Heredity (1981), **46** (1), 49-57 0018-067X/81/02600049 $02.00

NUMERICAL STUDIES OF THE FREQUENCY TRAJECTORIES IN THE PROCESS OF FIXATION OF NULL GENES AT DUPLICATED LOCI*

TAKEO MARUYAMA AND NAOYUKI TAKAHATA
National Institute of Genetics, Mishima, 411 Japan

Received 7.vii.80

SUMMARY

A new numerical method, based on the stochastic differential equations, has been applied to the problem of the fixation of null alleles at duplicated loci. The results reported here were obtained under the assumption that individuals carrying homozygous null genes at both of the duplicated loci are lethal, but all the other genotypes are normal. The mean and median time for null alleles to fix can vary from $2N$ to $100N$ or more, depending on the mutation rate and the population size, where N is the population size. Linkage has strong effects on the fate of null genes in the populations, if $2Nc$ is of the order of 100 or less (c = recombination value). When the linkage is loose the frequencies of null genes in the course of fixation stay in a region where the frequency of the null gene is very low at one of the two duplicated loci. If the two loci are tightly linked, the fate of null genes at duplicated loci is determined by the mutation pressure and the random genetic drift, and selection plays a minor role. In these cases, the fixation time is invariably reduced and the trajectories can cover the entire region where the sum of null allele frequencies does not exceed unity. This makes the prediction that there should be rapid evolution of linked silent genes and pseudo-genes, and that silent DNA might be found close to functional genes.

1. INTRODUCTION

THERE is a growing interest in duplicated gene loci and their evolutionary significance. Although the subject has a rather long history of experimental and theoretical work, it is only in the past few years that molecular probes of duplicated loci and of their function have become feasible. Isozyme techniques, applied to a number of species of animals and plants, has revealed gene duplication to be a widespread phenomenon (see particularly Ferris and Whitt, 1979). This research has stimulated theoretical work, and several different models have been postulated in attempts to gain a deeper understanding of the subject. Some of the theoretical studies dealing with fish data are Bailey *et al.* (1978), Ferris *et al.* (1979), Kimura and King (1979), Takahata and Maruyama (1979), Allendorf (1979), and Li (1980).

Since it is generally accepted that duplicated gene loci play a vital role in evolution of genomes and in creation of new functional genes, the theoretical study of this problem is one of the currently important subjects in population genetics. In this paper, using Itô (1944)'s stochastic integrals we present a powerful, new method for treating the subject and some results obtained thereby.

Consider a panmictic population of finite size and two loci, which have resulted from duplication. They can be either linked or unlinked. If

* Contribution No. 1333 from the National Institute of Genetics, Mishima, Shizouka-ken, 411 Japan.

necessary the population size can vary in time. Random mating is not absolutely necessary, but it is assumed mainly for mathematical simplicity and can be easily modified if needed. Assume two alleles at each of the two loci, and denote by A and a the alleles at one locus, and by B and b the alleles at the other locus. We will call A and B normal (or functional) genes, and a and b null (nonfunctional) genes. These null alleles may consist of many molecularly different forms, but we refer to them collectively as null genes. As in other theoretical studies of this subject, we also assume that allele A mutates to a and B to b at a fixed rate ν, and that reverse mutation does not occur or can be ignored. We designate the population size by N, and the recombination value by c. We denote the frequencies of gametes ab, aB, Ab and AB by x, y, z and w respectively, and the frequencies of a and b by p $(=x+y)$ and $q(=x+z)$.

In our previous paper, assuming $c=0{\cdot}5$, we have examined mainly the time of fixation and the level of heterozygosity for various selection schemes, and have made an attempt to reveal the general relationship among different models (Takahata and Maruyama, 1979). Our main aim here is to investigate the "trajectories" of the two dimensional vector representing frequencies p and q of null alleles a and b, on their way to fixation. This has been studied for deterministic models by Kimura and King (1979), and also by Allendorf (1979).

In the present paper, we will study the case of the double recessive lethal in which genotype $aabb$ is lethal, but all others are normal. We are aware that this is only one of the various conceivable selection schemes for genes at duplicated loci. It is obvious that this selection scheme represents the weakest selection as long as we assume at least one normal gene necessary. Therefore, we believe that the trajectories of this model will have the least restriction compared with that of models where some genotypes other than $aabb$ have deleterious effects. Hence the study of the trajectories for the case of double recessive lethal described above will provide a kind of upper bound for the other models, and for this reason we have chosen this selection scheme in the present paper.

2. Mathematical method

Time development of the gamete frequencies $x(t)$, $y(t)$, and $z(t)$ forms a Markov process which can be approximated by a diffusion process. Then the process has many equivalent representations in stochastic differential equations (Watanabe, 1971). Stochastic differential equations can be regarded as a limit of difference equations. The case being dealt with here is given by the following stochastic difference equations: $(D=xw-yz)$

$$\left.\begin{aligned}
\Delta x(t) &= e_{11}B_1(\Delta t)+e_{12}B_2(\Delta t)+e_{13}B_3(\Delta t)\\
&\quad +2N[\nu(y+z)-\{x^2(1-x)+cD\}/(1-x^2)]\Delta t\\
\Delta y(t) &= e_{21}B_1(\Delta t)+e_{22}B_2(\Delta t)+e_{23}B_3(\Delta t)\\
&\quad +2N[\nu(w-y)+\{x^2y+cD\}/(1-x^2)]\Delta t\\
\Delta z(t) &= e_{31}B_1(\Delta t)+e_{32}B_2(\Delta t)+e_{33}B_3(\Delta t)\\
&\quad +2N[\nu(w-z)+\{x^2z+cD\}/(1-x^2)]\Delta t\\
\Delta w(t) &= -\{\Delta x(t)+\Delta y(t)+\Delta z(t)\}.
\end{aligned}\right\} \qquad (1)$$

where $B_i(\Delta t)$ are independent Gaussian distributions each having variance equal to Δt, and $[e_{ij}]$ is a non-negative definite square root of the matrix

$$\begin{bmatrix} x(1-x) & -xy & -xz \\ -xy & y(1-y) & -yz \\ -xz & -yz & z(1-z) \end{bmatrix}.$$

As indicated above, the diffusion process governed by equation (1) has many representation forms with respect to the terms involving $B_i(\Delta t)$. Itoh (1979) has given a formula which does not involve taking a square root matrix $[e_{ij}]$, and therefore greatly simplifies the calculations. It is also possible to represent the process using a set of coefficients given by Pederson (1973). However they are mathematically equivalent. In doing calculations, we have realized that Δt in the difference equations (1) has to be small for accuracy. We chose Δt so that the maximum of the terms $2N[\cdots]\Delta t$ is equal to 0·0005 or if this makes Δt greater than 0·001, $\Delta t = 0{\cdot}001$. At any rate, as Δt decreases, the simulation based on equation (1) converges uniformly to a continuous diffusion process governed by the stochastic differential equations derived as a limit of (1). In this sense our method gives approximations tending toward the correct solution, and it is possible to obtain an approximation of any required accuracy by letting Δt be small. For the mathematical theory, readers may refer to Storokhod (1965), Wong and Zakai (1965), and McShane (1974). There are several other simulation methods (Bailey *et al.* 1978; Kimura 1980).

For each case of a given set of parameter values, simulations were repeated 100 times or more. In other words, 100 or more independent populations were simulated, and the entire trajectory for each population was recorded. Every path started from the same initial condition $p = q = 0$ and ended as soon as p or q became 1.

3. Trajectory and fixation time

There are three different parameters to vary in this model of linked loci, *i.e.*, the population size (N), mutation rate (v), and recombination value (c). We tried to examine a wide range of values for each of the three parameters, and to reveal general features of the problem. Considering linkage, the case $c = 0$ represents one extreme, and we intend to see how the trajectories and the fixation time change as c becomes large and eventually 0·5 (free recombination). The simulation results on the trajectories are presented in fig. 1. The dark area shown in fig. 1 indicates the region where the path of the two dimensional vector representing the frequencies of null alleles passed through at least once in the fixation process repeated 100 times. In making the observation of path, the entire area of unit square was divided into 100 on each axis. We have made in several cases a considerably larger number of repeats, but found that results were about the same as those calculated from 100 repeats. We regard this as evidence indicating sufficiency of our simulations.

The mean fixation time and the median which is defined as the time required for 50 per cent of populations to be fixed with null alleles at one of the duplicated loci are given in table 1.

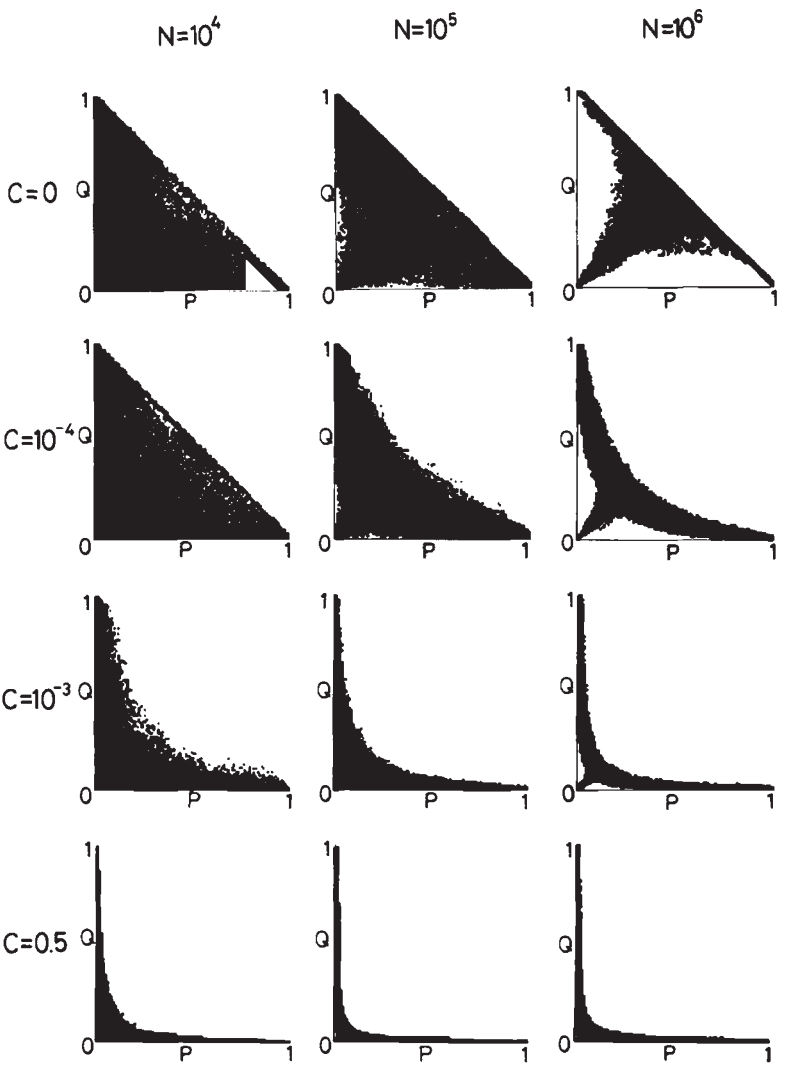

FIG. 1.—Dark areas indicate the regions where the trajectories of a two-dimensional vector representing the frequencies of null genes a and b have reached at least once in 100 repeats of independent sample paths (populations). In each diagram, p and q are the frequencies of a and b, respectively, and N = population size. In all the cases given in this figure the same mutation rate was used, $\nu = 10^{-6}$. In the figure, each horizontal row consisting of three diagrams represents cases of the same recombination value, and each vertical column represents cases of the same population size.

It is clear from the graphs in fig. 1 and the values in table 1 that linkage has a strong effect on these quantities. When $c = 0$ the frequencies of null genes at both loci can reach 0·5 simultaneously. Particularly if Nv is large, trajectories travel along the line $p = q$ to the point $p = q = 0{\cdot}5$, and then turn to either $p = 1$ or $q = 1$ along the line determined by $p + q = 1$, where p and q are the frequencies of a and b, *i.e.*, $p = x + y$ and $q = x + z$. The time required for the null genes to be fixed can be considerably faster than when

TABLE 1

Time required for 50 per cent of populations to reach the fixation of null alleles at one of the duplicated loci. (Numbers in parentheses are the average fixation times.) c, *the recombination value between the duplicated loci;* N, *population size;* v, *mutation rate from normal allele to null. The time is measured in units of* N *generations*

c	v	$N = 10^3$	10^4	10^5	10^6
0	10^{-5}	18·8 (25·0)	6·1 (9·5)	3·1 (3·3)	2·6 (3·3)
	10^{-6}	20·5 (29·6)	15·5 (21·3)	7·9 (9·4)	3·2 (3·6)
	10^{-7}	27·9 (38·7)	22·1 (31·7)	17·1 (24·2)	7·9(9·1)
0·0001	10^{-5}	16·8 (22·8)	8·4 (10·8)	5·1 (6·7)	5·8 (6·2)
	10^{-6}	28·0 (36·8)	17·9 (24·5)	11·6 (15·4)	8·8 (11·1)
	10^{-7}	30·6 (43·0)	20·5 (29·2)	20·9 (29·3)	15·4 (21·8)
0·001	10^{-5}	15·2 (20·9)	13·9 (16·4)	7·3 (10·6)	7·8 (9·0)
	10^{-6}	25·0 (36·9)	23·3 (31·4)	11·9 (15·7)	9·6 (13·6)
	10^{-7}	26·7 (43·5)	21·5 (33·9)	21·9 (29·3)	18·2 (25·4)
0·01	10^{-5}	18·2 (23·8)	16·6 (12·7)	10·0 (13·0)	9·2 (12·4)
	10^{-6}	21·5 (29·0)	17·4 (24·7)	14·2 (18·6)	14·1 (18·8)
	10^{-7}	24·2 (31·6)	25·7 (33·2)	23·4 (30·3)	33·3 (41·8)
0·1	10^{-5}	15·3 (22·9)	12·4 (16·2)	10·3 (14·2)	—
	10^{-6}	17·5 (22·3)	15·6 (24·1)	17·9 (22·0)	—
	10^{-7}	19·2 (27·1)	25·9 (36·7)	20·4 (28·6)	—
0·5	10^{-5}	22·7 (31·6)	11·4 (15·9)	10·8 (14·0)	14·8 (19·5)
	10^{-6}	31·7 (48·6)	20·6 (31·7)	23·0 (31·2)	27·5 (31·1)
	10^{-7}	44·2 (63·6)	30·7 (45·9)	31·2 (45·2)	31·1 (53·8)

$c = 0{\cdot}5$. If $Nc = 0$ and Nv is small, the trajectories can fill the entire region enclosed by lines $p = 0$, $q = 0$ and $p + q = 1$, and the reduction in the fixation time is not as drastic as that in cases with large Nv. Although we were not able to predict this before carrying out the simulations nor able to prove it mathematically, this seems intuitively plausible. Firstly, note that the gametes carrying only one null allele have no selective disadvantage. Therefore, mutation from the AB type to Ab or aB increases the frequency of a or b, while the steady loss of these gametes is from mutation to ab leading to lethal homozygotes $aabb$. This seems to imply that the frequencies of Ab and aB gametes can reach high values and both frequencies being very close to 0·5 is perhaps quite feasible, provided that recombination is absent or small so that the rate of producing ab gametes is not greater than that due to the mutation. If this argument is correct, selection will be very inefficient and the sum of the frequencies of Ab and aB gametes will be determined mostly by mutation and random genetic drift, as long as it stays less than unity. Under this hypothesis, it is easy to show that if Nv is small, the trajectories will cover the entire region surrounded by lines $p = 0$, $q = 0$ and $p + q = 1$, and if Nv is large, the trajectories will be restricted to the regions near lines determined by equations $p = q < 0{\cdot}5$ and $p + q = 1$. Since the selection is weak under these circumstances, the fixation time should be reduced. Particularly the reduction should be pronounced with large N for which the selection can act strongly against the fixation of null alleles if Nc and N are large.

As Nc values increase, the effect of linkage on the fate of null genes decreases gradually, and the processes become eventually cases with $c = 0{\cdot}5$ which are discussed below. Although it has not been possible to determine the transition pattern accurately, our simulation results appear to indicate

that in the present case of double recessive lethals, the linkage effect might be very small and insignificant if Nc is about 100 or larger, Li (1980).

When $c = 0{\cdot}5$, it is obvious that since the selection pressure is strong in a region where both p and q are large, a trajectory will be forced away from these areas and will stay in a region where either p or q is small. In other words, the trajectories will stay away from strong selection pressure and move to a region where it is relatively weak. From our experience with one-dimensional stochastic processes in population genetics, we conjectured that a trajectory rarely reaches areas where the selection pressure is large. Although processes in one and two dimensional space can be quite different, this seems to hold in our present cases. We could not obtain a satisfactory analytic solution to the problem, but we were able to reveal some of the interesting nature of the trajectories based on simulations. We have observed that all the trajectories stayed in a region where either p or q is very small. Three typical cases are presented in the bottom row of fig. 1 ($c = 0{\cdot}5$). As the mutation rate becomes small and population size becomes large, the permissible area for the populations becomes small and very close to one of the axes. Therefore, under the present model, and also probably under other models, if the two loci are not linked, the fixation of null alleles proceeds at only one of the duplicated loci, while the other remains almost entirely free of null genes. This is intuitively expected, and the present paper provides a quantitative confirmation.

4. Linkage disequilibrium

Since the formulation of the model in terms of the stochastic difference equation enables us to calculate the linkage disequilibrium between the two loci, we examined some of the characteristic features. We computed the commonly used disequilibrium measure, $D = (xw - yz)^2/pq(1-p)(1-q)$. Following the trajectory of each sample path, we measured the disequilibrium in two different ways. One was to measure the ratio D at every observation made at a time interval of $0{\cdot}1$, which is equivalent to $2N/10$ generations, and then calculate the average. The second was to measure the mean of $(xw - yz)^2$ and that of $pq(1-p)(1-q)$ separately, and then calculate the ratio of the two means. These measures of linkage disequilibrium were studied by Ohta and Kimura (1969) and Hill and Robertson (1968), in which these authors have obtained analytic solutions, in selectively neutral cases, for the second definition (see also Hill, 1974). We denote the former by σ_r^2 and the latter by σ_d^2.

The present study has shown that, as expected, the linkage disequilibrium is strong when Nc is small, particularly when $Nc = 0$, and it decreases gradually as Nc increases. As the value of Nc becomes 10 or larger the mean disequilibrium values become small and probably too small to be experimentally detectable. The values of σ_d^2 and σ_r^2 for the cases with $c \leqq 10^{-3}$ given in fig. 1 are presented in table 2. These values seem to represent some typical feature of the linkage disequilibrium associated with the process of the null allele fixation. We have also found that the values of σ_r^2 are often less than σ_d^2, and the ratio σ_r^2/σ_d^2 is approximately $0{\cdot}5$. However, the difference becomes small and ambiguous when the level of disequilibrium declines, as noted by Ohta and Kimura (1969).

TABLE 2

Linkage disequilibrium among the gametes carrying normal and null alleles at the duplicated loci. Letting x, y, z *and* w *be the frequencies of gametes* ab, aB, Ab *and* AB, $\sigma_r^2 \equiv$ *the average value of* $(xw-yz)^2/\{pq(1-p)(1-q)\}$ *and* $\sigma_d^2 \equiv$ *the average of* $(xw-yz)^2$ */the average of* $\{pq(1-p)(1-q)\}$, *where* $p = x+y$ (*frequency of* a) $q = x+z$ (*frequency of* b). *Cases given here correspond to those presented in fig.* 1, *excluding* $c = 0{\cdot}5$, *and thus* $v = 10^{-6}$. *Symbol* * *indicates a value which is not determined accurately, but the simulation suggested likely to be less than* 0·01.

N	10^4		10^5		10^6	
c	σ_d^2	σ_r^2	σ_d^2	σ_r^2	σ_d^2	σ_r^2
0	0·49	0·10	0·50	0·16	0·65	0·53
0·0001	0·20	0·07	0·05	0·02	0·02	0·01
0·001	0·04	0·05	*	*	*	*

The distribution of the disequilibrium values measured by σ_r^2 were also studied along the trajectories, and at several fixed values of p and q. We observed that the distribution of $(xw-yz)/\sqrt{pq(1-p)(1-q)}$ is rather strongly shifted toward negative values, particularly when $Nc = 0$.

5. REMARKS

Of course, the cases studied in the present paper are limited and thus may not reveal the entire nature of the linkage effect. However, when the selection is strong, the effect of linkage appears to be considerably larger than that in a system without selection pressure (Ohta and Kimura, 1971; Hill, 1975). As seen in table 1 and fig. 1, the fixation times and trajectories can be strongly affected by linkage only if Nc is of the order of 10 or less. At any rate, our simulation results have revealed an important fact that if the duplicated loci are linked, then the fixation of null alleles invariably occurs more rapidly.

Recent molecular study of genes suggests that unequal crossing over is often responsible for creating multigene families in which duplicated genes are tightly linked (Hood *et al.* 1975; Ohta, 1980). For instance, in the human globin genes, each of α and γ loci are duplicated, and the distance between the duplicated loci are about $3 \sim 4$ kilobases in both cases. Obviously, a distance of a few thousand nucleotide bases must mean a very tight linkage which might be of the same order as mutation rate. Interestingly, man, chimpanzee, gorilla and orangutan have similar duplications of α locus (Zimmer *et al.* 1980), and both of the human α loci are equally functional. These similar duplications in those primates may have resulted from independent unequal crossing overs after the species have separated, and thus the duplication of the human α is relatively recent. But another likely possibility is that the duplications have occurred prior to the species divergence. If the latter is the case, our results may suggest some pressure operating against null mutants even in heterozygotes, since these species are believed to have separated more than ten million years ago. On the other hand, there have been found several β-like sequences between the β and γ loci. These β-like sequences are most likely formed by changing duplicated genes into nonfunctional pseudogenes. Similar pseudogenes appear to exist in the neighborhood of the mouse α locus (Nishioka *et al.* 1980), and still other instances of similar situations have been reported (Fedoroff and Brown, 1978). Based on these observations and our results of simulation,

we would like to postulate a possible implication of the rapid fixation of null genes at closely linked, duplicated loci to evolution of silent genes and pseudogenes, which are usually functionless, but show strong homology and are linked to their normal genes. Since these silent genes and pseudogenes must have evolved from dispensable genes, thus existing in duplicated copies in a genome, the closer they are in linkage the easier it is for one of them to become nonfunctional. Therefore our finding appear to imply that there should be rapid evolution of linked silent genes and pseudogenes, and that silent DNA might be found close to functional genes.

Acknowledgements.—We are grateful to an anonymous referee, Drs J. F. Crow, Takeyuki Hida, T. Naglyaki, Tomoko Ohta and Akinobu Shimizu for their suggestions and comments, which contributed to substantial improvement of the paper. We also thank Mrs Sumiko Yano for her technical assistance.

6. References

ALLENDORF, F. W. 1979. Rapid loss of duplicate gene expression by natural selection. *Heredity, 43*, 247-258.

BAILEY, C. S., POULTER, R. T. M., AND STOCKWELL, P. A. 1978. Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. *Proc. Natl. Acad. Sci. U.S.A., 75*, 5575-5579.

FEDOROFF, N. V., AND BROWN, D. D. 1978. The nucleotide sequence of the repeating unit in oocyte 5s ribosomal DNA of *Xenopus laevis*. *Cold Spring Harbor Symposia* Vol. XLII, 1195-1200.

FERRIS, S. D., PORTNOY, S. L., AND WHITT, G. S. 1979. The role of speciation and divergence time in the loss of duplicate gene expression. *Theoret. Pop. Biol., 15*, 114-139.

FERRIS, S. D., AND WHITT, G. S. 1979. Evolution of the differential regulation of duplicate genes after polyploidization. *Journal of Molecular Evolution, 12*, 267-317.

HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity, 33*, 229-239.

HILL, W. G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoret. Pop. Biol., 8*, 117-126.

HILL, W. G., AND ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *Theoret. Pop. Biol., 38*, 226–231.

HOOD, L., CAMPBELL, J. H., AND ELGIN, S. C. R. 1975. The organization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics, 9*, 305-353.

ITÔ, K. 1944. Stochastic integral. *Proceedings of the Imperial Academy, 20*, 519-524.

ITOH. Y. 1979. Random collision process of oriented graph. *Institute of Statistical Mechanics* (Japan), *Research Memorandum, 154*.

KIMURA, M. 1980. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical and pseudo-sampling methods. *Proc. Natl. Acad. Sci. U.S.A., 77*, 522-526.

KIMURA, M., AND KING, J. L. 1979. Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proc. Natl. Acad. Sci. U.S.A., 76*, 2758-2861.

LI, W. H. 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics, 95*, 237-258.

MCSHANE, E. J. 1974. *Stochastic Calculus and Stochastic Models*. Academic Press, New York, San Francisco, London.

NISHIOKA, Y., LEDER, A., AND LEDER, P. 1980. An unusual alpha globin-like gene that has cleanly lost both globin intervening sequences. *Proc. Natl. Acad. Sci. U.S.A., 77*, 2806-2809.

OHTA, T. 1980. *Evolution and Variation of Multigene Families. Lecture Notes in Biomathematics* Vol. *37*, Springer-Verlag, Berlin, Heidelberg, New York.

OHTA, T., AND KIMURA, M. 1969. Linkage disequilibrium due to random genetic drift. *Genet. Res. Camb., 13*, 47-55.

OHTA, T., AND KIMURA, M. 1971. Linkage disequilibrium between two segregating nucleotide sites under steady flux of mutations in a finite population. *Genetics, 68,* 571-580.

PEDERSON, D. G. 1973. Note: An approximate method of sampling a multinomial population. *Biometrics, 29,* 814-821.

STOROKHOD, A. V. 1965. *Studies in the Theory of Random Processes.* Addison-Wesley Pubishing Company, Inc., Reading, Massachusetts.

TAKAHATA, N. AND MARUYAMA, T. 1979. Polymorphism and loss of duplicate gene expression: A theoretical study with application to tetraploid fish. *Proc. Natl. Acad. Sci. U.S.A., 76,* 4521-4525.

WATANABE, S. 1971. On stochastic differential equations for multi-dimensional diffusion processes with boundary conditions. *Journal of Mathematics of Kyoto University, 11,* 169-180.

WONG, E., AND ZAKAI, M. 1965. On the convergence of ordinary integrals to stochastic integrals. *Annals Mathematical Statistics, 36,* 1560-1564.

ZIMMER, E. A., MARTIN, S. L., BEVERLEY, S. M., KAN, Y. W., AND WILSON, A. C. 1980. Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc. Natl. Acad. Sci. U.S.A., 77,* 2158–2162.