Interacting Particle Systems for Random Genetic Drift

Yoshiaki Itoh, The Institute of Statistical Mathematics
and The Graduate University for Advanced Studies

1 Introduction

The diffusion approximation of a discrete model in population genetics is useful to get analytical solution. For many cases without analytical solution we use computer simulations by using stochastic difference equation for the diffusion approximation. However the discretization of time makes the trajectory go out of the boundary. We need some devices to solve this problem. The simulations by a dicrete system is sometimes useful. In section 2 we compare the two methods for the simulations of overdominance model in population genetics (Itoh (1984)). In section 3 to we introduce a model of speciation for which we could get an analytical solution (Itoh, Mallows, and Shepp (1998)).

2 Overdominance model

2.1 Stochastic difference equation

The method presented here is obtained by an approximate description of an interacting particle system for random genetic drift (Itoh (1973, 1979a, 1979b, 1984)), which is used by Maruyama and Nei (1982), Maruyama and Takahata (1981), Takahata (1981), Maruyama (1983), Nei et al.(1983), to discuss genetic variability maintained by mutation and overdominant selection in finite populations and is shown to be convenient. For the simulation studies, it is necessary to decompose a covariance matrix called drift matrix in population genetics. For general covariance matrix, Cholesky decomposition is usually used. For our case a stochastic model which has the same diffusion approximation to the original Fisher-Wright model automatically gives a decomposition of the drift matrix.

Our interacting particle model for overdominant selection gives a natural behaviour of trajectory at the boundary. We compared the heterozygosity obtained by our model with the result by Maruyama and Nei (1982), and found these two results agree well with each other.

In the Fisher-Wright model it is supposed that each of the genes of the next ${}_{\mathrm{m}}C_2$ generation is obtained by a random choice among the genes of the previous generation and that the whole population changes all at once. In Moran's model it is supposed that there are M individuals each formed from m alleles $A_1, A_2, \cdots, A_k$, and that at each instant at which the state of the model may change, one individual of the alleles,chosen at random,dies and is replaced by a new individual which is A_i with probability m_i/M, where m_i is the abundance of the allele A_i. It is supposed that the probability of any individual "dying" during an interval $(t, t+dt)$ and then being replaced by a new individual is λdt. Hence the mean number of such events in unit time is λM and the mean length of a generation is λ^{-1}. The following model is another reasonable one.

Consider a population of M particles each of which is one of k types, $A_1, A_2, \cdots, A_k$. The types may represent species, alleles, genotypes or other classification. We then consider interactions between particles,which are assumed to occur at the rate λdt per time interval $(t, t+dt)$ for each particle. If a pair of particles of different types i and j interact , then after the iteraction the both particles are the type i with probability $1/2$ and the type j with probability $1/2$. If the type of the interacting particles are the same, no change occurs. In this model two particles are chosen by random sampling without replacement at first, and from the two particles,two particles are chosen by random sampling with replacement.

We can approximate our model by a system of stochastic difference equations (1). In it, the relative abundance of type i increase by $\sqrt{x_i(t)x_j(t)}\Delta B_{ij}(t)$ by the interaction with j which results the decrease of the type j by $-\sqrt{x_i(t)x_j(t)}\Delta B_{ij}(t)$, where $c = \sqrt{\lambda/M}$. Hence our model automatically leads to the following equation (1), which has the drift matrix $c^2\{x_i(t)(\sigma_{ij} - x_j(t))\}\Delta t$ as covariances.

For $i, j = 1, 2, \cdots, m$, consider

$$\Delta x_i(t) = \sum_{i \neq j} c\sqrt{x_i(t)x_j(t)}\Delta B_{ij}(t), \qquad (1)$$

where $B_{ij}(t)(i > j)$ are mutually independent one dimensional normal random variable with the mean 0 and the variance t and $\Delta B_{ij}(t) = B_{ij}(t+\Delta) -$

$B_{ij}(t)$. Let

$$x_i(t+\Delta t) = x_i(t) + \Delta x_i(t) \quad \text{for } i = 1, 2, \cdots, m.$$

Then this difference scheme represent the random sampling drift of m alleles, $1, 2, \cdots, m$ whose relative abundances at time t are $x_1(t), x_2(t), \cdots, x_m(t)$ respectively.

Pederson (1973) gave a representation in which $x_i(t+\Delta t)$ is constructed from $x_i(t+\Delta t)$ for $j = 1, 2, \cdots, i-1$, and $x_j(x)$ for $j = 1, 2, \cdots, i$, as

$$x_i(t+\Delta t) = \frac{(1-\sum_{j=1}^{i-1} x_j(t+\Delta t))x_i(t)}{1-\sum_{j=1}^{i-1} x_j(t)} + c\Delta B_i(t)[x_i(t)(1-x_i(t))$$

$$-(\frac{x_i(t)}{1-\sum_{j=1}^{i-1} x_j(t)})^2 \sum_{j,k=1}^{i-1} x_j(t)(\delta_{jk} - x_k(t))]^{1/2}, \tag{2}$$

where $B_i(t), i = 1, 2, \cdots, m$, are mutually independent standard Bronian motion.

Our method requires ${}_{\mathrm{m}}C_2$ mutually independent normal random numbers for each step, while by the above method mutually independent m normal random numbers are used. The system of equations (1) is simple and the decomposision is explicitly given in it.

2.2 An interacting particle system

Here we introduce four-particle collision model to simulate overdominant selection model in population genetics.

Consider a random mating population of effective size N, and assume that selection and mutuation occur deterministically and that, after selection and mutuation, $2N$ gametes are randomly chosen for the next generation. If we assume that the fitness of heterozygotes is 1 for all pairs of alleles and $1-s$ for all homozygotes, and that every new mutation is different from the extent alleles. Then we have

$$E(\Delta x_i(t)) = 2Nx_i(t)\{-v + s(J - x_i(t))/(1-sJ)\}\Delta t$$

$$E(\Delta x_i(t)\Delta x_j(t)) = x_i(t)(\sigma_{ij} - x_j(t))(\Delta t) \tag{3}$$

By an appropriate scaling of time, where v is the mutation rate, $x_i(t)$ is the frequency of allele A_i at time t, $J = \sum_{i=1}^{n} x_i^2(t)$,and N is effective population size. The study of the allele frequency distribution in finite population for overdominance selection was initiated by Wright (1949). Maruyama and Nei (1982) used the form (1) (Itoh (1979b) to study various properties of overdominant selection in a finite population by computer simulations, simulating the stochastic differential equation with expectations and covariances given by equation (3).

In a population there are n particles of m types, $A_1, A_2, \cdots, A_m$.Consider the following four-particle random collision. Four particles are chosen from the population by random sampling without replacement, and let the four particles be A_i, A_j, A_k,and A_l. A_i and A_j from an individual A_iA_j, and A_k and A_l from A_kA_l.The A_iA_j and the A_kA_l collide and produce two A_iA_js with probability $1/2+s_{ij,kl}$ and two A_kA_ls with probability $1/2+s_{kl,ij}$, where $s_{ij,kl} = -s_{kl,ij}$, with

$$s_{ij,kl} = \begin{cases} s/2 & \text{if } i \neq j \text{ and } k = l \\ -s/2 & \text{if } i = j \text{ and } k \neq l \\ 0 & \text{if } i \neq j \text{ and } k \neq l \text{ or } i = j \text{ and } k = l, \end{cases}$$

and then the two A_iA_js (or the two A_iA_ls), split into two A_is and two A_js, (or two A_ks and two A_ls). Hence by the above collision A_i, A_j, A_k and A_l become two A_is and two A_js or two A_ks and two A_ls. We assume that a collision takes place in a time interval $[t, t + dt]$, with probability Cdt. Let the array of alleles frequencies be $\vec{X}= (X_1, X_2, \cdots, X_n)$ at time t. We caluculate the expectation $W(\Delta X_i(t))$ and covariance $E(\Delta X_i(t)\Delta X_j(t))$.

Hence we have approximately the following

$$E(\Delta x_\alpha) = \frac{4}{n} Csx_\alpha(\sum_k x_k^2 - x_\alpha)\Delta t$$

$$E(\Delta x_\alpha \Delta x_\beta) = \frac{4}{n^2} Cx_\alpha(\delta_{\alpha\beta} - x_\beta)\Delta t$$

where $x_\alpha = X_\alpha/n$, for $\alpha = 1, 2, \cdots, m$.

We assume that every mew mutation is different from the extant alleles and in a time unit Δt each of the n alleles is replaced by a mutant with probability $(4v/n)C\Delta t$. The four-particle collisions and mutations take place at random mutually independently. Hence we have

$$E(\Delta x_\alpha = \frac{4}{n} C x_\alpha \{-v + s(\sum_k x_k^2 - x_\alpha)\Delta t.$$

We choose $C = n^2/4$ and put $n = 2N$. We have

$$E(\Delta x_\alpha(t)) = 2N\{-v + s x_\alpha(\sum_k x_k^2 - x_\alpha)\}\Delta t$$

$$E(\Delta x_\alpha \Delta x_\beta) = x_\alpha(\delta_{\alpha\beta} - x_\beta)\Delta t$$

where $x_\alpha = X_\alpha/2N$ for $\alpha = 1, 2, \cdots, m$. The first equation is approximately equivalent with (4) when s is small. The variance caused by mutation is negligible. Hence we can use our random collision model as a simulation method for overdominance model when s is small.

A step consists of successive two stages. In the first stage a random collision of four particles takes place and in the next stage a mutation takes place with probability $4v$, that is to say, one of n particles is randomly chosen and replaced by a mutant with probability $4v$. We repeat this step one by one and take the time average $\bar{\mathrm{h}}$ of heterozygosity $h = 1 - \sum x_j^2$. Initially we set all of the n particles are of one type. We take the time average of heterozygosity over the last half duration of the total steps, that is, we take the time average from time $T/2$ to T, to get the average heterozygosity of the stationary state. Our results are compared with the results by Maruyama and Nei [10] in Table 1

Table 1. Comparison of values of heterozygosity, obtained by the two methods, stochastic differential equation and four-particle collision model (Itoh (1984))

2Ns	4Nv	$\bar{h}$		
		Stochastic differential equation Maruyama and Nei[10]	Random collision model n=2N=50 T=1,000,000	Random collision model n=2N=100 T=4,000,000
0	0.050	0.0476	0.0507	0.0457
5	0.0075	0.0450	0.0315	0.0295
0	0.100	0.0909	0.0838	0.0904
5	0.0175	0.0930	0.0879	0.1048
0	0.500	0.3333	0.3326	0.3276
5	0.150	0.3377	0.3642	0.3492
0	1.0	0.500	0.486	0.477
5	0.5	0.480	0.533	0.529
25	0.002	0.485	0.489	0.486
0	4.0	0.800	0.787	0.791
5	3.2	0.802	0.786	0.793
25	1.3	0.796	0.783	0.790

$\bar{h}$ for $2Ns = 0$ by Maruyama and Nei [10] are obtained by an analytical formula.

3 Model for speciation

(by Y. Itoh, C. Mallows and L. Shepp)

We introduce an analytical solution for a simple model of speciation (Itoh, Mallows, and Shepp (1998)). Suppose initially there are $N_i(0)$ particles at each vertex i of G, and that the particles interact to form a Markov chain: at each instant two particles are chosen at random, and if these are at *adjacent* vertices of G, one particle jumps to the other particle's vertex, each with probability 1/2. The process **N** enters a death state after a finite time when all the particles are in some *independent* subset of the vertices of G, i.e., a set of vertices with no edges between any two of them. The problem is to find the distribution of the death state, $\eta_i = N_i(\infty)$, as a function of $N_i(0)$.

We are able to obtain, for some special graphs, the *limiting* distribution of N_i if the total number of particles $N \to \infty$ in such a way that the fraction,

$N_i(0)/S = \xi_i$, at each vertex is held fixed as $N \to \infty$. In particular we can obtain the limit law for the graph, $S_2 :$ •——•——•, having 3 vertices and 2 edges.

For the complete graph, the model is that of Moran (1958) for the Fisher-Wright random sampling effect in population genetics. In the more general case the model might be applied to study speciation in biology as well as political positionings. For example consider a genetic system for m alleles $A_i, i = 1, 2, \cdots, m$, in which zygotes A_iA_j are fertile for $j = i-1, i, i+1$ and infertile for the other j. This problem was studied numerically by Nei, Maruyama and Wu (1983), considering the Fisher-Wright random sampling effect with some selection structure. Our present model has a random sampling effect depending on the structure of a graph, which could be a natural simplified model of the genetic problem. The graph R_{2k}, which is a regular polygon with $2k$ vertices and all edges present except those joining opposite vertices, is a special case of our genetic model.

Let G be any graph, and let $\sum\limits_{i \in G} \xi_i = 1$, $\xi_i \geq 0$, $i \in G$, be given. We will define $X_i(t)$, $i \in G$, $t \geq 0$, with $X_i(0) = \xi_i$, as the solution to the stochastic differential equation for $t \geq 0$,

$$dX_i = \sum_{j \in Ne_i} \sqrt{X_iX_j}dB_{ij}, \quad i \in G \tag{4}$$

where Ne_i is the set of neighbors of i in G, and B_{ij} are independent standard Wiener processes for *distinct* pairs $\{i, j\}$ and with the skew -symmetry property if the order is reversed,

$$B_{ji}(t) = -B_{ij}(t), \quad t \geq 0 \; .$$

Thus, it is clear that there exists a first time $\tau \geq 0$, for which $\{i : X_i(\tau) > 0\} = I(\tau)$ is an independent subset of G and $P(\tau < \infty) = 1$, i.e. the situation is the same for X as for N. Indeed if the total number of particles $N = \sum N_i(0)$ in the discrete process tends to infinity in such a way that $N_i(0)/N$ is held fixed for each $i \in G$, then the limiting process of $N(t)/N$ is X.

The probability of fixation can be obtained for the case for S_2. In this case we can use the resulting family of martingales to determine the limiting distribution explicitly. There is one martingale for each $n \geq 2$, given by

$$Y_n(t) = \sum_{i=1}^{n-1} \binom{n}{i}\binom{n-2}{i-1}(-1)^i X_1^i(t) X_2^{n-i}(t) \ . \tag{5}$$

We will use the martingale property:

$$EY_n(\tau) = EY_n(0) \tag{6}$$

to obtain the laws of $I(\tau)$ and $X(\tau)$ as a function of $\xi = X(0)$. With Y_n, define for any u the process

$$Z_u(t) = \sum_{n\geq 2} \frac{u^n}{n} Y_n(t), \quad t \geq 0 \ . \tag{7}$$

The following identity is valid for $|v| < 1/4$, $0 \leq x \leq 1$,

$$\sum_{n\geq 2} \frac{v^n}{n} \sum_{i=1}^{n-1} \binom{n}{i}\binom{n-2}{i-1}(-1)^i x^i (1-x)^{n-i} = xv + \frac{1-v}{2}(1-\sqrt{1+\frac{4xv}{(1-v)^2}}). \tag{8}$$

$$EZ_u(\tau) = \int_0^1 (xu + \frac{1-u}{2}(1-\sqrt{1+\frac{4xu}{(1-u)^2}}))\mu(dx) \ . \tag{9}$$

$$Z_u(0) = \xi_1 u + \frac{1-u(\xi_1+\xi_2)}{2}(1-\sqrt{1+\frac{4u\xi_1}{(1-u(\xi_1+\xi_2))^2}}). \tag{10}$$

where $\xi_i = X_i(0)$ and $\mu(dx) = P\{X_1(\tau) \in dx\}$.
From the analysis using the above two equations eq. (9) and eq.(10), we obtain

$$P\{X_1(\tau) = 0\} = \frac{1-\xi_1-\xi_2}{2}\left\{1 + \frac{1+\xi_1+\xi_2}{((1+\xi_1+\xi_2)^2 - 4\xi_2)^{1/2}}\right\} \ . \tag{11}$$

Obviously we have

$$P\{X_0(\tau) = 1\} = P\{X_1(\tau) = X_2(\tau) = 0\} = \xi_0 = 1 - \xi_1 - \xi_2 \tag{12}$$

By symmetry we have (interchanging 1 and 2) the point mass of μ at $x = 1$,

$$P\{X_1(\tau) = 1\} = \frac{1-\xi_1-\xi_2}{2}\left\{-1 + \frac{1+\xi_1+\xi_2}{((1+\xi_1+\xi_2)^2 - 4\xi_1)^{1/2}}\right\} \ . \tag{13}$$

The identity (8) has combinatorial meaning. It is obtained from an identy for generating function to enumerate plane unlabelled trees (Flajolet (1999)).

References

Flajolet, P. (1999) Personal communication.

Itoh, Y. (1973) On a ruin problem with interaction, Ann. Inst. Statist. Math., 25, 635-641.

Itoh, Y.(1979a). Random collision process on oriented graph, Research Memorandum No. 154, 1-20, The Institute of Statistical Mathematics, Tokyo.

Itoh, Y.(1979b). Random collision models in oriented graphs, J. Appl. Prob., 16, 36-44.

Itoh, Y. (1984) Random collision model for random genetic drift and stochastic difference equation, Ann. Inst. Statist. Math. 36, 353-362.

Itoh, Y., Mallows, C., and Shepp L. (1998) Explicit sufficient invariants for an interacting particle system, J. Appl. Prob., Vol 35, No.3.

Maruyama, T. (1983). Stochastic theory of population genetics, Bulletin of Mathematical Biology, 45, 521-554.

Maruyama, T. and Nei, M.(1982). Genetic variability maintained by mutation and overdominant selection in finite populations, Genetics, 98,441-459.

Maruyama, T. and Takahata, N.(1981). Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci, Heredity, 46, 49-57.

Nei, M.,Maruyama, T. and Wu, C. I.(1983). Models of evolution of reproductive isolation, Genetics, 103, 557-579.

Moran, P.A.P.(1958). Random processes in genetics, Proc. Phil. Soc.,54,60-71.

Pederson, D.G.(1973). An approximate method of sampling a multinomial population, Biometrics, 29, 814-821.

Takahata, N.(1981). Genetics variability and rate of gene substition in a finite population under mutation and fluctuating selection, Genetics, 98, 427-440.

Wright, S.(1949b). Adaptation and selection, Genetics, Paleontology and Evolution, (eds. G. L. Jepson, G. G. Simpson and T. Mayr), Princeton University Press, Princeton, New Jersey.