Korean J. Math. **27** (2019), No. 1, pp. 1–8
https://doi.org/10.11568/kjm.2019.27.1.1

ON THE REPRESENTATION OF PROBABILITY VECTOR WITH SPECIAL DIFFUSION OPERATOR USING THE MUTATION AND GENE CONVERSION RATE

Won Choi

Abstract. We will deal with an n locus model in which mutation and gene conversion are taken into consideration. Also random partitions of the number n determined by chromosomes with n loci should be investigated. The diffusion process describes the time evolution of distributions of the random partitions. In this paper, we find the probability of distribution of the diffusion process with special diffusion operator L_1 and we show that the average probability of genes at different loci on one chromosome can be described by the rate of gene frequency of mutation and gene conversion.

1. Introduction

Consider n locus model

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

so we find n genes on a chromosome. 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. If the partition X has α_i parts equal

Received August 20, 2018. Revised October 2, 2018. Accepted October 11, 2018.

2010 Mathematics Subject Classification: 92D10, 60H30, 60G44.

Key words and phrases: martingale problem, mutation and gene conversion, diffusion operator.

This research was supported by Incheon National University Research Grant, 2018-2019.

to i, then X describes that there exists α_i kinds of alleles occurring i loci for each i. The set of partitions of n with k parts is denoted by S_n. Let q_{ij} denote "mutation rate" or "gene conversion rate" from a partition X_i to another partition X_j per generation measured on the t time scale and p_i denotes 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

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

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

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

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. The operator L has the same form as the generator of the diffusion describing a $p(t)$-allele model incorporating mutation and random drift with single locus, but we could give a remark that the matrix q_{ij} depends on the combinatorial structure of the partitions.

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 process describes the time evolution of distributions of the random partitions. We will deal with an n locus model in which mutation and gene conversion are taken into consideration. Also random partitions of the number n determined by chromosomes with n loci should be investigated. In this paper, we show that the probability of types of chromosomes can be described by the function of rates of mutation and gene conversion.

2. Main Results

In order to consider an stochastic differential equation for $p(t)$, we need boundary conditions and regularity condition on the drift coefficients b_i.

[Assumption for $b_i(p)$] : $\{b_i(p)\}_{i\in S}$ is the set of real functions defined on K which satisfy the following conditions :

(i) $b_i(p) \geq 0$ if $p_i = 0$,
(ii) $\sum_{i\in S} b_i(p) = 0$ uniformly in $p \in K$,
(iii) there exists a matrix $\{c_{ij}\}_{i,j\in S}$ such that $c_{ij} \geq 0$ for every i and j of S, and

$$|b_i(p) - b_i(p')| \leq \sum_{j\in S} c_{ij}|p_j - p'_j|.$$

Suppose that $\{b_i(p)\}_{i\in S}$ satisfies the [Assumption for $b_i(p)$]. Then $p(t)$ is unique solution to stochastic differential equation

$$dp_i(t) = \sum_{k\in S} \alpha_{ik}(p(t))dB_k(t) + b_i(p(t))dt, \;\; i \in S$$

where

$$\alpha_{ij}(p) = (\delta_{ij} - p_i)\sqrt{\beta_j p_j}$$

and B_i are independent Brownian motions. Here $\{\beta_i\}$ is non-negative constant satisfying that $sup_i \beta_i < +\infty$

In order to construct the stochastic differential equation associated to mean vector, we need the following definition.

Definition. A sequence $\{X_1, X_2, \cdots, X_K, \cdots\}$ of partitions is called (X_1, X_K)-*chain* if X_{i+1} is a consequent of X_i by mutation or gene conversion for each $i = 1, 2, \cdots$.

We begin with the following Lemma.

Lemma 1.

$$\rho = \left(\frac{q_{12}}{q_{21}}\right)\left(\frac{q_{23}}{q_{32}}\right)\cdots\left(\frac{q_{K-1\ K}}{q_{K\ K-1}}\right) = 1.$$

Proof. See W. Choi ([2]). □

By Lemma 1, the value

$$\left(\frac{q_{12}}{q_{21}}\right)\left(\frac{q_{23}}{q_{32}}\right)\cdots\left(\frac{q_{K-1\ K}}{q_{K\ K-1}}\right)\cdots$$

does not depend on the choice of (X_1, X_K)-chain.

Let X be any partition of n and let $\{X_1, X_2, \cdots, X_i, \cdots\}$ be a $((n), X_i)$-chain. Put

$$P_X = \prod_{j=1}^{K-1}\left(\frac{q_{j\ j+1}}{q_{j+1\ j}}\right), \quad P_{(n)} = 1.$$

Let

$$K_1 = \{P = (P_X)_{X\in S_n} : \sum_{X\in S_n} P_X < +\infty\}$$

for the set S_n of partition and define a mapping $\bar{P}$ on K_1 called by probability vector

$$\bar{P}_X = \frac{P_X}{\sum_Y P_Y}.$$

Consider the solution to stochastic differential equation for $P_X(t)$

$$dP_X(t) = \sqrt{\beta_X P_X(t)}dB_X(t) + \tilde{b}_X(P(t))dt, \ \ i \in S \tag{1.1}$$

where

$$\tilde{b}_X(P(t)) = b_X(\bar{P}(t)) + c\bar{P}_X(t) + \bar{P}_X(t)(\beta_X - \sum_{X\in S_n}\bar{P}_X(t)\beta_X)$$

for a nonnegative constants c and β_X satisfying $c > (1/2)sup_{X\in S_n}\beta_X$.

It was shown easily that the existence and the uniqueness of solutions hold for the equation (1.1) when the set of drift coefficients $\{b_X(P)\}_{X\in S_n}$ satisfies the [Assumption for $b_X(P)$], not [Assumption for $\tilde{b}_X(P)$].([2])

Let L_1 be a second order differential operator on K_1

$$L_1 = \sum_{X,Y\in S_n}\tilde{a}_{XY}(P)\frac{\partial^2}{\partial P_X \partial P_Y} + \sum_{X\in S_n}\tilde{b}_X(P)\frac{\partial}{\partial P_X}$$

where

$$\tilde{a}_{XY} = \begin{cases}(\text{number of elements } S_n) \times \sqrt{\beta_X\beta_Y P_X(t)P_Y(t)} & \text{if } S_n \text{ is finite}\\ 0 & \text{if } S_n \text{ is infinite.}\end{cases}$$

W. Choi showed that the uniqueness of solution for the (K_1, L_1, P_0)-martingale problem holds.([3])

Consider a partition to be a sequence

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

If the partition X has α_i parts equal to i, then we write

$$X = [1^{\alpha_1}, 2^{\alpha_2}, \cdots, n^{\alpha_n}],$$

and let $\alpha(X) = \sum_i \alpha_i$. The probability that arbitrarily chosen k objects from n ones belong to the same kind is determined by the partition X. Therefore the probability is a function defined on S_n and denoted by G_{nk}. Assume that every mutation and gene conversion are new, and that each rate of gene frequencies is equal to $a(1/N)$ and $b(1/N)$ per generation, respectively. Here, N stands for the number of population and $1/N$ equals the time of one generation in diffusion models. Let the set $\{\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{id}\}$ be the collection of $\alpha_i > 0$. We suppose that firstly mutation occurs successively $\{\alpha(X) - 1\}$ times. that is,

$$X_1 = (n), X_2 = (n-1, 1), \cdots, X_{\alpha(X)} = (n - \alpha(X) + 1, 1, 1, \cdots, 1).$$

After that, gene conversion occurs successively $\{n - \alpha(X) - (i_d - 1)\}$ times.

Then we have;

Theorem 2. The vector $\bar{P}$ of stationary distribution of the diffusion process with operator L_1 can be written in the form

$$\bar{P}_X = \frac{n!}{\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)} \prod_{k=1}^{n} \frac{\theta^{\alpha_k}}{k^{\alpha_k}\alpha_k!}$$

where $\theta = a/b$.

Proof. The number of genes of each allele is increasing monotonously and the chain includes the partitions

$$(n - (\alpha(X) - 1) - (i_d - 1), i_d, 1, 1, \cdots, 1),$$

$$(n - (\alpha(X) - 1) - 2(i_d - 1)i_d, i_d, 1, 1, \cdots, 1),$$

$$\cdots$$

$$(n - (\alpha(X) - 1) - (\alpha_{i_d} - 1)(i_d - 1), i_d, \cdots, i_d, 1, 1, \cdots, 1),$$

$$(n - (\alpha(X) - 1) - (\alpha_{i_d} - 1)(i_d - 1) - \alpha_{i_{d-1}}(i_{d-1} - 1), i_d, \cdots, i_d, i_{d-1}, \cdots, i_{d-1}, 1, 1, \cdots, 1),$$

$$\cdots.$$

Since $\{X_1, X_2, \cdots, X_i, \cdots\}$ is a $((n), X_i)$-chain, the equalities

$$\prod_{j=1}^{K-1} q_{j\ j+1} = \frac{n!}{i_d!(i_d-1)!}\prod_{i=1}^{n}((i-1)!)^{\alpha_i}\frac{(\alpha(X)-1)!}{\alpha_i}a^{\alpha(X)-1}b^{n-\alpha(X)-i_d+1}$$

and

$$\prod_{j=1}^{K-1} q_{j+1\ j} = \frac{(n-1)!}{i_d!(i_d-1)!}\prod_{i=2}^{n}\alpha_i!\prod_{i=1}^{n}(i!)^{\alpha_i}(\alpha(X)-1)!\ b^{n-i_d}$$

hold. Therefore we can see that P_X is written as follows,

$$P_X = \frac{n}{\prod_{i=1}^{n}\alpha_i! i^{\alpha_i}}(\frac{a}{b})^{\alpha(X)-1}.$$

From the Riordan ([5]), we see that

$$\sum_X P_X = \frac{1}{(n-1)!}(\frac{a}{b}+1)(\frac{a}{b}+2)\cdots(\frac{a}{b}+n-1)$$

and

$$\bar{P}_X = \frac{n!}{\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)}\prod_{k=1}^{n}\frac{\theta^{\alpha_k}}{k^{\alpha_k}\alpha_k!}$$

□

Let a point $X_1 = (x_1, x_2, \cdots, x_k, \cdots) \in S_n$ be fixed and let $Y_1, Y_2, \cdots, Y_n$ be i.i.d. random variables such that $P(Y_k = j) = x_j,\ j = 1, 2, \cdots$. Let $\alpha_i(Y) = \alpha_i(Y_1, Y_2, \cdots, Y_n)$ be the cardinality of

$$\{j\ :\ i \text{ random variables of } Y_1,\ Y_2, \cdots, Y_n \text{ are equal to } j\}.$$

The partition

$$[1^{\alpha_1(Y)}, 2^{\alpha_2(Y)}, \cdots, n^{\alpha_n(Y)}]$$

induces a probability distribution P_{X_1}.

Then we meet with;

Theorem 3. The equality

$$\sum_{X\in S_n} P_{X1}G_{nk}(X) = \frac{b^{k-1}(k-1)!}{(a+b)(a+2b)\cdots(a+(k-1)b)}.$$

holds.

Proof. First, we note that

$$\sum_{X\in S_n} P_{X_1} G_{nk}(X)$$

is equal to probability that arbitrarily chosen k random variables $Y_{i_1}, Y_{i_2}, \cdots, Y_{i_k}$ from $Y_1, Y_2, \cdots, Y_n$ take the same values. Therefore we have

$$\begin{aligned}\sum_{X\in S_n} P_{X_1} G_{nk}(X) &= P(Y_1 = Y_2 = \cdots = Y_k)\\ &= \sum_{X\in S_k} P_{X_1} G_{kk}(X)\\ &= P_{X_1}((k)).\end{aligned}$$

By Theorem 2, last probability is

$$\frac{b^k k!}{a(a+b)(a+2b)\cdots(a+(k-1)b)} \cdot \frac{a/b}{k}$$

and we can see that Theorem 3 holds. □

Theorem 3 can be applied to many population and evolutionary genetic models. We conclude with the average probability of alleles which occupy x_1-loci, x_2-loci.

Example. The genes are contained in chromosomes. The existence of two alleles for given character, one inherited from each parent, parallels the existence of two chromosomes of each kind, also derived one from each parent. The two alleles for a character segregate in the formation of the gametes.([1]) The position that a gene has in a chromosome is known as its locus.
Letting $k = 2$ in Theorem 3, there exist alleles which occupy x_1-loci, x_2-loci and the average probability of genes at different loci on one chromosome is

$$\frac{b}{a+b}.$$

This means that the average probability of genes at different place on one chromosome is determined by the ratio of gene frequency of mutation and gene conversion and this probability is approximately the quotient of ratio of gene frequency of mutation and gene conversion.

We apply this theory to Mendel's experiment. In Mendel's experiment, Mendel studied the inheritance of seed shape by crossing plants yielding round seeds with plants yielding wrinkled seeds. Contrasting traits, such as the roundness or wrinkling of peas, are determined by gene. In this case, if a is the ratio of gene frequency of mutation and b is the ratio of frequency of gene conversion, then the probability of roundness or wrinkling at different place is

$$\frac{b}{a+b}.$$

If the rate of frequency of mutation approach 1, probability of genes at different place on one chromosome is approximately 1 and it is meant that genes at different place on one chromosome certainly occurs.

References

[1] F.J.Ayala, *Population and Evolutionary Genetics: A Primer*, The Benjamin/Cummings Publishing Company (1982),

[2] W. Choi, *The application of diffusion processes to population genetic model*, Bulletin of the Korean Mathematical Society **40** (4) (2003), 677–683.

[3] W.Choi and B.K.Lee *On the diffusion processes and their applications in population genetics*, J. Appl. Math. and computing. **15** (1–2) (2004), 415–423.

[4] W.Ewens *The sampling theory of selectively neutral alleles*, Theor. Pop. Biol. **3** (1972), 87–112.

[5] J.Riordan, *An introduction to combinatorial analysis*, John Wiley & Sons, (1958), 70–71.

Won Choi
Department of Mathematics
Incheon National University
Incheon 406-772, Republic of Korea
E-mail: choiwon@inu.ac.kr