Mathematical Population Genetics:

Lecture Notes

Cornell University,

June – July, 2006

Warren J Ewens

Preface

These notes should, ideally, be read before the Cornell 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 27 pages should be read before the meeting.

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,polymorphism, linkage, recombination.

Introduction

The historical background

Population genetics is the subject growing out of the amalgamation of the Darwinian theory of evolution and the Mendelian hereditary system. It is crucial to many current areas of science.

Darwin had no idea of the heredity mechanism, other than the vague concept that "children tend to be like their parents". It was thus very risky of him to put forward his evolutionary ideas in the absence of having this knowledge. Fortunately his instincts were so good that he made very few errors when he did this.

Perhaps the central theme in population genetics theory is the examination of the change in the genetic make-up of a population as time goes on as a result of selection, mutation, and similar factors. These notes are based on such an examination. However, they focus on the *random* changes in the genetic make-up of a population, due essentially to random sampling effects. The random choice of one of two genes to be passed on from parent to child, and the random events during life that ensure that two equally fit individuals do not necessarily have the same number of offspring, make this random, or stochastic, aspect of the theory an important component of population genetics theory.

Clearly selection is a major factor in the Darwinian theory. However, in these notes we assume that there are no selective differences between the different genotypes in a population - that is, we assume

selective neutrality throughout. On the other hand, much of the discussion in these notes relates to mutation.

Although these notes describe stochastic processes in evolutionary genetics, it is appropriate to start with a deterministic result that ignores the possibility of random changes in gene frequencies, since it is so important.

The Hardy–Weinberg law

We consider a random-mating diploid population in which there is no concept of two separate sexes (that is, a monoecious population) which is so large that genotype frequency changes may be treated as deterministic, and focus attention on some gene locus "A", at which two alleles may occur, namely A_1 and A_2. Suppose that in any generation the proportions of the three possible genotypes at this locus, namely A_1A_1, A_1A_2 and A_2A_2, are X, $2Y$, and Z, respectively. Since random mating obtains, the frequency of matings of the type $A_1A_1 \times A_1A_1$ is X^2, that of $A_1A_1 \times A_1A_2$ is $4XY$, 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 X^2, $4XY$, $4Y^2$, the frequency X' of A_1A_1 in the following generation is

$$X' = X^2 + \tfrac{1}{2}(4XY) + \tfrac{1}{4}(4Y^2) = (X+Y)^2. \tag{1}$$

Similar considerations give the frequencies $2Y'$ of A_1A_2 and Z' of A_2A_2 as

$$\begin{aligned} 2Y' &= \tfrac{1}{2}(4XY) + \tfrac{1}{2}(4Y^2) + 2XZ + \tfrac{1}{2}(4YZ) \\ &= 2(X+Y)(Y+Z), \qquad (2) \\ Z' &= \tfrac{1}{4}(4Y^2) + \tfrac{1}{2}(4YZ) + Z^2 = (Y+Z)^2. \qquad (3) \end{aligned}$$

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

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

and similarly it is found that $Y'' = Y'$, $Z'' = Z'$. 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

$$(Y')^2 = X'Z'. \tag{4}$$

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

$$Y^2 = XZ, \tag{5}$$

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 $X + 2Y + Z = 1$, only two of the frequencies X, $2Y$ and Z 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 of the allele A_1.

This is an important result, since it shows that for diploid populations such as those discussed above, it is sufficient to focus on allelic frequencies for much of the analysis. In doing this we will follow the conventions of population genetics and refer to these, somewhat illogically, as gene frequencies.

These conclusions may be summarized in the form of a theorem:

Theorem (Hardy–Weinberg). Under the assumptions stated, a population having genotypic frequencies X (of A_1A_1), $2Y$ (of A_1A_2) and Z (of A_2A_2) achieves, after one generation of random mating, stable genotypic frequencies x^2, $2x(1-x)$, $(1-x)^2$ where $x = X + Y$ and $1 - x = Y + Z$. If the initial frequencies X, $2Y$, Z are already of the

form x^2, $2x(1-x)$, $(1-x)^2$, then these frequencies are stable for all generations. The frequency x of the allele A_1 remains unchanged is *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 result is of great importance for the Darwinian theory, but we do not dwell on it now.

The Hardy-Weinberg law can be generalized in various ways, in particular to the case where more than two alleles are possible (for example in the ABO blood group system), but we do not consider these generalizations here.

As stated in the above, the Hardy-Weinberg law assumes that the population considered is infinite in size, so that that random, or stochastic, changes in gene (more strictly, allelic) frequencies are not allowed. However, all population sizes are, of course, finite, and thus the stochastic aspect of evolutionary population genetics must be investigated. **From now on, these notes focus entirely on stochastic processes in evolutionary genetics.**

The stochastic theory: Two alleles

The "simple" Wright-Fisher model

Basic theory

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 (1922, 1930, 1958) and Wright (1931), in developing the model considered in this section, are in fact those of this theory and its close relative, diffusion theory. Here we present some of the conclusions of Fisher and Wright in the terminology of Markov chains.

We consider, as the simplest possible case, a population of fixed size N. Suppose that the individuals in this population are monoecious, that no selective difference exist between the genotypes possible at the gene locus "A" under consideration, and that there is no mutation. Since each individual carries two genes at this locus (one maternally, the other paternally, derived) 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 model discussed below, although it was not written down explicitly by Fisher and Wright, was clearly 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, given that $X(t) = i$, the probability p_{ij} that $X(t+1) = j$ is assumed to be given by

$$p_{ij} = \binom{2N}{j} (i/2N)^j \{1 - (i/2N)\}^{2N-j}, \quad i, j = 0, 1, 2, \ldots, 2N. \quad (6)$$

We refer to the model (6) 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 nevertheless share the binomial sampling characteristic of (6), will all be referred to generically as "Wright–Fisher" models.

In the form of (6), 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, (6) 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 $X(\cdot)$. These are absorbing states of the Markov chain (6). Genetically this corresponds, of course, to the fact that since the model (6) 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 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.

It is easy enough to see that 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 the solution $\pi_j = j/(2N)$ satisfies the standard Markov chain fixation probability difference equations and the appropriate boundary conditions. Setting $j = X(0)$ 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

$$E\{X(t+1)/2N \mid X(t)\} = X(t)/2N, \tag{7}$$

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 fixation occurs. The most obvious quantity to evaluate

is the mean time $\bar{t}\{X(0)\}$ taken until $X(\cdot)$ reaches 0 or $2N$, starting from $X(0)$. As it happens, no simple explicit formula for this mean time is known, although a simple approximation, given later, is available. Fisher and Wright, no doubt noting this difficulty, paid comparatively little attention to the mean fixation time, concentrating on an approach centering around the leading nonunit eigenvalue of P. It follows immediately from (6) that if we put $x(t) = X(t)/2N$,

$$\mathrm{E}\big(x(t+1)\{1-x(t+1)\} \mid x(t)\big) = \{1-(2N)^{-1}\}x(t)\{1-x(t)\}, \quad (8)$$

so that the expected value of the heterozygosity measure $2x(\cdot)\{1-x(\cdot)\}$ decreases by a factor of $1-(2N)^{-1}$ each generation. It follows immediately that $1-(2N)^{-1}$ is an eigenvalue of the matrix P, and it is easy to show that it is the leading nonunit eigenvalue. We write the right and left eigenvectors corresponding to this eigenvalue as $\mathbf{r} = (r_0, r_1, r_2, \ldots, r_{2N})$, and $\boldsymbol{\ell}' = (\ell_0, \ell_1, \ell_2, \ldots, \ell_{2N})$ respectively. It follows from (8) that $\mathbf{r}'$ is proportional to the vector

$$\{0,\ 2N-1,\ 2(2N-2),\ 3(2N-3),\ \ldots,\ 2N-1,\ 0\}. \quad (9)$$

Unfortunately, no such simple formula is known for the left eigenvector $\boldsymbol{\ell}$. If we suppose that ℓ and $\mathbf{r}$ are normalized by the requirements

$$\sum_{k=1}^{2N-1} \ell_k = 1, \quad \sum_{k=0}^{2N} \ell_k r_k = 1, \quad (10)$$

then standard spectral theory shows that

$$\begin{aligned} p_{ij}(t) &= \mathrm{Prob}\{X(t) = j \mid X(0) = i\} \\ &= r_i \ell_j \{1-(2N)^{-1}\}^t + o\{1-(2N)^{-1}\}^t \ \text{ for } t \text{ large.} \end{aligned} \quad (11)$$

Equations (8) and (11) jointly provide much interesting information. It is clear that especially in a large population, the mean heterozygosity of the population decreases extremely slowly with time as a result of the random sampling effects implicit in the model under consideration. We conclude that although genetic variation must ultimately be lost under the model (6), the loss is usually very slow. This slow rate of loss may be thought of as a stochastic analogue of the "variation-preserving" property of infinite populations shown by the Hardy–Weinberg law. This conclusion can be generalized, taking into account complications brought about through variation in the population size, through geographical factors, through the existence of two sexes, and so on.

An ergodic argument

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 (6) applies. How much time will pass before the mutant is lost (probability $1 - (2N)^{-1}$) or fixed (probability $(2N)^{-1}$)? The mean number of generations $\bar{t}_1$ for one or other of these events may be written in the form

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j}, \tag{12}$$

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

$$\bar{t}_{1,j} \approx 2j^{-1}, \quad j = 1, 2, \ldots, 2N-1, \tag{13}$$

so that, using (12),

$$\bar{t}_1 \approx 2\big(\log(2N-1) + \gamma\big), \tag{14}$$

where γ is Euler's constant 0.5772 ....

There is an ergodic equivalent to the expressions in (12) and (13) which is perhaps of more interest than (12) and (13) 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 (6) 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 is of great interest to geneticists, and will be taken up later at much greater length.

Conditional processes

Consider now only those cases for which the number of A_1 genes eventually takes the value $2N$. What is the transition matrix of the

conditional process when the condition is made that eventually this happens?

We assume that the initial value of $X(\cdot) = i$. Recalling that we write $X(t)$ for the number of A_1 genes in generation t, we get

$$\begin{aligned} p^*_{ij} &= \text{Prob}\{X(t+1) = j \mid X(t) = i \text{ and eventually} X(\cdot) = 2N\} \\ &= \text{Prob}\{X(t+1) = j \text{ and eventually} X(\cdot) = 2N \mid X(t) = i\} \\ &\quad \div \text{Prob}\{\text{eventually} X(\cdot) = 2N \mid X(t) = i\} \\ &= p_{ij} j/i, \qquad (i, j = 1, 2, \ldots, 2N). \end{aligned} \tag{15}$$

Here p_{ij} is the (i, j) term in the Wright-Fisher transition matrix (6) and we have used the fact that when $X(\cdot) = i$, the probability that eventually $X(\cdot) = 2N$ is $i/(2N)$, as well as standard conditional probability arguments and the Markovian nature of $X(\cdot)$. Let $\widetilde{P}$ be the matrix derived from the Wright-Fisher transition matrix P by omitting the first row and first column and let

$$V = \begin{pmatrix} \pi_1 & & & \\ & \pi_2 & & \\ & & \ddots & 0 \\ 0 & & & \pi_{2N} \end{pmatrix}. \tag{16}$$

Then if $P^* = \{p^*_{ij}\}$, Eq. (15) shows that

$$P^* = V^{-1} \widetilde{P} V. \tag{17}$$

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

$$p^{*(n)}_{ij} = p^{(n)}_{ij} \pi_j / \pi_i, \tag{18}$$

a conclusion that can be reached directly. If $\bar{t}^*_{ij}$ is the conditional mean time, measured in generations, that $X(\cdot) = j$, given initially

$X(0) = i$, then

$$\begin{aligned} \bar{t}^*_{ij} &= \sum_{n=0}^{\infty} p_{ij}^{*(n)} \\ &= (j/i) \sum_{n=0}^{\infty} p_{ij}^{(n)} \\ &= \bar{t}_{ij} j/i. \end{aligned} \tag{19}$$

Further theoretical results

In this section we consider the mean time $\bar{t}_i$ until an absorbing state ($X(\cdot) = 0$ or $2N$) of the Markov chain describing the simple Wright-Fisher model, given that initially the number $X(0)$ of A_1 genes is i, and will also consider the mean number of times $\bar{t}_{ij}$ that $X(\cdot)$ takes the value j before an absorbing state is reached. While in principle these expressions can be found from standard theory, in practice solution of the equations that arise seems extremely difficult, and simple expressions for these mean times have not yet been found. On the other hand, it is possible to find a simple approximation for $\bar{t}_i$ by the following line of argument.

We put $i/M = x$, $j/M = x + \delta x$, and $\bar{t}_i = \bar{t}(x)$, and suppose $\bar{t}(x)$ is a twice differentiable function of a continuous variable x. Then from standard theory,

$$\bar{t}(x) = \sum \mathrm{Prob}\{x \to x + \delta x\} \bar{t}(x + \delta x) + 1 \tag{20}$$

$$= \mathrm{E}\{\bar{t}(x + \delta x)\} + 1 \tag{21}$$

$$\approx \bar{t}(x) + \mathrm{E}(\delta x)\{\bar{t}(x)\}' + \tfrac{1}{2}\mathrm{E}(\delta x)^2\{\bar{t}(x)\}'' + 1. \tag{22}$$

In the expression (22) the random variable is δx, all expectations are conditional on x and only the first three terms in an infinite Taylor series have been retained. Since from (6)

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

Eq. (22) gives

$$x(1 - x)\{\bar{t}(x)\}'' \approx -4N. \tag{23}$$

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

$$\bar{t}(p) \approx -4N\{p \log p + (1 - p) \log(1 - p)\}, \tag{24}$$

where $p = i/2N$ is the initial frequency of A_1. This can be shown to be a very accurate approximation to the true, but to this day unknown, mean fixation time.

(The value given in (24), and various other approximations given in these notes, are in effect "diffusion approximations", that is approximate expressions found by approximating a Markov chain by a diffusion process. Any reference to approximations in these notes refer to such approximations. Diffusion theory will be discussed in these lectures by Dr Griffiths.)

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, Eq. (24) reduces to

$$\bar{t}\{(2N)^{-1}\} \approx 2 + 2\log 2N \text{ generations}, \tag{25}$$

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

$$\bar{t}\left\{\tfrac{1}{2}\right\} \approx 2.8N \text{ generations}. \tag{26}$$

The value given in (25) is very close to Wright's and Fisher's approximation given in (14).

Suppose now the condition is made that A_1 eventually fixes. The possible values for X are $1, 2, 3, \ldots, 2N$ and Eq. (15) shows that the conditional transition probability p^*_{ij} is

$$\begin{aligned} p^*_{ij} &= \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(\frac{2N-i}{2N}\right)^{2N-j}\frac{j}{i} \\ &= \binom{2N-1}{j-1}\left(\frac{i}{2N}\right)^{j-1}\left(\frac{2N-i}{2N}\right)^{2N-j}. \end{aligned} \tag{27}$$

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. An argument parallel to that leading to (23) gives

$$(1-x)\{\bar{t}^*(x)\}' + \frac{1}{2}x(1-x)\{\bar{t}^*(x)\}'' = -2N \tag{28}$$

for the conditional mean time $\bar{t}^*(x)$ to fixation, given a current frequency of x. The solution of (28), subject to $\bar{t}^*(1) = 0$ and the requirement

$$\lim_{x\to 0}\bar{t}^*(x) \text{ is finite}, \tag{29}$$

and assuming initially $x = p$, is

$$\bar{t}^*(p) = -4Np^{-1}(1-p)\log(1-p). \tag{30}$$

We observe from this that

$$\bar{t}^*\{(2N)^{-1}\} \approx 4N - 2 \text{ generations}, \tag{31}$$
$$\bar{t}^*\left\{\tfrac{1}{2}\right\} \approx 2.8N \text{ generations}, \tag{32}$$
$$\bar{t}^*\{1-(2N)^{-1}\} \approx 2\log 2N \text{ generations}. \tag{33}$$

A result equivalent to (31) is that, given initially $2N$ -1 A_1 genes, the conditional mean number of generations until, loss of A_1, given that such loss will occur, is approximately

$$4N - 2 \text{ generations}. \tag{34}$$

Eq. (32) is to be expected from Eq. (26), 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, Eq. (31) and Eq. (33) provide new information and show 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.

As noted earlier, the initial analysis of the model (6) by Fisher and Wright paid particular attention to the leading eigenvalue of the transition matrix, regarded as a measure of the rate at which one or other allele is lost from the population. Although the eigenvalues of the transition matrix in (6) are of less use than expressions like (24) and (30) for investigating the length of tome that both alleles may be expected to remain in the population , they are nevertheless of some interest, so we now write down the formulae for these eigenvalues.

Since the matrix defined by the p_{ij} in (6) is the transition matrix of a Markov chain, it follows that one eigenvalue of the matrix is automatically 1. Denoting this eigenvalue by λ_0, the remaining eigenvalues, first derived by Feller (1951), are

$$\lambda_j = (2N)(2N-1)\ldots(2N-j+1)/(2N)^j, \quad j = 1, 2, \ldots, 2N. \tag{35}$$

This confirms the values $\lambda_1 = 1$ and $\lambda_2 = 1-(2N)^{-1}$ found earlier by other methods. We derive the eigenvalues in (35) later as particular cases of those for the Cannings model.

Mutation in the Wright-Fisher model

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 (6) by

$$p_{ij} = \binom{2N}{j} (\psi_i)^j (1-\psi_i)^{2N-j} \tag{36}$$

where $\psi_i = i(1-u)/2N$. Here it is certain that A_1 will be eventually lost from the population, and interest centers on properties of the time until this loss occurs, either using eigenvalues or mean time properties. An argument parallel to that leading to (23) shows that, to a first approximation, the mean time $\bar{t}(x)$ for the loss of A_1, given a current frequency x, satisfies

$$-4Nux\{\bar{t}(x)\}' + x(1-x)\{\bar{t}(x)\}'' = -4N. \tag{37}$$

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

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

is

$$\bar{t}(p) = \int_0^1 t(x,p)\,dx \text{ generations,} \tag{38}$$

where

$$\left.\begin{array}{lcll} t(x,p) & = & 4Nx^{-1}(1-\theta)^{-1}\{(1-x)^{\theta-1}-1\}, & 0 \le x \le p, \\ t(x,p) & = & 4Nx^{-1}(1-\theta)^{-1}(1-x)^{\theta-1}\{1-(1-p)^{1-\theta}\}, & p \le x \le 1, \end{array}\right\} \tag{39}$$

and $\theta = 4Nu$. (Formulae for the case $\theta = 1$ are found from (39) by standard limiting processes.)

It may be shown that with the definition of $t(x,p)$ in (39), $\bar{t}(p)$ may be written as

$$\bar{t}(p) = \sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}\Big(1-(1-p)^j\Big). \tag{40}$$

The function $t(x,p)$ in (39) is more informative than it initially appears since, as is shown 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.

There are two interesting special cases of (39). First, when $\theta = 2$,

$$\left.\begin{array}{lcll} t(x,p) & = & 4N, & 0 \le x \le p, \\ t(x,p) & = & 4Nx^{-1}(1-x)\{(1-p)^{-1}-1\}, & p \le x \le 1, \end{array}\right\} \quad (41)$$

and from this,

$$\bar{t}(p) = \frac{-4Np\log p}{1-p}, \quad (42)$$

a conclusion that can also be found directly from (40). Second, when $p = 1$, Eq. (40) gives immediately

$$\bar{t}(1) = \sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}. \quad (43)$$

A slightly more accurate approximation is

$$4N \sum_{j=1}^{2N} \{j(j+\theta-1)\}^{-1} \text{ generations.} \quad (44)$$

The case $\theta = 2$ is of some interest. For this value of θ the expression in (44) reduces to

$$4N - 2 \quad (45)$$

generations. This is identical to the conditional mean loss time, given initially $2N-1$ genes of the allele A_1, as given in (34). 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 $\theta = 2$ is that the entire properties of the two processes are, perhaps surprisingly, essentially 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 (36) by

$$\psi_i = \{i(1-u) + (2N-i)v\}/2N. \quad (46)$$

There now exists a stationary distribution $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_{2N})$ for the number of A_1 genes. No simple explicit form of this distribution is known. On the other hand, certain properties of this distribution can be extracted immediately from (36) and (46). The stationary distribution satisfies the equation $\boldsymbol{\phi}' = \boldsymbol{\phi}' P$, where P is defined by (36) and (46), 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

$$\mu = 2Nv/(u+v). \tag{47}$$

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

$$\sigma^2 = 4N^2uv/\{(u+v)^2(4Nu+4Nv+1)\} + \text{small order terms.} \tag{48}$$

The above values are sufficient to answer the question: "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 with respect to the stationary distribution, namely

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

If $u = v$, $4Nu = \theta$, Eq. (47) and Eq. (48) show that $\mathrm{E}(x) = \frac{1}{2}$ and, to a close approximation, $\mathrm{E}(x^2) = \frac{1}{4} + \frac{1}{4(2\theta+1)}$. Thus

$$\text{Prob (two genes of same allelic type)} \approx (1+\theta)/(1+2\theta). \quad (49)$$

This probability 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

$$F = (1+\theta)/(1+2\theta), \quad (50)$$

in agreement with (49).

The Cannings (Exchangeable) Model

No mutation

An important generalization of the Wright–Fisher model was introduced by Cannings (1974). As with the Wright-Fisher model, this model considers 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, assuming only that any subset of genes at time t has the same distribution of "descendant" genes at time $t+1$ as any other subset of the same size. More precisely, if the ith gene leaves y_i descendant genes, it

is required only that $y_1 + \ldots + y_{2N} = 2N$ and that the distribution of $(y_i, y_j, \ldots, y_k)$ be independent of $\{i, j, \ldots, k\}$. In particular all genes must have the same offspring probability distribution. This distribution must have mean 1 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 descendants. An example of this is discussed later.

Our first calculation concerning the Cannings model relates to eigenvalues. Let the genes be divided into two allelic classes, A_1 and A_2, and let X_t be the number of A_1 genes at time t. Then Cannings' first theorem is as follows:-

Theorem (Cannings). (No proof). 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

$$\lambda_0 = 1, \quad \lambda_j = \text{E}(y_1 y_2 \ldots y_j), \quad j = 1, 2, \ldots, 2N. \tag{51}$$

The Wright–Fisher model (6) is clearly a particular case of the Cannings model, since in the Wright-Fisher model (6) $(y_1, y_2, \ldots, y_{2N})$ have a symmetric multinomial distribution. (However, the Cannings model is far more general, and thus realistic, than the Wright–Fisher model.) 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 $\lambda_j, j = 1, 2, \ldots, 2N$ for the simple Wright-Fisher model is given by

$$\begin{aligned} \lambda_j &= \sum \ldots \sum y_1 y_2 \ldots y_j \binom{n}{\mathbf{y}} \left(\frac{1}{2N}\right)^{\sum y_i} \left(1 - \frac{j}{2N}\right)^{2N - \sum y_i} \\ &= (2N)(2N-1)\ldots(2N-j+1)/(2N)^j. \end{aligned} \tag{52}$$

This confirms the values given in (35), found originally by Feller (1951), using other methods.

The theorem shows that, for the Cannings model, the leading non-unit eigenvalue is $\lambda_2 = \text{E}(y_1 y_2)$ where y_i is the number of descendent genes of the ith gene in the population. Now $\sum y_j \equiv 2N$,

so that the variance of $(\sum y_j)$ is 0. Then by symmetry,

$$2N \text{ var}(y_i) + 2N(2N-1) \text{ covar}(y_i, y_j) = 0.$$

This implies that

$$\text{covar}(y_i, y_j) = -\sigma^2/(2N-1), \tag{53}$$

where $\sigma^2 = \text{var}(y_i)$. Immediately then,

$$\begin{aligned} \lambda_2 &= \text{E}(y_1 y_2) \\ &= \text{Covar}(y_1, y_2) + \text{E}(y_1)\text{E}(y_2) \\ &= 1 - \sigma^2/(2N-1). \end{aligned} \tag{54}$$

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 for this model,

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

agreeing with the "$j = 2$" case in the expression in Eq. (52).

Other properties of the Cannings model follow easily. For example, it is clear by symmetry that 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 A_1,

$$\begin{aligned} \text{Var}\{X(t+1) \mid X(t)\} &= \text{Var}(y_1 + \ldots + y_i) \\ &= i\sigma^2 + i(i-1) \text{ Covar}(y_1, y_2) \\ &= i(2N-i)\sigma^2/(2N-1), \end{aligned} \tag{55}$$

from Eq. (53). If $x(t) = X(t)/2N$, it follows that

$$\text{var}\{x(t+1) \mid x(t)\} = x(t)\{1 - x(t)\}\sigma^2/(2N-1). \tag{56}$$

Mutation

Suppose now that the genes in the population are divided into two allelic types, A_1 and A_2, and that if mutation does not exist the conditions for Cannings' first theorem hold. Now assume that A_1 mutates to A_2 at rate u, with reverse mutation at rate v. Write

$x_i = y_i + z_i$, where $y_i = 1$ or 0 depending on whether or not the ith gene at time t continues to exist at time $t+1$. Thus, $y_i = 0$ in the model (36), but we are considering now more general conditions than those specified by this equation. The variable z_i is the number of offspring genes from the ith gene at time t. If this gene is of type A_1, define z_{i1} as the (random) number of its A_1 (that is, non-mutated) offspring: z_{i1} has a distribution which depends on z_i. Similarly if the ith gene is of type A_2 let z_{i2} be the random number of its A_1 (that is mutant) offspring. Then we have

Theorem 2 (Cannings). The eigenvalues of the matrix P describing the stochastic behavior of the number of A_1 genes are $\lambda_0 = 1$ and

$$\lambda_j = \sum \Pr(z_1, \ldots, z_j) \left\{ \mathrm{E} \prod_{i=1}^{j} (y_i + z_{i1} - z_{i2} \mid z_1, \ldots, z_j) \right\},$$
$$(j = 1, 2, \ldots, 2N). \tag{57}$$

In the Wright-Fisher model defined by (36) and (46), $y_i \equiv 0$ and $z_1 \ \ldots \ z_j$ have a multinomial distribution with index $2N$ and common parameter $(2N)^{-1}$. Further, given z_i, z_{i1} and z_{i2} have binomial distributions with respective parameters $1-u$ and v. Thus

$$\mathrm{E}(z_{i1} - z_{i2} \mid z_i) = (1 - u - v) z_i$$

and

$$\begin{aligned} \lambda_j &= \sum \Pr(z_1, \ldots, z_j)(1-u-v)^j z_1 \ldots z_j \\ &= (1-u-v)^j \mathrm{E}(z_1 \ldots z_j) \qquad (58) \\ &= (1-u-v)^j \{2N(2N-1)\ldots(2N-j+1)/(2N)^i\}, \\ &\quad j = 1, 2, \ldots, 2N. \end{aligned}$$

The conclusion of (52) has been used in reaching this formula. 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. The eigenvalues (58) apply also in the one-way mutation model, for which we simply put $v = 0$ in (58).

The Moran Model

No mutation

In this section we consider a model due to Moran (1958) which, in contrast to the Wright-Fisher model and the Cannings model, has the advantage of allowing explicit expressions for many quantities of evolutionary interest, although, strictly, it applies only for haploid populations (where each individual has only one gene, rather than two, at the "A"locus under consideration).

Consider 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

$$p_{i,i-1} = i(2N - i)/(2N)^2. \tag{59}$$

Similar reasoning shows that

$$p_{i,i+1} = i(2N - i)/(2N)^2, \tag{60}$$

$$p_{i,i} = \{i^2 + (2N - i)^2\}/(2N)^2. \tag{61}$$

The matrix defined by these transition probabilities is a continuant, so that standard theory of can be applied to it. In standard continuant "birth-and-death" notation,

$$\lambda_i = \mu_i = i(2N - i)/(2N)^2, \quad i = 0, 1, 2, \ldots, 2N. \tag{62}$$

It follows that the probability π_i of fixation of A_1, given currently i A_1 individuals, is

$$\pi_i = i/2N, \tag{63}$$

and that, using notation developed above,

$$\begin{aligned} \bar{t}_{ij} &= 2N(2N - i)/(2N - j), \quad j = 1, 2, \ldots, i, \\ \bar{t}_{ij} &= 2Ni/j, \quad j = i + 1, \ldots, 2N - 1. \end{aligned} \tag{64}$$

Thus immediately

$$\begin{aligned}
\bar{t}_i &= 2N(2N-i)\sum_{j=1}^{i}(2N-j)^{-1} + 2Ni\sum_{j=i+1}^{2N-1} j^{-1}, && (65)\\
\bar{t}^*_{ij} &= 2N(2N-i)j/\{i(2N-j)\}, \quad j = 1,2,\ldots,i,\\
\bar{t}^*_{ij} &= 2N, \quad j = i+1,\ldots,2N-1, && (66)\\
\bar{t}^*_i &= 2N(2N-i)i^{-1}\sum_{j=1}^{i} j(2N-j)^{-1} + 2N(2N-i-1). && (67)
\end{aligned}$$

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

$$\bar{t}^*_1 = 2N(2N-1) \tag{68}$$

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

The eigenvalues of the matrix defined by (59) – (61) can be found by using Cannings' first 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 "descendants". Using the notation of Canings' first 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

$$\begin{aligned}
\lambda_j &= \mathrm{E}(y_1 y_2 \ \ldots \ y_j)\\
&= 2j(2N-j)/(2N)^2 + 1 - j(4N-j-1)/(2N)^2\\
&= 1 - j(j-1)/(2N)^2, \quad j = 1,2,\ldots,2N. && (69)
\end{aligned}$$

Various expressions for the corresponding eigenvectors have been given. We are particularly interested in the largest non-unit eigenvalue and its associated eigenvectors. The required eigenvalue is

$$\lambda_2 = 1 - 2/(2N)^2, \tag{70}$$

and elementary calculations show that the corresponding right eigenvector $\mathbf{r}$ and left eigenvector ℓ' are

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

Thus the asymptotic distribution of the number X_t of A_1 genes for large t, given $X_t \neq 0, 2N$, is uniform over the values $\{1, 2, 3, \ldots, (2N-1)\}$. The fact that λ_2 is very close to unity for large N agrees with the very large mean absorption times (65) for large N and intermediate values of i.

Mutation

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 is now amended to

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

Standard continuant matrix theory can now be used to find $\bar{t}_{ij}$ and thus $\bar{t}_i$. We do not present explicit expressions since it will be more useful (see below) to proceed via approximations. If mutation from A_2 to A_1 (at rate v) is also allowed, the model becomes

$$\begin{aligned} p_{i,i-1} &= \{i(2N-i)(1-v) + ui^2\}/(2N)^2 = \mu_i \\ p_{i,i+1} &= \{i(2N-i)(1-u) + v(2N-i)^2\}/(2N)^2 = \lambda_i \\ p_{i,i} &= 1 - p_{i,i-1} - p_{i,i+1}. \end{aligned} \tag{71}$$

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

$$\phi_j = \phi_0 \frac{(2N)!\Gamma\{j+A\}\Gamma\{B-j\}}{j!(2N-j)!\Gamma\{A\}\Gamma\{B\}} \tag{72}$$

where $\Gamma\{\cdot\}$ is the 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 $\phi_0 = \Gamma\{B\}\Gamma\{A+C\}/[\Gamma\{D\}\Gamma\{C\}]$. 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 (71), 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 (36) and (46) 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 matrix defined by (71) can be found by applying Cannings' second theorem. Here $y_i = 1$ unless the ith gene has been chosen to die, in which case $y_i = 0$. Similarly z_i, z_{i1} and z_{i2} are zero unless the ith gene has been chosen to reproduce. It is found after some calculation that $\lambda_0 = 1$ and

$$\lambda_j = 1 - \frac{j(u+v)}{(2N)} - \frac{j(j-1)(1-u-v)}{(2N)^2}, \quad j = 1, \ldots, 2N. \quad (73)$$

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 (58) for that model.

Some approximations

Several of the exact results found above for the Moran model are unwieldy, so we now give simple approximate expressions for them.

For the case where there is no mutation, is evident from Eq. (65) that

$$\bar{t}(p) \approx -(2N)^2\{p\log p + (1-p)\log(1-p)\}, \tag{74}$$

where $p = i/2N$. The similarity between this formula and (24) 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.

In the case of one-way mutation, approximate values for $\bar{t}_{ij}$ may be calculated from (71), and from these we obtain an approximate value for $\bar{t}_i$. This is

$$\begin{aligned}\bar{t}_i \;\;&\approx\;\; (2N)^2(1-\theta)^{-1}\Big(\int_0^p x^{-1}\{(1-x)^{\theta-1}-1\}dx \\ &\quad + \int_p^1 x^{-1}(1-x)^{\theta-1}\{1-(1-p)^{1-\theta}\}\,dx\Big)\end{aligned} \tag{75}$$

birth-death events, where $p = i/(2N)$, $x = j/(2N)$ and θ is defined for the (diffusion) approximation to this Moran model as $2Nu$. In the particular case $p = (2N)^{-1}$ this is, to a close approximation,

$$\bar{t}_i \approx 2N\big(1 + \int_{(2N)^{-1}}^p x^{-1}(1-x)^{\theta-1}\,dx\big) \tag{76}$$

birth-death events. When $\theta = 1$ the form of $\bar{t}_i$ may be found by application of L'Hospital's rule.

For the case of two-way mutation we put $x = j/(2N)$, $u = \alpha/(2N)$, $v = \beta/(2N)$ in (72)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 (72) becomes, approximately,

$$\phi_j \sim (2N)^{-1}\frac{\Gamma\{\alpha+\beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}}x^{\beta-1}(1-x)^{\alpha-1}, \tag{77}$$

at least for values of x not extremely close to 0 or 1. This approximation expression is far simpler than the exact value (72).

K-allele Models

The models considered so far can easily be extended to allow K different alleles at the locus in question, where K is an arbitrary positive integer. (For the ABO blood group system, for example, $K = 3$.) In this case the population configuration at any time can be described by a vector $(X_1, X_2, \ldots, X_K)$, where X_i is the number of genes of allelic type A_i. If we assume, as is usual, that $X_1 + X_2 + \ldots + X_K = 2N$, only $K - 1$ elements in the above vector are independent. It is however convenient, for reasons of symmetry, to retain all elements in the vector. The most interesting cases of these models arise when there is no mutation and the K allele generalization of the Wright-Fisher, the Cannings or the Moran model determines the evolution of the population. In this case any allele A_i can be treated on its own, all other alleles being classed simply as non-A_i, and much of the above theory can be applied. (On the other hand, one problem for which the above theory is inadequate is to find the mean time until loss of the first allele lost, the mean time until loss of the second allele lost, and so on. This is a more complex problem that we do not discuss.)

We consider in detail only the K-allele generalization of the model Wright–Fisher (6), namely

$$\begin{aligned} &\Pr\{Y_i \text{ genes of allele } A_i \text{ at time } t+1 \mid X_i \text{ genes of allele} \\ &\qquad i \text{ at time } t, \quad i = 1, 2, \ldots, K\} \\ &= \frac{(2N)!}{Y_1!Y_2!\ldots Y_K!}\psi_1^{Y_1}\psi_2^{Y_2}\ldots\psi_K^{Y_K} \end{aligned} \tag{78}$$

where $\psi_i = X_i/(2N)$. In this case the model (78) is in effect a Cannings model and a straightforward generalization of the theory for the Cannings model given above can be applied.

The eigenvalues of the matrix defined by (78) are precisely the values in Eq. (52), where now λ_j has multiplicity $(K+j-2)!/\{(K-2)!/j!\}$, $(j = 2, 3, \ldots, 2N)$. The eigenvalue $\lambda_0 = 1$ has total multiplicity K. These eigenvalues have the interesting interpretation (Littler, (1975)) that

$$\Pr\{\text{at least } j \text{ allelic types remain present at time } t\} \sim \text{const } \lambda_j^t. \tag{79}$$

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 (47) and (48). 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 $(K-1)^{-1}$, $(j \neq i)$. By symmetry, the mean number of genes of allelic type A_i alleles in the stationary distribution must be $2N/K$. 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 (50) we find, ignoring terms of order u^2, that

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

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

$$F \approx (K-1+\theta)/(K-1+K\theta). \tag{80}$$

This expression agrees with that in (50) for $K = 2$. For large K,

$$F \approx (1+\theta)^{-1}, \tag{81}$$

an expression we return to later.

These formulas demonstrates an important theme. In both formulas, 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 the values (52) if λ_i is multiplied by $\{1-uK(K-1)^{-1}\}^i$. The multiplicity of λ_i is $(i+K-2)!/\{i!(K-1)!\}$.

In view of the comments concerning the Cannings model made above it is plausible that Eqs. (80) and (81) hold with θ defined by $\theta = 4Nu/\sigma^2$. There is also a K-allele Moran model which allows various exact formulae, but we do not consider this here.

Infinitely Many Alleles Models

Introduction

In this section we consider 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 remaining models. This is not because it is more important than the other two, but arises for two reasons. The first is that calculations for this 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. This definition is given below in the section on the Cannings infinitely many alleles model.

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.

The infinitely many alleles model is central for the theory of molecular population genetics, for reasons discussed later.

The Wright–Fisher Infinitely Many Alleles Model

The Wright–Fisher infinitely many alleles model follows the generic binomial sampling characteristic of all Wright–Fisher models. The nature of the mutation mechanism, discussed above, implies that if the mutation rate (always to new allelic types) is u, and if in generation t there are X_i genes of allelic type A_i $(i = 1, 2, 3, \ldots)$, then the probability that in generation $t+1$ there will be Y_i genes of allelic type A_i, together with Y_0 new mutant genes, all of different novel allelic types, is

$$\text{Prob}\{Y_0, Y_1, Y_2, \ldots \mid X_1, X_2, \ldots\} = \frac{(2N)!}{\Pi Y_i!}\,\Pi\pi_i^{Y_i}, \qquad (82)$$

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

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: 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 (82) 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, some of the characteristics of which we now explore.

We consider first the probability that two genes drawn at random 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

$$F_2^{(t+1)} = (1-u)^2\big((2N)^{-1} + \{1-(2N)^{-1}\}F_2^{(t)}\big). \tag{83}$$

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

$$F_2 = \{1 - 2N + 2N(1-u)^{-2}\}^{-1} \approx (1+\theta)^{-1}, \tag{84}$$

where, as as is standard for Wright–Fisher models, $\theta = 4Nu$. This is identical to the limiting ($K \to \infty$) value in (81). 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.

Consider next 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

$$F_3^{(t+1)} = (1-u)^3(2N)^{-2}\big(1+3(2N-1)F_2^{(t)}+(2n-1)(2N-2)F_3^{(t)}\big). \tag{85}$$

At equilibrium $F_3^{(t+1)} = F_3^{(t)} = F_3$, and rearrangement in (85) yields

$$F_3 \approx 2(2+\theta)^{-1}F_2 \approx 2!/[(1+\theta)(2+\theta)]. \tag{86}$$

Continuing in this way we find

$$\begin{aligned} F_i^{(t+1)} &= (1-u)^i[(2N-1)(2N-2)\cdots(2N-i+1)(2N)^{1-i}F_i^{(t)} \\ &\quad + \text{terms in } F_{i-1}^{(t)},\ldots,F_2^{(t)}] \end{aligned} \tag{87}$$

and that for small values of i,

$$F_i \approx (i-1)!/[(1+\theta)(2+\theta)\cdots(i-1+\theta)]. \tag{88}$$

We can also interpret F_i as the probability that a sample of i genes contains only one allelic type, or, in other words, that the sample configuration is $\{i\}$. This conclusion may be used to find the probability of the sample configuration $\{i-1,1\}$. The probability that in a sample of i genes, the first $i-1$ genes are of one allelic type while the last gene is of a new allele type is $F_{i-1}-F_i$. The probability we require is, for $i \geq 3$, just i times this, or

$$\text{Prob}\{i-1,1\} = i\{F_{i-1}-F_i\} \approx i(i-2)!\theta/[(1+\theta)(2+\theta)\cdots(i-1+\theta)]. \tag{89}$$

For $i=2$ the required probability is

$$\text{Prob}\{1,1\} \approx \theta/(1+\theta). \tag{90}$$

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

$$\begin{aligned} F_{2,2}^{(t+1)} &= (1-u)^4(2N)^{-3}\big((2N-1)(2N-2)(2N-3)F_{2,2}^{(t)} \\ &\quad + 2(2N-1)(2N-2)F_{2,1}^{(t)} + 3(2N-1)F_{1,1}^{(t)}\big). \end{aligned} \tag{91}$$

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

$$F_{2,2} \approx (3+\theta)^{-1}F_{2,1} = 3\theta/\big((1+\theta)(2+\theta)(3+\theta)\big). \tag{92}$$

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 formula can be presented in various ways. Perhaps the most useful formula arises if we define $\mathbf{A} = (A_1, A_2, \ldots, A_n)$ as the vector of the (random) numbers of allelic types each of which is represented by exactly j genes in the sample. With this definition,

$$\text{Prob}(\mathbf{A} = \mathbf{a}) = \frac{n!\,\theta^{\sum a_j}}{1^{a_1}2^{a_2}\cdots n^{a_n}\,a_1!\,a_2!\cdots a_n!\,S_n(\theta)}\,. \tag{93}$$

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

$$S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1)$$

. The expression (93) was derived by Ewens (1972) and Karlin and McGregor (1972). It is necessary that $\sum jA_j = \sum ja_j = n$, and it is convenient to denote $\sum A_j$, the (random) number of different allelic types seen in the sample, by K, and $\sum_j a_j$, the corresponding observed number in a given sample, by k.

By suitable summation in (93) the probability distribution of the random variable K may be found as

$$\text{Prob}\,(K=k) = |S_n^k|\theta^k/S_n(\theta), \tag{94}$$

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 (94), the mean of K is

$$\text{E}(K) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1}\,, \tag{95}$$

the variance of K is

$$\text{var}(K) = \theta\sum_{j=1}^{n-1}\frac{j}{(\theta+j)^2}, \tag{96}$$

and the probability that $K = 1$ is

$$\frac{(n-1)!}{(\theta+1)(\theta+2)\cdots(\theta+n-1)}. \tag{97}$$

A formula equivalent to (93) is the following. Suppose that in the sample we observe k different allelic types. We label these in some arbitrary order as types $1, 2, \ldots, k$. Then the probability that $K = k$ and also that with the types labelling in the manner chosen, there are n_1, n_2, ..., n_k genes respectively observed in the sample of these various types, is

$$\frac{n!\theta^k}{k!n_1n_2\cdots n_kS_n(\theta)}\,. \tag{98}$$

We now turn to eigenvalue calculations.Equation (83) can be rewritten in the form

$$F_2^{(t+1)} - F_2^{(\infty)} = (1-u)^2\{1-(2N)^{-1}\}\{F_2^{(t)} - F_2^{(\infty)}\}, \tag{99}$$

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 (85) shows that a second eigenvalue is $(1-u)^3\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}$. Equations (87) and (91) 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

$$\lambda_i = (1-u)^i\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}\cdots\{1-(i-1)(2N)^{-1}\} \tag{100}$$

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 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 (36), with ψ_i defined immediately below (36). 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 $\mathrm{E}(K)/\mathrm{E}(T)$, where $\mathrm{E}(K)$ is

the mean number of alleles existing in each generation. It follows, by balancing the number of alleles gained each generation with the number lost, that at stationarity,

$$\mathrm{E}(K) = 2Nu\mathrm{E}(T). \tag{101}$$

An approximation to $\mathrm{E}(T)$ is found by putting $p = (2N)^{-1}$ in (39). This gives, to a close approximation,

$$\mathrm{E}(K) \approx \theta + \int_{(2N)^{-1}}^{1} \theta x^{-1}(1-x)^{\theta-1}\,dx. \tag{102}$$

A more detailed approximation is possible. If $\mathrm{E}(K(x_1, x_2))$ 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

$$\mathrm{E}(K(x_1, x_2)) \approx \int_{x_1}^{x_2} \theta x^{-1}(1-x)^{\theta-1}\,dx. \tag{103}$$

This equation can be used to confirm (95). An allele whose population frequency is x is observed in a sample of size n with probability $1-(1-x)^n$. From this and (103) it follows that the mean number of different alleles observed in a sample of size n is approximately

$$\int_0^1 \{1-(1-x)^n\}\theta x^{-1}(1-x)^{\theta-1}\,dx, \tag{104}$$

and the value of this expression is equal to that given in (95). The function

$$\phi(x) = \theta x^{-1}(1-x)^{\theta-1} \tag{105}$$

is called the "frequency spectrum"of the process considered, and will be discussed in detail by Dr Griffiths. Ignoring small-order terms, it has the (equivalent) interpretations that 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 frequency spectrum can be used to arrive at further results reached more laboriously by discrete distribution methods. Thus,

for example,

$$\begin{aligned}
&\text{Prob}\{\text{only one allele observed in a sample of } n \text{ genes}\} \\
&\quad \approx \ \theta \int_0^1 x^n \{x^{-1}(1-x)^{\theta-1}\}\, dx \\
&\quad = \ (n-1)!/\big((1+\theta)(2+\theta)\cdots(n-1+\theta)\big)
\end{aligned}$$

and this agrees with the expression in (88) with the notational change of n to i. More complex formulas such as (93) can be re-derived using multivariate frequency spectra, but we do not pursue the details.

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 K-allele model as discussed below (81). This imbalance agrees qualitatively with that found surrounding (81) for the K-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

$$\begin{aligned} & 1-\theta \int_{0.99}^{1} x^{-1}(1-x)^{\theta-1}\,dx \qquad (106) \\ \approx\ & 1-\theta \int_{0.99}^{1} (1-x)^{\theta-1}\,dx \\ =\ & 1-(0.01)^{\theta}. \end{aligned}$$

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.

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 (43), or more accurately in (44).

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

$$\theta = 4N_e u, \qquad (107)$$

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

The Cannings Infinitely Many Alleles Model

The reproductive mechanism in the nonoverlapping 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, to apply for 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 the 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 (71), 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 there will exist a concept of the stationary distribution of allelic configurations. The possible configurations of the process are the same as those for those models, but for the Moran model an exact population probability can be given for each configuration. Suppose that β_j $(j = 1, 2, \ldots, 2N)$ is the number of allelic types with exactly j representative genes in the population, so that $\sum j\beta_j = 2N$. The quantity β_j is the population analogue of the sample number α_j in (93). The exact stationary distribution of the population configuration process is

$$\text{Prob}(\beta_1, \beta_2, \ldots, \beta_{2N}) = \frac{(2N)!\ \theta^{\sum \beta_j}}{1^{\beta_1} 2^{\beta_2} \cdots (2N)^{\beta_{2N}}\ \beta_1!\beta_2! \cdots \beta_{2N}!\ S_{2N}(\theta)}\,. \tag{108}$$

Here $S_j(\cdot)$ is defined below (93) and θ is defined for this model by

$$\theta = 2Nu/(1-u). \tag{109}$$

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 (108) is of exactly the same form as (93), with n replaced by $2N$ and α_j by β_j. Thus several of the calculations arising from (93) are exact for the Moran population process. For example, the distribution of the number K_{2N} of allelic types in the population is given exactly by (94), with n replaced by $2N$. Thus, immediately from (94), the probability that $K_{2N} = 1$ is, exactly,

$$\frac{(2N-1)!}{(1+\theta)(2+\theta)\cdots(2N-1+\theta)}\,. \tag{110}$$

The mean of K_{2N} is given by (95), with n replaced by $2N$ and θ defined by (109), and the variance K_{2N} is

$$\mathrm{var}(K_{2N}) \;=\; \theta \sum_{j=1}^{2N-1} \frac{j}{(\theta+j)^2}. \tag{111}$$

An exact expression is available for the Moran model (discrete) frequency spectrum, for which (105) gives the approximate Wright–Fisher model formula. To find this we consider first the "two-allele" model (71). 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

$$\mu(T) = (2N+\theta) \sum_{j=1}^{2N} j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \tag{112}$$

The form of ergodic argument that led to (102) 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

$$\theta \sum_{j=1}^{2N} j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \tag{113}$$

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 (113) 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

$$\theta\, j^{-1} \left(\binom{2N}{j} \binom{2N+\theta-1}{j}^{-1} \right), \quad j = 1, 2, \ldots, 2N. \qquad (114)$$

A standard asymptotic formula for the gamma function for large N shows the parallel between this exact expression with the diffusion theory frequency spectrum (105).

Various special cases of (114) are of interest. For example, when $\theta = 1$, (114) simplifies to j^{-1}, in which form the parallel with the Wright-Fisher approximation (105) 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 (112) 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 (112) 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 (39), is also $\log(2N)$ generations, but this formal equality is misleading because of the different definitions of θ in the two cases.

Second, the expression (113) 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 (95), 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

$$C = \sum_{j=0}^{2N-1} \binom{2N+\theta-1}{j} \Big(\binom{2N-1}{j} \Big)^{-1}. \tag{115}$$

This is a different probability than the probability that, at any specified time, the population is "quasi-fixed" for one or other allele. This latter probability is given by the $j = 2N$ term in the exact Moran frequency spectrum (114), namely

$$\frac{\theta}{2N} \Big(\binom{2N+\theta-1}{2N} \Big)^{-1}, \tag{116}$$

or, more simply,

$$\frac{(2N-1)!}{(1+\theta)(2+\theta)\cdots(2N-1+\theta)}. \tag{117}$$

To illustrate the difference between the probability defined by (115) and the probability defined by (117), when $\theta = 1$ the former probability is

$$\frac{1}{2N}\big(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2N}\big)^{-1} \tag{118}$$

while the latter probability is $\frac{1}{2N}$.

(Parenthetically, no exact probability for quasi-fixation of one or other allele is known for the Wright-Fisher model. The most accurate approximation available is that of Watterson (1975), namely

$$\exp(-0.1003\theta)\Gamma(1+\theta)(2N)^{-\theta}. \tag{119}$$

This expression gives numerical values quite close to those given by (117) for a wide range of values of θ, although the different definitions of θ for Moran and Wright-Fisher models must be kept in mind when making this comparison.)

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 (114) by $j/(2N)$. This gives

$$\theta(2N)^{-1}\left(\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\right) \tag{120}$$

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

Third, (113) 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, (113) shows that the Harris probability of polymorphism is

$$1-\theta\sum_{j=M+1}^{2N} j^{-1}\left(\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\right). \tag{121}$$

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 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 (43), or more accurately in (44). In the case of the Moran model an exact calculation is available, namely that the required mean number of birth and death events is

$$2N(2N+\theta)(\theta-1)^{-1}\sum_{j=1}^{2N} j^{-1}\left(1-\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}\right). \tag{122}$$

(A formula different from (122), found by applying l'Hôpital's rule, applies for the case $\theta = 1$.) In the case $\theta = 2$, the expression (122)

gives, exactly, $8N^2(N+1)/(2N+1)$, or about $4N^2$, birth and death events. This can be thought of as corresponding to $4N$ "generations", which appears to agree closely with the Wright–Fisher approximation in (45). This agreement is, however, misleading, since the definitions of θ differ in the two models.

There are several further comments to make about (122). First, The typical (jth) term in (122) 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 (122) gives more information than might otherwise be thought.

Second, although the identity is not immediately obvious, the expression in (122) is identical to the expression

$$2N(2N+\theta)\sum_{j=1}^{2N}\frac{1}{j(j+\theta-1)}. \tag{123}$$

We shall see later that the individual terms in the sum also have an important interpretation, in this case concerning the past history of the population rather than its future evolution.

Third, recalling the definition $\theta = 2Nu/(1-u)$ for the Moran model, the expression in (123) may be written equivalently as

$$\sum_{j=1}^{2N}\frac{1}{v_j+w_j}, \tag{124}$$

where

$$v_j = \frac{ju}{2N}, \quad w_j = \frac{j(j-1)(1-u)}{(2N)^2}. \tag{125}$$

We shall later see why expressions of the form defined by (124) and (125) arise.

Further exact Moran model results, relating to "time" and "age" properties, will be discussed later.

Complications and the effective population size

Introduction

All the theory described above (and also that described later) makes a large number of assumptions, genetical, modelling 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 modelling 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 Wright-Fisher model is 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 $= 1 - (2N)^{-1}$,

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

(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.

In view of these properties it is perhaps natural, if the Wright–Fisher model (6) is to be used as a standard, to define the effective population size in diploid models that are more complicated and

realistic then (6) in the following way:

$$N_e^{(e)} = \text{eigenvalue effective population size} = \tfrac{1}{2}(1-\lambda_{\max})^{-1}, \quad (126)$$

$$N_e^{(i)} = \text{inbreeding effective population size} = (2\pi_2)^{-1}, \quad (127)$$

$$N_e^{(v)} = \text{variance effective population size} = \frac{x(t)\{1-x(t)\}}{2\ \mathrm{var}\{x(t+1) \mid x(t)\}}\,. \quad (128)$$

Here $\lambda_{\max}$ is the largest nonunit eigenvalue of the transition matrix of the model considered and π_2 is 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 the more complicated 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.

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 (54) and (126) show immediately that for these models, the eigenvalue effective population size $N_e^{(e)}$ is given by

$$N_e^{(e)} = \big(N - \tfrac{1}{2}\big)/\sigma^2, \quad (129)$$

where σ^2 is the variance in the number of offspring genes from any given gene. Equations (56) and (128) show that the variance effective population size $N_e^{(v)}$ is given by

$$N_e^{(v)} = \big(N - \tfrac{1}{2}\big)/\sigma^2. \quad (130)$$

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

$$\sum_{i=1}^{2N} m_i(m_i-1)/\{2N(2N-1)\}. \quad (131)$$

The probability π_2 in (127) is the expected value of this random variable. Now m_i has mean unity and variance σ^2, so that, on taking expectations, $\pi_2 = \sigma^2/(2N-1)$. From this,

$$N_e^{(i)} = \left(N - \tfrac{1}{2}\right)/\sigma^2. \tag{132}$$

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 (107), it is plausible that the various Wright–Fisher infinitely many alleles model results apply for the nonoverlapping 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

$$\theta = 4Nu/\sigma^2. \tag{133}$$

As stated earlier, the definition of θ given in (133) 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 (126)–(128), 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 each time unit. Since $k = 2N$ for models where generations do not overlap, this leaves (126)–(128) unchanged for such models. For the Moran model, where $k = 1$, this convention yields

$$N_e^{(e)} = N_e^{(i)} = N_e^{(v)} = \tfrac{1}{2}N. \tag{134}$$

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 (24) 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

$$\bar{t}(p) \approx -(4N-2)\{p\log p + (1-p)\log(1-p)\}/\sigma^2, \tag{135}$$

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 (74). 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.

Diploid organisms

So far we have ignored the diploid nature of most organisms of interest, and we now consider a definition of effective population size for the diploid case. We do this here for a Cannings model, and devise an inbreeding effective population number that allows for the diploid nature of the organisms in the population. This number will be denoted $N_e^{(id)}$, and is defined as the reciprocal of the probability that two genes taken at random in generation $t+1$ are descended from the same 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 (131), the probability of this event can be written as the expected value of

$$\sum_{i=1}^{N}(m_i + m_{N+i})(m_i + m_{N+i} - 1)/\{2N(2N-1)\}. \tag{136}$$

It is not hard to see this leads to

$$N_e^{(id)} = \frac{4N-2}{\sigma_d^2 + 2}\,, \tag{137}$$

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 monoecious 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

$$\sigma^2 = \frac{\sigma_d^2 + 2}{4}, \tag{138}$$

where σ^2 is the Cannings model gene "offspring number" variance, and from this it follows that the expressions in (132) and (137) 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 (6).

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

$$X_1(t+1) = i(t+1) + j(t+1), \tag{139}$$

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

$$E[Y\{X_1(t+1), X_2(t+1)\} \mid X_1(t), X_2(t)] = \lambda Y(t) \tag{140}$$

for some constant λ. 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

$$\begin{aligned} Y(X_1, X_2) &= \tfrac{1}{2}C\{X_1(2N_1 - X_1)(2N_1)^{-2} + X_2(2N_2 - X_2)(2N_2)^{-2}\} \\ &\quad + \{1 - (X_1 - N_1)(X_2 - N_2)N_1^{-1}N_2^{-1}\}, \end{aligned} \tag{141}$$

where

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

With this definition the eigenvalue λ becomes

$$\lambda = \tfrac{1}{2}[1 - (4N_1)^{-1} - (4N_2)^{-1} + \{1 + N^2(4N_1N_2)^{-2}\}^{1/2}], \tag{142}$$

or approximately

$$\lambda \approx 1 - (N_1 + N_2)(8N_1N_2)^{-1}. \tag{143}$$

From this result and (126) it follows that to a close approximation,

$$N_e^{(e)} = 4N_1N_2N^{-1}. \tag{144}$$

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

$$N_e^{(i)} = (2\pi_2)^{-1} \approx 4N_1N_2N^{-1}. \tag{145}$$

The variance effective population size cannot be found so readily, and indeed strictly it is impossible to use (128) 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 (128), 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 (128), as

$$N_e^{(v)} = 4N_1N_2N^{-1}. \tag{146}$$

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 K 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 (140). It is found, after some trial and error, that a suitable function $Y(t)$ is

$$\begin{aligned} Y(t) &= [A - D + \{(A-D)^2 + 4BC\}^{1/2}] \sum_i X_i(t)\{2N - X_i(t)\} \\ &\quad + 2B \sum_{i \neq j} \sum X_i(t)\{2N - X_j(t)\}, \end{aligned} \tag{147}$$

where

$$\begin{aligned} A &= (4N^2 + H^2K^2 + K^2H - 2N - 4NKH)/4N^2, \\ B &= (4KN - K^2H - K^2)/(4N^2), \\ C &= (4HKN - K^2H^2 - K^2H)/(4N^2), \\ D &= (4N^2 + HK^2 + K^2 - 4HK)/(4N^2). \end{aligned}$$

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

$$\lambda = \tfrac{1}{2}\big(A + D + \{(A-D)^2 + 4BC\}^{1/2}\big). \tag{148}$$

If small-order terms are ignored, this yields eventually

$$N_e^{(e)} \approx N(H+1)\{1 + (2K(H+1))\}^{-1}\} \tag{149}$$

for large H and K. This equation is in fact accurate to within 10% even for $H = K = 1$, 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 K, since the act of migration is irrelevant to the computation of its numerical value. Thus immediately from (132)

$$N_e^{(i)} = \big\{N(H+1) - \tfrac{1}{2}\big\}/\{1 - (2N)^{-1}\}, \tag{150}$$

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 K, it differs slightly from $N_e^{(e)}$.

Because of these two results, one may be tempted to ignore geographical sub-division in modelling 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 (6),

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

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,

$$N_e^{(e)} \approx k\{N_1^{-1} + \cdots + N_k^{-1}\}^{-1}. \tag{151}$$

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.

Molecular population genetics

Introduction

The infinitely many alleles model was 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.

Another model inspired by this knowledge is the "infinitely many sites" model, described in detail in these lectures by Dr Griffiths. However, some aspects of this model are discussed below.

There are three points where the mathematical population genetics theory based on nucleotide frequencies differs from the classical theory based on gene frequencies. First, the molecular theory is dynamic, in contrast to the often static classical theory. Mutations are usually seen as leading to new allelic types rather than back to currently or previously existing types, since it is plausible that most nucleotide mutations will lead to sequences not currently existing in the population. Both the infinitely many alleles and the infinitely many sites models were originally proposed with this view in mind.

Second, while the classical theory concerns the evolution of genes given labels "A_1", "A_2", etc., at the molecular level the actual genetic material is known, so that the symbols a, g, c, and t refer to specific rather than type entities. The fact that the theory thus concerns ultimate and real entities is of great importance. Perhaps

the most important point derives from the fact that molecular considerations often lead to *retrospective* rather than *prospective* evolutionary questions. Classical population genetics theory was largely prospective: Given reasonable numerical values for various genetic parameters, the main aim was to show that evolution as a genetic process could and would occur. A hundred years ago such an undertaking was required. It is, however, no longer necessary to do this, and it now appears more useful to attempt to describe the course that evolution has taken by a retrospective analysis, and thus to gain empirical insight into evolutionary questions. This change of viewpoint has also led to the introduction of statistical methods for analyzing current genetical data, discussed below. The current emphasis on statistical inference procedures is perhaps the most important new direction in the theory in recent times. Knowledge of the actual genetic material is essential for these inferences, and the entire retrospective analysis must therefore be carried out in the framework of molecular population genetics.

Finally, most of the retrospective theory is inferential, and for reasons of practicality relies for the inferences made on the data available from a *sample* of genes taken from the population of interest, rather than from data from the entire population. We denote throughout the number of genes in this sample by n. Almost all results given below relate to such a sample.

We consider first the Wright–Fisher infinitely many alleles model. The properties of a sample of n genes under this model are best summarized through the (approximating) partition formula (93). This leads to the distribution of the number K_n of different allelic types observed in the sample as given in (94) and thus to the mean of K_n as given by (95).

There is currently much interest in estimating the parameter θ. Equations (93) and (94) show jointly that the conditional distribution of the vector $\mathbf{A} = (A_1, A_2, \ldots, A_n)$ defined before (93), given the value of K_n, is

$$\text{Prob}\{\mathbf{A} = \mathbf{a} | K_n = k\} = \frac{n!}{|S_n^k| \; 1^{a_1} 2^{a_2} \cdots n^{a_n} \, a_1! a_2! \cdots a_n!}\,, \quad (152)$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$. Equation (152) 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. The main inferential procedure deriving from (152) is the testy of hypothesis that the alleles in the sample are selectively equivalent. The fact that, in practice, θ is unknown does not matter for such inferences, since the conditional distribution (152) is independent of θ, and forms the "null hypothesis" distribution of the allelic partition. Tests of the neutrality hypothesis are discussed in a separate section below.

It is also possible to use (94) to estimate θ or more generally any function of θ. Since K_n is a sufficient statistic for θ we can use the probability distribution in (94) 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

$$K_n = \frac{\hat{\theta}_K}{\hat{\theta}_K} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 1} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 2} + \cdots + \frac{\hat{\theta}_K}{\hat{\theta}_K + n - 1}\,. \tag{153}$$

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

$$k_n = \frac{\hat{\theta}_k}{\hat{\theta}_k} + \frac{\hat{\theta}_k}{\hat{\theta}_k + 1} + \frac{\hat{\theta}_k}{\hat{\theta}_k + 2} + \cdots + \frac{\hat{\theta}_k}{\hat{\theta}_k + n - 1}\,. \tag{154}$$

Numerical calculation of the estimate $\hat{\theta}_k$ using (154) is usually necessary.

The estimator implied by (153) 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 $g(K_n)$, (94) 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 (94). 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 $k_n = 2$,

$$g(2) = \frac{\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}. \tag{155}$$

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 (153), even though this estimator is slightly biased.

Geneticists sometimes prefer to estimate $(1+\theta)^{-1}$ by f, defined in the notation of (98) by

$$f = \sum_i \frac{n_i^2}{n^2}. \tag{156}$$

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

$$\hat{\theta}_f = f^{-1} - 1, \tag{157}$$

is biased and has mean square error approximately six or eight times larger than that of $\hat{\theta}$.

An approximation for the mean square error (MSE) of the estimator $\hat{\theta}_K$ as defined by (153) is found as follows. Writing the right-hand side of (153) as $\psi(\hat{\theta}_K)$, we have $K_n = \psi(\hat{\theta}_K)$ and also, from (95), $\mathrm{E}(K_n) = \psi(\theta)$. Thus by subtraction,

$$K_n - \mathrm{E}(K_n) = \psi(\hat{\theta}_K) - \psi(\theta).$$

A first-order Taylor series approximation for the right-hand side is $(\hat{\theta}_K - \theta)\psi'(\theta)$, so that

$$K_n - \mathrm{E}(K_n) \approx (\hat{\theta}_K - \theta)\psi'(\theta).$$

Squaring and taking expectations, we get

$$\mathrm{MSE}(\hat{\theta}_K) \approx \frac{\mathrm{var}(K_n)}{\psi'(\theta)^2}. \tag{158}$$

The variance of K_n is given in (96), and it is immediate that

$$\psi'(\theta) = \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}. \tag{159}$$

This leads to

$$\mathrm{MSE}(\hat{\theta}_K) \approx \frac{\theta}{\sum_{j=1}^{n-1} \frac{j}{(j+\theta)^2}}. \tag{160}$$

The approximation (160) appears to be quite accurate.

Exact sample results for the Moran model can be obtained rapidly, since under the Moran infinitely many alleles model, (93) holds exactly if θ is defined by (109). This is in contrast to the situation for the Wright–Fisher model, where (93) is only an approximation. Thus with the Moran model definition of θ, (94), (95), (96), (97), and (98) are all exact, as is also the conditional distribution formula (152) that derives from them. It is interesting to ask why these formulas hold exactly in the Moran model, not only in a sample but also in the population, and also why sample formulas and population formulae are identical, with the replacement of n for $2N$. Coalescent theory, which we now turn to, explains this fact.

The coalescent

Introduction

The concept that is most frequently used for inferential and other purposes in population genetics is the Kingman coalescent (Kingman (1982)) . In this section a description of the most immediate properties of the simple coalescent process is given. Complications that in practice must be taken into account (for example changes over time in the size of the population under consideration) are not discussed. One of the main values of the coalescent is to provide a coherent framework within which to view various properties of the models considered above.

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

$$\frac{b}{a+b}\Big(\frac{a}{a+b}\Big)^j, \quad j = 0, 1, \ldots . \tag{161}$$

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

$$(1,2), \quad (3), \quad (4,5), \quad (6,7), \quad (8). \tag{162}$$

Such a time τ is shown in Figure 1.

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 ignored, and given that the process is in ξ at time τ,

$$\text{Prob (process in } \eta \text{ at time } \tau + \delta\tau) = \delta\tau, \tag{163}$$

and if j is the number of equivalence classes in ξ,

$$\text{Prob (process in } \xi \text{ at time } \tau + \delta\tau) = 1 - \frac{j(j-1)}{2}\delta\tau. \tag{164}$$

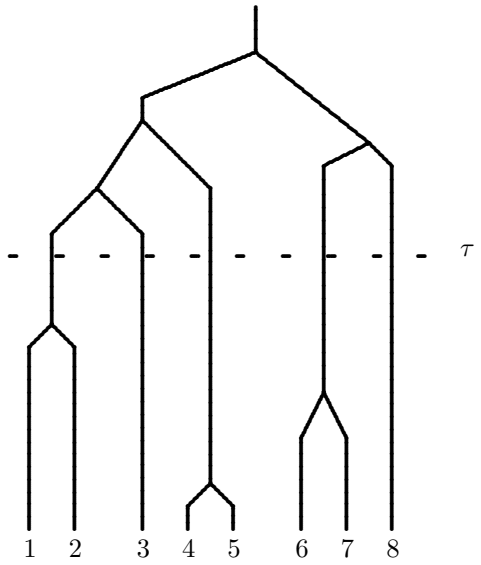

Figure 1: The coalescent

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 (163) and (164)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 $E(T_{\mathrm{MRCAS}})$ is, immediately,

$$E(T_{\mathrm{MRCAS}}) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2\sum_{j=1}^{n-1} \frac{1}{j(j+1)}. \qquad (165)$$

This time is $2 - 2/n$ coalescence time units, and it requires a multiplicative scaling factor of $2N$ to convert to a "generations" basis when applied to the Wright–Fisher model. In other words, $E(T_{\mathrm{MRCAS}}) = 4N - 4N/n$ generations.

It is clear from (165) that about half this mean time relates to the final coalescence of two lines of ascent into one. This observation gives some idea of the shape of the coalescent tree: The long arms tend to arise when there is a very small number of genes in the ancestry of the sample.

The times $T_j, j = 1, 2, \ldots, n-1$, are independent, so that the variance of T_{MRCAS} is the sum of the variances of the T_j. Standard calculations show that this is approximately $4\pi^2/3 - 12$, or about 1.16, (squared) time units. This implies a standard deviation of about 2.16 generations.

The complete distribution of T_{MRCAS} is also known (Tavaré (2004)). However the expression is complicated and we do not reproduce it here, other than to note the inequalities

$$e^{-t} \leq \mathrm{Prob}\,(T_{\mathrm{MRCAS}} > t) \leq e^{-3t}.$$

If the above theory were to apply to the entire population of genes in a Wright–Fisher model, the mean $E(T_{\mathrm{MRCAP}})$ of the total time to arrive at the most recent ancestor gene of all the genes in the population (MRCAP) would be found by putting $n = 2N$, to get

$$E(T_{\mathrm{MRCAP}}) = 4N - 2 \tag{166}$$

generations. Although coalescent theory does not apply directly to the entire population, the mean number of generations given in (166) is essentially correct. The reason for this is implicit in an observation made above, that the long arms in any coalescent process tend to arise when the number of genes in the ancestry of the genes considered is small, and for such small numbers the assumptions for the coalescent process hold.

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 2 describes such an ancestry, identical to that of Figure 1 but with crosses to indicate mutations.

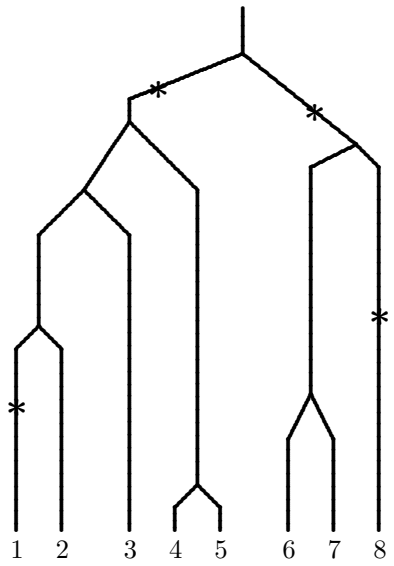

Figure 2: The coalescent with mutations

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 3 (describing the same ancestry as that in Figure 2).

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

$$\frac{1}{2}j(j-1)\delta\tau + \frac{1}{2}j\theta\delta\tau \;=\; \frac{1}{2}j(j+\theta-1)\delta\tau, \tag{167}$$

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

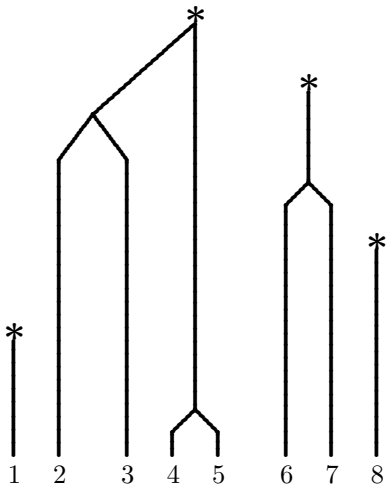

Figure 3: Tracing back to, and stopping at, mutational events

of a mutation.

If a defining event is a coalescence of two lines of ascent, the 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

$$2\sum_{j=1}^{n}\frac{1}{j(j+\theta-1)} \tag{168}$$

coalescent time units. The value in (168) 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 (168), 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 (165) to the expression in (168). 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 (165) and (168) 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 (95). 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 (96).

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 (94).

The complete distribution of the allelic configuration in the sample as given in (93) 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 (93). That is, (93) 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

$$\prod_{j=1}^{n-1} \frac{j}{(j+\theta)} \;=\; \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}, \tag{169}$$

and this agrees, as it must, with the expression in (97).

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

$$E(T_{\mathrm{MRCAS}}) = \sum_{j=2}^{n} \frac{(2N)^2}{j(j-1)} = (2N)^2 \left(1 - \frac{1}{n}\right) \tag{170}$$

birth and death events. In the particular case $n = 2N$ this is

$$E(T_{\mathrm{MRCAP}}) = 2N(2N-1) \tag{171}$$

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

$$\mathrm{var}(T_{\mathrm{MRCAS}}) = \sum_{j=2}^{n} \frac{(2N)^4}{j^2(j-1)^2} - \sum_{j=2}^{n} \frac{(2N)^2}{j(j-1)}. \tag{172}$$

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

$$\frac{ju}{2N} + \frac{j(j-1)(1-u)}{(2N)^2} = \frac{2Nju + j(j-1)(1-u)}{(2N)^2}. \tag{173}$$

We write the left-hand side as $v_j + w_j$, where v_j and w_j are defined in (125). 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,

$$\sum_{j=1}^{n} \frac{1}{v_j + w_j}, \tag{174}$$

where v_j and w_j are defined in (125).

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,

$$\sum_{j=1}^{n} \frac{\theta}{j-1+\theta}, \tag{175}$$

as given by (95). 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 (94), 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.

"Age" and "frequency" results

Many elegant results are found when considering the ages of alleles and the frequencies of alleles when ordered by their ages. In this section we consider both "frequency" and "age" results, starting with the former.

Approximate Wright-Fisher model "frequency" results

We consider first approximate results applying for the Wright-Fisher model. These follow largely from the so-called GEM distribution, named for Griffiths, (1980), Engen (1975) and McCloskey (1965), who established its salient properties. This distribution can be found in the following way.

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. The frequency spectrum shows immediately that the (random) frequency of the allele determined by this randomly chosen gene is

$$f(x) = \theta(1-x)^{\theta-1}. \tag{176}$$

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 (relative) frequency of the allelic type of this gene among the genes remaining at this stage is also given by (176). All genes of this second allelic type are now also removed from the population. A third gene is

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 (relative) frequency of the allelic type of any gene just drawn among the genes left when the draw takes place is given by (176). 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 $j = 2, 3, \ldots$,

$$w_j = (1 - x_1)(1 - x_2) \cdots (1 - x_{j-1})x_j. \tag{177}$$

where the x_j are independent random variables, each having the distribution (176). The random vector $(w_1, w_2, \ldots)$ is then defined as having the GEM distribution.

All the alleles in the population at any time eventually leave the population, through the joint processes of mutation and random drift, and any allele with current population frequency x survives the longest with probability x. That is, the GEM arises when alleles are labelled 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 labelled 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 formulas derives directly from the simplicity and tractability of the GEM distribution. Here are two examples. First, the GEM distribution shows immediately that the mean population frequency of the oldest allele in the population is

$$\theta \int_0^1 x(1 - x)^{\theta - 1} dx = \frac{1}{1 + \theta}, \tag{178}$$

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 (97). 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 $n/(n+\theta)$.

A question of some interest is to find the probability that the oldest allele in the population is also the most frequent. By reversibility arguments 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 not user-friendly. A lower bound for the mean frequency of the most frequent allele is $(1/2)^\theta$, which is useful for small θ but not of much value for larger θ, and an upper bound is $1-\theta(1-\theta)\log 2$. When $\theta = 1$ this mean 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.

Exact Moran model "frequency" results

It will be expected that exact results, corresponding to those given above for the Wright-Fisher model, hold for the Moran model, with θ defined, as always for this model, as $2Nu/(1-u)$. The first of these is an exact representation of the GEM distribution, analogous to (177). 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

$$N_i = 1 + M_i, \quad i = 1, 2, \ldots, \tag{179}$$

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 iid continuous

random variables each having the density function (176). Eventually the sum $N_1 + N_2 + \cdots + N_k$ reaches the value $2N$ and the process then 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 $1/\{1+(2N-1)u\}$, which is very close to the approximation $1/\{1+\theta\}$.

If there is only one allele in the population, 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

$$\text{Prob}\,(M_1 = 2N-1) = \theta \int_0^1 x^{2N-1}(1-x)^{\theta-1}\,dx.$$

This probability also follows immediately from the fact that an allele arising $2N$ times in the population must be the oldest allele, and is given more simply (see (110)) as

$$\frac{(2N-1)!}{(1+\theta)(2+\theta)\cdots(2N-1+\theta)}\,. \tag{180}$$

More generally, Kelly (1977) has shown that the probability that the oldest allele is represented by j genes in the population is

$$\frac{\theta}{2N}\binom{2N}{j}\binom{2N+\theta-1}{j}^{-1}, \tag{181}$$

or more simply,

$$\frac{\theta(2N-1)(2N-2)\cdots(2N-j+1)}{(2N+\theta-j)(2N+\theta-j+1)\cdots(2N+\theta-1)}. \tag{182}$$

The case $j = 2N$ given in (180) is clearly a particular example of this, and the mean number $(2N+\theta)/(1+\theta)$ follows from (182).

Approximate Wright-Fisher model "age" results

We now turn to "age" questions, considering first approximate Wright-Fisher results.. Some for these follow immediately from the previous

calculations. For example, the mean time for all the alleles existing in the population at any time to leave the population is given in (44), and by reversibility this is the mean time, into the past, that the oldest of these originally arose by mutation. This is then the mean age of the oldest allele in the population, given on a "generations" basis. In other words,

$$\text{mean age of oldest allele} \;=\; \sum_{j=1}^{2N} \frac{4N}{j(j+\theta-1)} \text{ generations.} \quad (183)$$

In the case $\theta = 2$, this mean age is very close to $4N - 2$, that is, to the conditional mean fixation time (45).

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 Wright–Fisher model approximation for this is found immediately from (40) as

$$\sum_{j=1}^{\infty} \frac{4N}{j(j+\theta-1)}\Big(1-(1-p)^j\Big). \quad (184)$$

This is clearly a generalization of the expression in (183), since if $p = 1$, only one allele arises in the population, and it must then be the oldest allele.

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? With a change of notation, the density function of the frequency p of the allelic type of the randomly chosen gene is, from (176), $f(p) = \theta(1-p)^{\theta-1}$. The mean age $\bar{t}(p)$ of an allele with frequency p is, by reversibility, given by (40). The required probability is

$$\theta \int_0^1 \bar{t}(p)(1-p)^{\theta-1}\,dp, \quad (185)$$

and use of (40) for $\bar{t}(p)$ shows that this reduces to $2/\theta$ time units, or for the Wright–Fisher model, $1/u$ generations. This conclusion may also be derived by looking backward to the past and using coalescent arguments. It is also an immediate result. Looking backward 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 $u(1-u)^{j-1}, \quad j = 1, 2, \ldots$, and the mean of this (geometric) distribution is $1/u$.

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 obtained d by Donnelly and Tavaré (1986), who found 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) $n_{(1)}, n_{(2)}, \ldots, n_{(k)}$. This probability is

$$\frac{\theta^k (n-1)!}{S_n(\theta) n_{(k)} (n_{(k)} + n_{(k-1)}) \cdots (n_{(k)} + n_{(k-1)} + \cdots + n_{(2)})}, \tag{186}$$

where $S_j(\theta)$ is defined below (93).

Griffiths and Tavaré (1998) give the Laplace transform of the distribution of the age of an allele observed b times in a sample of n genes, together with a limiting Laplace transform for the case in which θ approaches 0. These results show that the diffusion approximation for the mean age of such an allele is

$$\sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)} \left(1 - \frac{(n-b+\theta)_{(j)}}{(n+\theta)_{(j)}} \right) \tag{187}$$

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 (184), and it converges to (184) as $n \to \infty$ with $b = np$.

Our final calculation concerns the mean age of the oldest allele in the sample. For the Wright–Fisher model this mean age is

$$4N \sum_{j=1}^{n} \frac{1}{j(j+\theta-1)}. \tag{188}$$

For the case $n = 2N$ this is the value given in (44), and for the case $n = 1$ it reduces to the value $1/u$ given above.

Exact Moran model age results

We now turn to exact Moran model "age" results. The exact result corresponding to (183) for the Moran model is given in (122),

or equivalently in (123), being almost exactly $4N^2$ birth and death events. The exact Moran model calculation corresponding to (184) follows from the mean persistence time found eventually using standard continuant Markov chain theory and (71). The Moran model calculations parallel to those leading from (185) use the exact frequency spectrum (113) and the exact mean age deriving from (71). However, a direct argument parallel to that given for the Wright–Fisher model shows that the exact mean time, measured in birth and death events, is $2N/u$.

The expression (186) 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))

$$\frac{\theta}{n}\binom{n}{j}\binom{n+\theta-1}{j}^{-1}. \tag{189}$$

This is identical to the expression (181) if $2N$ is replaced by n in the latter.

Further exact Moran model results provide connections between the oldest allele in the sample and the oldest allele in the population. 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 when $n = 1$ it reduces to a result found above that a randomly selected gene is of the oldest allelic type in the population. (The Wright-Fisher model approximation to this probability, found by letting $N \to \infty$, is $n/(n+\theta)$.)

A further result is that 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 Wright-Fisher model 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 formulas extending these. He showed, for example, that the probability that the oldest allele in the population is observed j times in the sample is

$$\frac{\theta}{n+\theta}\binom{n}{j}\binom{n+\theta-1}{j}^{-1}, \quad j = 0, 1, 2, \ldots, n. \tag{190}$$

This is, of course, closely connected to the Kelly result (189). 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 (190) shows that this conditional probability and that in (189) are identical.

The result corresponding to (188) for the Moran model is

$$2N(2N + \theta) \sum_{j=1}^{n} \frac{1}{j(j + \theta - 1)} \tag{191}$$

birth and death events, with (of course) θ defined as $2Nu/(1 - u)$. When $n = 1$ this reduces to the calculation $2N/u$ found above. When $n = 2N$ it is identical to (123) and, less obviously, to the expression given in (122).

As noted in the discussion following (123), the expression in (191) may be written equivalently as

$$\sum_{j=1}^{n} \frac{1}{v_j + w_j}, \tag{192}$$

where v_j and w_j are defined in (125). Coalescent arguments explain why the mean age of the oldest allele in a sample can be expressed in this form and why the mean age of the oldest allele in the population can similarly be expressed in the form defined by (124) and (125).

Testing neutrality

Introduction

Almost all the theory discussed so far, and in particular all of the coalescent theory described, 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 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 (152) conform reasonably to what is expected under neutrality, that is, under the formula (152), 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 (98) and to assess whether these conform reasonably to their conditional probability given n and k, namely,

$$\text{Prob}(n_1, n_2, \ldots, n_k | k) = \frac{n!}{|S_n^k| k! n_1 n_2 \cdots n_k} \,. \tag{193}$$

The Ewens and the Watterson testing procedures differ only in the test statistic employed, and here we discuss only the (superior) Watterson procedure. This uses as test statistic the observed sample

homozygosity, defined as

$$f = \sum_{j=1}^{k} \frac{n_j^2}{n^2}. \tag{194}$$

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 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 (193). In practice, difficulties arise with the mathematical calculations because of the form of the distribution (193), 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 (193) 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 (193): 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 1) in Table 2.

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).

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 1: *Drosophila* sample data

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 2: Sample statistics, means, variances, and probabilities for the data of Table 11.1.

For each set of data we compute f, the observed homozygosity. Then the exact neutral theory probability P (given in Table 2) that the homozygosity is more extreme than its observed value may be calculated (except for the *D. simulans* case where the computations are prohibitive). The simulated probabilities P_{sim} are also given in Table 2; 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 (152) as

$$\mathrm{E}\big(A_i|k,n\big) = \frac{n!}{i(n-i)!}\,\frac{|S_{k-1}^{n-i}|}{|S_k^n|}. \tag{195}$$

In this formula the S_j^i are values of Stirling numbers of the first kind as discussed after (94). 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 (71) shows that given that $k = 10$ and $n = 21$,

$$\mathrm{E}\big(A_i \mid k = 10, n = 21\big) = \frac{21!}{i(21-i)!} \frac{|S_9^{21-i}|}{|S_{10}^{21}|}, \tag{196}$$

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 (196) is given in Table 3. 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 3: 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 (152) to be

$$\text{Prob}(A_1 = a|k, n) = \sum_{j=a}^{k-1} (-1)^{j-a} \frac{|S_{k-1}^n|}{a!(j-a)!|S_k^n|}. \tag{197}$$

Here S_i^j is again a Stirling number. The conditional mean of A_1 is $|S_{k-1}^n|/|S_k^n|$, and the distribution (197) is approximately Poisson, with this mean. This observation enables a rapid approximate assessment of whether the number of singleton alleles is a significantly large one, assuming selective neutrality.

Tests based on the infinitely many sites model

Introduction

Dr Griffiths has discussed the infinitely many sites model in detail, 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 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. The basic idea of the Tajima test is to form a statistic whose numerator is the difference between two such unbiased estimators and whose denominator is an 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. To find the sampling properties of these statistics it is necessary first to discuss properties of the various unbiased estimators of θ used in them. The theory described below relates to the Wright-Fisher model testing procedure. A parallel theory applies for other models.

Estimators of θ

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 (discussed by Dr Griffiths) shows that the mean of S_n is given by

$$\theta \sum_{j=1}^{n-1} 1/j = g_1\theta,$$

where

$$g_1 \;=\; \sum_{j=1}^{n-1} \frac{1}{j}. \tag{198}$$

We note for future reference that the variance of S_n is

$$\mathrm{var}(S_n) \;=\; g_1\theta + g_2\theta^2, \tag{199}$$

where

$$g_2 \;=\; \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{200}$$

Clearly an unbiased estimator of θ is

$$\hat{\theta}_S = \frac{S_n}{g_1}. \tag{201}$$

Equation (199) implies that the variance of $\hat{\theta}_S$ is

$$\text{var}(\hat{\theta}_S) = \frac{\theta}{g_1} + \frac{g_2\theta^2}{g_1^2}\,. \tag{202}$$

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

$$T = \frac{\sum_{i<j} T(i,j)}{\binom{n}{2}}. \tag{203}$$

Since this is also an unbiased estimator of θ, we think of it as forming the unbiased estimator $\hat{\theta}_T$, defined by

$$\hat{\theta}_T = \frac{\sum_{i<j} T(i,j)}{\binom{n}{2}}. \tag{204}$$

This estimator of θ was proposed by Tajima (1983). It is a poor estimator of θ in that its variance, namely,

$$\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 = b_1\theta + b_2\theta^2, \tag{205}$$

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.

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

$$D = \frac{\hat{\theta}_T - \hat{\theta}_S}{\sqrt{\hat{V}}}, \tag{206}$$

where $\hat{V}$ is an unbiased estimate of the variance of $\hat{\theta}_T - \hat{\theta}_S$ and is defined in (208) below. Tajima showed, by using adroit coalescent arguments, that the variance V of $\hat{\theta}_T - \hat{\theta}_S$ is

$$V = c_1\theta + c_2\theta^2, \tag{207}$$

where

$$c_1 = b_1 - \frac{1}{g_1}, \quad c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{g_2}{g_1^2}.$$

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 (207). Elementary statistical theory shows that this function is

$$\hat{V} = \frac{c_1 S}{g_1} + \frac{c_2 S(S-1)}{g_1^2 + g_2}. \tag{208}$$

This is then used in the D statistic given in (206) 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 and its mean is not zero, nor is its variance 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.

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 $2S/n$, and the numerator in D is then $\{(2/n) - (1/g_1)\}S$. In this case the value of D approaches a, defined by

$$a = \frac{\{(2/n) - (1/g_1)\}\sqrt{g_1^2 + g_2}}{\sqrt{c_2}}, \tag{209}$$

as the value of S 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

$$b = \frac{\{(n/2(n-1)) - (1/a_1)\}\sqrt{g_1^2 + g_2}}{\sqrt{c_2}} \tag{210}$$

when n is even and the value of S 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

$$f(D) = \frac{\Gamma(\alpha+\beta)(b-D)^{\alpha-1}(D-a)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)(b-a)^{\alpha+\beta-1}}, \tag{211}$$

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.

References

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).

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).

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.: The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25**, 123–159 (1987).

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.: Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

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.: *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.: Heterosis or neutrality? *Genetics* **85**, 789–814 (1977).

Watterson, G.A.: The homozygosity test of neutrality. *Genetics* **88**, 405–417 (1978).

Wright, S.: Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).