The "Hitchhiking Effect" Revisited

Norman L. Kaplan,* Richard R. Hudson† and Charles H. Langley‡

**Statistics and Biomathematics Branch and ‡Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, and †Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717*

Manuscript received February 28, 1989
Accepted for publication September 12, 1989

ABSTRACT

The number of selectively neutral polymorphic sites in a random sample of genes can be affected by ancestral selectively favored substitutions at linked loci. The degree to which this happens depends on when in the history of the sample the selected substitutions happen, the strength of selection and the amount of crossing over between the sampled locus and the loci at which the selected substitutions occur. This phenomenon is commonly called hitchhiking. Using the coalescent process for a random sample of genes from a selectively neutral locus that is linked to a locus at which selection is taking place, a stochastic, finite population model is developed that describes the steady state effect of hitchhiking on the distribution of the number of selectively neutral polymorphic sites in a random sample. A prediction of the model is that, in regions of low crossing over, strongly selected substitutions in the history of the sample can substantially reduce the number of polymorphic sites in a random sample of genes from that expected under a neutral model.

THE effect of a strongly selected allele at one locus on the frequencies of neutral alleles at a linked locus, commonly referred to as the "hitchhiking effect," has been studied by a number of authors (*e.g.*, KOJIMA and SCHAEFFER 1967; MAYNARD SMITH and HAIGH 1974; OHTA and KIMURA 1975). MAYNARD SMITH and HAIGH presented a deterministic analysis that supported their contention that the hitchhiking effect of linked favorably selected variants on neutral polymorphisms is an important mechanism for reducing heterozygosity. They also showed that the hitchhiking effect on the "half-life" of a neutral polymorphism depends on the strength of selection and the rate of crossing over. OHTA and KIMURA presented a stochastic analysis of the hitchhiking effect, which they interpreted as refuting the contentions of MAYNARD SMITH and HAIGH. The major distinction between these two investigations was that MAYNARD SMITH and HAIGH considered the consequences of a favorably selected substitution on preexisting linked neutral polymorphism, while OHTA and KIMURA chose to examine the impact on neutral mutations arising while the selected allele is on its way to fixation (OHTA and KIMURA 1976; HAIGH and MAYNARD SMITH 1976). Neither side in this controversy presented a complete steady state analysis with recurring selected substitutions, neutral mutation, random genetic drift and crossing over.

With the advent of molecular population genetics, the question of the evolutionary effects of strongly selected substitutions merits a reexamination. The large number of polymorphisms found in a small stretch of DNA among randomly chosen individuals from large outbreeding natural populations should make it possible to investigate the consequences of hitchhiking with more precision and power. What is needed is an analysis that predicts the hitchhiking effect on experimentally measurable quantities derived from random samples of genes. The results of MAYNARD SMITH and HAIGH (1974) on the hitchhiking effect on the half-life of a neutral polymorphism cannot be readily applied.

A natural summary statistic for DNA sequence data is S, the number of polymorphic nucleotide sites in a sample of genes. The distribution of S depends on the assumed underlying population genetics model. For selectively neutral, infinite sites models both with and without recombination, the distributional properties of S at equilibrium are well characterized (*e.g.*, WATTERSON 1975; KINGMAN 1982; HUDSON 1983; TAVARÉ 1984). We will refer to such selectively neutral loci as "isolated, selectively neutral loci." More recently, the distribution of S has been studied for a class of infinite sites models where selectively neutral sites are linked to a locus at which natural selection leads to polymorphism, *e.g.*, strong balancing selection (KAPLAN, DARDEN and HUDSON 1988; HUDSON and KAPLAN 1988). The effect in this case is also an example of hitchhiking, but heterozygosity is increased rather than decreased. HUDSON and KAPLAN (1988) determined the distribution of S and quantified

The publication costs of this article were partly defrayed by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

the hitchhiking effect on the moments of S.

In this paper we study the distribution of S for a class of models where selectively neutral sites are linked to loci where strongly selected mutations have rapidly swept through the population at different times in the past. All of the previous work on hitchhiking has focused on analyzing the effect on neutral polymorphism of a selectively favored allele while it is in a state of transient polymorphism. The distribution of S, on the other hand, is dependent on the population dynamics since the most recent common ancestor of the sample, and so advantageous substitutions occurring at different times in the past at different loci can affect the distribution of S. In order to quantify this effect on the distribution of S, it is necessary to consider the coalescent process which describes the genealogical history of a random sample of genes at a selectively neutral region of the genome which is linked to a locus that is not selectively neutral. This stochastic process was studied by HUDSON and KAPLAN (1988) and so their results can be directly applied. The analysis is then extended to the case where selected substitutions of newly arising mutants or very rare variants are assumed to occur randomly in the history of the sample at random locations in the genome. The details are given in the next section (THEORY).

For the models considered, the degree to which the distribution of S is affected by ancestral, selected substitutions at linked loci is determined by the parameters $\theta = 4N\mu$, $C = 2Nc$, $\alpha = 2Ns$ and $\Lambda = 2N\lambda$, where N is the diploid population size, μ is the expected number of selectively neutral mutations per nucleotide site, per chromosome, per generation, c is the expected number of crossovers per nucleotide site, per genome, per generation, s is the selective advantage of the favored allele per generation and λ is the expected number of selected substitutions per nucleotide site, per generation (Table 1). The parameter λ may or may not be a function of $2N$, depending on assumptions about the process generating selected substitutions. In the third section (CALCULATIONS) some numerical calculations are presented to illustrate the behavior of the mean of S for a sample of size 2 for different values of the parameters.

Finally in the last section (DISCUSSION) the results are compared with available data. A conclusion of this investigation is that hitchhiking may well reduce standing variation (as measured in number of polymorphic sites) in samples from natural populations. At the very least it is the most plausible explanation for the reduced restriction site polymorphism observed in regions of the genome with restricted crossing over per physical length (STEPHAN and LANGLEY 1989; AGUADÉ, MIYASHITA and LANGLEY 1989).

TABLE 1

Definition of parameters

N	Diploid population size
μ	Expected number of selectively neutral mutations per nucleotide site per chromosome per generation
s	Selective advantage of the favored allele per generation
c	Expected number of crossovers per nucleotide site per genome per generation
θ	$4N\mu$
α	$2Ns$
C	$2Nc$
Case of single selected substitution	
$2N\tau$	Ancestral generation that a strongly selected allele (destined for fixation) was introduced into the population
$2N\tau_f$	Ancestral generation that a strongly selected allele (introduced into the population in generation $2N\tau$) fixed in the population
η	$\tau - \tau_f$
R	Expected number of crossovers between the neutral region and selected locus per genome per $2N$ generations
Case of multiple selected substitutions	
λ	Expected number of strongly selected substitutions per nucleotide site per generation
Λ	$2N\lambda$
Λ_r	Λ/C
Λ_{MAX}	The upper bound on Λ_r
M	Maximal distance (measured in units of recombinational distance) from the neutral region that a selected locus can be and have a hitchhiking effect on the neutral region

THEORY

We begin by considering the evolutionary effects of the substitution of a newly arising selectively favored mutant. We refer to the present population as generation zero, and the population i generations back in time as the i ancestral generation. Suppose in a randomly mating diploid population of size N, a strongly selected allele that is destined to fix in the population was introduced in the $2N\tau$ ancestral generation at a locus, and that it fixes in the population in the $2N\tau_f$ ancestral generation ($\tau_f < \tau$). The wildtype is denoted by b, the selectively favored allele by B and the frequency of the B allele in the t ancestral generation by $X(t)$, $t > 0$. We assume that the fitnesses of the three genotypes bb, bB and BB are 1, $1 + s$ and $1 + 2s$, ($s > 0$) (*i.e.*, selection is additive). This model was also considered by MAYNARD SMITH and HAIGH (1974) and OHTA and KIMURA (1975).

If the population size, $2N$, is large and selection is strong [$\alpha = 2Ns$ is large, typically $10^3 \leq \alpha \leq 10^{-2}(2N)$], then it is well known (KURTZ 1971; NORMAN 1974) that so long as the X process stays away from the boundaries 0 and 1, it can be treated as deterministic and is approximately equal to the solution of an appropriate differential equation. More specifically, if $\varepsilon > 0$ and small, then with high probability, $X(t)$ can be written as

$$X(t) = \begin{cases} X(t) & 0 \leq t \leq \tau(1-\varepsilon) \\ x(t) & \tau(1-\varepsilon) \leq t \leq \tau(\varepsilon) \\ X(t) & t < \tau(\varepsilon), \end{cases} \quad (1)$$

where

$$\tau(1-\varepsilon) = \inf\{t: X(t) = 1-\varepsilon\},$$
$$\tau(\varepsilon) = x^{-1}(\varepsilon),$$

and $x(t)$ satisfies the differential equation

$$\frac{dx(t)}{dt} = -\alpha x(t)(1-x(t)),$$
$$x(\tau(1-\varepsilon)) = 1-\varepsilon, \qquad t \geq \tau(1-\varepsilon). \quad (2)$$

The magnitude of the X process while the B allele is on its way to fixation is essentially deterministic, *i.e.*, near the boundary 0, X is near 0, near the boundary 1, X is near 1 and on the interior X is determined by the differential equation in (2). The choice of ε depends on the magnitude of α; the larger α is, the smaller ε can be. Since the probability that the mutant goes extinct equals $1-2s$, and the frequency process of the mutant in its early stages can be approximated by a branching process, ε is chosen so that $(1-2s)^{2N\varepsilon} \approx e^{-2\alpha\varepsilon} \approx 0$. It therefore suffices to choose $\varepsilon = 5/\alpha$.

The only randomness of the X process of any consequence is at the boundaries 0 and 1. This randomness affects how long the X process remains near the boundaries, and so it cannot be ignored when calculating the expected time for the favored allele to fix in the population, conditional on the event that it does fix. This expected time is relevant to later results.

Let $\eta = \tau - \tau_f$. Since the derivation of the diffusion estimate of the conditional expectation of η assumes that $s = \alpha/2N$ is small, there is reason to suspect the diffusion estimate [EWENS (1979 Eq. 5.52] when s is large, *e.g.*, $s = 10^{-2}$. To check its adequacy we derive the following estimate. It follows from the representation in (1) that the time that it takes the X process to go from $1-\varepsilon$ to ε can be assumed to be deterministic and equal to

$$\int_{\varepsilon}^{1-\varepsilon} \frac{dx}{\alpha x(1-x)} = \frac{-2\ln\varepsilon}{\alpha}.$$

To estimate the conditional expectation of the time that the X process is less than ε, we use simulation techniques. This is feasible because, when $X(t)$ is small, the process, $\{2NX(\tau - t)\}$, can be approximated by a branching process whose offspring distribution is Poisson with mean $1+s$ (EWENS 1979). Since the branching process goes extinct with probability $1-2s$ (EWENS 1979), a large number of simulations are required to obtain a sample path that eventually reaches ε. Thus the estimate is based on 50 sample paths that ultimately reach $\varepsilon = 5/\alpha$. If $X(t)$ is near 1, then the process, $\{2N(1-X(\tau-t))\}$, can be approximated by a branching process whose offspring distribution is Poisson with mean $1-s$. Therefore, the conditional expectation of the time that the X process is greater than $1-\varepsilon$ can also be estimated using simulations. Since the subcritical branching process goes extinct with probability one, there is no difficulty with the sample size. In Table 2 the estimate of the conditional expectation of η, $E^*(\eta)$, is given for different values of α ($2N = 10^8$ and $\varepsilon = 5/\alpha$). In all cases, $E^*(\eta)$ agrees with the diffusion estimate.

TABLE 2

Values of M, Λ_{MAX}, $I_{22}(M)$ and $E^*(\eta)$ for different values of α assuming $2N = 10^8$ and $\delta = 0.01$

α /	M	Λ_{MAX}	$I_{22}(M)$	$E^*(\eta)$
10^3	3.7×10^2	1.2×10^{-3}	8.8×10^1	1.6×10^{-2} (1×10^{-3})
10^4	2.6×10^3	5.3×10^{-4}	5.8×10^2	2.1×10^{-3} (1×10^{-4})
10^5	1.9×10^4	3.2×10^{-4}	4.4×10^3	2.6×10^{-4} (1×10^{-5})
10^6	1.5×10^5	2.8×10^{-4}	3.6×10^4	3.0×10^{-5} (1×10^{-6})

$E^*(\eta)$ is given for $\epsilon = 5/\alpha$. The number in the parentheses is the sample standard deviation of the time (measured in units of $2N$ generations) that the frequency process spends less than $5/\alpha$ for the 50 sample paths that reach $5/\alpha$.

The value of $E^*(\eta)$ is negligible if α is large. For example, as α increases from 10^3 to 10^6, $E^*(\eta)$ decreases from 1.6×10^{-2} to 3.0×10^{-5}. Thus, for strong selection, fixation is virtually instantaneous when time is measured in units of $2N$ generations.

Suppose we take a random sample of n genes ($n \geq 2$) and consider a small region of the genome of L nucleotide sites. Let S denote the number of selectively neutral polymorphic sites in the sample in the region. If μ, the expected number of selectively neutral mutations per nucleotide site, per chromosome, per generation is sufficiently small, then with high probability, at most one selectively neutral mutation will have occurred at each nucleotide site since the most recent common ancestor. In this case the distribution of S can be approximated as

$$P(S = k) = \int_0^{\infty} e^{-\mu t} \frac{(\mu t)^k}{k!}\, dF(t), \qquad k \geq 0, \quad (4)$$

where

$$F(t) = P\left(\sum_{i=1}^{L} t_i \leq t\right), \qquad t \geq 0,$$

and t_i is the sum of the lengths (measured in generations) of all the branches of the ancestral tree describing the genealogical history of the sample for the ith nucleotide site, $1 \leq i \leq L$ (WATTERSON, 1975). For (4) to hold, it is necessary that $\mu \max_{1 \leq i \leq L}(t_i)$ be small with high probability. If the L nucleotides are completely linked, then there is only one ancestral tree and so $\sum_{i=1}^{L} t_i = LT$, where T is the size of the ancestral tree of a single nucleotide site. Unless otherwise stated, we will assume that the region of L nucleotides (in which

selectively neutral mutations occur) is completely linked. In this case it is appropriate to think of the region as a single locus.

In view of (4), the distribution of S follows directly from the distribution of T, and so it suffices to study the distributional properties of T. For example,

$$E(S) = L\mu E(T)$$

and

$$\mathrm{Var}(S) = L\mu E(T) + (L\mu)^2\mathrm{Var}(T).$$

The formula for $E(S)$ holds even if there is recombination between the L nucleotide sites. This is not the case however, for $\mathrm{Var}(S)$ (HUDSON 1983).

Assuming an isolated selectively neutral locus, WATTERSON (1975) showed that T (measured in $2N$ generations) can be represented as

$$T = \sum_{j=2}^{n} jY(j), \tag{5}$$

where the $\{Y(j)\}$ are independent random variables, and for large N the distribution of $Y(j)$ is approximately negative exponential with parameter $(j(j-1))/2$, $2 \leq j \leq n$. The time to the most recent common ancestor of the sample, T_0, can also be represented as

$$T_0 = \sum_{j=2}^{n} Y(j). \tag{6}$$

The occurrence of a selected substitution at some time in the past can affect the distribution of T and consequently the distribution of S. Our goal is to quantify this effect, and to do this we will use the results of HUDSON and KAPLAN (1988) on the coalescent process for a sample of genes at a selectively neutral locus that is linked to a locus at which selection is operating.

Each ancestor of each sampled gene (referred to as an ancestral gene) is linked to either a B allele or a b allele. We therefore define $Q(t) = (i, j)$ if, in the t ancestral generation, $t > 0$, i of the ancestral genes of the sample are linked to a B allele and j to a b allele, $1 \leq i+j \leq n$. Since the B allele is fixed in the population at the time of sampling, it is necessarily the case that $Q(0) = (n, 0)$. In Figure 1 a possible realization of the Q process for a sample of size 4 is described.

The Q process is a jump process and because the number of ancestral genes cannot increase, this process eventually reaches either of the two states (0, 1) or (1, 0), *i.e.*, there is a single ancestor of the sample and it is linked to either a b allele or a B allele. The ancestral generation in which this first occurs, T_0, is the generation that has the most recent common ancestor of the sample.

HUDSON and KAPLAN showed that, when $2N$ is large and time is measured in units of $2N$ generations, the distribution of the Q process when conditioned on the ancestral frequency process, $\{X(t),\ t > 0\}$, of the selected allele B, can be approximated by the time inhomogeneous Markov jump process described below. From now on time will be measured in units of $2N$ generations unless otherwise specified.

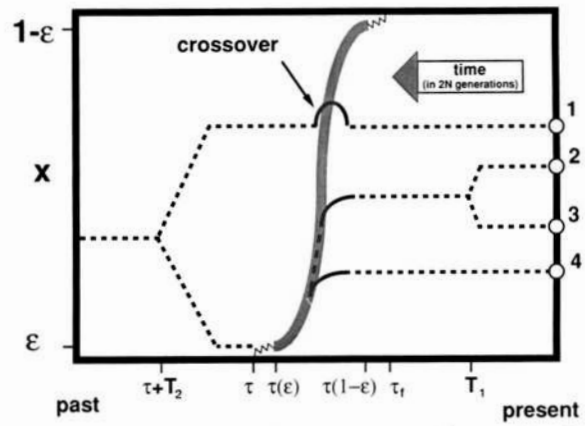

FIGURE 1.—A realization of the Q process of a sample of four genes at a selectively neutral region linked to another locus at which a selectively favored mutation (destined to fix at time τ_f) arose at time τ in the past (time measured in $2N$ generations). The most recent time that the Q process changes value [(4, 0) to (3, 0)] is T_1, the occurrence of the most recent common ancestor of sampled genes 2 and 3. During the selective phase [$0 < X(t) < 1$, $\tau_f < t < \tau$] the ancestral gene of sampled gene 1 crosses over to a wild type or b bearing chromosome, and the ancestral genes of sampled genes 3 and 4 have a most recent common ancestor [$Q(\tau) = (1, 1)$] and so $Q(\tau+) = (0, 2)$. Finally, at time $\tau + T_2$, the most recent common ancestor of the sample occurs [$Q(\tau + T_2) = (0, 1)$].

If, for any $t > 0$, $Q(t) = (i, j)$, then, according to HUDSON and KAPLAN (1988), the probability that the process jumps to a different state before $t + \Delta$ equals

$$h_{ij}(X(t))\Delta + O(\Delta^2)$$

as Δ approaches 0, where

$$h_{ij}(x) = \left(\frac{\binom{i}{2}}{x} + iR(1-x)\right)\chi(x > 0) + \left(\frac{\binom{j}{2}}{1-x} + jRx\right)\chi(x < 1), \tag{7}$$

and R is the expected number of crossovers between the neutral region and the selected locus per genome, per $2N$ generations. Also, $\chi(x > 0) = 1$ if $x > 0$ and equals 0 otherwise. $\chi(x < 1)$ is defined similarly. Finally $\binom{i}{2}$ is set equal to 0 if $i < 2$.

The only states that the Q process can jump to from (i, j) are $(i-1, j)$, $(i, j-1)$, $(i+1, j-1)$ and $(i-1, j+1)$. The first two states represent common ancestor events and the latter two crossover events.

The probabilities of these four jumps are

$$q_{i-1,j}(X(t)) = \frac{\binom{i}{2}\chi(X(t) > 0)}{X(t)h_{ij}(X(t))},$$

$$q_{i,j-1}(X(t)) = \frac{\binom{j}{2}\chi(X(t) < 1)}{(1 - X(t))h_{ij}(X(t))}, \qquad (8)$$

$$q_{i+1,j-1}(X(t)) = \frac{jRX(t)\chi(X(t) < 1)}{h_{ij}(X(t))}$$

and

$$q_{i-1,j+1}(X(t)) = \frac{iR(1 - X(t))\chi(X(t) > 0)}{h_{ij}(X(t))}.$$

The description of the Q process is complete except for one thing. Since the favored allele is introduced into the population in the τ ancestral generation, it is always the case that

$$Q(\tau+) = (0, 1 + j) \quad \text{if} \quad Q(\tau) = (1, j), \qquad j \geq 1.$$

The dynamics of the Q process can be described as follows. Going back in time until the τ_f ancestral generation, the Q process behaves as a coalescent process for a sample of size n at an isolated, selectively neutral locus, since there is no polymorphism at the linked selected locus. If $n - k$ coalescent events occur before the τ_f ancestral generation, then $Q(\tau_f) = (k, 0)$, *i.e.*, the selective phase is entered having k ancestral genes that are linked to the favored B allele. During the selective phase additional coalescent events can occur, as well as crossovers, and so the number of ancestral genes at the end of the selective phase (the τ ancestral generation), equals j where $1 \leq j \leq k$ (*i.e.*, $Q(\tau+) = (0, j)$). After the selective phase, the Q process again behaves as a coalescent process for a sample at an isolated, selectively neutral locus, but now the sample is of size j (Figure 1). The selected substitution thus accelerates the coalescing process, and the degree to which this takes place depends on α and R.

It is only during the selective phase that the complexities of the Q process are relevant. If selection is strong enough, then the time, η, of the selective phase, when measured in units of $2N$ generations, is negligible with high probability. Thus, the only quantities related to the selected event that need to be calculated in order to completely specify the distribution of the size of the ancestral tree, T, are the probabilities of entering the selective phase with k ancestral genes and exiting it with j ancestral genes, $1 \leq j \leq k \leq n$. We denote these probabilities by

$$P_{kj} = P(Q(\tau+) = (0, j) \mid Q(\tau_f) = (k, 0)) \quad 1 \leq j \leq k \leq n.$$

It is important to remember that the P_{kj} are unconditional probabilities, and so are calculated by averaging over all possible realizations of the X process.

All the other quantities that are needed to specify the distribution of T are computed assuming that the Q process behaves as a coalescent process for a sample at an isolated, selectively neutral locus. For example, to calculate the expectation of T, we write T as

$$T = T\chi(T_0 \leq \tau_f) + \sum_{k=2}^{n}\sum_{j=1}^{k} T\chi(Q(\tau_f) = (k, 0))\chi(Q(\tau+) = (0, j)). \qquad (9)$$

Let $T(t)$ denote the size of the ancestral tree for the first t ancestral generations. If η is negligible, then it follows from (9) that the expectation of T can be written as

$$E(T) = E_n(T, T_0 \leq \tau_f) + \sum_{k=2}^{n}\sum_{j=1}^{k}\left(J_{nk}(\tau_f) + L_{nk}(\tau_f)P_{kj}\sum_{i=1}^{j-1}\frac{2}{i}\right), \qquad (10)$$

where

$$E_n(T, T_0 \leq \tau_f) = E(T, T_0 \leq \tau_f \mid Q(0) = (n, 0)),$$

$$L_{nk}(t) = P(Q(t) = (k, 0) \mid Q(0) = (n, 0)),$$

$$J_{nk}(t) = E(T(t)\chi(Q(t) = (k, 0)) \mid Q(0) = (n, 0)),$$

$$t \geq 0 \quad \text{and} \quad 2 \leq k \leq n,$$

and each of the above three quantities is computed assuming that the Q process is for a sample at an isolated, selectively neutral locus. Since η is negligible, $\tau_f \approx \tau$ and so τ_f can be treated as a constant.

It is not difficult to show from the properties of the coalescent process of a sample at an isolated, selectively neutral locus that the $\{J_{nk}(t), t \geq 0\}$ and $\{L_{nk}(t), t \geq 0\}$ satisfy the following system of differential equations (GRIFFITHS 1980):

$$\frac{dJ_{nk}(t)}{dt} = -\binom{k}{2}J_{nk}(t) + \binom{k+1}{2}J_{nk+1}(t) + kL_{nk}(t),$$

$$\frac{dL_{nk}(t)}{dt} = -\binom{k}{2}L_{nk}(t) + \binom{k+1}{2}L_{nk+1}(t)$$

$$t \geq 0, \quad 2 \leq k \leq n - 1,$$

and

$$J_{nn}(t) = nte^{-\binom{n}{2}t} \quad \text{and} \quad L_{nn}(t) = e^{-\binom{n}{2}t}, \qquad t \geq 0. \qquad (11)$$

The initial values are $J_{nk}(0) = 0$, $2 \leq k \leq n$ and $L_{nk}(0) = 0$, $1 \leq k \leq n - 1$ and $L_{nn}(0) = 1$. To complete the

calculation of $E(T)$, we note that

$$E_n(T,\ T_0 \leqslant \tau_f) = \int_0^{\tau_f} J_{n2}(t)\ dt.$$

The system of equations in (11) has been solved explicitly (GRIFFITHS 1980). The solution, however, is complicated and for moderately sized samples, it is just as easy to solve the system of differential equations numerically.

The simplest case to evaluate $E(T)$ is for samples of size 2. Indeed,

$$E_2(T,\ T_0 \leqslant \tau_f) = \int_0^{\tau_f} 2ue^{-u}\ du$$
$$= 2(1 - (1 + \tau_f)e^{-\tau_f}),$$
$$J_{22}(\tau_f) = 2\tau_f e^{-\tau_f} \quad \text{and} \quad L_{22}(\tau_f) = e^{-\tau_f}.$$

Thus

$$\begin{aligned} E(T) &= 2(1 - (1 + \tau_f)e^{-\tau_f}) \\ &\quad + 2\tau_f e^{-\tau_f} + 2e^{-\tau_f}P_{22} \\ &= 2(1 - e^{-\tau_f} + e^{-\tau_f}P_{22}). \end{aligned} \tag{12}$$

We now turn our attention to the calculation of the $\{P_{kj}\}$. Define for $\tau_f \leqslant t \leqslant \tau$ and $1 \leqslant i + j \leqslant k$, $k \geqslant 2$, the probabilities,

$$P_{k,i,j}(t) = P(Q(t) = (i, j) \mid Q(\tau_f) = (k, 0)).$$

It follows from the definition of the Q process that

$$P_{kj} = P_{k,0,j}(\tau) + P_{k,1,j-1}(\tau).$$

Let $\varepsilon > 0$. If ε is sufficiently small, then it is not difficult to show from the Markov structure of the Q process that there is low probability of the Q process changing its state in the time interval $(\tau_f,\ \tau(1 - \varepsilon))$. Thus

$$P(Q(\tau(1 - \varepsilon)) = (k, 0) \mid Q(\tau_f) = (k, 0)) \approx 1,$$

and so for $t \geqslant \tau(1 - \varepsilon)$

$$P_{k,i,j}(t) \approx P(Q(t) = (i, j) \mid Q(\tau(1 - \varepsilon)) = (k, 0)).$$

On the interval $(\tau(1 - \varepsilon),\ \tau(\varepsilon))$ the X process can be treated as deterministic (see (1)), and so the Q process is a time inhomogeneous Markov process. We can therefore numerically solve the associated Kolmogorov forward differential equations to evaluate the $\{P_{k,i,j}(t),\ \tau(1 - \varepsilon) \leqslant t \leqslant \tau(\varepsilon)\}$. These equations are

$$\begin{aligned} \frac{dP_{k,i,j}(t)}{dt} &= -h_{ij}(x(t))P_{k,i,j}(t) + \frac{i(i + 1)}{2x(t)} P_{k,i+1,j}(t) \\ &\quad + \frac{j(j + 1)}{2(1 - x(t))} P_{k,i,j+1}(t) \\ &\quad + (i + 1)R(1 - x(t))P_{k,i+1,j-1}(t) \\ &\quad + (j + 1)Rx(t)P_{k,i-1,j+1}(t), \end{aligned}$$

where we recall that $x(t)$ is defined by

$$\frac{dx(t)}{dt} = -\alpha x(t)(1 - x(t))$$
$$\tau(1 - \varepsilon) \leqslant t \leqslant \tau(\varepsilon) \quad \text{and} \quad 2 \leqslant i + j \leqslant k. \tag{13}$$

The initial conditions are $P_{k,k,0}(\tau(1 - \varepsilon)) = 1$ and $P_{k,i,j}(\tau(1 - \varepsilon)) = 0$ otherwise, $P_{k,i,j}(t) = 0$ for all t if $i + j > k$, and $x(\tau(1 - \varepsilon)) = 1 - \varepsilon$.

Finally we need to relate the $\{P_{k,i,j}(\tau(\varepsilon))\}$ to the $\{P_{k,i,j}(\tau)\}$. If all the $\{P_{k,i,j}(\tau(\varepsilon)),\ 2 \leqslant i \leqslant k\}$ are negligible, *i.e.*, at time $\tau(\varepsilon)$ at most one ancestral gene is linked to a B allele, then

$$P_{kj} = P_{k,1,j-1}(\tau(\varepsilon)) + P_{k,0,j}(\tau(\varepsilon)). \tag{14}$$

If $\varepsilon = 5/\alpha$, then the $\{P_{k,i,j}(\tau(\varepsilon)),\ 2 \leqslant i \leqslant k\}$ are not always negligible. In this case there is still an opportunity for recombination to occur in the interval $(\tau(\varepsilon),\ \tau)$ and so P_{kj} as defined in (14) is an underestimate of the true value.

There are two ways to proceed. One is to extend the solution of (13) for larger values of t until the $\{P_{k,i,j}(t),\ 2 \leqslant i \leqslant k\}$ are negligible. Since we are treating the frequency of the X process deterministically in a region where it should be treated stochastically, this approach leads to an overestimate of P_{kj} [see MAYNARD SMITH and HAIGH (1974) for a discussion of this issue]. Alternatively, we can treat the X process stochastically and use simulation methods. More specifically, when $X(t)$ is small, the process, $\{2NX(\tau - t)\}$, behaves as a Poisson branching process whose offspring distribution has mean $1 + s$. Thus the frequency process can be simulated and the $\{P_{k,i,j}(t)\}$ evaluated conditional on the sample path of the frequency process. In this case the conditional Q process is a time inhomogeneous Markov chain. This method is clearly more involved and will only be carried out for a few cases in order to examine how much the first approach overestimates the true value.

The quantity, P_{kj}, is a function of R and for what is to follow, it is convenient to denote this by $P_{kj}(R)$. For $k \geqslant 2$, the asymptotic behavior of $P_{kk}(R)$, as R converges to 0 or ∞ is easy to describe. If R is very small, then crossing over is a rare event and so $P_{kk}(R)$ converges to 0 as R goes to 0 (*i.e.*, the k sampled genes remain linked to the selected allele and so the most recent common ancestor of the sample occurs during the selective phase). Alternatively, if R is very large, then the Q process behaves as a coalescent process for a sample of size k at an isolated, selectively neutral locus, and so $P_{kk}(R)$ converges to $E(e^{-\binom{k}{2}\eta})$ as R converges to ∞. If α is large, then $E(e^{-\binom{k}{2}\eta}) \approx 1$ and thus, in this case nothing happens during the selective phase.

As a concrete example, we consider a sample of size 2. The equations in (13) are

$$\frac{dP_{2,2,0}(t)}{dt} = -\left(\frac{1}{x(t)} + 2R(1 - x(t))\right)P_{2,2,0}(t) + Rx(t)P_{2,1,1}(t),$$

$$\frac{dP_{2,1,1}(t)}{dt} = -RP_{2,1,1}(t) + 2Rx(t)P_{2,0,2}(t) + 2R(1 - x(t))P_{2,2,0}(t)$$

$$\frac{dP_{2,0,2}(t)}{dt} = -\left(\frac{1}{1 - x(t)} + 2Rx(t)\right)P_{2,0,2}(t) + R(1 - x(t))P_{2,1,1}(t)$$

and

$$\frac{dx(t)}{dt} = -\alpha x(t)(1 - x(t)), \qquad \tau(1 - \varepsilon) \leq t \leq \tau(\varepsilon), \quad (15)$$

where R, the recombination parameter and α, the selection parameter are defined in Table 1. The initial conditions are

$$P_{2,2,0}(\tau(1 - \varepsilon)) = 1,$$

$$P_{2,1,1}(\tau(1 - \varepsilon)) = P_{2,0,2}(\tau(1 - \varepsilon)) = 0 \quad \text{and}$$

$$x(\tau(1 - \varepsilon)) = 1 - \varepsilon.$$

There is no possibility of solving (15) analytically and so it must be done numerically.

The quantity P_{22} can also be related to the change in expected heterozygosity at the neutral locus. Since the duration of the selective phase is so short, the only way that two genes, sampled just after fixation, can be heterozygous is if they do not have a common ancestor during the selective phase (the probability of this event is P_{22}) and their ancestral genes just prior to the introduction of the selected mutation are different (the probability of this event is assumed to be H_0). These two events are independent and so

$$H_\infty = H_0 P_{22}, \quad (16)$$

where H_∞ equals the expected heterozygosity at the neutral locus after fixation.

Up until now we have only examined the evolutionary effect of one ancestral selected substitution. Now we consider recurring selected substitutions. We recall some notation introduced earlier. Let $\Lambda = 2N\lambda$ where λ is the expected number of selected substitutions per nucleotide site, per generation. Also, let $C = 2Nc$ where c is the expected number of crossovers per nucleotide site, per genome, per generation.

Selected substitutions are assumed to occur according to a time-homogeneous Poisson process with mean Λ (time measured in units of $2N$ generations), and the location of each substitution is randomly distributed in the genome. If the physical distance of a selected substitution from the neutral region is m (measured in base pairs (bp)), then its recombinational distance from the neutral region is defined to be Cm, which equals the expected number of crossovers between the selected substitution and the neutral region per genome per $2N$ generations. It is convenient for what follows to measure Λ in units of C, and so we define $\Lambda_r = \Lambda/C$.

Since $P_{nn}(Cm)$ converges to $E(e^{-\binom{n}{2}\eta}) \approx 1$ as m goes to infinity, a selected substitution whose recombinational distance (Cm) from the neutral region is large, will have negligible effect on the coalescent process of the sample. We therefore restrict our attention to selected substitutions whose recombinational distances from the neutral region are less than some large value M, a value to be specified. (Equivalently, selected substitutions are assumed to have a physical distance less than M/C bp from the neutral region.)

In order to use the theory developed for a single selected substitution, it is necessary that at any time at most one selected allele, with recombinational distance less than M from the neutral region, is on its way to fixation. If Λ_r is small enough, then selected substitutions are infrequent and so the previous theory will apply. The question of how small Λ_r must be will be considered shortly.

M was chosen so that selected substitutions whose recombinational distances from the neutral region are greater than M have negligible effect on the coalescent process of the sample. It is reasonable to assume that this is still the case, even if several of these distant selected variants are simultaneously going to fixation.

We now define for a sample of size n at the selectively neutral locus, the analog of the Q process defined earlier, assuming that the selected substitutions are sufficiently infrequent. For $t \geq 0$ (measured in $2N$ generations) let $Q(t) = k$ if in the t ancestral generation there are k ancestral genes, $1 \leq k \leq n$. There are two ways that the Q process can change its state: either by the occurrence of a common ancestor or the occurrence of a selected substitution and at least one coalescent event occurs during the selective phase. If $Q(t) = k$, $2 \leq k \leq n$, then it is not difficult to show using previous arguments and properties of the Poisson process that as Δ goes to 0

$$P(Q(t + \Delta) = k \mid Q(t) = k) = 1 - \left(\binom{k}{2} + 2\Lambda_r I_{kk}(M)\right)\Delta + O(\Delta^2),$$

$$P(Q(t + \Delta) = k - 1 \mid Q(t) = k) = \left(\binom{k}{2} + 2\Lambda_r I_{kk-1}(M)\right)\Delta + O(\Delta^2)$$

and

$$P(Q(t + \Delta) = j \mid Q(t) = k) = 2\Lambda_r I_{kj}(M)\Delta + O(\Delta^2), \qquad 1 \leq j \leq k - 2,$$

where

$$I_{kk}(M) = \int_0^M (1 - P_{kk}(u))\, du$$

and

$$I_{kj}(M) = \int_0^M P_{kj}(u)\, du, \qquad 1 \leq j \leq k - 1.$$

Thus, the Q process is a Markov process, where the holding time in state k, $2 \leq k \leq n$, is negative exponential with parameter $\binom{k}{2} + 2\Lambda_r I_{kk}(M)$, and the jump probabilities are

$$q_{kk-1} = \frac{\binom{k}{2} + 2\Lambda_r I_{kk-1}(M)}{\binom{k}{2} + 2\Lambda_r I_{kk}(M)}$$

and

$$q_{kj} = \frac{2\Lambda_r I_{kj}(M)}{\binom{k}{2} + 2\Lambda_r I_{kk}(M)}, \qquad 1 \leq j \leq k - 2. \quad (17)$$

To calculate the mean and variance of T one can use a renewal argument similar to that used by KAPLAN, DARDEN and HUDSON (1988). For a sample of size 2 these calculations are simple since T has an exponential distribution with parameter $(1 + 2\Lambda_r I_{22}(M))/2$. Thus, for a sample of size 2

$$E(T) = \frac{2}{1 + 2\Lambda_r I_{22}(M)}$$

and

$$\mathrm{Var}(T) = \left(\frac{2}{1 + 2\Lambda_r I_{22}(M)}\right)^2. \quad (18)$$

If $2\Lambda_r I_{22}(M)$ is much smaller than 1, then $E(T) \approx 2$ which is its value for an isolated, selectively neutral locus. On the other hand, if $2\Lambda_r I_{22}(M)$ is much larger than 1, then $E(T) \approx 2/\Lambda_r I_{22}(M)$, which is less than 2. Thus, the larger $\Lambda_r I_{22}(M)$ is, the stronger the effect of hitchhiking on $E(T)$.

It remains to specify M and the range of values of Λ_r. To demonstrate how this is done, we consider the case of a sample of size 2. It follows from (17) that the distribution of the Q process is dependent on M only through the upper limit of the integral $I_{22}(M)$. For any $M^* > M$ we have

$$\int_0^{M^*} (1 - P_{22}(u))\, du = (1 + \Delta(M, M^*)) \int_0^M (1 - P_{22}(u))\, du,$$

where

$$\Delta(M, M^*) = \frac{\int_M^{M^*} (1 - P_{22}(u))\, du}{\int_0^M (1 - P_{22}(u))\, du}.$$

If $\Delta(M, M^*)$ is small, say less than a specified δ, then increasing M to M^* will not change the Q process very much. Also, we want M to be as small as possible so as not to be too restrictive on the range of values of Λ_r. Thus, we choose M so that

$$M = \inf\left\{M' : \sup_{M^* > M'} \Delta(M', M^*) < \delta\right\}. \quad (19)$$

Since the physical size of the genome is finite, M^* has a finite upper bound, M_f, and so M is well defined. Furthermore, it is not difficult to show that M satisfies

$$\int_0^M (1 - P_{22}(u))\, du = \frac{1}{1 + \delta} \int_0^{M_f} (1 - P_{22}(u))\, du. \quad (20)$$

Finally, we consider the possible range of values of Λ_r. For the Markov process defined in (17) let K_0' denote the number of selected substitutions up until T_0', the first time that $Q(t) = 1$. On the set Ω, where none of the K_0' selected variants occurring before T_0' are simultaneously going to fixation, $T_0 = T_0'$. Since the probability of Ω is a decreasing function of Λ_r, it suffices that $\Lambda_r < \Lambda_{\mathrm{MAX}}$, where

$$\Lambda_{\mathrm{MAX}} = \sup\{\Lambda_r : P(\Omega) = 1 - \delta\},$$

where δ is a small positive number.

We now calculate $P(\Omega)$. The probability that a selective substitution occurs before a coalescent event is $2\Lambda_r M/(1 + 2\Lambda_r M)$. Furthermore, the probability that no additional selective substitutions occur during the selective phase and that during the selective phase there is a common ancestor is $pI_{22}(M)/M$, where

$$p = E(e^{-2M\Lambda_r \eta}) \approx 1 - 2M\Lambda_r E(\eta),$$

remembering that the expectation is conditional on the selected allele fixing in the population.

Also, if no common ancestor occurs during the selective phase, then the entire process repeats itself. Thus

$$P(\Omega) = \frac{1}{1 + 2\Lambda_r M} + \frac{2\Lambda_r M}{1 + 2\Lambda_r M} \frac{pI_{22}(M)}{M} + \frac{2\Lambda_r M}{1 + 2\Lambda_r M} p\left(1 - \frac{I_{22}(M)}{M}\right) P(\Omega),$$

and so

$$P(\Omega) = \frac{1 + 2\Lambda_r I_{22}(M)p}{1 + 2\Lambda_r I_{22}(M)p + 2\Lambda_r M(1-p)}$$
$$\approx 1 - \frac{2\Lambda_r M(1-p)}{1 + 2\Lambda_r I_{22}(M)}. \quad (21)$$

Since p is close to 1, $1 - p \approx 2\Lambda_r M E(\eta)$. Hence, Λ_{MAX} satisfies

$$\frac{(2\Lambda_{MAX}M)^2 E^*(\eta)}{1 + 2\Lambda_{MAX} I_{22}(M)} = \delta, \quad (22)$$

where δ is a small positive number and $E^*(\eta)$ is the estimate of $E(\eta)$.

The renewal argument used to derive (21) can also be used to obtain the expectation of K_0, so long as $\Lambda_r \leq \Lambda_{MAX}$. Indeed,

$$E(K_0) = \frac{2\Lambda_r M}{1 + 2\Lambda_r M}\frac{pI_{22}(M)}{M} + \frac{2\Lambda_r M}{1 + 2\Lambda_r M} p\left(1 - \frac{I_{22}(M)}{M}\right)(1 + E(K_0))$$
$$= \frac{2\Lambda_r I_{22}(M)p}{1 + 2\Lambda_r I_{22}(M)p + 2\Lambda_r M(1-p)}$$
$$\approx \frac{2\Lambda_r M}{1 + 2\Lambda_r I_{22}(M)}.$$

The analysis up until now has assumed that the initial frequency of the favored allele is $1/2N$ as it would be in the case of a newly arising mutant that is destined to fix. If the favored allele is a rare existing variant that for some reason becomes selectively favored, then the analysis presented above still holds.

CALCULATIONS

For simplicity, we only examine a sample of size 2. Similar calculations can be made for larger samples. When considering the hitchhiking effect caused by the substitution of a single, rare, selected allele, the distribution of the Q process for a sample of size 2, only depends on $P_{22}(R)$, where R is the expected number of crossovers per genome per $2N$ generations between the selectively neutral region and the selected locus. In Figure 2, $P_{22}(R)$ is plotted as a function of R for different values of α. To calculate $P_{22}(R)$ we used formula (14) with $\varepsilon = 10^2/2N$ and the population size, $2N = 10^8$. This value of ε is small enough so that $P_{2,2,0}(\tau(\varepsilon))$ is negligible. Calculations not presented here show that the curves in Figure 2 are fairly insensitive to $2N$ so long as $2N \geq 100\alpha$.

Since $5/\alpha > 10^2/2N$, the curves in Figure 2 overestimate $P_{22}(R)$. In order to examine how much of an overestimation there is, a simulation was performed as described in the theory section. For each value of α, $P_{22}(R)$ was evaluated for $R = 10$, $R = 100$ and $R = 1000$. For each set of parameter values 50 sample paths of the X process that reach $5/\alpha$ were simulated.

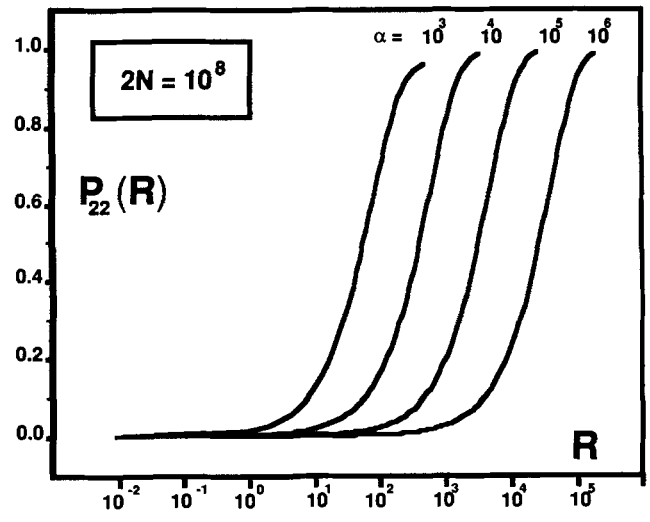

FIGURE 2.—$P_{22}(R)$, the probability of escaping the hitchhiking effect for a sample of size 2, is plotted against R, the expected number of crossovers between the selectively neutral region and the selected locus per genome per $2N$ generations for various values of α, where $2N = 10^8$ (see text for explanation).

In all cases the estimate of $P_{22}(R)$ differed from the value plotted in Figure 2 by less than 2%. For example, if $\alpha = 10^4$ and $R = 100$, then the plotted value of $P_{22}(R)$ is 0.158, while the estimate obtained from the simulation is 0.155. Thus the error caused by using the deterministic model for the frequency of the B allele when the frequency is small, is negligible.

As the amount of crossing over increases between the neutral region and the selected locus, the probability of no common ancestor occurring during the selected phase, $P_{22}(R)$, would be expected to increase. It is clear from Figure 2 that $P_{22}(R)$ is an increasing function of R. Increasing α decreases η, the time of the selective phase, and so one would expect that $P_{22}(R)$ decreases as a function of α. This behavior is also seen in Figure 2. Another way to interpret this observation is that the larger α is, the larger the region of the genome that is affected by the selected substitution.

The curves in Figure 2 are all similar in shape and they appear to be equal distance from each other, suggesting that $P_{22}(R)$ may be a function of R/α and so independent of population size. Direct calculation however, shows that this is only approximately true.

The expectation, $E(T)$, whose formula is given in (12), is plotted in Figure 3a for different values of R, and in Figure 3b for different values of α as a function of τ (measured in $2N = 10^8$ generations), the ancestral time when the selected mutation was introduced into the population (or when the rare allele becomes selectively favored). In Figure 3a, $\alpha = 10^4$ and in Figure 3b, $R = 10$. It is not difficult to show from Figure 2 and Equation (12) that $E(T)$ is an increasing function of τ and R and decreasing function of α. In particular, as is seen in Figure 3, a and b, $E(T)$ will differ

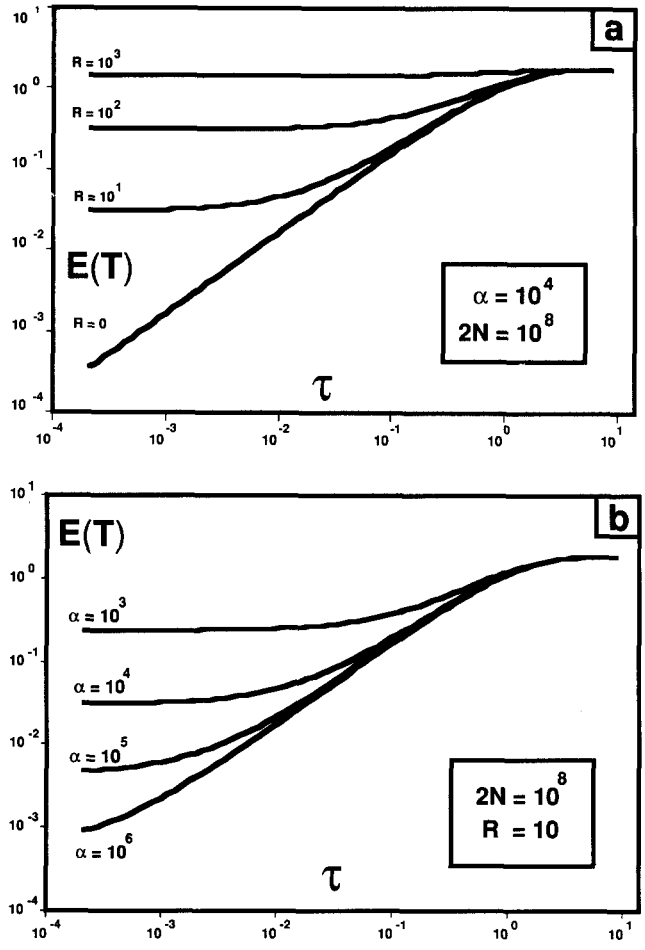

FIGURE 3.—$E(T)$, the expected size (measured in $2N$ generations) of the ancestral tree of a sample of two genes at a selectively neutral region that is linked to a selected locus (Equation 12), is plotted against τ, the ancestral time of fixation of the selected substitution, **a**. For different values of R, the expected number of crossovers between the neutral region and the selected locus per genome per $2N$ generations ($\alpha = 10^4$), and **b**. For different values of selection, α ($R = 10$); (see text for explanation).

significantly from 2, its neutral value, if $\tau \leq 0.1$ and $R/\alpha \leq 0.01$. This means that the expected level of variation will be substantially reduced for all the sites within a physical distance of $(0.01)\alpha/C$ base pairs of a locus at which a selected substitution has recently occurred. For example, if $2N = 10^8$, $s = 10^{-4}$ and $c = 10^{-8}$, then the width of the affected region is only about 200 bp. But if $s = 10^{-3}$ and $c = 10^{-8}$, then the expected variation is reduced in a region about 2000 bp wide.

In Table 2 the values of M (Equation 20), Λ_{MAX} (Equation 22) and $I_{22}(M)$ are given for different values of α ($2N = 10^8$ and $\delta = 0.01$). The value of M_f in (20) is chosen so that $1 - P_{22}(M_f)$ is within 1% of its limit, $1 - E(e^{-\eta}) \approx E^*(\eta)$. The quantities M and $I_{22}(M)$ increase more or less linearly with α, and the ratio $I_{22}(M)/M \approx 0.24$, independent of the value of α. The product $2M\Lambda_{MAX}$ varies between 0.8 and 84 as α varies between 10^3 and 10^7. All the quantities in Table 2 are insensitive to $2N$ so long as $\alpha \leq 10^{-2}(2N)$ (calculations not shown).

The major goal of this paper is to determine the consequence of hitchhiking resulting from recurring selected substitutions on standing, selectively neutral variation at the DNA level. In Figure 4 $E(T)$, (Equation 18), is plotted as a function of Λ_r for different values of α with $2N = 10^8$. Since $I_{22}(M)$ is insensitive to $2N$ (for fixed α), the same is true of $E(T)$. Even for small values of Λ_r the hitchhiking effect can reduce $E(T)$ substantially from 2 (its expectation for an isolated, selectively neutral locus) for large values of α, *e.g.*, if $\alpha \geq 10^5$, and $0.0002 \leq \Lambda_r \leq \Lambda_{MAX}$, then $E(T) \leq 0.7$. Since the expected number of polymorphic sites, $E(S)$, is proportional to $E(T)$, it is clear that the hitchhiking effect associated with the rapid fixation of selected mutants (or very rare alleles) can substantially reduce the expected number of polymorphic sites in a sample from that expected in the absence of selection.

The expected number of selected substitutions in a region of size $2M$ until the most recent common ancestor of the sample, $E(K_0)$, can be calculated from Equation 23. The values of $E(K_0)$ corresponding to Λ_{MAX} range from 0.7 (for $\alpha = 10^3$) to 4.1 (for $\alpha = 10^6$). It is interesting to note that this latter value is near the maximum value of $E(K_0)$, $M/I_{22}(M)$, which is about 5 and independent of α.

DISCUSSION

The analysis of the theoretical population genetics model of selectively neutral molecular variation under the forces of mutation and random genetic drift (KIMURA 1983; GILLESPIE 1987) has yielded many important and useful predictions. The ability of the theory to explain much of the observed variation within and between species has led many to accept the proposition that most molecular polymorphism within and divergence between species is of no phenotypic consequence to the fitness of the organisms. Two critical assumptions of the neutral theory are that selected variants are so rare that they comprise a minute portion of molecular genetic variation and that their dynamics have negligible effects on the dynamics of the preponderant, neutral variation. It is this second assumption that we have investigated.

MAYNARD SMITH and HAIGH (1974) studied the effect of a single selected substitution of a newly arising mutant (or rare variant) on a neutral polymorphism, and showed that the hitchhiking effect of a single selected substitution can substantially reduce heterozygosity at a linked selectively neutral polymorphic locus. The experimental data motivating their investigation was the mounting evidence of allozyme polymorphism in the early 1970s. Today more de-

tailed information on molecular genetic variation in natural populations is accumulating and so what is needed is an analysis of the cumulative hitchhiking effect of multiple, linked selected substitutions that affords predictions about DNA sequence variation in samples from natural populations.

For DNA sequence data, a natural summary statistic is S, the number of polymorphic sites in the sample. To study the consequences of hitchhiking on the distribution of S, it suffices to consider the effect of recurring, linked selected substitutions on the coalescent process for a random sample of genes at a selectively neutral locus. This approach differs from more classical population genetics theory (in particular the approach of MAYNARD SMITH and HAIGH) in that the genealogical history of a sample is the focus rather than population variables or statistics thereof, *e.g.*, heterozygozity or half-life of neutral polymorphism.

The expected time to the common ancestor of two randomly sampled genes at a selectively neutral locus that is linked to loci at which strongly selected substitutions have occurred in the history of the sample is our primary interest. Our results can in principal, be generalized to study higher moments and/or larger samples. However, the calculations become much more complicated. Our analysis is an extension of that of KAPLAN, DARDEN and HUDSON (1988) and HUDSON and KAPLAN (1988) and proceeds in two steps. Using their techniques, we first quantify the effect of single, linked ancestral, selected substitution on the time to the common ancestor. The magnitude of the effect on the expected time depends on the population size, when the selected substitution occurred, the strength of the selection and finally the amount of crossing over between the two loci. Next we identify the domain of the parameters where these results can be extended to the case of recurring selected substitutions at random distances from the selectively neutral region of interest. The critical assumption for this extension is that selected substitutions near the neutral region be sufficiently rare so that it is unlikely that in any generation more than one would be polymorphic in the population.

The parameter λ, the expected number of selected substitutions per nucleotide site per generation, has two interpretations. On one hand, it can reflect the rate at which the environment changes causing previously rare deleterious alleles to be selectively favored. In this case λ and Λ_r are constant and do not depend on the population size. Alternatively, the fixation of favored variants may be mutation-limited. In this case, λ can be approximated as $(2N\mu_s)(2s)$, where μ_s is the mutation rate to favorable variants per nucleotide site per chromosome per generation, each of which is assumed to have selective advantage s and $2s$ in the heterozygote and homozygote, respectively.

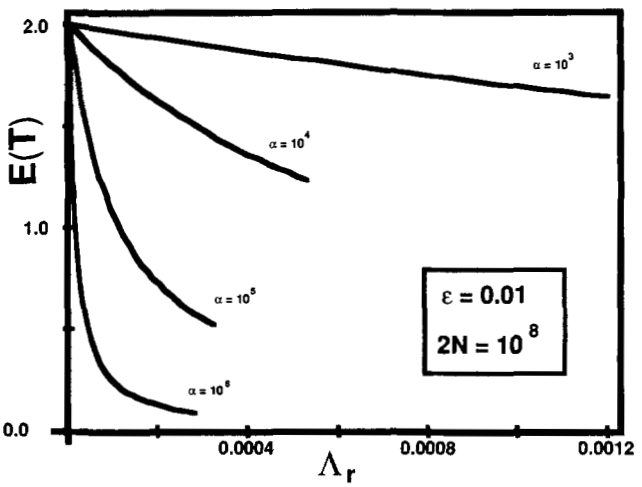

FIGURE 4.—$E(T)$, the expected size (measured in $2N$ generations) of the ancestral tree of a sample of 2 genes at a selectively neutral region assuming recurring selected substitutions (Equation 16), is plotted as a function of Λ_r, for different values of α where $2N = 10^8$ (see text for explanation). Each curve is plotted for the values of Λ_r between 0 and Λ_{MAX} (see Table 2).

Thus λ and Λ_r are proportional to $2N$, assuming μ_s, s and c are fixed.

The hitchhiking effect of recurring, linked selected substitutions on T and consequently on S, can be substantial. As Figure 4 shows the expected time to the common ancestor of a sample of two genes can be more than an order of magnitude below its predicted value under the neutral theory. Furthermore, this effect is sensitive to the strength of selection and the rate of crossing over.

Decreasing c, the rate of crossing over per nucleotide site, per genome, per generation, increases Λ_r and so as seen in Figure 4, decreases $E(T)$. Thus in regions of low crossing over per physical length, one would expect that the number of polymorphic sites in a random sample would be less than that predicted by the neutral theory. The recent data of STEPHAN and LANGLEY (1989) and AGUADÉ, MIYASHITA and LANGLEY (1989) surveying the levels of restriction map variation in chromosomal regions with reduced levels of crossing over per physical length support this prediction.

MAYNARD SMITH and HAIGH (1974) investigated the reduction in heterozygosity due to the substitution of a linked favored mutation, B. They developed a two locus deterministic model that provided the limiting value of heterozygosity as a function of the initial frequency, R_0 of the neutral allele, A, the coefficient of selection, s, and the recombination fraction, r (*i.e.*, $r = R/2N$). Their calculation of heterozygosity is conditional on the selected mutation occurring on a chromosome bearing a, the alternative allele to A. By weighting their expression by the probability $(1 - R_0)$ that the selected mutation occurs on a chromosome bearing an a allele and adding it to the symmetrical

expression multiplied by the probability (R_0) that the selected mutation occurs on an A bearing chromosome, the following unconditioned expression for heterozygosity can be obtained:

$$H_\infty = (1 - R_0)Q_\infty(ab)(1 - Q_\infty(ab)) + R_0Q_\infty(Ab)(1 - Q_\infty(Ab)), \quad (24)$$

where $Q_\infty(ab)$ and $Q_\infty(Ab)$ are the frequencies of the AB allele after fixation, conditional on the selected mutation occurring on an ab or Ab chromosome, respectively. The subsequent calculations of MAYNARD SMITH and HAIGH (1974) that are based on the ratio H_∞/H_0 (*e.g.*, Equation 29) are also conditioned on the selected mutation occurring on an a bearing chromosome and so they too need to be modified appropriately. EWENS (1979, pp. 206) used the expression for H_∞ in (24) in his discussion of OHTA and KIMURA's (1975) analysis of the hitchhiking effect.

In the diploid case MAYNARD SMITH and HAIGH (1974) were unable to obtain closed form expressions for $Q_\infty(ab)$ and $Q_\infty(Ab)$, while in the haploid case they showed that

$$Q_\infty(ab) = R_0D,$$

where

$$D = r(1 - X(0)) \sum_{n=0}^{\infty} \left(\frac{(1 - r)^n}{1 - X(0) + X(0)(1 + s)^{n+1}} \right),$$

and r is the recombination fraction between the neutral locus and the selected locus per genome per generation. We note by symmetry that

$$Q_\infty(Ab) = (1 - R_0)D.$$

If we substitute the expressions for $Q_\infty(ab)$ and $Q_\infty(Ab)$ in (24), then

$$\frac{H_\infty}{H_0} = D(2 - D).$$

The diploid and haploid cases are effectively the same when s is small, and so one might expect from (16) that $P_{22} \approx D(2 - D)$. Numerical calculations, however, show P_{22} is smaller than $D(2 - D)$. One reason why P_{22} is less than $D(2 - D)$ is because Equation 24 is only valid if the variance of the frequency of the AB allele after fixation is negligible, and this in general is not the case. Ignoring this variability leads to an overestimate of the ratio, H_∞/H_0. A discrepancy between P_{22} and $D(2 - D)$ exists even when P_{22} is very small, the case where one might expect the variance of the AB allele after fixation to be negligible. For example, if $2N = 10^8$, $s = 10^{-3}$ and $r = 10^{-6}$, then $P_{22} = 2.19 \times 10^{-6}$ and $D(2 - D) = 3.6 \times 10^{-6}$. Part of the problem in this case may be that MAYNARD SMITH and HAIGH's calculation of $Q_\infty(ab)$ uses a deterministic model for the frequency of B allele regardless of whether the frequency is very small or very large. This, as pointed out by MAYNARD SMITH and HAIGH, results in an overestimate of $Q_\infty(ab)$.

The simplest prediction of the neutral theory of molecular evolution is that expected heterozygosity should increase with population size. The apparent failure of allozyme survey results (LEWONTIN 1974) to support this prediction stimulated MAYNARD SMITH and HAIGH (1974) to investigate the hitchhiking effect. Their results certainly indicate that the hitchhiking effect "can account for the uniformity of heterozygosity between species." Our analysis of the hitchhiking effect on the expected number of segregating sites in samples from natural populations supports their contention. In the absence of selection $E(T) = 2$ and so $E(S)$, which equals $2N\mu E(T)$, is proportional to $2N$. In the presence of hitchhiking, $E(T)$ is a decreasing function of α and Λ_r (Figure 4), and so $E(S)$ is no longer proportional to $2N$. For example, suppose $s = 10^{-3}$, $2N = 10^8$ ($\alpha = 10^5$) and Λ_r is large enough so that there is a hitchhiking effect on $E(T)$, say 2.5×10^{-5} ($E(T) = 1.64$). If λ reflects the rate of change of the environment (*i.e.*, it does not depend on $2N$ and so Λ_r is a constant), then $E(S)$ increases fivefold as $2N$ increases tenfold from 10^8 to 10^9. Alternatively, if the fixation of favored mutations is mutation-limited (*i.e.*, λ and Λ_r are proportional to $2N$), then $E(S)$ actually decreases by more than 35% when $2N$ increases from 10^8 to 10^9.

The hitchhiking effect on $E(T)$ can be substantial, even when the fraction of substitutions that are selectively fixed, f_s is negligible as assumed in the neutral theory of molecular evolution. For example, if $c = 10^{-8}$, $\mu = 10^{-9}$ (reasonable values), $2N = 10^9$, $s = 10^{-3}$ and $\Lambda_r = 2.5 \times 10^{-4}$, then $E(T) = 0.096$, and

$$f_s = \lambda/(\mu + \lambda) \approx C\Lambda_r/2N\mu = 2.5 \times 10^{-3}.$$

Thus on average, one out of every 400 substitutions is selectively fixed, while $E(T)$ [and thus $E(S)$] is only 1/20 of its value under strict selective neutrality.

The analysis we have presented requires several assumptions. The requirement that α is large ensures that we can model the frequency process at the selected locus deterministically. This is not a stringent assumption since there is no hitchhiking effect when α is small. A second and more important limitation concerns the magnitude of Λ_r. The results are only valid when Λ_r is sufficiently small so that it can be assumed that at any one time at most only one selected, linked substitution is on its way to fixation. This constraint on Λ_r is a technical requirement and has no biological counterpart. Actual values of Λ_r in natural populations could well exceed this artificial upper bound. The consequences of hitchhiking in natural population with the larger values of Λ_r requires further investigation.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

GILLESPIE, J. H., 1987 Molecular evolution and the neutral allele theory. Oxf. Surv. Evolutionary Biol. **4:** 10–37.

GRIFFITHS, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Popul. Biol. **17:** 37–50.

HAIGH, J., and J. MAYNARD SMITH, 1976 The hitchhiking effect—a reply. Genet. Res. **27:** 85–87.

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

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

KAPLAN, N. L., T. DARDEN and R. R. HUDSON, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KINGMAN, J. F. C., 1982 The coalescent. Stochast. Proc. Appl. **13:** 235–248.

KOJIMA, K. I., and H. E. SCHAEFFER, 1967 Survival process of linked genes. Evolution **21:** 518–531.

KURTZ, T. G., 1971 Limit theorems for sequences of jump Markov processes approximating ordinary differential equations. J. Appl. Prob. **8:** 344–356.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favorable gene. Genet. Res. **23:** 23–35.

NORMAN, M. F., 1974 A central limit theorem for Markov processed that move by small steps. Ann. Prob. **2:** 1065–1074.

OHTA, T., and M. KIMURA, 1975 The effect of a selected linked locus on heterozygosity of neutral alleles (the hitchhiking effect). Genet. Res. **25:** 313–325.

OHTA, T., and M. KIMURA, 1976 Hitchhiking effect—a counter reply. Genet. Res. **28:** 307–308.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the *x* chromosome in three *Drosophila ananassae* populations. I. Contrast between *vermilion* and *forked* loci. Genetics **121:** 89–99.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Pop. Biol. **26:** 119–164.

WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Pop. Biol. **10:** 256–276.

Communicating editor: M. TURELLI