Bull. Korean Math. Soc. **42** (2005), No. 2, pp. 397–404

ON THE LIMITING DIFFUSION OF SPECIAL DIPLOID MODEL IN POPULATION GENETICS

WON CHOI

ABSTRACT. In this note, we characterize the limiting diffusion of a diploid model by defining the discrete generator for the rescaled Markov chain. We conclude that this limiting diffusion model is with uncountable state space and mutation selection and special "mutation or gene conversion rate".

1. Introduction

Consider n locus model

$$X = (x_1, x_2, \cdots, x_d) \in R^d.$$

A partition X describes a state of a chromosome and X means that there exist d kinds of alleles which occupy x_1 loci, x_2 loci, $\cdots$, x_d loci. Let p_i denote the frequency of chromosome of type X_i.

Let S be a countable set. In population genetics theory we often encounter diffusion process on the domain

$$K = \{p = (p_i)_{i \in S} \colon p_i \geq 0, \sum_{i \in S} p_i = 1\}.$$

We suppose that the vector $p(t) = (p_1, p_2, \cdots)$ of gene frequencies varies with time t.

Let L be a second order differential operator on K

$$L = \sum_{i,j \in S} a_{ij}(p) \frac{\partial^2}{\partial p_i \partial p_j} + \sum_{i \in S} b_i(p) \frac{\partial}{\partial p_i}$$

Received April 27, 2004.

2000 Mathematics Subject Classification: 60G44, 60H30, 92D10.

Key words and phrases: diploid model, martingale problem, mutation or gene conversion rate, mutation selection, discrete generator for the rescaled Markov chain, limiting diffusion.

This research was supported by University of Incheon Research Grant, 2004–2005.

with domain $C^2(K)$, where $\{a_{ij}\}$ is a real symmetric and non-negative definite matrix defined on K and $\{b_i\}$ is an measurable function defined on K. The coefficient $\{a_{ij}\}$ comes from chance replacement of individuals by new ones after random mating and $\{b_i\}$ is represented by the addition of "mutation or gene conversion rate" and the effect of natural selection.

We assume that $\{a_{ij}\}$ and $\{b_i\}$ are continuous on K. Let $\Omega = C([0,\infty) : K)$ be the space of all K-valued continuous function defined on $[0,\infty)$. A probability P on $(\Omega, \mathcal{F})$ is called a solution of the (K, L, p)-*martingale problem* if it satisfies the following conditions,

(1) $P(p(0) = p) = 1$.
(2) denoting $M_f(t) = f(p(t)) - \int_0^t Lf(p(t))ds$, $(M_f(t), \mathcal{F}_t)$ is a P-martingale for each $f \in C^2(K)$.

The diffusion operator L was first introduced by Gillespie[5] in case that the partition consists of two points. Choi[2] tried to apply the stochastic differential equation for multi-allelic model. Also, Choi and Lee[1] showed the uniqueness of martingale problem associated with mean vector and obtained a complete description of ergodic property by using of the semigroup method. A key point of their work was that the (K, L, p)-martingale problem in population genetics model is related to simpler stochastic differential equation, so they found various diffusion properties for multi-allelic model.

The many diffusion model in population genetics was that each individual of some "type" and the set of S of types is finite. The case in which S is uncountably infinite, however, requires a different approach. The key idea is to topologize S and replace K by $\mathcal{P}(S)$, the set of Borel probability measures on S with the topology of weak convergence.

If there is the empirical distribution μ of the N genotypes in the parent generation, then the empirical distribution of the N genotypes in the offspring generation is determined from μ in the four steps, corresponding, respectively, to reproduction and selection, recombination, mutation, and random sampling. In particular, an infinite number of zygotes are produced in Hardy-Weinberg form [6] as the initial step in the life cycle.

In this note, we consider a special case of diploid models. We can identify and characterize the limiting diffusion of this diploid model by defining this discrete generator for the rescaled Markov chain. We conclude that the limiting diffusion model is with mutation selection and special "mutation or gene conversion rate".

2. Main results

We start with diploid model. For each positive integer M, let ω_M be a positive, symmetric, bounded, Borel function on S^2, let $R_M((p,q), dx \times dy)$ be a one-step transition function on $S^2 \times \mathcal{B}(S^2)$ satisfying

$$R_M((p,q), dx \times dy) = R_M((q,p), dy \times dx),$$

and $Q_M(p, dx)$ be a one-step transition function on $S \times \mathcal{B}(S)$.

Let N be the diploid population size. We consider $M = 2N$ gametes and the mapping $\eta_M : S^M \to \mathcal{P}(S)$ by letting

$$\eta_M(p_1, p_2, \cdots, p_M) = \frac{1}{M}(\delta_{p_1} + \delta_{p_2} + \cdots + \delta_{p_M}). \tag{2.1}$$

Here $\delta_p \in \mathcal{P}(S)$ denotes the unit mass at $p \in S$. The state space for this model is

$$\mathcal{K}_M(S) = \eta_M(S^M).$$

Given $\mu \in \mathcal{P}(S)$, we define $\mu_1 \in \mathcal{P}(S^2)$ and $\mu_2, \mu_3 \in \mathcal{P}(S)$ by

$$\mu_1(dp \times dq) = \omega_M(p,q)\mu^2(dp \times dq)/\langle \omega_M, \mu^2 \rangle, \tag{2.2}$$

$$\mu_2(dx) = \int_{S^2} R_M((p,q), dx \times S)\, \mu_1(dp \times dq), \tag{2.3}$$

$$\mu_3(dx) = \int_S Q_M(p, dx)\mu_2(dp). \tag{2.4}$$

The Markov chain has one-step transition function $P_M(\mu, d\theta)$ on $\mathcal{K}_M(S) \times (\mathcal{K}_M(S))$ defined by

$$P_M(\mu, \cdot) = \int_{S^M} (\mu_3)^M (dp_1 \times dp_2 \times \cdots \times dp_M)\, \delta_{\eta_M(p_1, p_2, \cdots, p_M)}(\cdot).$$

We start with

LEMMA 1. *Let $\{\nu_\tau^{(M)}, \tau \in Z_+ \ (M = 1, 2, \cdots)\}$ be a sequence of diploid models as described. Assuming weak convergence of initial distributions, it is hold that*

$$\{\nu_{[Mt]}^{(M)}, t \geq 0\} \ \Rightarrow \ \{\mu_t, t \geq 0\} \ \text{as} \ M \to \infty,$$

where $\{\mu_t, t \geq 0\}$ is a diffusion process in $\mathcal{P}(S)$.

Proof. This result follows easily from using of a special case of Wright-Fisher models. See Ethier and Kurtz [4]. □

We can identify and characterize the limiting diffusion of the diploid model by thinking of Lemma 1.

In order to consider a limiting diffusion, we define the discrete generator $\mathcal{L}_M$ for the M-th rescaled Markov chain :

$$(\mathcal{L}_M\phi)(\mu) = M\int_{\mathcal{P}_M(S)}(\phi(\nu)-\phi(\mu))P_M(\mu,\nu)$$

where P_M is given in the diploid models as described above.

We restrict our attention to test functions ϕ of the form

$$\phi(\nu) = \beta_1\langle f_1,\nu\rangle\cdots\beta_k\langle f_k,\nu\rangle,\ \ \phi(\mu) = \langle f_1,\mu\rangle\cdots\langle f_k,\mu\rangle$$

where $f_1,\cdots,f_k \in \mathcal{B}(S)$ and $\{\beta_i\}$ is non-negative constant satisfying that $\sup_i \beta_i < +\infty$. Assume that "mutation or gene conversion rate" is

$$\sum_{k\in S}\beta_k\langle f_i,\mu\rangle - \beta_i - \beta_j \ \text{ for every } \ i<j,$$

in the diploid models as described above. This means that mutations or gene conversions occur with particular rate in case of $i<j$.

Then we have

THEOREM 2. *Suppose that there exist a selection function σ on S^2 and bounded linear operator A, B on $\mathcal{B}(S)$ such that*

$$\omega_M(p,q) = 1+\frac{1}{M}\sigma(p,q)+o(M^{-1}), \tag{2.5}$$

$$\int_S f(x)R_M((p,q),dx\times S) = f(p)+\frac{1}{M}(Bf)(p,q)+o(M^{-1}), \tag{2.6}$$

$$\int_S f(x)Q_M(p,dx) = f(x)+\frac{1}{M}(Af)(p)+o(M^{-1}). \tag{2.7}$$

Then there exist a_{f_i,f_j}, $b_{f_i}\in\mathcal{B}(\mathcal{P}(S))$ such that

$$(\mathcal{L}_M\phi)(\mu)\ \to\ (\mathcal{L}\phi)(\mu)\ \text{ as }\ M\to\infty$$

uniformly in $\mu\in\mathcal{K}_M(S)$, where

$$(\mathcal{L}\phi)(\mu) = \sum_{1\le i<j\le k} a_{f_i,f_j}\prod_{l:l\neq i,j}\langle f_l,\mu\rangle + \sum_{i=1}^{k} b_{f_i}\prod_{l:l\neq i}\langle f_l,\mu\rangle.$$

Proof. By using of test function $\phi(\nu) = \beta_1\langle f_1,\nu\rangle\cdots\beta_k\langle f_k,\nu\rangle$, $\phi(\mu) = \langle f_1,\mu\rangle\cdots\langle f_k,\mu\rangle$, we have

$$(L_M\phi)(\mu) = \int_{\mathcal{K}_M(S)}\left\{\prod_{i=1}^{k}\beta_i\langle f_i,\nu\rangle - \prod_{i=1}^{k}\langle f_i,\mu\rangle\right\}P_M(\mu,d\nu)$$

$$= M\left\{ \int_{S^M} \prod_{i=1}^{k} \beta_i \langle f_i, \eta_M(p_1, \cdots, p_M)\rangle (\mu_3)^M (dp_1 \times \cdots \times dp_M) \right.$$
$$\left. - \prod_{i=1}^{k} \langle f_i, \mu\rangle \right\}$$
$$= M\left\{ \int_{S^M} M^{-k} \sum_{i_1=1}^{M} \cdots \sum_{i_k=1}^{M} (\beta_1 \cdots \beta_k) f_1(p_{i_1}) \cdots \right.$$
$$\left. f_k(p_{i_k}) (\mu_3)^M (dp_1 \times \cdots \times dp_M) - \prod_{i=1}^{k} \langle f_i, \mu\rangle \right\}$$

uniformly in $\mu \in \mathcal{P}(S)$.

Since we assume that "mutation or gene conversion rate" is

$$\sum_{k \in S} \beta_k \langle f_i, \mu\rangle - \beta_i - \beta_j \quad \text{for every } i < j$$

in a special case of the diploid model, we obtain from (2.1) and (2.4)

$$(L_M \phi)(\mu)$$
$$= M\left\{ M^{-k} \frac{M!}{(M-k+1)!} \sum_{1 \le i < j \le k} \beta_i \langle f_i f_j, \mu_3\rangle \right.$$
$$\left(\sum_{k \in S} \beta_k \langle f_i, \mu\rangle - \beta_i - \beta_j \right) \prod_{l: l \ne i,j} \langle f_l, \mu_3\rangle$$
$$\left. + M^{-k} \frac{M!}{(M-k)!} \prod_{i=1}^{k} \langle f_i, \mu_3\rangle - \prod_{i=1}^{k} \langle f_i, \mu\rangle + O(M^{-2}) \right\}$$
$$= \sum_{1 \le i < j \le k} \beta_i (\langle f_i f_j, \mu_3\rangle - \langle f_i, \mu_3\rangle \langle f_j, \mu_3\rangle)$$
$$\left(\sum_{k \in S} \beta_k \langle f_i, \mu\rangle - \beta_i - \beta_j \right) \prod_{l: l \ne i,j} \langle f_l, \mu_3\rangle$$
$$+ \sum_{i=1}^{k} M(\langle f_i, \mu_3\rangle - \langle f_i, \mu\rangle) \prod_{l: l<i} \langle f_l, \mu\rangle \prod_{l: l>i} \langle f_l, \mu_3\rangle + O(M^{-1})$$

uniformly in $\mu \in \mathcal{P}(S)$.

On the other hand, by (2.2), (2.3), (2.4), and (2.5), (2.6), (2.7), it follows

$$\begin{aligned}\langle f, \mu_3 \rangle &= \left\langle f + \frac{1}{M} Af, \mu_2 \right\rangle + o(M^{-1}) \\ &= \left\langle \left(f + \frac{1}{M} Af \right) \circ \pi + \frac{1}{M} B \left(f + \frac{1}{M} Af \right), \mu_1 \right\rangle + o(M^{-1}) \\ &= \langle f, \mu \rangle + \frac{1}{M} \{ \langle Af, \mu \rangle + \langle Bf, \mu^2 \rangle \\ &\qquad + \langle (f \circ \pi)\sigma, \mu^2 \rangle - \langle f, \mu \rangle \langle \sigma, \mu^2 \rangle \} + o(M^{-1})\end{aligned}$$

for all f, uniformly in $\mu \in \mathcal{P}(S)$, where π is the projection of S^2 onto its first coordinate.

Therefore, letting

$$a_{f_i, f_j} = \beta_i \langle f_i f_j, \mu \rangle - \langle f_i, \mu \rangle \langle f_j, \mu \rangle \left(\sum_{k \in S} \beta_k \langle f_i, \mu \rangle - \beta_i - \beta_j \right)$$

and

$$b_{f_i} = \langle Af_i, \mu \rangle + \langle Bf_i, \mu^2 \rangle + \langle (f_i \circ \pi)\sigma, \mu^2 \rangle - \langle f_i, \mu \rangle \langle \sigma, \mu^2 \rangle,$$

we have

$$(\mathcal{L}_M \phi)(\mu) = (\mathcal{L}\phi)(\mu) + o(1)$$

uniformly in $\mu \in \mathcal{P}(S)$. □

Theorem 2 is generalized in the following

COROLLARY 3. *Suppose the conditions* (2.2), (2.3), *and* (2.4) *are satisfied and* ϕ *have the form*

$$\phi(\mu) = F(\langle f_1, \mu \rangle, \langle f_2, \mu \rangle, \cdots, \langle f_k, \mu \rangle) = F(\langle \mathbf{f}, \mu \rangle)$$

where $F \in C^2(\mathbf{R}^k)$. *Then there exist* a_{f_i, f_j}, $b_{f_i} \in \mathcal{B}(\mathcal{P}(S))$ *such that*

$$\lim_{M \to \infty} (\mathcal{L}_M \phi)(\mu) = \sum_{1 \le i < j \le k} a_{f_i, f_j} F_{z_i z_j}(\langle \mathbf{f}, \mu \rangle) + \sum_{i=1}^{k} b_{f_i} F_{z_i}(\langle \mathbf{f}, \mu \rangle)$$

uniformly in $\mu \in \mathcal{K}_M(S)$, *where* F_{z_i} *and* $F_{z_i z_j}$ *mean the partial derivative with respect to* i *and* i, j, *respectively.*

Proof. If we let

$$\phi(\mu) = F(\langle f_1, \mu\rangle, \langle f_2, \mu\rangle, \cdots, \langle f_k, \mu\rangle) = F(\langle \mathbf{f}, \mu\rangle)$$

$$a_{f_i,f_j} = \beta_i\langle f_i f_j, \mu\rangle - \langle f_i, \mu\rangle\langle f_j, \mu\rangle \left(\sum_{k\in S} \beta_k \langle f_i, \mu\rangle - \beta_i - \beta_j\right)$$

$$b_{f_i} = \langle Af_i, \mu\rangle + \langle Bf_i, \mu^2\rangle + \langle (f_i \circ \pi)\sigma, \mu^2\rangle - \langle f_i, \mu\rangle\langle \sigma, \mu^2\rangle,$$

this result is immediate from a second order Taylor expansion with Theorem 2. □

EXAMPLE. (From countable model to uncountably infinite model) We let the coefficients $\{b_i(p)\}$ have the form $b_i(p) = g_i(p) + h_i(p)$ in the case in which S is countable, where

$$g_i(p) = \sum_{k=1}^{d} p_k q_{ki},$$

$$q_{ij} \geq 0 \text{ for } i \neq j, \quad q_{ii} = -\sum_{j\neq i} q_{ij}.$$

For $i \neq j$, q_{ij} represents a rate of change of type X_i to type X_j.

For a countable number of types $X_1, X_2, \cdots$, a standard model for incorporating natural selection is to take

$$h_i(p) = p_i \left(\sum_{k\in S} p_k m_{ki} - \sum_{j,\ l\in S} p_j p_l m_{jl}\right). \tag{2.8}$$

The interpretation is that $X_1, X_2, \cdots$ are possible alleles carried by a gamete at some gene locus, and m_{ij} is a fitness coefficient of the genotype (X_i, X_j) such that $m_{ji} = m_{ij}$. If p_i is the frequency of type X_i, then $p_i p_j$ represents the frequency of (X_i, X_j).

Let m be a symmetric function on $C(S\times S)$. We denote $m(X_i, X_j)$ by m_{ij}. Let H be a function with the following property : given any finite set $\bar{S} = \{X_1, X_2, \cdots,\}$ of chromosome of type X_i, there exist $h_1, h_2, \cdots$ such that

$$H(\mu, \sigma) = \sum_{j\in\bar{S}} h_j(p)\sigma(X_j) \tag{2.9}$$

for each $\mu = \sum_{i\in\bar{S}} p_i \delta_{X_i}$ and $\sigma \in C(S)$.

From (2.8) and (2.9) we find that the correct choice for H is

$$H(\mu, \sigma) = \langle m\sigma, \mu\otimes\mu\rangle - \langle \sigma, \mu\rangle\langle m, \mu\otimes\mu\rangle \tag{2.10}$$

where $\mu \otimes \nu$ denote the product measure on $S \times S$. If $m_{ij} = \bar{m}_i + \bar{m}_j$, then (2.8) simplifies to

$$h_i(p) = p_i \left(\bar{m}_i - \sum_{j \in S_1} p_j \bar{m}_j \right).$$

Correspondingly, if $m(X_1, X_2) = \bar{m}(X_1) + \bar{m}(X_2)$, then (2.10) simplifies to

$$H(\mu, \sigma) = \langle \bar{m}\sigma, \mu \rangle - \langle \bar{m}, \mu \rangle \langle \sigma, \mu \rangle.$$

In order to apply Dawson's theorem [3], we need a continuous mapping $\mu \to f_\mu$ from $\mathcal{P}(s)$ to $C(S)$ such that

$$H(\mu, \alpha) = \langle f_\mu \alpha, \mu \rangle - \langle f_\mu, \mu \rangle \langle \alpha, \mu \rangle.$$

Therefore, we conclude that the model of Theorem 2 is with mutation selection and "mutation or gene conversion rate" of

$$\sum_{k \in S} \beta_k \langle f_i, \mu \rangle - \beta_i - \beta_j \text{ for every } i < j,$$

in the diploid models as described above.

ACKNOWLEDGEMENT. This paper was written during the author's stay at University of Iowa, U.S.A. I wish to thank the University of Iowa, in particular professor Jian Huang, for his support, friendship and hospitality.

References

[1] W. Choi and B. K. Lee, *On the diffusion processes and their applications in population genetics*, J. Appl. Math. Comput. **15** (2004), no. 1–2.

[2] W. Choi, *The application of stochastic analysis to countable allelic diffusion model*, Bull. Korean Math. Soc. **41** (2004), no. 2, 337–345.

[3] D. A. Dawson, *Geostochastic calculus*, Canad. J. Statist. **6** (1978), 143–168.

[4] S. N. Ethier and T. G. Kurtz, *The infinitely-many-alleles model with selection as a measure-valued diffusion*, in stochastic methods in biology, proceedings, Nagoya, Japan, 1985.

[5] J. H. Gillespie, *Natural selection for within-generation variance in offspring number,* Genetics, **76** (1974), 601–606.

[6] J. Roughgarden, *Theory of population genetics and evolutionary ecology : An Introduction*, Macmillan Publishing Co., 1979.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF INCHEON, INCHEON 402-749, KOREA
E-mail: choiwon@incheon.ac.kr