Reprinted from THEORETICAL POPULATION BIOLOGY
All Rights Reserved by Academic Press, New York and London

Vol. 39, No. 1, February 1991
Printed in Belgium

A Multi-dimensional Coalescent Process Applied to Multi-allelic Selection Models and Migration Models

JODY HEY*

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138

Received February 13, 1990

For a sample of two genes from a population divided into an arbitrary number of allele classes, a general mathematical framework is developed to address the expectation and variance of the time of the most recent common ancestor. Depending on the meaning of allele classes and the manner in which genes can change among them, this framework can be applied to a diversity of population genetic models. By adoption of the infinite sites model, the effect on heterozygosity is modelled for balancing selection among allele classes, mutation between allele classes, migration among populations, and gene conversion between loci. Most results are described for a continuous time approximation to a discrete generation model. It is also shown how the discrete generation model can be used to study the hitch-hiking effect of favorable mutations.

1. INTRODUCTION

The genealogical history of a random sample of genes at a locus can, in the absence of recombination, be described with a binary tree. Thus one approach towards modelling the distribution of genetic variation in a sample of genes is to develop a model of the distribution of tree lengths that could give rise to the sample. For example, under an infinite sites model (Kimura, 1969), where all mutations are unique and neutral, the expected number of segregating sites in a sample of genes is proportional to the expected length of the tree of that sample. In the special case when only two genes have been sampled, the proportion of segregating sites is equivalent to the heterozygosity per site.

The distribution of tree lengths can often be obtained via a coalescent, a family of stochastic processes so named because they describe the times at which sampled genes are joined by common ancestry (Kingman, 1982a, b).

* Current address: Dept. of Biological Sciences, Rutgers University, Nelson Labs, P.O. Box 1059, Piscataway, NJ 08855.

0040-5809/91 $3.00

Kaplan, Darden, and Hudson (1988) introduced a two dimensional coalescent process to study models where the population can be divided into two classes of alleles. In the traditional or one dimensional coalescent, all pairs of genes are equally likely to have had a common ancestor. In the two dimensional coalescent, genes can only coalesce with other genes of the same allele class. Two genes of different classes may coalesce if one makes a transition to the allele class of the other. In the models of Kaplan, Darden, and Hudson (1988) these transitions are actually mutations. In a companion paper, Hudson and Kaplan (1988) showed how the same model can be used to study variation at a locus linked to a locus undergoing balancing selection. In this view, each gene in the sample is linked to one of the two alleles at the selected locus, and transitions between the two linkage states occurs via recombination.

This report contains extensions of the two dimensional models and shows how the approach is readily extended to populations with more than two classes of alleles. All of these results are for samples of two genes and thus are useful for depicting expected heterozygosity.

2. Theory

2.1. *Discrete Generations*

The basic model is haploid, for tractability, but can be applied without alteration to several diploid models, as will be shown. Consider a single locus A in a population of N haploid individuals. Let there be m allele classes, $A_1, ..., A_m$, with fixed frequencies $p_1, ..., p_m$. The number of A_i alleles is then $N_i = Np_i$. Each generation the population gives rise to an infinite number of gametes at which time gametes undergo "switching" among allele classes. The next generation is formed by the random sampling of N individuals from the gamete pool, allele frequencies being kept constant (i.e., N_i gametes are sampled from class A_i in the gamete pool). The meaning of "switching" depends on the model, but it can be operationally described with the quantity f_{ij}, the probability that a randomly sampled A_j allele was descended from an A_i allele of the previous generation. In general f_{ij} will be of order $1/N$ for $i \neq j$ and

$$f_{ii} = 1 - \sum_{i \neq j}^{m} f_{ij}.$$

Thus

$$\sum_{i=1}^{m} f_{ij} = 1.$$

A general model allows population size, allele frequency, and switching parameters to vary with time, but with one exception, all results presented here require that these parameters remain constant. In particular, note that the assumption of constant allele frequencies implies some type of stabilizing force (e.g., balancing selection among allele classes).

Consider a single locus for which a sample of two gene copies is drawn from the population. We would like to know the distribution of T, the number of generations between the time the sample is taken and the time of the most recent common ancestor of the two sampled genes. What follows is a description of a discrete Markov chain in which $R(t)$ denotes the allele classes of the ancestors of the sampled genes t generations prior to the time the sample is taken. At the time the sample is taken, $t = 0$, $R(0) = (ij)$ indicates that one of the sample genes was in allele class i and the other in allele class j. With this notation $R(t) = (ij)$ is equivalent to $R(t) = (ji)$. At $t = 0$ there are m homoallelic states (i.e., both genes in the sample are in the same allele class) and $m(m-1)/2$ distinct heteroallelic states (i.e., the genes are in different allele classes). For $t > 0$, $R(t)$ describes the state of the ancestors of the sample t generations ago. The number of states now includes all homoallelic and heteroallelic states as well as m states of the form $R(t) = (iO)$, which indicates a common ancestor of the sampled genes in allele class i in generation t.

We define the probabilities for all states of the system in generation $t+1$ conditioned on the states in generation t.

For $0 < i, j, k \leqslant m$,

$$P\{R(t+1) = (iO) \mid R(t) = (jk)\} = \frac{f_{ij} f_{ik}}{N_i},$$

$$P\{R(t+1) = (ii) \mid R(t) = (jk)\} = f_{ij} f_{ik} \left(1 - \frac{1}{N_i}\right),$$

$$P\{R(t+1) = (iO) \mid R(t) = (jO)\} = f_{ij},$$

and

$$P\{R(t+1) = (ij) \mid R(t) = (kO)\} = 0.$$

For $0 < i, j, k, l \leqslant m$, where $i \neq j$,

$$P\{R(t+1) = (ij) \mid R(t) = (kl)\} = f_{ik} f_{jl} + f_{il} f_{jk}.$$

The transition from generation t to generation $t+1$ can be expressed as a matrix equation,

$$x(t+1) = Ax(t). \tag{1}$$

The number of dimensions of matrix A is the total number of possible states, $m(m+3)/2$. For example, when $m=2$,

$$x(t)=\begin{pmatrix} P\{R(t)=(10)\} \\ P\{R(t)=(11)\} \\ P\{R(t)=(12)\} \\ P\{R(t)=(22)\} \\ P\{R(t)=(20)\} \end{pmatrix}$$

and

$$A=\begin{pmatrix} f_{11} & f_{11}^2/N_1 & f_{12}f_{11}/N_1 & f_{12}^2/N_1 & f_{12} \\ 0 & f_{11}^2(1-1/N_1) & f_{12}f_{11}(1-1/N_1) & f_{12}^2(1-1/N_1) & 0 \\ 0 & 2f_{11}f_{21} & f_{11}f_{22}+f_{12}f_{21} & 2f_{22}f_{12} & 0 \\ 0 & f_{21}^2(1-1/N_2) & f_{22}f_{12}(1-1/N_2) & f_{22}^2(1-1/N_2) & 0 \\ f_{21} & f_{21}^2/N_2 & f_{21}f_{22}/N_2 & f_{22}^2/N_2 & f_{22} \end{pmatrix}. \quad (2)$$

It is evident that the common ancestor states are absorbing states and that

$$\lim_{t\to\infty} \sum_{i=1}^{m} P\{R(t)=(iO)\}=1.$$

With an initial vector $x(0)$ describing the state of the sample, the state of the system in generation t can be described by

$$x(t)=A^t x(0).$$

The probability that the system moved into the closed set of common ancestor states in a particular generation is obtained from the difference between successive generations of the total probability that the system is not in a common ancestor state. One method is to use a matrix derived from A by deleting the rows and columns associated with transitions to the common ancestor states. Let $\hat{A}$ be a matrix of dimension $m(m+1)/2$ containing transition probabilities between non-common ancestor states. For Eq. (2), $\hat{A}$ is the central 3 by 3 matrix contained in A. Let $\hat{x}(t)$ contain the probabilities associated with non-absorbing states listed in $x(t)$, and let $P_c(t)$ denote the probability that the system moved into the closed set of common ancestor states in generation t. This is obtained by summing the elements of a vector. For example when $m=2$, the sum, $P_c(t)$, is equal to

$$[1 \quad 1 \quad 1](\hat{A}^{(t-1)}-\hat{A}^t)\,\hat{x}(0). \quad (3)$$

This expression can be used for finding the expectation and variance of the time of most recent common ancestor, particularly if the eigensystem of $\hat{A}$

is found. Let $\hat{A}=\Theta\Lambda\Theta^{-1}$, where Θ is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. Then

$$P_c(t)=[1 \quad 1 \quad 1](\Theta[\Lambda^{(t-1)}-\Lambda^t]\,\Theta^{-1})\,\hat{x}(0). \quad (4)$$

Approximate calculations of the expectation and variance are made by iterative calculation until the cumulative probability is greater than some point very near one.

2.2. *Continuous Generations*

An efficient alternative to the calculation of discrete time Markov chains is to approximate the process with a continuous time pure jump process. This approach requires an infinitesimal parameter $q_{(ij)(kl)}$, the instantaneous rate at which the system moves from sample state (ij) to sample state (kl). These parameters depend, in turn, on the rates of transition among allele classes. Using the same notation as for the discrete model, let f_{ij} be the rate at which an allele of class j is descended from class i. Note the difference between the discrete generation case in which the probabilities of switching sum to one and the continuous case in which the transition rates sum to zero. This is because, in the continuous case,

$$\sum_{i\neq j}^{m} f_{ij}=-f_{ii}.$$

The values for $q_{(ij)(kl)}$ depend on the identities of i, j, k, and l. By assuming that only one of the alleles of the sample can change class in the instant described by q, let $q=0$ when neither i nor j is equal to either k or l.

For transitions from homoallelic to heteroallelic states,

$$q_{(ii)(ij)}=2f_{ji}, \qquad i\neq j.$$

For transitions from heteroallelic to homoallelic states,

$$q_{(ij)(ii)}=f_{ij}, \qquad i\neq j.$$

For transitions between heteroallelic states,

$$q_{(ij)(jk)}=f_{ik}, \qquad i\neq k.$$

To simplify the model, the common ancestor states are lumped into a single state C. It is assumed that the system cannot move from a heteroallelic state into state C in one step. Therefore,

$$q_{(ij)C}=0, \qquad i\neq j.$$

For homoallelic states

$$q_{(ii)C} = 1/N_i.$$

The total rate of leaving a state is

$$q_{ii} = \sum_{j \neq i}^{m} 2f_{ji} + \frac{1}{N_i},$$

for a homoallelic state, and

$$q_{ij} = \sum_{k=1}^{m} f_{kj} + \sum_{k=1}^{m} f_{ki},$$

for a heteroallelic state.

The probability of transition from state (ij) to state (kl) is

$$Q_{(ij)(kl)} = \frac{q_{(ij)(kl)}}{q_{ij}}.$$

In summary, the amount of time the system state (meaning the state of the original sample and its ancestors) remains unchanged is described by an exponential distribution with parameter determined by the transition rates away from that sample state. If the system is in a homoallelic state then the sample may move to a heteroallelic state or to state C, in which case no further transitions occur. If the system is in a heteroallelic state then it may move to either one of two homoallelic states or to another heteroallelic state. In other words, the process is only finished when the system moves into C, and it can only move to C from a homoallelic state.

Simple expressions for the expectation and the variance of T, the time to the most recent common ancestor (i.e., the time required for the system to move into C) can be found by exploiting the Markov property. Since the progress of the system at any time depends only on its state at the time and not on its history, the process can be viewed as a sum of independent exponential distributions.

The expected time to the most recent common ancestor is given by

$$E_{ii}(T) = \frac{1}{q_{ii}} + \sum_{j \neq i}^{m} Q_{(ii)(ij)} E_{ij}(T) \tag{5a}$$

and

$$E_{ij}(T) = \frac{1}{q_{ij}} + \sum_{k \neq j}^{m} Q_{(ij)(ik)} E_{ik}(T) + \sum_{k \neq i}^{m} Q_{(ij)(kj)} E_{kj}(T), \tag{5b}$$

for homoallelic states and heteroallelic states, respectively. The subscript of E denotes the value of $R(0)$, the state of the sample at $t = 0$.

For the variance,

$$V_{ii}(T) = \frac{1}{q_{ii}^2} + \sum_{j \neq i}^{m} Q_{(ii)(ij)} V_{ij}(T) + \sum_{j \neq i}^{m} Q_{(ii)(ij)} E_{ij}^2(T) - \left(\sum_{j \neq i}^{m} Q_{(ii)(ij)} E_{ij}(T) \right)^2 \tag{6a}$$

and

$$V_{ij}(T) = \frac{1}{q_{ij}^2} + \sum_{k \neq j}^{m} Q_{(ij)(ik)} V_{ik}(T) + \sum_{k \neq i}^{m} Q_{(ij)(kj)} V_{kj}(T) + \sum_{k \neq j}^{m} Q_{(ij)(ik)} E_{ik}^2(T) + \sum_{k \neq i}^{m} Q_{(ij)(kj)} E_{kj}^2(T) - \left(\sum_{k \neq j}^{m} Q_{(ij)(ik)} E_{ik}(T) + \sum_{k \neq i}^{m} Q_{(ij)(kj)} E_{kj}(T) \right)^2. \tag{6b}$$

The expectation for a sample in which the two genes are drawn randomly from the entire population is

$$E(T) = \sum_{i=1}^{m} p_i^2 E_{ii}(T) + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} 2p_i p_j E_{ij}(T).$$

The variance for a random population sample includes the variance both from within and from among sample states,

$$V(T) = \sum_{i=1}^{m} [p_i^2 (V_{ii}(T) + (E_{ii}(T) - E(T))^2)] + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} [2p_i p_j (V_{ij} + (E_{ij}(T) - E(T))^2)].$$

In the special case when all transition rates between allele classes are equal ($f_{ij} = f$, for $i \neq j$) and all allele frequencies are equal ($p_i = 1/m$), solutions are further simplified. Under these conditions, all homoallelic states are equivalent and all heteroallelic states are equivalent, and (5) reduces to just two equations:

$$E_{ii}(T) = \frac{1}{2f(m-1) + m/N} + \frac{2f(m-1)}{2f(m-1) + m/N} E_{ij}(T)$$

and

$$E_{ij}(T)=\frac{1}{2f}+E_{ii}(T).$$

These simplify to

$$E_{ii}(T)=N$$

and

$$E_{ij}(T)=\frac{1}{2f}+N.$$

The variance under these conditions simplifies in a similar fashion:

$$V_{ii}(T)=N^2+\frac{N(m-1)}{fm}$$

and

$$V_{ij}(T)=\frac{1}{4f^2}+N^2+\frac{N(m-1)}{fm}.$$

The result in which the expectation for a homoallelic state depends only on the total population size, and not on the switching rates or the number of allele classes, was also found by Slatkin (1987) and Strobeck (1987), for migration models, and Kaplan and Hudson (1987), for gene conversion models.

For a random sample from the population

$$E(T)=N+\frac{m-1}{2fm}$$

and

$$V(T)=\frac{1}{4f^2}+N^2+\frac{N(m-1)}{fm}-\frac{1}{4f^2m^2}.$$

Alternatively these expressions can be rearranged to yield the expectation

$$E\left(\frac{T}{N}\right)=1+\frac{m-1}{2Nfm}, \tag{7}$$

in units of N generations, and the variance,

$$V\left(\frac{T}{N}\right)=1+\frac{m-1}{Nfm}+\frac{m^2-1}{4N^2f^2m^2}, \tag{8}$$

in units of N^2 generations. This is the conventional format for continuous time coalescent results. It is simpler by virtue of replacing three variables (T, N, and f) with two (T/N and Nf). In this context Nf will generally be of order 1.

The probability density function of the time of common ancestry can be obtained from a Kolmogorov backward equation. Solution of this differential equation would be impractical for large m were it not that the matrix of transition rates becomes increasingly sparse as m increases. The total number of cells in the matrix is $(m(m+1)/2)^2$ and the proportion that contain zeros is

$$\frac{(m-1)(m^2-m+2)}{m(m+1)^2}.$$

This quantity is equal to 0.333 for $m=3$, 0.489 for $m=5$, and 0.684 for $m=10$.

For $m=2$, all of the results for continuous time were obtained by Kaplan, Darden, and Hudson (1988). These authors also provide expressions for the expectation and variance of tree length for sample sizes greater than two in a two dimensional model.

3. Models

Three general classes of models, which differ in the manner in which f_{ij} is defined, can be addressed with the mathematical framework that has been developed.

3.1. *Class 1 models*

In this case, the switching parameters are initially described in terms of the destination of genes leaving an allele class as might occur in a model of mutation or migration. Let u_{ij} be the proportion of genes in class i in generation $t+1$ that leave descendants in class j in generation t. Then the probability that a randomly sampled A_j allele was descended from an A_i allele of the previous generation is

$$f_{ij}=\frac{p_iu_{ij}}{p_j}.$$

In this view the size of each allele class is determined by the proportion of genes that switch classes. The equilibrium frequencies are found from the matrix of u_{ij}. For $m=2$, the equilibrium frequencies are

$$p_i=\frac{u_{ji}}{(u_{ij}+u_{ji})}$$

and

$$p_j=\frac{u_{ij}}{(u_{ij}+u_{ji})}.$$

Class 1 models include the case where u_{ij} represents the probability that an A_i allele mutates to an A_j allele. Thus, this model allows one to investigate the common ancestry time of two genes, each drawn from any of m allele classes, where mutation occurs among alleles. In this context, the assumption of constant allele frequencies requires that there also be some form of balancing selection among alleles.

A very different biological model that also fits in this framework is one in which the population is divided into m subpopulations and u_{ij} represents the probability that a gene from subpopulation i migrates to subpopulation j. In contrast to the mutation model, genes are not labeled by allele class but rather by the subpopulation in which they occur. It is necessary, in this case, to assume some type of density dependent force so that subpopulation sizes are constant.

3.2. *Class 2 models*

For many models it may be useful to define f_{ij} directly. In this situation, there is no need to define u_{ij} as for class 1 models and there is no necessary relation between f_{ij} and N_i.

This framework applies to a migration model in which a constant proportion, f_{ij}, of the genes in subpopulation j are replaced by genes from subpopulation i each generation. As with the class 1 migration model, some sort of density dependent force, acting to maintain constant subpopulation sizes, is assumed.

Class 2 models also include a very different case, that of gene conversion among loci. Let m represent the number of unlinked loci among which gene conversion can occur, and let f_{ij} represent the proportion of genes at locus j converted by genes at locus i each generation. Clearly the total number of copies is the same for each locus. In this view N, the total number of gene copies, is not the population size but rather it is the population size times m. This model also requires that the population size be constant.

3.3. *Class 3 models*

Up to this point the models have been haploid. No modifications are necessary to apply the framework to diploid models so long as the f_{ij} pertain to a haploid stage of the life cycle. Thus the migration models described are most appropriate when considering gamete migration (e.g., pollen dispersal) and when considering genes transmitted by only one sex. For diploid migration models, one needs to consider that there will be an added variance to the waiting time between transitions due to the constraint of genes switching in pairs. However, the migration models described here are a very good approximation of the diploid case.

Class 3 models are explicitly diploid and apply to the case of recombination between allele classes. Hudson and Kaplan (1988) showed that the two dimensional coalescent could be used to model tree lengths for genes sampled from a neutral locus that is linked to another locus at which two alleles, A_1 and A_2, occur in a balanced polymorphism. At the neutral locus each gene copy must be coupled with one of the A alleles. In the multi-dimensional case, m alleles segregating at locus A are maintained at fixed frequencies by balancing selection. In this view each class represents a state of linkage to one of the alleles at locus A. Switching, via recombination, is a function of the recombination rate and the chance that an allele occurs in a heterozygote.

Let there be m alleles maintained in a stable balanced polymorphism. Then u_{ij}, the probability that a descendent of a gene from class i in generation $t+1$ was in class j in generation t, depends on three things: the probability that an A_i allele forms a zygote with an A_j allele; the relative fitness of an A_iA_j heterozygote; and the probability of a recombination event between the two loci. Then

$$u_{ij}=\frac{w_{ij}}{w_i}p_j r, \qquad i\neq j,$$

where w_{ij} is the fitness of an A_iA_j heterozygote; w_i is the mean fitness of individuals with i alleles, which is necessarily equal to the mean fitness of the population, $\bar{w}$, at equilibrium; and r is the recombination rate per generation between locus A and the neutral locus under consideration.

The probability, f_{ij}, that a randomly sampled gene coupled to an A_j allele in the current generation was coupled to an A_i in the previous generation is

$$\frac{u_{ij}p_i}{p_j}=\frac{w_{ij}}{\bar{w}}p_i r.$$

In their two allele model, Hudson and Kaplan (1988) considered weak

selection. By setting w_{12} equal to $\bar{w}$ (i.e., $f_{12}=p_1 r$), the model becomes identical to Hudson and Kaplan's extension to the work of Kaplan, Darden, and Hudson (1988), for a sample size of two.

It is apparent that strong selection will not have a great effect on outcomes. For example, consider a two allele model in which homozygotes are lethal. Then $w_{12}/\bar{w}=2$. In effect the recombination rate is doubled relative to the case in which selection is very weak. As observed by Strobeck (1983), the effect of selection against homozygotes can be viewed as simply increasing the recombination rate. If both homozygotes had only half the fitness of a heterozygote then $w_{12}/\bar{w}=1.33$. It is expected that as more allele classes are added $w_{ij}/\bar{w}$ will become closer to 1 because an increasing proportion of allele pairs are heterozygotes. In general a model that ignores fitness effects should be adequate unless selection is very strong.

It is important to remember that when a model with strong selection is desired, and the approximation of allowing all genotypes to be equally fit is not used, then the equilibrium allele frequencies are determined by the selection coefficients.

4. Applications

4.1. *Discrete Generations*

In Table I are shown some comparisons between a three class discrete time model and continuous time approximations for several values of N and r. The differences are very slight.

In contrast to the continuous time approximation, however, the discrete model does not require that N, p_i, and f_{ij} be constants. In particular, (1) is appropriate even if the parameters are functions of time. As an example, consider a model with 2 gene classes, A_1 and A_2, in which the A_2 gene class began as a mutation in generation $t=\tau$ with frequency $p_2=1/N$ and has just reached $p_2=1-1/N$ at $t=0$. By using a deterministic model of the fixation process, which is appropriate if selection is strong, and a model with switching via recombination (class 3 model) the hitch-hiking effect of a favorable gene on linked neutral loci can be examined. The reason strong selection ($Ns \gg 1$) is required is that the model assumes the parameters $N(t)$, $p_i(t)$, and $f_{ij}(t)$ have no variance for a given t.

If A_2 increased in frequency under genic selection with a coefficient of s then the dependency of frequency on time can be expressed by

$$p_2(t+1)=\frac{p_2(t)[2-s(1+p_2(t))]}{2(1-sp_2(t))}.$$

TABLE I

Results from Discrete (D) and Continuous (C) Models

Model	p_1	p_2	N	r	Nr	$E\left(\frac{T}{N}\right)$	$V\left(\frac{T}{N}\right)$
D	0.33	0.33	10	0.01	0.1	10.84	214.2
D	0.33	0.33	1000	0.0001	0.1	10.99	219.4
C	0.33	0.33	—	—	0.1	11.00	221.0
D	0.2	0.3	10	0.001	0.01	89.13	17502
D	0.2	0.3	1000	0.00001	0.01	89.26	17543
C	0.2	0.3	—	—	0.01	89.34	17647
D	0.2	0.3	10	0.01	0.1	9.73	187.9
D	0.2	0.3	1000	0.0001	0.1	9.86	192.4
C	0.2	0.3	—	—	0.1	9.87	193.6
D	0.2	0.3	10	0.1	1	1.78	3.82
D	0.2	0.3	1000	0.001	1	1.91	4.56
C	0.2	0.3	—	—	1	1.91	4.60
D	0.01	0.1	100	0.001	0.1	3.11	43.11
D	0.01	0.1	1000	0.0001	0.1	3.11	43.19
C	0.01	0.1	—	—	0.1	3.11	43.32

Note. All examples are based on a model of recombination among 3 allele classes maintained by weak balancing selection. Thus $p_3=1-p_1-p_2$. In the continuous case, calculations were made using (5) and (6). Calculations for the discrete model were made using a three allele version of Eq. (4). Discrete model calculations were halted when the cumulative probability of coalescence for the least frequent sample state was greater then 0.9999.

The calculation of the expected time to the most recent common ancestor at the neutral locus has two components. The first contribution to the expectation is from times between $t=0$ and $t=\tau$ (i.e., between the time of mutation and the time of fixation). Let $J(t)$ be a 3 by 3 matrix of the form taken by $\hat{A}$ in Eq. (3). Let the switching rate, f_{ij}, be replaced by $f_{ij}(t)=p_i(t)\,r$. Then let

$$G(t)=\prod_{i=1}^{t} J(i).$$

The probability that the system moved into a common ancestor state in generation t is analogous to Eq. (4),

$$P_c(t)=[1 \quad 1 \quad 1](G(t-1)-G(t))\,\hat{x}(0).$$

In this case $\hat{x}(0)$ necessarily contains a 1 in the position corresponding to sample state (22) and zeros elsewhere. This is because at the time the sample is taken, the population is fixed for the A_2 allele.

It follows that the probability of the system moving into a common ancestor state at some time $t \leqslant \tau$ is

$$\sum_{t=1}^{\tau} P_c(t),$$

and the contribution to the expected time of most recent common ancestor is

$$\sum_{t=1}^{\tau} tP_c(t). \tag{9}$$

Since the coalescent process reverts to a one dimensional case for $t > \tau$, the expected time to most recent common ancestor given that the most recent common ancestor did not occur between $t=0$ and $t=\tau$ is $N+\tau$. This quantity (which under strong selection will be very near N) times the probability that the most recent common ancestor did not occur between $t=0$ and $t=\tau$ makes up the second term in the calculation:

$$(N+\tau)\left[1-\sum_{t=1}^{\tau} P_c(t)\right]. \tag{10}$$

The sum of (9) and (10) provides the expected time since the most recent common ancestor for a locus some recombination distance, r, from locus A.

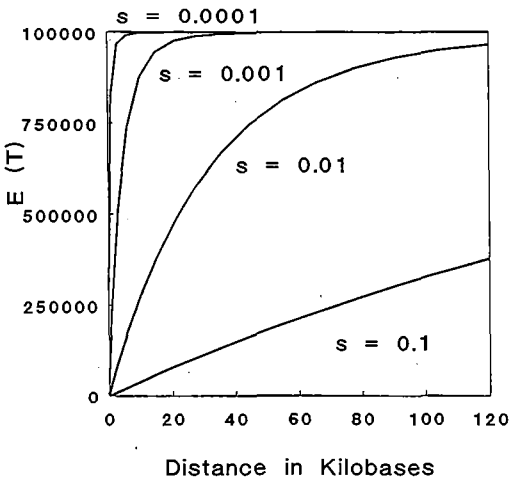

FIG. 1. The hitch-hiking effect of a favorable mutation on the expected common ancestry time, in generations, at a linked neutral locus. The parameters were $N = 10^6$ and $r = 10^{-5}$ per kilobase pair per generation. The results for four different selection coefficients are shown.

Examples of these calculations are provided in Fig. 1. As expected, the hitchhiking effect, expressed as reduced common ancestry time relative to N generations, increases with s and decreases with r. To consider the A locus itself, both r and the second term of the calculation are equal to zero.

The deterministic description of the trajectory of p_2 may not be appropriate for values near 0 or 1, at which times p_2 behaves stochastically. In their similar treatment of the hitchhiking effect, Kaplan, Hudson, and Langley (1989) discuss this issue in depth and describe simulations that support their results.

The results presented here were also checked with simulations. Numerous selection events were simulated in the manner of Kimura and Ohta (1968), and the allele frequencies during the fixation process were stored. Samples of two genes were then repeatedly coalesced on the arrays of allele frequencies that had been generated during the selection simulations. The coalescent simulations followed the method developed by Hudson (1983). The results (not shown) were very similar to those from the calculations.

4.2. *Continuous Generations*

All three classes of models can be addressed for the case of two alleles by applying the results of Kaplan, Darden, and Hudson (1988) and Hudson and Kaplan (1988). Therefore, rather than focus on the effect of varying the product Nf, the effect of increasing the number of allele classes will be examined.

4.2.1. *Class* 1 *Models.* For models in which switching is first described by the destination of genes leaving an allele class (i.e., when f_{ij} and p_i are determined by u_{ij}), the effect of additional allele classes is most easily examined for the case when all u_{ij} are equal (i.e., $u_{ij} = u$, for $i \neq j$). In this case $f_{ij} = u$, for $i \neq j$, and all p_i are equal to $1/m$. Equations (7) and (8) can be used for the expectation and the variance of common ancestry times.

4.2.2. *Class* 2 *Models.* When f_{ij} are defined directly and do not determine allele frequencies, a diversity of island and stepping-stone migration models can be developed. In essence, one can solve for the expected divergence between two genes drawn from any two of an arbitrary number of populations. Each of the $m(m-1)$ migration rates can take on any value between zero and one.

One example of population structure will be developed. Consider a linear stepping-stone model in which each subpopulation exchanges genes only with the two neighboring subpopulations. The terminal subpopulations can exchange genes only with their single neighboring subpopulations. For simplicity assume that every subpopulation is the same size and that all non-zero migration rates are identical. When chains of different lengths are

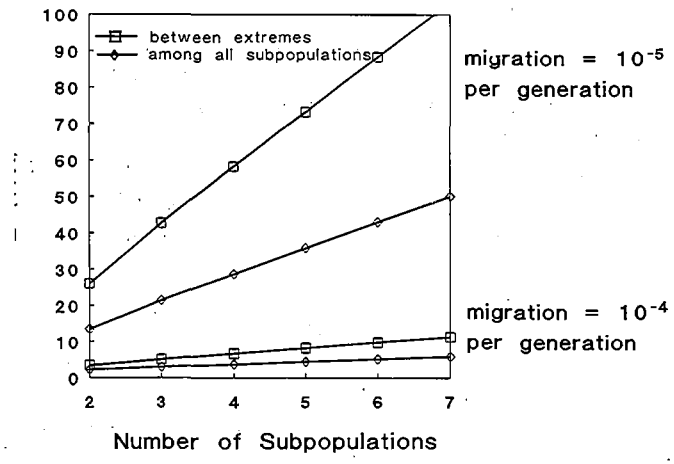

FIG. 2. Results of a linear stepping-stone migration model. This is a class 2 model with all subpopulations having 1000 individuals and the total population size, N, equal to 1000 times the number of subpopulations. Results are shown for two migration rates. In each case, the expected time to the most recent common ancestor, in units of N generations, is shown for the expectation of a random sample from the grand population and for a sample in which the two genes are drawn from the opposite extremes of the chain. The system of equations in (5) were used for the calculations.

contrasted, migration rates and the size of each subpopulation are unchanged. Thus, the total population size is a multiple of chain length. Figure 2 presents results for samples drawn randomly from the entire chain and for samples in which the two genes are drawn from opposite extremes of the chain. For this example, expected common ancestry time increases almost linearly with chain length and is nearly double for samples with genes drawn from opposite extremes of the chain relative to randomly drawn samples.

4.2.3. *Class 3 Models.* As in the case of class 1 models, a model of linkage to a balanced polymorphism is most readily examined when all rates and frequencies are equal. In this case, because $f_{ij} = p_i r = r/m$, Eq. (7) and (8) take the forms

$$E\left(\frac{T}{N}\right) = 1 + \frac{m-1}{2Nr}, \tag{11}$$

and

$$V\left(\frac{T}{N}\right) = 1 + \frac{m-1}{Nr} + \frac{m^2-1}{4N^2r^2}, \tag{12}$$

respectively.

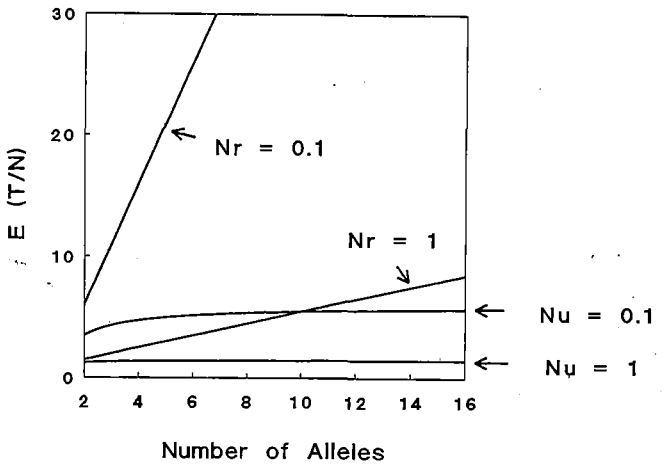

FIG. 3. Expected time to most recent common ancestor in units of N generations is plotted for a class 1 model and for a class 3 model. The results for the class 1 model, indicated by $N\mu$, were calculated with Eq. (7). The results for the class 3 model, indicated by Nr, were calculated with Eq. (11).

Some examples, using Eq. (11), are given in Fig. 3, where they are contrasted with examples from a class 1 model, using Eq. (7). In contrast to the results of class 1 models, the expectation in a class 3 model shows a linear association with m. In class 1 models, heteroallelic samples can switch to a homoallelic state at any time, but in class 3 models switching can only occur in heterozygotes, via recombination, and the frequency of any particular heterozygote drops sharply as m increases.

5. DISCUSSION

Coalescent models have proven useful for a variety of population genetic questions. In addition to being analytically accessible, the coalescent approach is eminently suited to simulation studies (Hudson, 1983). An attraction of coalescent models for empirical workers is that the models and outcomes are phrased explicitly in terms of the properties of samples rather than entire populations.

It is shown in this report how models of balancing selection, migration, mutation, gene conversion, and, with some modifications, genetic hitchhiking can be addressed with a common mathematical framework.

The models developed here are restricted to samples of size two. In this case the expectation of the total length of the genealogical tree of a sample is twice $E(T)$ and the variance of tree length is four times $V(T)$. These

quantities can be used for predicting divergence between the two genes of a sample. Consider an infinite sites model (Kimura, 1969) in which the expected number of mutations per generation is μ. Let S be the number of sites that differ between the two genes of a sample. Then

$$E(S) = 2\mu E(T),$$

and

$$V(S) = 2\mu E(T) + 4\mu^2 V(T).$$

Let S_{tot} be the total number of sites in the gene (e.g., the number of base pairs). If $S_{\text{tot}} \gg E(S)$ then the infinite sites approximation can be used to estimate heterozygosity. Let H represent heterozygosity per site. Then

$$E(H) \approx \frac{E(S)}{S_{\text{tot}}}$$

and

$$V(H) \approx \frac{V(S)}{S_{\text{tot}}^2}.$$

Although $E(H)$ does not depend on sample size, $V(H)$ for a sample size of two is only useful as an upper bound when considering larger samples. It is also important to note that all of the expressions for variance contained in this report necessarily include both sampling variance and stochastic (among population) variance (Tajima, 1983).

If one is primarily interested in the expectation, the assumption of no recombination within a locus can be relaxed. Allowing recombination among genes of the same allele class can have a large effect on $V(S)$, via dissipation of the stochastic component (Hudson, 1983), but $E(S)$ is not affected.

Acknowledgments

This research was supported by National Institutes of Health Award GM12043. I thank R. R. Hudson, N. Kaplan, and S. Otto for comments on the manuscript and R. C. Lewontin for comments as well as for discussions that inspired much of the research.

References

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination, *Theor. Pop. Biol.* **23**, 183–201.

Hudson, R. R., and Kaplan, N. L. 1988. The coalescent process in models with selection and recombination, *Genetics* **120**, 819–829.

Kaplan, N. L., Darden, T., and Hudson, R. R. 1988. The coalescent process in models with selection, *Genetics* **120**, 831–840.

Kaplan, N. L., and Hudson, R. R. 1987. On the divergence of genes in multigene families, *Theor. Pop. Biol.* **31**, 178–194.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. 1989. The "hitch-hiking effect" revisited, *Genetics* **123**, 887–899.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations, *Genetics* **61**, 893–903.

Kimura, M., and Ohta, T. 1968. The average number of generations until fixation of a mutant gene in a finite population, *Genetics* **61**, 763–771.

Kingman, J. F. C. 1982a. On the genealogy of large populations, *J. Appl. Probab. A* **19**, 27–43.

Kingman, J. F. C. 1982b. The coalescent, *Stochastic Process. Appl.* **13**, 235–248.

Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population, *Theor. Pop. Biol.* **32**, 42–49.

Strobeck, C. 1983. Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement, *Genetics* **103**, 545–555.

Strobeck, C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision, *Genetics* **117**, 149–153.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**, 437–460.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium