A. Shimizu
Nagoya Math. J.
Vol. 119 (1990), 81-92

A MEASURE VALUED DIFFUSION PROCESS DESCRIBING AN n LOCUS MODEL INCORPORATING GENE CONVERSION

AKINOBU SHIMIZU

§ 1. Introduction

Probability measure valued diffusion processes have been discussed by many authors, in connection with population genetics. Most papers studying probability measure valued diffusions are mainly concerned with the ones describing single locus models. In this paper, we will discuss a measure valued diffusion describing an n locus model. Random sampling, mutation and gene conversion, a kind of interaction between loci, which was introduced and investigated by T. Ohta in [5], [6], will be taken into consideration.

The first aim of this paper is to give a mathematical justification to the Ohta's results. Let $\mathbf{E}$ be the set $[0, 1]^n$ in R^n. Here, the interval $[0, 1]$ describes the set of alleles, and a point of $\mathbf{E}$, which is an n-ple of alleles, means a chromosome with n loci. The bounded operator B_1 introduced in § 2 describes mutation of the neutral infinitely many allele model in n locus case. A positive constant v is mutation rate. We assume that mutation occurs independently at each locus. The operator B_2 defined in § 2 describes the Ohta's gene conversion. A positive constant λ stands for gene conversion rate. Let $\mathscr{P}(\mathbf{E})$ be the space of probability measures on $\mathbf{E}$. We consider the $\mathscr{P}(\mathbf{E})$-valued diffusion process with the generator $\mathbf{G}$ given in § 2. The known results on the diffusion with the generator $\mathbf{G}$ will be stated in Propositions 2.1 and 2.2. For simplicity, it will be omitted to explain the reason why our diffusion process is the stochastic process existing behind the Ohta's arguments. See the paper [8], where the author explained the reason by means of giving a discrete model describing the Ohta's model in 2 locus case with the diffusion approximation. The argument on discrete models and the diffusion approximation in n locus case is essentially the same as in [8].

Received December 2, 1988.

The quantity $c_1(t)$ defined in § 3 is *the average identity probability of genes at different loci.* The quantities $f(t)$ and $c_2(t)$ are *the average probability of allelic identity* and *the average identity probability of two genes taken from different loci of two homologous chromosomes of the population* respectively. In this paper, the quantities $c_1(t)$, $c_2(t)$ and $f(t)$ are defined in terms of the first and the second moments of the measure valued diffusion. We will show that these quantities satisfy the system of ordinary differential equations

$$\begin{aligned}
(d/dt)c_1(t) &= -2(\lambda + v)c_1(t) + 2\lambda\,, \\
(d/dt)f(t) &= 1 - (1 + 2v + 2(n-1)\lambda)f(t) + 2(n-1)\lambda c_2(t)\,, \\
(d/dt)c_2(t) &= c_1(t) + 2\lambda f(t) - (1 + 2v + 2\lambda)c_2(t)\,.
\end{aligned}$$

The relation between our results and the Ohta's results will be explained at the end of § 3.

The second aim is to give another proof of the formula given by the author in [9], which has been further investigated by G.A. Watterson [10]. We consider the average probability at stationarity that we find β_l-kinds of alleles appearing l times, $l = 1, 2, \cdots, n$, in randomly chosen one chromosome. The author showed in [9] that the probability is given by

$$\{n!/\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)\}\prod_{j=1}^{n}\{\theta^{\beta_j}/j^{\beta_j}\beta_j!\}\,,$$

where $\theta = v/\lambda$. In [9], the author discussed a diffusion process taking values in probability distributions on the Young diagrams. The proof does not explain the reason why the sampling formula similar to the well-known Ewens one holds in this case. Here, we will try to explain the reason, giving the proof in terms of the measure valued diffusion.

The most important problem on our diffusion in the application to population genetics is to count the average actual number of alleles existing in a finite population at stationarity. However, it seems rather difficult, and it is still open.

§ 2. Measure valued diffusion process describing an n locus model

In the following, for a topological space $\boldsymbol{X}$, $C(\boldsymbol{X})$ denotes the space of bounded continuous functions on $\boldsymbol{X}$, and $\mathscr{B}(\boldsymbol{X})$ stands for the space of bounded Borel functions on $\boldsymbol{X}$. Let $\mathbf{N}$ be the set of natural numbers. For $k \in \mathbf{N}$, $\boldsymbol{X}^k$ denotes the k-fold direct product of $\boldsymbol{X}$.

Let $\mathbf{E}$ be the set $[0, 1]^n$ in R^n. Let B_1 be the bounded operator on the space $\mathscr{B}(\mathbf{E})$ given by

$$(2.1)\qquad B_1 f(x_1, x_2, \cdots, x_n) = v\sum_{j=1}^{n}\left\{\int_0^1 f(x_1, x_2, \cdots, x_n)dx_j - f(x_1, x_2, \cdots, x_n)\right\},$$

where $(x_1, x_2, \cdots, x_n) \in \mathbf{E}$, $f \in \mathscr{B}(\mathbf{E})$. The operator B_1 for $n = 1$ has already been discussed in Chapter 10 of [1]. Next, we introduce another bounded operator B_2 on $\mathscr{B}(\mathbf{E})$. Define B_2 by

$$(2.2)\qquad B_2 f(x_1, x_2, \cdots, x_n) = \lambda\sum_{j_1, j_2; j_1 \neq j_2}\{\psi_{j_1 j_2} f(x_1, x_2, \cdots, x_n) - f(x_1, x_2, \cdots, x_n)\}.$$

Here, $\psi_{j_1 j_2}$ is the operator of $\mathscr{B}(\mathbf{E})$ to $\mathscr{B}(\mathbf{E})$, by which the variable x_{j_2} is replaced by the variable x_{j_1}:

$$\psi_{j_1 j_2} f(x_1, x_2, \cdots, x_n) = f(x_1, \cdots, x_{j_2-1}, x_{j_1}, x_{j_2+1}, \cdots, x_n).$$

The operator B_2 in 2 locus case was discussed by the author [8]. Define B by

$$(2.3)\qquad B = B_1 + B_2.$$

Note that the operator B generates a Feller semigroup on $C(\mathbf{E})$.

Let $\mathscr{P}(\mathbf{E})$ be the space of probability measures on $\mathbf{E}$, and

$$\mathscr{D} = \{\phi \in C(\mathscr{P}\mathbf{E})):\ \phi(\mu) = F(\langle f_1, \mu\rangle, \cdots, \langle f_k, \mu\rangle),\ k \in \mathbf{N},\ F \text{ is a polynomial on } R^k,\ f_1, \cdots, f_k \in C(\mathbf{E})\},$$

and

$$\mathscr{D}^+ = \{\phi \in \mathscr{B}(\mathscr{P}(\mathbf{E})):\ \phi(\mu) = \langle \boldsymbol{f}, \mu^k\rangle,\ k \in \mathbf{N},\ f \in \mathscr{B}(\mathbf{E}^k)\},$$

where $\mu^k \in \mathscr{P}(\mathbf{E}^k)$ denotes the k-fold product measure for $\mu \in \mathscr{P}(\mathbf{E})$. Note that $\mathscr{D} \subset \mathscr{D}^+$.

For $\phi \in \mathscr{D}^+$ of the form $\phi(\mu) = \langle \boldsymbol{f}, \mu^k\rangle$, define

$$(2.4)\qquad \mathbf{G}\phi(\mu) = \sum_{1 \leq i < j \leq k}(\langle \Psi_{ij}\boldsymbol{f}, \mu^{k-1}\rangle - \langle \boldsymbol{f}, \mu^k\rangle) + \sum_{i=1}^{k}\langle B^{(i)}\boldsymbol{f}, \mu^k\rangle,$$

where the operators Ψ_{ij}: $\mathscr{B}(\mathbf{E}^k) \to \mathscr{B}(\mathbf{E}^{k-1})$ and $B^{(i)}$: $\mathscr{B}(\mathbf{E}^k) \to \mathscr{B}(\mathbf{E})$ are given by

$$\Psi_{ij}\boldsymbol{f}(X_1, X_2, \cdots, X_{k-1}) = \boldsymbol{f}(X_1, \cdots, X_{j-1}, X_i, X_j, \cdots, X_{k-1}),$$

and

$$B^{(i)}\boldsymbol{f}(X_1, X_2, \cdots, X_k) = B[\boldsymbol{f}(X_1, \cdots, X_{i-1}, \cdot, X_{i+1}, \cdots, X_k)](X_i),$$

for $f \in \mathscr{B}(\mathbf{E}^k)$, and $X_m \in \mathbf{E}$, $m = 1, 2, \cdots, k$. The operator Ψ_{ij} replaces the variable X_j of f by X_i, and it changes the numbering of the variables X_{j+m} to X_{j+m-1} for $m = 1, 2, \cdots, k-j$. To operate $B^{(i)}$ to f means to operate B regarding the k-variable function f as a function with one variable X_i. For $k = 1$, the first term of the right-hand side of (2.4) is equal to zero.

Set $\mathbf{A} = \{(\phi, G\phi)\colon \phi \in \mathscr{D}\}$ and $\mathbf{A}^+ = \{(\phi, G\phi)\colon \phi \in \mathscr{D}^+\}$. The space of continuous functions $\omega\colon [0, \infty) \to \mathscr{P}(\mathbf{E})$ is denoted by $C([0, \infty), \mathscr{P}(\mathbf{E}))$. The $C([0, \infty), \mathscr{P}(\mathbf{E}))$-martingale problem for $\mathbf{A}^+$ which we will discuss is formulated as follows. A stochastic process $\{\mu(t), t \geqq 0\}$ with sample paths in $C([0, \infty), \mathscr{P}(\mathbf{E}))$ is called a solution to the $C([0, \infty), \mathscr{P}(\mathbf{E}))$-martingale problem for $\mathbf{A}$ (or for $\mathbf{A}^+$) if

$$\phi_1(\mu(t)) - \int_0^t \phi_2(\mu(s))ds \tag{2.5}$$

is a martingale with respect to $\sigma(\mu(s)\colon 0 \leq s \leq t)$ for any $(\phi_1, \phi_2) \in \mathbf{A}$ (or $\mathbf{A}^+$ respectively). The next statement is found in Ethier and Griffiths [3].

PROPOSITION 2.1. *The martingale problem for* $\mathbf{A}$ *mentioned above is well posed. That is, there exists a solution* $\{\mu(t), t \geqq 0\}$ *of the martingale problem for* $\mathbf{A}$ *with any initial distribution* $\mu(0) = \mu_0$, *and every solution with arbitrarily given initial distribution induces the same distribution on* $C([0, \infty), \mathscr{P}(\mathbf{E}))$. *The solution* $\{\mu(t), t \geqq 0\}$ *is also the solution of the martingale problem for* $\mathbf{A}^+$, *and the martingale given by* (2.5) *for* $(\phi_1, \phi_2) \in \mathbf{A}^+$ *has sample paths belonging to* $C([0, \infty), R)$ *with probability* 1.

Hence, the measure valued process $\{\mu(t), t \geqq 0\}$ is a diffusion process. The set of purely atomic measure on $\mathbf{E}$ is denoted by $P_a(\mathbf{E})$. Then, we have the next proposition.

PROPOSITION 2.2. *The diffusion process* $\{\mu(t), t \geqq 0\}$ *with arbitrarily given initial distribution satisfies*

$$P[\mu(t) \in P_a(\mathbf{E}) \text{ for any } t > 0] = 1. \tag{2.6}$$

The process $\{\mu(t), t \geqq 0\}$ *has a unique stationary distribution, which is denoted by* $\tilde{\mu}$ *and*

$$\tilde{\mu}(P_a(\mathbf{E})) = 1. \tag{2.7}$$

Theorem 2.4 *in* [2] *implies* (2.6). *The proof of the ergodocity is essentially the same as the ones in* [4], [7] *and* [3], *and it is omitted.*

§ 3. Average identity probabilities introduced by T. Ohta

From now on, we will discuss the diffusion process $\{\mu(t),\ t \geqq 0\}$ defined in the previous section.

Define

$$f_{i_1 i_2} = \chi_{\{x_{i_1} = x_{i_2}\}}(x_1, x_2, \cdots, x_n) \in \mathscr{B}(\mathbf{E}), \tag{3.1}$$

$$\phi^1_{i_1 i_2}(\mu) = \langle f_{i_1 i_2}, \mu \rangle \qquad \text{for } i_1 \leqq i_2, \tag{3.2}$$

and

$$\phi^1(\mu) = 2 \sum_{(i_1, i_2): i_1 < i_2} \phi^1_{i_1 i_2}(\mu). \tag{3.3}$$

Define $\bar{\phi}^1(\mu)$ and $c_1(t)$ by

$$\begin{aligned} \bar{\phi}^1(\mu) &= \{1/n(n-1)\}\phi^1(\mu) \\ &= \left\{1 \Big/ \binom{n}{2}\right\} \sum_{i_1 < i_2} \phi^1_{i_1 i_2}(\mu), \end{aligned} \tag{3.4}$$

and

$$c_1(t) = E[\bar{\phi}^1(\mu(t))]. \tag{3.5}$$

Operate the generator G to $\phi^1_{i_1 i_2}(\mu)$, $i_i < i_2$, then we have

$$\begin{aligned} G\phi^1_{i_1 i_2}(\mu) &= v\left(\left\langle \sum_{j=1}^n \int_0^1 f_{i_1 i_2} dx_j,\ \mu \right\rangle - n\langle f_{i_1 i_2}, \mu \rangle\right) \\ &\quad + \lambda \sum_{1 \leqq j_1 < j_2 \leqq n} (\langle \psi_{j_1 j_2} f_{i_1 i_2}, \mu \rangle + \langle \psi_{j_2 j_1} f_{i_1 i_2}, \mu \rangle - 2\langle f_{i_1 i_2}, \mu \rangle) \\ &= -2v\langle f_{i_1 i_2}, \mu \rangle \\ &\quad + \lambda \sum_{1 \leqq j_1 < j_2 = i_1 \leqq n} (\langle \psi_{j_1 i_1} f_{i_1 i_2}, \mu \rangle - \langle f_{i_1 i_2}, \mu \rangle) \\ &\quad + \lambda \sum_{1 \leqq j_1 = i_1 < j_2 \leqq n} (\langle \psi_{j_2 i_1} f_{i_1 i_2}, \mu \rangle - \langle f_{i_1 i_2}, \mu \rangle) \\ &\quad + \lambda \sum_{1 \leqq j_1 < j_2 = i_2 \leqq n} (\langle \psi_{j_1 j_2} f_{i_1 i_2}, \mu \rangle - \langle f_{i_1 i_2}, \mu \rangle) \\ &\quad + \lambda \sum_{1 \leqq j_1 = i_2 < j_2 \leqq n} (\langle \psi_{j_2 i_2} f_{i_1 i_2}, \mu \rangle - \langle f_{i_1 i_2}, \mu \rangle). \end{aligned} \tag{3.6}$$

Here, we have used the following simple properties.

$$\text{(a)} \quad \int_0^1 f_{i_1 i_2} dx_j = \begin{cases} 0 & \text{if } j = i_1 \text{ or } i_2 \\ f_{i_1 i_2} & \text{otherwise}, \end{cases} \tag{3.7}$$

$$\text{(b)} \quad \psi_{j_1 j_2} f_{i_1 i_2} = f_{i_1 i_2} \qquad \begin{array}{l} \text{if } j_1 = i_i < j_2 \neq i_2 \\ \text{or } i_1 < i_2 = j_1 < j_2, \end{array} \tag{3.8}$$

$$\text{(c)} \quad \psi_{j_2 j_1} f_{i_1 i_2} = f_{i_1 i_2} \qquad \begin{array}{l} \text{if } j_1 < j_2 = i_1 < i_2 \\ \text{or } i_1 \neq j_1 < j_2 = i_2, \end{array} \tag{3.9}$$

and

(3.10) (d) $\psi_{j_1j_2}f_{i_1i_2} = \psi_{j_2j_1}f_{i_1i_2} = f_{i_1i_2}$,
if $j_1 \neq i_1$, $j_1 \neq i_2$, $j_2 \neq i_1$ and $j_2 \neq i_2$.

Noting the facts that $\psi_{ki_1}f_{i_1i_2} = f_{ki_2}$, $\psi_{ki_2}f_{i_1i_2} = f_{i_1k}$, and that $f_{i_1i_1} = f_{i_2i_2} = 1$, we get

(3.11) The right-hand side of (3.6) $= -2v < f_{i_1i_2}, \mu\rangle$
$+ \lambda \sum_k \{(\langle f_{ki_2}, \mu\rangle - \langle f_{i_1i_2}, \mu\rangle) + (\langle f_{i_1k}, \mu\rangle - \langle f_{i_1i_2}, \mu\rangle)\}$.

Note that (3.11) also holds for $i_1 > i_2$. Summing up the both sides of (3.11) on i_1 and i_2 satisfying $i_1 \neq i_2$, we obtain

$$\begin{aligned} G\phi^1(\mu) &= -2v\phi^1(\mu) + \lambda\{2n(n-1) - 2\phi^1(\mu)\} \\ &= -2(\lambda + v)\phi^1(\mu) + 2\lambda n(n-1). \end{aligned} \tag{3.12}$$

Hence, that

$$G\bar{\phi}^1(\mu) = -2(\lambda + v)\bar{\phi}^1(\mu) + 2\lambda.$$

Furthermore, Proposition 2.1 implies that $E[G\bar{\phi}^1(\mu(s))]$ is continuous in s, and that $[E\bar{\phi}^1(\mu(t))]$ is differentiable in t.

Thus, we obtain the next statement.

THEOREM 3.1. *The average probability of allelic identity* $c_1(t)$, *defined by* (3.1)–(3.5), *satisfies the ordinary differential equation*

$$(d/dt)c_1(t) = -2(\lambda + v)c_1(t) + 2\lambda. \tag{3.13}$$

Let $\mu^2 = \mu \times \mu$ be the direct product of $\mu \in \mathscr{P}(\mathbf{E})$, and let $g_{ij}(X_1, X_2)$ be a function on $\mathbf{E}^2$ given by

$$\begin{aligned} &g_{ij}(X_1, X_2) = g_{ij}(x_1^1, \cdots, x_n^1, x_1^2, \cdots, x_n^2) = \chi_{A_{ij}}(X_1, X_2), \\ &(X_1, X_2) = (x_1^1, \cdots, x_n^1, x_1^2, \cdots, x_n^2) \in \mathbf{E}^2, \end{aligned} \tag{3.14}$$

where $A_{ij} = \{(X_1, X_2)\colon x_i^1 = x_j^2\}$. Obviously, $g_{ij} \in \mathscr{B}(\mathbf{E}^2)$. Define $\phi_{ij}^2(\mu)$, $\bar{\phi}^2(\mu)$, $\tilde{\phi}^2(\mu)$, $f(t)$ and $c_2(t)$ by

$$\phi_{ij}^2(\mu) = \langle g_{ij}, \mu^2\rangle, \tag{3.15}$$

$$\bar{\phi}^2(\mu) = (1/n)\sum_i \phi_{ii}^2, \tag{3.16}$$

$$\tilde{\phi}^2(\mu) = (1/n(n-1))\sum_{(l,k):l\neq k} \langle g_{lk}, \mu^2\rangle, \tag{3.17}$$

$$f(t) = E[\bar{\phi}^2(\mu(t))] \tag{3.18}$$

and

$$c_2(t) = E[\tilde{\phi}^2(\mu(t))]. \tag{3.19}$$

Now, we will try to derive the equation which the quantities $f(t)$ and $c_2(t)$ satisfy.

By the definition of the generator G, we see

$$G\phi_{ij}^2(\mu) = \langle \Psi_{12} g_{ij}, \mu \rangle - \phi_{ij}^2(\mu) + \sum_{l=1}^{2} \langle B^{(l)} g_{ij}, \mu \rangle .$$

Note the following facts. The equalities

$$\langle B^{(1)} g_{ij}, \mu \rangle = \langle B g_{ij}(\cdot, X_2), \mu \rangle = -v\phi_{ij}^2(\mu) + \lambda \sum_l (\langle g_{lj}, \mu^2 \rangle - \langle g_{ij}, \mu^2 \rangle),$$

and

$$\langle B^{(2)} g_{ij}, \mu \rangle = \langle B g_{ij}(X_1, \cdot), \mu \rangle = -v\phi_{ij}^2(\mu) + \lambda \sum_l (\langle g_{il}, \mu^2 \rangle - \langle g_{ij}, \mu^2 \rangle),$$

hold. Besides, we see that

$$\Psi_{12} g_{ij} = 1 \qquad \text{for each } i,$$

and that

$$\Psi_{12} g_{ij} = f_{ij} = \chi_{\{x_i = x_j\}}(x_1, \cdots, x_n) \in \mathscr{B}(\mathbf{E}),$$

for (i, j) such that $i \neq j$.

Thus we can calculate $G\phi_{ii}^2(\mu)$, and obtain

$$G\phi_{ii}^2(\mu) = 1 - (1 + 2v + 2(n-1)\lambda)\phi_{ii}^2(\mu) + \lambda \sum_{l: l \neq i} (\langle g_{li}, \mu^2 \rangle + \langle g_{il}, \mu^2 \rangle).$$

Hence, we get

$$G \sum_i \phi_{ii}^2(\mu) = n - (1 + 2v + 2(n-1)\lambda) \sum_i \phi_{ii}^2(\mu) + 2\lambda \sum_{(l,k): l \neq k} \langle g_{lk}, \mu^2 \rangle . \tag{3.20}$$

By (3.16), (3.17) and (3.20), we obtain

$$G\bar{\phi}^2(\mu) = 1 - (1 + 2v + 2(n-1)\lambda)\bar{\phi}^2(\mu) + 2(n-1)\lambda\tilde{\phi}^2(\mu) . \tag{3.21}$$

Since we have

$$G\phi_{ij}^2(\mu) = \phi_{ij}^1(\mu) - (1 + 2v)\phi_{ij}^2(\mu) + \lambda\{\textstyle\sum_l (\phi_{lj}^2(\mu) - \phi_{ij}^2(\mu) + \phi_{il}^2(\mu) - \phi_{ij}^2(\mu))\},$$

for (i, j) such that $i \neq j$, we get

$$\begin{aligned} Gn(n-1)\tilde{\phi}^2(\mu) &= G \textstyle\sum_{(i,j): i \neq j} \phi_{ij}^2(\mu) = \sum_{(i,j): i \neq j} \phi_{ij}^1(\mu) - (1 + 2v) \sum_{(i,j): i \neq j} \phi_{ij}^2(\mu) \\ &\quad + \lambda\{\textstyle\sum_j \sum_l \sum_{i: i \neq j} \phi_{lj}^2(\mu) + \sum_i \sum_l \sum_{j: j \neq i} \phi_{il}^2(\mu) - 2n \sum_{(i,j): i \neq j} \phi_{ij}^2(\mu)\} \\ &= \textstyle\sum_{(i,j): i \neq j} \phi_{ij}^1(\mu) - (1 + 2v) \sum_{(i,j): i \neq j} \phi_{ij}^2(\mu) \\ &\qquad + \lambda\{2(n-1) \sum_i \sum_j \phi_{ij}^2(\mu) - 2n \sum_{(i,j): i \neq j} \phi_{ij}^2(\mu)\} \\ &= n(n-1)\bar{\phi}^1(\mu) - (1 + 2v)n(n-1)\tilde{\phi}^2(\mu) \\ &\qquad + \lambda\{2n(n-1)\bar{\phi}^2(\mu) - 2n(n-1)\tilde{\phi}^2(\mu)\} . \end{aligned}$$

Thus, we obtain

$$G\tilde{\phi}^2(\mu) = \bar{\phi}^1(\mu) - (1 + 2v + 2\lambda)\tilde{\phi}^2(\mu) + 2\lambda\bar{\phi}^2(\mu). \tag{3.22}$$

Therefore, by (3.21) and (3.22), we get the next theorem.

THEOREM 3.2. *The quantities $f(t)$ and $c_2(t)$, defined by* (3.14)–(3.19), *satisfy the system of ordinary differential equations*

$$\begin{aligned}(d/dt)f(t) &= 1 - (1 + 2v + 2(n-1)\lambda)f(t) + 2(n-1)\lambda c_2(t), \\ (d/dt)c_2(t) &= c_1(t) + 2\lambda f(t) - (1 + 2v + 2\lambda)c_2(t).\end{aligned} \tag{3.23}$$

At the end of this section, we will explain the relation between our results (3.13), (3.23) and Ohta's results [5]. In our formulation, roughly speaking, the quantities f, c_1 and c_2 change in one generation $(1/2N)$ times of their derivatives, where N stands for the population size in the discrete model. Using Ohta's notation, we see

$$\begin{aligned}(d/dt)f(t) &= 2N\Delta f, \\ (d/dt)c_1(t) &= 2N\Delta c_1, \\ (d/dt)c_2(t) &= 2N\Delta c_2,\end{aligned}$$

where Δ denotes the change per one generation. Since the mutation rate in one generation is roughly equal to v $(1/2N)$ in our case, the rate v should be replaced by $2N\,v$ in Ohta's discussion, where the parameter v in [5] means the mutation rate in one generation. As for the rate of gene conversion, $(n-1)\lambda$ in our equation should be replaced by $2N\,\lambda$. Then, we have

$$\begin{aligned}\Delta f &= -\{2v + (1/2N) + 2\lambda\}f + 2\lambda c_2 + (1/2N), \\ \Delta c_1 &= -\{2v + 2(\lambda/(n-1))\}c_1 + 2(\lambda/(n-1)), \\ \Delta c_2 &= 2(\lambda/(n-1))f + (1/2N)c_1 - \{2v + (1/2N) + 2(\lambda/n-1))\}c_2.\end{aligned}$$

These equations are just the same as (3), (5) and (7) in [5], when we do not consider interchromosomal crossing-over.

§ 4. Sampling formula similar to the Ewens one

Let $\beta_1, \cdots, \beta_n$ be non-negative integers such that $\sum_l l\beta_l = n$. A point $(x_1, x_2, \cdots, x_n)$ in $\mathbf{E}$ is defined to be belonging to $\mathbf{E}_\beta$, $\beta = \{\beta_1, \cdots, \beta_n\}$, if and only if there exist distinct $\sum_l \beta_l$ real numbers $y_1, y_2, \cdots, y_{\sum \beta_l} \in [0, 1]$ such that

$$\#\{i\colon x_i = y_{\iota(l)+k}\} = l \qquad \text{for } k = 1, \cdots, \beta_l,$$

where $t(l)$ equals $\sum_{i=1}^{j=l-1}\beta_i$ for $l \geqq 2$ and $t(1) = 0$. The formula stated in the introduction can be formulated as follows.

$$\text{(4.1)}\quad \int \mu(\mathbf{E}_\beta)\tilde{\mu}(d\mu) = \{n!/\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)\}\prod_{j=1}^{n}\{\theta^{\beta_j}/j^{\beta_j}\beta_j!\},$$

where $\theta = v/\lambda$. The proof of (4.1) in this section is essentially due to a private discussion with Professor S.N. Ethier.

First, recall the single locus case which was discussed in [1]. Let $\mathbf{E}_0 = [0, 1]$, $\mu \in \mathscr{P}(\mathbf{E}_0)$, and $f \in \mathscr{B}(\mathbf{E}_0^k)$. Define $\phi(\mu) = \langle f, \mu^k\rangle$, where μ^k is the k-fold direct product of μ. The diffusion process taking values in $\mathscr{P}(\mathbf{E}_0)$ describing the so-called infinitely many neutral allele model has the generator G given by

$$\begin{aligned} G\phi(\mu) &= \sum_{i<j}(\langle \Psi_{ij}f, \mu^{k-1}\rangle - \langle f, \mu^k\rangle) \\ &\quad + v\sum_{l=1}^{k}\left(\left\langle \int_0^1 f dx_l, \mu^k\right\rangle - \langle f, \mu^k\rangle\right) \quad \text{(Chapter 10 in [1])}. \end{aligned}$$

Let $A_1, A_2, \cdots, A_L$ be a measurable partition of $\mathbf{E}_0$ such that each A_l has the mass $(1/L)$ with respect to the Lebesgue measure. Let h_l be the indicator function of the set A_l, $l = 1, 2, \cdots, L$, and put

$$\phi_{\boldsymbol{\alpha}}(\mu) = \langle h_1, \mu\rangle^{\alpha_1}\langle h_2, \mu\rangle^{\alpha_2}\cdots\langle h_L, \mu\rangle^{\alpha_L},$$

where α_l, $l = 1, 2, \cdots, L$, are non-negative integers, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_L)$ and $\sum_l \alpha_l = k$. Then, we have

$$\begin{aligned} \text{(4.2)}\quad G\phi_{\boldsymbol{\alpha}}(\mu) &= (1/2)\sum_{l=1}^{L}\alpha_l(\alpha_l - 1)\phi_{\boldsymbol{\alpha}-\boldsymbol{\varepsilon}_l}(\mu) - \{k(k-1)/2\}\phi_{\boldsymbol{\alpha}}(\mu) \\ &\quad + v\sum_{l=1}^{L}[\alpha_l\{(1/L)\phi_{\boldsymbol{\alpha}-\boldsymbol{\varepsilon}_l}(\mu) - \phi_{\boldsymbol{\alpha}}(\mu)\}] \\ &= \sum_{l=1}^{L}\alpha_l\{(\alpha_l - 1)/2 + (v/L)\}\phi_{\boldsymbol{\alpha}-\boldsymbol{\varepsilon}_l}(\mu) - k\{(k-1)/2 + v\}\phi_{\boldsymbol{\alpha}}(\mu), \end{aligned}$$

where $\boldsymbol{\varepsilon}_i = (\delta_{li})_{l=1,\cdots,L}$, and δ_{li} stands for Kronecker's δ. Let $\tilde{\mu}$ be the stationary distribution of the $\mathscr{P}(\mathbf{E}_0)$-valued diffusion process. Since $\int G\phi_{\boldsymbol{\alpha}}(\mu)\tilde{\mu}(d\mu) = 0$, we obtain by (4.2)

$$\text{(4.3)}\quad \sum_{l=1}^{L}\alpha_l\{(\alpha_l - 1)/2 + (v/L)\}\hat{\phi}_{\boldsymbol{\alpha}-\boldsymbol{\varepsilon}_l}(\tilde{\mu}) - k\{(k-1)/2 + v\}\hat{\phi}_{\boldsymbol{\alpha}}(\tilde{\mu}) = 0,$$

where $\hat{\phi}_{\boldsymbol{\alpha}}(\tilde{\mu}) = \int \phi_{\boldsymbol{\alpha}}(\mu)\tilde{\mu}(d\mu)$.

Hence we see

$$\hat{\phi}_{\boldsymbol{\alpha}}(\tilde{\mu}) = \{1/k(k-1+2v)\}\sum_{l=1}^{L}\alpha_l\{(\alpha_l - 1) + (2v/L)\}\hat{\phi}_{\boldsymbol{\alpha}-\boldsymbol{\varepsilon}_l}(\tilde{\mu}).$$

Noting that

$$\int \phi_{\varepsilon_i}(\mu)\tilde{\mu}(d\mu) = 1/L\,, \tag{4.4}$$

by (4.3) we obtain the well-known result

$$\int \phi_{\alpha}(\mu)\tilde{\mu}(d\mu) = \{\Gamma(2v)/\Gamma(2v/L)^L\} \prod_{l=1}^{L} \Gamma(\alpha_l + 2(v/L))/\Gamma(2v+k)\,. \tag{4.5}$$

Let $\beta_1, \cdots, \beta_n$ be non-negative integers such that $\sum_l l\beta_l = n$, as the biginning of this section. Now consider the next problem. What is the average probability at stationarity that we find β_l-kinds of alleles appearing l times, $l = 1, 2, \cdots, n$, in randomly chosen n genes? The answer to this question is the well-known Ewens sampling formula. That is, the probability is given by (4.1) with $\Theta = 2v$. The proof of this formula can be found in Chapter 10 in [1], which is a little complicated. The Ewens' sampling formula can be shown directly from (4.5) by modifying the proof in [1]. We will omit the details here, because it seems known.

Now, consider the n locus model. First take a partition of the set of loci. Let $\{S_i\}_{i=0,1,2,\cdots,L}$ be a family of disjoint subsets of $\{1, 2, \cdots, n\}$ such that S_i has α_i elements for each i. Here, α_i are non-negative integers and $\sum_{l=0}^{L} \alpha_l = n$. Note that S_i may be empty. Define f_{α} by

$$f_{\alpha}(S_0, S_1, \cdots, S_L) = \prod_{i=1}^{L} \prod_{l\in S_i} h_i(x_l)\,, \tag{4.6}$$

where h_i is the indicator function of the set A_i for each i, $i = 1, 2, \cdots, n$. Obviously, $f_{\alpha} \in \mathscr{B}(\mathbf{E})$, $\mathbf{E} = [0, 1]^n$, and f_{α} does not depend on the variables x_l, $l \in S_0$. Define the degree of f_{α} by

$$\deg f_{\alpha} = \sum_{i=1}^{L} \alpha_i\,.$$

Put

$$\phi_{\alpha}(S_0, S_1, \cdots, S_L)(\mu) = \langle f_{\alpha}(S_0, S_1, \cdots, S_L), \mu\rangle \in \mathscr{B}(\mathscr{P}(\mathbf{E}))\,. \tag{4.7}$$

Define the degree of $\phi_{\alpha}(\mu)$ by the degree of f_{α}. When μ is fixed, $\phi_{\alpha}(\mu)$ means the probability that we find genes belonging to A_i at loci belonging to S_i for $i \geqq 1$. Let ε_i be the vector with components 0 or 1 such that only the i-th coordinate equals 1. Put

$$\phi_{\alpha-\varepsilon_i,l} = \langle f_{\alpha-\varepsilon_i}(S_0 \cup \{l\}, S_1, \cdots, S_i - \{l\}, \cdots, S_L), \mu\rangle\,,$$

for $l \in S_i$, and

$$\phi_{\alpha,l,m,i} = \langle f_\alpha(S_0 \cup \{m\} - \{l\}, S_1, \cdots, S_i \cup \{l\} - \{m\}, \cdots, \cdots, S_L), \mu \rangle ,$$

for $l \in S_0$, and $m \in S_i$, $i \geqq 1$. Then, we get

$$\begin{aligned} (4.8) \quad G\phi_\alpha(\mu) = {} & v \sum_{i=1}^{L} \sum_{l \in S_i} \{(1/L)\phi_{\alpha-\varepsilon_i,l}(\mu) - \phi_\alpha(\mu)\} \\ & + \lambda[\sum_{i=1}^{L} (\alpha_i - 1) \sum_{l \in S_i} \phi_{\alpha-\varepsilon_i,l}(\mu) \\ & + \sum_{i=1}^{L} \sum_{l \in S_0} \sum_{m \in S_i} \phi_{\alpha,l,m,i}(\mu) - \{k(k-1) + \alpha_0 k\}\phi_\alpha(\mu)] , \end{aligned}$$

if $\deg \phi_\alpha(\mu) = k$. Note the fact that any $h \in \mathscr{B}(\mathbf{E})$ satisfies

$$\begin{aligned} (4.9) \quad & \int \exp\{\langle h(x_1, x_2, \cdots, x_n), \mu \rangle\} \tilde{\mu}(d\mu) \\ & = \int \exp\{\langle h(x_{\sigma(1)}, x_{\sigma(2)}, \cdots, x_{\sigma(n)}), \mu \rangle\} \tilde{\mu}(d\mu) , \end{aligned}$$

for any permutation σ of $\{1, 2, \cdots, n\}$, which is deduced from the symmetry of the generator G and the property of the stationary distribution $\tilde{\mu}$. Define

$$\hat{\phi}_\alpha = \int \phi_\alpha(\mu) \tilde{\mu}(d\mu) ,$$

and $\hat{\phi}_{\alpha-\varepsilon_i,l}$, $\hat{\phi}_{\alpha,l,m,i}$ in the same manner, then we see by the above fact (4.9) that $\hat{\phi}_{\alpha-\varepsilon_i,l}$ is independent of l for each i, and that $\hat{\phi}_{\alpha,l,m,i}$ is independent of l, m, and i. Hence, they can be written by $\hat{\phi}_{\alpha-\varepsilon_i}$ and $\hat{\phi}_\alpha$ respectively. Combining this with (4.8) and $\int G\phi_\alpha(\mu)\tilde{\mu}(d\mu) = 0$, we obtain

$$\begin{aligned} (4.10) \quad & v \sum_{i=1}^{L} \alpha_i \{(1/L)\hat{\phi}_{\alpha-\varepsilon_i} - \hat{\phi}_\alpha\} + \lambda \{\sum_{i=1}^{L} \alpha_i(\alpha_i - 1)\hat{\phi}_{\alpha-\varepsilon_i} - k(k-1)\hat{\phi}_\alpha\} \\ & = \lambda[\sum_{i=1}^{L} \alpha_i \{(\alpha_i - 1) + (v/\lambda)/L\}\hat{\phi}_{\alpha-\varepsilon_i} - \{k(k - 1 + v/\lambda)\hat{\phi}_\alpha] \\ & = 0 , \end{aligned}$$

for ϕ_α with degree k. Making use of (4.9), we see that

$$(4.11) \qquad \hat{\phi}_{\varepsilon_i} = 1/L, \qquad \text{for each } i \geqq 1.$$

Note that (4.10) and (4.11) have the same form as (4.3) and (4.4). If we replace $2v$ in (4.3) by v/λ, then we get (4.10). This is the essential part of our argument.

Thus we obtain the next theorem.

THEOREM 4.1. *The average of* $\phi_\alpha(S_0, S_1, \cdots, S_L)(\mu)$, *defined by* (4.6) *and* (4.7), *with respect to* $\tilde{\mu}$ *is equal to*

$$\{\Gamma(\theta)/\Gamma(\theta/L)^L\} \prod_{l=1}^{L} \Gamma(\alpha_l + \theta/L)/\Gamma(\theta + k) ,$$

where θ *is* v/λ.

By the argument similar to the single locus case, we can see that Theorem 4.1 implies the formula (4.1) given at the beginning of this section.

REFERENCES

[1] S. N. Ethier and T. G. Kurtz, Markov processes: Characterization and convergence, John Wiley & Sons, New York, 1986.
[2] ——, The infinitely many alleles model with selection as a measure-valued diffusion, Lecture Notes in Biomathematics, **70** (1987), 72–86.
[3] S. N. Ethier and R. C. Griffiths, The infinitely-many-sites model as a measure-valued diffusion, Ann. Prob., **15** (1987), 515–545.
[4] W. H. Fleming and M. Viot, Some measure valued Markov processes in population genetics theory, Indiana Univ. Math. J., **28** (1979), 817–843.
[5] T. Ohta, Allelic and nonallelic homology of a supergene family, Proc. Natl. Acad. Sci. USA, **79** (1982), 3251–3254.
[6] ——, On the evolution of multigene families. Theor. Pop. biol., **23** (1983), 216–240.
[7] T. Shiga, An interaction system in population genetics, J. Math. Kyoto Univ., **20** (1980), 213–242.
[8] A. Shimizu, Diffusion approximation of an infinite allele model incorporating gene conversion, Population genetics and Molecular Evolution, eds. T. Ohta and K. Aoki. Japan Sci. Soc. Press, Tokyo/Springer-Verlag, Berlin, 1985.
[9] ——, Stationary distribution of a diffusion process taking values in probability distributions on the partitions, Lecture Notes in Biomathematics, **70** (1987), 100–114.
[10] G. A. Watterson, Allele frequencies in multigene families. Theoretical Population Biology, **35** (1989), 142–160.

Department of Mathematics
Nagoya Institute of Technology
Gokiso, Showa-ku
Nagoya 466, Japan