Adv. Appl. Prob. **36,** 1175–1197 (2004)
Printed in Northern Ireland

SOLVING THE FISHER–WRIGHT AND COALESCENCE PROBLEMS WITH A DISCRETE MARKOV CHAIN ANALYSIS

SAMUEL R. BUSS,* *University of California, San Diego*
PETER CLOTE,** *Boston College*

Abstract

We develop a new, self-contained proof that the expected number of generations required for gene allele fixation or extinction in a population of size n is $O(n)$ under general assumptions. The proof relies on a discrete Markov chain analysis. We further develop an algorithm to compute expected fixation or extinction time to any desired precision. Our proofs establish $O(nH(p))$ as the expected time for gene allele fixation or extinction for the Fisher–Wright problem, where the gene occurs with initial frequency p and $H(p)$ is the entropy function. Under a weaker hypothesis on the variance, the expected time is $O(n(p(1-p))^{1/2})$ for fixation or extinction. Thus, the expected-time bound of $O(n)$ for fixation or extinction holds in a wide range of situations. In the multi-allele case, the expected time for allele fixation or extinction in a population of size n with n distinct alleles is shown to be $O(n)$. From this, a new proof is given of a coalescence theorem about the mean time to the most recent common ancestor (MRCA), which applies to a broad range of reproduction models satisfying our mean and weak variation conditions.

Keywords: Fisher–Wright model; diffusion equation; population genetics; mitochondrial Eve; Markov chain; mean stopping time; coalescence; martingale

2000 Mathematics Subject Classification: Primary 60J05
Secondary 60J20; 60J22; 60J70

1. Introduction

Fisher [10] and Wright [27], [28] considered the following problem in population genetics. In a fixed-size population of n haploid individuals carrying only gene alleles A and a, what is the expected number of generations before all n individuals carry only allele A, or all n individuals carry only allele a?

Formally, we assume neutral selection and that the successive generations are discrete, nonoverlapping, and of fixed size n. If one generation contains i haploid members with allele A and $n-i$ with allele a, then the conditional probability that the next generation has exactly j members with allele A equals

$$p_{i,j} = \binom{n}{j}\left(\frac{i}{n}\right)^j\left(1-\frac{i}{n}\right)^{n-j}. \tag{1}$$

Received 12 June 2003; revision received 9 June 2004.
* Postal address: Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA.
Email address: sbuss@ucsd.edu
Supported in part by NSF grants DMS-9803515 and DMS-0100589.
** Postal address: Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.
Email address: clote@bc.edu

Modeling this as a Markov chain $\mathcal{M}$ having states $0, \ldots, n$, where state i represents the situation that exactly i alleles are of type A, it is clear that

$$\lim_{t\to\infty} \Pr[\mathcal{M} \text{ is in either state 0 or } n \text{ at time } t] = 1$$

or, in other words, that eventually only one allele will be present in a population. The first time t at which $\mathcal{M}$ is in state 0 or n is called the *absorption time* or the *stopping time*, where time is measured in terms of the number of generations.

A series of contributions by Fisher [10], Wright [27], [28], Kimura [12]–[14], Karlin and McGregor [11], Watterson [25], and Ewens [5] solved the Fisher–Wright problem via the diffusion equation, a differential equation giving a continuous approximation to the Fisher–Wright process. (See [7], [22], [26] for a comprehensive overview of the diffusion equation approach and the rigorous justification of its applicability to the discrete Fisher–Wright problem.) Kimura [12], Watterson [25] and Ewens [5] established the mean stopping time (in generations) for the diffusion equation associated with the Fisher–Wright process, as

$$-2n(p \ln p + (1-p)\ln(1-p)) = 2nH(p)$$

when starting from an initial allele frequency of $p = i/N$, where $H(p)$ is the base e entropy function. (The Fisher–Wright problem is generally stated for n diploid individuals, whereas we are working with n haploid individuals. Thus, in our setting, the quantity n replaces the quantity $2n$ in the usual formulation of the Fisher–Wright problem.) Ewens has shown that this is an upper bound for the mean stopping time for the discrete Fisher–Wright problem and, in [6], estimated the error in the diffusion equation approximation with a logarithmic additive term.

This article presents a new, self-contained proof that the expected number of generations required for gene allele fixation or extinction is $O(nH(p))$. Of course, this result is a weakening of the just-discussed approximation by the diffusion equation. However, the diffusion approximation proofs are long and arduous. In contrast, our proofs avoid the diffusion equation approach altogether and work solely with discrete Markov chains.

Our proof methods can be applied to a wide range of Markov models: our only assumptions are a *mean condition* and a *variation condition* as defined in Section 2. For instance, S. Cook and C. Rackoff (private communication, 2001) suggested the following hypergeometric model for reproduction. (Schensted [21] earlier studied the hypergeometric model. A generalized hypergeometric model has been suggested by Möhle [20].) Assume that a population consists of discrete, nonoverlapping generations of fixed size n. The $(i+1)$th generation is obtained from the ith generation as follows: each parent individual in the ith generation has two offspring, yielding $2n$ potential offspring; a randomly chosen set of n of these potential offspring survives to be the $(i+1)$th generation. Thus, each parent individual will have zero, one or two offspring survive in the succeeding generation. The hypergeometric model is based on sampling from a set of size $2n$ without replacement, whereas the Fisher–Wright model uses sampling with replacement. For the hypergeometric model, if a generation contains i individuals with allele A, then the probability that the next generation contains j individuals with allele A is equal to

$$q_{i,j} = \frac{\binom{2i}{j}\binom{2n-2i}{n-j}}{\binom{2n}{n}}. \tag{2}$$

The hypergeometric process satisfies the mean and variation conditions, and thus our theorems imply that it has expected absorption time $O(nH(p))$.

Since our approach is different from the traditional one, it is likely that our results cover some cases which cannot be covered by the diffusion equation approach. In fact, we prove similar stopping time results of the form $O(n\sqrt{p(1-p)})$ for Markov chains satisfying a weaker assumption on the variance of the number of individuals with a given allele. This weaker assumption is defined in Section 2 as the *weak variation condition*. The intuitive difference between the 'weak variation condition' and the 'variation condition' is that the weak variation condition allows for less variance in the sizes of subpopulations that have density close to 0 or 1, i.e. alleles which are carried by nearly all or nearly none of the population may have less variance. These theorems need no assumptions about exchangeability.

As part of our Markov chain analysis, Section 4 develops an algorithm to compute to any desired precision the expected absorption time as a function of i and n.

Our work was originally motivated by a question of Cook and Rackoff concerning the expected time for mitochondrial monomorphism to occur in a fixed-size population. Their question was motivated by Cann *et al.* [2], who used branching process simulation studies of Avise *et al.* [1] as corroborating mathematical justification for the 'mitochondrial Eve hypothesis' that there is a 200 000 year old mitochondrial ancestor of all of current humanity. Avise *et al.* showed by computer simulation that if expected population size remains constant, then with probability 1 all members of some future generation will be mitochondrially monomorphic. Simulating n independent branching processes, each involving a Poisson random variable with mean 1, Avise *et al.* [1] cited $4n$ as an upper bound on the expected number of generations to absorption.

The mitochondrial Eve question is a special case of the coalescence problem. The coalescent was introduced by Kingman [15]–[17] as a method for estimating the most recent common ancestor (MRCA) of a population. (See [23], [24] for an overview of applications of the coalescent.) The fundamental result for the coalescent is that if a population evolves with discrete nonoverlapping generations of fixed size n with neutral selection, each child having only one parent, then it is expected that the $2n$th generation before the present contains a common ancestor for the entire present generation. There are a number of proofs of the coalescent theorem in the literature, e.g. [4], [18], [20]. Many of these are based on approximating an evolutionary discrete death process with a continuous Markov process, but Möhle [20] proved some tight bounds using a discrete Markov analysis, giving tight bounds of the form $2N_e(1-1/n)$ in many cases, including the binomial and hypergeometric processes. Here N_e is the *effective population size* and equals the inverse of the coalescent probability; N_e equals n in the binomial case and $2n-1$ in the hypergeometric case.

In Section 3, we use the combinatorial Markov chain analysis to give a new proof of the $O(n)$ coalescent theorem, based on the mean and weak variation conditions plus an additional assumption that lets us define the time-reversal of an evolutionary process. For instance, the binomial and hypergeometric processes, as well the generalized binomial and hypergeometric processes of Möhle [20], are proved to have $O(n)$ expected number of generations before an MRCA is reached. Unlike earlier proofs, our proof is based on the forward absorption times of the Fisher–Wright problem, rather than on an analysis of the death process.

2. Preliminaries and definitions

We restrict attention to haploid individuals, who carry a single set of genetic material and receive their genes from a single parent. However, our results can apply to diploid genes as well, as long as the genes are not sex- or reproduction-linked. We assume there is no mutation and that the evolutionary process is time-homogeneous.

The populations are assumed to consist of discrete, nonoverlapping generations, each of size n. Each individual carries one of two alleles, A or a. If i individuals of a generation have allele A, then $n - i$ have allele a and the generation is in 'state i'. The population is modeled by a Markov chain $(Q, \boldsymbol{M})$, where $Q = \{0, 1, 2, \ldots, n\}$ is the set of states and $\boldsymbol{M} = (m_{i,j})$ is a stochastic transition matrix. The *transition probability* $m_{i,j}$ is the probability that state i is followed immediately by state j. We often drop the Q from the notation and refer to $\boldsymbol{M}$ as a Markov chain.

We have $m_{0,0} = 1$ and $m_{n,n} = 1$ since the states 0 and n are *absorbing* states in which all the individuals are carrying the same allele. The Markov chain can be viewed as having stopped once it enters one of these two states. The *mean stopping time* or the *expected absorption time* is the expected number of generations until an absorbing state is reached. The mean stopping time is a function of the initial state.

The Fisher–Wright problem concerns bounding the mean stopping time. For the traditional Fisher–Wright problem, the transition probabilities are defined by $m_{i,j} = p_{i,j}$ where the $p_{i,j}$ are defined in (1). Sometimes, the probabilities $q_{i,j}$ from (2) are used instead. To generalize to a wider range of transition probabilities, we define a mean condition and two conditions on the variance of the probabilities.

Definition 1. A Markov chain satisfies the *mean condition* if, for all $n \geq 3$ and for all i such that $0 \leq i < n/2$,

$$\sum_{j=0}^{n} j m_{i,j} \leq i$$

and, symmetrically, for all i such that $n \geq i > n/2$,

$$\sum_{j=0}^{n} j m_{i,j} \geq i.$$

If i individuals have allele A, then the quantity $\sum_j j m_{i,j}$ is the expected number of individuals with allele A in the next generation. The intuition behind the mean condition is that the Markov chain does not have a tendency to drift towards the state $n/2$.

In the case of neutral selection, $\sum_j j m_{i,j} = i$ for all i, and the Markov chain is called a *martingale*.

Theorem 1. *The Markov chain with transition probabilities (1) or (2) satisfies the mean condition and is a martingale.*

This theorem is well known, and its proof is omitted.

In addition to the mean condition, we need conditions that bound the variance in the population state below. Recall that the standard deviation of the binomial distribution (1) is equal to $\sqrt{i(n-i)/n}$.

Definition 2. Define $\sigma_{i,n} = \sqrt{i(n-i)/n}$. A family of Markov chains $\boldsymbol{M}_n$ satisfies the *variation condition* provided that there are constants $\delta, \varepsilon > 0$ such that, for all $n \geq 2$ and i such that $0 < i < n$,

$$\sum_{|j-i| > \delta \sigma_{i,n}} m_{i,j} > \varepsilon. \tag{3}$$

The intuition is that the variation condition forces the state of the population to vary noticeably between generations; this, plus the mean condition, is enough to ensure that an absorbing state is reached in a finite amount of time.

The condition (3) can equivalently be stated as

$$\Pr[|C_i - i| > \delta\sigma_{i,n}] > \varepsilon,$$

where C_i is a random variable distributed according to the ith row of $\boldsymbol{M}$; note that C_i can be interpreted as the number of children of i individuals. This means that there is nonnegligible probability that the number of individuals with allele A changes by at least $\delta\sigma_{i,n}$. Since $\sigma_{i,n}$ is the standard deviation of the binomial process, we expect the binomial process to fulfill the variation condition. The intuition is that a process satisfies the variation condition provided its standard deviation is proportional to $\sigma_{i,n}$ or larger. (However, this is not a mathematically equivalent condition.)

The definition of the mean condition is stated in terms of families of Markov chains in order to allow us to prove results that hold asymptotically in n. For example, the transition probabilities given in (1) and (2) specify families of Markov chains, one for each value of $n \geq 1$. It is important for the definition that δ, ε are fixed constants, independent of i and n.

Theorem 2. *The binomial distribution (1) satisfies the variation condition.*

Theorem 3. *The hypergeometric distribution (2) satisfies the variation condition.*

These theorems are presumably not new; however, the authors have not been able to find any place where they are proved in this strong form. The usual proofs of the de Moivre–Laplace theorem for the binomial and hypergeometric distributions (cf. [9]) prove the theorems only for fixed values of $p = i/n$, whereas we need the theorems to hold for all values of i and n. We thus include the proofs of these theorems, but relegate them to Appendix A.

Our main theorems will also hold under the following weak form of the variation condition.

Definition 3. Let

$$\sigma'_{i,n} = \left(\frac{i}{n}\right)^{3/4} \left(\frac{n-i}{n}\right)^{3/4} n^{1/2}.$$

The *weak variation condition* holds for a family of Markov chains $\boldsymbol{M}_n$, provided there are constants $\delta, \varepsilon > 0$ such that, for all $n \geq 2$ and i such that $0 < i < n$,

$$\sum_{|j-i|>\delta\sigma'_{i,n}} m_{i,j} > \varepsilon.$$

Since $\sigma'_{i,n} \leq \sigma_{i,n}$, the weak variation condition is less restrictive than the variation condition. Thus, the variation condition implies the weak variation condition.

The bound $\sigma'_{i,n}$ for the weak variation condition differs most from the bound $\sigma_{i,n}$ when i is close to 0 or close to n. For these values of i, a Markov chain that satisfies the weak variation condition (but not the variation condition) is allowed to have variance substantially less than the variance of the binomial distribution. Hypothetically speaking, such a Markov chain could arise in situations where alleles that occur rarely have some reproductive advantage. For example, this can occur with human intervention, where efforts are made to preserve a rare subspecies; or it might occur if scarcity of the allele conveys some reproductive advantage.

The mean condition implies that 0 and n are absorbing states. The weak variation condition implies that no other state is absorbing.

We can now state our main theorems for populations with two alleles. Let D_i denote the expected stopping time for a Markov chain started in state i. (Although D_i depends on both i and n, we suppress the 'n' in the notation.)

Theorem 4. *Suppose that the Markov chain $\boldsymbol{M} = (m_{i,j})$ satisfies the mean condition and the variation condition. Then the expected stopping time D_i is bounded by*

$$D_i = O(nH(i/n)),$$

where $H(p) = -p \ln p - (1-p)\ln(1-p)$ is the entropy function.

Theorem 5. *Suppose that the Markov chain $\boldsymbol{M} = (m_{i,j})$ satisfies the mean condition and the weak variation condition. Then the expected stopping time D_i is bounded by*

$$D_i = O(\sqrt{i(n-i)}).$$

We prove these theorems below, in Sections 4–6. First, however, Section 3 proves a corollary about the n-allele situation and the MRCA.

3. The multi-allele case

This section extends the expected stopping time theorems to the case where there are n distinct alleles in the population. From this it proves upper bounds for expected time to coalescence under fairly general conditions.

The Markov chain model generalizes straightforwardly to multiple alleles; namely, in the multi-allele setting, a *state* consists of the numbers of individuals with each allele. The mean condition and the weak variation condition need to be redefined to apply multi-allelle evolution. For this, let $\mathcal{A}$ be any set of alleles. Let $n_{\mathcal{A}}(t)$ be the number of individuals in generation t that carry an allele from $\mathcal{A}$. Define $p_{i,j}^{\mathcal{A}}$ to be the conditional probability

$$p_{i,j}^{\mathcal{A}} = \Pr[n_{\mathcal{A}}(t+1) = j \mid n_{\mathcal{A}}(t) = i].$$

The mean condition is satisfied by the multi-allele Markov chain provided that, for every set $\mathcal{A}$ of alleles, the transition probabilities $p_{i,j}^{\mathcal{A}}$ satisfy the mean condition. The multi-allele Markov chain satisfies the (weak) variation condition provided that there are fixed constants δ, ε such that, for every $\mathcal{A}$, the probabilities $p_{i,j}^{\mathcal{A}}$ satisfy the (weak) variation condition with those values of δ, ε.

The binomial and hypergeometric processes were earlier defined for populations with two alleles, but the definitions extend naturally to the multi-allele setting. In the multi-allele setting, the binomial process is as follows: each individual in generation $i+1$ receives its allele from an independently randomly chosen individual in generation i. The hypergeometic process is now defined by letting each individual in generation i have two offspring and then selecting a randomly chosen set of n of the offspring to survive as the next generation. Clearly the multi-allele binomial and hypergeometric processes both satisfy the (multi-allele) mean and variation conditions; indeed, for any set $\mathcal{A}$, the probability $p_{i,j}^{\mathcal{A}}$ is exactly equal to the probability (1) for the multi-allele binomial process, and to (2) for the hypergeometric process.

To better understand the generality of the mean and weak variation conditions, we define some processes that satisfy these conditions yet are not martingales. The intuition will be that frequently occurring alleles confer a reproductive advantage. Let $\alpha(i)$ be a nondecreasing, positive function, which is intended to represent the relative reproductive advantage when there

are i individuals with a given allele. Let $\lambda(i) = i\alpha(i)$. The *binomial α-advantage* process is defined as follows: if there are i individuals with allele a, then each individual in the next generation has allele a with probability proportional to $\lambda(i)$. (It is also possible to define hypergeometric versions of the α-advantage process.) If there are just two alleles, this means that the probabilities for the binomial α-advantage process are defined by

$$r_{i,j} = \binom{n}{j}\left(\frac{\lambda(i)}{\lambda(i)+\lambda(n-i)}\right)^j\left(\frac{\lambda(n-i)}{\lambda(i)+\lambda(n-i)}\right)^{n-j}.$$

For two alleles, these processes satisfy the mean and variation conditions.

The binomial α-advantage process generalizes in the obvious way to multi-alleles. However, this does not necessarily satisfy the multi-allele mean condition; for example, when $\alpha(1) = 1$ and $\alpha(n/3) = 2$ and there are $n/3$ individuals with a common allele a, the remaining individuals have distinct alleles, and $\mathcal{A} = \{a\}$. Consider instead a thresholded version of the α-advantage process where the reproductive advantage applies only to alleles that form a majority of the population. For this, we require $\alpha(i) = 1$ for $i < n/2$, and $\alpha(i) > 1$ for $i \geq n/2$. The multi-allele process defined with such an α can be shown to satisfy the mean and variation conditions, but is not a martingale.

Theorem 6. *Suppose that a population begins with n individuals with distinct alleles, and evolves according to a Markov chain that satisfies the multi-allele mean and weak variation conditions. Then the expected stopping time is $O(n)$.*

Proof. Consider any set $\mathcal{A}$ of alleles. Let $\overline{\mathcal{A}}$ be the complement of $\mathcal{A}$; thus $\{\mathcal{A}, \overline{\mathcal{A}}\}$ forms a partition of the alleles. We say that $\mathcal{A}$ has *stopped* when either all individuals carry an allele from $\mathcal{A}$ or all individuals carry an allele from $\overline{\mathcal{A}}$. By Theorem 5, the expected stopping time for $\mathcal{A}$ is less than cn for some constant c. It follows that the probability that $\mathcal{A}$ stops in less than $4cn$ generations is greater than $\frac{3}{4}$.

There are 2^n sets $\mathcal{A}$. We wish to find the expected time at which more than one half of the sets $\mathcal{A}$ have stopped. We claim that the considerations in the last paragraph imply that, with probability at least $\frac{1}{2}$, more than half of the sets $\mathcal{A}$ have stopped by time $4cn$. To prove this claim, note that if α is the probability that at least a fraction β of the sets $\mathcal{A}$ have stopped by time $4cn$, then, for some $\mathcal{A}$,

$$\Pr[\mathcal{A} \text{ has not stopped by time } 4cn] \geq (1-\alpha)(1-\beta).$$

The claim is proved by setting $\alpha = \frac{1}{2} = \beta$.

Repeating this argument shows that, with probability at least $\frac{3}{4}$, more than half of the sets $\mathcal{A}$ are stopped by time $8cn$. More generally, with probability at least $1 - 1/2^i$, more than half of the sets $\mathcal{A}$ are stopped by time $4icn$. Therefore, the expected time before which more than half of the sets $\mathcal{A}$ are stopped, is bounded by

$$\sum_{i=1}^{\infty} \frac{4icn}{2^i} = 8cn.$$

To complete the proof of the theorem, we claim that once a generation is reached where more than half of the sets $\mathcal{A}$ are stopped, then all the individuals carry the same allele. To prove this claim, let A_1 and A_2 be two alleles. We say that a set $\mathcal{A}$ separates A_1 from A_2 if $A_1 \in \mathcal{A}$ and

$A_2 \in \overline{\mathcal{A}}$ or if, vice versa, $A_1 \in \overline{\mathcal{A}}$ and $A_2 \in \mathcal{A}$. If we choose a random set $\mathcal{A}$, the probability that it separates A_1 from A_2 is exactly $\frac{1}{2}$. Since more than one half of the sets $\mathcal{A}$ are stopped, there must therefore be some stopped $\mathcal{A}$ that separates A_1 from A_2. Thus, at least one of A_1 or A_2 has disappeared from the population. Since this argument applies to any pair of alleles A_1, A_2, it follows that there is only one allele left in the population.

Note that the above proof depends only on the fact that, for any subpopulation, the expected time for it to either become the entire population, or to be eliminated, is $O(n)$ generations. Thus, the property of Theorem 6 is a robust phenomenon.

Theorem 6 is similar to a coalescence theorem. The viewpoint of the coalescence problem is that the current generation is the end of an evolutionary process, and we then consider which evolutionary sequences could have led to the current generation. That is, unlike the Fisher–Wright problem, where we consider the evolution of future generations, the coalescence problem considers the possible past evolutionary processes. The *expected coalescence time* is defined to be the expected number of generations elapsed since all individuals in the present generation had a common ancestor.

The usual assumption for coalescence is that the individuals in a generation choose their parents at random from the previous generation and, as Kingman [17] notes, this is mathematically equivalent to the binomial probabilities (1). To generalize this to other evolutionary processes, such as the hypergeometric process, it is necessary to define the time-reversals of the processes. We do this only for processes of the following type.

Definition 4. A Markov process $\mathcal{M}$ on n individuals is *controlled by function probabilities* provided that there is a probability distribution $P(f)$ on the functions $f : [n] \to [n]$, where $[n] = \{1, 2, \ldots, n\}$, such that the process $\mathcal{M}$ evolves as follows: given generation t containing individuals numbered $1, 2, \ldots, n$, choose a random function f according to the distribution P and, for each i, let the ith individual of generation $t+1$ inherit the allele of the $f(i)$th individual of generation t.

As defined, these Markov processes differ from our previous notion of a Markov process since the individuals are numbered or indexed. To revert to the previous kind of Markov process, the individuals could be randomly permuted in every evolutionary step. Equivalently, the probability distribution on functions could be required to be invariant under permutations of the domain and range of the function, that is, it could be required that $P(\pi \circ f) = P(f) = P(f \circ \pi)$ for any function $f : [n] \to [n]$ and any one-to-one function $\pi : [n] \to [n]$.

The binomial process can be defined as a Markov process controlled by function probabilities by letting $P(f)$ be equal to $1/n^n$, i.e. each function is equally likely. The hypergeometric process can likewise be defined as being controlled by function probabilities; namely, to choose a random f, choose uniformly at random a one-to-one function $m : [n] \to [2n]$, and then choose f so that $f(x) = \lfloor m(x)/2 \rfloor$. In the hypergeometric case, the functions f do not all have equal probabilities.

Theorem 7. *Let a multi-allele Markov process be controlled by function probabilities and satisfy the multi-allele mean and weak variation conditions. Then the expected coalescence time is* $O(n)$.

Proof. This is an immediate corollary of Theorem 6. Note that the evolution through a series of k generations can be represented by a sequence of k functions $f_1, \ldots, f_k$. The probability of this evolutionary sequence is the product $\prod_i P(f_i)$. By Theorem 6, the expected value of k such that $f_1, \ldots, f_k$ causes coalescence is $O(n)$.

The function probability model for Markov processes is quite general; for instance, it includes the examples of processes for which Möhle [20, Section 5] has proved coalescence theorems (with the exception of the more slowly evolving Moran model). However, as formulated above, the function probability model applies mainly to martingales, especially if the functions are required to be invariant under permutations of the range and domain. This is somewhat unexpected since, if the mean condition holds, the failure of the martingale property would be expected to only decrease the mean stopping time. It would be worthwhile to have more general techniques for formulating the time-reversal of an evolutionary process.

The method of duality, used by Möhle [19], is another technique for relating forward and backward evolutionary processes.

4. Calculating stopping times

We now discuss an algorithm for calculating the exact stopping times for a Markov chain, for particular values of n. In addition, we develop some properties of the stopping time that are needed later for the proofs of the stopping time theorems.

We could try running random trials to try to determine stopping times experimentally. This turns out to be difficult, since the stopping times have a fairly large standard deviation and a large number of trials are needed to accurately measure them. Instead, we describe an iterative algorithm that converges to the values D_i, the stopping times for a population of size n that starts with i individuals with allele A and the rest with a.

Let $\boldsymbol{M}^k$ be the kth power of the Markov chain matrix $\boldsymbol{M}$. The entries of $\boldsymbol{M}^k$ are denoted by $m_{i,j}^{(k)}$ and are transition probabilities for k-generation steps. We let $\boldsymbol{M}^\infty$ equal the limit of $\boldsymbol{M}^k$ as $k \to \infty$; its entries are $m_{i,j}^\infty = \lim_k m_{i,j}^{(k)}$. The matrices $\boldsymbol{M}^k$ and $\boldsymbol{M}^\infty$ are also stochastic. A state $i_0 \in Q$ is said to be *absorbing* if $m_{i_0,i_0} = 1$. State j is said to be *accessible* from state i or, equivalently, i can *reach* j, if there exists a t for which $m_{i,j}^{(t)} > 0$. The mean and (weak) variation conditions imply that 0 and n are the only absorbing states.

It is common to analyze Markov chains using eigenvalues and eigenvectors of the transition matrix $\boldsymbol{M}$. Indeed, this has been done by Feller [8] for the binomial probabilities and by Cannings [3] for more general transition matrices under the assumption of exchangeability. We will not use eigenvectors or eigenvalues however.

For a Markov chain $\mathcal{M}$ with state set $Q = \{0, \ldots, n\}$, define the *mean stopping time vector* $\boldsymbol{D} = (D_0, \ldots, D_n)$ so that D_i is the expected number of transitions before $\mathcal{M}$, starting in state i, enters an absorbing state. Since 0 and n will be the only absorbing states, $D_0 = 0 = D_n$. From state i, in one step the Markov chain makes a transition to state j with probability $m_{i,j}$; hence,

$$D_i = \begin{cases} 0 & \text{if } i \in \{0, n\}, \\ 1 + \sum_{j=0}^{n} m_{i,j} D_j & \text{if } 0 < i < n. \end{cases} \tag{4}$$

Proposition 1. *If $\boldsymbol{M} = (m_{i,j})$ satisfies the mean condition, and 0 and n are the only absorbing states, then there is at most one vector $\boldsymbol{D}$ that satisfies (4).*

Before proving Proposition 1, we establish the following three lemmas. (Results similar to Lemmas 1 and 2 can be found in Möhle [19].)

Lemma 1. *Suppose that $\boldsymbol{M}$ satisfies the mean condition, and that 0 and n are the only absorbing states.*

(a) *If $0 < i < n/2$ then there is a $j < i$ such that $m_{i,j} \neq 0$.*

(b) *If $n > i > n/2$ then there is a $j > i$ such that $m_{i,j} \neq 0$.*

(c) *If $i = n/2$ then there is a $j \neq i$ such that $m_{i,j} \neq 0$.*

Proof. Part (c) is obvious from the fact that $n/2$ is not absorbing. To prove (a), suppose that $0 < i < n/2$. Since state i is not absorbing, $m_{i,i} \neq 1$. By the stochastic property, there are values $j \neq i$ such that $m_{i,j} > 0$. By the mean property, not all of these values of j are greater than or equal to i. This proves (a). Part (b) is proved similarly.

Lemma 2. *Suppose that $\boldsymbol{M}$ satisfies the mean condition, and that 0 and n are the only absorbing states. Then there is a value r, with $1 \leq r \leq n/2$, such that $m_{i,0}^{(r)} + m_{i,n}^{(r)} \neq 0$ for all i. That is, from any starting state i, there is nonzero probability of reaching an absorbing state within $\lfloor n/2 \rfloor$ steps.*

Proof. This is a simple consequence of Lemma 1. For $i = i_0 < n/2$, let $i_0 > i_1 \geq i_2 \geq \cdots$ be chosen so that, for all k, if $i_k > 0$, then $m_{i_k,i_{k+1}} \neq 0$ and $i_{k+1} < i_k$. Then, clearly, $m_{i,i_k}^{(k)} \neq 0$ for all $k > 0$. Since the values i_k are decreasing down to zero, the sequence reaches zero in at most i steps. Thus, for $i < n/2$, $m_{i,0}^{(i)} \neq 0$. A symmetric argument shows that, for $i > n/2$, $m_{i,n}^{(n-i)} \neq 0$. Similarly, when n is even, at least one of $m_{n/2,0}^{(n/2)}$ or $m_{n/2,n}^{(n/2)}$ is nonzero.

Lemma 3. *Let $\boldsymbol{M}$ satisfy the mean condition, and let 0 and n be the only absorbing states. Then the entries of $\boldsymbol{M}^\infty$ are 0 except in the first and last columns.*

Lemma 3 implies that an absorbing state is eventually reached with probability 1.

Proof. This follows from Lemma 2. Choose $r \leq n/2$ and

$$\varepsilon = \min\{m_{i,0}^{(r)} + m_{i,n}^{(r)} : 0 \leq i \leq n\},$$

so that $\varepsilon > 0$. Then $(\boldsymbol{M}^r)^k = \boldsymbol{M}^{kr}$ has the property that all its entries outside the first and last columns are bounded above by $(1-\varepsilon)^k$. More precisely, for each row i, $\sum_{j=1}^{n-1} m_{i,j} \leq (1-\varepsilon)^k$. From this, it is immediate that the limit of these matrices exists and has zero entries everywhere except in the first and last columns.

We are now ready to prove Proposition 1.

Proof of Proposition 1. Suppose that $\boldsymbol{A} = (A_0, \ldots, A_n) \neq \boldsymbol{D}$ also satisfies the equation (4); hence, $A_0 = 0 = A_n$, so $A_0 = D_0$, $A_n = D_n$. Define $\boldsymbol{C} = (C_0, \ldots, C_n)$ by $C_i = A_i - D_i$. It follows that $C_0 = 0 = C_n$ and, for i such that $0 < i < n$,

$$C_i = \sum_{j=0}^{n} m_{i,j}(A_j - D_j) = \sum_{j=0}^{n} m_{i,j} C_j.$$

In other words, $\boldsymbol{C} = \boldsymbol{M}\boldsymbol{C}$. By induction on k, $\boldsymbol{C} = \boldsymbol{M}^k\boldsymbol{C}$. By taking the limit as $k \to \infty$, $\boldsymbol{C} = \boldsymbol{M}^\infty\boldsymbol{C}$. Then, by Lemma 3, $\boldsymbol{C}$ is the zero vector; that is, $\boldsymbol{A} = \boldsymbol{D}$.

Proposition 1 established the uniqueness of a solution to the equation (4). The next proposition establishes the existence of a solution and will form the basis of our algorithm for computing mean stopping times for particular values of n.

Define $\boldsymbol{E}^{(s)} = (E_0^{(s)}, \ldots, E_n^{(s)})$ by setting $E_i^{(0)} = 0$ when $0 \le i \le n$, and

$$E_i^{(s+1)} = \begin{cases} 0 & \text{if } i \in \{0, n\}, \\ 1 + \sum_{j=0}^{n} m_{i,j} E_j^{(s)} & \text{if } 0 < i < n. \end{cases} \tag{5}$$

This is more succinctly expressed by letting $\mathbf{1}_0 = (0, 1, 1, \ldots, 1, 1, 0)$ be the column vector of length $n + 1$, consisting of 1s, with the exception of the first and last coordinates, and defining

$$\begin{aligned} \boldsymbol{E}^{(0)} &= (0, 0, \ldots, 0, 0), \\ \boldsymbol{E}^{(s+1)} &= \mathbf{1}_0 + \boldsymbol{M}\boldsymbol{E}^{(s)}. \end{aligned} \tag{6}$$

Proposition 2. *If $\boldsymbol{M}$ satisfies the mean condition, and 0 and n are the only absorbing states, then the sequence $(\boldsymbol{E}^{(s)})_s$ is, in each component, nondecreasing and bounded. The limit $\boldsymbol{E}^{(\infty)}$ satisfies (4) and, hence, equals the vector $\boldsymbol{D}$ of mean stopping times.*

Proof. We prove that $\boldsymbol{E}_i^{(s+1)} \ge \boldsymbol{E}_i^{(s)}$ for all i by induction on s. For the base case, $E_0^{(0)} = E_0^{(1)} = 0 = E_n^{(0)} = E_n^{(1)}$ and $E_i^{(0)} = 0 < 1 = E_i^{(1)}$ for $0 < i < n$. For the inductive case, $E_i^{(s+1)} - E_i^{(s)} = \sum_{j=1}^{n-1} m_{i,j}(E_j^{(s)} - E_j^{(s-1)})$, where, by the induction hypothesis, $E_j^{(s)} - E_j^{(s-1)} \ge 0$. Since $0 \le m_{i,j} \le 1$, the claim follows.

We now show the existence of an upper bound L such that $E_i^{(s)} \le L$ when $0 \le i \le n$ and $s \ge 0$. By Lemma 2, for $r = \lfloor n/2 \rfloor$ there is an $\varepsilon > 0$ such that

$$m_{i,0}^{(r)} + m_{i,n}^{(r)} \ge \varepsilon$$

whenever $0 \le i \le n$. From the proof of Lemma 3, $\boldsymbol{M}^{rk}\mathbf{1}_0 \le (1 - \varepsilon)^k \mathbf{1}_0$. It follows that, for $t = rm + s$, where $0 \le s < r$,

$$\begin{aligned} \boldsymbol{E}^{(t)} = \sum_{i=0}^{t-1} \boldsymbol{M}^i \mathbf{1}_0 &\le \left(\sum_{i<mr} \boldsymbol{M}^i + \boldsymbol{M}^{mr} \sum_{i<s} \boldsymbol{M}^i \right) \mathbf{1}_0 \\ &\le \left(\sum_{i<r} \boldsymbol{M}^i \right) \left(\sum_{j<m} \boldsymbol{M}^{rj} \right) \mathbf{1}_0 + \sum_{i<s} \boldsymbol{M}^i \boldsymbol{M}^{mr} \mathbf{1}_0 \\ &\le \left(\sum_{i<r} \boldsymbol{M}^i \right) \left(\sum_{j<m} (1-\varepsilon)^j \right) \mathbf{1}_0 + \sum_{i<s} \boldsymbol{M}^i (1-\varepsilon)^m \mathbf{1}_0 \\ &\le \frac{1}{\varepsilon} \left(\sum_{i<r} \boldsymbol{M}^i \right) \mathbf{1}_0 . \end{aligned}$$

This provides an explicit upper bound.

For each i, the values $E_i^{(s)}$ form a nondecreasing sequence bounded above, so the limit $E_i^{(\infty)} = \lim_{s\to\infty} E_i^{(s)}$ exists. Applying limits to both sides of (6), it follows that $\boldsymbol{E}^{(\infty)}$ satisfies (4); hence, by Proposition 1, it must be equal to $\boldsymbol{D}$.

Proposition 3. *Let $\boldsymbol{M} = (m_{i,j})$ satisfy the mean condition, and let 0 and n be the only absorbing states. Suppose that $F_0 = 0 = F_n$ and that, when $0 < i < n$, $F_i \geq 0$ and*

$$F_i \geq 1 + \sum_{j=0}^{n} m_{i,j} F_j. \tag{7}$$

Then $F_i \geq D_i$ whenever $0 \leq i \leq n$.

Proof. By induction on $s \geq 0$, we show that $\boldsymbol{F} \geq \boldsymbol{E}^{(s)}$. This is clear for $\boldsymbol{E}^{(0)}$, and in the inductive case, by (7), we have $\boldsymbol{F} - \boldsymbol{E}^{(s+1)} \geq \boldsymbol{M}(\boldsymbol{F} - \boldsymbol{E}^{(s)})$. The entries of $\boldsymbol{M}$ are nonnegative, so, applying the induction hypothesis, the proof is complete.

From Propositions 2 and 3, the following algorithm is guaranteed to correctly compute the mean stopping time to any precision ε.

Algorithm 1. (*Mean stopping time.*) Let $\boldsymbol{M} = (m_{i,j})$ be the transition matrix of a Markov chain that satisfies the mean condition, and for which 0 and n are the only absorbing states.

```
ε = 0.001; s = 0; E = (0, 0, ..., 0, 0); finished = false
while ( not finished ) {
    E' = 1_0 + ME; max = max_{0≤i≤n}(E'_i − E_i)
    if (max < ε) {        // if appear to have accuracy ε
      F = E + ε 1_0; F' = 1_0 + MF
      min = min_{0≤i≤n}(F_i − F'_i)
      if (min ≥ 0 ) finished = true
    }
}
return E
```

The algorithm admits an obvious extension to multiple alleles; however, the algorithm's space requirement for ℓ alleles is $O(n^{\ell-1})$ plus the space, if any, needed to store the transition matrix.

By Proposition 3, Theorem 4 can be proved by showing that there is a fixed constant $c > 0$ (with c independent of n) such that, when $0 < i < n$,

$$cnH(i/n) \geq 1 + \sum_{j=0}^{n} m_{i,j} cnH(j/n).$$

Letting $\alpha = 1/c$, this means that Theorem 4 follows from the following lemma.

Lemma 4. *Suppose that the transition probabilities satisfy the mean and the variation conditions. Then there is a constant $\alpha > 0$ (independent of n) such that, when $0 < i < n$,*

$$nH(i/n) \geq \alpha + \sum_{j=0}^{n} m_{i,j} nH(j/n). \tag{8}$$

Similarly, to prove Theorem 5, it suffices to prove the next lemma.

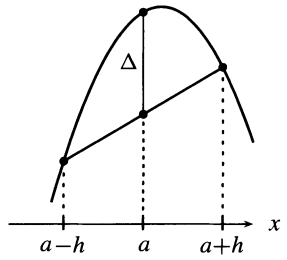

FIGURE 1: The function f and a secant line. The left-hand side of the inequality (10) is equal to Δ, the vertical distance between $f(x)$ and the secant line at $x = a$.

Lemma 5. *Suppose that the transition probabilities satisfy the mean and the weak variation conditions. Then there is a constant $\alpha > 0$ (independent of n) such that, when $0 < i < n$,*

$$\sqrt{i(n-i)} \geq \alpha + \sum_{j=0}^{n} m_{i,j}\sqrt{j(n-j)}. \tag{9}$$

These lemmas are proved in Section 6.

5. Some lemmas on secants and tangents

5.1. On vertical distance from a secant

Consider a function f, let $h > 0$, and consider the secant to $f(x)$ at the points $x = a \pm h$, as shown in Figure 1. The next theorem gives a lower bound on the difference Δ between the value of $f(a)$ and the y-coordinate of the secant line at $x = a$ (see Figure 1). We will state the theorem only for the situation where the second derivative of f is concave down, but of course it could be generalized somewhat.

Theorem 8. *Let $f : \mathbb{R} \to \mathbb{R}$ have continuous second derivative and assume that its second derivative is concave down. Let $a, h \in \mathbb{R}$ with $h > 0$. Then*

$$f(a) - \frac{f(a+h) + f(a-h)}{2} > -\frac{h^2}{2} f''(a). \tag{10}$$

Proof. Fix a and let $g(h)$ be equal to the left-hand side of the inequality (10). Clearly $g(0) = 0$. In addition, its first derivative satisfies

$$\begin{aligned} g'(h) &= -\tfrac{1}{2}(f'(a+h) - f'(a-h)) = -\frac{1}{2}\int_{-h}^{h} f''(a+x)\,\mathrm{d}x \\ &= -\frac{1}{2}\int_{0}^{h} (f''(a+x) + f''(a-x))\,\mathrm{d}x \\ &\geq -\frac{1}{2}\int_{0}^{h} 2f''(a)\,\mathrm{d}x \quad \text{(by the concavity of } f'') \\ &= -hf''(a). \end{aligned}$$

Therefore,

$$g(h) = \int_0^h g'(y)\,\mathrm{d}y \geq \int_0^h -yf''(a)\,\mathrm{d}y \geq -\frac{h^2}{2} f''(a),$$

and the theorem is proved.

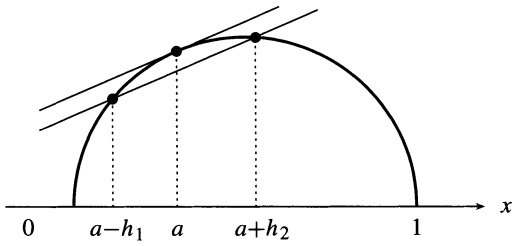

FIGURE 2: Illustration of Theorem 9.

5.2. On secants parallel to tangents of $\sqrt{p(1-p)}$

Let f henceforth be the function $f(p) = \sqrt{p(1-p)}$, defined on the interval $[0, 1]$. We would like to establish the following theorem about parallel secant lines and tangent lines to f. For future reference, we compute the first and second derivatives of f:

$$f'(x) = \frac{1-2x}{2\sqrt{x(1-x)}} \quad \text{and} \quad f''(x) = \frac{-1}{4(x(1-x))^{3/2}}.$$

It is easy to verify that the second derivative is concave down.

Theorem 9. *Suppose that* $0 \leq a - h_1 < a < a + h_2 \leq 1$. *Suppose further that*

$$f'(a) = \frac{f(a+h_2) - f(a-h_1)}{h_2 + h_1}. \tag{11}$$

Then $h_1 \leq 3h_2$ *and* $h_2 \leq 3h_1$.

The situation of Theorem 9 is sketched in Figure 2. The equation (11) says that the slope of the secant line containing the points $(a - h_1, f(a - h_1))$ and $(a + h_2, f(a + h_2))$ is equal to the slope of the line tangent to f at $f(a)$. There are several simple observations to make. First, by the concavity of f, if the values of a and h_2 are fixed, then there is at most one value for $h_1 \in [0, 1]$ (respectively, $h_2 \in [0, 1]$) such that (11) holds; equally, if the values of a and h_1 are fixed, then there is at most one value for $h_2 \in [0, 1]$ such that (11) holds. Second, since f' is a strictly decreasing function, the value of a is uniquely determined by the values of $a - h_1$ and $a + h_2$. Third, the theorem is easily seen to be true for $a = \frac{1}{2}$ since, in that case, $h_1 = h_2$. Fourth, since f is symmetric around $a = \frac{1}{2}$, with $f(x) = f(1 - x)$, it suffices to prove that $h_2 \leq 3h_1$, as $h_1 \leq 3h_2$ will then follow by symmetry.

Before starting the proof of Theorem 9, it is useful to note that the graph of the function $f(x) = \sqrt{x(1-x)}$ forms the upper half of the circle of radius $\frac{1}{2}$ with center at the point $(\frac{1}{2}, 0)$. To prove this, just note that

$$\sqrt{x(1-x)} = \sqrt{(\tfrac{1}{2})^2 - (x - \tfrac{1}{2})^2}.$$

Therefore, the situation of Theorem 9 is as illustrated in Figure 3. In the figure, the center of the semicircle is labeled P and the three points on the graph of $f(x)$ at $x = a - h_1$, a, and $a + h_2$ are labeled A, B and C. In order for the secant $\overline{AC}$ to be parallel to the tangent at B, it is necessary and sufficient that the angles $\angle APB$ and $\angle BPC$ are equal to the same value, θ. We also let φ be the angle $\angle OPA$, where O is the point $(0, 0)$.

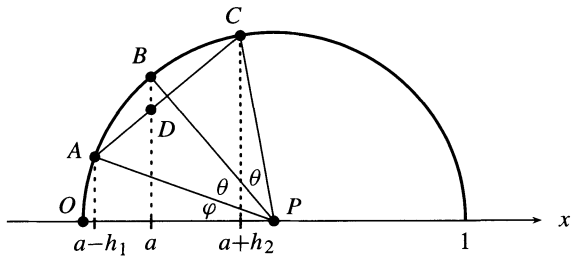

FIGURE 3: The situation in the proof of Theorem 9.

We first prove that $h_2 < 3h_1$, under the assumption that $\varphi = 0$: this is the same as the assumption that $a - h_1 = 0$. Clearly, it is sufficient to prove that

$$\frac{h_1 + h_2}{h_1} < 4.$$

Since $\varphi = 0$, we have $h_1 = a = (1 - \cos\theta)/2$ and $h_1 + h_2 = (1 - \cos(2\theta))/2$. So,

$$\frac{h_1 + h_2}{h_1} = \frac{1 - \cos(2\theta)}{1 - \cos\theta} = \frac{2(1 - \cos^2\theta)}{1 - \cos\theta} = 2(1 + \cos\theta),$$

which is clearly bounded by 4 since $\theta < \pi/2$.

It remains to prove the theorem for nonzero values of φ. However, we claim that the $\varphi = 0$ case is actually the worst case. To see this, consider Figure 3 again. Suppose that we keep the angle θ fixed and let φ vary: then the values a, h_1, and h_2 and the points A, B, C, and D are all functions of φ. We use $|AD|$ to denote the distance from A to D. Then, by similar triangles,

$$\frac{|DC|}{|AD|} = \frac{h_2}{h_1}.$$

Clearly, $|DC|$ is a decreasing function of φ and $|AD|$ is an increasing function of φ, so the value of the fraction h_2/h_1 decreases as φ increases. We have already proved that $h_2/h_1 < 3$ for $\varphi = 0$, so this holds for all values of φ. This completes the proof of Theorem 9.

The following is a corollary of Theorems 8 and 9 and the fact that f'' is concave down.

Corollary 1. *Let f, a, h_1, and h_2 be as in the hypothesis of Theorem 9. Then*

$$f(a) - \frac{h_2 f(a - h_1) + h_1 f(a + h_2)}{h_1 + h_2} \geq -\frac{h_i^2}{18} f''(a)$$

for both $i = 1$ and $i = 2$.

Proof. Let $h_{\min} = \min\{h_1, h_2\}$. Consider the two secant lines S_1 and S_2, where S_1 is secant to the graph $f(x)$ at $x = a - h_1$ and $x = a + h_2$, and S_2 is the secant line at $x = a - h_{\min}$ and $x = a + h_{\min}$. The value $(h_2 f(a - h_1) + h_1 f(a + h_2))/(h_1 + h_2)$ is the y-coordinate of the secant line S_1 at $x = a$. By the fact that f is concave down, the secant line S_2 is above the line S_1 when $a - h_{\min} < x < a + h_{\min}$, and, in particular, S_2 is above S_1 at $x = a$. By Theorem 8,

$$f(a) - \frac{h_2 f(a - h_1) + h_1 f(a + h_2)}{h_1 + h_2} \geq -\frac{h_{\min}^2}{2} f''(a).$$

By Theorem 9, for $i = 1, 2$, we have $h_i \leq 3h_{\min}$, and this proves the corollary.

5.3. On secants parallel to tangents of the entropy function

The entropy function $H(p) = -p\ln(p) - (1-p)\ln(1-p)$ is defined on the interval $[0, 1]$. Its first and second derivatives are

$$H'(p) = -\ln(p) + \ln(1-p) = \ln\left(\frac{1-p}{p}\right) = \ln\left(\frac{1}{p} - 1\right),$$
$$H''(p) = \frac{-1}{1-p} - \frac{1}{p} = \frac{-1}{(1-p)p}.$$

It is easy to check that $H'(p)$ is strictly decreasing, and is concave up for $p \leq \frac{1}{2}$ and concave down for $p \geq \frac{1}{2}$. Also, $H''(p)$ is concave down.

The next theorem states that $H(p)$ satisfies the same kind of property that Theorem 9 established for $\sqrt{p(1-p)}$.

Theorem 10. *Suppose that* $0 \leq a - h_1 < a < a + h_2 \leq 1$. *Suppose further that*

$$H'(a) = \frac{H(a+h_2) - H(a-h_1)}{h_2 + h_1}. \tag{12}$$

Then $h_1 \leq (\mathrm{e}-1)h_2$ *and* $h_2 \leq (\mathrm{e}-1)h_1$, *where* e *is the base of the natural logarithm.*

Proof. We shall prove a weaker form of this theorem, namely that there is a constant c such that $h_1 \leq ch_2$ and $h_2 \leq ch_1$, and then appeal to experimental results obtained by graphing functions in MATHEMATICA® to conclude that $c = \mathrm{e} - 1$ works.

The entropy function $H(p)$ is qualitatively similar to the function $f(p) = \sqrt{p(1-p)}$ in that it is concave down, is zero at $p = 0$ and $p = 1$, and attains its maximum at $p = \frac{1}{2}$. So, Figure 2 can also serve as a qualitative illustration of Theorem 10. We introduce new variables, and let $r = a - h_1$ and $s = a + h_2$. Suppose that the values r, a, s satisfy (12), that is, they satisfy

$$H'(a) = \frac{H(s) - H(r)}{s - r}. \tag{13}$$

Then the values r, a, and s are dependent in that any two of them determine the third. To prove the theorem, we must give upper and lower bounds on the ratio $h_2/h_1 = (s-a)/(a-r)$. The values of r and s come from the set where $0 \leq r < s \leq 1$. The problem is that this set is not compact by virtue of having an open boundary along the line $r = s$. Even worse, the ratio is essentially discontinuous at $r = s = 0$ and at $r = s = 1$. For our proof, we will examine the values of the ratio along the line $r = s$ and at $r = 0$ (the case of $s = 1$ is symmetric), and argue by compactness of the remaining values of r and s that the ratio attains a finite maximum value and a positive minimum value.

We assume without loss of generality that $a < \frac{1}{2}$. When also $s < \frac{1}{2}$, we have $h_1 < h_2$ by the fact that $H'(p)$ is concave up for $p \leq \frac{1}{2}$ and the fact that $H'(a)$ is the average slope of $H(p)$ for $p \in [r, s]$. Thus, for $s < \frac{1}{2}$, it suffices to prove that $h_1/(h_1 + h_2) \geq 1/\mathrm{e}$ and thereby obtain that $h_2 \leq (\mathrm{e}-1)h_1$.

We start by proving the theorem in the case that $r = 0$. In this case,

$$\frac{H(s) - H(0)}{s - 0} = \frac{H(s)}{s} = \frac{-s\ln(s) - (1-s)\ln(1-s)}{s}$$
$$= \ln(s^{-1}(1-s)^{-(1-s)/s}).$$

To find the value of a such that $H'(a)$ equals this last value, we need to solve

$$\frac{1}{a} - 1 = s^{-1}(1-s)^{-(1-s)/s}.$$

We are really interested in the value of a/s since, with $r = 0$, we have $a = h_1$ and $s = h_1 + h_2$, and we need to establish $a/s \geq 1/\mathrm{e}$. Solving for a/s gives

$$\frac{a}{s} = \frac{(1-s)^{(1-s)/s}}{s(1-s)^{(1-s)/s} + 1}. \tag{14}$$

It is easy to check that $\lim_{s\to 0^+}(1-s)^{1/s} = 1/\mathrm{e}$. Therefore, as $s \to 0^+$, the quantity (14) approaches the limit $1/\mathrm{e}$.

Now consider the case when $r = 0$ and $0 < s \leq 1$ (so $a_1 \leq \frac{1}{2}$). Let $R(s) = a/s = Y/(sY+1)$, where $Y(s) = (1-s)^{(1-s)/s}$. The first derivative of R is

$$R' = \frac{Y' - Y^2}{(sY+1)^2}, \tag{15}$$

with

$$Y' = Y\left(-\frac{1}{s^2}\ln(1-s) - \frac{1}{s}\right).$$

Note that the numerator of (15) is equal to

$$Y\left[-\frac{1}{s^2}\ln(1-s) - \frac{1}{s} - (1-s)^{(1-s)/s}\right]. \tag{16}$$

The power series expansion for $\ln(1-s)$ shows that $\ln(1-s) < -s - s^2$ when $0 < s < 1$. Thus, (16) is positive and, hence, $R(s)$ is increasing and $1/e < R(s) \leq 1$ when $0 < s \leq 1$.

We have proved the $r = 0$ case of the theorem (and, by symmetry, the $s = 1$ case). We now consider the case where $r \approx s$. First, if $a = \frac{1}{2}$, then of course $s - a = a - r$, so $h_1 = h_2$ and the theorem is satisfied. More generally, compactness and continuity considerations imply that there is a $\delta > 0$ such that, if $|a - \frac{1}{2}| < \delta$, then $1/(\mathrm{e}-1) < (s-a)/(a-r) < \mathrm{e} - 1$. (If $\delta = \frac{1}{2}$ works, we are done, but for now we just know that there is some such $\delta > 0$.) Now, fix a value of $a < \frac{1}{2} - \delta$. We consider values of r and s that correspond to this value for a. Again, $h_2 = s - a$ and $h_1 = a - r$. We are thinking of h_1 and h_2 increasing in such a way that a stays fixed.

We claim that, as h_1 and h_2 increase, the ratio h_2/h_1 is increasing, at least for h_2 and h_1 not too large. In order to prove this, it is equivalent to prove that

$$\frac{\mathrm{d}h_2/\mathrm{d}h_1}{h_2/h_1} > 1.$$

With a fixed, taking the first derivative of (12) gives

$$0 = -\frac{\mathrm{d}h_1 + \mathrm{d}h_2}{(h_1+h_2)^2}(H(a+h_2) - H(a-h_1)) + \frac{H'(a+h_2)\,\mathrm{d}h_2 + H'(a-h_1)\,\mathrm{d}h_1}{h_1+h_2}.$$

So, using (12) again and multiplying by $h_1 + h_2$,

$$0 = -(\mathrm{d}h_1 + \mathrm{d}h_2)H'(a) + H'(a+h_2)\,\mathrm{d}h_2 + H'(a-h_1)\,\mathrm{d}h_1.$$

Algebraic manipulation transforms this into

$$\frac{\mathrm{d}h_2/\mathrm{d}h_1}{h_2/h_1} = \frac{h_1(H'(a-h_1) - H'(a))}{h_2(H'(a) - H'(a+h_2))}. \tag{17}$$

The derivative H' is concave up and decreasing on $[0, \frac{1}{2}]$; thus, when $a-h_1 \geq 0$ and $a+h_2 < \frac{1}{2}$,

$$h_1(H'(a-h_1) - H'(a)) > 2\int_{a-h_1}^{a} (H'(x) - H'(a))\,\mathrm{d}x, \tag{18}$$

$$h_2(H'(a) - H'(a+h_2)) < 2\int_{a}^{a+h_2} (H'(a) - H'(x))\,\mathrm{d}x. \tag{19}$$

Since $\int_{a-h_1}^{a} H'(x)\,\mathrm{d}x = H(a) - H(a-h_1)$ and $\int_{a}^{a+h_2} H'(x)\,\mathrm{d}x = H(a+h_2) - H(a)$ and using (12), the right-hand sides of (18) and (19) are equal. Therefore, (17) is greater than 1. This shows that the ratio h_1/h_2 is decreasing as long as $a + h_2 \leq \frac{1}{2}$ and $a - h_1 \geq 0$. If the Markov chain reaches $a - h_1 = 0$ with $a + h_2 \leq \frac{1}{2}$, then since we have already proved that $h_1/h_2 > 1/\mathrm{e}$ at $a - h_1 = 0$, it follows that $h_1/h_2 > 1/\mathrm{e}$ for all values of h_1 and h_2 for this a.

On the other hand, if the Markov chain stops with $a + h_2 = \frac{1}{2}$ and $a - h_1 > 0$, it is sufficient to prove the following fact: for all r, a, and s with $r \leq s - \delta$, we have $h_1/h_2 > 1/\mathrm{e}$. Now the set of points r, s, with $0 \leq r \leq s - \delta$ and $s \leq 1$, is compact, and the ratio h_1/h_2 is a continuous positive function of r and s. Thus, the ratio attains a minimum on this set. By graphing the function with MATHEMATICA it is seen that h_1/h_2 is bounded below by $1/\mathrm{e}$, by a fair margin. Thus, we have proved the theorem. (If the reader dislikes the use of MATHEMATICA here, then he or she can take this as a proof that there exists some constant $c > 0$, rather than as a proof that $c = \mathrm{e} - 1$ works.)

Corollary 2. *Let H, a, h_1, and h_2 be as in the hypothesis of Theorem 10. Then*

$$H(a) - \frac{h_2 H(a-h_1) + h_1 H(a+h_2)}{h_1 + h_2} \geq -\frac{h_i^2}{2(\mathrm{e}-1)^2} H''(a)$$

for both $i = 1$ and $i = 2$.

The proof of this corollary is identical to the proof of Corollary 1.

6. Proofs of the main theorems

This section presents the proofs of Lemmas 4 and 5, thus completing the proofs of Theorems 4 and 5.

6.1. Proof of Lemma 5, the weak variation condition lemma

Let $f(x) = \sqrt{x(1-x)}$. Dividing (9) by n, we need to prove that, for some $\alpha > 0$,

$$f\left(\frac{i}{n}\right) \geq \frac{\alpha}{n} + \sum_{j=0}^{n} m_{i,j} f\left(\frac{j}{n}\right) \quad \text{whenever } 0 < i < n. \tag{20}$$

Fix i and assume without loss of generality that $i \leq n/2$; a symmetric argument will work for $i \geq n/2$. It will help to work with vectors in $\mathbb{R}^2$, and we define $\boldsymbol{P}_j$ to equal the following point (or vector) in $\mathbb{R}^2$:

$$\boldsymbol{P}_j = \left(j, f\left(\frac{j}{n}\right)\right).$$

Consider the summation

$$\boldsymbol{P} = \sum_{j=0}^{n} m_{i,j} \boldsymbol{P}_j$$

(although $\boldsymbol{P}$ depends on i, we suppress any mention of i in the notation). We want to establish an upper bound on the second coordinate of $\boldsymbol{P}$. First, however, consider the first component of $\boldsymbol{P}$. The mean condition implies that $\sum_{j=0}^{n} j m_{i,j} \leq i$ since $i \leq n/2$ (except that, if this condition fails for $i = n/2$, then $i = n/2$ has to be handled in the symmetric argument for the case $i \geq n/2$). Therefore, the first coordinate of $\boldsymbol{P}$ is less than or equal to i.

To bound the second coordinate, let $\mathcal{J}$ be the set of values j such that $|j - i| > \delta\sigma'_{i,n}$, where δ is the value from the weak variation condition. Then

$$\boldsymbol{P} = \sum_{j \notin \mathcal{J}} m_{i,j} \boldsymbol{P}_j + \sum_{j \in \mathcal{J}} m_{i,j} \boldsymbol{P}_j.$$

Let $a = i/n$ and let T be the line tangent to the graph of $f(x)$ at $x = a$. Set $h_2 = \delta\sigma'_{i,n}/n$ and then choose h_1 so that the secant line S which is secant to f at $x = a - h_1$ and $x = a + h_2$ is parallel to T. That is to say, we are in the situation of Theorem 9. Thus, since $a \leq \frac{1}{2}$, we have $h_1 < h_2 \leq 3h_1$.

As f is concave down, geometric considerations imply that each point $\boldsymbol{P}_j$ is on or below the tangent line T. Also, for every $j \in \mathcal{J}$, either $j/n < a - h_1$ or $j/n > a + h_2$. Therefore, again since f is concave down, for every $j \in \mathcal{J}$, the point $\boldsymbol{P}_j$ is on or below the secant line S. By the weak variation condition,

$$\sum_{j \in \mathcal{J}} m_{i,j} \geq \varepsilon,$$

so in particular, the total weight of the points $\boldsymbol{P}_j$ which lie below the secant line S is greater than or equal to ε. Define R to be the line parallel to T and S, lying between those lines, so that the distance from T to R is equal to ε times the distance from T to S. Since all the points $\boldsymbol{P}_j$ are on or below T, and the sum of the coefficients of $\boldsymbol{P}_j$ below the line S is no less than ε, the point $\boldsymbol{P}$ lies on or below the line R.

Let $\pi_2(\boldsymbol{P})$ denote the y-component of $\boldsymbol{P}$, i.e. the value of the summation in (20). Since the slope of R is nonnegative and the first coordinate of $\boldsymbol{P}$ is no greater than i, the value $f(i/n) - \pi_2(\boldsymbol{P})$ is greater than or equal to ε times the vertical distance between $f(x)$ and the secant line S at $x = a = i/n$. Thus, by Corollary 1,

$$\begin{aligned} f\left(\frac{i}{n}\right) - \pi_2(\boldsymbol{P}) &\geq -\varepsilon \frac{(h_2)^2}{18} f''(a) = -\varepsilon \frac{(\delta\sigma'_{i,n}/n)^2}{18} f''(a) \\ &= -\varepsilon \frac{\delta^2 i^{3/2}(n-i)^{3/2}}{18n^4} \cdot \frac{-1}{4(i(n-i)/n^2)^{3/2}} = \frac{\varepsilon\delta^2}{72} n^{-1}. \end{aligned}$$

To establish (20) and finish the proof of Lemma 5, choose $\alpha = \varepsilon\delta^2/72$.

6.2. Proof of Lemma 4, the variation condition lemma

The proof of Lemma 4 is similar to the proof of Lemma 5, but uses $H(p)$ in place of $f(p)$. We indicate only the changes in the proof. This includes defining $\mathcal{J}$ to be the set of values j such that $|j - i| > \delta\sigma_{i,n}$, and letting $h_2 = \delta\sigma_{i,n}/n$. At the end of the proof, the calculations

change. Using Corollary 2, we have

$$H\left(\frac{i}{n}\right) - \pi_2(\boldsymbol{P}) \geq -\varepsilon \frac{(h_2)^2}{2(\mathrm{e}-1)^2} H''(a) = -\varepsilon \frac{(\delta\sigma_{i,n}/n)^2}{2(\mathrm{e}-1)^2} H''(a)$$
$$= -\varepsilon \frac{\delta^2 i(n-i)}{2(\mathrm{e}-1)^2 n^3} \cdot \frac{-1}{(1-i/n)i/n} = \frac{\varepsilon\delta^2}{2(\mathrm{e}-1)^2} n^{-1}.$$

To finish the proof of Lemma 4, set $\alpha = \mathrm{e}\delta^2/2(\mathrm{e}-1)^2$.

Appendix A. Proofs of variation conditions

We prove that the variation condition holds for both the binomial and the hypergeometric distributions. Fix a Markov chain with transition matrix $\boldsymbol{M}$ on states $0, \ldots, n$, and fix i with $1 \leq i \leq n-1$. Define

$$a_k = m_{i,i+k}$$

for all k such that $0 \leq i + k \leq n$. We say that the *unimodal property* holds provided that $a_k \geq a_{k+1}$ for all $k \geq 0$ and that $a_k \geq a_{k-1}$ for all $k \leq 0$.

Lemma 6. *Suppose that $\boldsymbol{M}$ is a transition matrix satisfying the unimodal property. For each i, let $k_0 = \lceil \sigma_{i,n} \rceil$ (we suppress in the notation the dependence of k_0 on i). Suppose that there is a constant $\alpha > 0$ such that, for all i,*

$$a_{k_0} > \alpha a_0 \quad \textit{and} \quad a_{-k_0} > \alpha a_0.$$

Then the variation condition holds with any $\delta < \frac{1}{2}$ and $\varepsilon = \alpha/(1+\alpha)$.

Proof. Fix i with $1 \leq i \leq n-1$. We need to show that

$$\frac{\sum_{\{k:|k|>\delta\sigma_{i,n}\}} a_k}{\sum_k a_k} > \frac{\alpha}{1+\alpha}. \tag{21}$$

First consider values of a_k for nonnegative values of k. By the unimodal property,

$$\sum_{0\leq k\leq \delta\sigma_{i,n}} a_k \leq \sum_{0\leq k\leq \delta\sigma_{i,n}} a_0 \leq \frac{k_0}{2} a_0.$$

Similarly,

$$\sum_{k>\delta\sigma_{i,n}} a_k \geq \sum_{\delta\sigma_{i,n}<k\leq k_0} a_{k_0} \geq \frac{k_0}{2} a_{k_0} > \alpha \frac{k_0}{2} a_0.$$

Therefore,

$$\frac{\sum_{k>\delta\sigma_{i,n}} a_k}{\sum_{k\geq 0} a_k} > \frac{\alpha(k_0/2)a_0}{(k_0/2)a_0 + \alpha(k_0/2)a_0} = \frac{\alpha}{1+\alpha}.$$

A similar argument shows that

$$\frac{\sum_{(-k)>\delta\sigma_{i,n}} a_k}{\sum_{(-k)\geq 0} a_k} > \frac{\alpha}{1+\alpha}.$$

The previous two equations imply the desired condition (21).

A.1. Proof of Theorem 3

Let $q_{i,j}$ be the hypergeometric probabilities given in (2). Fix n and some $i \in \{1, \ldots, n-1\}$. Let $\sigma = \sqrt{i(n-i)/n}$ and let $k_0 = \lceil \sigma \rceil$. Let $a_{i,k} = q_{i,i+k}$. By Lemma 6, it will suffice to show that $a_{k_0}/a_0 > \alpha$ and $a_{-k_0}/a_0 > \alpha$ for some constant α. By the symmetry of the hypergeometric probabilities, $a_k = a_{-k}$, so we may assume without loss of generality that $i \le n/2$, and prove only $a_{k_0}/a_0 > \alpha$. An easy calculation shows that

$$\frac{a_k}{a_{k-1}} = \frac{(i-k+1)(n-i-k+1)}{(i+k)(n-i+k)}. \tag{22}$$

With $k = 1, 2$, this is

$$\frac{a_1}{a_0} = \frac{i(n-i)}{(i+1)(n-i+1)} \quad \text{and} \quad \frac{a_2}{a_1} = \frac{(i-1)(n-i-1)}{(i+2)(n-i+2)},$$

respectively. For $k \le k_0$,

$$\begin{aligned}
\frac{a_k}{a_0} &= \frac{a_1}{a_0}\frac{a_2}{a_1}\cdots\frac{a_k}{a_{k-1}} \\
&= \frac{i(i-1)(i-2)\cdots(i-k+1)\cdot(n-i)(n-i-1)\cdots(n-i-k+1)}{(i+1)(i+2)(i+3)\cdots(i+k)\cdot(n-i+1)(n-i+2)\cdots(n-i+k)} \\
&= \frac{i}{i+k}\cdot\frac{i-1}{i+k-1}\cdots\frac{i-k+1}{i+1}\cdot\frac{n-i}{n-i+k}\cdot\frac{n-i-1}{n-i+k-1}\cdots\frac{n-i-k+1}{n-i+1} \\
&= \left(\prod_{j=0}^{k-1}\frac{i-j}{i-j+k}\right)\left(\prod_{j=0}^{k-1}\frac{n-i-j}{n-i-j+k}\right) \\
&> \left(\frac{i+1-k}{i+1}\right)^k\left(\frac{n-i-k+1}{n-i+1}\right)^k = \left(1-\frac{k}{i+1}\right)^k\left(1-\frac{k}{n-i+1}\right)^k \\
&> \exp\left(-\frac{\beta k^2}{i+1}\right)\exp\left(-\frac{\beta k^2}{n-i+1}\right) \quad (\text{where } \beta = 2\ln 2) \\
&> \exp\left(-\frac{\beta k^2}{i}\right)\exp\left(-\frac{\beta k^2}{n-i}\right) = \exp\left(-\frac{\beta n k^2}{i(n-i)}\right).
\end{aligned}$$

The inequality introducing the factor β deserves justification. Note that $k/(i+1) \le k_0/(i+1) \le \frac{1}{2}$ for all i and n. Likewise, $k/(n-i+1) \le \frac{1}{2}$. The inequality follows from the fact that $(1-c)^c < e^{-\beta c}$ when $0 < c \le \frac{1}{2}$.

Consider the case of $k = k_0$. Note that $k_0 \le i$ since $k_0 = \lceil\sqrt{i(n-i)/n}\rceil \le \lceil\sqrt{i}\rceil$. Also, $nk_0^2/(i(n-i)) \le 1$ since $k_0 \ge \sigma = \sqrt{i(n-i)/n}$. Thus,

$$\frac{a_{k_0}}{a_0} > e^{-\beta} = \frac{1}{4}.$$

This completes the proof of the theorem.

A.2. Proof of Theorem 2

Consider the binomial probabilities $p_{i,j}$ as defined by (1). Fixing i and letting $a_k = p_{i,i+k}$, we have

$$\frac{a_k}{a_{k-1}} = \frac{(n-i-k+1)i}{(i+k)(n-i)} \quad \text{and} \quad \frac{a_{-k}}{a_{-(k-1)}} = \frac{(i-k+1)(n-i)}{(n-i+k)i}.$$

Both the quantities are clearly less than the corresponding ratio (22) obtained for the hypergeometric probabilities. Hence, by the previous proof,

$$\frac{a_{k_0}}{a_0} > \frac{1}{4} \quad \text{and} \quad \frac{a_{-k_0}}{a_0} > \frac{1}{4},$$

and, by Lemma 6, the proof is completed.

Acknowledgements

We thank W. J. Ewens, I. Abramson, and an anonymous referee for helpful comments.

References

[1] AVISE, J., NEIGEL, J. AND ARNOLD, J. (1984). Demographic influences on mitochondrial DNA lineage survivorship in animal populations. *J. Molecular Evolution* **20,** 99–105.
[2] CANN, R., STONEKING, M. AND WILSON, A. (1987). Mitochondrial DNA and human evolution. *Nature* **325,** 31–36.
[3] CANNINGS, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* **6,** 260–290.
[4] DONNELLY, P. (1991). Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics. *Ann. Prob.* **19,** 1102–1117.
[5] EWENS, W. J. (1963). The mean time for absorption in a process of genetic type. *J. Austral. Math. Soc.* **3,** 375–383.
[6] EWENS, W. J. (1964). The pseudo-transient distribution and its uses in genetics. *J. Appl. Prob.* **1,** 141–156.
[7] EWENS, W. J. (1979). *Mathematical Population Genetics*. Springer, Berlin.
[8] FELLER, W. (1951). Diffusion processes in genetics. In *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, ed. J. Neyman, University of California Press, Berkeley, CA, pp. 227–246.
[9] FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edn. John Wiley, New York.
[10] FISHER, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
[11] KARLIN, S. AND MCGREGOR, J. (1966). The number of mutant forms maintained in a population. In *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, Vol. 4, University of California Press, Berkeley, CA, pp. 403–414.
[12] KIMURA, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Acad. Sci. USA* **41,** 144–150.
[13] KIMURA, M. (1962). On the problem of fixation of mutant genes in a population. *Genetics* **47,** 713–719.
[14] KIMURA, M. (1964). Diffusion models in population genetics. *J. Appl. Prob.* **1,** 177–232.
[15] KINGMAN, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.
[16] KINGMAN, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, eds G. Koch and F. Spizzichino, North-Holland, Amsterdam, pp. 97–112.
[17] KINGMAN, J. F. C. (1982). On the genealogy of large populations. In *Essays in Statistical Science* (J. Appl. Prob. Spec. Vol. **19A**), eds J. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, pp. 27–43.
[18] MÖHLE, M. (1998). Robustness results for the coalescent. *J. Appl. Prob.* **35,** 438–447.
[19] MÖHLE, M. (1999). The concept of duality and applications to Markov processes arising in neutral population genetics models. *Bernoulli* **5,** 761–777.
[20] MÖHLE, M. (2004). The time back to the most recent common ancestor in exchangeable population models. *Adv. Appl. Prob.* **36,** 78–97.
[21] SCHENSTED, I. (1958). Appendix: Model of subnuclear segregation in the macronucleus of ciliates. *Amer. Naturalist* **92,** 161–170.
[22] TAKAHATA, N. (ed.) (1994). *Population Genetics, Molecular Evolution, and The Neutral Theory: Selected Papers*. University of Chicago Press.
[23] TAVARÉ, S. (1995). Calibrating the clock: using stochastic processes to measure the rate of evolution. In *Calculating the Secrets of Life. Applications of the Mathematical Sciences in Molecular Biology*, eds E. S. Lander and M. S. Waterman, National Academy Press, Washington, DC, pp. 114–152.
[24] TAVARÉ, S. (1997). Ancestral inference from DNA sequence data. In *Case Studies in Mathematical Modeling: Ecology, Physiology, and Cell Biology*, eds H. G. Othmer *et al.*, Prentice-Hall, Upper Saddle River, NJ, pp. 91–96.
[25] WATTERSON, G. (1962). Some theoretical aspects of diffusion theory in population genetics. *Ann. Math. Statist.* **33,** 939–957. (Correction: **34** (1963), 352.)

[26] WATTERSON, G. (1996). Motoo Kimura's use of diffusion theory in population genetics. *Theoret. Pop. Biol.* **49,** 154–158.

[27] WRIGHT, S. (1945). The differential equation of the distribution of gene frequencies. *Proc. Nat. Acad. Sci. USA* **31,** 382–389.

[28] WRIGHT, S. (1949). Adaptation and selection. In *Genetics, Paleontology and Evolution*, eds G. Jepson, G. Simpson, and E. Mayr, Princeton University Press, pp. 365–389.