数理解析研究所講究録
第762巻 1991年 7-18

Genealogical Process in Models with Migration and Selection

Masaru Iizuka （飯塚 勝）
General Education Course, Chikushi Jogakuen Junior College
Ishizaka 2-12-1, Dazaifu-shi, Fukuoka-ken 818-01, Japan

Summary

Statistical properties of the process describing the genealogical history of a random sample of genes are obtained for two population genetics models with population subdivision and migration. The first model is that without selection. The second model is that with selection, recombination and neutral mutation. The calculations are greatly simplified if migration is strong in the second model.

§1. Introduction

Restriction map data and DNA sequence data for a sample of genes from a population provide information about genetic variability at nucleotide level. An important summary statistic of these type of data is S, the number of segregating sites in the samlpe. Note that $S = S_{sel} + S_{neu}$, where S_{sel} and S_{neu} is the number of segregating sites resulting from mutations with and without selective effect, respectively. The distribution of S_{sel} is difficult to derive. The distribution of S_{neu}, on the other hand, is analytically more tractable. If S_{sel} is negligible compared to S_{neu}, e.g. selection acts on a few nucleotide sites, then the statistical properties of S can be inferred from those of S_{neu}. The distribution of S_{neu} can be represented as

$$(1) \quad P(S_{neu} = k) = \int_0^\infty \frac{(\mu t)^k}{k!} e^{-\mu t} F(dt),$$

where μ is the rate of neutral mutation per gene per generation,

$F(t) = P(T \leq t)$, $t \geq 0$ and T is the sum of the lengths (mesured in generations) of all the branches of the ancestral tree describing the genealogy of the sample. It follows from (1) that

(2) $E[S_{neu}] = \mu E[T]$,

(3) $Var[S_{neu}] = \mu E[T] + \mu^2 Var[T]$,

(4) $P(S_{neu} = 0) = E[e^{-\mu T}]$.

For selectively neutral models the stochastic process describing the genealogical history of the sample has been characterized (Watterson 1975; Kingman 1982a,b; Tavare 1984), and properties of the distribution of T have been determined.

Since many natural populations are geographically subdivided, it is important to analyse population genetic models, focusing on the genealogical process of samples of genes from a subdivided population. In the following sections, a neutral model with population subdivision and a selection model with population subdivision are analysed.

§2. A neutral model with population subdivision

Suppose a randomly mating diploid population of size N has two subpopulations, subpopulation 1 of size $N_1 = fN$ and subpopulation 2 of size $N_2 = (1-f)N$, $0<f<1$. Let m_i be the migration rate from subpopulation i to the other subpopulation per generation ($i=1,2$). Assume that $m_i = \lambda_i/(2N) + o(1/N)$ ($i=1,2$). Since the ancestral genes can be located in subpopulation 1 and 2, the genealogical process is a two dimensional process. Suppose that n neutral genes are chosen at random from the $0th$ generation and let $Q(0) = (i, j)$ if the sample consists of i genes from subpopulation 1 and j genes from subpopulation 2 ($0 \leq i, j \leq n$, $i + j = n$). For $t<0$, $Q(t)$ denotes the number of the ancestral genes of the sample located in subpopulation 1 and 2 in generation t. Neglecting the

quantities higher than $1/N$, the process, $Q(t)$, $t \leq 0$, is the following Markov chain.

$$(5) \quad P(Q(t-1)=(i-1,j)\,|\,Q(t)=(i,j)) = \binom{i}{2}/(2fN),$$

$$(6) \quad P(Q(t-1)=(i,j-1)\,|\,Q(t)=(i,j)) = \binom{j}{2}/\{2(1-f)N\},$$

$$(7) \quad P(Q(t-1)=(i-1,j+1)\,|\,Q(i,j)) = i\lambda_2/(2N),$$

$$(8) \quad P(Q(t-1)=(i+1,j-1)\,|\,Q(i,j)) = j\lambda_1/(2N),$$

$$(9) \quad P(Q(t-1)=(i,j)\,|\,Q(i,j)) = 1 - \{\binom{i}{2}/f + \binom{j}{2}/(1-f) + i\lambda_2 + j\lambda_1\}/(2N),$$

where $\binom{i}{2} = i!/\{(i-2)!2!\}$, $(i \geq 2)$, $= 0$, $(i<2)$.

Let $T(i,j)$ be the sum of the lengths (measured in units of $2N$ generations) of all the branches of the ancestral tree, assuming that the sample consisted of i genes located in population 1 and j genes located in population 2 $(i + j \geq 2)$. As a direct consequence of the Markov property of $Q(t)$ process, we have

$$(10) \quad T(i,j) = (i + j)T_{ij} + T(Z_{ij}),$$

where T_{ij} is the holding time in state (i,j), Z_{ij} is the random state to which the process moves and if $Z_{ij} = (k,l)$, then $T(Z_{ij})$ is an independent random variable having the same distribution as $T(k,l)$. Let M_{ij} denote the mean of $T(i,j)$ and V_{ij} its variance. Then

$$(11) \quad M_{ij} = (i+j)/h_{ij} + q_{i-1,j}M_{i-1,j} + q_{i,j-1}M_{i,j-1} + q_{i-1,j+1}M_{i-1,j+1} + q_{i+1,j-1}M_{i+1,j-1},$$

$$(12) \quad V_{ij} = \{(i+j)/h_{ij}\}^2 + q_{i-1,j}V_{i-1,j} + q_{i,j-1}V_{i,j-1} + q_{i-1,j+1}V_{i-1,j+1} + q_{i+1,j-1}V_{i+1,j-1} + q_{i-1,j}M_{i-1,j}^2 + q_{i,j-1}M_{i,j-1}^2 + q_{i-1,j+1}M_{i-1,j+1}^2 + q_{i+1,j-1}M_{i+1,j-1}^2 - (q_{i-1,j}M_{i-1,j} + q_{i,j-1}M_{i,j-1} + q_{i-1,j+1}M_{i-1,j+1} + q_{i+1,j-1}M_{i+1,j-1})^2,$$

where

$$(13) \quad h_{ij} = \binom{i}{2}/f + \binom{j}{2}/(1-f) + i\lambda_2 + j\lambda_1,$$

$$(14) \quad q_{i-1,j} = \binom{i}{2}/\{fh_{ij}\},$$

$$(15) \quad q_{i,j-1} = \binom{j}{2}/\{(1-f)h_{ij}\},$$

$$(16) \quad q_{i-1,j+1} = i\lambda_2/h_{ij},$$

$$(17) \quad q_{i+1,j-1} = j\lambda_1/h_{ij},$$

and $M_{01} = M_{10} = V_{01} = V_{10} = 0$.

In the following in this section, we consider the case of $n = 2$. In this case, we have

$$(18) \quad M_{20} = 2f(1 + \lambda_2 M_{11})/(1 + 2\lambda_2 f),$$

$$(19) \quad M_{02} = 2(1-f)(1 + \lambda_1 M_{11})/\{1 + 2\lambda_1(1-f)\},$$

$$(20) \quad M_{11} = 2\{1 + \lambda_1 f/(1 + 2\lambda_2 f) + \lambda_2(1-f)/[1 + 2\lambda_1(1-f)]\}/\{\lambda_1/(1 + 2\lambda_2 f) + \lambda_2/[1 + 2\lambda_1(1-f)]\}.$$

As the strong migration limit ($\lambda_1, \lambda_2 \to +\infty$), we have $M_{20}, M_{02}, M_{11} \sim 2(\lambda_1 + \lambda_2)^2 f(1-f)/\{\lambda_1^2(1-f) + \lambda_2^2 f\}$. As the weak migration limit ($\lambda_1, \lambda_2 \to 0$), we have $M_{20} \sim 2f\{1 + 2\lambda_2/(\lambda_1 + \lambda_2)\}$, $M_{02} \sim 2(1-f)\{1 + 2\lambda_1/(\lambda_1 + \lambda_2)\}$, $M_{11} \sim +\infty$. Note that in the case of complete isolation ($\lambda_1 = \lambda_2 = 0$), we have $M_{20} = 2f$, $M_{02} = 2(1-f)$ and $M_{11} = +\infty$. Here, we consider the symmetric case ($f = 1/2$ and $\lambda = \lambda_1 = \lambda_2$). For $\lambda > 0$, we have $M_{20} = M_{02} = 2$, $M_{11} = 2 + 1/\lambda$, $V_{20} = V_{02} = 2(2 + 1/\lambda)$ and $V_{11} = 2(2 + 1/\lambda) + 1/\lambda^2$. For $\lambda = 0$, we have $M_{20} = M_{02} = 1$, $M_{11} = +\infty$, $V_{20} = V_{02} = 1$ and $V_{11} = +\infty$.

Slatkin (1987) and Strobeck (1987) found for the infinite site model (Kimura 1969) that the average number of sites that differ in two genes

chosen from the same subpopulation depends on the size of the total population and not on either the migration rate or the size of each subpopulation. The number of subpopulations is m (≥ 2) in the both studies and some conditions on the migration are assumed. Here, we consider the case of $m = 2$ and derive the results of Slatkin (1987) and Strobeck (1987) by means of the genealogical process approach, that is, by means of the results that we have obtained above in this section. First consider the case of Slatkin (1987), where the assumption on the migration is $\lambda = \lambda_1 = \lambda_2$ ($m_1 = m_2$). Let N_a be the harmonic mean of $N_1 = Nf$ and $N_2 = N(1-f)$.

$$(21) \quad 1/N_a = (1/N_1 + 1/N_2)/2,$$

that is, $N_a = 2f(1-f)N$. Averaging M_{20} and M_{02} by $1/N_1$ and $1/N_2$, we have $M_0 = N_a(M_{20}/N_1 + M_{02}/N_2)/2 = 4N_a/N$. Let $K_0 = 2NuM_0$, where u is the mutation rate for each site per generation. We have $K_0 = 8N_a u$, which is the same result as Equation (17) in Slatkin (1987) in the case of $m = 2$. Next consider the results of Strobeck (1987), where a conservative migration is assumed. In the case of $m = 2$, this assumption is $m_1 N_1 = m_2 N_2$ which is equivalent to $\lambda_1 f = \lambda_2(1-f)$. Averaging M_{20} and M_{02} by N_1 and N_2, we have $M = \{N_1/(N_1 + N_2)\}M_{20} + \{N_2/(N_1 + N_2)\}M_{02} = 2$. Let $\bar{\xi}_{ii} = 2NuM$. We have $\bar{\xi}_{ii} = 4Nu$, which is the same result as Strobeck (1987) in the case of $m = 2$.

For further results on genealogical processes with population subdivision and no selection, see Takahata and Nei (1985), Takahata (1988) and Tajima (1989).

§3. A selection model with population subdivision

Suppose a randomly mating diploid population of size N has two subpopulations, subpopulation 1 of size $N_1 = Nf$ and subpopulation 2 of size $N_2 = N(1-f)$, $0<f<1$. Let m_i be the migration rate from subpopulation

i to the other subpopulation per generation (i=1,2). Consider a selected locus with two alleles, A_1 and A_2. Let u_i be the mutation rate from A_i to the other allele per generation (i=1,2). The relative fitness of A_iA_j is denoted by w_{ij} (i,j=1,2). We denote the gene frequency of A_1 in subpopulation i at generation t by $x_i(t)$ (i=1,2) and put $x(t) = (x_1(t),x_2(t))$. We are interested in a neutral gene linked to the selected locus. Let r be the recombination rate between this neutral gene and the selected locus. We assume that $m_i = \lambda_i/(2N) + o(1/N)$, $u_i = v_i/(2N) + o(1/N)$, $w_{ij} = 1 + O(1/N)$ and $r = R/(2N) + o(1/N)$, where λ_i, $v_i > 0$ and $R \geq 0$ (i,j=1,2). Since the ancestral genes can be linked to A_1 and A_2 and located in subpopulation 1 and 2, the genealogical process is a four dimensional process. Suppose that n neutral genes are chosen at random from the $0th$ generation and let $Q(0) = (i_1,j_1,i_2,j_2)$ if the sample consists of i_k genes from subpopulation k and linked to A_1 and j_k genes from subpopulation k and linked to A_2 ($0 \leq i_k,j_k \leq n$, k = 1,2, $i_1+j_1+i_2+j_2=n$). For $t<0$, $Q(t)$ denotes the number of the ancestral genes of the sample located in subpopulation 1 and 2 and linked to A_1 and A_2 in generation t. Neglecting the quantities higher than $1/N$ and given $x(t)$, $t \leq 0$, the process, $Q(t)$, $t \leq 0$, is the following Markov chain (Kaplan *et al.* 1988; Hudson and Kaplan 1988).

$$(22)\quad p(i_1-1,j_1,i_2,j_2) = \binom{i}{2}1/\{2Nfx_1(t-1)\},$$

$$(23)\quad p(i_1,j_1-1,i_2,j_2) = \binom{j}{2}1/\{2Nf[1-x_1(t-1)]\},$$

$$(24)\quad p(i_1,j_1,i_2-1,j_2) = \binom{i}{2}2/\{2N(1-f)x_2(t-1)\},$$

$$(25)\quad p(i_1,j_1,i_2,j_2-1) = \binom{j}{2}2/\{2N(1-f)[1-x_2(t-1)]\},$$

$$(26)\quad p(i_1+1,j_1-1,i_2,j_2) = j_1\psi_{11}(t-1)x_1(t-1)/\{2N[1-x_1(t-1)]\},$$

$$(27)\quad p(i_1-1,j_1+1,i_2,j_2) = i_1\psi_{21}(t-1)[1-x_1(t-1)]/\{2Nx_1(t-1)\},$$

$$(28)\quad p(i_1,j_1,i_2+1,j_2-1) = j_2\psi_{12}(t-1)x_2(t-1)/\{2N[1-x_2(t-1)]\},$$

$$(29)\quad p(i_1,j_1,i_2-1,j_2+1) = i_2\psi_{22}(t-1)[1-x_2(t-1)]/\{2Nx_2(t-1)\},$$

$$(30)\quad p(i_1+1,j_1,i_2-1,j_2) = i_2\lambda_1 x_1(t-1)/\{2Nx_2(t-1)\},$$

$$(31)\quad p(i_1-1,j_1,i_2+1,j_2) = i_1\lambda_2 x_2(t-1)/\{2Nx_2(t-1)\},$$

$$(32)\quad p(i_1,j_1+1,i_2,j_2-1) = j_2\lambda_1[1-x_1(t-1)]/\{2N[1-x_2(t-1)]\},$$

$$(33)\quad p(i_1,j_1-1,i_2,j_2+1) = j_1\lambda_2[1-x_2(t-1)]/\{2N[1-x_1(t-1)]\},$$

$$\begin{aligned}(34)\quad p(i_1,j_1,i_2,j_2) = 1 - \{&p(i_1-1,j_1,i_2,j_2) + p(i_1,j_1-1,i_2,j_2)\\ &+ p(i_1,j_1,i_2-1,j_2) + p(i_1,j_1,i_2,j_2-1) + p(i_1+1,j_1-1,i_2,j_2)\\ &+ p(i_1-1,j_1+1,i_2,j_2) + p(i_1,j_1,i_2+1,j_2-1) + p(i_1,j_1,i_2-1,j_2+1)\\ &+ p(i_1+1,j_1,i_2-1,j_2) + p(i_1-1,j_2,i_2+1,j_2) + p(i_1,j_1+1,i_2,j_2-1)\\ &+ p(i_1,j_1-1,i_2,j_2+1)\},\end{aligned}$$

where we put $\psi_{1k}(t) = v_1 + R[1 - x_k(t)]$, $\psi_{2k}(t) = v_2 + Rx_k(t)$, k=1,2, and

$$(35)\quad p(h_1,k_1,h_2,k_2) = P(Q(t-1)=(h_1,k_1,h_2,k_2)\,|\,Q(t)=(i_1,j_1,i_2,j_2),x(t-1)).$$

Transitions resulting from coalescence are given by (22)~(25). Those resulting from mutation and recombination are given by (26)~(29). Those resulting from migration are given by (30)~(33).

In the following, we consider the case where the frequencies of alleles in the selected locus are tightly regulated (Kaplan *et al.* 1988). In other words, selection is so strong that the frequencies of alleles in the selected locus can be regarded as constats; $x(t) = x = (x_1,x_2)$ where x_1 and x_2 are constants ($0\le x_1,x_2\le 1$) for all t.

Let $T(i_1,j_1,i_2,j_2)$ be the sum of lengths (measured in units of $2N$ generations) of all the branches of the ancestral tree, assuming that the sample consisted of i_k genes that are linked to A_1 and located in subpopulation k and j_k genes that are linked to A_2 and located in

subpopulation k (k=1,2, $i_1+j_1+i_2+j_2 \geq 2$). As a direct consequence of the Markov property of $Q(t)$ process, we have

$$(36) \quad T(i_1,j_1,i_2,j_2) = (i_1+j_1+i_2+j_2)T_{i_1j_1i_2j_2} + T(Z_{i_1j_1i_2j_2}),$$

where $T_{i_1j_1i_2j_2}$ is the holding time in state (i_1,j_1,i_2,j_2), $Z_{i_1j_1i_2j_2}$ is the random state to which the process moves and if $Z_{i_1j_1i_2j_2} = (h_1,k_1,h_2,k_2)$ then $T(Z_{i_1j_1i_2j_2})$ is an independent random variable having the same distribution as $T(h_1,k_1,h_2,k_2)$. Let $M_{i_1j_1i_2j_2}$ denote the mean of $T(i_1,j_1,i_2,j_2)$. Then

$$\begin{aligned}(37) \quad M_{i_1j_1i_2j_2} &= (i_1+j_1+i_2+j_2)h_{i_1j_1i_2j_2}(x) + q_{i_1-1,j_1i_2j_2}(x)M_{i_1-1,j_1i_2j_2} \\ &+ q_{i_1,j_1-1,i_2j_2}(x)M_{i_1,j_1-1,i_2j_2} + q_{i_1j_1,i_2-1,j_2}(x)M_{i_1j_1,i_2-1,j_2} \\ &+ q_{i_1j_1i_2,j_2-1}(x)M_{i_1j_1i_2,j_2-1} + q_{i_1+1,j_1-1,i_2j_2}(x)M_{i_1+1,j_1-1,i_2j_2} \\ &+ q_{i_1-1,j_1+1,i_2j_2}(x)M_{i_1-1,j_1+1,i_2j_2} \\ &+ q_{i_1j_1,i_2+1,j_2-1}(x)M_{i_1j_1,i_2+1,j_2-1} \\ &+ q_{i_1j_1,i_2-1,j_2+1}(x)M_{i_1j_1,i_2-1,j_2+1} \\ &+ q_{i_1+1,j_1,i_2-1,j_2}(x)M_{i_1+1,j_1,i_2-1,j_2} \\ &+ q_{i_1-1,j_1,i_2+1,j_2}(x)M_{i_1-1,j_1,i_2+1,j_2} \\ &+ q_{i_1,j_1+1,i_2,j_2-1}(x)M_{i_1,j_1+1,i_2,j_2-1} \\ &+ q_{i_1,j_1-1,i_2,j_2+1}(x)M_{i_1,j_1-1,i_2,j_2+1},\end{aligned}$$

where

$$(38) \quad h_{i_1j_1i_2j_2}(x) = \binom{i}{2}1/(fx_1) + \binom{j}{2}1/\{f(1-x_1)\} + \binom{i}{2}2/\{(1-f)x_2\}$$

$$+ \binom{j}{2}2/\{(1-f)(1-x_2)\} + j_1\psi_{11}(x)x_1/(1-x_1) + i_1\psi_{21}(x)(1-x_1)/x_1$$
$$+ j_2\psi_{12}(x)x_2/(1-x_2) + i_2\psi_{22}(x)(1-x_2)/x_2 + i_2\lambda_1x_1/x_2 + i_1\lambda_2x_2/x_1$$
$$+ j_2\lambda_1(1-x_1)/(1-x_2) + j_1\lambda_2(1-x_2)/(1-x_1),$$

(39) $$q_{i_1-1,j_1i_2j_2}(x) = \binom{i}{2}1/\{fx_1h_{i_1j_1i_2j_2}(x)\},$$

(40) $$q_{i_1,j_1-1,i_2j_2}(x) = \binom{j}{2}1/\{f(1-x_1)h_{i_1j_1i_2j_2}(x)\},$$

(41) $$q_{i_1j_i,i_2-1,j_2}(x) = \binom{i}{2}2/\{(1-f)x_2h_{i_1j_1i_2j_2}(x)\},$$

(42) $$q_{i_1j_1i_2,j_2-1}(x) = \binom{j}{2}2/\{(1-f)(1-x_2)h_{i_1j_1i_2j_2}(x)\},$$

(43) $$q_{i_1+1,j_1-1,i_2j_2}(x) = j_1\psi_{11}(x)x_1/\{(1-x_1)h_{i_1j_1i_2j_2}(x)\},$$

(44) $$q_{i_1-1,j_1+1,i_2j_2}(x) = i_1\psi_{21}(x)(1-x_1)/\{x_1h_{i_1j_1i_2j_2}(x)\},$$

(45) $$q_{i_1j_1,i_2+1,j_2-1}(x) = j_2\psi_{12}(x)x_2/\{(1-x_2)h_{i_1j_1i_2j_2}(x)\},$$

(46) $$q_{i_1j_1,i_2-1,j_2+1}(x) = i_2\psi_{22}(x)(1-x_2)/\{x_2h_{i_1j_1i_2j_2}(x)\},$$

(47) $$q_{i_1+1,j_1,i_2-1,j_2}(x) = i_2\lambda_1x_1/\{x_2h_{i_1j_1i_2j_2}(x)\},$$

(48) $$q_{i_1-1,j_1,i_2+1,j_2}(x) = i_1\lambda_2x_2/\{x_1h_{i_1j_1i_2j_2}(x)\},$$

(49) $$q_{i_1,j_1+1,i_2,j_2-1}(x) = j_2\lambda_1(1-x_1)/\{(1-x_2)h_{i_1j_1i_2j_2}(x)\},$$

(50) $$q_{i_1,j_1-1,i_2,j_2+1}(x) = j_1\lambda_2(1-x_2)/\{(1-x_1)h_{i_1j_1i_2j_2}(x)\},$$

$\psi_{1k}(x) = \nu_1 + R(1-x_k)$, $\psi_{2k}(x) = \nu_2 + Rx_k$ $(k=1,2)$ and
$M_{1000} = M_{0100} = M_{0010} = M_{0001} = 0$.

In the following, we consider the case of $n = 2$ $(i_1+j_1+i_2+j_2=2)$ for simplicity. In this case, we have ten linear equations for M_{2000}, M_{0200}, M_{0020}, M_{0002}, M_{1100}, M_{1010}, M_{1001}, M_{0110}, M_{0101} and M_{0011}, which can be solved analytically or numerically. The expression of the solution is very complicated. In the case of strong migration, however, the structure

of the Markov chain is simplified considerably after some approximation. We will mention briefly this asymptotic analysis.

In the case of strong migration, we are interested in the linkage of the ancestral genes to the selected locus but are not interested in their location. For this reason we introduce the following stochastic process, $W(t)$, $t\leq 0$.

(51) $W(t) = [i_1+i_2, j_1+j_2]$ if $Q(t) = (i_1, j_1, i_2, j_2)$.

Note that $W(t)$ is not a Markov chain. For example, this can be easily seen by the following relation.

(52) $P(W(t-1)=[1,1]|W(t)=[2,0])$

$= P(W(t-1)=[1,1]|Q(t)=(2,0,0,0))P(Q(t)=(2,0,0,0)|W(t)=[2,0])$

$+ P(W(t-1)=[1,1]|Q(t)=(1,0,1,0))P(Q(t)=(1,0,1,0)|W(t)=[2,0])$

$+ P(W(t-1)=[1,1]|Q(t)=(0,0,2,0))P(Q(t)=(0,0,2,0)|W(t)=[2,0])$.

Let $Q_1(t)$, $t\leq 0$ be a Markov chain induced by $Q(t)$ by putting the right-hand-sides of (22)~(29) equal to 0. $Q_1(t)$ is a pure migration Markov chain. Let $U(i_1, j_1, i_2, j_2)$ be the stationary probability for $Q_1(t)$. On the other hand, let $Q_2(t)$, $t\leq 0$ be a Markov chain induced by $Q(t)$ by putting $m_1 = m_2 = 0$. $Q_2(t)$ is a Markov chain without migration. In the strong migration limit, we can replace $P(Q(t)=(i_1, j_1, i_2, j_2)|W(t)=[i_1+i_2, j_1+j_2])$ by $U(i_1, j_1, i_2, j_2)$ and $P(W(t-1)=[i_1+i_2, j_1+j_2]|Q(t)=(i_1, j_1, i_2, j_2))$ by $P(W(t-1)=[i_1, j_1, i_2, j_2]|Q_2(t)=(i_1, j_1, i_2, j_2))$ as an approximation. We denote this stochastic process by $R(t)$, $t\leq 0$. It is easily seen that $R(t)$, $t\leq 0$ is a Markov chain on $\{[i,j]|i,j\geq 0,\ 1\leq i+j\leq n\}$. Let $T[i,j]$ be the sum of the lengths (measured in units of $2N$ generations) of all the branches of ancestral tree, assuming that sample consisted of i genes linked to A_1 and j genes linked to A_2. As a direct consequence of the Markov property

of $R(t)$ process, we have

(53) $T[i,j] = (i+j)T_{ij} + T(Z_{ij})$,

where T_{ij} is the holding time in state $[i,j]$, Z_{ij} is the random state to which the process moves and if $Z_{ij} = [h,k]$ then $T(Z_{ij})$ is an independent random variable having the same distribution as $T[h,k]$. Let $M_{[i,j]}$ denote the mean of $T[i,j]$. We can obtain the equations for $T[2,0]$, $T[0,2]$ and $T[1,1]$. For the detailed analysis and an application to the DNA sequence data from the alcohol dehydrogenase (*Adh*) region of *Drosophila melanogaster*, see Kaplan *et al.* (1991).

References

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.

Kaplan,N.L., R.R.Hudson and M.Iizuka (1991). The coalescent process in models with selection, recombination and geographic subdivision. *Genet. Res., Camb.* to appear.

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

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

Kingman,J.F.C., (1982b). The coalescent. *Stochastic Processes Appl.* **13**, 235-248.

Kreitman,M., (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412-417.

Slatkin,M., (1987). The average number of sites separating DNA sequences

drawn from a subdivided population. *Theor. Popul. Biol.* **32**, 42-49.

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

Tajima,F., (1989). DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* **123**, 229-240.

Takahata,N., (1988). The coalescent in two partially isolated diffusion populations. *Genet. Res., Camb.* **52**, 213-222.

Takahata,N. and M.Nei, (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325-344.

Tavare,S., (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theor. Popul. Biol.* **26**, 119-164.

Watterson,G.A., (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **10**, 256-276.