descendants of the same parent gene. Similarly, $\mathrm{var}^*\{x(t+1)\}$ is the conditional variance of the frequency of A_1 in generation $t+1$ in this model, given the value of this frequency in generation t.

A fourth concept of effective population size, namely the mutation effective size, is also possible, but we do not consider this concept here. A more general concept of effective population size is the "coalescent" effective population size. This will be discussed by Dr. Krone.

Application to the Cannings model

In this section we consider the application of the effective population size concept for the Cannings model, and limit attention for the moment to those versions of the model where generations do not overlap. Equations (84) and (175) show immediately that for these models, the eigenvalue effective population size $N_e^{(e)}$ is given by

$$N_e^{(e)} = \left(N - \tfrac{1}{2}\right)/\sigma^2, \tag{178}$$

where σ^2 is the variance in the number of offspring genes from any given gene. Equations (86) and (177) show that the variance effective population size $N_e^{(v)}$ is given by

$$N_e^{(v)} = \left(N - \tfrac{1}{2}\right)/\sigma^2. \tag{179}$$

A value for $N_e^{(i)}$ can be found in the following way. Suppose that the ith gene in generation t leaves M_i offspring genes in generation $t+1$, ($\sum M_i = 2N$). Then the probability, given $M_1, \ldots, M_{2N}$, that two genes drawn at random in generation $t+1$ are descendants of the same gene is

$$\sum_{i=1}^{2N} M_i(M_i - 1)/\{2N(2N-1)\}. \tag{180}$$

The probability π_2^* is the expected value of this random variable. Now M_i has mean unity and variance σ^2, so that, on taking expec-

tations, $\pi_2^* = \sigma^2/(2N-1)$. From this,

$$N_e^{(i)} = \big(N - \tfrac{1}{2}\big)/\sigma^2. \tag{181}$$

It follows from these various equations that for the Cannings model, all three effective population sizes are equal.

One application of this conclusion is the following. If leading terms only are retained, all three definitions of the effective population size in the Cannings model are N/σ^2. From the remarks surrounding (134), it is plausible that the various Wright–Fisher infinitely many alleles model results apply for the non-overlapping generation Cannings model if θ is defined wherever it occurs by $4N_e u$. That is, to a close approximation, we define θ for the Cannings model by

$$\theta = 4Nu/\sigma^2. \tag{182}$$

As stated earlier, the definition of θ given in (182) is to be used whenever the Cannings model is discussed.

Application to the Moran model

The three definitions of the effective population size given above are not appropriate for models where generations overlap. If we write N_e for any one of the effective population sizes defined in above, it seems reasonable for such models to define the effective population size as $N_e k/(2N)$, where k is the number of individuals to die at each time unit. Since $k = 2N$ for models where generations do not overlap, this leaves the definitions of the effective population size unchanged for such models. For the Moran model, where $k = 1$, this convention yields

$$N_e^{(e)} = N_e^{(i)} = N_e^{(v)} = \tfrac{1}{2}N. \tag{183}$$

The equations show that the effective population size in the Moran model is half that in the Wright–Fisher model. We now discuss the reason for this.

Arguments parallel to those leading to (47) show that if two alleles A_1 and A_2 are allowed in the population, the mean time until fixation of one or other allele in the Cannings model is

$$\bar{t}(p) \approx -(4N-2)\{p\log p + (1-p)\log(1-p)\}/\sigma^2, \tag{184}$$

where p is the initial frequency of A_1 and σ^2 is defined above. This formula explains the factor of 2 discussed after equation (102) and in several other places. In the Wright–Fisher model $\sigma^2 \approx 1$ while in the Moran model $\sigma^2 \approx 2/(2N)$. Setting aside the factor $2N$ as explained by the conversion from generations to birth and death events, it is clear that the crucial factor is the difference between the two models in the variance in offspring distribution. This explains the "factor of 2" difference between the expressions in (47) and (102), and other similar pairs of equations, (see for example the calculations two paragraphs after (169)), once the factor of $2N$ has been set aside to allow for the conversion between birth-death events and generations.

Diploid organisms

So far we have largely ignored the diploid nature of most organisms of interest, and have in effect considered a "population" of $2N$ genes, rather than a population of N individuals, each carrying 2 genes. We now consider a definition of effective population size where we focus on the N individuals rather than the $2N$ genes. We shall do this for a Cannings model, so that the results found also apply to the Wright-Fisher model.

Our aim is to devise an inbreeding effective population number that is focused around the N diploid individuals in the population. This number will be denoted $N_e^{(id)}$, ('d" for diploid, "i" for inbreeding), and is defined as the reciprocal of the probability that two genes (corresponding to one individual) taken at random in generation $t+1$ are descended from the same diploid individual in generation t. This is tantamount, in the Cannings model, to selecting two genes at random in generation t and asking whether the two genes drawn at random in generation $t+1$ are both descended from one or other or both of these. In the notation of (180), the probability of this event can be written as the expected value of

$$\sum_{i=1}^{N}(m_i + m_{N+i})(m_i + m_{N+i} - 1)/\{2N(2N-1)\}. \tag{185}$$

It is not hard to see this leads to

$$N_e^{(id)} = \frac{4N-2}{\sigma_d^2 + 2}\,, \tag{186}$$

where σ_d^2 is the variance of the number of offspring genes from each (diploid) individual. It is therefore necessary to extend the Cannings model to the diploid case. We define a diploid Cannings model as one for which the concept of exchangeability relates to diploid individuals. We also assume that the gene transmitted by any individual to any offspring is equally likely to be each of the two genes in that individual, is independent of the gene(s) transmitted by this individual to any other offspring, and is also independent of the genes transmitted by any other individual. With these conventions it can be shown that

$$\sigma^2 = \frac{\sigma_d^2 + 2}{4}, \tag{187}$$

where σ^2 is the Cannings model gene "offspring number" variance, and from this it follows that the expressions in (181) and (186) are identical.

More realistic Wright-Fisher models

We turn next to the second class of models where a definition of effective population size is useful, namely those Wright–Fisher models which attempt to incorporate biological complexity more than does the simple Wright–Fisher model (35).

The first model considered allows for the existence of two sexes. Suppose in any generation there are N_1 diploid males and N_2 diploid females, with $N_1 + N_2 = N$. The model assumes that the genetic make-up of each individual in the daughter generation is found by drawing one gene at random, with replacement, from the male pool of genes, and similarly one gene with replacement from the female pool. If $X_1(t)$ represents the number of A_1 genes among males in generation t and $X_2(t)$ the corresponding number among females, then $X_1(t+1)$ can be represented in the form

$$X_1(t+1) = i(t+1) + j(t+1), \tag{188}$$

where $i(t+1)$ has a binomial distribution with parameter $X_1(t)/(2N_1)$ and index N_1, and $j(t+1)$ has a binomial distribution with parameter $X_2(t)/(2N_2)$ and index N_1. A similar remark applies to $X_2(t+1)$, where now the index is N_2 rather than N_1. Evidently the pair $\{X_1(t),\, X_2(t)\}$ is Markovian, and there will exist a transition matrix whose leading nonunit eigenvalue we require to find so that we can calculate $N_e^{(e)}$.

To do this it is necessary to find some function $Y(X_1, X_2)$ which is zero in the absorbing states of the system, positive otherwise, and for which

$$E[Y\{X_1(t+1), X_2(t+1)\} \mid X_1(t), X_2(t)] = \lambda^* Y(t) \tag{189}$$

for some constant λ^*. The value λ^* is then the largest non-unit eigenvalue of the Markov chain transition matrix defined by the pair $\{X_1(t),\ X_2(t)\}$. Such a function always exists, but some trial and error is usually necessary to find it. In the present case it is found, after much labor, that a suitable function is

$$\begin{aligned} Y(X_1, X_2) &= \tfrac{1}{2}C\{X_1(2N_1 - X_1)(2N_1)^{-2} + X_2(2N_2 - X_2)(2N_2)^{-2}\} \\ &\quad + \{1 - (X_1 - N_1)(X_2 - N_2)N_1^{-1}N_2^{-1}\}, \end{aligned} \tag{190}$$

where

$$C = \tfrac{1}{2}\{1 + (1 - 2N_1^{-1} - 2N_2^{-1})^{1/2}\}.$$

With this definition the eigenvalue λ^* is given by

$$\lambda^* = \tfrac{1}{2}[1 - (4N_1)^{-1} - (4N_2)^{-1} + \{1 + N^2(4N_1N_2)^{-2}\}^{1/2}], \tag{191}$$

or approximately

$$\lambda^* \approx 1 - (N_1 + N_2)(8N_1N_2)^{-1}. \tag{192}$$

From this result and (175) it follows that to a close approximation,

$$N_e^{(e)} = 4N_1N_2N^{-1}. \tag{193}$$

If $N_1 = N_2$ $(= \frac{1}{2}N)$, then $N_e^{(e)} \approx N$, as we might expect, while if N_1 is very small and N_2 is large, $N_e^{(e)} \approx 4N_1$. This latter value is sometimes of use in certain animal-breeding programs.

The inbreeding population size is found much more readily. Two genes taken at random in any generation will have identical parent genes if both are descended from the same "male" gene or both from the same "female" gene. The probability of identical parentage is thus

$$\pi_2^* = \tfrac{1}{2}\frac{N-1}{2N-1}\{(2N_1)^{-1} + (2N_2)^{-1}\},$$

and from this it follows that

$$N_e^{(i)} = (2\pi_2^*)^{-1} \approx 4N_1N_2N^{-1}. \tag{194}$$

The variance effective population size cannot be found so readily, and indeed strictly it is impossible to use (177) to find such a quantity, since an equation of this form does not exist in the two-sex case we consider. The fraction of A_1 genes is not a Markovian variable and in particular, using the notation of (177), the variance of $x(t+1)$ cannot be given in terms of $x(t)$ alone. This indicates a real deficiency in this mode of definition of effective population size. On the other hand, sometimes there exists a "quasi-Markovian" variable exists in terms of which a generalized expression for the variance effective population size may be defined. In the present case the weighted fraction of A_1 genes, defined as

$$x(t) = X_1(t)/(4N_1) + X_2(t)/(4N_2)$$

has the required quasi-Markovian properties, and

$$\text{var}\{x(t+1) \mid x(t)\} = x(t)\{1 - x(t)\}N(8N_1N_2)^{-1} + O(N_1^{-2}, N_2^{-2}).$$

From this a generalized variance effective population size may be defined, in conjunction with (177), as

$$N_e^{(v)} = 4N_1N_2N^{-1}. \tag{195}$$

Thus for this model, $N_e^{(e)} \approx N_e^{(i)} \approx N_e^{(v)}$, although strict equality does not hold for any of these relations.

We return now to the case of a monoecious population and consider complications due to geographical structure. A simplified model for this situation which, despite its obvious biological unreality, is useful in revealing the effect of population subdivision, has been given by Moran (1962).

It is supposed that the total population, of size $N(H+1)$, is subdivided into $H+1$ sub-populations each of size N, and that in each generation G genes chosen at random migrate from subpopulation i to subpopulation j for all i, j $(i \neq j)$. Suppose that in subpopulation i there are $X_i(t)$ A_1 genes in generation t. There is no single Markovian variable describing the behavior of the total population, but the quantities $X_i(t)$ are jointly Markovian, and to find $N_e^{(e)}$ it is necessary to find some function $Y(t) = Y\{X_1(t), \ldots, X_{H+1}(t)\}$ obeying an equation parallel to (189). It is found, after some trial

and error, that a suitable function $Y(t)$ is

$$\begin{aligned} Y(t) &= [A - D + \{(A-D)^2 + 4BC\}^{1/2}] \sum_i X_i(t)\{2N - X_i(t)\} \\ &\quad + 2B \sum_{i \neq j} \sum X_i(t)\{2N - X_j(t)\}, \end{aligned} \tag{196}$$

where

$$\begin{aligned} A &= (4N^2 + H^2G^2 + G^2H - 2N - 4NGH)/4N^2, \\ B &= (4GN - G^2H - G^2)/(4N^2), \\ C &= (4HGN - G^2H^2 - G^2H)/(4N^2), \\ D &= (4N^2 + HG^2 + G^2 - 4HG)/(4N^2). \end{aligned}$$

With this definition of $Y(t)$, the eigenvalue λ^* satisfying

$$\mathrm{E}\{Y(t+1) \mid X_1(t), \ldots, X_{H+1}(t)\} = \lambda^* Y(t)$$

is

$$\lambda^* = \tfrac{1}{2}\big(A + D + \{(A-D)^2 + 4BC\}^{1/2}\big). \tag{197}$$

If small-order terms are ignored, this yields eventually

$$N_e^{(e)} \approx N(H+1)\{1 + (2G(H+1))\}^{-1}\} \tag{198}$$

for large H and G. This equation is in fact accurate to within 10% even for $H = G = 1$, and it thus reveals that population subdivision leads to only a slight increase in the eigenvalue effective population size compared to the value $N(H+1)$ obtaining with no subdivision.

The inbreeding effective population size $N_e^{(i)}$ can be found most efficiently by noting that it is independent of G, since the act of migration is irrelevant to the computation of its numerical value. Thus immediately from (181)

$$N_e^{(i)} = \big\{N(H+1) - \tfrac{1}{2}\big\}/\{1 - (2N)^{-1}\}, \tag{199}$$

since each gene produces a number of offspring according to a binomial distribution with index $2N$ and parameter $(2N)^{-1}$. This value clearly differs only trivially from the true population size $N(H+1)$ and, for small H and G, it differs slightly from $N_e^{(e)}$.

Because of these two results, one may be tempted to ignore geographical sub-division in modeling evolutionary population genetic processes.

The computation of $N_e^{(v)}$ is beset with substantial difficulties since there exists no scalar Markovian variable for the model. Indeed, unless migration rates are of a large order of magnitude, there is not even a "quasi-Markovian" variable. Because of this no satisfactory value for $N_e^{(v)}$ has yet been put forward for the geographical structure case.

We consider finally a population whose size assumes cyclically the sequence of values $N_1, N_2, N_3, \ldots, N_k, N_1, N_2, \ldots$. There is no unique value of $N_e^{(e)}$, $N_e^{(i)}$ or $N_e^{(v)}$ in this case, and it is convenient to extend our previous definition to cover k consecutive generations of the process. If the population size in generation $t+k$ is N_i, it is easy to see that if $X(t)$ is the number of A_1 genes in generation t, and in each generation reproduction occurs according to the model (35),

$$E[X(t+k)\{2N_i - X(t+k)\} \mid X(t)] = X(t)\{2N_i - X(t)\} \prod_{i=1}^{k} \{1-(2N_i)^{-1}\}.$$

Defining now $N_e^{(e)}$ by the equation

$$\{1-(2N_e^{(e)})^{-1}\}^k = \prod_{i=1}^{k} \{1-(2N_i)^{-1}\},$$

it is clear that if k is small and the N_i large,

$$N_e^{(e)} \approx k\{N_1^{-1} + \cdots + N_k^{-1}\}^{-1}. \tag{200}$$

Thus the eigenvalue effective population size is effectively the harmonic mean of the various population sizes taken during the k-generation cycle. A parallel formula holds for $N_e^{(i)}$, although here it is easier to work through the probability $Q(t+k)$ that two genes in generation $t+k$ do *not* have the same ancestor in generation t. Clearly

$$Q(t+k) = \{1-(2N_{i-1})^{-1}\}Q(t+k-1),$$

and iteration over k generations gives

$$Q(t+k) = \prod_{i=1}^{k} \{1-(2N_i)^{-1}\}Q(t).$$

Elementary calculations now show that $N_e^{(i)}$ is also essentially equal to the harmonic mean of the various population sizes. Again, if $x(t)$ is the fraction of A_1 genes in generation t,

$$\mathrm{var}\{x(t+k) \mid x(t)\} = \tfrac{1}{2}k\{N_1^{-1}+N_2^{-1}+\cdots+N_k^{-1}\}x(t)\{1-x(t)\}+0(N_i^{-2}).$$

This shows that to a suitable approximation, $N_e^{(v)}$ is also the harmonic mean of the various population sizes.

We conclude this section by noting that many problems exist with the concept of the effective population size. Perhaps the most notable is the following. The expression "effective population size" is widely used in areas associated with population genetics, especially in connection with the evolution of the human population, by authors who appear to have no idea of its intimate connection to the Wright-Fisher model or of the fact that different concepts of the effective population size exist. The numerical values given by these different concepts can differ widely for a population whose size increases with time, as with the human population, so that many dubious claims about the "effective size" of the human population at some given time in the past exist in the literature.

Selection

So far we have totally ignored the possibility of natural selection in our stochastic models. Since selection is a central feature of evolutionary theory, we now have to discuss it in conjunction with stochastic models. We consider in turn the Wright–Fisher model and the Moran model.

The Wright-Fisher model with selection

Suppose then that both selection exists, that the genotypes A_1A_1, A_1A_2, and A_2A_2 have fitnesses given by (8). In view of (7) a reasonable Wright–Fisher model incorporating selection is found by assuming that the transition matrix for the number of A_1 individuals is (63), where now

$$\psi_i = \bar{w}^{-1}\{w_{11}x^2 + w_{12}x(1-x)\}, \tag{201}$$

where $x = i/2N$ and $\bar{w} = w_{11}x^2 + 2w_{12}x(1-x) + w_{22}(1-x)^2$. The main qualitative property of this model is that one or other absorb-

ing state, $X(\cdot) = 0$, $X(\cdot) = 2N$, is eventually reached. Despite this, essentially no quantitative results are known concerning the stochastic behavior of the model, and the best that can be done is to consider approximations. We do this in the next section by using diffusion theory, and for the moment foreshadow this approach by deriving an approximate formula for the probability that eventually $X(\cdot) = 2N$.

It is convenient to use the notation (see (9)) $w_{11} = 1 + s$, $w_{12} = 1+sh$ and $w_{11} = 1$, where we assume that s is of order N^{-1} and h is of order 1. Put $\alpha = 2Ns$ and, in (20), write $i = 2Nx$, $j = 2N(x+\delta x)$. Then this equation may be written

$$\begin{aligned}\pi(x) &= \sum \text{Prob}(x \to x+\delta x)\pi(x+\delta x) \\ &\approx \sum \text{Prob}(x \to x+\delta x)\big\{\pi(x) + \delta x \frac{d\pi(x)}{dx} + \tfrac{1}{2}(\delta x)^2 \frac{d^2\pi}{dx^2}\big\} \\ &= \pi(x) + \text{E}(\delta x)\frac{d\pi(x)}{dx} + \tfrac{1}{2}\text{E}(\delta x)^2 \frac{d^2\pi}{dx^2}.\end{aligned}$$

Under the assumptions we have made,

$$\begin{aligned}\text{E}(\delta x) &= (2N)^{-1}\alpha x(1-x)\{x + h(1-2x)\} + O(N^{-2}), \\ \text{E}(\delta x)^2 &= (2N)^{-1}x(1-x) + O(N^{-2}).\end{aligned}$$

Thus to the order of approximation we use, these calculations give

$$2\alpha\{x + h(1-2x)\}\frac{d\pi(x)}{dx} + \frac{d^2\pi}{dx^2} = 0.$$

The solution of this equation, subject to the obvious boundary conditions $\pi(0) = 0$, $\pi(1) = 1$, is

$$\pi(x) = \int_0^x \psi(y)\,dy \Big/ \int_0^1 \psi(y)\,dy, \tag{202}$$

where

$$\psi(y) = \exp\big(-\alpha y\{2h + y(1-2h)\}\big).$$

In the particular case $h = \frac{1}{2}$, for which the heterozygote is intermediate in fitness between the two homozygotes, this reduces to

$$\pi(x) = \{1 - \exp(-\alpha x)\}/\{1 - \exp(-\alpha)\}. \tag{203}$$

It is of interesting to use this approximate formula to get some idea of the effect of the selective differences on the probability of fixation of A_1. Suppose for example that $N = 10^5$, $s = 10^{-4}$, and $x = 0.5$. Then $\alpha = 20$ and, from (203), $\pi(0.5) = 0.999955$. By contrast, for $s = 0$ we have $\pi(0.5) = 0.5$. Evidently the rather small selective advantage 0.0001, which is too small to be observed in laboratory experiments, is nevertheless large enough in evolutionary terms to have a significant effect on the fixation probability. Clearly this occurs because, while selection might have only a minor effect in any generation, the number of generations until fixation occurs is so very large that the cumulative effect of selection is considerable.

To find further results concerning the selective theory of the Wright–Fisher model we will later turn to the vehicle that is most convenient to find these results, namely diffusion theory.

The Moran model with selection

Selection can be incorporated into the Moran model (92) - (94) by assuming differential birth rates or differential death rates. The two approaches give similar results so we consider here only the case where death rates differ. To do this we suppose that if at any time there are i A_1 genes in the population the probability that the next individual chosen to die is A_1 is

$$\mu_1 i/\{\mu_1 i + \mu_2(2N - i)\}. \tag{204}$$

If $\mu_1 = \mu_2$ there is no selection while if $\mu_1 < \mu_2$ the allele A_1 has a selective advantage over A_2. It follows that the transition matrix for the number of A_1 individuals has elements

$$\begin{aligned} p_{i,i-1} &= \mu_1 i(2N - i)/[2N\{\mu_1 i + \mu_2(2N - i)\}], && (205)\\ p_{i,i+1} &= \mu_2 i(2N - i)/[2N\{\mu_1 i + \mu_2(2N - i)\}], && (206)\\ p_{i,i} &= 1 - p_{i,i-1} - p_{i,i+1}. && (207) \end{aligned}$$

The matrix defined by these equations is a continuant, and the theory for these can be applied immediately. This theory shows that the probability π_i of eventual fixation of A_1, given an initial number of i A_1 individuals, is

$$\pi_i = \{1 - (\mu_1/\mu_2)^i\}/\{1 - (\mu_1/\mu_2)^{2N}\}. \tag{208}$$

If now $\mu_1/\mu_2 = 1 - \frac{1}{2}s$, where s is small and positive, A_1 has a slight selective advantage over A_2 and (208) can be approximated by

$$\pi(x) \approx \{1 - \exp(-\tfrac{1}{2}\alpha x)\}/\{1 - \exp(-\tfrac{1}{2}\alpha)\}, \qquad (209)$$

where $x = i/2N$ and $\alpha = 2Ns$. This formula differs from (203) by a factor of 2 in the exponents. This is not because the selective differences differ by a factor of 2, since indeed they do not, but from a more deep-rooted difference between the two models which we examine elsewhere.

It is possible to use formulae for continuant Markov chains to get expressions for the mean absorption time, conditional mean absorption times, and so on. We do not do this here since the formulae become very unwieldy and uninformative, and since also we later consider simple approximations for these quantities. It may finally be remarked that no formula is known for the eigenvalues of the matrix defined by (205).

Diffusion theory

Introduction

In previous sections we encountered some difficulty in deriving explicit formulae for several quantities of evolutionary interest, particularly when the population behavior was described by the Wright–Fisher model (35) or any of its generalizations. Even for the Moran model, where explicit formulae can often be found, the effects of the genetic parameters are sometimes obscured by the complexities of the expressions that arise. For both these reasons it would be most useful to us if we could approximate these quantities by reasonably accurate expressions which are comparatively simple, and which display explicitly the effects of the various genetic parameters. Fortunately there exists a general approach which very often does all this for us, namely in approximating the discrete process by a continuous-time continuous-space diffusion process.

A substantial and mathematically deep theory of diffusion processes exists. Our approach to diffusion processes does not, however, proceed through this theory, being often rather intuitive and avoiding theoretical niceties. We shall in particular assume without question the existence of a unique diffusion process having certain

properties that we require. We first consider the elements of the theory divorced from specific genetical applications, and later apply this theory to a variety of genetical models.

The forward and backward Kolmogorov equations

We consider a discrete Markov chain with state space $\{0, 1, 2, \ldots, 2N\}$, transition matrix $P = \{p_{ij}\}$ and initial value k for the random variable whose properties are described by this Markov chain. For convenience we write $p_{ki}^{(t)}$ as $f(i; k, t)$, so that

$$f(j; k, t+1) = \sum_{i=0}^{2N} f(i; k, t) p_{ij}. \tag{210}$$

We re-scale the space axis by a factor $(2N)^{-1}$ and consider the new variables

$$x = i(2N)^{-1}, \quad x + \delta x = j(2N)^{-1}, \tag{211}$$

and write $p = k(2N)^{-1}$. In all applications of interest to us, $\mathrm{E}(\delta x|x) = 0((2N)^{-\gamma})$ and $\mathrm{Var}(\delta x|x) = 0((2N)^{-\gamma})$, where $\gamma = 1$ or 2; now change the time scale so that possible changes in the random variable can occur at time points δt, $2\delta t$, $3\delta t, \ldots$, where $\delta t = (2N)^{-\gamma}$. The re-scaled process is of course essentially identical to the original process and in particular is still a discrete process. Nevertheless we feel that as $2N \to \infty$ the process converges in some way to a continuous-time continuous-space diffusion process, and our aim is to identify this diffusion process and to discover some of its properties.

Suppose that in the discrete process the moments of the change δx, given the current value x at time t, satisfy the equations

$$\mathrm{E}(\delta x) = a(x)\delta t + o(\delta t), \tag{212}$$

$$\mathrm{Var}(\delta x) = b(x)\delta t + o(\delta t), \tag{213}$$

$$\mathrm{E}(|\delta x|^3) = o(\delta t). \tag{214}$$

Here $a(x)$ and $b(x)$ are assumed to be functions of x but not of t. We write (210) in the form

$$f(x + \delta x; p, t + \delta t) = \int f(x; p, t) f(x + \delta x; x, \delta t)\, dx,$$

where here and below all integrals have terminals 0 and 1. We now formally expand on both sides as Taylor series in δt and δx. Using equations (212) – (214) and retaining leading terms only, we eventually arrive at the equation

$$\frac{\partial f(x;t)}{\partial t} = -\frac{\partial}{\partial x}\{a(x)f(x;t)\} + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\{b(x)f(x;t)\}. \tag{215}$$

This is the forward Kolmogorov (Fokker–Planck or diffusion) equation and is of fundamental importance in the theory of population genetics. This formal procedure can be justified by more advanced mathematical diffusion theory.

Since small $\delta t \to 0$ corresponds to large $2N$, we now assume that there exists a diffusion process on $[0, 1]$ that satisfies (212)–(214) and possesses a density function $f(x;t)$ which satisfies (215). We expect this process to approximate the original discrete process in the sense that, for $0 < g < h < 1$,

$$\int_g^h f(x;t)\,dx \tag{216}$$

provides a good approximation to the probability that the original unscaled discrete random variable is between $2Ng$ and $2Nh$ at time $(2N)^\gamma t$.

In the procedure leading to (215) little mention was been made of the initial value p of the diffusion variable, and p does not appear explicitly in (215). However, the function $f(x;t)$ should be written more fully $f(x;p,t)$, since the solution of the equation depends on the value of p. There is, however, a second equation that makes a more explicit and indeed fundamental use of the value of p. If we consider instead of the time points $(0,\, t,\, t+\delta t)$ the new time points $(0,\, \delta t,\, t+\delta t)$, we arrive at the equation

$$f(x;p,t+\delta t) = \int g(\delta p;p) f(x;p+\delta p,t) d(\delta p). \tag{217}$$

Here δp is the change in the value of the random variable in the time interval $(0, \delta t)$ and $g(\delta p;p)$, its probability density function. Expanding the integrand as above and retaining leading terms, we arrive at the equation

$$\frac{\partial f(x;p,t)}{\partial t} = a(p)\frac{\partial f(x;p,t)}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 f(x;p,t)}{\partial p^2}\,. \tag{218}$$

This is the backward Kolmogorov equation, which for several purposes is more useful than the forward equation (215).

Some care must be exercised in the interpretation of equation (218). As stated above, the density function $f(x;p,t)$ depends on p, and all that is claimed is that, as a function of p, this density function satisfies equation (218). The statement sometimes made that (218) implies a time reversal and that p is a random variable with x fixed is incorrect: the random variable in equation (218) is the current gene frequency x.

An explicit solution of (215), or of (218), can sometimes be achieved, as we see later. The solution is usually of the eigenfunction expansion form

$$f(x;p,t) = \sum_{i=1}^{\infty} g_i(x,p) \exp(-\lambda_i t), \tag{219}$$

where the λ_i $(0 \leq \lambda_1 < \lambda_2 < \lambda_3 \ldots)$ are eigenvalue constants and the $g_i(x,p)$, the associated eigenfunctions. This form of solution is clearly analogous to the spectral expansion of a Markov chain n-step transition matrix , a parallel we examine in more detail in particular cases. Remarkably, a considerable amount of information concerning the diffusion process (215) can be found without computing the explicit solution (219), as we now see.

Fixation probabilities

In this and the next three sections we assume without question the existence of a diffusion process on $[0,1]$ satisfying (212)–(214) and admitting a density function satisfying (215) and (218).

An equation parallel to (218) can be found by replacing $f(x;p,t)$ by $F(x;p,t)$ throughout, where

$$F(x;p,t) = \int_0^x f(y;p,t)\, dy, \tag{220}$$

so that

$$\frac{\partial F(x;p,t)}{\partial t} = a(p)\frac{\partial F(x;p,t)}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 F(x;p,t)}{\partial p^2}\,. \tag{221}$$

Suppose now that both $x = 0$ and $x = 1$ are absorbing states of the diffusion process. From equation (221) we arrive at the equation

$$\frac{\partial P_0(p;t)}{\partial t} = a(p)\frac{\partial P_0(p;t)}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 P_0(p;t)}{\partial p^2}\,, \tag{222}$$

where $P_0(p;t)$ is the probability that absorption has occurred at $x = 0$ at or before time t. The same equation holds for the probability $P_1(p;t)$ that absorption has occurred at $x = 1$ at or before time t. Although $P_0(p;t)$ and $P_1(p;t)$ obey the same equation, their values differ due to different boundary conditions. By letting $t \to \infty$, the probability $P_0(p)$ that absorption ever occurs at $x = 0$ satisfies the equation

$$0 = a(p)\frac{dP_0(p)}{dp} + \tfrac{1}{2}b(p)\frac{d^2P_0(p)}{dp^2}. \tag{223}$$

Since $P_0(p)$ clearly satisfies the boundary conditions $P_0(0) = 1$, $P_0(1) = 0$, it is straightforward to solve equation (223) explicitly to get

$$P_0(p) = \int_p^1 \psi(y)\,dy \Big/ \int_0^1 \psi(y)\,dy, \tag{224}$$

where

$$\psi(y) = \exp\bigl(-2\int^y \{a(z)/b(z)\}\,dz\bigr). \tag{225}$$

Similarly the probability $P_1(p)$ that absorption eventually occurs at $x = 1$ is found to be

$$P_1(p) = \int_0^p \psi(y)\,dy \Big/ \int_0^1 \psi(y)\,dy. \tag{226}$$

We have already found these formulae as approximations to the values in a finite Markov chain equations (202) and (203), where a different notation was used, and without reference to diffusion processes. Although we have carried out a scaling of the time axis in passing from the original Markov chain to the diffusion process, there is no need to re-scale the values (224) and (226) when using them as approximations in the Markov chain. This is no longer true for questions concerning the time until absorption, as we now see.

Fixation time properties

We start by assuming that both $x = 0$ and $x = 1$ are absorbing barriers and consider the mean time until one or other boundary is reached in the diffusion process of interest. Equation (222) and the corresponding equation for $x = 1$ show that if $\phi(t; p)$ is the density function of the time t until absorption occurs, then $\phi(t; p)$ satisfies the equation

$$\frac{\partial \phi(t;p)}{\partial t} = a(p)\frac{\partial \phi(t;p)}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 \phi(t;p)}{\partial p^2}\,. \tag{227}$$

Then

$$\begin{aligned} -1 &= -\int_0^\infty \phi(t;p)\,dt \\ &= -[t\phi(t;p)]_0^\infty + \int_0^\infty t\frac{\partial \phi}{\partial t}\,dt \\ &= 0 + \int_0^\infty t\left\{a(p)\frac{\partial \phi}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 \phi}{\partial p^2}\right\}dt \end{aligned}$$

so that

$$-1 = a(p)\frac{d\bar{t}(p)}{dp} + \tfrac{1}{2}b(p)\frac{d^2\bar{t}(p)}{dp^2}\,, \tag{228}$$

providing an interchange in the order of integration and differentiation is justified, that the mean fixation time is finite, and that $t\phi(t; p) \to 0$ as $t \to \infty$. Here

$$\bar{t}(p) = \int_0^\infty t\phi(t;p)\,dt \tag{229}$$

is the mean time until one or other absorbing boundary is reached, given the initial frequency p. The solution of (228), subject to the boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$, is best expressed in the form

$$\bar{t}(p) = \int_0^1 t(x;p)\,dx, \tag{230}$$

where

$$t(x;p) = 2P_0(p)[b(x)\psi(x)]^{-1}\int_0^x \psi(y)\,dy, \quad 0 \le x \le p, \quad (231)$$

$$t(x;p) = 2P_1(p)[b(x)\psi(x)]^{-1}\int_x^1 \psi(y)\,dy, \quad p \le x \le 1. \quad (232)$$

For the original Markov chain we approximate the mean absorption time by

$$(2N)^{\gamma}\bar{t}(p). \quad (233)$$

The representation (230) suggests a more detailed examination of the function $t(x;p)$. This function has the interpretation that

$$\int_{x_1}^{x_2} t(x;p)\,dx \quad (234)$$

is the mean time in the diffusion process that the random variable spends in the interval (x_1, x_2) before absorption. Correspondingly, we approximate the mean number of times in the Markov chain that the discrete random variable takes the value j $(= 2Nx)$ before absorption by

$$\bar{t}_{k,j} \approx (2N)^{\gamma-1}t(x;p). \quad (235)$$

It is possible to derive higher moments of the absorption time. For example, we have

$$\begin{aligned} -2\bar{t}(p) &= -2\int_0^\infty t\phi(t;p)\,dt \\ &= -[t^2\phi(t;p)]_0^\infty + \int_0^\infty t^2\frac{\partial\phi}{\partial t}\,dt \\ &= \int_0^\infty \left\{a(p)\frac{\partial t^2\phi(t;p)}{\partial p} + \tfrac{1}{2}b(p)\frac{\partial^2 t^2\phi(t;p)}{\partial p^2}\right\}dt \\ &= a(p)\frac{dS(p)}{dp} + \tfrac{1}{2}b(p)\frac{d^2S(p)}{dp^2}\,, \end{aligned} \quad (236)$$

where $S(p)$ is the second moment of the absorption time. In this procedure we have formally interchanged the order of integration and differentiation. Equation (236) can be solved for $S(p)$, subject to the boundary conditions $S(0) = S(1) = 0$, and hence a formula for the variance of the absorption time can be found.

The above formulae require modification when there is only one absorbing state. We do not go into details here and only state the conclusions. If 0 is the only absorbing state (230) continues to hold, but $t(x;p)$ must be redefined as

$$t(x;p) = 2\big(b(x)\psi(x)\big)^{-1} \int_0^x \psi(y)\,dy, \quad 0 \le x \le p, \qquad (237)$$

$$t(x;p) = 2\big(b(x)\psi(x)\big)^{-1} \int_0^p \psi(y)\,dy, \quad p \le x \le 1. \qquad (238)$$

Similarly, when 1 is the only absorbing state we have

$$t(x;p) = 2\big(b(x)\psi(x)\big)^{-1} \int_p^1 \psi(y)\,dy, \quad 0 \le x \le p, \qquad (239)$$

$$t(x;p) = 2\big(b(x)\psi(x)\big)^{-1} \int_x^1 \psi(y)\,dy, \quad p \le x \le 1. \qquad (240)$$

In both cases equations (233) and (234) hold.

The stationary distribution

We have assumed above that in the Markov chain we are interested in there has existed at least one absorbing state. In several cases of interest there are no absorbing states, and there exists a stationary distribution $\{\phi_j\}$ for the number of A_1 genes. Since an explicit expression for this distribution has not been found in many examples of genetic interest, we aim in this section to approximate this distribution by finding the stationary distribution of the approximating diffusion process. It will turn out that this leads to a very simple form for this approximating distribution in which the effects of the general parameters are clearly displayed.

Our starting point is the forward Kolmogorov equation in (215). If we integrate throughout formally with respect to x, there results eventually

$$\frac{\partial}{\partial t}[1 - F(x;t)] = a(x)f(x;t) - \tfrac{1}{2}\,\frac{\partial\{b(x)f(x;t)\}}{\partial x}\,. \tag{241}$$

Here $F(x;t)$ is the distribution function

$$F(x;t) = \int_0^x f(y;t)\,dy. \tag{242}$$

This formal derivation suggests that the right-hand side in (241) is the rate of flow of probability (from left to right) across the point x at time t. This interpretation can be verified, and we thus call the right-hand side in (241) the probability flux of the diffusion process. If a stationary distribution $f(x)$ exists this probability flux will be zero if $f(x;t)$ is replaced by $f(x)$, so that the stationary distribution satisfies the equation

$$-a(x)f(x) + \tfrac{1}{2}\,\frac{d\{b(x)f(x)\}}{dx} = 0. \tag{243}$$

Integration shows that the solution of this equation is

$$f(x) = \text{const}[b(x)]^{-1}\exp\Big(2\int^x a(y)/b(y)\,dy\Big), \tag{244}$$

where the constant is allocated so that

$$\int_0^1 f(x)\,dx = 1. \tag{245}$$

So far as the original Markov chain is concerned, our interpretation is that the diffusion approximation to the stationary probability that the random variable in the Markov chain lies in $[2Nx_1, 2Nx_2]$ is given by

$$\Pr\{2Nx_1 \le X \le 2Nx_2\} \approx \int_{x_1}^{x_2} f(x)\,dx. \tag{246}$$

This approximation turns out to be satisfactory except when $x_1 \approx 0$ or $x_2 \approx 1$, in which case special arguments, which we shall consider later, are needed.

Conditional processes

In this section we consider diffusion processes where 0 and 1 are both absorbing barriers. It is often of interest to single out those diffusions for which a nominated absorbing barrier is eventually reached, and we do this by the theory of conditional processes. For definiteness we assume the barrier in question is $x = 1$, although we shall also give some formulae applying when it is $x = 0$.

Since there can be no stationary distribution for such conditional processes, and since also there is no interest in fixation probabilities, interest centers almost entirely on properties of the time until fixation. Regarding the diffusion as an approximation to a Markov chain, it is clear from (28) that the sojourn time function (231) and (232) should be replaced by

$$t^*(x;p) = t(x;p)P_1(x)/P_1(p). \tag{247}$$

This gives

$$t^*(x;p) = 2P_0(p)P_1(x)[P_1(b)b(x)\psi(x)]^{-1}\int_0^x \psi(y)\,dy,\ 0 \le x \le p, \tag{248}$$

$$t^*(x;p) = 2P_1(x)[b(x)\psi(x)]^{-1}\int_x^1 \psi(y)\,dy, \quad p \le x \le 1. \tag{249}$$

We consistently use the asterisk notation $(*)$ to denote functions computed conditional on eventual absorption at $x = 1$ and, below, the double asterisk notation $(**)$ when conditioning on eventual absorption at $x = 0$. Thus conditional on eventual absorption at $x = 0$, the sojourn time function is, by arguments parallel to those just given,

$$t^{**}(x;p) = 2P_0(p)[b(x)\psi(x)]^{-1}\int_0^x \psi(y)\,dy, \quad 0 \le x \le p, \tag{250}$$

$$t^{**}(x;p) = 2P_0(x)P_1(p)[P_0(p)b(x)\psi(x)]^{-1}\int_x^1 \psi(y)\,dy,\ p \le x \le 1. \tag{251}$$

Equation (27) suggests an even stronger result than these, namely that the conditional density functions $f^*(x;p,t)$ and $f^{**}(x;p,t)$ of

the diffusion variable at time t satisfy

$$f^*(x;p,t) = f(x;p,t)P_1(x)/P_1(p), \quad (252)$$
$$f^{**}(x;p,t) = f(x;p,t)P_0(x)/P_0(p). \quad (253)$$

It is clear that equations (248)–(251) can be used immediately to find the conditional mean times before absorption.

We now indicate another way in which these conditional mean times can be derived, namely by finding the conditional process analogues to the Kolmogorov equations (215) and (218). To do this we must find the conditional process drift and diffusion coefficients analogous to those defined by (212) and (213). Let A be the event that absorption eventually occurs at $x = 1$ and $p^*(x \to x + \delta x)$ be the conditional probability density, given A, of a transition from x to $x + \delta x$ in time δt. Then

$$\begin{aligned} p^*(x \to x + \delta x) &= p(x \to x + \delta x \text{ and } A)/\text{Prob}(A) \\ &= p(x \to x + \delta x)P_1(x + \delta x)/P_1(x) \\ &\approx p(x \to x + \delta x)[1 + \delta x P_1'(x)/P_1(x)], \end{aligned}$$

where we use the dash notation ($'$) to refer to differentiation with respect to x. Hence, in an obvious notation,

$$\begin{aligned} a^*(x)\delta t &= \int (\delta x)p^*(x \to x + \delta x)d(\delta x) \\ &\approx \int (\delta x)p(x \to x + \delta x)[1 + (\delta x)P_1'(x)/P_1(x)]d(\delta x) \\ &= \{a(x) + b(x)P_1'(x)/P_1(x)\}\delta t. \end{aligned}$$

Thus it follows that

$$a^*(x) = a(x) + b(x)P_1'(x)/P_1(x). \quad (254)$$

It is found similarly that

$$b^*(x) = b(x). \quad (255)$$

In the case of the Wright–Fisher model, with no selection or mutation, so that $a(x) = 0, b(x) = x(1 - x), P_1(x) = x$, equations (254) and (255) give

$$a^*(x) = 1 - x, \quad b^*(x) = x(1 - x). \quad (256)$$

These values have already been used in equation (55), and was found for that equation by a process different from the above. In the same model, when the condition is made that the allele of interest is eventually lost,

$$a^{**}(x) = -x, \quad b^{**}(x) = x(1-x). \tag{257}$$

The arguments leading to these formulae can be made more rigorous by suitable handling of small-order terms. The conditional density $f^*(x;p,t)$ now satisfies the forward equation

$$\frac{\partial f^*(x;p,t)}{\partial t} = -\frac{\partial\{a^*(x)f^*(x;p,t)\}}{\partial x} + \tfrac{1}{2}\,\frac{\partial^2\{b^*(x)f^*(x;p,t)\}}{\partial x^2} \tag{258}$$

and the backward equation

$$\frac{\partial f^*(x;p,t)}{\partial t} = a^*(p)\frac{\partial f^*(x;p,t)}{\partial p} + \tfrac{1}{2}b^*(p)\frac{\partial^2 f^*(x;p,t)}{\partial p^2}\,. \tag{259}$$

Using (252), (254) and (255) it is easy to check that these are consistent with (215) and (218). The conditional mean absorption time may now be found by using $a^*(x)$ and $b^*(x)$ in (239) and (240), and the resulting value agrees with that found from (248) and (249). This final approach is more general in that it uses the defining equations (258) and (259) and thus can be used to find higher moments of the conditional absorption time. We take this point up later when considering specific applications.

Parallel calculations apply, with the obvious changes, to find the conditional density function $f^{**}(x;p,t)$ when the condition is made that the allele of interest is eventually lost from the population.

Some diffusion process theory

As mentioned above, there exists a deep mathematical theory of diffusion processes. In this section we consider those parts of the theory that are of use to us in genetic processes. Because the random variable of interest to us is the frequency of some allele, we consider only diffusion processes on the interval $[0,1]$.

The drift and diffusion functions $a(x)$ and $b(x)$ were introduced in equations (213) and (212). They may be used to define the im-

portant functions $p(x)$ and $m(x)$, defined respectively by

$$p(x) = \int_c^x \exp\bigl(-2\int^y a(z)/b(z)\,dz\bigr)dy, \tag{260}$$

$$m(x) = 2\int_c^x \{b(y)\}^{-1}\exp\bigl(2\int^y a(z)/b(z)\,dz\bigr)dy, \tag{261}$$

for some arbitrary constant c. Up to a linear transform, $p(x)$ is identical to the fixation probability $P_1(x)$. A diffusion is said to be on its natural scale if $p(x) = x$, which, from (260), is equivalent to $a(x) = 0$. For any diffusion not on its natural scale it is possible to find a transformed random variable (indeed the transformation is $x \to p(x)$) that is, and this explains the intimate link between $p(x)$ and $P_1(x)$. For this reason, $p(x)$ is called the *scale function* of the diffusion process. The function $m(x)$ is called the *speed function* of the process.

The functions $p(x)$ and $m(x)$ are central to many properties of diffusion processes, and we now show how they can be used to elucidate boundary behavior. Let r be an arbitrary point in $(0, 1)$ and s be one or other boundary point (that is $s = 0$ or $s = 1$). From $p(x)$ and $m(x)$ we compute the functions

$$u(s) = \int_r^s m(x)dp(x), \tag{262}$$

$$v(s) = \int_r^s p(x)dm(x). \tag{263}$$

The nature of the boundary s is exhibited as follows:

$u(s)$	$v(s)$	boundary type	accessible?	absorbing?
$< \infty$	$< \infty$	regular	yes	no
$< \infty$	$= \infty$	exit	yes	yes
$= \infty$	$< \infty$	entrance	no	no
$= \infty$	$= \infty$	natural	no	yes

(264)

A boundary is accessible if there exists positive probability that it can be reached in finite time from a given interior point, and is absorbing if the process remains forever at the boundary if it should reach it. We later given genetic examples of some of these various boundaries.

Applications of diffusion theory in genetics

In this section we apply some of the diffusion theory considered in the previous section to various Markov chain models arising in population genetics in order to arrive at various conclusions of evolutionary interest.

Our first aim is to see how the behavior of a given Markov chain can be mimicked by a diffusion process on $[0, 1]$. To do this it is convenient to start with the general Wright–Fisher model specified by (63) and (201). In this model the variable considered is the number j of A_1 genes in a diploid population of fixed size N and thus has state space $\{0, 1, 2, \ldots, 2N\}$. To work with a variable whose state space is closer to that of the diffusion process, we consider instead the fraction x of A_1 genes in the population, whose state space is $\{0, (2N)^{-1}, \ldots, 1\}$. We assume the notation x for the frequency of A_1 throughout, and also write p for the initial frequency of A_1.

So far as other notation is concerned, it is convenient to adopt the notation given in (9) that the genotype fitnesses are denoted by

$$w_{11} = 1 + s, \quad w_{12} = 1 + sh, \quad w_{22} = 1. \tag{265}$$

Further, when mutation exists, we assume mutation rates u (from $A_1 \to A_2$) and v (from $A_2 \to A_1$).

The diffusion model we concentrate on requires that s, u and v are all $0(N^{-1})$. We make this assumption throughout, and then put

$$\alpha = 2Ns, \quad \beta_1 = 2Nu, \quad \beta_2 = 2Nv \tag{266}$$

where α, β_1, and β_2 are all $0(1)$. Then standard binomial formulae for the model (63) show that

$$\begin{aligned} \mathrm{E}(\delta x \mid x) &= \big(\alpha x(1-x)\{x + h(1-2x)\} - \beta_1 x + \beta_2(1-x)\big)(2N)^{-1} \\ &\quad + o(N^{-1}), \\ \mathrm{Var}(\delta x \mid x) &= x(1-x)(2N)^{-1} + o(N^{-1}), \\ \mathrm{E}\{|\delta x|^3\} &= o(N^{-1}). \end{aligned} \tag{267}$$

These moments fit into the format (212)–(214) provided we choose

$$\begin{aligned} \delta t &= (2N)^{-1}, && (268)\\ b(x) &= x(1-x), && (269)\\ a(x) &= \alpha x(1-x)\{x + h(1-2x)\} - \beta_1 x + \beta_2(1-x). && (270) \end{aligned}$$

The requirement (268) is met by taking unit time in the diffusion process to correspond to $2N$ generations in the Markov chain. It is important to keep this scaling in mind when considering the relation between "time" properties in the diffusion process and those in the Markov chain. We now consider some properties of the diffusion process on $[0, 1]$ with drift and diffusion coefficients given respectively by (270) and (269).

Before proceeding we observe that in practical applications the idealized model (63) will probably have to be replaced by something more complex, perhaps one or other of the models discussed above in connection with effective population sizes. At the end of the next section we pursue this point for one particular such complex model. Although the theory is by no means clear, it seems likely that all the diffusion results given below will continue to hold, at least to a good approximation, when N is replaced by the variance effective population size $N_e^{(v)}$. Except for the case considered at the end of the next section we make no further explicit mention of this point in these notes.

The first step in discussing properties of the diffusion process with the drift and diffusion coefficients (269) and (270) is to compute the scale function and speed measure of the process, defined by (260) and (261). These become

$$p(x) = \int_c^x y^{-2\beta_2}(1-y)^{-2\beta_1} \exp\{\alpha(2h-1)y^2 - 2\alpha h y\}\, dy, \quad (271)$$

$$m(x) = 2\int_c^x y^{2\beta_2 - 1}(1-y)^{2\beta_1 - 1} \exp\{-\alpha(2h-1)y^2 + 2\alpha h y\}\, dy, \quad (272)$$

for an arbitrary constant c. We first use these expressions to consider boundary behavior. Use of (271) and (272) in (262) and (263) shows that near $x = 1$, the functions $u(x)$ and $v(x)$ take the form (for

$\beta_1 \neq \frac{1}{2}$)

$$u(x) = A + 0(1-x)^{1-2\beta_1}, \quad v(x) = B + 0(1-x)^{2\beta_1}.$$

Here A and B are constants whose precise values are unimportant. It follows that $v(x)$ is always finite at $x = 1$, but that $u(x)$ is finite at this point only if $\beta_1 < \frac{1}{2}$. From this we conclude that the boundary $x = 1$ is regular (accessible but non-absorbing) if $\beta_1 < \frac{1}{2}$ and entrance (inaccessible and non-absorbing) if $\beta_1 > \frac{1}{2}$. The same conclusion holds for the boundary $x = 0$, with β_2 replacing β_1. The values of α and h are irrelevant to these boundary descriptions. The case $\beta_1 = \frac{1}{2}$ is easily handled separately.

The intuitive meaning of these conclusions is clear enough. If the mutation rate from A_1 to A_2 and the population size are jointly large enough there is zero probability that the frequency of A_1 can ever achieve the value unity. Of course this conclusion applies for the diffusion process and it not true for the Markov chain (63).

If $\beta_1 = 0$ the boundary $x = 1$ is found to be exit (accessible and absorbing), and this again accords with what we expect since, if the boundary is reached, the absence of mutation from A_1 to A_2 means that the frequency of A_1 remains forever at unity. The fact that the boundary is accessible is less obvious intuitively – it is possible that a natural boundary is absorbing but not accessible, that is for which there is zero probability that it is reached by diffusion from within $(0, 1)$.

The functions $p(x)$ and $m(x)$ are also central to the calculation of fixation probabilities and stationary distributions respectively, when these are appropriate. We defer consideration of these until we take up specific cases later.

We conclude this section by emphasizing that our main interest is in Markov chain models such as (63), and we view diffusion processes mainly as approximations to these. Usually the approximations are excellent, but in some instances, particularly near the boundaries $x = 0, x = 1$ they are less so, and for these cases some care is needed in proceeding.

No selection or mutation

When there is no selection or mutation the model defined by (63) and (201) reduces to (35). Rather complete knowledge of the diffu-

sion approximation to this model is available, and in this section we explore this in some detail. Clearly we have

$$a(x) = 0, \quad b(x) = x(1-x), \tag{273}$$

and the forward equation becomes

$$\frac{\partial f(x;t)}{\partial t} = \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\{x(1-x)f(x;t)\}. \tag{274}$$

The solution of this equation, and others more complex, was achieved in a series of papers by Kimura (1955a, b, c, 1956, 1957). The explicit solution of (274), subject to the requirement $x = p$ when $t = 0$, is

$$\begin{aligned} f(x;p,t) &= \sum_{j=1}^{\infty} \frac{4(2j+1)p(1-p)}{j(j+1)} T^1_{j-1}(1-2p)T^1_{j-1}(1-2x) \\ &\quad \times \exp\left\{-\tfrac{1}{2}j(j+1)t\right\}. \end{aligned} \tag{275}$$

Here $T^1_{j-1}(x)$ is a Gegenbauer polynomial defined in terms of the hypergeometric function by

$$T^1_{j-1}(x) = \tfrac{1}{2}j(j+1)F\big(j+2, 1-j, 2, \tfrac{1}{2}(1-x)\big),$$

so that in particular

$$T^1_0(x) = 1, \quad T^1_1(x) = 3x. \tag{276}$$

The speed measure $m(x)$ for the coefficients (273) is such that

$$w(x) = dm(x)/dx = 2x^{-1}(1-x)^{-1}. \tag{277}$$

The probabilities $P_0(t)$ and $P_1(t)$ that the diffusion has reached 0 or 1 respectively by time t are

$$\begin{aligned} P_0(t) &= 1 - p + \sum_{j=1}^{\infty}(2j+1)p(1-p)(-1)^j F(1-j, j+2, 2, 1-p) \\ &\quad \times \exp\left(-\tfrac{1}{2}j(j+1)t\right), \end{aligned} \tag{278}$$

$$\begin{aligned} P_1(t) &= p + \sum_{j=1}^{\infty}(2j+1)p(1-p)(-1)^j F(1-j, j+2, 2, p) \\ &\quad \times \exp\left(-\tfrac{1}{2}j(j+1)t\right). \end{aligned} \tag{279}$$

The probability of ultimate fixation at $x = 1$ can be found by letting $t \to \infty$ in (279) or else by computing (226), with $\psi(x)$ defined by (225) and (273). Evidently $\psi(x) = 1$ and hence

$$\text{Prob(ultimate fixation at } x = 1) = p. \tag{280}$$

The mean fixation time can be found from (231) and (232). These equations give

$$\begin{aligned} \bar{t}(x;p) &= 2(1-p)/(1-x), \qquad 0 \le x \le p, \\ \bar{t}(x;p) &= 2p/x, \qquad\qquad\quad p \le x \le 1, \end{aligned} \tag{281}$$

so that the mean absorption time is

$$\bar{t}(p) = -2\{p \log p + (1-p)\log(1-p)\} \tag{282}$$

time units, or $-4N\{p \log p + (1-p)\log(1-p)\}$ generations. This agrees with the value (47) found without recourse to diffusion processes, and yields (49) and (50) as cases of particular interest.

The variance of the absorption time can be found, by further calculations, to be

$$4\big(p \int_p^1 \lambda(x)\,dx - (1-p)\int_0^p \lambda(x)\,dx\big) - \bar{t}(p)^2, \tag{283}$$

where

$$\lambda(x) = -2\int^x [(1-y)^{-1}\log y + y^{-1}\log(1-y)]\,dy. \tag{284}$$

The value (283) is in terms of (squared) time units and must be multiplied by $4N^2$ to be brought to a (squared) generation basis.

The complete distribution of the absorption time is implicit in (278) and (279), since

$$\text{Prob\{absorption time} \le t\} = P_0(t) + P_1(t). \tag{285}$$

Because of the form of the solutions (278) and (279), this expression is of most use when t is large. This solution may be supplemented by an asymptotic expansion the accuracy of which is best for small values of t. This asymptotic expansion, together with (285), then

yields a rather complete picture of the distribution of the absorption time.

What do these diffusion results mean for the Markov chain model (35)? The fixation probability (280) is exactly correct for this model, since we have seen that this value can be reached directly. The mean absorption time approximation has been confirmed. We have, however, arrived at the more detailed information, from (281) and (235), that if the initial number of A_1 genes in the Markov chain model is k, the mean number of generations for which this number assumes the value j, before reaching 0 or $2N$, is approximately

$$\begin{aligned} \bar{t}_{k,j} &= 2(2N-k)/(2N-j), \qquad j \leq k, \\ \bar{t}_{k,j} &= 2k/j, \qquad\qquad\qquad\quad j \geq k. \end{aligned} \tag{286}$$

The particular case $k = 1$, of particular interest to Fisher and Wright, gives $\bar{t}_{1,j} = 2j^{-1}$, in agreement with (52).

We turn now to the diffusion process spectral expansion (275). Recalling the difference in time scale between the Markov chain (35) and the diffusion process (274), it is clear that the expression $\exp\{-\frac{1}{2}j(j+1)\}$ is the analogue (see (82)) of the Markov chain n-step eigenvalue

$$[(1-\frac{1}{2N})(1-\frac{2}{2N})\ldots(1-\frac{j}{2N})]^{2N} \tag{287}$$

which is approximately

$$\exp -\{1+2+\ldots+j\} = \exp -\tfrac{1}{2}\{j(j+1)\}. \tag{288}$$

There is also a parallel between the eigenfunctions in (275), but we do not pursue the details of this.

We consider now processes conditional on the event that a specified boundary is eventually reached, and recover various formulae found earlier by other methods. We suppose for definiteness that $x = 1$ is the absorbing state ultimately reached. Equations (252) and (275) show that the density function of x at time t is

$$\begin{aligned} f^*(x;p,t) &= \sum_{j=1}^{\infty} \frac{4(2j+1)x(1-p)}{j(j+1)} T_{j-1}^1(1-2p) T_{j-1}^1(1-2x) \\ &\quad \times \exp\left\{-\tfrac{1}{2}j(j+1)t\right\}. \end{aligned} \tag{289}$$

For large t and small p this gives

$$f^*(x;p,t) \sim 6x\exp(-t), \tag{290}$$

so that

$$\lim_{t\to\infty} f(x \mid x \neq 0,1, \text{eventual fixation at } x=1) = 2x. \tag{291}$$

The functions $t^*(x)$ defined in (248) and (249) become

$$\begin{aligned} t^*(x) &= 2(1-p)x/\{p(1-x)\}, \qquad 0 \le x \le p, \\ t^*(x) &= 2, \qquad\qquad\qquad\qquad\quad p \le x \le 1. \end{aligned} \tag{292}$$

The conditional mean absorption time, found by integration, is then

$$t^*(p) = -2p^{-1}(1-p)\log(1-p). \tag{293}$$

In the Markov chain (35) this suggests the approximation that, if k is the initial value of the Markov variable,

$$\begin{aligned} t^*_{k,j} &\approx 2(2N-k)j/k(2N-j), \qquad j \le k, \\ t^*_{k,j} &\approx 2, \qquad\qquad\qquad\qquad\quad j \ge k, \end{aligned} \tag{294}$$

and a conditional mean fixation time of $-4Np^{-1}(1-p)\log(1-p)$ generations. One interesting case arises when $k=1$, so that there is initially only one gene of the allele of interest. One case of this concerns a unique selectively neutral new mutant destined for fixation. Equation (294) shows that, on average, this allele spends two generations at each possible frequency value, $(k = 1, 2, \ldots, 2N-1$, so that if t_1^* is the conditional mean fixation time,

$$\bar{t}_1^* = 4N - 2 \tag{295}$$

generations. It is instructive to see how easily information about the conditional process can be found from information concerning the unconditional process.

The value given in (295) is identical to the calculation given in (131), and it is interesting to discuss why this is so. Both expressions are identical to the conditional mean loss time, given initially $2N-1$ genes of the allele A_1. The reason why the unconditional mean time in the mutation process and the conditional mean time in the non-mutation process are essentially identical for the case $\theta = 2$ can be

seen from the fact that for $\theta = 2$, the drift and diffusion coefficients $a(x)$ and $b(x)$, given in (313) are identical to the conditional process drift and diffusion coefficients given in (257). Identical arguments show that the same mean time applies when there is a single initial A_1 gene when the condition is made, in the no mutation case, that A_1 eventually fixes in the population. This mean time then has the interpretation as the mean time back to the most recent common ancestor gene of all genes in the current population, as is discussed by Dr. Joyce when considering coalescent theory.

The conditional variance of the absorption time can be found by solving (236), subject to appropriate boundary conditions. Here we must use the conditional process drift coefficient

$$a^*(x) = 1 - x$$

rather than the unconditional value. It is found that

$$\begin{aligned}(\sigma^*)^2(p) &= 8\Big[\pi^2/6 + p^{-1}(1-p)\log(1-p)\{1 - (2p)^{-1}(1-p)\log(1-p)\} \\ &\quad - \sum_{j=1}^{\infty} p^j/j^2\Big]. \end{aligned} \tag{296}$$

In the limiting case $p \to 0$ this gives

$$(\sigma^*)^2 \approx 8[\pi^2/6 - 1.5] \approx 1.16, \tag{297}$$

or, for the process (35), $4.64N^2$ (squared generations). The complete distribution of the conditional absorption time can be found immediately from (279). We have

$$\begin{aligned}&\text{Prob}\{\text{absorption at } x = 1 \text{ before time } t \mid \text{eventual absorption at } x = 1\} \\ &= \frac{\text{Prob } \{ \text{ absorption at } x = 1 \text{ before time } t\}}{\text{Prob}\{ \text{ eventual absorption at } x = 1\}} \\ &= 1 + \sum_{j=1}^{\infty}(2j+1)(1-p)F(j+2, 1-j, 2, p)(-1)^i \exp\{-j(j+1)t\}.\end{aligned} \tag{298}$$

The expressions (293) and (297) can in principle be found from this distribution, but it is far simpler to arrive at them in the manner we have shown earlier.

We consider briefly the case where we condition on eventual loss of the allele A_1. Since for this case $a(x) = 0, b(x) = x(1-x), P_0(x)1-$

x, the analogue of equation (254) gives $a^{**} = -x$. The analogue of equation (293) is, in terms of generations, $t^{**} = -4N(p \log p)/(1-p)$. This is identical to the value given in equation (74), and this is not surprising since the value of the drift coefficient a^{**} given above is identical to that given in equation (313) below for the one-way mutation case when $\theta = 2$. Since the diffusion coefficients are also the same in the two cases, the entire stochastic behavior of the conditional process without mutation and the unconditional process with mutation (for $\theta = 2$) are identical. This seems for the moment to be no more than a curiosity, but it will turns out to have a more interesting interpretation when considering age and retrospective properties in genetic processes.

Selection

Fixation probabilities

Suppose now that the three genotypes have fitnesses given by (265). Assuming no mutation, the drift coefficient (270) becomes

$$a(x) = \alpha x(1-x)\{x + h(1-2x)\}. \tag{299}$$

From this the scale function and speed measure are calculated as

$$p(x) = \int_{c_1}^{x} \psi(y)\, dy, \tag{300}$$

$$m(x) = 2 \int_{c_2}^{x} y^{-1}(1-y)^{-1}\{\psi(y)\}^{-1}\, dy, \tag{301}$$

where

$$\psi(y) = \exp \alpha\{(2h-1)y^2 - 2hy\}. \tag{302}$$

Both boundaries $x = 0$, $x = 1$ are exit, and the probability that one or other boundary is eventually reached is unity. The respective probabilities are given by (224) and (226), with $\psi(y)$ defined by (302).

These expressions simplify significantly only in the case of no dominance ($h = \frac{1}{2}$), for which

$$P_1(p) = \{1 - \exp(-\alpha p)\}/\{1 - \exp(-\alpha)\}. \tag{303}$$

		$p = 0.001$			$p = 0.5$		
		$N = 10^4$	$N = 10^5$	$N = 10^6$	$N = 10^4$	$N = 10^5$	$N = 10^6$
	0.01	0.181	0.865	1.000	1.000	1.000	1.000
	0.001	0.020	0.181	0.865	1.000	1.000	1.000
s	0.0001	0.002	0.020	0.181	0.731	1.000	1.000
	0.00001	0.001	0.002	0.020	0.525	0.731	1.000

Table 1: Values of $P_1(p)$, for various values of N, s, and p, calculated from (303)

This agrees with the approximation (203) found without using diffusion methods. Some numerical values calculated from (303) are given in Table 1.

The conclusions to be drawn from this table are obvious enough. When N, s and p are jointly sufficiently large, fixation of the favored allele is essentially certain: this occurs roughly when $Nsp > 5$. As N, s or p decreases, the fixation probability decreases, and if $Ns < 0.1$ it does not differ (relatively) by more than 10% from the neutral value p. Perhaps the most striking conclusion is the very strong effect of selection in influencing fixation probabilities: as noted below equation (203), selective differences far too small to be found in the laboratory can nevertheless have a decisive effect on evolutionary behavior, at least in populations that are not too small. The same conclusion holds, at least qualitatively, when there is no dominance (that is $h \neq \frac{1}{2}$), although some minor modifications to the numerical values are necessary, especially when dominance is complete ($h = 0$ or $h = 1$). Even in the overdominant case ($sh > s > 0$) fixation of one of other allele is certain although this will normally take an extremely long time, and in practical terms one must then question the appropriateness of the assumptions made, in particular that there is no mutation and that the population size, selective differences, and dominance relationship remain unchanged throughout the entire fixation process.

The complete solution of the forward equation (215), with $a(x)$ and $b(x)$ defined by (299) and (269), is very complex. Nevertheless solutions were found by Kimura (1955a, b, c; 1957) initially for the no dominance case and subsequently for the general case. Unfortunately the very complexity of the solutions makes examination of their implications difficult.

Fixation time properties

Despite the very complex form of $f(x;t)$ referred to at the end of the previous section, a rather simple expression exists for the function $\bar{t}(x;p)$, defined in equation (234), and since this function summarizes perhaps the most important features of the transient behavior of the process, we now compute it for the selective model we are considering. All that is required to do this is to substitute (299) and (269) into the general formulae (231) and (232). For $h = \frac{1}{2}$ we get

$$\begin{aligned} \bar{t}(x;p) &= 2P_0(p)\{\alpha x(1-x)\}^{-1}\{\exp(\alpha x) - 1\}, && 0 \le x \le p \\ \bar{t}(x;p) &= 2P_1(p)\{\alpha x(1-x)\}^{-1}\{1 - \exp(-\alpha(1-x))\}, && p \le x \le 1 \end{aligned} \tag{304}$$

where $P_1(p)$ is found from (303) and $P_0(p) = 1 - P_1(p)$. For the Markov chain defined by (63) and (201), this implies that the mean number of generations for which there are $j = 2Nx$ A_1 alleles, given an initial number $k = 2Np$, is approximately

$$\begin{aligned} \bar{t}_{k,j} &= 2\{\exp(-2p) - \exp(-\alpha)\}\{\exp(\alpha x) - 1\} \\ &\qquad \times [\alpha x(1-x)\{1 - \exp(-\alpha)\}]^{-1} \quad (j \le k), \\ \bar{t}_{k,j} &= 2\{1 - \exp(-\alpha p)\}\{1 - \exp -\alpha(1-x)\} \\ &\qquad \times [\alpha x(1-x)\{1 - \exp(-\alpha)\}]^{-1} \quad (j \ge k). \end{aligned} \tag{305}$$

The mean time for fixation is found jointly from (230) and (304), but unfortunately no explicit evaluation of the integrals is possible, and numerical computation is necessary. There is, however, one case where useful progress can be made. If α and p are jointly sufficiently large so that fixation of the favored allele can be taken as being almost certain, we get

$$\begin{aligned} \bar{t}(x;p) &\approx 0, && 0 \le x \le p \\ \bar{t}(x;p) &\approx 2\{\alpha x(1-x)\}^{-1}, && p \le x \le 1 - 4\alpha^{-1}, \\ \bar{t}(x;p) &\approx 2\{\alpha x(1-x)\}^{-1}\{1 - \exp -\alpha(1-x)\}, && 1 - 4\alpha^{-1} \le p \le 1. \end{aligned} \tag{306}$$

The first equation shows that in the case considered, the frequency of the favored allele spends negligible time less than its initial value. The second equation is perhaps the most interesting. Converting to generations, it implies that in the Markov chain the mean time spent in the frequency range (x_1, x_2) where $p \le x_1 < x_2 \le 1 - 4\alpha^{-1}$,

is approximately

$$\int_{x_1}^{x_2} \left\{\tfrac{1}{2}sx(1-x)\right\}^{-1} dx$$

generations. This is identical to the value (14) found for the corresponding deterministic process, and we can conclude that the behavior of the process $(p, 1-4\alpha^{-1})$ is "quasi-deterministic". When the frequency exceeds $1-4\alpha^{-1}$ the deterministic value no longer gives an adequate guide to the stochastic behavior. In particular, the mean number of generations (in the Markov chain) for which $x = 1 - i(2N)^{-1}$, for small integers i, is essentially equal to the "neutral" value 2. This is severely overestimated by the deterministic formula, and clearly at this stage of the process selective forces have become of secondary importance, and random sampling almost wholly determines the gene frequency behavior.

For general values of h in $(0,1)$ the expressions (231) and (232) do not simplify readily. However, the general behavior just noted for the no dominance case continues to apply. Quasi-deterministic behavior obtains for sufficiently large p and α, at least until the frequency x of A_1 approaches unity, when selective forces once more can be ignored. The value of x where this occurs will depend to some extent on the level of dominance but will not differ materially from the value $1-4\alpha^{-1}$ found in the no dominance case.

The value $k = 1$ is of particular interest. Here we may approximate the probability $P_1\{(2N)^{-1}\}$ by $(2N)^{-1}\big(\int_0^1 \psi(y)\,dy\big)^{-1}$.

We turn now to mean absorption times conditional on eventual fixation (or loss) of a specified allele. The formulae appropriate to calculate this are (248) and (249) or (250) and (251). Perhaps the case of greatest interest is when $0 < h < 1$ and the condition is made that the favored allele fixes. When $h = \frac{1}{2}$, (248) and (249) give

$$\begin{aligned} t^*(x;p) &= 2e^{-\alpha x}\{1-e^{-\alpha(1-p)}\}\{e^{\alpha x}-1\}^2 \\ &\quad \times [\alpha x(1-x)\{1-e^{-\alpha}\}\{e^{\alpha p}-1\}]^{-1}, \quad 0 \le x \le p, \end{aligned} \tag{307}$$

$$t^*(x;p) = 2\{e^{\alpha x}-1\}\{e^{\alpha(1-x)}-1\}[\alpha x(1-x)\{e^{\alpha}-1\}^{-1}, \quad p \le x \le 1. \tag{308}$$

Similarly, if the condition is made that eventually A_1 is lost,

$$t^{**}(x;p) = 2\{e^{\alpha x}-1\}\{e^{\alpha(1-x)}-1\}[\alpha x(1-x)\{e^{\alpha}-1\}]^{-1}, \quad 0 \le x \le p, \tag{309}$$

$$t^{**}(x;p) = 2e^{-\alpha(1-x)}\{1-e^{-\alpha p}\}\{e^{\alpha(1-x)}-1\}^2 \times [\alpha x(1-x)\{1-e^{-\alpha}\}\{e^{\alpha(1-p)}-1\}]^{-1}, \quad p \le x \le 1. \tag{310}$$

There are several interesting points about these equations. First, the value of $t^*(x;p)$ for $x \ge p$ is identical to that of $t^{**}(x;p)$ for $x \le p$. This can be explained using time-reversal properties. The second point concerns the nature of the formula for $t^*(x;p)$ for very small p, or correspondingly $t^{**}(x;p)$ for very large p, and is relevant when considering a selectively favored new mutant destined for fixation. We observe that $t^*(x;p)$ is symmetric about $x = 0.5$; the mean time spent in any interval $(x, x+\delta x)$ is the same as the mean time spent in $(1-x-\delta x,\, 1-x)$. Even more surprisingly, $t^*(x;p)$ remains unchanged if α is replaced by $-\alpha$, so that a selectively disadvantageous mutant, if destined for fixation, spends as much time, on the average, in any frequency range as a corresponding selectively advantageous mutant destined for fixation. This remarkable fact will be reconsidered later in the light of time-reversal properties. It is indeed easy to see that the entire behavior of the conditional process is independent of the sign of s, since the diffusion coefficient $b^*(x)$, calculated from (255) and (269), is independent of s while the drift coefficient $a^*(x)$, calculated from (254), (299), and (303), is

$$a^*(x) = \tfrac{1}{2}\alpha x(1-x)/\tanh\big(\tfrac{1}{2}\alpha x\big).$$

Clearly $a^*(x)$ is independent of the sign of α. However, despite the symmetry of $t^*(x)$ around $x = \frac{1}{2}$, it is not true that $a^*(x) = a^*(1-x)$.

For arbitrary levels of dominance, (249) shows that with $p = (2N)^{-1}$,

$$t^*(x;(2N)^{-1}) = 2\big(b(x)\psi(x)\int_0^1 \psi(y)\,dy\big)^{-1}\int_0^x \psi(y)\,dy\int_x^1 \psi(y)\,dy, \tag{311}$$

where $\psi(y)$ is defined by (302). If this expression is written more fully as $t^*(x;\alpha,h,(2N)^{-1})$, it follows that

$$t^*(x;\alpha,h,(2N)^{-1}) = t^*(1-x;\alpha,1-h,(2N)^{-1}). \tag{312}$$

This implies that conditional mean fixation time properties for a favored allele are the same as those for the corresponding disadvantageous allele, provided the dominance relation is reversed. This generalizes the conclusion just reached for the case of no dominance.

One-way mutation

We turn now to the diffusion approximation to the one-way mutation Markov chain model (63), where A_1 mutates to A_2 (at rate u), with no reverse mutation. The drift and diffusion coefficients for the diffusion process approximating this Markov chain are

$$a(x) = -\tfrac{1}{2}\theta x, \quad b(x) = x(1-x), \tag{313}$$

where $\theta = 4Nu$. Clearly A_1 eventually becomes lost from the population and interest centers entirely on properties of the time until this loss occurs. These properties are defined in large measure by the function $t(x;p)$, and insertion of the coefficients (313) into (237) and (238) gives this function immediately. The values so calculated are given in (66), where allowance must be made for the fact that the time-scale assumed there assumes unit time for one generation. Perhaps the case of most interest is when $p = (2N)^{-1}$, so that, to a close approximation, the mean time that A_1 exists in the population is

$$\bar{t}\{(2N)^{-1}\} \approx \int_{(2N)^{-1}}^{1} 2y^{-1}(1-y)^{\theta-1}\,dy \tag{314}$$

generations. This is of order $2\log(2N)$ generations for moderate values of θ: a new mutation A_1 will not, on average, remain in the population for very long, or to attain a high frequency, if there is no recurrent mutation $A_2 \to A_1$.

The process we are considering, since it admits the possibility of only two alleles, is perhaps of limited interest. However, several of its properties throw considerable light on important features of the infinitely many alleles model (119). Some of these were already given in (125) and (126). It is clear that in the infinitely many alleles model we may normally expect several low-frequency alleles in the population. For example if $\theta = 1$, $2N = 10^6$, there will typically be about fifteen alleles present in the population at any time, and of these about ten will have a frequency less than 0.01. If θ is

small enough the most likely situation is where there is one allele at high frequency together with several alleles at a very low frequency. This is confirmed by the calculations following from the expression (128). Thus for $\theta = 0.1$ the probability that there exists an allele with frequency greater than 0.99 is about 0.63. For larger values of θ ($\theta > 1$ approximately) it becomes rather unlikely that such a high-frequency allele will exist, and the most likely configuration is one where a number of alleles exist at low but unequal frequencies. In all cases the least likely situation is one where two, three or four alleles exist with approximately equal frequencies. These arguments suggest an approach to testing whether a neutral model such as (119) and not, for example, one involving selection is adequate to explain observed allelic frequencies. This approach is discussed later in these notes.

There are two further points that are of interest in considering the model (119) and its evolutionary behavior. The first concerns the nature of the boundary $x = 1$ for the two allele model originally considered. Use of (313) in (264) shows that this boundary is entrance if $\theta \geq 1$. This implies that it is impossible to reach this boundary by diffusion from the interior of $(0, 1)$ in this case. It is therefore impossible to consider behavior conditional on the requirement that this boundary is reached, and further it is unnecessary to impose the condition that the boundary is not reached and then consider conditional behavior: this latter condition is already implicit and formulae such as (314) apply immediately. When $\theta < 1$ the boundary $x = 1$ is regular and hence attainable and now new behavior arises under the condition that this boundary is not reached. Again assuming $p = (2N)^{-1}$ we find that (314) must be replaced, conditional on $x = 1$ not being reached, by

$$\bar{t}\{(2N)^{-1}\} = \int_{(2N)^{-1}}^{1} 2y^{-1}(1-y)^{1-\theta}\, dy \tag{315}$$

generations. The integrand in (315) has the usual interpretation that its integral over any frequency range provides the mean time that the allele frequency spends in this range before the allele is eventually lost.

The second point concerns the frequency of the most frequent allele. The argument that led to (128) shows that for $0.5 \leq x \leq 1$ the

probability density function of the frequency of the most frequent allele in the infinitely many alleles model is, at equilibrium,

$$f(x) = \theta x^{-1}(1-x)^{\theta-1}. \tag{316}$$

For values of x less than 0.5 a deeper argument is clearly required: nevertheless the probability density function of the most frequent allele can be found for these values also.

Two-way mutation

Suppose now in the model (63) that mutation both from A_1 to A_2 (at rate u) and from A_2 to A_1 (at rate v) occurs, with no selection. As we have already seen, there will now exist a stationary distribution for the frequency x of A_1 for which we already have an exact expression (77) for the mean and an approximation expression (78) for the variance.

Our aim now is to approximate the entire distribution by diffusion methods. The drift and diffusion coefficients are found from (270) (putting $\alpha = 0$) and (269) and then (244) leads to the stationary distribution

$$f(x) = \frac{\Gamma\{2\beta_1 + 2\beta_2\}}{\Gamma\{2\beta_1\}\Gamma\{2\beta_2\}} x^{2\beta_2-1}(1-x)^{2\beta_1-1}. \tag{317}$$

The mean and variance of this distribution are $\beta_2/(\beta_1+\beta_2)$ and $\beta_1\beta_2/\{(\beta_1+\beta_2)^2(2\beta_1+2\beta_2+1)\}$ respectively, and these agree with the exact and approximate values given in (77) and (78), once allowance is made for a change of scale.

This stationary distribution allows a third derivation of equations (79) and (80). If $u = v$ and $4Nu = \theta$, then $2\beta_1 = 2\beta_2 = \theta$ and thus

$$f(x) = \frac{\Gamma(2\theta)}{\Gamma(\theta)\Gamma(\theta)}\, x^{\theta-1}(1-x)^{\theta-1}.$$

From this, the probability that two genes drawn at random from the population are of the same allelic type is

$$\int_0^1 f(x)\{x^2 + (1-x)^2\} = \frac{1+\theta}{1+2\theta}\,, \tag{318}$$

in agreement with (79) and (80).

The general shape of the curve (317) is clear enough. For small β_1 and β_2, that is small mutation rates and/or population sizes, most of the probability mass is in the extremes of the distribution, so that the most likely situation is one where one or other allele is at a low frequency or is even temporarily absent from the population. When β_1 and β_2 are large the variance becomes small and only small deviations are likely from the mean.

When selection is also allowed, together with two-way mutation, there will still exist a stationary distribution, although its form is naturally more complicated than that in (317). Use of the complete expressions (269) and (270) gives, from (244), the formula

$$f(x) = \text{const } x^{2\beta_2-1}(1-x)^{2\beta_1-1}\exp\{2\alpha hx - \alpha(2h-1)x^2\} \quad (319)$$

for this distribution, where the constant is a function of β_1, β_2, α, and h and may be found in principle by normalization.

Time-reversal and age properties

It is interesting that information about the past behavior of diffusion processes allowing a stationary distribution can be obtained by determining properties of the future behavior. We should therefore be able to use some of the conclusions reached above to discuss past behavior of various processes, and in particular to find properties of the "age" of an allele.

The time-reversal property states that for any diffusion on $[0,1]$ admitting a stationary distribution, the probability of any sample path leading from x (at time 0) to y (at time t) is equal to that of the "mirror-image" path leading from y (at time $-t$) to x (at time 0). Unfortunately this observation is not immediately useful for several questions of interest in population genetics, since these questions refer to processes for which either the boundary $\{0\}$, or the boundary $\{1\}$, or both, are accessible absorbing states of the diffusion process, and thus for which no stationary distribution exists. This problem can be overcome in the following way.

Suppose that $\{0\}$ is an absorbing state but that $\{1\}$ is not: this will occur in practice, for example, if there is mutation from A_1 to A_2 but no reverse mutation. Now introduce mutation from A_2 to A_1 at rate ϵ: a stationary distribution now exists and reversibility arguments apply. In particular, given a current value x for the

frequency of A_1, the distribution of the time (in the future) until $\{0\}$ is next reached is identical to that of the time (in the past) that it was last left. Now let $\epsilon \to 0$: the distribution of the time (in the future) until the frequency reaches 0 converges to that applying when $\epsilon = 0$. The limiting distribution is then identical to the age distribution of an allele which arose as a unique new mutation and whose current frequency is x. This argument can be made more precise by introducing a "return" process whereby the frequency of A_1 is returned from 0 to δ ($\delta > 0$) whenever 0 is reached: in practice we put $\delta = (2N)^{-1}$ to correspond to the frequency of a new mutant. We now give some examples of the conclusions reached by this argument.

Consider first the case of no selection or mutation. Assume the allele A_1 arose by a unique mutation in an otherwise pure A_2A_2 population and is now observed with frequency x. The distribution of its age is thus the distribution of its time until loss, conditional on the event that eventual loss does occur. This distribution can be found by centering attention on A_2 (with current frequency $1-x$) rather than A_1, and is then given by (298) with $p = 1-x$. The mean age can be found either through this distribution or alternatively by replacing p by $1-x$ in (293). This leads to a neutral theory mean age of $-4Nx(1-x)^{-1}\log x$ generations. The variance of the age is found by putting $p = 1-x$ in (296).

A parallel formula can be found when we assume fitness values $1+s$ for A_1A_1, $1+\frac{1}{2}s$ for A_1A_2 and 1 for A_2A_2. Use of (309) (with $p = x$) shows that the mean age of A_1, given that it is currently observed with frequency x, is

$$\int_0^x 4N[\alpha\{e^\alpha - 1\}]^{-1}\{e^{\alpha y} - 1\}\{e^{\alpha(1-y)} - 1\}\{y(1-y)\}^{-1}\,dy$$
$$+ \int_x^1 4N\{1 - e^{-\alpha x}\}[\alpha\{1 - e^{-\alpha}\}\{e^{\alpha(1-x)} - 1\}]^{-1}e^{-\alpha(1-y)}\{e^{\alpha(1-y)} - 1\}^2$$
$$\times \{y(1-y)\}^{-1}\,dy \qquad (320)$$

generations. This converges to the neutral theory expression as $\alpha \to 0$, as we expect, and the form of the integrand allows calculation of the mean time, in the past, that the frequency of A_1 assumed a value in any arbitrary interval (y_1, y_2).

Suppose now A_1 mutates to A_2 at rate u with no reverse mutation. If one initial A_1 gene occurred by a unique mutation and the frequency of A_1 is currently observed at x, the mean age of A_1 is, from (66),

$$4N(1-\theta)^{-1}\int_0^x y^{-1}\{(1-y)^{\theta-1}-1\}\,dx$$
$$+\,4N(1-\theta)^{-1}\{1-\{(1-x)^{1-\theta}\}\int_x^1(1-y)^{\theta-1}\,dy \quad (321)$$

generations. A case of particular interest is that for which $x = 1$, corresponding to temporary fixation of A_1: this evaluation is allowed only when $\theta < 1$.

It is also possible to consider the mean age of A_1 conditional on the requirement that the frequency of A_1 was never unity in the past. This is identical to the mean time for loss of A_1 given that its future frequency never achieves the value unity. This is given by (321) for $\theta > 1$, since then the condition that the frequency of A_1 never reaches unity is automatically satisfied. For $\theta < 1$ the probability that the frequency of A_1 never reaches unity given a current value x is found from (224) to be $(1-x)^{1-\theta}$. Use of (250) and (251) then shows that the conditional mean age of A_1 is

$$\int_0^x 4Ny^{-1}(1-\theta)^{-1}\{1-(1-y)^{1-\theta}\}\,dy$$
$$+\int_x^1 4Ny^{-1}(1-\theta)^{-1}(1-y)^{1-\theta}\{(1-x)^{\theta-1}-1\}\,dy. \quad (322)$$

This reduces to the expression (315) for $x = (2N)^{-1}$.

More "age" properties of this model will be considered by Dr. Joyce.

Multi-allele diffusion processes

In this section we consider diffusion approximations to finite Markov chain M-allele models of the form (115).

The simplest version of the model (115) arises when the function ψ_i takes the value $X_i/2N$. It is clear that in this Markov chain model the probability of fixation of any allele is initial frequency, and we also have the eigenvalue formula (116) concerning the rate of decrease of the probability that j or more alleles exist at time t. To obtain further results we turn to the diffusion approximation to (115).

We write $x_i = X_i/2N$ $(i = 1, 2, \ldots, M-1)$ and let δx_i be the change in x_i from one generation to the next. Then elementary theory shows that, given $x_1, \ldots, x_{M-1}$,

$$\mathrm{E}(\delta x_i) = 0, \quad \mathrm{Var}(\delta x_i) = (2N)^{-1} x_i(1 - x_i),$$

and

$$\mathrm{Covar}(\delta x_i, \delta x_j) = -(2N)^{-1} x_i x_j.$$

These values lead to the following partial differential equation for the joint density function $f = f(x_1, \ldots, x_{M-1}; t)$ of $x_1, \ldots, x_{M-1}$ at time t, where unit time corresponds to $2N$ generations:

$$\begin{aligned} \frac{\partial f}{\partial t} &= \tfrac{1}{2} \sum_{i=1}^{M-1} \frac{\partial^2 f}{\partial x_i^2} \{x_i(1 - x_i)\} \\ &\quad - \sum_{i<j} \sum \frac{\partial^2 f}{\partial x_i \partial x_j} \{x_i x_j\}. \end{aligned}$$

This is a generalization of equation (274) and admits an eigenfunction solution generalizing (275). The corresponding backward equation is

$$\frac{\partial f}{\partial t} = \tfrac{1}{2} \sum_i p_i(1 - p_i) \frac{\partial^2 f}{\partial p_i^2} - \sum_{i<j} \sum p_i p_j \frac{\partial^2 f}{\partial p_i \partial p_j},$$

where p_i is the initial value of x_i. This equation may be used to find various fixation probabilities. The probability π $(= \pi(p_1, p_2, \ldots, p_{M-1}))$ of any fixation event satisfies

$$\tfrac{1}{2} \sum p_i(1 - p_i) \frac{d^2 \pi}{dp_i^2} - \sum_{i<j} \sum p_i p_j \frac{d^2 \pi}{dp_i dp_j} = 0, \qquad (323)$$

subject to the appropriate boundary conditions. For example, the probability that A_i eventually fixes satisfies (323) together with the

boundary conditions

$$\pi(p_1,\ldots,p_{M-1}) = 1 \qquad \text{if} \quad p_i = 1,$$
$$\pi(p_1,\ldots,p_{M-1}) = 0 \qquad \text{if} \quad p_j + p_m + \ldots + p_u = 1 \quad (j,m,\ldots,u \neq i).$$

The solution of these equations is $\pi = p_i$, which we know also to be exactly correct for the model (115) with $\psi_i = X_i/2N$. Suppose now that we wish to find the probability π that ultimately A_i and A_j are the last two alleles to exist. Here the boundary conditions are

$$\pi(p_1,\ldots,p_{M-1}) = 1 \qquad \text{if} \quad p_i + p_j = 1,$$
$$\pi(p_1,\ldots,p_{M-1}) = 0 \qquad \text{if} \quad p_m + p_s + \ldots + p_u = 1 \quad (m,s,\ldots,u \neq i,j),$$

and the solution of (323) satisfying these conditions is

$$\pi = p_i p_j \{(1-p_i)^{-1} + (1-p_j)^{-1}\}.$$

In the case $M = 3$ this shows, for example, that the probability that A_1 is the first allele lost is

$$p_2 p_3 \{(1-p_2)^{-1} + (1-p_3)^{-1}\}.$$

Similar probabilities may be found for other fixation events.

We turn now to questions concerning the time until various fixation events occur. The development is easiest when $M = 3$, so we discuss the analysis in detail in this case only, and quote results for larger values of M.

For the case $M = 3$, define T_i as the time required until exactly i $(i = 1, 2)$ alleles exist in the population. We first find an expression for $\mathrm{E}(T_1)$. Conditional on the event that A_1 is the last remaining allele, the mean of T_1 is

$$\mathrm{E}(T_1 \mid A_1) = 2p_1^{-1}(1-p_1)\log(1-p_1),$$

from (293). Since the probability is p_1 that indeed A_1 is the last remaining allele we have

$$\mathrm{E}(T_1) = -2[(1-p_1)\log(1-p_1)+(1-p_2)\log(1-p_2)+(1-p_3)\log(1-p_3)].$$

Clearly this value can be extended immediately to the case of an arbitrary number M alleles to get

$$\mathrm{E}(T_1) = -2\sum(1-p_i)\log(1-p_i). \tag{324}$$

It is equally straightforward to use the analysis leading to (296) to find an expression for the variance of T_1.

In the three-allele case we find $\mathrm{E}(T_2)$ as follows. The event $T_2 \leq t$ implies that, at time t, at least one p_i value is zero. Standard probabilistic formulae for unions of events give

$$\mathrm{E}(T_2) = -2\left[\sum p_i \log p_i + \sum (1-p_i)\log(1-p_i)\right]. \qquad (325)$$

In the particular case $p_i = 1/3$ these formulae give

$$\mathrm{E}(T_1) \sim 3.2N \text{ generations}, \quad \mathrm{E}(T_2) \sim 1.6N \text{ generations}.$$

Littler (1975) shows that in the M-allele case,

$$\begin{aligned} \mathrm{E}(T_i) &= -2\Big(\sum_{s=1}^{i}(-1)^{i-s}\binom{M-1-s}{i-s}\Big(\sum(1-p_{i_1}-\ldots-p_{i_s}\Big) \\ &\quad \times \log(1-p_{i_1}-\ldots-p_{i_s})\Big)\Big) \qquad (326) \end{aligned}$$

where the inner sum is taken over all possible values $1 \leq i_1 < i_2 < \ldots < i_s \leq M$. This reduces to (324) when $i = 1$ and generalizes (325) to arbitrary M when $i = 2$. It is of some interest to note that if $p_i = M^{-1}$,

$$\lim_{M\to\infty} \mathrm{E}(T_j) = 2/j, \quad j = 1, 2, \ldots . \qquad (327)$$

These conclusion may be compared to the "eigenvalue" expression (116), and this comparison shows that the eigenvalues give a poor indication of the way in which $\mathrm{E}(T_j)$ changes as a function of j.

Inference procedures

Estimating θ

The notes above show that the parameter θ arises in many population genetics formulae, particularly those in the infinitely many alleles model. (As discussed by Dr. Joyce, it also arises frequently in the theory of the infinitely many sites model.) In this section we discuss the properties of various estimators of this parameter.

Estimating θ in the infinitely many alleles model

For the Wright–Fisher infinitely many alleles model, equations (143) and (145) show jointly that to a close approximation, the conditional distribution of the vector $\mathbf{A} = (A_1, A_2, \ldots, A_n)$ defined before equation (143), given the value of K_n, is

$$\text{Prob}\{\mathbf{A} = \mathbf{a} | K_n = k\} = \frac{n!}{|S_n^k| \; 1^{a_1} 2^{a_2} \; \ldots \; n^{a_n} \, a_1! a_2! \; \ldots \; a_n!} \,, \quad (328)$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$. This conditional distribution is exact for the Moran model, and we use it as the basis of the theory for estimating θ in infinitely many alleles models.

Equation (328) implies that K_n is a *sufficient statistic* for θ. Standard statistical theory then shows that, once the observed value k_n of K_n is given, no further information about θ is provided by the various a_j values, so that all inferences about θ should be carried out using the observed value k_n of K_n only. This includes estimation of θ or of any function of θ.

Since K_n is a sufficient statistic for θ we can use the probability distribution in equation (145) directly to find the maximum likelihood estimator $\hat{\theta}_K$ of θ. It is found that this estimator is the implicit solution of the equation

$$K_n = \frac{\hat{\theta}_K}{\hat{\theta}_K} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 1} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 2} + \cdots + \frac{\hat{\theta}_K}{\hat{\theta}_K + n - 1} \,. \quad (329)$$

Given the observed value k_n of K_n, the corresponding maximum likelihood estimate $\hat{\theta}_k$of θ is found by solving the equation

$$k_n = \frac{\hat{\theta}_k}{\hat{\theta}_k} + \frac{\hat{\theta}_k}{\hat{\theta}_k + 1} + \frac{\hat{\theta}_k}{\hat{\theta}_k + 2} + \cdots + \frac{\hat{\theta}_k}{\hat{\theta}_k + n - 1} \,. \quad (330)$$

Numerical calculation of the estimate $\hat{\theta}_k$ using (330) is usually necessary.

The estimator implied by (329) is biased, and it is easy to show that there can be no unbiased estimator of θ. On the other hand there exists an unbiased estimator of the population homozygosity probability $1/(1+\theta)$. If this estimator is denoted by $g(K_n)$, equation (145) shows that

$$\sum_{k=1}^{n} \frac{|S_n^k| \theta^k g(k)}{S_n(\theta)} = \frac{1}{1+\theta} \,,$$

where $|S_n^k|$ is the absolute value of a Stirling number, defined below equation (145). From this,

$$\sum_{k=1}^{n} |S_n^k| \theta^k g(k) = \theta(\theta+2)(\theta+3)\cdots(\theta+n-1)\,.$$

Since this is an identity for all θ, the expression for $g(k)$ for any observed value k_n of K_n can be found by comparing the coefficients of θ^k on both sides of this equation. In particular, when $k_n = 2$,

$$g(2) = \frac{\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}. \tag{331}$$

Unbiased estimation of $1/(1+\theta)$ for values of k_n larger than 2 is complicated, and it is then probably more convenient to use instead the estimator $(1+\hat{\theta}_K)^{-1}$, where $\hat{\theta}_K$ is found from (329), even though this estimator is slightly biased.

It is sometimes preferred to estimate $(1+\theta)^{-1}$ by f, defined in the notation of equation (149) by

$$f = \sum n_i^2/n^2. \tag{332}$$

This is a poor estimate in that it uses precisely that part of the data that is *least* informative about $(1+\theta)^{-1}$. The estimate of θ derived from f, namely

$$\hat{\theta}_f = f^{-1} - 1, \tag{333}$$

has been shown to be strongly biased and to have mean square error approximately six or eight times larger than that of $\hat{\theta}$.

More generally, the only functions of θ allowing unbiased estimation are linear combinations of functions of the form

$$\{(a+\theta)(b+\theta)\cdots(c+\theta)\}^{-1}\,, \tag{334}$$

where $a, b, \ldots c$ are integers with $1 \le a < b \ldots < c \le n-1$. While this fact derives mathematically from the form of the probability distribution (145), an argument in support of it, from an empirical sampling point of view, is as follows.

Suppose for example that $k_n = 2$ and write the unordered numbers of genes of the two alleles observed as N_1 and $n - N_1$. The

probability distribution of the pair $(N_1, n - N_1)$ is identical to that of N_1, and (328) shows that this is

$$\text{Prob}(N_1 = n_1) = \frac{n!}{|S_n^2| n_1 (n - n_1)} \,. \tag{335}$$

Given the observed values n_1 and $n - n_1$, the probability that two genes taken at random are of the same allelic type is

$$\frac{\binom{n_1}{2} + \binom{n-n_1}{2}}{\binom{n}{2}}.$$

Multiplying this expression by the right-hand side in (335) and summing over all possible values of n_1 gives the estimator (331). A similar argument can be used to justify the fact that any function of the form in (334) admits unbiased estimation.

We now consider an approximation for the mean square error (MSE) of the estimator $\hat{\theta}_K$ as defined by (329). Writing the right-hand side of (329) as $\psi(\hat{\theta}_K)$ we have $K_n = \psi(\hat{\theta}_K)$ and also, from (146), $\text{E}(K_n) = \psi(\theta)$. Thus by subtraction

$$K_n - \text{E}(K_n) = \psi(\hat{\theta}_K) - \psi(\theta).$$

A first-order Taylor series approximation for the right-hand side is $(\hat{\theta}_K - \theta)\psi'(\theta)$, so that

$$K_n - \text{E}(K_n) \approx (\hat{\theta}_K - \theta)\psi'(\theta).$$

Squaring and taking expectations, we get

$$\text{MSE}(\hat{\theta}_K) \approx \text{Var}(K_n)/\psi'(\theta)^2. \tag{336}$$

The variance of K_n is given in (147), and it is immediate that

$$\psi'(\theta) = \sum_{j=1}^{n-1} \frac{j}{(\theta + j)^2}. \tag{337}$$

This leads to

$$\text{MSE}(\hat{\theta}_K) \approx \theta / \sum_{j=1}^{n-1} \frac{j}{(j + \theta)^2}. \tag{338}$$

The approximation (338) can be shown to be quite accurate, and we use it later when comparing estimation of θ in the infinitely many alleles and the infinitely many sites models.

Estimators of θ in the infinitely many sites model

In this section we consider properties of two statistics that in the neutral case are both unbiased estimators of the parameter θ. As discussed above, the theory considered in this section concerns only the case of completely linked segregating sites.

The first unbiased estimator of θ that we consider is that based on the number S_n of segregating sites. Standard theory (as discussed by Dr Joyce) shows that the mean of S_n is given by

$$\theta \sum_{j=1}^{n-1} 1/j = g_1 \theta,$$

where

$$g_1 \ = \ \sum_{j=1}^{n-1} \frac{1}{j}. \tag{339}$$

We note for future reference that the variance of S_n is

$$\mathrm{var}(S_n) \ = \ g_1\theta + g_2\theta^2, \tag{340}$$

where

$$g_2 \ = \ \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{341}$$

Clearly an unbiased estimator of θ is

$$\hat{\theta}_S = \frac{S_n}{g_1}. \tag{342}$$

Equation (340) implies that the variance of $\hat{\theta}_S$ is

$$\mathrm{var}(\hat{\theta}_S) = \frac{\theta}{g_1} + \frac{g_2\theta^2}{g_1^2}\,. \tag{343}$$

The second unbiased estimator of θ is found as follows. Suppose that the nucleotide sequences i and j in the sample are compared and differ at some random number $T(i,j)$ of sites. Then $T(i,j)$ is an unbiased estimator of θ. It is natural to consider all $\binom{n}{2}$ possible comparisons of two nucleotide sequences in the sample and to form the statistic

$$T = \frac{\sum_{i<j} T(i,j)}{\binom{n}{2}}. \tag{344}$$

Since this is also an unbiased estimator of θ, we think of it as forming the unbiased estimator $\hat{\theta}_T$, defined by

$$\hat{\theta}_T = \frac{\sum_{i<j} T(i,j)}{\binom{n}{2}}. \tag{345}$$

This estimator of θ was proposed by Tajima (1983). It is a poor estimator of θ in that its variance, namely,

$$\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 = b_1\theta + b_2\theta^2, \tag{346}$$

does not approach 0 as the sample size n increases. (b_1 and b_2 are implicitly defined in this equation.) However, our interest here in this estimator is that it forms part of a hypothesis testing procedure, and not as a possible estimator of θ.

Comparison of the properties of the "alleles-based" and the "sites-based" estimators of θ

It is interesting to compare the estimates of θ based on k (the number of alleles observed in a sample) and s, (the number of segregating sites (SNPs) observed in the same sample). To do this we consider the five (very short) genes below.

T G T A T G C C T G C
T G T A T G C C T G C
T G T A T G C C C G C
T G T A T G C C T G C
T C T A T G C C T G C

Genes 1,2 and 4 are identical, while the remaining two genes are different from all the others. Thus k, the number of different alleles in the sample, is 3. Sites 2 and 9 are segregating, so that $s = 2$.

From (330), the estimate θ_k is the solution of the equation

$$3 = \frac{\hat{\theta}_k}{\hat{\theta}_k} + \frac{\hat{\theta}_k}{\hat{\theta}_k+1} + \frac{\hat{\theta}_k}{\hat{\theta}_k+2} + \frac{\hat{\theta}_k}{\hat{\theta}_k+3} + \frac{\hat{\theta}_k}{\hat{\theta}_k+4}\,, \tag{347}$$

and numerical methods give $\theta_k \approx 2.21$.

From (339) and (342), the estimate θ_s is

$$\frac{2}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} \approx 0.969.$$

θ	0.5	1.0	3.0	5.0
$n = 50$	0.902	0.874	0.891	0.928
$n = 100$	0.918	0.903	0.960	1.038
$n = 500$	0.943	0.942	1.047	1.178

Table 2: Values of the ratio $\mathrm{Var}(\hat{\theta}_S)/\mathrm{MSE}(\hat{\theta}_K)$ for selected values of n and θ.

It is interesting to compare the approximate (but accurate) the mean square error of the standard "alleles-based" estimator given in (338) to the variance (and also the mean square error) of the unbiased "sites-based" estimator given in (401). Table 2 compares these two for a variety of combinations of n and θ. In some cases the "alleles-based" estimator has the smaller mean square error, while for other cases the "sites-based" estimator has the smaller mean square error.

Testing neutrality

Introduction

Almost all the theory discussed so far assumes selective neutrality at the gene locus considered. In this section we consider the question: May we in fact reasonably assume selective neutrality at this gene locus?

The hypothesis of selective neutrality is more frequently called the "non-Darwinian" theory, and was promoted mainly by Kimura (1968). Under this theory it is claimed that, whereas the gene substitutions responsible for obviously adaptive and progressive phenomena are clearly selective, there exists a further class of gene substitutions, perhaps in number far exceeding those directed by selection, that have occurred purely by chance stochastic processes. A better name for the theory would thus be the "extra-Darwinian" theory, although here we adhere to the standard expression given above.

In a broader sense, the theory asserts that a large fraction of currently observed genetic variation between and within populations is nonselective. In this more extreme sense the theory has been described as the "neutral alleles" theory, although this term and the term "non-Darwinian" have been used interchangeably in the

literature and will be so used here.

This theory has, of course, been controversial, not only among theoreticians but also among practical geneticists, and the question whether certain specific substitutions have been neutral has been argued for decades. We do not refer here to the extensive literature on this matter.

In statistical terms the neutral theory is the "null hypothesis" to be tested, and all calculations given here assume that this null hypothesis is true. Most tests in the current literature relate to "infinitely many sites" data: here we consider both these tests and those tests that use "infinitely many alleles" data.

Tests of selection based on the infinitely many alleles model

The first objective tests of selective neutrality based on the infinitely many alleles model were put forward by Ewens (1972) and Watterson (1977). The broad aim of both tests was to assess whether the observed values $\{a_1, \ldots, a_n\}$ in (328) conform reasonably to what is expected under neutrality, that is, under the formula (328), given the sample size n and the observed number k of alleles in the sample. It is equivalent to use the observed numbers $\{n_1, \ldots, n_k\}$ defined in connection with (149) and to assess whether these conform reasonably to their conditional probability given n and k, namely,

$$\text{Prob}(n_1, n_2, \ldots, n_k | k) = \frac{n!}{|S_n^k| k! n_1 n_2 \cdots n_k}\,. \tag{348}$$

The Ewens and the Watterson testing procedures differ only in the test statistic employed, and here we discuss only the Watterson procedure. This uses as test statistic the observed sample homozygosity, defined as

$$f = \sum_{j=1}^{k} \frac{n_j^2}{n^2}. \tag{349}$$

The first aim is to establish what values of f will lead to rejection of the neutral hypothesis. Clearly, f will tend to be smaller under selection favoring heterozygotes than under neutrality, since this form of selection will tend to equalize allele frequencies compared to that expected for the neutral case, thus tending to decrease f. If we expect one high-frequency "superior" allele and a collection of

Species	n	k	n_1	n_2	n_3	n_4	n_5	n_6	n_7
willistoni	582	7	559	11	7	2	1	1	1
tropicalis	298	7	234	52	4	4	2	1	1
equinoxalis	376	5	361	5	4	3	3		
simulans	308	7	91	76	70	57	12	1	1

Table 3: *Drosophila* sample data

low-frequency deleterious alleles, f will tend to exceed its neutral theory value. Thus the hypothesis of neutrality is rejected if f is "too small" and also if f is "too large".

To determine how large or small f must be before neutrality is rejected, it is necessary to find its neutral theory probability distribution. This may be found in principle from (348). In practice, difficulties arise with the mathematical calculations because of the form of the distribution (348), and other procedures are needed.

For any observed data set $\{n_1, \ldots, n_k\}$, a computer-intensive exact approach proceeds by taking n and k as given, and summing the probabilities in (348) over all those n_1, $n_2, \ldots, n_k$ combinations that lead to a value of f more extreme than that determined by the data. This procedure is increasingly practicable with present-day computers, but will still be difficult in practice if an extremely large number of sample points is involved.

An approximate approach is to use a computer simulation to draw a large number of random samples from the distribution in (348): Efficient ways of doing this are given by Watterson (1978). If a sufficiently large number of such samples is drawn, a reliable empirical estimate can be made of various significance level points. This was done by Watterson (1978): see his Table 1.

The simulation method allows calculation of tables of $\mathrm{E}(f|k)$ and $\mathrm{var}(f|k)$ for various k and n values, which are of independent interest and are given (for the data of Table 3) in Table 4.

We illustrate this test of neutrality by applying it to particular data. The data concern numbers and frequencies of different alleles at the Esterase-2 locus in various *Drosophila* species and are quoted by Ewens (1974) and Watterson (1977).

For each set of data we compute f, the observed homozygosity. Then the exact neutral theory probability P (given in Table 4) that the homozygosity is more extreme than its observed value may be

Species	f	$\mathrm{E}(f)$	$\mathrm{var}(f)$	P	P_{sim}
willistoni	0.9230	0.4777	0.0295	0.007	0.009
tropicalis	0.6475	0.4434	0.0253	0.130	0.134
equinoxalis	0.9222	0.5654	0.0343	0.036	0.044
simulans	0.2356	0.4452	0.0255		0.044

Table 4: Sample statistics, means, variances, and probabilities for the data of Table 3.

calculated (except for the *D. simulans* case where the computations are prohibitive). The simulated probabilities P_{sim} are also given in Table 4; these are in reasonable agreement with the exact values. The conclusion that we draw is that significant evidence of selection appears to exist in all species except *D. tropicalis*.

We next outline two procedures based on the sample "frequency spectrum". Define A_i as the (random) number of alleles in the sample that are represented by exactly i genes. For given k and n, the mean value of A_i can be found directly from (328) as

$$\mathrm{E}\big(A_i|k,n\big) = \frac{n!}{i(n-i)!}\,\frac{|S_{k-1}^{n-i}|}{|S_k^n|}. \tag{350}$$

In this formula the S_j^i are values of Stirling numbers of the first kind as discussed after (145). The array of the $\mathrm{E}(A_i|k,n)$ values for $i = 1, 2, \ldots, n$ is the sample conditional mean frequency spectrum, and the corresponding array of observed values a_i is the observed conditional frequency spectrum. The first approach that we outline is an informal one, consisting of a simple visual comparison of the observed and the expected sample frequency spectra. Coyne (1976) provides an illustration of this approach. In Coyne's data, $n = 21$, $k = 10$, and

$$n_1 = n_2 = \cdots = n_9 = 1, \quad n_{10} = 12.$$

Direct use of (106) shows that given that $k = 10$ and $n = 21$,

$$\mathrm{E}\big(A_i \mid k = 10, n = 21\big) = \frac{21!}{i(21-i)!}\,\frac{|S_9^{21-i}|}{|S_{10}^{21}|}\,, \tag{351}$$

and this may be evaluated for $i = 1, 2, \ldots, 12$, the only possible values in this case. A comparison of the observed a_i values and

the expected values calculated from (351) is given in Table 5. It appears very difficult to maintain the neutral theory in the light of this comparison.

	i											
a_i	1	2	3	4	5	6	7	8	9	10	11	12
E	5.2	2.1	1.1	0.7	0.4	0.2	0.1	0.1	0.0	0.0	0.0	0.0
O	9	0	0	0	0	0	0	0	0	0.0	0.0	1

Table 5: Comparison of expected (E) and observed (O) sample frequency spectra.

A second approach provides a formal test of hypothesis, but focuses only on the number A_1 of singleton alleles in the sample. This procedure originally assumed selective neutrality and was used to test for a recent increase in the mutation rate. However, it may equally well be used as a test of neutrality itself if a constant mutation rate is assumed, especially for any test in which the alternative selective hypothesis of interest would lead to a large number of singleton alleles. The procedure may be generalized by using as test statistic the total number of singleton, doubleton, ..., j-ton alleles, leading to a test in which the selective alternative implies a significantly large number of low-frequency alleles. A parallel procedure, using the frequency of the most frequent allele in the data, may also be used.

We describe here only the test based on the number A_1 of singleton alleles. The total number k of alleles in the sample is taken as given, and the test is based on the neutral theory conditional distribution of A_1, given k and n. (It is assumed, as is always the case in practice, that n strictly exceeds k.) This conditional distribution is independent of θ and is found from (328) to be

$$\text{Prob}(A_1 = a|k,n) = \sum_{j=a}^{k-1} (-1)^{j-a} \frac{|S_{k-1}^n|}{a!(j-a)!|S_k^n|}. \tag{352}$$

Here S_i^j is again a Stirling number of the first kind. The conditional mean of A_1 is $|S_{k-1}^n|/|S_k^n|$, and the distribution (352) is approximately Poisson, with this mean. This observation enables a rapid approximate assessment of whether the number of singleton alleles is significantly large, assuming selective neutrality.

Tests based on the infinitely many sites model

Introduction

Dr Joyce has introduced the infinitely many sites model, and here we use results for that model which relate to testing the neutrality hypothesis. Since the complete nucleotide (i.e. DNA) sequences of genes are now available in large numbers, and since these data represent an ultimate state of knowledge of the gene, tests of neutrality based on infinitely many sites data are increasingly popular. Although several tests have been proposed that use infinitely many sites data, here we focus on what is by far the most popular of these, namely the Tajima (1989) test. The theory for this test is based on the Watterson (1975) infinitely many sites theory, which assumes complete linkage (that is, no recombination) between sites. It is therefore assumed throughout that the data at hand conform to this assumption. In practice this might mean that the DNA sequences in the data relate to a single gene.

As for tests using infinitely many alleles theory, discussed above, it is assumed in all the calculations in this section that selective neutrality holds, so that these can be thought of as "null hypothesis" calculations.

We assume a sample of n aligned sequences. The number S_n of sites segregating in the sample is not a sufficient statistic for the central parameter θ describing the stochastic behavior of the evolution of these sequences. Indeed, there is no simple nontrivial sufficient statistic for θ for this case. This implies that no direct analogue of the exact infinitely many alleles tests is possible.

On the other hand, in the infinitely many sites model there are several unbiased estimators of θ when neutrality holds, as discussed above. The basic idea of the Tajima test is to form a statistic whose numerator is the difference between two estimators that are unbiased under selective neutrality, and whose denominator is a neutral theory estimate of the standard deviation of this difference. Although under neutrality these two observed values of these estimators should tend to be close, since they are both unbiased estimators of the same quantity, under selection they should tend to differ, since the estimators on which they are based tend to differ under selection, and in predictable ways. Thus values of the statistic formed sufficiently far from zero lead to rejection of the neutrality hypothesis.

The details of the Tajima test, using this fact, are now discussed.

The Tajima test

The Tajima test in effect compares the values of $\hat{\theta}_T$ and $\hat{\theta}_S$, defined above. Specifically, the procedure is carried out in terms of the statistic D, defined by

$$D = \frac{\hat{\theta}_T - \hat{\theta}_S}{\sqrt{\hat{V}}}, \tag{353}$$

where $\hat{V}$ is an unbiased estimate of the variance of $\hat{\theta}_T - \hat{\theta}_S$ and is defined in (355) below. Tajima showed, by using adroit coalescent arguments, that the variance V of $\hat{\theta}_T - \hat{\theta}_S$ is

$$V = c_1\theta + c_2\theta^2, \tag{354}$$

where

$$c_1 = b_1 - \frac{1}{g_1}, \quad c_2 = b_2 - \frac{n+2}{g_1 n} + \frac{g_2}{g_1^2}.$$

Since this variance depends on θ, any estimate of this variance depends on a choice of an estimate of θ.

The variance of the estimator $\hat{\theta}_S$ decreases to 0 as the sample size increases (although the decrease is very slow), so the Tajima procedure is to estimate the variance of $\hat{\theta}_T - \hat{\theta}_S$ by the function of S that provides an unbiased estimator of the variance (354). Elementary statistical theory shows that this function is

$$\hat{V} = \frac{c_1 S}{g_1} + \frac{c_2 S(S-1)}{g_1^2 + g_2}. \tag{355}$$

This is then used in the D statistic given in (353) above.

The next task is to find the null hypothesis distribution of D. Although D is broadly similar in form to a z-score, it does not have a normal distribution under the null hypothesis of selective neutrality. Further, under this hypothesis, its mean is not zero and its variance is not 1, since the denominator of D involves a variance estimate rather than a known variance. Further, the distribution of D depends on the value of θ, which is in practice unknown. Thus there is no null hypothesis distribution of D invariant over all θ values, and in general little theoretical knowledge is available about the null hypothesis distribution of D.

The Tajima procedure approximates the null hypothesis distribution of D in the following way. First, the smallest value that D can take arises when there is a singleton nucleotide at each site segregating. In this case $\hat{\theta}_T$ is $2S_n/n$, and the numerator in D is then $\{(2/n) - (1/g_1)\}S_n$. In this case the value of D approaches a, defined by

$$a = \frac{\{(2/n) - (1/g_1)\}\sqrt{g_1^2 + g_2}}{\sqrt{c_2}}, \tag{356}$$

as the value of S_n approaches infinity.

The largest value that D can take arises when there are $n/2$ nucleotides of one type and $n/2$ nucleotides of another type at each site (for n even) or when there are $(n-1)/2$ nucleotides of one type and $(n+1)/2$ nucleotides of another type at each site (for n odd). In this case the value of D approaches b, defined by

$$b = \frac{\{(n/2(n-1)) - (1/a_1)\}\sqrt{g_1^2 + g_2}}{\sqrt{c_2}} \tag{357}$$

when n is even and the value of S_n approaches infinity. A similar formula applies when n is odd.

Second, it is assumed, as an approximation, that the mean of D is 0 and the variance of D is 1. Finally, it is also assumed that the density function of D is the generalized beta distribution over the range (a, b), defined by

$$f(D) = \frac{\Gamma(\alpha+\beta)(b-D)^{\alpha-1}(D-a)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)(b-a)^{\alpha+\beta-1}}, \tag{358}$$

with the parameters α and β chosen so that the mean of D is indeed 0 and the variance of D is indeed 1. This leads to the choice

$$\alpha = -\frac{(1+ab)b}{b-a}, \quad \beta = \frac{(1+ab)a}{b-a}.$$

This approximate null hypothesis distribution is then used to assess whether any observed value of D is significant.

The various approximations listed above have been examined in detail in the literature. It appears that the Tajima procedure is often fairly accurate, although examples can be found where this is not so. We do not pursue these matters here.

Complications

The tests of selective neutrality discussed above have all assumed that a stationary situation exists. However, stationarity typically takes a long time to be reached in genetic processes, so it is necessary to consider some time-dependent results and then to see how much these are relevant to tests of neutrality. The results we use are due to Griffiths (1979a, b).

Griffiths' calculations concern the number and frequencies of alleles observed in a sample of n genes. These of course depend on the initial population frequencies chosen as well as on the mutation rate. At one extreme one can assume that initially only one allelic type exists in the population and at the other extreme that $2N$ allelic Many of these properties are found using the time-dependent frequency spectrum $\phi_t(x)$, which has the form

$$\phi_t(x) = \theta x^{-1}(1-x)^{\theta-1}\big(1 + \sum_{i=2}^{\infty} \lambda_i(t)\psi_i(x,\theta))g_i(p_1, p_2, \ldots)\big) \quad (359)$$

where the $\lambda_i(t)$ are eigenvalues whose values are given below, $\psi(x,\theta)$ is a function of x only and θ and $g_i(p_1, p_2, , \ldots)$ is a complicated function of the initial allelic frequencies $p_1, p_2, \ldots$. The rate of convergence of this frequency spectrum to the stationary spectrum $\theta x^{-1}(1-x)^{\theta-1}$ depends on the eigenvalues $\lambda_j(t)$, which are given by

$$\lambda_i(t) = \exp\{-\frac{1}{2}j(j-1+\theta)t\}, \quad j = 2, 3, 4, \ldots, \quad (360)$$

and in particular on the largest eigenvalue $\exp\{-(1+\theta)t\}$. These eigenvalues are the limiting case of the discrete configuration values given in (123) as $N \to \infty, u \to 0$, with $4Nu = \theta$ held fixed.

The mean number of alleles in a sample of n genes can be found, following the same argument as that leading to (150), by evaluation of

$$\int_0^1 \{1 - (1-x)^n\}\phi_t(x)\, dx. \quad (361)$$

An explicit expression for this mean is given by Griffiths (1979b, equation (2.10)), who also provides numerical calculations for various r, θ, t, and p_j values. We reproduce some representative calculations in Table 6 for two cases, first where there exists initially a

		t							
		0.2		0.5		1.0		∞	
		(i)	(ii)	(i)	(ii)	(i)	(ii)	(i) &	(ii)
	0.1	1.31	10.12	1.40	4.62	1.47	2.77	1.57	
θ	1.0	4.03	12.39	4.89	7.64	5.49	6.34	5.88	
	1.5	5.51	13.62	6.74	9.25	7.54	8.18	7.90	

Table 6: Mean number of alleles observed in a sample of 200 genes for various θ, t values. Unit time = $2N$ generations. Case (i): one initial allele. Case (ii): many initial alleles of equal frequency. From Griffiths (1979b)

single allele in the population and second where there exist initially many alleles of equal frequency. We observe that in the former case the approach to the equilibrium point appears rather more rapid than in the latter. Griffiths also found properties of two samples, one in each of two sub-populations, which split apart some time in the past. In particular he gives formulae the mean number of alleles common to the two samples at time t after the split and the joint probability distributions of the sample frequencies of these alleles.

As has been seen above, tests of selective neutrality often reduce to a comparison of properties of the number of alleles, or of segregating sites, in a sample to some measure of population homozygosity (or, equivalently, heterozygosity). Unfortunately, the properties of the two measures under selection are often similar to their properties in a selectively neutral case where the population has recently expanded in size after going through a bottleneck, or at the end of a selectively induced replacement process at a locus closely linked to the neutral locus. Thus these tests of selection can be rendered invalid at times closely following such historical events. Table 6 can be used to find various properties of the number of alleles a sample following a bottleneck or a selective sweep, since we might assume, to a close approximation, that only one allele survives a tight bottleneck or a selective sweep. Table 6 then shows, for example, that when $\theta = 1$ the mean number of alleles in a sample of 200 genes is 4.89 when N generations have passed after the bottleneck or selective sweep, about 83% of its stationary mean value of 5.88.

The properties of the sample homozygosity should be close to those of the population homozygosity. We take take 0 to be the time of the bottleneck and the population homozygosity at this time to be 1. Denoting the mean homozygosity at time t diffusion time

units by $F^{(t)}$, equation (122) shows that

$$F^{(t)} = \frac{1}{1+\theta} + \frac{\theta}{1+\theta}\exp^{-(1+\theta)t}. \tag{362}$$

Thus $F^{(t)}$ depends only on the leading eigenvalue in the set (360) whereas the mean number of alleles depends on all the eigenvalues. When $\theta = 1$ the value of $F^{(t)}$ arising N generations after the bottleneck is 0.684, so that the mean heterozygosity at this time is 0.316. This is about 63% of its stationary value. The comparison of this with the corresponding value for the mean number of alleles in the sample is then relevant to the effect of a bottleneck on a test for selective neutrality conducted N generations after the bottleneck or the selective sweep.

References

References for all the papers referred to in the notes, together with some other useful references, are given here.

Abramowitz, M., Stegun, I.A.: *Handbook of Mathematical Functions.* New York: Dover Publ. Inc., 1965.

Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach 1. Haploid models. *Adv. Appl. Prob.* **6**, 260–290 (1974).

Coyne, J.A.: Lack of genetic similarity between two sibling species of *Drosophila* as revealed by varied techniques. *Genetics* **84**, 593–607 (1976).

Donnelly, P.J.: Partition structure, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theoret. Pop. Biol.* **30**, 271–288 (1986).

Donnelly, P.J., Tavaré, S.: The ages of alleles and a coalescent. *Adv. Appl. Prob.* **18**, 1–19 (1986).

Donnelly, P.J., Tavaré, S.: Coalescents and genealogical structure under neutrality. In: *Annual Review of Genetics*, Campbell, A., Anderson, W., Jones, E. (eds.), pp 401–421. Palo Alto, Annual Reviews Inc., (1995).

Engen, S.: A note on the geometric series as a species frequency model. *Biometrika* **62**, 694–699 (1975).

Ewens, W.J.: The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3**, 87–112 (1972).

Ewens, W.J.: Mathematical and statistical problems arising in the non-Darwinian theory. *Lectures on Mathematics in the Life Sciences* **7**, 25–42 (1974).

Ewens, W.J.: *Mathematical Population Genetics*, Second edition, Springer NY, (2004).

Feller, W.: Diffusion processes in genetics. In *Proc. 2nd Berkeley Symp. on Math. Stat. and Prob.* Neyman, J. (ed.), pp. 227–246. Berkeley: University of California Press, (1951).

Fisher, R. A.: On the dominance ratio. *Proc. Roy. Soc. Edin.* **42**, 321–341 (1922).

Fisher, R. A.: *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press, (1930a).

Fisher, R. A.: *The Genetical Theory of Natural Selection* (second revised edit.). New York: Dover, (1958).

Fu, Y.-X., Li, W.-H.: Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).

Griffiths, R. C.: Exact sampling distributions from the infinite neutral alleles model. *Adv. Appl. Prob.* **11**, 326–354, (1979a).

Griffiths, R. C.: A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Prob.* **11**, 310–325, (1979b).

Griffiths, R.C. unpublished notes, (1980).

Griffiths, R.C., Tavaré, S.: Ancestral inference in population genetics. *Statist. Sci* **9**, 307–319, (1994).

Griffiths, R.C., Tavaré, S.: Unrooted tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127**, 77–98, (1995).

Griffiths, R.C., Tavaré, S.: Computational methods for the coalescent. *IMA Vol. Math. Applic.* **87**, 165–182, (1997).

Griffiths, R.C., Tavaré, S.: The age of a mutation in a general coalescent tree. *Stochastic Models* **14** 273–298 (1998).

Griffiths, R.C., Tavaré, S.: The ages of mutations in gene trees. *Ann. Appl. Probab.* **9**, 567–590, (1999).

Griffiths, R.C., Tavaré, S.: The genealogy of a neutral mutation. In *Highly Structured Stochastic Systems*. Green, P., Hjort, N., Richardson, S. (eds.), 393–412 (2003).

Harris, H: *The Principles of Human Biochemical Genetics* (third revised edition). Amsterdam: Elsevier, (1980).

Hoppe, F.: Polya-like urns and the Ewens sampling formula. *J.Math. Biol.* **20**, 91–99 (1984).

Hoppe, F.: Size-biased sampling of Poisson–Dirichlet samples with an application to partition structures in population genetics. *J. Appl. Prob.* **23**, 1008–1012 (1986).

Hoppe, F.: The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25**, 123–159 (1987).

Karlin, S., McGregor, J.: Direct product branching processes and related induced Markoff chains. I. Calculations of rates of approach to homozygosity. In: *Bernoulli (1723), Bayes (1773), Laplace (1813)*: Anniv. Vol., LeCam, L., Neyman, J., (eds.), pp. 111–145. Berlin, Heidelberg, New York: Springer, 1965.

Karlin, S., McGregor, J.L.: Addendum to a paper of W. Ewens. *Theoret. Pop. Biol.* **3**, 113–116 (1972).

Kelly, F.P.: On stochastic population models in genetics. *J. Appl. Prob.* **13**, 127–131 (1976).

Kelly, F.P.: Exact results for the Moran neutral allele model. *J. Appl. Prob.* **9**, 197–201 (1977).

Kimura, M.: Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci.* **41**, 144–150 (1955a).

Kimura, M.: Random drift in a multi-allelic locus. *Evolution* **9**, 419–435 (1955b).

Kimura, M.: Stochastic processes and distribution of gene frequencies under natural selections. *Cold Spring Harbor on Quant. Biol.* **20**, 33–53 (1955c).

Kimura, M.: Random genetic drift in a tri-allelic locus — exact solution with a continuous model. *Biometrics* **12**, 57–66 (1956a).

Kimura, M.: A model of a genetic system which leads to closer linkage under natural selection. *Evolution* **10**, 278–287 (1956b).

Kimura, M.: Some problems of stochastic processes in genetics. *Ann. Math. Stat.* **28**, 882–901 (1957).

Kimura, M.: Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

Kimura, M.: The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics* **61**, 893 (1969).

Kimura, M.: Theoretical foundations of population genetics at the molecular level. *Theoret. Pop. Biol.* **2**, 174–208 (1971).

Kimura, M., Crow, J. F.: The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738 (1964).

Kingman, J. F. C.: Random discrete distributions. *J. Roy. Stat. Soc. B.* **37**, 1–22 (1975).

Kingman, J. F. C.: Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A* **361**, 1–20 (1978).

Kingman, J.F.C.: The coalescent. *Stoch. proc. Applns.* **13**, 235–248, (1982).

Littler, R. A.: Loss of variability at one locus in a finite population. *Math. Bio.* **25**, 151–163 (1975).

McCloskey, J.W.: A model for the distribution of individuals by species in an environment. Unpublished PhD. thesis, Michigan State University, (1965).

Moran, P. A. P.: Random processes in genetics. *Proc. Camb. Phil. Soc.* **54**, 60–71 (1958).

Moran, P. A. P., Watterson, G. A.: The genetic effects of family structure in natural populations. *Aust. J. Biol. Sci.* **12**, 1–15 (1958).

Moran, P.A.P.: *The Statistical Processes of Evolutionary Theory*, Oxford: Clarendon Press, (1962).

Tajima, F.: Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).

Tajima, F.: Statistical methods for testing the neutral mutations hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

Tavaré, S.: Lines of descent and genealogical processes, and their application in population genetics models. *Theoret. Pop. Biol.* **26**, 119–164, (1984).

Tavaré, S.: Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–86 (1986).

Tavaré, S.: The age of a mutant in a general coalescent tree. *Stoch. Models* **14**, 273–295 (1998).

Tavaré, S.: Ancestral inference in population genetics. In *Proceedings of Saint Flour Summer School in Probability and Statistics*, (2004).

Watterson, G.A.: On the number of segregating sites in genetic models without recombination. *Theoret. Pop. Biol.* **7**, 256–276 (1975).

Watterson, G. A.: Reversibility and the age of an allele I. Moran's infinitely many neutral alleles model. *Theoret. Pop. Biol.* **10**, 239–253 (1976).

Watterson, G.A.: Heterosis or neutrality? *Genetics* **85**, 789–814 (1977).

Watterson, G.A.: The homozygosity test of neutrality. *Genetics* **88**, 405–417 (1978).

Watterson, G. A.: Lines of descent and the coalescent. *Theoret. Pop. Biol.* **26**, 239–253 (1984).

Watterson, G. A.: Guess, H. A.: Is the most frequent allele the oldest? *Theoret. Pop. Biol.* **11**, 141–160 (1977).

Wright, S.: Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).

Mathematical Population Genetics

Introduction to the Retrospective View of Population Genetics Theory

Lecture Notes

Guanajuato, March 2009

Paul Joyce

Age-ordered alleles: frequencies and ages

Introduction

The current direction of interest in population genetics is a retrospective one, looking backwards to the past rather than looking forward into the future. This change of direction is largely spurred by the large volume of genetic data now available at the molecular level and a wish to infer the forces that led to the data observed. This data driven modern perspective will be the focus of the notes that follow.

The material in this section should provide a useful transition from the prospective view to the retrospective view. The age of an allele refers to the past yet many of the results on ages of alleles are derived using the prospective machinery developed in the lectures of Dr. Ewens.

The material in this section covers both sample and population formulae relating to the infinitely many alleles model. Some results are diffusion approximations, and for them the definition of θ depends on the population model implicitly discussed. Various formulae for the Moran model are exact.

Frequencies

We first discuss allelic frequencies, for which finding "age" properties amounts to finding size-biased properties. Kingman's (1975) Poisson–Dirichlet distribution, which arises in various allelic frequency calculations, is not user-friendly. This makes it all the more interesting that a *size-biased* distribution closely related to it, namely the GEM distribution, named for Griffiths, (1980), Engen (1975) and McCloskey (1965), who established its salient properties, is both simple and elegant. More important, it has a central interpretation with respect to the ages of the alleles in a population. We now describe this distribution.

Suppose that a gene is taken at random from the population. The probability that this gene will be of an allelic type whose frequency in the population is x is just x. In other words, alleles are sampled by this choice in a size-biased way. The frequency spectrum (127) shows the probability that there exists an allele in the population with frequency between x and $x+\delta x$. It follows that the probability

the gene chosen is of this allelic type is $\theta x^{-1}(1-x)^{\theta-1}x\delta x = \theta(1-x)^{\theta-1}\delta x$. From this, the density function $f(x)$ of the frequency of this allele is given by

$$f(x) = \theta(1-x)^{\theta-1}. \tag{363}$$

Suppose now that all genes of the allelic type just chosen are removed from the population. A second gene is now drawn at random from the population and its allelic type observed. The frequency of the allelic type of this gene among the genes remaining at this stage can be shown to also be given by (363). All genes of this second allelic type are now also removed from the population. A third gene then drawn at random from the genes remaining, its allelic type observed, and all genes of this (third) allelic type removed from the population. This process is continued indefinitely. At any stage, the distribution of the frequency of the allelic type of any gene just drawn among the genes left when the draw takes place is given by (363). This leads to the following representation. Denote by w_j the original population frequency of the jth allelic type drawn. Then we can write $w_1 = x_1$, and for $j = 2, 3, \ldots$,

$$w_j = (1-x_1)(1-x_2)\cdots(1-x_{j-1})x_j, \tag{364}$$

where the x_j are independent random variables, each having the distribution (363). The random vector $(w_1, w_2, \ldots)$ then has the GEM distribution.

All the alleles in the population at any time eventually leave the population, through the joints processes of mutation and random drift, and any allele with current population frequency x survives the longest with probability x. That is, since the GEM distribution was found according to a size-biased process, it also arises when alleles are labeled according to the length of their future persistence in the population. Reversibility arguments then show that the GEM distribution also applies when the alleles in the population are labeled by their age. In other words, the vector $(w_1, w_2, \ldots)$ can be thought of as the vector of allelic frequencies when alleles are ordered with respect to their ages in the population (with allele 1 being the oldest).

The elegance of many age-ordered formulae derives directly from the simplicity and tractability of the GEM distribution. We now give two examples. First, the GEM distribution shows immediately that

the mean population frequency of the oldest allele in the population is

$$\theta \int_0^1 x(1-x)^{\theta-1} = 1/(1+\theta), \tag{365}$$

and more generally that the mean population frequency of the jth oldest allele in the population is

$$\frac{1}{1+\theta}\Big(\frac{\theta}{1+\theta}\Big)^{j-1}.$$

Second, the probability that a gene drawn at random from the population is of the type of the oldest allele is the mean frequency of the oldest allele, namely $1/(1+\theta)$, as just shown. More generally the probability that n genes drawn at random from the population are all of the type of the oldest allele is

$$\theta \int_0^1 x^n(1-x)^{\theta-1}\,dx = \frac{n!}{(1+\theta)(2+\theta)\cdots(n+\theta)}.$$

The probability that n genes drawn at random from the population are all of the same unspecified allelic type is

$$\theta \int_0^1 x^{n-1}(1-x)^{\theta-1}\,dx = \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n+\theta-1)},$$

in agreement with (148). From this, given that n genes drawn at random are all of the same allelic type, the probability that they are all of the allelic type of the oldest allele is $n/(n+\theta)$.

The GEM distribution is of course a diffusion approximation and the above results are diffusion approximations. The distribution has a number of interesting mathematical properties. It is invariant under size-biased sampling, and this property has been used by Hoppe (1987) to derive the frequency spectrum (127). It also has important properties with respect to the concepts of random deletions and non-interference, which were also exploited by Hoppe (1986).

It will be expected that various exact results hold for the Moran model, with θ defined as $2Nu/(1-u)$. The first of these is an exact representation for of the GEM distribution, analogous to (364). This has been provided by Hoppe (1987). Denote by $N_1, N_2, \ldots$ the numbers of genes of the oldest, second-oldest, ... alleles in the population. Then $N_1, N_2, \ldots$ can be defined in turn by

$$N_i = 1 + M_i, \quad i = 1, 2, \ldots, \tag{366}$$

where M_i has a binomial distribution with index $2N - N_1 - N_2 - \cdots - N_{i-1} - 1$ and parameter x_i, where $x_1, x_2, \ldots$ are independently and identically continuous random variables each having the density function (363). Eventually $N_1 + N_2 + ...N_k = 2N$ and the process stops, the final index k being identical to the number K_{2N} of alleles in the population.

It follows directly from this representation that the mean of N_1 is

$$1 + (2N-1)\theta \int_0^1 x(1-x)^{\theta-1}\, dx \;=\; \frac{2N+\theta}{1+\theta}.$$

The mean of the proportion $N_1/(2N)$ is $1/\{1 + (2N-1)u\}$, and is very close to the diffusion approximation $1/\{1+\theta\}$.

If there is only one allele in the population, so that the population is monomorphic, this allele must be the oldest one in the population. The above representation shows that the probability that the oldest allele arises $2N$ times in the population is

$$\text{Prob}\,(M_1 = 2N-1) = \theta \int_0^1 x^{2N-1}(1-x)^{\theta-1}\, dx,$$

and this reduces to the monomorphism probability (153).

More generally, Kelly (1977) has shown that for the Moran model, the probability that the oldest allele is represented by j genes in the sample is given exactly by

$$\frac{\theta}{2N}\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}. \tag{367}$$

The case $j = 2N$ considered above is a particular example of (367), and the mean number $(2N+\theta)/(1+\theta)$ follows from (367).

We now turn again to approximations deriving from diffusion methods. A question of some interest is to find the probability that the oldest allele in the population is also the most frequent. By time reversibility arguments (not discussed in detail here) this is also the probability that the most frequent allele in the population will survive the longest into the future, and in turn this is the mean of the frequency of the most frequent allele. Unfortunately, the distribution of the frequency of the most frequent allele is very complicated, taking different functional forms in the intervals $(1, 1/2), (1/2, 1/3), (1/3, 1/4), \ldots$. However, the frequency spectrum

allows one immediate calculation. If there is an allele in the population with frequency in $1/2, 1)$, it must be the most frequent allele. Thus the frequency spectrum (127) shows that the density function of the most frequent allele is $\theta x^{-1}(1-x)^{\theta-1}$ in the interval $(1/2, 1)$. Thus a lower bound for the mean frequency of the most frequent allele is

$$\int_{1/2}^{1} x[\theta x^{-1}(1-x)^{\theta-1}]dx = (1/2)^{\theta},$$

which is useful for small θ but not of much value for larger θ. Numerical calculations are given by Watterson and Guess (1977) for a range of θ values, who provide also the upper bound $1-\theta(1-\theta)\log 2$. For example, when $\theta = 1$ the mean frequency of the most frequent allele is 0.624, which may be compared with the mean frequency of the oldest allele (which must be less than the mean frequency of the most frequent allele) of 0.5.

Ages

We now turn to "age" questions. Some for these follow immediately from our previous calculations. For example, the mean time for all alleles existing in the population at any time to leave the population is given in (130), and again by reversibility arguments this is the mean time, into the past, that the oldest of these originally arose by mutation. This is is then the mean age of the oldest allele in the population, given on a "generations" basis. Since we refer to this calculation with reference to the mean age of the oldest allele in the population we repeat it here, with this new interpretation:

$$\text{Mean age of oldest allele } = \sum_{j=1}^{2N} \frac{4N}{j(j+\theta-1)} \text{ generations.} \quad (368)$$

In the case $\theta = 2$ this mean age is very close to $4N - 2$, that is to the conditional mean fixation time (295). The exact result corresponding to (368) for the Moran model is given in (169), or equivalently in (166), and is almost exactly $4N^2$ birth-death events when $\theta = 2Nu/(1-u) = 2$. This is close to the conditional mean fixation time given in (101), and the reason for these identities is discussed below equation (130).

In employing the argument leading to (368) we in effect use a result of Watterson and Guess (1977) and Kelly (1977), stating that

not only the mean age of the oldest allele, but indeed the entire probability distribution of its age, is independent of its current frequency and indeed of the frequency of all alleles in the population.

We next ask: "If an allele is observed in the population with frequency p, what is its mean age?" By reversibility, this is the mean time $\bar{t}(p)$ that it persists in the population, and in the diffusion approximation to the Wright–Fisher model this is found immediately from (68) as

$$4N \sum_{j=1}^{\infty} \{j(j+\theta-1)\}^{-1} \Big(1-(1-p)^j\Big). \tag{369}$$

This is clearly a generalization of the expression in (73), to which it reduces when $p = 1$, since if $p = 1$ only one allele arises in the population, and it must then be the oldest allele. A parallel exact calculation for the Moran model follows from the mean persistence time found eventually using (105).

A question whose answer follows from the above calculation is the following: "If a gene is taken at random from the population, what is the diffusion approximation for the mean age of its allelic type?" Changing notation, the density function of the frequency p of the allelic type of the randomly chosen gene is, from (363), $f(p) = \theta(1-p)^{\theta-1}$. The mean age $\bar{t}(p)$ of an allele with frequency p is, by reversibility, given by (68). The required probability is

$$\theta \int_0^1 \bar{t}(p)(1-p)^{\theta-1}\, dp, \tag{370}$$

and use of (68) for $\bar{t}(p)$ shows that this reduces to $2/\theta$ diffusion time units, or for the Wright–Fisher model, $1/u$ generations. This conclusion may also be derived by looking backwards to the past and using the coalescent. However, we shall not derive it this way since it is an immediately result. Looking backwards to the past, the probability that the original mutation creating the allelic type of the gene in question occurred j generations in the past is clearly $u(1-u)^{j-1}$, $(j = 1, 2, \ldots)$, and the mean of this (geometric) distribution is $1/u$.

An exact calculation parallel to this is possible for the Moran model, using the exact frequency spectrum (156) and the exact mean age deriving from (105). However a direct argument parallel to that

just given for the Wright–Fisher model shows that the exact mean time, measured in birth-death events, is $2N/u$.

We turn now to sample properties, which are in practice more important than population properties. The most important sample distribution concerns the frequencies of the alleles in the sample when ordered by age. This distribution was found by Donnelly and Tavaré (1986), who showed that the probability that the number K_n of alleles in the sample takes the value k, and that the age-ordered numbers of these alleles in the sample are, in age order, $n_{(1)}, n_{(2)}, \ldots, n_{(k)}$, is

$$\frac{\theta^k (n-1)!}{S_n(\theta) n_{(k)} (n_{(k)} + n_{(k-1)}) \cdots (n_{(k)} + n_{(k-1)} + \cdots n_{(2)})}, \tag{371}$$

where $S_n(\theta)$ is defined below equation (143). This formula can be found in several ways, one being as the size-biased version of Equation (149). The expression (371) is exact for the Moran model with θ defined as $2Nu/(1-u)$.

Several results concerning the oldest allele in the sample can be found from this formula, or in some cases more directly by other methods. For example, the probability that the oldest allele in the sample is represented by j genes in the sample is (Kelly, (1976))

$$\frac{\theta}{n} \binom{n}{j} \binom{n+\theta-1}{j}^{-1}. \tag{372}$$

This is identical to the expression (367) if $2N$ is replaced by n in the latter.

Further results provide connections between the oldest allele in the sample to the oldest allele in the population. Some of these results are exact for a Moran model and others are the corresponding diffusion approximations. For example, Kelly (1976) showed that in the Moran model, the probability that the oldest allele in the population is observed at all in the sample is $n(2N+\theta)/[2N(n+\theta)]$. This is equal to 1, as it must be, when $n = 2N$, and for the case $n = 1$ reduces to a result found above that a randomly selected gene is of the oldest allelic type in the population. The diffusion approximation to this probability, found by letting $N \to \infty$, is $n/(n+\theta)$.

A further result is that in the Moran model, the probability that a gene seen j times in the sample is of the oldest allelic type in the population is $j(2N+\theta)/[2N(n+\theta)]$. Letting $N \to \infty$, the diffusion

approximation for this probability is $j/(n+\theta)$. When $n = j$ this is $j/(j+\theta)$, a result found above found by other methods.

Donnelly (1986)) provides further formulae extending these. He showed, for example, that the probability that the oldest allele in the population is observed j times in the sample is

$$\frac{\theta}{n+\theta}\binom{n}{j}\binom{n+\theta-1}{j}^{-1}, \quad j = 0, 1, 2, \ldots, n. \tag{373}$$

This is of course closely connected to the Kelly result (372). For the case $j = 0$ this probability is $\theta/(n+\theta)$, confirming the complementary probability $n/(n+\theta)$ found above. Conditional on the event that the oldest allele in the population does appear in the sample, a straightforward calculation using (373) shows that this conditional probability and that in (372) are identical.

Griffiths and Tavaré (1998) give the Laplace transform of the distribution of that age of an allele observed b times in a sample of n genes, together with a limiting Laplace transform for the case when θ approaches 0. These results show, for the Wright–Fisher model, that the diffusion approximation for the mean age of such an allele is

$$\sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}\Big(1 - \frac{(n-1-b+\theta-1)_{(j)}}{(n-1+\theta-1)_{(j)}}\Big) \tag{374}$$

generations, where $a_{(j)}$ is defined as $a_{(j)} = a(a+1)\cdots(a+j-1)$. This is the sample analogue of the population expression in (369), and converges to (369) as $n \to \infty$ with $b = np$.

In the particular case $\theta = 2$, which we have considered several times above, the expression in (374) simplifies to

$$\frac{4Nb}{n-b}\sum_{j=b+1}^{n} j^{-1}. \tag{375}$$

Under the limiting process $n \to \infty$ with $b = np$ this approaches the expression in (74). This is as expected, since when $\theta = 2$, (74) is by reversibility arguments also the mean age of an allele observed with frequency p in the population.

Our final calculation concerns the mean age of the oldest allele in the sample. For the Wright–Fisher model the diffusion approxi-

mation for this mean age is

$$4N \sum_{j=1}^{n} \frac{1}{j(j+\theta-1)}. \tag{376}$$

For the case $n = 2N$ this is the value given in (130) and for the case $n = 1$ it reduces to the value $1/u$ given above. The corresponding exact result for the Moran model is

$$2N(2N+\theta) \sum_{j=1}^{n} \frac{1}{j(j+\theta-1)} \tag{377}$$

birth-death events, with (of course) θ defined as $2Nu/(1-u)$. When $n = 1$ this reduces to the calculation $2N/u$ given above. When $n = 2N$ it is identical to (166) and, less obviously, to the expression given in (169).

The expression in (376) may be written equivalently as

$$\sum_{j=1}^{n} \frac{1}{v_j + w_j}, \tag{378}$$

where

$$v_j = \frac{ju}{2N}, \quad w_j = \frac{j(j-1)(1-u)}{(2N)^2}. \tag{379}$$

These expressions follow the pattern of (167) and (168).

The coalescent

Introduction

In 1982 John Kingman, inspired by his friend Warren Ewens, took to heart the advice of Danish philosopher Soren Kierkegaard and realized that "Life can only be understood backwards, but it must be lived forwards." Applying this perspective to the world of population genetics led him to the development of the *coalescent*, a mathematical model for the evolution of a sample of individuals drawn from a larger population. The coalescent has come to play a fundamental role in our understanding of population genetics and has been at the heart of a variety of widely-employed analysis methods. For this it also owes a large debt to Richard Hudson, who arguably wrote

the first paper about the coalescent that the non-specialist could easily understand. Here we introduce the coalescent, summarize its implications, and survey its applications.

The central intuition of the coalescent is driven by parallels with pedigree-based designs. In those studies, the shared ancestries of the sample members, as described by the pedigree, are used to inform any subsequent analysis, thereby increasing the power of that analysis. The coalescent takes this a step further by making the observation that there is no such thing as unrelated individuals. We are all related to some degree or other. In a pedigree the relationship is made explicit. In a population-based study the relationships are still present, albeit more distant, but the details of the pedigree are unknown. However, it remains the case that analyses of such data are likely to benefit from the presence of a model that describes those relationships. The coalescent *is* that model.

Motivating problem

Human evolution and the infinitely- many-sites model

One of the signature early applications of the coalescent was to inference regarding the early history of humans. Several of the earliest data-sets consisted of short regions of mitochondrial DNA [mtDNA] or Y chromosome. Since mtDNA is maternally inherited it is perfectly described by the original version of the coalescent, with its reliance upon the existence of a single parent for each individual and its recombination-free nature. To motivate what follows, here we consider one of those early data sets.

The data in the following example comes from Ward *et. al.* (1991). The data analysis and mathematical modeling comes from a paper by Griffiths and Tavaré (1994).

Mitochondria DNA (mtDNA) comprises only about 0.00006% of the total human genome, but the contribution of mtDNA to our understanding of human evolution far outweighs its minuscule contribution to our genome. Human mitochondrial DNA, first sequenced by Anderson *et.al.* (1981), is a circular double-stranded molecule about 16,500 base pairs in length, containing genes that code for 13 proteins, 22 tRNA genes and 2 rRNA genes. Mitochondria live outside the nucleus of cells. One part of the molecule, the control

region (sometimes referred to as the D-loop), has received particular attention. The region is about 1,100 base pairs in length.

As the mitochondrial molecule evolves, mutations result in the substitution of one of the bases A,C,G or T in the DNA sequence by another one. Transversions, those changes between purines (A,G) and pyrimidines (C,T), are less frequent than transitions, the changes that occur between purines or between pyrimidines.

It is known that base substitutions accumulate extremely rapidly in mitochondrial DNA, occurring at about 10 times the rate of substitutions in nuclear genes. The control region has an even higher rate, perhaps on order of magnitude higher again. This high mutation rate makes the control region a useful molecule with which to study DNA variation over relatively short time spans, because sequence differences will be found among closely related individuals. In addition, mammalian mitochondria are almost exclusively maternally inherited, which makes these molecules ideal for studying the maternal lineages in which they arise. This simple mode of inheritance means that recombination is essentially absent, making inferences about molecular history somewhat simpler than in the case of nuclear genes.

In this example, we focus on mitochondrial data sampled from a single North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island. Based on the archaeological records (cf. Dewhirst, 1978), it is clear that there is a remarkable cultural continuity from earliest levels of occupation to the latest. This implies not only that there was no significant immigration into the area by other groups, but that subsistence pattern and presumably the demographic size of the population has also remained roughly constant for at least 8,000 years. Based on the current size of the population that was sampled, there are approximately 600 women of child bearing age in the traditional Nuu-Chah-Nulth population.

The original data, appearing in Ward *et. al.* (1991) comprised a sample of mt DNA sequences from 63 individuals. The sample approximated a random sample of individuals in the tribe, to the extent to which this can be experimentally arranged. Each sequence is the first 360 basepair segment of the control region. The region comprises 201 pyrimidine sites and 159 purine sites; 21 of the pyrimidine sites are variable (or segregating), that is, not identical in all 63 sequences in the sample. In contrast, only if 5 of the purine sites

	1	1	2	2	3			1	1	1	1	1	2	2	2	2	3	3	
Position	0	9	5	9	4	8	9	2	4	6	6	9	3	6	7	7	1	3	
	6	0	1	6	4	8	1	4	9	2	6	4	3	7	1	5	9	9	
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Lineage Frequencies
Lineage																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	6
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Table 7: North American Indian tribe from Vancouver Island. The data consist of 55 individuals 352 sites. There are 18 segregating sites in all; 13 of the sites are pyrimidines, and 5 purines.

are variable. There are 28 distinct DNA sequences (hereafter called lineages) in the data. Because, no transversions are seen in these data each DNA site is binary, having just two possible bases at each site.

To keep the presentation simple, we focus on one part of the data that seems to have a relatively simple mutation structure. We shall assume that substitutions at any nucleotide position can occur only once in the ancestry of the molecule. This is called **the infinitely-many-sites assumption.** Hence we have eliminated lineages in which substitutions are observed to have occurred more than once. The resulting subsample comprises 55 of the original 63 sequences, and 352 of the original 360 sites. Eight of the pyrimidine segregating sites were removed resulting in a set of 18 segregating sites in all; 13 of these sites are pyrimidines, and 5 are purines. These data are given in Table 7, subdivided into sites containing purines and pyrimidines. Each row of the table represents a distinct DNA

sequence, and the frequency of these lineages are given in the right most column of the table.

What structure do these sites have? Because of the infinitely-many-sites assumption, the pattern of segregating sites tells us something about the mutations that have occurred in the history of the sample. Next we consider an ancestral process that could have given rise the observed pattern of variability. This is called the coalescent.

The coalescent process, which we will discuss in some detail as the course goes on, is a way to describe the ancestry of the sample. The coalescent has a very simple structure. Ancestral lines going backward in time coalesce when they have a common ancestor. Coalescence occur only between pairs of individuals. This process may also be thought of as generating a binary tree, with the leaves representing the sample sequences and the vertices where ancestral lines coalesce. The root of the tree is the most recent common ancestor (MRCA) of the sample.

Example 1 To get an idea for building trees from sequences we begin with a simple example. Consider just the segregating purines of lineage b, c, d, e from Table 7. Below is this reduced data set.

	Site	1	2	3	4
lineage					
b		A	G	G	A
c		G	A	G	G
d		G	G	A	G
e		G	G	G	A

Suppose $G\ G\ G\ G$ is the ancestral sequence to the four lineages. Figure 1 represents one possible evolutionary scenario connecting the individuals in the sample. The $\bullet$ represents a mutation. The number next to the $\bullet$ represents the position of the mutation. We read the tree diagram in Figure 1 as follows. Start at the ancestral tip of the tree. The first event to occur is a split in the ancestral line. Next a mutation occurs and a G mutates to an A at position 4. This mutation is passed on to lineage b and e. The tree splits

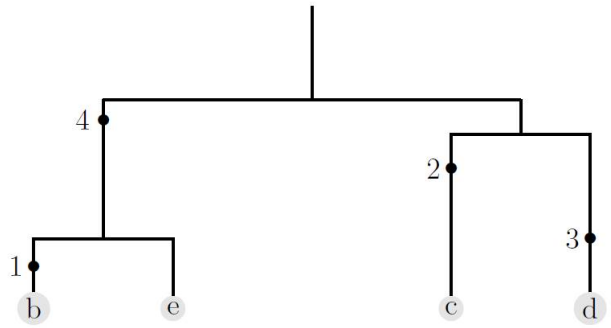

Figure 1: A tree consistent with the data in Example 1

again and a mutation occurs at site 2 in lineage c. Next, a mutation occurs at site 3 in lineage d. The final split in the tree separates lineage b and e. The last evolutionary event is a mutation at site 1 in lineage b.

A coalescent tree consistent with the data in Table 7 is given below

A rooted and unrooted gene tree consistent with the data in Table 7 is given below

Exercise 1 Convince yourself that Figure 2. represents a coalescent tree that is consistent with the data given in Table 7. Use the most frequently occurring basepair at each site as the ancestral sequence. Describe each event that lead to the sample. Construct another coalescent tree that is consistent with the data.

Exercise 2 Since mutations can occur only once in a given site, there is an ancestral type and a mutant type at each segregating site. For the moment assume we know which is which, and label the ancestral type is 0 and the mutant type as 1. To fix ideas, take each column of the data in Table 7 and label the most commonly occurring base as 0, the other as 1. Construct a matrix of 0's and 1's for the data in Table 7 in the manner described above.

The matrix of 0's and 1's can be represented by a rooted tree

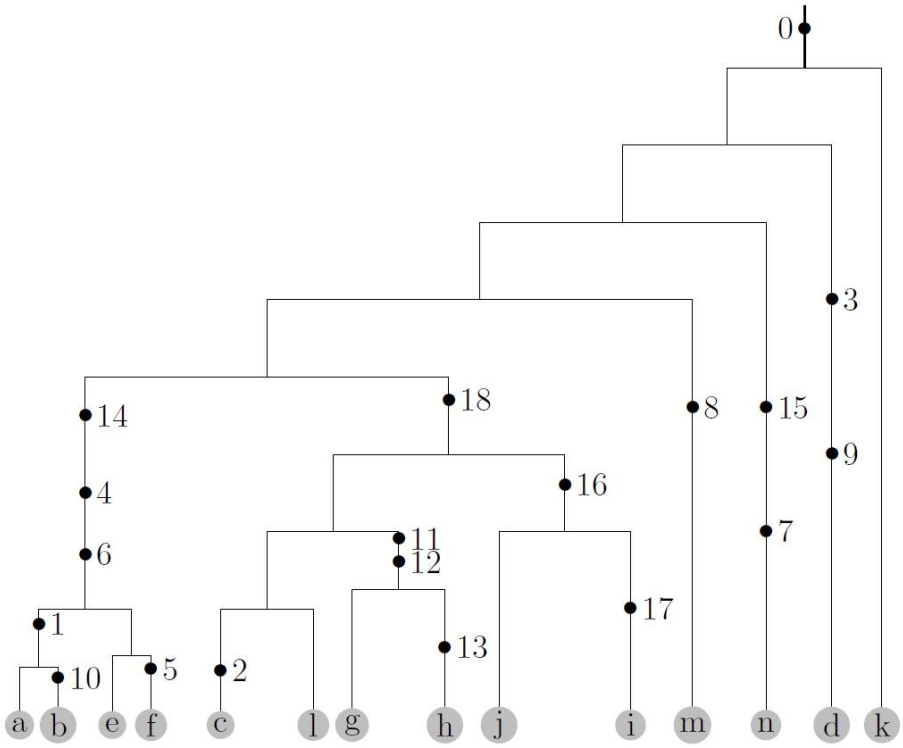

Figure 2: A tree consistent with the data in Table 7

by labelling each distinct row by a sequence of mutations up to the common ancestor. These mutations are the vertices in the tree. This rooted tree is a condensed description of the coalescent tree with its mutations, and it has no time scale in it. Figure 3 represents a rooted condensed tree consistent with the data in table 1.

Exercise 3 Verify that Figure 3 is consistent with the data. Take lineages a, b, c, d, j and draw the rooted condensed tree for this subset of individuals.

Of course, in practice we never know which type at a site is ancestral. All that can be deduced then from the data are the number of segregating sites between each pair of sequences. In this case the data is equivalent to an *unrooted* tree whose vertices represent

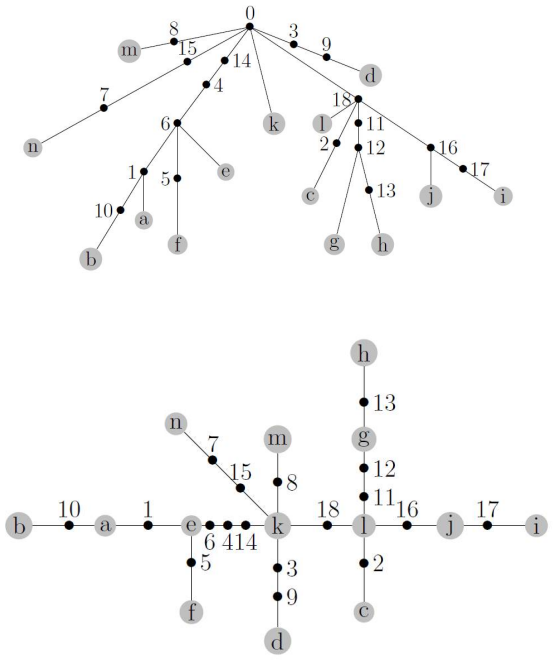

Figure 3: Rooted gene tree Figure 3a (top), Unrooted tree Figure 3b (bottom)

distinct lineages and whose edges are labeled by mutations between lineages. The unrooted tree corresponding to the rooted tree in Figure 3a is shown in Figure 3b. All possible rooted trees may be found from an unrooted tree by placing the root at a vertex or between mutations, then reading off mutation paths between lineages and the root.

Exercise 4 Construct three rooted trees consistent with the unrooted tree in Figure 3b.

We will return to this example later in the course. At that time we will address the problem of estimating the mutation rate, and predicting the time back to the most recent common ancestor. The example illustrates how to connect DNA sequence data to the ancestry of the individuals in the population.

Two technical results

It is convenient to start with two technical results, one of which will be relevant for approximations in the coalescent associated with the Wright-Fisher model, and by implication the Cannings model, while the other will be relevant for exact Moran model calculations.

We consider first a Poisson process in which events occur independently and randomly in time, with the probability of an event in $(t, t+\delta t)$ being $a\delta t$. (Here and throughout we ignore terms of order $(\delta t)^2$.) We call a the rate of the process. Standard Poisson process theory shows that the density function of the (random) time X between events, and until the first event, is $f(x) = a\ e^{-ax}$, and thus that the mean time until the first event, and also between events, is $1/a$.

Consider now two such processes, process (a) and process (b), with respective rates a and b. From standard Poisson process theory, given that an event occurs, the probability that it arises in process (a) is $a/(a+b)$. The mean number of "process (a)" events to occur before the first "process (b)" event occurs is a/b. More generally, the probability that j "process (a)" events occur before the first "process (b)" event occurs is

$$\frac{b}{a+b}\Big(\frac{a}{a+b}\Big)^j, \quad j = 0, 1, \ldots . \tag{380}$$

The mean time for the first event to occur under one or the other process is $1/(a+b)$. Given that this first event occurs in process (a), the conditional mean time until this first event occurs is equal to the unconditional mean time, namely $1/(a+b)$. The same conclusion applies if the first event occurs in process (b).

Similar properties hold for the geometric distribution. Consider a sequence of independent trials and two events, event A and event B. The probability that one of the events A and B occurs at any trial is $a+b$. The events A and B cannot both occur at the same trial, and given that one of these events occurs at trial i, the probability that it is an A event is $a/(a+b)$.

Consider now the random number of trials until the first event occurs. This random variable has geometric distribution, and takes the value i, $i = 1, 2, \ldots$, with probability $(1-a-b)^{i-1}(a+b)$. The mean of this random variable is thus $1/(a+b)$. The probability that the first event to occur is an A event is $a/(a+b)$. Given that the

first event to occur is an A event, the mean number of trials before the event occurs is $1/(a+b)$. In other words, this mean number of trials applies whichever event occurs first. The similarity of properties between the Poisson process and the geometric distribution is evident.

Approximate results for the Wright-Fisher model - no mutation

With the above results in hand, we first describe the general concept of the coalescent process. To do this, we consider the ancestry of a sample of n genes taken at the present time. Since our interest is in the ancestry of these genes, we consider a process moving backward in time, and introduce a notation acknowledging this. We consistently use the notation τ for a time in the past before the sample was taken, so that if $\tau_2 > \tau_1$, then τ_2 is further back in the past than is τ_1.

We describe the common ancestry of the sample of n genes at any time τ through the concept of an equivalence class. Two genes in the sample of n are in the same equivalence class at time τ if they have a common ancestor at this time. Equivalence classes are denoted by parentheses: Thus if $n = 8$ and at time τ genes 1 and 2 have one common ancestor, genes 4 and 5 a second, and genes 6 and 7 a third, and none of the three common ancestors are identical, the equivalence classes at time time τ are

$$(1,2), \quad (3), \quad (4,5), \quad (6,7), \quad (8). \tag{381}$$

Such a time τ is shown in Figure 4.

We call any such set of equivalence classes an equivalence relation, and denote any such equivalence relation by a Greek letter. As two particular cases, at time $\tau = 0$ the equivalence relation is $\phi_1 = \{(1),(2),(3),(4),(5),(6),(7),(8)\}$, and at the time of the most recent common ancestor of all eight genes, the equivalence relation is $\phi_n = \{(1,2,3,4,5,6,7,8)\}$. The coalescent process is a description of the details of the ancestry of the n genes moving from ϕ_1 to ϕ_n.

Let ξ be some equivalence relation, and η some equivalence relations that can be found from ξ by amalgamating two of the equivalence classes in ξ. Such an amalgamation is called a coalescence, and the process of successive such amalgamations is called the coalescence process. It is assumed that, if terms of order $(\delta\tau)^2$ are

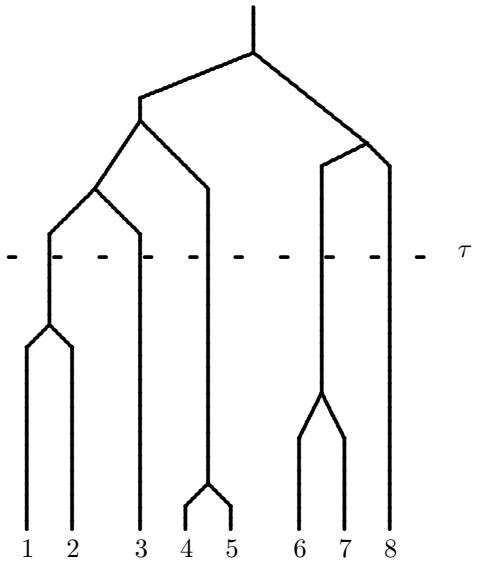

Figure 4: The coalescent

ignored, and given that the process is in ξ at time τ,

$$\text{Prob (process in } \eta \text{ at time } \tau + \delta\tau) = \delta\tau, \tag{382}$$

and if j is the number of equivalence classes in ξ,

$$\text{Prob (process in } \xi \text{ at time } \tau + \delta\tau) = 1 - \frac{j(j-1)}{2}\delta\tau. \tag{383}$$

The above assumptions are clearly approximations for any discrete-time process, but they are precisely the assumptions needed for the Wright-Fisher approximate coalescent theory.

The coalescent process defined by (382) and (383)consists of a sequence of $n-1$ Poisson processes, with respective rates $j(j-1)/2$, $j = n, n-1, \ldots, 2$, describing the Poisson process rate at which two of these classes amalgamate when there are j equivalence classes in the coalescent. Thus the rate $j(j-1)/2$ applies when there are j ancestors of the genes in the sample for $j < n$, with the rate $n(n-1)/2$ applying for the actual sample itself.

The Poisson process theory outlined above shows that the time T_j to move from an ancestry consisting of j genes to one consisting of $j-1$ genes has an exponential distribution with mean $2/\{j(j-1)\}$.

Since the total time required to go back from the contemporary sample of genes to their most recent common ancestor is the sum of the times required to go from j to $j-1$ ancestor genes, $j = 2, 3, \ldots, n$, the mean $E(T_{\text{MRCAS}})$ is, immediately,

$$T_{MRCAS} = T_n + T_{n-1} + \cdots + T_2 \tag{384}$$

It follows that

$$\begin{aligned} E(T_{MRCAS}) &= \sum_{k=2}^{n} E(T_k) \\ &= \sum_{k=2}^{n} \frac{2}{k(k-2)} \\ &= 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) \\ &= 2\left(1 - \frac{1}{n}\right) \end{aligned} \tag{385}$$

Therefore

$$1 = E(T_2) \leq E(T_{MRCAS}) < 2$$

Note that T_{MRCAS} is close to 2 even for moderate n.

Example 2 Again consider a sample of $n = 30$ Nuu-Chah females in a population of $N = 600$. The mean time to a common ancestor of the sample is $2(1 - \frac{1}{30}) = 1.933$ (1160 generations) and the mean time to a common ancestor of the population is $2(1 - \frac{1}{600}) = 1.997$ (1198 generations). The mean difference between the time for a sample of size 30 to reach a MRCA, and the time for the whole population to reach its MRCA is 0.063, which is about 38 generations.

Warning The above calculations are not based on any of the base-pair sequence information in the Nuu-Chah data set. They can only be viewed as crude guesses as to what one might expect from an

unstructured randomly mating population. We will see later that our predictions can be refined once we fit the data to the model.

Note that T_2 makes a substantial contribution to the sum in (385) for T_{MRCAS}. For example, on average for over half the time since its MRCA, the sample will have exactly two ancestors.

Further, using independence of the T_k,

$$\begin{aligned}\mathrm{Var}(T_{MRCAS}) &= \sum_{k=2}^{n} \mathrm{Var}(T_k) \\ &= \sum_{k=2}^{n} \left(\frac{2}{k(k-1)}\right)^2 \\ &= 8\sum_{k=1}^{n-1} \frac{1}{k^2} - 4\left(1-\frac{1}{n}\right)\left(3+\frac{1}{n}\right).\end{aligned}$$

It follows that

$$1 = \mathrm{Var}(T_2) \leq \mathrm{Var}(T_{MRCAS}) \leq \lim_{n\to\infty} \mathrm{Var}(T_{MRCAS}) = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

Exercise 5 Calculate the mean and standard deviation of the time to the MRCA of a population of $N = 600$. Express your answer in units of generations.

Lineage Sorting–an application in phylogenetics

Now focus on two particular individuals in the sample and observe that if these two individuals do not have a common ancestor at t, the whole sample cannot have a common ancestor. Since the two individuals are themselves a random sample of size two from the population, we see that

$$P(T_{MRCAS} > t) \geq P(T_2 > t) = e^{-t},$$

it can be shown that

$$P(T_{MRCAS} > t) \leq \frac{3(n-1)}{n+1} e^{-t} \tag{386}$$

and so

$$e^{-t} \leq P(T_{MRCAS} > t) \leq 3e^{-t} \tag{387}$$

The coalescent provides information on the history of genes within a population or species; by contrast, phylogenetic analysis studies the relationship between species. Central to a phylogenetic analysis of molecular data is the assumption that all individuals within a species have coalesced to a common ancestor at a more recent time point than the time of speciation, see Figure 5 for an illustration. If this assumption is met then it does not matter which homologous DNA sequence region is analyzed to infer the ancestral relationship between species. The true phylogeny should be consistently preserved regardless of the genetic locus used to infer the ancestry. If there is a discrepancy between the inferred phylogeny at one locus versus another then that discrepancy can be explained by the stochastic nature of statistical inference. However, the within species ancestry and the between species ancestry are not always on different time scales and completely separable. It is possible that a particular homologous region of DNA used to produce a phylogeny between species could produce a different phylogeny than a different homologous region and the difference is real (see Figure 6). One explanation of this phenomena is called *lineage sorting* and it occurs when the time to speciation is more recent than the time to the most recent common ancestry of the gene. This makes it appear like two sub-populations from the same species are more distantly related than two distinct species.

However, the coalescent model can actually help determine if lineage sorting is plausible. For example, if based on external evidence, (possibly fossil evidence) the time to speciation is at least u generations into the past, then it is reasonable to ask, how likely is it that a population has not reached a common ancestor by time u. Converting from generations to coalescent the time scale, define $t = u/2N_e$. If T_{MRCAS} is the time it takes a population to reach a common ancestor, then we can use equation (387) to determine if lineage sorting is a reasonable explanation. If $3e^{-t}$ is small, then coalescent time scale and the phylogenetic time scales are likely to be different and lineage sorting is likely not to be the appropriate explanation. Thus another implication of coalescent theory is that the it provides appropriate insight as to how distantly related genes are within a species, which can help resolve issues in phylogenetic

analysis.

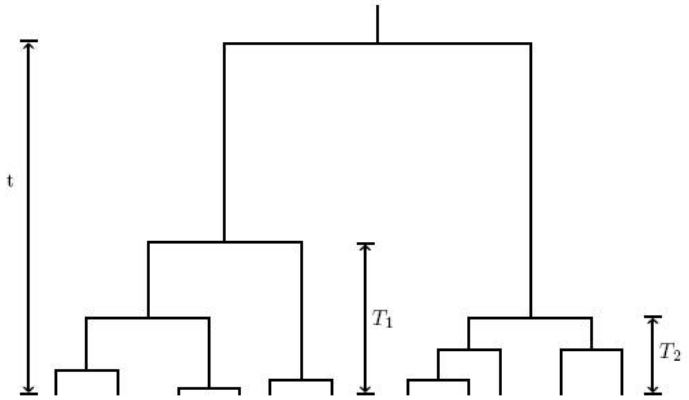

Figure 5: Population coalescence does not predate speciation

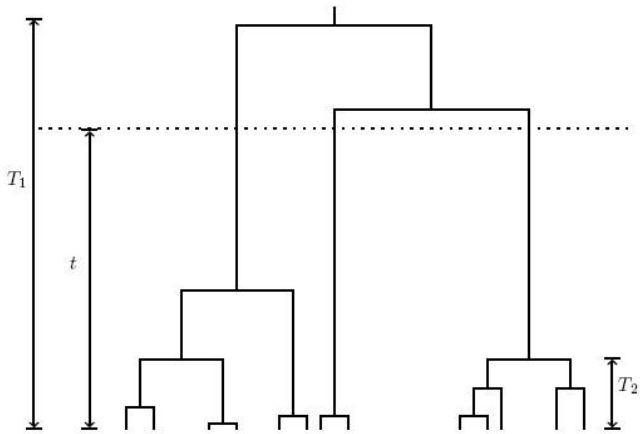

Figure 6: Population coalescence predates speciation

Approximate results for the Wright-Fisher model with mutation

We now introduce mutation, and suppose that the probability that any gene mutates in the time interval $(\tau + \delta\tau, \tau)$ is $(\theta/2)\delta\tau$. All mutants are assumed to be of new allelic types. Following the coalescent paradigm, we trace back the ancestry of a sample of n genes

to the mutation forming the oldest allele in the sample. As we go backward in time along the coalescent, we shall encounter from time to time a "defining event", taken either as a coalescence of two lines of ascent into a common ancestor or a mutation in one or other of the lines of ascent. Figure 7 describes such an ancestry, identical to that of Figure 4 but with crosses to indicate mutations.

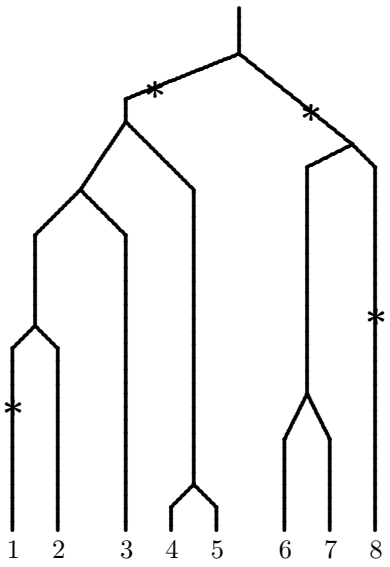

Figure 7: The coalescent with mutations

We exclude from further tracing back any line in which a mutation occurs, since any mutation occurring further back in any such line does not appear in the sample. Thus any such line may be thought of as stopping at the mutation, as shown in Figure 8 (describing the same ancestry as that in Figure 7).

If at time τ there are j ancestors of the n genes in the sample, the probability that a defining event occurs in $(\tau, \tau + \delta\tau)$ is

$$\frac{1}{2}j(j-1)\delta\tau + \frac{1}{2}j\theta\delta\tau \;=\; \frac{1}{2}j(j+\theta-1)\delta\tau, \tag{388}$$

the first term on the left-hand side arising from the possibility of a coalescence of two lines of ascent, and the second from the possibility of a mutation.

If a defining event is a coalescence of two lines of ascent, the

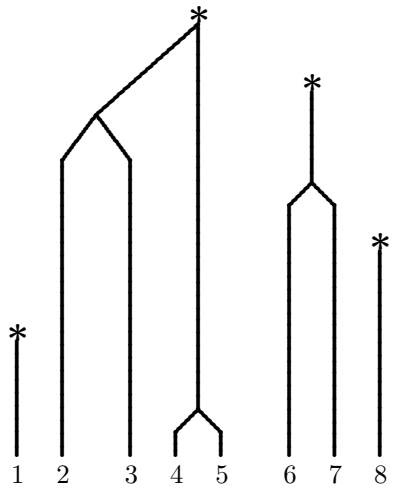

Figure 8: Tracing back to, and stopping at, mutational events

number of lines of ascent clearly decreases by 1. The fact that if a defining event arises from a mutation we exclude any further tracing back of the line of ascent in which the mutation arose implies that the number of lines of ascent also decreases by 1. Thus at any defining event the number of lines of ascent considered in the tracing back process decreases by 1. Given a defining event leading to j genes in the ancestry, the Poisson process theory described above shows that, going backward in time, the mean time until the next defining event occurs is $2/\{j(j+\theta-1)\}$, and that the same mean time applies when we restrict attention to those defining events determined by a mutation.

Thus starting with the original sample and continuing up the ancestry until the mutation forming the oldest allele in the sample is reached, we find that the mean age of the oldest allele in the sample is

$$2\sum_{j=1}^{n}\frac{1}{j(j+\theta-1)} \tag{389}$$

coalescent time units. The value in (389) must be multiplied by $2N$ to give this mean age in terms of generations.

The time backward until the mutation forming the oldest allele in the sample, whose mean is given in (389), does not necessarily trace back to, and past, the most recent common ancestor of the genes in the sample (MRCAS), and will do so only if the allelic type of the MRCAS is represented in the sample. This observation can be put in quantitative terms by comparing the MRCAS given in (385) to the expression in (389). For small θ, the age of the oldest allele will tend to exceed the time back to the MRCAS, while for large θ, the converse will tend to be the case. The case $\theta = 2$ appears to be a borderline one: For this value, the expressions in (385) and (389) differ only by a term of order n^{-2}. Thus for this value of θ, we expect the oldest allele in the sample to have arisen at about the same time as the MRCAS.

The competing Poisson process theory outlined above shows that, given that a defining event occurs with j genes present in the ancestry, the probability that this is a mutation is $\theta/(j-1+\theta)$. Thus the mean number of different allelic types found in the sample is

$$\sum_{j=1}^{n} \frac{\theta}{j-1+\theta},$$

and this is the value given in (146). The number of "mutation-caused" defining events with j genes present in the ancestry is, of course, either 0 or 1, and thus the variance of the number of different allelic types found in the sample is

$$\sum_{j=1}^{n} \left(\frac{\theta}{j-1+\theta} - \frac{\theta^2}{(j-1+\theta)^2} \right).$$

This expression is easily shown to be identical to the variance formula (147).

Even more than this can be said. The probability that exactly k of the defining events are "mutation-caused" is clearly proportional to $\theta^k/\{\theta(\theta+1)\cdots(\theta+n-1)\}$, the proportionality factor not depending on θ. Since this is true for all possible values of θ and since the sum of the probabilities over $k = 1, 2, \ldots, n$ must be 1, the probability distribution of the number of different alleles in the sample must be given by (145).

The complete distribution of the allelic configuration in the sample as given in (143) is not so simply derived. Kingman (1982), to

whom coalescent theory is due, employed the full machinery of the coalescent process, together with a combinatorial argument considering all possible paths from ϕ_n to ϕ_1, to derive (143). That is, (143) derives immediately from, and is best thought of as a consequence of, the coalescent properties of the ancestry of the genes in the sample.

The sample contains only one allele if no mutants occurred in the coalescent after the original mutation for the oldest allele. Moving up the coalescent, this is the probability that all defining events before this original mutation is reached are amalgamations of lines of ascent rather than mutations. The probability of this is

$$\prod_{j=1}^{n-1} \frac{j}{(j+\theta)} = \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}, \tag{390}$$

and this agrees, as it must, with the expression in (148).

The length of a coalescent tree is defined to be the sum of all of its branch lengths and is denoted by L_n which can be determined from the coalescent times as follows

$$L_n = \sum_{j=2}^{n} jT_j,$$

where the random variable T_j are independent and have exponential distribution with rate parameter $j(j-1)/2$. If S_n denotes the total number of mutations on the genealogical tree back to the MRCA of a sample of size n, then conditional on L_n, S_n has a Poisson

distribution with mean $\theta L_n/2$. It follows that

$$\begin{aligned} E(S_n) &= E(E(S_n|L_n)) \\ &= E(\theta L_n/2) \\ &= \frac{\theta}{2} E(\sum_{i=2}^{n} iT_i) \\ &= \frac{\theta}{2} \sum_{i=2}^{n} iE(T_i) \\ &= \frac{\theta}{2} \sum_{i=2}^{n} i \frac{2}{i(i-1)} \\ &= \theta \sum_{j=1}^{n-1} \frac{1}{j} \end{aligned} \tag{391}$$

Notice that for large n then $E(S_n) \sim \theta \log n$.

Example 3 We calculate the mean number of mutations for various sample sizes, when $\theta = 4$,

Sample size n	θ	$E(S_n)$
10	4	9.21
20	4	11.98
40	4	14.76
45	4	15.23
50	4	15.65
60	4	16.38
100	4	18.42

Recall that $E(S_n)$ is the average number of mutations accumulated by a sample of size n under the neutral coalescent model. Any given realization of evolution will produce an S_n that varies around the expected value. The standard deviation of S_n tells you how much variation to expect. The formula for standard deviation is

$\text{STDEV}(S_n) = \sqrt{\text{Var}(S_n)}$. We will now calculate $\text{Var}(S_n)$. Because S_n arises as a result of two random processes, we need to account for both processes in our calculation of the variance of S_n. Below is the formula that is needed to calculate $\text{Var}(S_n)$.

$$\text{Var}(S_n) = E(\text{Var}(S_n|L_n)) + \text{Var}(E(S_n|L_n)). \tag{392}$$

One way to interpret Equation (392) is as follows. There are two sources of variation. One is due to fluctuations inherent in the coalescent process and the other is due to the fluctuations inherent in the Poisson mutation process. The term $E(\text{Var}(S_n|L_n))$ can be thought as the contribution of the variance of S_n attributed to the Poisson mutation process. $\text{Var}(E(S_n|L_n))$ may be though of as the amount variation due to the coalescent process. Therefore,

$$\begin{aligned}
\text{Var}(S_n) &= E(\text{Var}(S_n|L_n)) + \text{Var}(E(S_n|L_n)) \\
&= E\left(\frac{\theta}{2}L_n\right) + \text{Var}\left(\frac{\theta}{2}L_n\right) \\
&= \frac{\theta}{2}E(L_n) + \frac{\theta^2}{4}\text{Var}(L_n) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\text{Var}\left(\sum_{i=2}^{n} iT_i\right) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\sum_{i=2}^{n} i^2\text{Var}\left(T_i\right) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\sum_{i=2}^{n} i^2\frac{4}{i^2(i-1)^2} \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}
\end{aligned}$$

For large n then

$$\text{Var}(S_n) = \theta\log n + 2\theta^2$$

Below is a table means and standard deviations

n	θ	$E(S_n)$	Stdev(S_n)
10	4	9.21	6.00
20	4	11.98	6.10
40	4	14.76	6.20
45	4	15.23	6.21
50	4	15.65	6.23
60	4	16.38	6.25
100	4	18.42	6.32

Exact results for the Moran model - no mutation

We now turn to exact coalescent results for the Moran model. These are found in a manner similar to that used above, with the time unit used corresponding to the time between one birth and death event and the next.

As we did for the Wright–Fisher model, we first consider the coalescent process itself. Here, however, we use a coalescent theory that is not only exact, but that also applies for a sample of any size, and in particular to the entire population of genes itself. This implies that all results deriving from coalescent theory, for example the topology of the coalescent tree, are identical to corresponding results for the exact Moran model coalescent process.

It is convenient to think of a gene that does not die in a birth and death event as being its own descendant after that event has take place. Consider, then, a sample of n genes, where n is not restricted to be small and could be any number up to and including the entire population size of $2N$. As we trace back the ancestry of these n genes we will encounter a sequence of coalescent events reducing the size of the ancestry to $n-1, n-2, \ldots$ genes and eventually to one gene, the most recent common ancestor of the sample. Suppose that in this process we have just reached a time when there are exactly j genes in this ancestry. These will be "descendants" of $j-1$ parental genes if one of these parents was chosen to reproduce and the offspring is in the ancestry of the sample of n genes. The probability of this event is $j(j-1)/(2N)^2$. With probability $1-j(j-1)/(2N)^2$ the number of ancestors remains at j. It follows that, as we trace back the ancestry of the genes, the number T_j of birth and death events between the times when there are j ancestor genes and $j-1$ ancestor genes has,

exactly, a geometric distribution with parameter $j(j-1)/(2N)^2$ and thus with mean $(2N)^2/\{j(j-1)\}$. From this, the mean of the time T_{MRCAS} until the most recent common ancestor of all the genes in the sample is given by

$$E(T_{\mathrm{MRCAS}}) = \sum_{j=2}^{n} \frac{(2N)^2}{j(j-1)} = (2N)^2 \left(1 - \frac{1}{n}\right) \tag{393}$$

birth and death events. In the particular case $n = 2N$ this is

$$E(T_{\mathrm{MRCAP}}) = 2N(2N-1) \tag{394}$$

birth and death events.

Since the various T_j's are independent, the variance of T_{MRCAP} is the sum of the variances of the T_j's. This is

$$\mathrm{var}(T_{\mathrm{MRCAS}}) = \sum_{j=2}^{n} \frac{(2N)^4}{j^2(j-1)^2} - \sum_{j=2}^{n} \frac{(2N)^2}{j(j-1)}. \tag{395}$$

The complete distribution of T_{MRCAP} can be found, but the resulting expression is complicated and is not given here.

Exact results for the Moran model with mutation

We now introduce mutation. Consider again a sample of n genes and the sequence of birth and death events that led to the formation of this sample. We again trace back the ancestry of the n genes in the sample, and consider some birth and death event when this ancestry contains $j-1$ genes. With probability $j/2N$ the newborn created in the population at this birth and death event is in the ancestry of the sample, and with probability u is a mutant. That is, the probability that at this birth and death event a new mutant gene is added to the ancestry of the sample is $ju/(2N)$. As for the Wright–Fisher model, we trace back upward along the lines of ascent from the sample, and do not trace back any further any line of ascent at a time when a new mutant arises in that line, so that at any mutation, the number of lines of ascent that we consider decreases by 1.

A further decrease can occur from a coalescence for which the addition of a newborn to the ancestry of the sample does not produce a mutant offspring gene. If at any time there are j lines in the ancestry, the probability of a coalescence not arising from a mutant newborn is $j(j-1)(1-u)/(2N)^2$.

It follows from the above that the number of lines of ascent from the sample will decrease from j to $j-1$ at some birth and death event with total probability

$$\frac{ju}{2N} + \frac{j(j-1)(1-u)}{(2N)^2} = \frac{2Nju + j(j-1)(1-u)}{(2N)^2}. \tag{396}$$

We write the left-hand side as $v_j + w_j$, where v_j and w_j are defined in (168). The number of birth and death events until a decrease in the number of lines of ascent from j to $j-1$ follows a geometric distribution with parameter $v_j + w_j$. It follows from the competing geometric theory given above that the mean number of birth and death events until the number of lines of ascent decreases from j to $j-1$ is $1/(v_j + w_j)$, and that this mean applies whatever the reason for the decrease. Tracing back to the mutation forming the oldest allele in the sample, we see that the mean age of this oldest allele is, exactly,

$$\sum_{j=1}^{n} \frac{1}{v_j + w_j}, \tag{397}$$

where v_j and w_j are defined in (168).

The probability that a decrease in the number of ancestral lines from j to $j-1$, given that such a decrease occurs, is $v_j/(v_j + w_j)$, or, using the Moran model definition of θ, more simply as $\theta/(j-1+\theta)$. The mean number of different alleles in the sample is thus, exactly,

$$\sum_{j=1}^{n} \frac{\theta}{j-1+\theta}, \tag{398}$$

as given by (146). Extending this argument as for the Wright–Fisher case, the exact distribution of the number of alleles in the sample is found to be given by (145), as expected.

The complete distribution of the sample allelic configuration, as with the Wright–Fisher model, requires a full description of the coalescent process.

The argument just used, while expressed as one concerning a sample of genes, applies equally for the entire population of genes. This occurs because, even in the entire population, at most one coalescent event can occur at each birth and death event. Thus all the exact sample Moran model results found by coalescent arguments apply for the population as a whole, with n being replaced by $2N$.

This explains the identity of the form of many exact Moran model sample and population formulas.

Estimating the parameter θ

We have been investigating the properties of the neutral coalescent. We have been focusing our efforts on answering the following question: if the neutral model for evolution with constant mutation rate is a reasonable model, what can we expect the ancestry of a sample to look like? We found that under neutrality coalescence occur at the rate of $n(n-1)/2$ where n is the sample size. This means that on average, coalescence occur quickly in the recent past and then very slowly in the more distant past, as the number of ancestors becomes small. In fact, on average, half the time back to MRCA is T_2 the time for last two ancestors to coalesce. We found that the average number of mutations back to the MRCA is proportional to the mutation parameter θ and inversely proportional to $\log n$.

Of course, averages tell only part of the story. There is a fair amount of variation about the average. To get a handle on the variation, we calculated the variance and standard deviation for T_n, the time back to the MRCA, and the variance and standard deviation of S_n, the number of mutations back to the MRCA of a sample of size n.

We now want to shift the focus from mathematical modelling to statistical inference. Rather than ask, 'for a given mutation parameter, θ, what can we say about the ancestry of the sample, we now ask the more relevant question, for a given sample, what can we say about the population. In particular, what is our best estimate for θ based on information in a sample.

Watterson's estimator

Under the assumptions of the infinite sites model, the number of segregating sites is exactly the total number of mutations S_n since the MRCA of the sample. Recall that

$$E(S_n) = a_n\theta \tag{399}$$

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ and

$$\text{Var}(S_n) = a_n\theta + b_n\theta^2 \tag{400}$$

where $b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$.

Define

$$\hat{\theta}_S = \frac{S_n}{a_n}.$$

This is called the segregating sites estimator for θ and goes back to a paper by Watterson (1975). Note that it follows from (399) that $E(\hat{\theta}_S) = \theta$. Estimators of this type are called unbiased. It follows from (400) that

$$\text{Var}(\hat{\theta}_S) = \frac{1}{a_n}\theta + \frac{b_n}{a_n^2}\theta^2 \tag{401}$$

It is easy to see that $\text{Var}(\hat{\theta}_S) \to 0$ as $n \to \infty$. Estimators with this property are said to be consistent. This means that one can attain any level of precision desired by choosing the sample size sufficiently large. However, don't expect the precision to be much better than half the size of the estimate, unless you require ridiculously large sample sizes.

Example 4 If it is known that θ is no bigger than 6, using the segregating sites estimator for θ, how large a sample is required to insure that the error of the estimate is less than or equal to 1?
Soln. Lets assume that the error of an estimate is 2 standard deviations. If 2 standard deviations is 1 unit, then we want to choose a sample size so that the standard deviation of the estimate for θ is less than or equal to .5. Using a conservative initial guess for θ to be 6 we have

$$.5 = \sqrt{\frac{6}{a_n} + \frac{36b_n}{a_n^2}}$$

We wish to solve for n in the above equation. To simplify matters lets replace b_n with its upper bound of 2. Therefore

$$\begin{aligned} .25 &= \frac{6}{a_n} + \frac{72}{a_n^2} \\ .25a_n^2 &= 6a_n + 72 \end{aligned}$$

Solving the above quadratic equation gives $a_n \approx 32$, implying $n \approx 1.73 \times 10^{14}$. However, if you require a standard deviation for the estimate to be 2, then the sample size required for this level of precision is just $n = 158$.

The expressions in (334) shows that θ does not admit unbiased estimation using infinitely many alleles data. By contrast, it is clear from the above that θ admits unbiased estimation in the infinitely many sites case, using estimators based on S_n. This makes it all the more remarkable that, whereas the infinitely many alleles quantity K_n is a sufficient statistic for θ in that model, S_n is *not* a sufficient statistic for θ in the infinitely many sites model. This implies that in the infinitely many sites model, the data in a sample of genes beyond that given by S_n can in principle be used to provide better estimation of θ than than provided through S_n.

Note that the variance (401) is of order $1/\log n$ and is thus quite large even for large n. The same is true of the variance of the estimator (338). This implies that neither K_n nor S_n provides reliable estimation of θ. A variance of order $1/\log n$, rather than the classic statistical order $1/n$, arises in both cases because of the dependence between the genes in the sample arising from their common ancestry.

Finally we can compare the variance (401) of $\hat{\theta}_S$ with the approximate infinitely many alleles mean square error (MSE) of $\hat{\theta}_K$, given in (338). This comparison shows that the variance of $\hat{\theta}_S$ is sometimes less than, and sometimes more than, the approximating MSE of $\hat{\theta}_K$ given in (338). For small θ the two expressions are, as we expect, quite close. For $\theta \leq 1$ the variance of $\hat{\theta}_S$ is always less than the approximating MSE of $\hat{\theta}_K$, being about 94% of the approximating MSE when $\theta = 1$, $n = 100$. Further, the variance of $\hat{\theta}_S$ is always less than the approximating MSE when $n \leq 50$, but for $n = 51$ it is possible to find values of θ for which the reverse is true. When $\theta = 5$, $n = 500$, the variance of $\hat{\theta}_S$ is about 18% larger than the approximating MSE of $\hat{\theta}_K$.

It is in principle possible to employ more detailed "sites" data to find a better estimator of θ than that provided by using only S_n, which ignores aspects of these more detailed data. This matter has been discussed at length in the literature. Optimal estimation in statistics arises through the method of maximum likelihood, and thus the aim is to find the likelihood of a sample of n genes, the data in this sample involving not only the value of S_n but the complete

configuration of the nucleotides at the various segregating sites.

Pairwise differences

Recall that θ is the expected number of mutations separating two individuals. So a natural way to estimate θ is to calculate number of mutations separating individuals two at a time and average over all pairs. This may be thought of as a sample average used to estimate a population average. To calculate this we take individuals two at a time. Denote by

$$S_{ij} = \text{Number of mutations separating individuals i and j.}$$

Under the infinite sites assumption, we can calculate S_{ij} from a sample by calculating the number of segregating sites between sequences i and j. If we average S_{ij} over all pairs (i, j), this is called the average number of pairwise differences. We denote the average number of pairwise differences by.

$$D_n = \frac{2}{n(n-1)} \sum_{i \leq j} S_{ij}.$$

Note that we can think of individuals (i, j) as sample of size 2, therefore

$$E(S_{ij}) = E(S_2) = \theta.$$

Therefore,

$$E(D_n) = \frac{2}{n(n-1)} \sum_{i \leq j} E\left(S_{ij}\right) = \theta$$

Thus, D_n is an unbiased estimator. Tajima (1981) was the first to investigate the properties of D_n. We will refer to $\hat{\theta}_T = D_n$. It is interesting to note that $\hat{\theta}_T$ has very poor statistical properties. In fact, $\hat{\theta}_T$ has higher variance than any of the other estimators we will consider. Why does an estimator that seems so natural have such poor properties? The answer lies in the fact that their is dependence in the data generated by the common ancestral history. This means that S_{ij} and S_{kl} are positively correlated random variables. As a result the precision of the estimator D_n will be low.

In fact,

$$\text{Var}(D_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \tag{402}$$

The details for deriving $\text{Var}(D_n)$ are left as an exercise (se below) Note that

$$\lim_{n\to\infty} \text{Var}(D_n) = \frac{\theta}{3} + \frac{2}{9}\theta^2.$$

The pairwise difference estimate is not consistent. The square root of the above limit represents the optimal precision one can obtain, regardless of sample size, using the pairwise difference estimator.

Exercise 6

1. Show that

$$\begin{aligned} E(D_n^2) &= \frac{1}{n^2(n-1)^2}\left[2n(n-1)E(S_{12}^2) + 4n(n-1)(n-2)E(S_{12}S_{13})\right. \\ &\quad \left. +n(n-1)(n-2)(n-3)E(S_{12}S_{34})\right]. \end{aligned}$$

2. Show that $E(S_{12}^2) = 2\theta^2 + \theta$.
3. It can be shown that

$$E(S_{12}S_{13}) = \frac{4\theta^2}{3} + \frac{\theta}{2},$$

and

$$E(S_{12}S_{34}) = \frac{11\theta^2}{9} + \frac{\theta}{3}.$$

Use these results to calculate $E(D_n^2)$

4. Show that

$$\text{Var}(D_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

Likelihood and Efficiency

In the last section we considered two estimators for θ that were based on summary statistics. Noticed that the segregating sites method performed better than the pairwise difference method. However, both estimators tend to have fairly high variance. The theory of mathematical statistics provides us with a lower bound on the variance of all unbiased estimators. This lower bound is called the Cramèr-Rao lower bound. Efficiency of an estimator is defined to be the variance of an estimator relative to the minimum variance possible. In this section we begin with some general results from mathematical statistics. In particular we establish the Cramèr-Rao

lower bound. We then calculate this lower bound in the context of the neutral coalescent model.

General Set up Let $X_1, X_2, \cdots, X_n$ be a sample of size n with

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n; \theta).$$

We wish to estimate the parameter θ. An estimate of θ is a function of the data. Let $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, \ldots, X_n)$ be an estimator of θ. An unbiased estimator has the property that

$$E(\hat{\theta}) = \theta.$$

Result 1

$$E\left(\frac{\partial}{\partial\theta}\log f(X_1, X_2, \ldots, X_n; \theta)\right) = 0$$

Proof.

Define

$$u(x_1, x_2, \ldots, x_n; \theta) = \frac{\partial}{\partial\theta}\log f(x_1, x_2, \ldots, x_n; \theta)$$

which can also be written as

$$u(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{f(x_1, x_2, \ldots, x_n; \theta)}\frac{\partial}{\partial\theta} f(x_1, x_2, \ldots, x_n; \theta).$$

If we define a random variable $U = u(X_1, X_2, \ldots, X_n; \theta)$ then

$$\begin{aligned} E(U) &= \sum u(x_1, x_2, \ldots, x_n; \theta) f(x_1, x_2, \ldots, x_n) \\ &= \sum \frac{\partial}{\partial\theta} f(x_1, x_2, \ldots, x_n; \theta) \\ &= \frac{\partial}{\partial\theta} \sum f(x_1, x_2, \ldots, x_n; \theta) \\ &= \frac{\partial}{\partial\theta} 1 \\ &= 0 \end{aligned}$$

Result 2

$$\text{Var}\left(\frac{\partial}{\partial\theta}\log f(X_1, X_2, \ldots, X_n;\theta)\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, \ldots, X_n;\theta)\right)$$

The proof is left as an exercise

Cramèr-Rao Lower Bound. If $\hat{\theta}$ is an unbiased estimator of θ then

$$\text{Var}(\hat{\theta}) \geq \frac{1}{-E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, \ldots, X_n;\theta)\right)}$$

Proof. Note that

$$\theta = E(\hat{\theta}) = \sum \hat{\theta}(x_1, x_2, \ldots, x_n) f(x_1, \ldots, x_n;\theta)$$

Differentiating the above equation with respect to θ gives

$$\begin{aligned} 1 &= \sum \hat{\theta}(x_1, x_2, \ldots, x_n)\frac{\partial}{\partial\theta} f(x_1, \ldots, x_n;\theta) \\ &= \sum \hat{\theta}(x_1, x_2, \ldots, x_n) u(x_1, x_2, \ldots, x_n;\theta) f(x_1, \ldots, x_n;\theta) \\ &= E(U\hat{\theta}) \\ &= \text{Cov}(U, \hat{\theta}) \end{aligned}$$

The last line follows from the fact that Cov $(U, \hat{\theta}) = E(U\hat{\theta}) - E(U)E(\hat{\theta})$. Recall from Result 1 that $E(U) = 0$.

Because the correlation coefficient is always between ± 1, it follows that

$$\text{Var}(\hat{\theta})\text{Var}(U) \geq [\text{Cov}(U, \hat{\theta})]^2.$$

Therefore

$$\text{Var}(\hat{\theta})\text{Var}(U) \geq 1$$

implying

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\text{Var}(U)}.$$

It follows from Result 2 that $\text{Var}(U) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, \ldots, X_n;\theta)\right)$.

Lower bound for the Variance of the mutation parameter θ

Suppose that we assume that every mutation that separates all individuals at a particular locus in the population is revealed, and the full ancestry is resolved. Further assume that the number of mutations between each coalescent event is observable. Define Y_j to be the number of mutations that occur during the time the sample has j distinct ancestors. Therefore, $P(Y_j = y_j)$ is the probability that y_j mutations occur before a coalescence. This is analogous to flipping an (unfair) coin and asking what is the probability of getting y_j tails before a heads. This produces the well known geometric distribution given by

$$P(Y_j = y_j) = \left(\frac{\theta}{j-1+\theta}\right)^{y_j} \left(\frac{j-1}{j-1+\theta}\right).$$

Because of independence we can write,

$$\begin{aligned} f(y_2, y_3, \ldots, y_n; \theta) &= P(Y_2 = y_2, Y_2 = y_3, \cdots, Y_n = y_n; \theta) \\ &= \prod_{j=2}^{n} P(Y_j = y_j) \\ &= \prod_{j=2}^{n} \left(\frac{\theta}{j-1+\theta}\right)^{y_j} \left(\frac{j-1}{j-1+\theta}\right) \end{aligned} \tag{403}$$

For notational convenience we will denote the likelihood by

$$L_n(\theta) = f(Y_2, Y_3, \ldots, Y_n; \theta)$$

It is easy to check that

$$\frac{\partial^2}{\partial \theta^2} \log L_n = -\frac{S_n}{\theta^2} + \sum_{j=2}^{n} \frac{Y_j + 1}{(j-1+\theta)^2}$$

so that

$$-E\left(\frac{\partial^2}{\partial\theta^2}\log L_n\right) = \frac{\sum_1^{n-1}\frac{1}{j}}{\theta^2} - \sum_{j=2}^{n}\left(\frac{\theta}{j-1}+1\right)\frac{1}{(j-1+\theta)^2}$$

$$= \frac{1}{\theta}\sum_{j=1}^{n-1}\frac{1}{j} - \sum_{j=1}^{n-1}\frac{1}{(j(j-\theta)}$$

$$= \sum_{j=1}^{n-1}\frac{1}{j-\theta}$$

Hence the variance of any unbiased estimators $\hat{\theta}$ of θ satisfies

$$\mathrm{Var}(\hat{\theta}) \geq \frac{\theta}{\sum_{j=1}^{n-1}\frac{1}{j+\theta}} \equiv \mathrm{Var}(\hat{\theta}_F)$$

Note that $\sum_{j=1}^{n-1} 1/(\theta+j) \approx \log(\theta+n)$. So as n-the number of individuals in the sample becomes large, the variance of the estimator will decrease at a very slow rate. The above Cramér-Rao lower bound on the variance shows that among unbiased estimators the best one can do is this lower bound.

This result is due to Fu and Li (1993). We will refer to the optimal estimator of Fu and Li as $\hat{\theta}_F$. The standard deviation efficiency for the Watterson's segregating sites estimator $\hat{\theta}_S$ and the Tajima's pairwise differences estimator $\hat{\theta}_T$ is given by

$$\sqrt{\frac{\mathrm{Var}(\hat{\theta}_F)}{\mathrm{Var}(\hat{\theta}_S)}}$$

and

$$\sqrt{\frac{\mathrm{Var}\hat{\theta_F}}{\mathrm{Var}\hat{\theta_T}}}$$

respectively. What follows are some plots of the standard deviation relative efficiency for the pairwise difference and segregating sites estimators.

While it is true that both the 'best' estimator and the segregating sites estimator have variance that converges to zero at rate $\log n$,

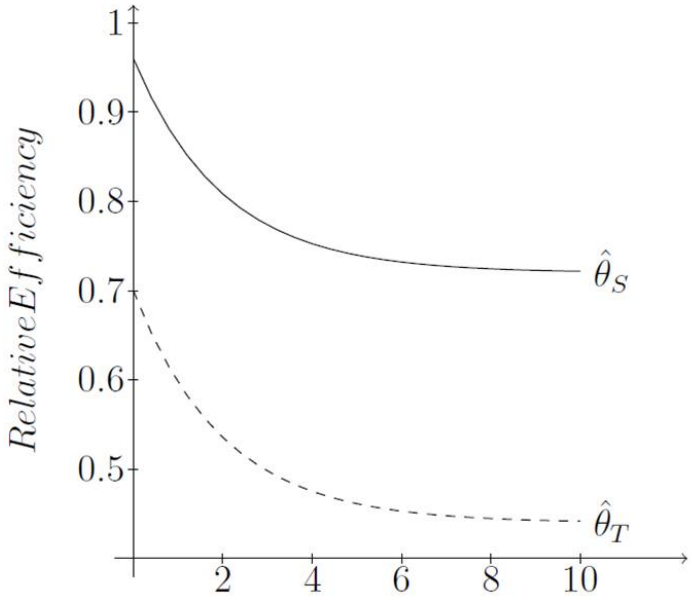

Figure 9: Relative efficiency of the pairwise differences estimator $\hat{\theta}_T$ (red), versus the relative efficiency of the segregating sites estimator $\hat{\theta}_S$ (green). $0 < \theta < 10$ and $n = 50$

the graphs in Figure 10 show that extremely large sample sizes are required before the segregating sites variance comes close to that of the optimal estimator. However, the Fu estimator is based on a likelihood (equation 403) that requires knowing the number of mutations between coalescent events. This is unobservable. To obtain a maximum likelihood estimate based on observed data, we need to consider a more computationally intensive approach. To gain some appreciation for the amount of computation required to implement a maximum likelihood approach, we begin by considering the full likelihood on a very small data set.

A numerical example using a small data set

Consider the following simple example. We have three sequences and four segregating sites and each sequence has multiplicity unity. Using the binary code discussed in the exercises we describe the data

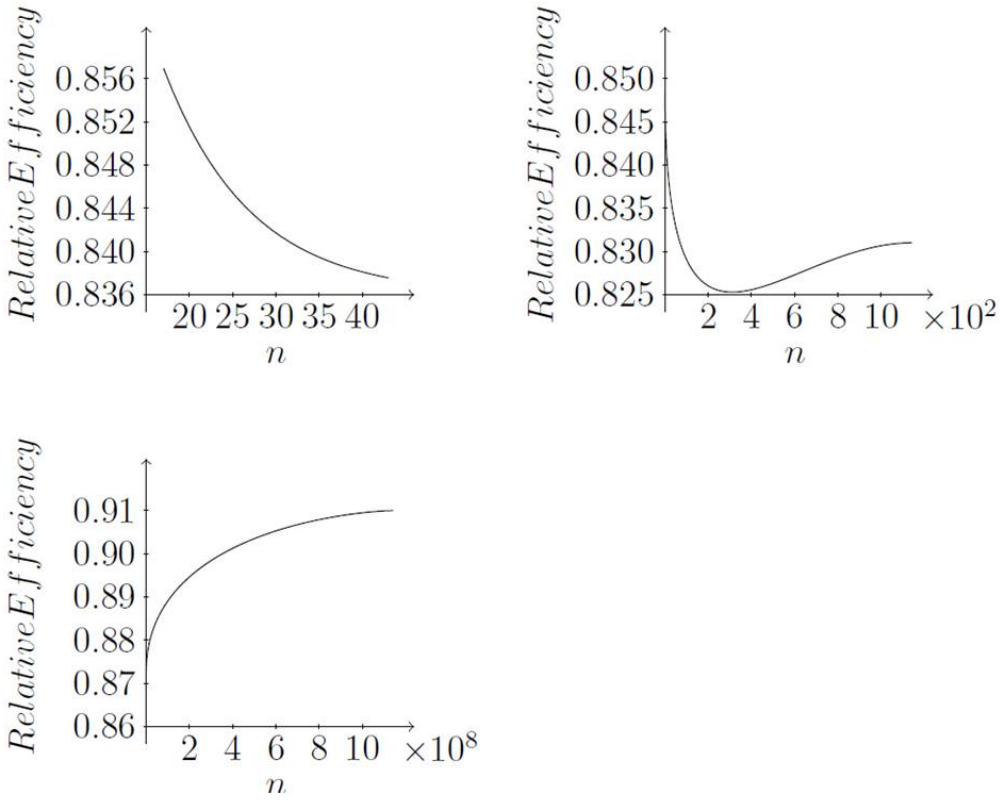

Figure 10: Relative efficiency of the segregating sites estimator $\hat{\theta}_S$ as a function of sample size. Small sample size (left top), moderate sample size (right top) and large sample size (left bottom)

as follows.

$$\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}$$

For convenience, label the segregating sites 1,2,3 and 4 from left to right. There are five possible labeled rooted trees constructed from the unrooted genealogy. These five rooted gene trees for this data are shown in Figure 11. The possible coalescent trees producing Figure 11 are given in Figure 12.

Let T_3 be the time during which the sample has three ancestors, and T_2 the time during which it has two. By considering the Poisson nature of the mutations along the edges of the coalescent tree, the

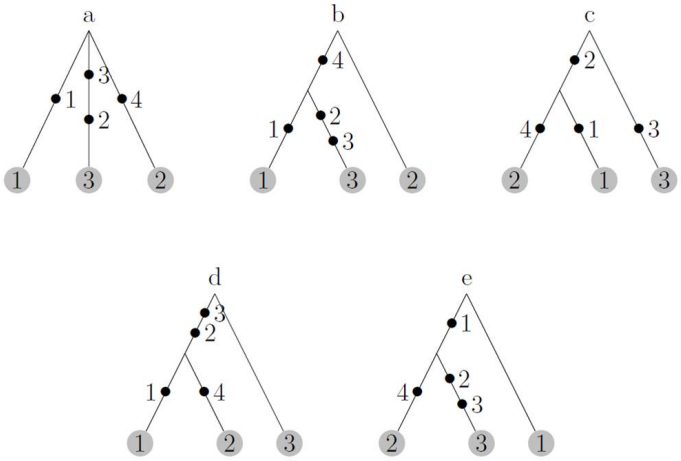

Figure 11: Gene trees consistent with the 4 segregating sites

probability of each type of tree can be calculated. For example, the probability $p_{(1a)}$ of the first labelled tree (a) is

$$\begin{aligned} p_{(a1)} &= E\left[\left(e^{-\theta T_3/2}\frac{\theta T_3}{2}\right)^2 e^{-\theta T_2/2} e^{-\theta(T_2+T_3)/2}\frac{1}{2!}(\theta(T_2+T_3)/2)^2\right] \\ &= \frac{\theta^4}{32}E\left[e^{-\theta(3T_3/2+T_2)}T_3^2(T_2+T_3)^2\right] \\ &= \frac{\theta^4(17\theta^2+46\theta+32)}{27(\theta+1)^3(\theta+2)^5} \end{aligned}$$

We must now do a similar calculation for each of the remaining five coalescent trees and sum the results. While it is indeed possible to calculate the likelihood explicitly for this extremely small data set, it is clear that a more feasible approach will be required for more realistic data sets. You can see that the number of coalescent trees consistent with the data will grow rapidly as we increase the number of sequences.

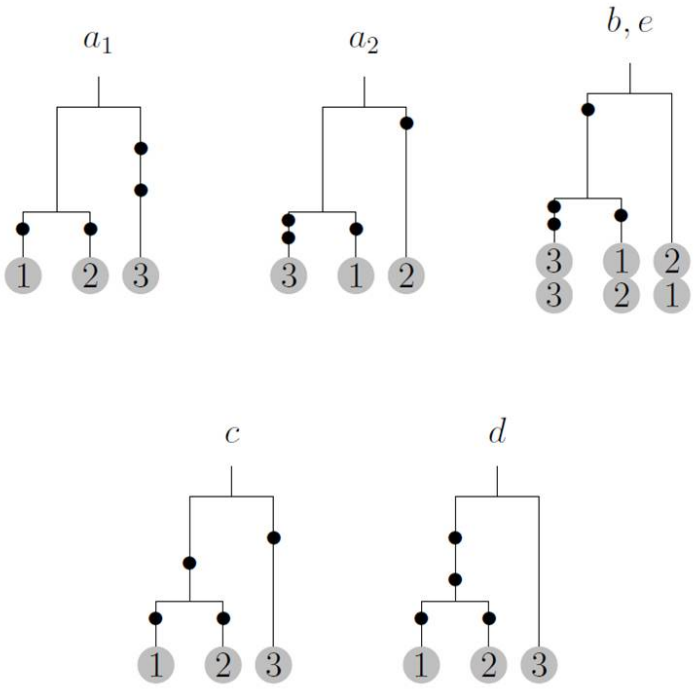

Figure 12: Coalescent trees consistent with the genetrees

Computationally intensive methods

It is not an exaggeration to say that Markov Chain Monte Carlo (MCMC) methods have revolutionized statistics and are at the heart of many computationally intensive methods. So it may be surprising to note that the most commonly used MCMC method, called the Metropolis Hastings Algorithm, is only three lines of code and the mathematical argument that justifies its legitimacy is only four lines long. In fact, the ease at which one can produce an MCMC algorithm to address a particular statistical problem can be seen as a drawback. The simplicity of the algorithm often leads individuals to try MCMC as there first method toward a solution. However, MCMC should be the algorithm of last resort. If all else fails, use MCMC. The reason for this is that the MCMC algorithm is plagued with tricky convergence issues and requires extensive diagnostics before one can reliably trust the answer. However, even with all its potential drawbacks and pitfalls, it is still an incredibly useful tool

in statistics.

Since a good deal of this course involves various types of Markov processes, it is worth pointing out the distinction between the Markov processes discussed in detail by Dr. Ewens and Markov Chain Monte Carlo methods discussed here. The typical approach to stochastic mathematical modeling is to begin with a probabilistic description of the phenomena of interest. In much of this course we are concerned with how population factors effect genetic variation over evolutionary time. Examples of mathematical descriptions that address this problem include the Moran Model, the Wright-Fisher Model and the general Cannings Model. These are all Markov models. Within the context of these models we are interested in long term behavior which often leads to a stationary distribution of the process of interest. The natural progression of ideas starts with a Markov model and from this we derive the stationary distribution.

However, MCMC reverses this logical progression and so initially may seem somewhat contrived. Rather than start with a model and then produce a stationary distribution as your final answer, in MCMC you start with the what we will call the target probability distribution and then devise a Markov chain whose stationary distribution returns you to the probability distribution you started with. This begs the question, if you know the answer to begin with, why go through the trouble of devising a Markov chain with a stationary distribution that returns you back to where you started? There are at least two good answers. 1) There is a difference between knowing the target probability distribution and being able to simulate data according to that target distribution. The MCMC algorithm is about simulating data. 2) The most important reason is in most applications you only know the target distribution up to a constant of integration. That constant of integration is often difficult to compute. If $\pi(x)$ is the target distribution, then MCMC only requires that you can write down the likelihood ratio of $\pi(x)/\pi_(y)$, where the constant of integration cancels.

For a given Markov chain there is at most one stationary distribution, but for a given stationary distribution there many Markov chains. The art of MCMC is picking the right chain. Since we get to choose the Markov chain in MCMC and the Markov chain is just a device for simulating from complex probability distributions, we might as well pick one for which it is easy to establish stationarity. A

reversible Markov chain is the simplest choice. A reversible Markov Chain with transition probabilities p_{ij} has stationary probabilities π_i if they satisfy

Detailed Balance Equations given by

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{404}$$

This means that in the long run the Markov chain visits state i followed by state j with the same probability as it visits state j followed by state i.

Metropolis Hastings Algorithm

Object Simulate a Markov chain with stationary distribution $\pi_i,\ i = 1, \ldots, m$ where m is the total number of possibilities. Typically m is quite large.

Method

1. **Propose a move** from state i to state j with probability q_{ij}.
2. **Accept the move** from i to j with probability

$$a_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

3. **Move** with probability $p_{ij} = q_{ij} a_{ij}$

Then $p_{ij} = q_{ij} a_{ij}$ are the transition probabilities having stationary distribution $\pi_1, \pi_2, \ldots, \pi_m$.

Many papers and textbooks will state that it is easy to show that the Metropolis Hastings algorithm follows the Detailed Balance Equations. However, since it is only four lines of mathematics it is worth taking the time to actually show that in fact the above algorithm does satisfy the Detailed Balance Equations. Below we do just that.

With out loss of generality assume that $a_{ij} < 1$ then $a_{ji} = 1$.

(This is the key observation). Now

$$\begin{aligned}\pi_i p_{ij} &= \pi_i q_{ij} a_{ij} \\ &= \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \\ &= \pi_j q_{ji} = \pi_j q_{ji} a_{ji} \\ &= \pi_j p_{ji}.\end{aligned}$$

Likelihood and the missing data problem

In many situations the data presents an incomplete picture because the probability of observing the data depends on unobservable random variables which we call missing data. It is often quite an easy matter to write down a likelihood function for the joint distribution of the observed data together with the missing data. To get the marginal distribution of the observed data alone you must 'average out' the missing data. This will often involve integration over a high dimensional space or a sum over a unfathomably large set of possibilities. Mathematicians realized long ago that summation and integration are really the same problem. However, that realization does not make the missing data problem any easier. Two methods for attacking the missing data problem will be presented here. They are: Markov Chain Monte Carlo (MCMC) (in particular the the Metropolis Hastings algorithm) and important sampling. Each presents different solutions to averaging out the missing data. In some sense they are more about numerical integration than they are about statistics.

General Setup

For modelling the ancestry of a sample of n individuals using the coalescent process, let D be the observed DNA sequences. What is missing is G, the true genealogy of the sample. Let θ be the mutation parameter. We will assume there is a tractable formula for the joint distribution of D and G. That is

$$P(D, G|\theta) = P(D|G, \theta)P(G|\theta)$$

where explicit formula exist for $P(D|G,\theta)$ and $P(G|\theta)$ and so

$$P(D|\theta) = \sum_G P(D|G,\theta)P(G|\theta)$$

Note that the dimension of the space is large. That is there are an enormous number of possible genealogies G in the sum.

Naive Simulation

The simplest simulation method is based on the law of large number. We begin by simulating multiple realizations of the genealogies $G_1, G_2, \cdots, G_L$ using the distribution $P(G|\theta_0)$. For a particular value θ_0. It follows from the law of large numbers that that

$$P(D|\theta_0) = E_G(P(D|G,\theta_0)) \approx \frac{1}{L}\sum_{i=1}^{L} P(D|G_i,\theta_0).$$

The main problem with this approach is that most terms in the sum are very close to zero, and in fact many of them may be identically zero. The above approach requires that one simulate coalescent trees with mutations (this is relatively easy to do) then calculate the probability of the data given that tree topology. Most tree topologies are inconsistent with the data and so $P(D|G) = 0$ for a large number of G. This suggest that in order to generate tree topologies consistent with the data we would prefer to simulate missing data according to $P(G|D,\theta)$. Suppose for a moment that this is possible and $G_1, G_2, \ldots, G_L$ are independent copies of G drawn according to the posterior distribution $P(G|D,\theta_0)$ for a particular value of θ_0. Then

$$\begin{aligned}\frac{P(D|\theta)}{P(D|\theta_0)} &= \sum_G \frac{P(G|D,\theta)P(D|\theta)}{P(G|D,\theta_0)P(D|\theta_0)} P(G|D,\theta_0) \\ &\approx \frac{1}{L}\sum_{i=1}^{L} \frac{P(G_i|D,\theta)P(D|\theta)}{P(G_i|D,\theta_0)P(D|\theta_0)} \qquad (405) \\ &= \frac{1}{L}\sum_{i=1}^{L} \frac{P(D|G_i,\theta)P(G_i|\theta)}{P(D|G_i,\theta_0)P(G_i|\theta_0)}\end{aligned}$$

Notice that the above formula is a likelihood ratio $\frac{P(D|\theta)}{P(D|\theta_0)}$ and not the likelihood itself. Since the denominator is a fixed constant that does not vary with θ. The maximum likelihood estimate using $\frac{P(D|\theta)}{P(D|\theta_0)}$ will be the same as the mle using $P(D|\theta)$.

The next thing to notice is that one needs only to simulate genealogies for a single value θ_0 to obtain a likelihood curve over a range of θ values. We call θ_0 the driving value. However, the further θ is from θ_0 the poorer the approximation in (405)

Unfortunately, it is impossible to devise a scheme to simulate independent copies of G_i but we can simulate correlated copies of G_i from the posterior distribution $P(G_i|D)$ via the Metropolis Hastings algorithm.

For the coalescent model, the states in our process are all possible coalescent trees consistent with our observed data. The stationary probabilities are given by $\pi(G) = P(G|D,\theta_0)$ Note that for any two genealogies G_1 and G_2 we have

$$\frac{\pi(G_1)}{\pi(G_2)} = \frac{P(G_1|D,\theta_0)}{P(G_2|D,\theta_0)} = \frac{P(D|G_1\theta_0)P(G_1|\theta_0)}{P(D|G_2,\theta_0)P(G_2|\theta_0)}.$$

While we do not have an explicit expression for the conditional probability $P(G|D,\theta_0)$ we do have an explicit formula for the likelihood ratio $\frac{P(G_1|D,\theta_0)}{P(G_2|D,\theta_0)}$

Important Sampling

The second approach to the problem is to simulate the missing genealogies according to a distribution that is (in some sense) close to $P(G|D,\theta)$, call this distribution $Q(G|D,\theta)$.

In this situation we simulate $G_1, G_2, \ldots, G_L$ according to the distribution $Q(G|D,\theta_0)$. Note that

$$\begin{aligned} P(D|\theta) &= E_Q\left(\frac{P(D|G,\theta)P(G|\theta)}{Q(G|D,\theta_0)}\right) \\ &= \sum_G \frac{P(D|G,\theta)P(G|\theta)}{Q(G|D,\theta_0)} Q(G|D,\theta_0) \qquad (406) \\ &\approx \frac{1}{L}\sum_{i=1}^{L} \frac{P(D|G_i,\theta)P(G_i|\theta)}{Q(G_i|D,\theta_0)} \end{aligned}$$

The above is called important sampling. The idea of the Tavaré Griffiths important sampling scheme is as follows. Starting with the observed sample, consider the most recent event that could have given rise to the current data. That event was either a coalescence or a mutation. Choose one of these according to some 'reasonable probability distribution.' Proceed one more step back into the past and pick one of the possible evolutionary events. Continue choosing until you have chosen a complete genealogical history for your data set. You have now chosen a genealogy according the the proposal distribution $Q(G|D,\theta_0)$. Repeat this process multiple times and use equation (406) to approximate the likelihood.

Exercise Estimating the time to a common ancestor conditional on the number of observed segregating sites using MCMC and the coalescent

The goal of this problem is to use a MCMC procedure to estimate the mean time to the most recent common ancestor for the Nuu-Chah-Nulth Indian Data. For simplicity we will summarize the data by using only the total number of segregating sites. We will assume that θ is known. Note that the distribution of the number of segregating sites is independent of the shape of the tree and it is only affected by the length of the tree. So the procedure can be roughly described in the following steps.

1. Start with a sequence of 'current' coalescent times.
2. Propose local changes to the coalescent times. Call these the 'proposed' coalescent times.
3. Decide whether to accept the proposed coalescent times or keep the current coalescent times by comparing the likelihood of observing the data under each using a version of MCMC called the Metropolis Hasting Algorithm.
4. Calculate $T_{mrca} = T_2 + T_3 + ...T_n$. Save this result
5. Repeat steps 2 through 4 M times and average the saved results .

Below we outline how to accomplish each part of the above procedure.

1. Starting Sequence of Coalescent times

Start with the mean coalescent times. Let $T_i^{(0)} = \frac{2}{i(i-1)}$. So $T_2^{(0)} = 1,\ T_3^{(0)} = 1/3,\ T_4^{(0)} = 1/6$ and so on. Let $L_0 = \sum iT_i^{(0)}$ be the initial length of the tree.

2. Proposed Coalescent Times

Pick a coalescent time X, where the probability that $X = i$ is $P(X = i) = iT_i/L$. Replace T_X with T'_X where T'_X is generated according to an exponential distribution with mean $2/(X(X-1))$. Define $L' = 2T_2 + 3T_3 + \cdots + XT'_X + \cdots + nT_n$ as the proposed length of the coalescent tree.

3. MCMC

If s is the observed number of segregating sites and L is the length of the tree, then s has a Poisson distribution. That is

$$p(s|L) = e^{-\frac{\theta}{2}L}\frac{((\theta/2)L)^s}{s!}.$$

If L is the current tree length and L' defined in 2. is the proposed tree length, then comparing the relatively likelihood of the data under

the two tree lengths leads to the following acceptance probability

$$A = \min\left\{1, \frac{e^{-\frac{\theta}{2}L'}(\theta L')^s(XT'_X/L')}{e^{-\frac{\theta}{2}L}(\theta L)^s(XT_X/L)}\right\} = \min\left\{1, e^{\frac{\theta}{2}(L-L')}(L'/L)^{s-1}(T'_X/T_X)\right\}$$

Write a short program to estimate the mean time to the most recent common ancestor conditional on observing 18 segregating sites for the sample of 55 sequences given in the Nuu-Chah-Nulth data set. Use the segregating sites estimate for θ that you calculated in the previous homework.

Software review

Simulation software

One of the main uses of the coalescent is as a method for efficient simulation of data-sets. As such it can be used as a tool in power studies, or for evaluating the efficiency of methods that estimate parameters from genetic data. In this section we list just some of the software available. We begin with programs that simulate the full coalescent model. However, there has been a recent trend to develop algorithms that approximate the coalescent in order to improve computational efficiency in contexts that had previously been intractable (such as for genome-wide data), so we go on to include examples of this trend. For a more full review of this field, see [11].

A nice place to start is

http://www.brics.dk/ compbio/coalescent/

This website has a number of interesting demonstrations on how the coalescent works.

Below is a list the coalescent-based simulators:

- By far the most popular coalescent simulation software is `ms`, due to Richard Hudson [24]. This allows simulation of the coalescent for a variety of differing demographic scenarios. More latterly, the software has been broadened to include recombination and gene conversion hotspots, in the form of the `msHot` software of Hellenthal & Stephens [20]. Both are available at http://home.uchicago.edu/∼ rhudson1/source/mksamples.html.

- The `SelSim` software of Spencer & Coop [64] allows for coalescent-based simulation of populations experiencing natural selection and recombination.

  Available at: http://www.stats.ox.ac.uk/mathgen/software.html.

- Users wishing to simulate more complex demographic settings might make use of `SIMCOAL 2.0`, a package due to Laval & Excoffier [32], which allows for arbitrary patterns of migration within complex demographic scenarios.

  Available at: http://cmpg.unibe.ch/software/simcoal2/.

- The `GENOMEPOP` software of Cavajal-Rodriguez [3] also allows for complex demographic scenarios, but is aimed at simulating coding regions. It is available at: http://darwin.uvigo.es/.

- In [33], Li & Stephens introduced an urn-model that approximates the coalescent. The goal is to produce data that will closely approximate that resulting from the coalescent, but at much greater computational efficiency. While no software is available, this elegant construction has been used to simulate data for power studies (*e.g.,* [8]), and forms the back-bone for data imputation schemes [59, 35].

- Another approximation to the coalescent was introduced by McVean & Cardin [38] and Marjoram & Wall [37]. Software for the latter (`FastCoal`) is available at

  http://chp200mac.hsc.edu/Marjoram/Software.html.

We now list a couple of the forward-simulation algorithms:

- `simuPOP` is a program due to Peng & Kimmel [52] that allows a good degree of flexibility via the use of user-written Python scripts. It is available at: http://simupop.sourceforge.

- The `FREGENE` software of Hoggart et al., [21] uses a re-scaling of population size to provide extremely efficient forward simulation of large data-sets. It is available at http://www.ebi.ac.uk/projects/BARGEN.

Parameter Estimation Software

One use for the coalescent is as a simulation tool (see previous section). However, it is also widely-used as the foundation for model-

based analysis, for example in parameter estimation. An early approach centered around rejection methods, where data are simulated under a variety of parameter values, and then the parameter value that generated each particular instance of those data-sets is *accepted* if the data matches that seen in an observed data-set of interest; otherwise the generating parameter is *rejected*. Taking a Bayesian perspective, the set of accepted parameter values then forms an empirical estimate of the posterior distribution of the parameter conditional on the data. However, in practical applications, the probability of simulating data identical to the observed data is vanishingly small, even if the correct parameter value is used. This has provoked a move towards so-called *Approximate Bayesian Computation*, in which the requirement for an exact match is relaxed. There has been widespread interest in this development in recent years, but here, as in most examples discussed in this section, there is little off-the-shelf software. For most applications users must write their own code!

A related methodology is that of Markov chain Monte Carlo, Metropolis-Hastings sampling. Here, at least, there is custom software in the form of the comprehensive `LAMARC` package of Felsenstein *et al.*. This is available from http://evolution .gs.washington.edu/lamarc/ and can be used to estimate a variety of population demographics parameters, such as mutation, recombination and migration rates. There are also a large number of importance sampling algorithms in existence, which again estimate a variety of population demographics parameters. A good example is the `GENETREE` software of Griffiths *et al.*, which can be found at http://www.stats.ox.ac.uk/ griff/software.html.

Bibliography

[1] Balloux, F.: EASYPOP (version 1.7): a computer program for population genetics simulations. J. Hered. **92**, 301–301 (2001)

[2] Cann, R., Stoneking, M., Wilson, A.: Mitochondrial DNA and human evolution. Nature **325**, 31–36 (1987)

[3] Carvajal-Rodriguez, A.: Genomepop: A program to simulate genomes in populations. BMC Bioinformatics **9**(1), 223 (2008)

[4] Cheverud, J.: A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**, 52–58 (2001)

[5] Cooper, G., Amos, W., Hoffman, D., Rubinsztein, D.: Network analysis of human Y microsatellite haplotypes. Hum. Mol. Genet. **5**, 1759–1766 (1996)

[6] Di Rienzo, A., Wilson, A.C.: Branching pattern in the evolutionary tree for human mitochondrial dna. Proc. Nat. Acad. Sci. **88**, 1597–1601 (1991)

[7] Dorit, R.L., Akashi, H., Gilbert, W.: Absense of polymorphism at the ZFY locus on the human Y chromosome. Science **268**, 1183–1185 (1995)

[8] Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P., Morris, A.P.: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. **75**, 35–43 (2004)

[9] Eswaran, V., Harpending, H., Rogers, A.: Genomics refutes an exclusively african origin of humans. Journal of Human Evolution **49**, 1–18 (2005)

[10] Excoffier, L.: Human demographic history: refining the recent african origin model. Current Opinion in Genetics & Development **12**, 675–682 (2002)

[11] Excoffier, L., Heckel, G.: Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet **7**(10), 745–758 (2006)

[12] Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L.: Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA **104**, 17,614–17,619 (2007)

[13] Fisher, R.A.: The Genetical Theory of Natural Selection. Clarendon Press (1930)

[14] Garrigan, D., Hammer, M.: Reconstructing human origins in the genomic era. Nat Rev Genet **7**, 669–680 (2006)

[15] Garrigan, D., Hammer, M.: Ancient lineages in the genome: A response to fagundes et al. Proceedings of the National Academy of Sciences **105**, E3 (2008)

[16] Green, R., Krause, J., Ptak, S., Briggs, A., Ronan, M.: Analysis of one million base pairs of neanderthal dna. Nature **444**, 330–336 (2006)

[17] Griffiths, R.C., Marjoram, P.: An ancestral recombination graph. In: P. Donnelly, S. Tavaré (eds.) Progress in Population Genetics and Human Evolution, *IMA Volumes in Mathematics and its Applications*, vol. 87, pp. 100–117. Springer Verlag (1997)

[18] Hammer, M.: A recent common ancestry for the human Y chromosome. Nature **378**, 376–378 (1995)

[19] Hein, J., Schierup, M.H., Wiuf, C.: Gene Genealogies, Variation and Evolution. Oxford University Press, Oxford (2005)

[20] Hellenthal, G., Stephens, M.: mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23**, 520–521 (2007)

[21] Hoggart, C.J., Chadeau-Hyam, M., Clark, T.G., Lampariello, R., Whittaker, J.C., Iorio, M.D., Balding, D.J.: Sequence-level population simulations over large genomic regions. Genetics **177**, 1725–1731 (2007)

[22] Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. Theor. Popn. Biol. **23**, 183–201 (1983)

[23] Hudson, R.R.: Gene genealogies and the coalescent process. In: D. Futuyma, J. Antonovics (eds.) Oxford Surveys in Evolutionary Biology, vol. 7, pp. 1–44 (1990)

[24] Hudson, R.R.: Generating samples under a Wright-Fisher neutral model. Bioinformatics **18**, 337–338 (2002)

[25] Hudson, R.R., Kaplan, N.L.: The coalescent process in models with selection and recombination. Genetics **120**, 831–840 (1988)

[26] Huentelman, M., Craig, D., Shieh, A., Corneveaux, J.: Sniper: improved snp genotype calling for affymetrix 10k genechip microarray data. BMC Genomics **6**, 149 (2005)

[27] Jobling, M., Tyler-Smith, C.: Fathers and sons: the Y chromosome and human evolution. Trends in Genetics **11**, 449–456 (1995)

[28] Kingman, J.F.C.: The coalescent. Stoch. Proc. Applns. **13**, 235–248 (1982)

[29] Kingman, J.F.C.: Exchangeability and the evolution of large populations. In: G. Koch, F. Spizzichino (eds.) Exchangeability in probability and statistics, pp. 97–112. North-Holland Publishing Company (1982)

[30] Kingman, J.F.C.: On the genealogy of large populations. J. Appl. Prob. **19A**, 27–43 (1982)

[31] Krone, S.M., Neuhauser, C.: Ancestral processes with selection. Theor. Popn. Biol. **51**, 210–237 (1997)

[32] Laval, G., Excoffier, L.: Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided .... Bioinformatics **20**, 2485–2487 (2004)

[33] Li, N., Stephens, M.: Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics **165**, 2213–2233 (2003)

[34] Liang, L., Zollner, S., Abecasis, G.R.: Genome: a rapid coalescent-based whole genome simulator. Bioinformatics **23**, 1565–1567 (2007)

[35] Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet **39**, 906–913 (2007)

[36] Marjoram, P., Donnelly, P.: Pairwise comparison of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**, 673–683 (1994)

[37] Marjoram, P., Wall, J.D.: Fast "coalescent" simulation. BMC Genetics **7:16** (2006)

[38] McVean, G.A.T., Cardin, N.J.: Approximating the coalescent with recombination. Phil. Trans. R. Soc. B **360**, 1387–1393 (2005)

[39] Minichiello, M., Durbin, R.: Mapping trait loci by use of inferred ancestral recombination graphs. The American Journal of Human Genetics (2006)

[40] Molitor, J., Marjoram, P., Thomas, D.: Application of Bayesian clustering via Voronoi tesselations to the analysis of haplotype risk and gene mapping. Am. J. Hum. Genet. **73**, 1368–1384 (2003)

[41] Molitor, J., Marjoram, P., Thomas, D.: Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. Gen. Epi. **25(2)**, 95–105 (2003)

[42] Morris, A.P., Whittaker, J.C., Balding, D.J.: Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. **70**, 686–707 (2002)

[43] Moskvina, V., Schmidt, K.M.: On multiple-testing correction in genome-wide association studies. Genet Epidemiol **32**(6), 567–573 (2008)

[44] Navarro, A., Barton, N.H.: The effects of multilocus balancing selection on neutral variability. Genetics **161**(2), 849–63 (2002)

[45] Neuhauser, C., Krone, S.M.: The genealogy of samples in models with selection. Genetics **145**, 519–534 (1997)

[46] Noonan, J., Coop, G., Kudaravalli, S., Smith, D.: Sequencing and Analysis of Neanderthal Genomic DNA. Science **314**, 1113–1118 (2006)

[47] Nordborg, M.: Coalescent theory. In: D.J. Balding, M.J. Bishop, C. Cannings (eds.) Handbook of Statistical Genetics, pp. 179–208. John Wiley & Sons, Inc., New York (2001)

[48] Nordborg, M., Innan, H.: The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics **163**, 1201–1213 (2003)

[49] Nyholt, D.: A simple correction for multiple testing for SNPs in linkage disequilibrium with each other. Am J Hum Genet **74**, 765–769 (2004)

[50] Nyholt, D.: Evaluation of Nyholt's procedure for multiple testing correction - author's reply. Hum Hered **60**, 61–62 (2005)

[51] Padhukasahasram, B., Marjoram, P., Wall, J.D., Bustamante, C.D., Nordborg, M.: Exploring population genetic models with recombination using efficient forward-time simulations. Genetics **178**(4), 2417–2427 (2008)

[52] Peng, B., Kimmel, M.: simuPOP: a forward-time population genetics simulation environment. Bioinformatics **21**(3686-3687) (2005)

[53] Plagnol, V., Wall, J.: Possible ancestral structure in human populations. PLoS Genet **2**(e105) (2006)

[54] Portin, P.: Evolution of man in the light of molecular genetics: a review. part i. our evolutionary history and genomics. Hereditas **144**, 80–95 (2007)

[55] Portin, P.: Evolution of man in the light of molecular genetics: a review. part ii. regulation of gene function, evolution of speech and of brains. Hereditas **145**, 113–125 (2008)

[56] Relethford, J.H.: Genetic evidence and the modern human origins debate. Heredity **100**(6), 555–563 (2008)

[57] Salyakina, D., Seaman, S.R., Browning, B.L., Dudbridge, F., Müller-Myhsok, B.: Evaluation of nyholt's procedure for multiple testing correction. Hum Hered **60**, 19–25 (2005)

[58] Saunders, I.W., Tavaré, S., Watterson, G.A.: On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16**, 471–491 (1984)

[59] Servin, B., Stephens, M.: Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet **3**, e114 (2007)

[60] Slade, P.F.: Simulation of 'hitch-hiking' genealogies. J. Math. Biol. **42**, 41–70 (2001)

[61] Slade, P.F.: The structured ancestral selection graph and the many-demes limit. Genetics **169**(2), 1117–1131 (2005)

[62] Slatkin, M.: Simulating genealogies of selected alleles in a population of variable size. Genetics Research **78**, 49–57 (2001)

[63] Slatkin, M., Hudson, R.R.: Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**, 555–562 (1991)

[64] Spencer, C.C.A.: Selsim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20**(18), 3673–3675 (2004)

[65] Stringer, C., Andrews, P.: Genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origina of modern humans. Science **239**, 1263–1268 (1988)

[66] Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P.: Inferring coalescence times for molecular sequence data. Genetics **145**, 505–518 (1997)

[67] Templeton, A.: Genetics and recent human evolution. Evolution **61**, 1507–1519 (2007)

[68] Templeton, A.R.: Haplotype trees and modern human origins. Yrbk Phys Anthropol **48**, 33–59 (2005)

[69] Templeton, A.R., Maxwell, T., Posada, D., Stengard, J.H., Boerwinkle, E., Sing, C.F.: Tree scanning: A method for using haplotype trees in phenotype/genotype association studies. Genetics **169**, 441–453 (2005)

[70] The International HapMap Consortium: A haplotype map of the human genome. Nature **437**, 1299–1320 (2005)

[71] Toleno, D., Morrell, P., Clegg, M.: Error detection in snp data by considering the likelihood of recombinational history implied by .... Bioinformatics **23**, 1807–1814 (2007)

[72] Waldron, E.R.B., Whittaker, J.C., Balding, D.J.: Fine mapping of disease genes via haplotype clustering. Genet. Epi. **30** (2006)

[73] Wallace, D.: 1994 Willliam Alan Award Address - Mitochondrial DNA variation in human evolution, degenerative disease, and aging. Am. J. Hum. Genet. **57**, 201–223 (1995)

[74] Whitfield, L.S., Sulston, J.E., Goodfellow, P.N.: Sequence variation of the human y chromosome. Nature **378**, 379–380 (1995)

[75] Wills, C.: When did Eve live? An evolutionary detective story. Evolution **49**, 593–607 (1995)

[76] Wiuf, C., Hein, J.: The ancestry of a sample of sequences subject to recombination. Genetics **151**, 1217–1228 (1999)

[77] Wiuf, C., Hein, J.: Recombination as a point process along sequences. Theor. Popul. Biol. **55**, 248–259 (1999)

[78] Wright, S.: Evolution in mendelian populations. Genetics **16**, 97–159 (1931)