The Coalescent Process in Models With Selection and Recombination

Richard R. Hudson[1] and Norman L. Kaplan

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received March 7, 1988
Revised copy accepted July 14, 1988

ABSTRACT

The statistical properties of the process describing the genealogical history of a random sample of genes at a selectively neutral locus which is linked to a locus at which natural selection operates are investigated. It is found that the equations describing this process are simple modifications of the equations describing the process assuming that the two loci are completely linked. Thus, the statistical properties of the genealogical process for a random sample at a neutral locus linked to a locus with selection follow from the results obtained for the selected locus. Sequence data from the alcohol dehydrogenase (Adh) region of *Drosophila melanogaster* are examined and compared to predictions based on the theory. It is found that the spatial distribution of nucleotide differences between Fast and Slow alleles of Adh is very similar to the spatial distribution predicted if balancing selection operates to maintain the allozyme variation at the Adh locus. The spatial distribution of nucleotide differences between different Slow alleles of Adh do not match the predictions of this simple model very well.

IN a companion study, KAPLAN, DARDEN and HUDSON (1988) described the process of the genealogical history of a random sample of genes at a locus which is not selectively neutral. They showed that this process has a relatively simple representation and they developed methods for calculating the moments of such quantities as T, the sum of the lengths of the branches of the ancestral tree.

The purpose of this paper is to study the distribution of the coalescent process for a random sample of genes at a selectively neutral locus which is linked to a locus which is not selectively neutral. If the two loci are completely linked, then the genealogical history of the neutral locus is the same as that of the selected locus and so the results of the companion study apply. On the other hand, if the two loci are completely unlinked, then the distribution of the coalescent process for the neutral locus is unaffected by the selected locus. In the THEORY section the distribution of the coalescent process for the neutral locus is determined for arbitrary rates of recombination.

The coalescent process for a sample of genes at two selectively neutral loci which are linked to a selected locus is also investigated. This distribution is needed to calculate such quantities as the variance of the number of segregating sites at m ($m \geq 1$) selectively neutral loci which are linked to a selected locus.

As an application, we consider the sequence data of KREITMAN (1983) which encompasses the *Adh* locus of *Drosophila melanogaster*. Several studies of variation at this locus suggest the presence of a balanced polymorphism at the *Adh* locus (*e.g.*, OAKSHOTT *et al.* 1982; KREITMAN and AGUADÉ 1986; HUDSON, KREITMAN and AGUADÉ 1987). This possibility is examined in light of the results in this paper.

THEORY

The coalescent process for a selectively neutral m-loci model with recombination has been studied by HUDSON (1983) and KAPLAN and HUDSON (1985). The genealogy for m linked loci is a collection of m correlated ancestral trees. For finite rates of recombination the topologies and the lengths of the branches of different ancestral trees are correlated because of linkage. In the neutral case, HUDSON (1983) described the structure of this more complicated coalescent process and showed how to simulate it. KAPLAN and HUDSON (1985) extended HUDSON's work and established recursive equations for calculating the moments of the number of segregating sites at the m linked loci.

In this section we study the coalescent processes for one and two selectively neutral loci which are linked to a locus which is not selectively neutral. These two cases are necessary to calculate such quantities as the mean and variance of the number of selectively neutral segregating sites.

We begin the analysis by considering one neutral locus. It is assumed that at the selected locus, A, there are two alleles, A_1 and A_2. The neutral locus is denoted by B. For generation t, $X_{11}(t)$, $X_{12}(t)$ and $X_{22}(t)$ represent the frequencies of A_1A_1, A_1A_2 and A_2A_2 diploids, respectively, in a population of size N. The frequency of the A_1 allele in generation t is denoted by $X(t)$. It is

[1] Present address: Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717.

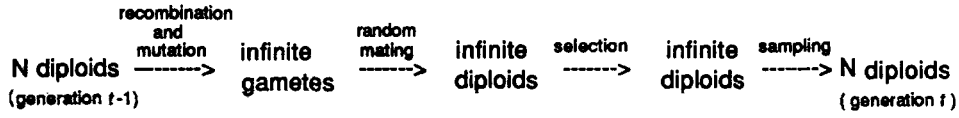

FIGURE 1.—The life cycle.

assumed that the population has achieved stationarity and so the current generation from which the sample is taken is denoted as the 0th generation. The time parameter t thus takes on both positive and negative values, where negative generation times denote ancestral generations and positive generation times future generations.

Each generation the daughter population is obtained by random sampling with replacement after mutation, recombination and selection have occurred. The life cycle of the process is shown in Figure 1. The fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are w_{11}, w_{12} and w_{22}, respectively, and the mean fitness is denoted by $\bar{w}(t)$. The rates of mutation at the selected locus, A, are u (A_1 to A_2) and v (A_2 to A_1) and the rate of recombination between the A and B locus is r. It is assumed that

$$w_{11} = w_{12} = w_{22} = 1 + O\left(\frac{1}{N}\right),$$

$$u = \frac{\beta_1}{2N} + O\left(\frac{1}{N^2}\right), \qquad v = \frac{\beta_2}{2N} + O\left(\frac{1}{N^2}\right)$$

and

$$r = \frac{R}{2N} + O\left(\frac{1}{N^2}\right),$$

where $\beta_1 > 0$, $\beta_2 > 0$ and $R > 0$.

We now consider the structure of the coalescent process for a sample of genes at the neutral B locus. Since the ancestral genes of the sample are linked to either an A_1 or A_2 allele, the coalescent process is two dimensional. If n genes are chosen at random from the 0th generation, then we let $Q(0) = (i, j)$ if i of the sampled genes are linked to an A_1 allele and j to an A_2 allele, $0 \le i, j \le n$, $i + j = n$. For $t < 0$, $Q(t)$ denotes the numbers of ancestral genes linked to A_1 and A_2 alleles in generation t.

By its very definition the Q process is a jump process. We define $T_1, T_2, \ldots$ to be the number of generations between successive jumps and $Z_1, Z_2, \ldots$ the successive random states to which the process moves. The Q process can therefore be represented as

$$Q(t) = \begin{cases} Q(0) & \text{if } |t| < T_1 \\ Z_k & \text{if } S_k \le |t| \le S_k + T_{k+1} \end{cases}$$

where $S_k = \sum_{i=1}^{k} T_i$, $k \ge 1$.

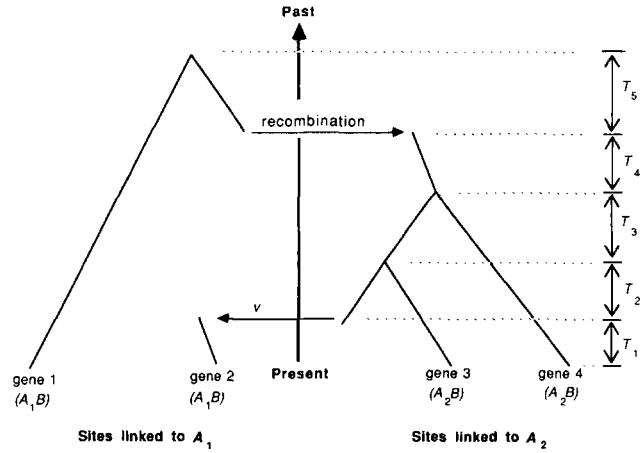

FIGURE 2.—A realization of the coalescent process for a sample of size four. At generation zero (the present) the Q process is at state (2, 2), *i.e.*, two of the B genes are linked to A_1 alleles and two are linked to A_2 alleles. The first change of state, as one follows the process back in time, is at time $-T_1$ (*i.e.*, T_1 generations before present), at which time the Q process moves to state (1, 3). This change of state is the result of the mutation of an ancestral A_2 allele to an A_1 allele and is indicated by a v in the figure. At time $-(T_1 + T_2)$ and time $-(T_1 + T_2 + T_3)$ common ancestors of genes linked to A_2 alleles occur, and so the Q process moves to (1, 2) and then to (1, 1). At time $-(T_1 + T_2 + T_3 + T_4)$ the Q process moves to (2, 0) as a result of a recombination event in which an ancestral A_1B parent produces an A_2B offspring. And finally, at time $-(T_1 + T_2 + T_3 + T_4 + T_5)$, the most recent common ancestor of the sample at locus B occurs, *i.e.*, the Q process moves to (1, 0).

Since the number of ancestral genes does not increase as one goes back in time, the Q process eventually reaches either of the two states (1, 0) or (0, 1), *i.e.*, there is a single ancestor of the sample at the B locus and it is linked to an A_1 allele or an A_2 allele. The ancestral generation in which this first occurs is that generation which has the most recent common ancestor of the sample at the B locus. An example of a realization of the coalescent process is shown in Figure 2.

The joint distribution of the $\{T_i\}$ and $\{Z_i\}$ can be computed using the arguments of the preceding paper. The only difference is in the calculation of the quantity $f_{A_j}(A_i,t)$ which in this paper denotes the probability that a randomly chosen gene at locus B from generation t is linked to an A_i allele and its parental gene from generation $t - 1$ is linked to an A_j allele, $1 \le i, j \le 2$. For the life cycle in Figure 1, standard population genetic arguments can be used

to show that

$$f_{A_1}(A_1, t) = \frac{1}{\bar{w}(t-1)}\left(X(t-1)^2 w_{11} + X(t-1)(1-X(t-1))w_{12}\right) + O\left(\frac{1}{N}\right)$$

$$= X(t-1) + O\left(\frac{1}{N}\right),$$

$$f_{A_1}(A_2, t) = \frac{1}{\bar{w}(t-1)}\left(uX(t-1)^2 w_{11} + (u+r)X(t-1)(1-X(t-1))w_{12}\right) + O\left(\frac{1}{N^2}\right)$$

$$= \frac{X(t-1)(\beta_1 + R(1-X(t-1)))}{2N} + O\left(\frac{1}{N^2}\right),$$

$$f_{A_2}(A_2, t) = \frac{1}{\bar{w}(t-1)}\left((1-X(t-1))^2 w_{22} + X(t-1)(1-X(t-1))w_{12}\right) + O\left(\frac{1}{N}\right)$$

$$= 1 - X(t-1)\, O\left(\frac{1}{N}\right)$$

and

$$f_{A_2}(A_1, t) = \frac{1}{\bar{w}(t-1)}\left(v(1-X(t-1))^2 w_{22} + (v+r)X(t-1)(1-X(t-1))w_{12}\right) + O\left(\frac{1}{N^2}\right)$$

$$= \frac{(1-X(t-1))(\beta_2 + RX(t-1))}{2N} + O\left(\frac{1}{N^2}\right).$$

Thus, all the formulas of KAPLAN, DARDEN and HUDSON (1988) apply providing that β_1 and β_2 are replaced by

$$\beta_1(x) = \beta_1 + R(1-x) \quad \text{and} \quad \beta_2(x) = \beta_2 + Rx. \quad (1)$$

The coalescent process for a random sample at the neutral B locus thus behaves much like the coalescent process for a random sample at the selected locus and so the conclusions of the preceding paper apply. For example, if the frequency of A_1 is maintained by selection at x_0, *i.e.*, the allelic frequencies are tightly regulated, then the mean and the variance of T for the selectively neutral B locus are different from their neutral values only if both $\beta_1(x_0)$ and $\beta_2(x_0)$ are small. (See Table 1 of KAPLAN, DARDEN and HUDSON 1988.) Since $\beta_1(x_0)$ is greater than $R(1-x_0)$ and $\beta_2(x_0)$ is greater than Rx_0, the mean and the variance of T will therefore look neutral if R is greater than, say, two. The distribution of T can also be calculated when the allelic frequencies are not tightly regulated by replacing β_1 and β_2 by $\beta_1(x)$ and $\beta_2(x)$ in Equation 20 of KAPLAN, DARDEN and HUDSON (1988).

We now consider the coalescent process for two neutral loci, B and C, which are linked to a selected locus, A. This process is more complicated than the previous case, since one now needs to keep track of whether or not the genes at the B and C loci are ancestral to the genes at the two loci in the sample. There are six different types of ancestral chromosomes: A_1BC, A_1B–, A_1–C, A_2BC, A_2B– and A_2–C. The dash indicates that the gene at that locus is not ancestral to the sample. The effect of recombination depends on whether the selected locus, A, lies to the right of, to the left of, or between the two neutral loci. For the sake of definiteness, we will assume that the relative positions of the three loci are ABC, and the modifications for the other case will be indicated. The rates of recombination between loci A and B and A and C are denoted by r_{AB} and r_{AC}, respectively. As before it is assumed that

$$r_{AB} = \frac{R_{AB}}{2N} + O\left(\frac{1}{N^2}\right) \quad \text{and} \quad r_{AC} = \frac{R_{AC}}{2N} + O\left(\frac{1}{N^2}\right),$$

where $R_{AC} \geq R_{AB} \geq 0$.

Despite the increased complexity of the ancestral history, the same kinds of arguments as before can be used to show that when time is scaled in units of $2N$ generations and the frequencies of the alleles at the A locus are tightly regulated, the coalescent process behaves like a Markov jump process (KARLIN and TAYLOR 1981). To specify the parameters of this process we need to introduce some notation. Let $Q(t) = (i, j)$ where $i = (i_1, i_2, i_3)$, $j = (j_1, j_2, j_3)$ and i_1, i_2, i_3, j_1, j_2 and j_3 are the numbers of A_1BC, A_1B–, A_1–C, A_2BC, A_2B– and A_2–C ancestral genes in generation t. When the Q process changes state, each of the components of i and j will either increase by 1, decrease by 1 or remain the same. To simplify the notation only those components which change will be indicated.

There are three different ways that the coalescent process can change state: coalescence, mutation and recombination. A common ancestor can occur only between ancestral genes which are linked to the same allele at the A locus, and so there are 8 transitions resulting from coalescent events. There are 6 transitions which result from mutation events and 10 trans-

TABLE 1

Possible transitions from (i, j), involving genes linked to the A_1 allele, and their probabilities (up to order $1/N$)

Transition	Probability
1. Transitions resulting from coalescence for loci configurations ABC, BCA and BAC	
$(i_1 - 1)$	$\frac{\binom{i_1}{2}}{2Nx_0}$
$(i_2 - 1)$	$\frac{\binom{i_2}{2}}{2Nx_0} + \frac{i_1 i_2}{2Nx_0}$
$(i_3 - 1)$	$\frac{\binom{i_3}{2}}{2Nx_0} + \frac{i_1 i_3}{2Nx_0}$
$(i_1 + 1, i_2 - 1, i_3 - 1)$	$\frac{i_2 i_3}{2Nx_0}$
2. Transitions resulting from mutation for loci configurations ABC, BCA and BAC	
$(i_1 - 1, j_1 + 1)$	$\frac{i_1\beta_2(1 - x_0)}{2Nx_0}$
$(i_2 - 1, j_2 + 1)$	$\frac{i_2\beta_2(1 - x_0)}{2Nx_0}$
$(i_3 - 1, j_3 + 1)$	$\frac{i_3\beta_2(1 - x_0)}{2Nx_0}$
3. Transitions resulting from recombination for loci configurations ABC and BCA	
$(i_1 - 1, j_1 + 1)$	$\frac{i_1 R_{AB}(1 - x_0)}{2N}$
$(i_2 - 1, j_2 + 1)$	$\frac{i_2 R_{AB}(1 - x_0)}{2N}$
$(i_3 - 1, j_3 + 1)$	$\frac{i_3 R_{AB}(1 - x_0)}{2N}$
$(i_1 - 1, i_2 + 1, j_3 + 1)$	$\frac{i_1(R_{AC} - R_{AB})(1 - x_0)}{2N}$
$(i_1 - 1, i_2 + 1, i_3 + 1)$	$\frac{i_1(R_{AC} - R_{AB})(1 - x_0)}{2N}$

itions which result from recombination. The conditional probabilities (up to order $1/N$) of the 12 transitions involving genes linked to the A_1 allele are given in Table 1. The conditional probabilities of the other 12 transitions involving genes linked to the A_2 allele can be obtained from Table 1 by replacing i by j, β_1 by β_2 and x_0, the frequency of the A_1 allele, by $1 - x_0$.

It has already been pointed out that the effect of recombination is influenced by the relative positions of the three loci. In Table 2 are given the conditional probabilities (up to order $1/N$) of the 5 transitions involving genes linked to the A_1 allele which result from recombination, assuming the selected locus, A, lies between the two neutral loci.

For ease of notation let the 24 transitions be labeled from 1 to 24 in some specified order. Also, the conditional probability of the kth transition is denoted by $p_k/2N$, $1 \le k \le 24$. For any i, j let

$$h_{ij} = \sum_{k=1}^{24} p_k \quad \text{and} \quad q_{ij}(k) = \frac{p_k}{h_{ij}}, \qquad 1 \le k \le 24.$$

TABLE 2

Possible transitions from (i, j), involving genes linked to the A_1 allele, and their probabilities (up to order $1/N$): transitions resulting from recombination for loci configuration BAC

Transition	Probability
$(i_2 - 1, j_2 + 1)$	$\frac{i_2 R_{AB}(1 - x_0)}{2N}$
$(i_3 - 1, j_3 - 1)$	$\frac{i_3 R_{AC}(1 - x_0)}{2N}$
$(i_1 - 1, i_3 + 1, j_2 + 1)$	$\frac{i_1 R_{AB}(1 - x_0)}{2N}$
$(i_1 - 1, i_2 + 1, j_3 + 1)$	$\frac{i_1 R_{AC}(1 - x_0)}{2N}$
$(i_1 - 1, i_2 + 1, i_3 + 1)$	$\frac{i_1(R_{AB} + R_{AC})x_0}{2N}$

Finally, we are in a position to give the parameters of the Markov jump process. Indeed, the holding time in state (i, j) has a negative exponential distribution with parameter h_{ij} and when a jump does occur, the probability that it is the kth transition equals $q_{ij}(k)$, $1 \le k \le 24$.

Let T_B and T_C denote the sum of the lengths (measured in units of $2N$ generations) of the branches of the ancestral trees for locus B and C, respectively. The mean and variance of T_A and T_B can be calculated using the recursive equations of KAPLAN, DARDEN and HUDSON (1988). To compute the expectation of such quantities as $T_B T_C$ and $e^{-(\theta_B T_B + \theta_C T_C)/2}$, $\theta_B > 0$, $\theta_C > 0$, we need to use the coalescent process for two loci. The expectation of $T_B T_C$ is needed to compute the variance of the number of segregating sites at the two loci. If $\theta_B = 4N\mu_B$ and $\theta_C = 4N\mu_C$, where μ_B and μ_C are the neutral mutation rates at the B and C locus, respectively, then the expectation of $e^{-(\theta_B T_B + \theta_C T_C)/2}$ is the probability that there are no segregating sites at either of the two loci. This expectation is required for some of the calculations of the next section.

For any i, j let

$$\begin{aligned} M_B(i, j) &= E(T_B | Q(0) = (i, j)), \\ M_C(i, j) &= E(T_C | Q(0) = (i, j)), \\ M_{BC}(i, j) &= E(T_B T_C | Q(0) = (i, j)), \\ H_{BC}(i, j) &= E(e^{-(\theta_B T_B + \theta_C T_C)/2} | Q(0) = (i, j)) \end{aligned} \tag{2}$$

and

$$i_B = i_1 + i_2 + j_1 + j_2, \qquad i_C = i_1 + i_3 + j_1 + j_3.$$

It follows from the Markov structure of the coalescent process that

$$M_{BC}(i, j) = \frac{2i_B i_C}{h_{ij}^2} + E\left(\left(\frac{i_B M_C(Z_1) + i_C M_B(Z_1)}{h_{ij}}\right) + M_{BC}(Z_1) | Q(0) = (i, j)\right) \tag{3}$$

and

$$H_{BC}(i, j) = \frac{2h_{ij}E(H_{BC}(Z_1) \mid Q(0) = (i, j))}{i_B\theta_B + i_C\theta_C + 2h_{ij}}, \qquad (4)$$

where the distribution of Z_1 is given by the $\{q_{ij}(k), 1 \leq k \leq 24\}$.

AN APPLICATION

Recent studies of polymorphism at the alcohol dehydrogenase (*Adh*) locus of *Drosophila melanogaster* have suggested that natural selection maintains variation at this locus (*e.g.*, OAKESHOTT *et al.* 1982; KREITMAN and AGUADÉ 1986). HUDSON, KREITMAN and AGUADÉ (1987) have shown that the levels of polymorphism within *D. melanogaster* and the amount of divergence between *D. melanogaster* and *Drosophila sechelia* at the *Adh* locus and a flanking region are not compatible with an equilibrium neutral model of molecular evolution. They suggested that balancing selection acting on the Fast/Slow electrophoretic polymorphism (at codon 192) of the *Adh* gene might account for the high observed level of polymorphism of silent sites in the coding region. With the theory presented in the previous section we can begin to examine this hypothesis. Using the *Adh* sequence data of KREITMAN (1983) we can compare the observed level of variation at different points along the sequence to the level predicted by a model with balancing selection operating on the Fast/Slow polymorphism of the *Adh* gene.

The data of KREITMAN (1983) consists of the sequences of 11 cloned *D. melanogaster Adh* genes. In this sample of eleven sequences, 43 polymorphic nucleotide sites were observed, only one of which results in an amino-acid polymorphism. That amino acid polymorphism is responsible for the electrophoretic variants, Fast and Slow, commonly found in *D. melanogaster* populations. Six of the 11 sequences code for the Slow variant and will be referred to as Slow sequences. The other five sequences will be referred to as Fast sequences.

The goal of this section is to address the following questions: (1) If the Fast/Slow polymorphism of *Adh* is a balanced polymorphism such as that considered in the THEORY section, then what spatial distribution of neutral variation is expected in this region of the genome? and (2) How does the actual spatial distribution of variation in the *Adh* region compare with the theoretical prediction? To examine the spatial distribution of the polymorphic sites, a "sliding window" method is used. That is, at each nucleotide site, the amount of variation expected and observed is calculated for a small window centered on the nucleotide site.

Three different quantities were calculated to characterize the variability in a window at each nucleotide site. These were $\pi_{FS}(k)$, the average number of pairwise differences between Fast and Slow sequences in the window centered on nucleotide k, $\pi_{SS}(k)$, the average number of pairwise differences between Slow sequences in the window centered on nucleotide k, and $\pi_{FF}(k)$, the average number of pairwise differences between Fast sequences in the window centered on nucleotide k. Numbering the sequences from 1 to 11 and letting $d_{ij}(k)$ denote the number of nucleotide differences between sequence i and sequence j in the window centered on nucleotide k, the three measures of variability can be written as follows:

$$\pi_{FS}(k) = \frac{1}{n_F n_S} \sum_{i \in \mathbf{F}, j \in \mathbf{S}} d_{ij}(k),$$

$$\pi_{SS}(k) = \frac{1}{n_S(n_S - 1)} \sum_{i \neq j \in \mathbf{S}} d_{ij}(k),$$

and

$$\pi_{FF}(k) = \frac{1}{n_F(n_F - 1)} \sum_{i \neq j \in \mathbf{F}} d_{ij}(k), \qquad (5)$$

where n_F and n_S are the number of Fast sequences and Slow sequences, respectively, and **F** and **S** denote the set of Fast sequences and Slow sequences, respectively.

Since the region sequenced contains protein coding sequences as well as introns and other noncoding sequences, the level of constraint presumably varies considerably. To take at least partial account of the different levels of constraints in these regions, the size of the window was varied so as to keep the number of possible silent changes in the window constant. At noncoding sites and intron sites all changes are considered silent, and in the coding region a silent change is a change which does not affect the amino acid sequence. [This is equivalent to adjusting the window size so that the window contains a constant "effective" number of silent sites, as defined by KREITMAN (1983).]

In Figures 3, 4 and 5 the observed values of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ are plotted as a function of k for the Adh sequence data using a window size of 150 possible silent nucleotide changes. This window size corresponds to 50 base pairs in noncoding regions. Three interesting features of the data are the rather low values of $\pi_{FF}(k)$ in the coding region of the gene, and the the very high levels of $\pi_{FS}(k)$ in a small region encompassing the Fast/Slow polymorphism, and a somewhat smaller peak in values of $\pi_{SS}(k)$ in the same region.

If each nucleotide site is treated as an individual locus and if it is assumed that the allelic frequencies at position 2 of codon 192 are maintained by strong balancing selection, then the theory of the previous section can be used to calculate the expectation and variance of $\pi_{FF}(k)$, $\pi_{FS}(k)$ and $\pi_{SS}(k)$. These calculations require that values be assigned to the parameters, β_1, β_2, x_0. For each nucleotide site, we also require

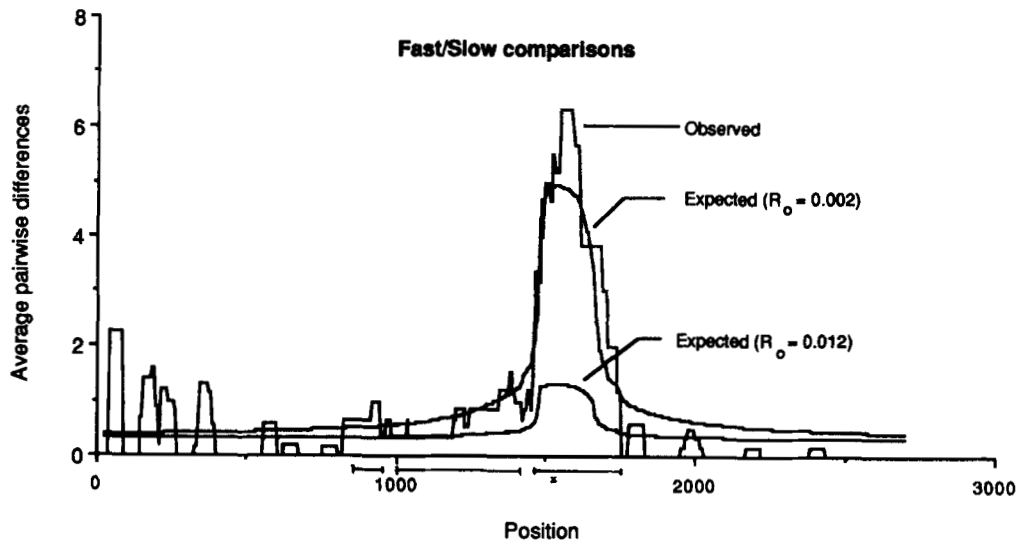

FIGURE 3.—The expected and observed number of differences between Fast and Slow sequences in a "sliding window," ($\pi_{FS}(k)$), plotted as a function of the nucleotide position, k. The coding exons are shown by the bold lines below the position axis of the plot. The Fast/Slow polymorphism is at position 1552. For these calculations, it was assumed that $\theta_0 = 0.006$, and $\beta_1 = \beta_2 = 0.001$. To obtain the top expected curve, it was assumed that $R_0 = 0.002$, and for the bottom expected curve, it was assumed that $R_0 = 0.012$. The width of the window was adjusted so that there were always 150 possible silent changes in the window. Thus, for example, in noncoding regions the window was 50 base pairs wide.

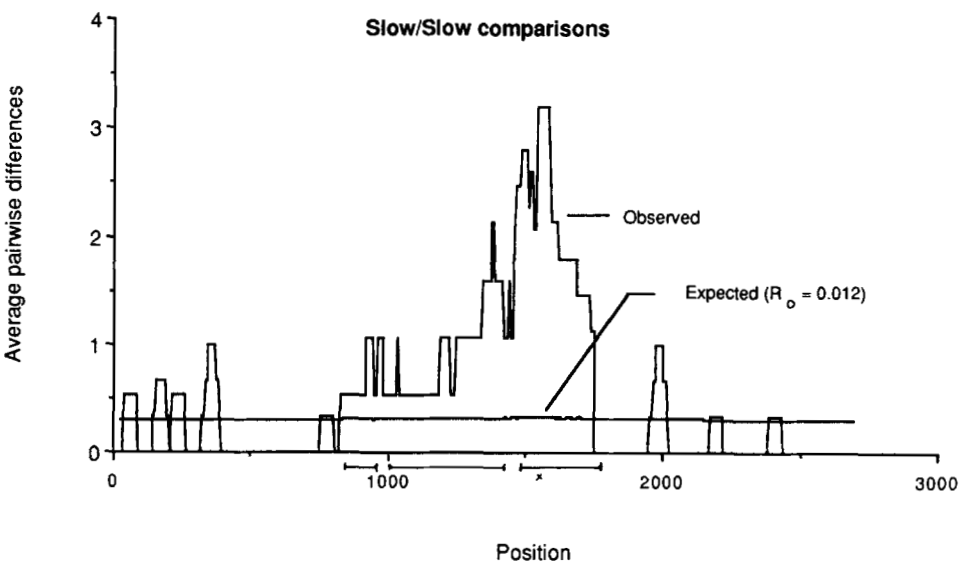

FIGURE 4.—The expected and observed number of differences between Slow sequences in a "sliding window," ($\pi_{SS}(k)$), plotted as a function of the nucleotide position, k. For these calculations, it was assumed that $R_0 = 0.012$. All other parameters were as in Figure 3.

the value of $\theta = 4N\mu$, where μ is the neutral mutation rate at the site, and for each pair of sites, we require the value of $R_{ij} = 2Nr_{ij}$, where r_{ij} is the recombination rate between site i and site j.

It is assumed that the neutral mutation rate is the same at all noncoding sites and all fourfold degenerate sites. (Fourfold degenerate sites are coding sites at which all four nucleotides result in the same amino acid. At such sites all mutations are silent, that is, not amino acid changing.) Denote $4N$ times this neutral mutation rate by θ_0. At sites where only one of the three possible mutations is a silent mutation, it is assumed that θ is $\theta_0/3$. Similarly, at sites where two of three possible changes are silent, θ is assumed to be $2\theta_0/3$. Estimates of θ have been obtained from restriction mapping studies of *Adh* and other loci in *D. melanogaster*. The heterozygosity per nucleotide (which is equivalent to θ when θ is small) has been estimated to be 0.006 for a region 13 kb long that includes the *Adh* locus (LANGLEY, MONTGOMERY and QUATTLEBAUM 1982; AQUADRO *et al.* 1986). KREITMAN and AGUADÉ (1986) estimated the heterozygosity per nucleotide to be 0.004 for a 4-kb region located just 5′ to the Adh coding region and 0.006 in the coding region. These estimates must be considered as average values over regions containing coding sequence, introns and other noncoding sequences. To calculate the expectations under the selective model we have assumed θ_0 equals 0.006.

We have assumed that only lysine and threonine are permissible at the selected site (codon 192) so only one of the three possible mutations at this site leads to the other selected allele. The mutations that change lysine to threonine and threonine back to lysine are the second position transversions $A \rightarrow C$ and $C \rightarrow A$, respectively. If noncoding sites and fourfold degenerate sites are on average only slightly constrained, then θ_0 equals approximately $4N$ times the spontaneous mutation rate. If mutations are equally likely to each of the other three nucleotides, then a plausible

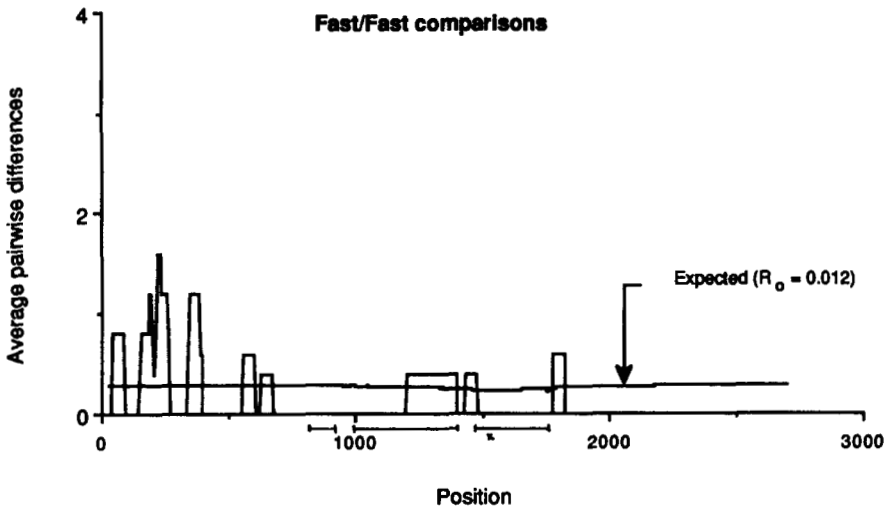

FIGURE 5.—The expected and observed number of differences between Fast sequences in a "sliding window," ($\pi_{FF}(k)$), plotted as a function of the nucleotide position, k. For these calculations, it was assumed that $R_0 = 0.012$. All other parameters were as in Figure 3.

value for β_1 and β_2 is $\beta_1 = \beta_2 = \theta_0/6 = 0.001$. (Recall that β_1 and β_2 equal $2N$ times the mutation rate to the other allele.) If considerable selective constraint operates on noncoding and silent sites then β_1 and β_2 would be somewhat larger than $\theta_0/6$. If the spontaneous mutations are strongly biased against transversions, then β_1 and β_2 should be somewhat smaller than $\theta_0/6$. For our calculations we have assumed $\beta_1 = \beta_2 = \theta_0/6 = 0.001$.

Since x_0, the frequency of the Slow variant, varies with geographic location (OAKESHOTT *et al.* 1982), it is not clear what value to assign to x_0. A more realistic model must take into account this geographic structure, but here we merely assume that x_0 is equal to 0.7, a value obtained for a sample of *D. melanogaster* from Raleigh, North Carolina (KREITMAN and AGUADÉ 1986).

Finally, we assume that $R_{ij} = |i - j|R_0$, that is, that recombination is uniform in the region. It remains only to assign a value to R_0. The value of this parameter is perhaps the most uncertain of all. Recombination per base pair has been estimated for several regions of the *D. melanogaster* genome to be approximately 10^{-8} per generation in females (CHOVNICK, GELBART and MCCARRON 1977). The neutral mutation rate has been estimated to be approximately 5×10^{-9} per year in many organisms. If we assume that *D. melanogaster* has 4 generations per year then the ratio of recombination per generation to the neutral mutation rate per generation is approximately $(10^{-8}/2)/(5 \times 10^{-9}/4) = 4$, where the factor of ½ in the numerator comes from the fact that recombination does not occur in males. This implies that R_0 is approximately 4 times $\theta_0/2$ or 0.012.

To calculate the expectation and variance of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ we assumed that the variation at each nucleotide site could be approximated by an infinite-allele model rather than a finite-allele model which would have been more realistic. This assumption is appropriate if θ is sufficiently small, so that with high probability at most one mutation can occur at a site in the history of the sample. The calculation of the expectation and variance of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ is described in the APPENDIX. The expectations are plotted in Figures 3, 4 and 5, together with the observed values. Two different values of R_0 were used, both the *a priori* guess, 0.012, and a smaller value, 0.002 chosen to produce a better fit to the observations. For R_0 equal to 0.012, the mean (and variance) of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ at the selected site are 1.3 (22.0), 0.33 (0.58), and 0.23 (0.83), respectively. For R_0 equal to 0.002, the mean (and variance) of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ at the selected site are 4.9 (101), 0.32 (0.47), and 0.21 (0.84), respectively.

Note that a remarkably good fit is obtainable, but that the recombination parameter required for this fit is about one-sixth of our *a priori* guess. The large uncertainty of the true value of R_0 and the high variance of $\pi_{FS}(k)$ near the selected site make the significance of this discrepancy difficult to interpret. The mean and variance of $\pi_{FS}(k)$ at the selected site for a completely neutral model with the same mutation parameter (and ignoring the Fast/Slow polymorphism) are 0.30 and 0.13 for R_0 equal to 0.012, and 0.30 and 0.14 for R_0 equal to 0.002.

The peak in the plot of observed values of $\pi_{SS}(k)$ near the selected site was not expected under the balancing selection model (Figure 4). Adjusting the parameter R_0 has a relatively small effect on the expectation of $\pi_{SS}(k)$ and no value of this parameter results in a peak such as that observed. This high variability among Slow alleles of *Adh* was noted before (KREITMAN and AGUADÉ 1986; HUDSON, KREITMAN and AGUADÉ 1987). It is possible that the presence of a second, and in this case silent, selected polymorphism would account for this high variability among Slow alleles. The possibility that this peak is not a statistically significant departure from the expectation

under the balancing selection model also needs consideration.

DISCUSSION

In the companion study (KAPLAN, DARDEN and HUDSON 1988), the coalescent process was studied for a locus completely linked to a locus at which selection operates. In this paper, the more general situation is considered in which a neutral locus is not completely linked to the locus at which selection operates. In this case, it is found that the coalescent process is similar in form to the process without recombination. The equations describing the process for the model with recombination can be obtained from those of the no-recombination model by replacing the mutation parameters (β_1 and β_2) with simple functions ($\beta_1(x)$ and $\beta_2(x)$ defined by Equation 1) that depend on the mutation rates, the recombination rates, the population size and the frequency of the A_1 allele at the selected locus.

STROBECK (1983) has studied the linkage disequilibrium and homozygosity at a neutral locus linked to a chromosomal arrangement that is maintained in the population by strong selection. The quantities $H_m(2, 0)$, $H_m(0, 2)$, and $H_m(1, 1)$, defined in the APPENDIX, are equivalent to STROBECK's identity coefficients Φ_{11}, Φ_{22}, Φ_{12}. If β_1 and β_2 equal zero, then $H_m(2, 0)$, $H_m(0, 2)$, and $H_m(1, 1)$ are identical to STROBECK's Φ_{11}, Φ_{22}, Φ_{12} as given by his Equation 7.

Under the no-recombination model with tight regulation of the frequencies of the alleles at the selected locus, it was shown by KAPLAN, DARDEN and HUDSON (1988) that the moments of T, the sum of the lengths of the branches of the ancestral tree, do not differ from their values under the neutral model, unless β_1 and β_2 are small. It follows that, for a site not completely linked to the selected site, the moments of T are approximately the same as they are under the neutral model, unless the mutation parameters and the quantities, Rx_0 and $R(1 - x_0)$, are small, say less than one. (Recall that R is $2N$ times the recombination rate between the neutral site and the site where selection operates.) Only sites very tightly linked to the site at which selection operates are expected to have significantly larger ancestral trees than are expected under the neutral model.

The physical size of the region in which neutral variation is significantly elevated depends on the population size and the rate of recombination per base pair. For what appear to be reasonable estimates of these parameters in Drosophila, the size of this region is quite small, on the order of a few hundred base pairs. This means that one needs high resolution techniques, such as sequencing or four-cutter restriction mapping to detect such narrow regions of high variability. Six-cutter surveys may not give sufficient resolution to detect such regions. In species with lower population size the expected length of the region with increased heterozygosity is larger, but with smaller population size the amount of nucleotide variation is also expected to be smaller so high resolution techniques may still be required.

HUDSON, KREITMAN and AGUADÉ (1987) previously found that the variation in the *Adh* region of *D. melanogaster* in conjunction with between species divergence data were incompatible with the neutral model. As shown in Figures 3, 4 and 5, a model in which balancing selection maintains the Fast/Slow amino acid polymorphism of *Adh*, predicts many aspects of the within species variation in the *Adh* region of *D. melanogaster*. For example, the balancing selection model predicts that comparisons between Slow and Fast sequences will show a narrow region of high variability. The predicted location and width of this region of high variability are similar to the location and width of the observed region of high variability in *D. melanogaster*.

Other aspects of the data do not match the predictions so well. With an *a priori* estimate of the recombination parameter, the predicted magnitude of the increase in variability is lower than observed. However, the predicted variance of this magnitude is large, so that the observed peak in variation is not incompatible with the *a priori* estimate of the recombination parameter. There is also great uncertainty in the *a priori* estimate of the recombination parameter. With lower values of the recombination parameter, the match between predicted and observed is good. Comparisons of different sequences bearing the Slow allele, show unexpected high levels of variability in the region of the Fast/Slow polymorphism. This may not be a significantly high level of variability, since under the selection model, the variance of the level of variability is high near the site of the balanced polymorphism. An alternative explanation is that selection is maintaining variation at both the Fast/Slow site and an additional site located near to the Fast/Slow site.

There are several areas for further theoretical research. We have made simplifying assumptions concerning the neutral mutation rates at noncoding sites, intron sites, and at coding sites of different degeneracy. It may be important to make more realistic assumptions about these neutral mutation rates, perhaps by incorporating between species divergence information into the analysis. We have assumed that selection coefficients remain constant, but certainly it is likely that selection coefficients vary in time and space. Properties of such models merit consideration. The effects of partial isolation of subpopulations on the coalescent process also need investigation. The analysis of models with more than one locus with selection may also be useful, particularly for the interpretation of the *Adh* data. Finally, statistical hypothesis tests are needed to distinguish between the neutral model and alternative selective models. It is our expectation that

the coalescent process for selective models will be useful in developing such tests.

LITERATURE CITED

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

CHOVNICK, A., W. GELBART and M. MCCARRON, 1977 Organization of the Rosy locus in *Drosophila melanogaster*. Cell **11:** 1–10.

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

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

KAPLAN, N. L., and R. R. HUDSON, 1985 The use of sample genealogies for studying a selectively neutral M-loci model with recombination. Theor. Popul. Biol. **28:** 382–396.

KARLIN, S., and H. M. TAYLOR, 1981 *A Second Course in Stochastic Processes*. Academic Press, New York.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KREITMAN, M., and M. AGUADÉ, 1986 Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics **114:** 93–110.

LANGLEY, C. H., E. A. MONTGOMERY and W. F. QUATTLEBAUM, 1982 Restriction map variation in the Adh region of Drosophila. Proc. Natl. Acad. Sci. USA **79:** 5631–5635.

OAKESHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON and G. K. CHAMBERS, 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on three continents. Evolution **36:** 86–96.

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

Communicating editor: B. S. WEIR

APPENDIX

In this appendix, a method is outlined for calculating the mean and variance of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$, which are defined by Equation 5. It is clear from the definitions, that

$$E(\pi_{FS}(k)) = E(d_{ij}(k))$$

$$= E\left(\sum_{m\in W_k} (1 - \delta_{ij}(m))\right)$$

where sequence i is a Slow sequence and sequence j is a Fast sequence, $\delta_{ij}(m)$ is one if site m is the same in sequence i and sequence j, and zero otherwise, and W_k is the set of sites in the window centered at k. The expection of $\delta_{ij}(m)$ is the probability of identity of site m in sequence i (a Slow sequence) and sequence j (a Fast sequence), which we will denote by $H_m(1, 1)$. Thus, it follows that

$$E(\pi_{FS}(k)) = \sum_{m\in W_k} (1 - H_m(1, 1)).$$

Similarly,

$$E(\pi_{FF}(k)) = \sum_{m\in W_k} (1 - H_m(0, 2))$$

and

$$E(\pi_{SS}(k)) = \sum_{m\in W_k} (1 - H_m(2, 0)),$$

where $H_m(0, 2)$ is the probability of identity at site m in two Fast sequences and $H_m(2, 0)$ is the probability of identity at site m in two Slow sequences.

If the variation at each site m, is described by a neutral infinite-allele model, then

$$H_m(1, 1) = E(e^{-\theta_m T_m/2}),$$

where T_m is twice the time in units of $2N$ generations to the common ancestor of a Fast and a Slow sequence at site m, and the expectation is over the distribution of T_m. Similar expressions exists for $H_m(0, 2)$ and $H_m(2, 0)$. In the case of a tightly regulated equilibrium, such as that considered in KAPLAN, DARDEN and HUDSON (1988) these identity coefficients, $H_m(1, 1)$, $H_m(0, 2)$ and $H_m(2, 0)$, satisfy the following system of three linear equations analogous to (4):

$$H_m(1, 1) = \frac{2h_{11}(x_0)}{2\theta + 2h_{11}(x_0)} \cdot \left[\frac{\beta_1(x_0)x_0}{(1 - x_0)h_{11}(x_0)} H_m(2, 0) + \frac{\beta_2(x_0)(1 - x_0)}{x_0 h_{11}(x_0)} H_m(0, 2)\right]$$

$$H_m(0, 2) = \frac{2h_{02}(x_0)}{2\theta + 2h_{02}(x_0)} \cdot \left[\frac{2\beta_1(x_0)x_0}{(1 - x_0)h_{02}(x_0)} H_m(1, 1) + \frac{1}{1 - x_0}\right]$$

$$H_m(2, 0) = \frac{2h_{20}(x_0)}{2\theta + 2h_{20}(x_0)} \cdot \left[\frac{2\beta_2(x_0)(1 - x_0)}{x_0 h_{20}(x_0)} H_m(1, 1) + \frac{1}{x_0}\right]$$

where

$$h_{11}(x) = \frac{\beta_1(x)x}{1 - x} + \frac{\beta_2(x)(1 - x)}{x},$$

$$h_{02}(x) = \frac{2\beta_1(x)x}{1 - x} + \frac{1}{1 - x},$$

$$h_{20}(x) = \frac{1}{x} + \frac{2\beta_2(x)(1 - x)}{x},$$

and $\beta_1(x)$ and $\beta_2(x)$ are given by Equation 1.

The variances of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ are more complicated to calculate. To see this, we consider the

calculation of $E(\pi_{FS}(k)^2)$. From the definition of $\pi_{fS}(k)$,

$$E(\pi_{FS}(k)^2) = \frac{1}{(n_F n_S)^2} E\left[\left(\sum_{i\in F, j\in S} d_{ij}(k)\right)^2\right].$$

By expanding the squared term on the right hand side, we get

$$E\left[\left(\sum_{i\in F, j\in S} d_{ij}(k)\right)^2\right] = E\left[\sum_{i,j} d_{ij}^2(k) + \sum_{i\neq i', j\neq j'} d_{ij}(k)d_{i'j'}(k) + \sum_{i,j\neq j'} d_{ij}(k)d_{ij'}(k) + \sum_{i\neq i', j} d_{ij}(k)d_{i'j}(k)\right]$$

in which i and i' refer to Fast sequences and j and j' refer to Slow sequences. The expectations on the right hand side can be expressed in terms of identity coefficients for samples of size two, three, and four for one or two sites. These identity coefficients satisfy systems of linear equations such as (4) and can be evaluated numerically, although in some cases the number of variables is as large as 20.