J. Mol. Evol. 9, 313–322 (1977)

Journal of
Molecular Evolution

Simulation Results with Stepwise Mutation Model and Their Interpretations*

Ranajit Chakraborty

Center for Demographic and Population Genetics, University of Texas Health Science Center,
P.O. Box 20334, Houston, Texas 77025, USA

Summary. Monte Carlo simulations are performed to compare the predictions based on the two presently used theoretical models for studying genetic variations in natural populations, the infinite allele model and the stepwise mutation model. Distribution of heterozygosity is noticed to be similar under these models until the product of population size and mutation rate is large. It is seen that electromorphs with high population frequency usually contain older alleles (at the codon level) than an electromorph of low population frequency. The interpretations of these results in explaining the allelic variations at electrophoretic level is also discussed.

Key words: Electrophoresis – Heterozygosity – No. of alleles – Protein polymorphism – Simulation.

Introduction

Ohta and Kimura (1973) proposed the stepwise mutation model to explain the genetic variability in natural populations detectable by electrophoresis. Since then the mathematical properties of this model have been under extensive investigation (e.g., Nei and Chakraborty, 1973; Ohta and Kimura, 1974; Wehrhahn, 1975; Brown et al., 1975; Avery, 1975; Kimura and Ohta, 1975; Li, 1976; Moran, 1975; etc.). However, this new model turns out to be mathematically much less tractable than the infinite allele model, originally proposed by Wright (1948) and later developed extensively by Kimura and Crow (1964). Monte Carlo simulations of Ohta and Kimura (1974) have shown that the pattern of allelic distribution can be considerably different under the above two models. However, they were more concerned with the ratio of observed and effective number of alleles in a population maintained by mutation-drift balance.

In the present paper some more simulation results are presented which deal with other aspects of comparison between the two models. Specifically, the questions to which empirical solutions are sought here are:

* Research supported by U.S. Public Health Service General Research Support Grant 5 SO 5 RR 07148 from the University of Texas Health Science Center, Graduate School of Biomedical Sciences, Houston, Texas

1) What is the distribution of heterozygosity when variation is measured at the electrophoretic level?
2) What is the sampling behavior of the actual number of electromorphs in a sample?
3) Does the oldest allele belong always to the most frequent electromorphic class?

Similar problems are also studied by Ewens and Gillespie (1975) with the infinite allele model. Therefore, here the emphasis will be on the qualitative as well as quantitative differences between the answers to see the effect of electrophoretically silent mutations. In another paper we have elaborated the problem of occurrences of electrophoretically silent alleles by studying the average number of alleles (at the codon level) present per electromorph in a random sample of electromorphs of various frequency classes (Nei and Chakraborty, 1976). Ewens and Gillespie briefly discuss the effect of allelic undetectability. Nevertheless, the problem requires further elucidation since electrophoresis appears to be the only practical means presently available of studying genetic variability on a reasonably large scale.

The Simulation Method

The infinite allele model seems to be appropriate if allelic variants are identified at the nucleotide or amino acid (codon) level, wheras the stepwise mutation model refers to the allelic variation detectable by electrophoresis, at least as a rough approximation. The present simulation considers the genetic variation at the two levels simultaneously. More explicitly, I assume that at the codon level every mutation yields a novel allele so that allelic states are actually from an infinite allele-state-space. When intracistronic recombination is ignored, mutations can be assumed to be independently accumulating in different sites. Mutational inputs are taken as Poisson in the present set of simulations. Thus, if u denotes the mutation rate at the codon level per locus per generation, the probability that x mutations occur in a given cistron (locus) during a single generation is given by $e^{-u}u^{x}/x!$. To save computer time, high mutations rates are used (.002, .008, and .08) in the present simulation. However, the effective population sizes used were the same ($N = 50$) in all cases. At the electrophoretic level each allele was represented as one of the infinite series of mobility (electromorph in the terminology of King and Ohta, 1975) states (..., $-2, -1, 0, 1, 2$, ...) and it was assumed that each mutation results in a state (mobility) change of -1, 0, and 1 with probabilities β, α, and $\beta(\alpha + 2\beta = 1)$, respectively. The value of 2β has been estimated to be about 1/4, so that I used $\beta = 1/8$. Namely, each new mutation was assumed to change the electrophoretic mobility one step in the positive direction with probability 0.125, one step in the negative direction with probability 0.125, and with the remaining probability (0.75) it did not affect the mobility at all. All mutations were assumed to be selectively neutral in every set of simulations. After the introduction of mutations, 100 gametes (50 individuals) were sampled to produce the next generation. In the inital generation all gametes were identical and occupied the electromorph state 0. The process of mutational changes and sampling was continued until the equilibrium between mutation and genetic drift was attained. The process was carried out approximately for $10N/(4Nu + 1)$ generations, since the eventual rate of approach to the equilibrium heterozygosity is $2u + 1/2N$ (Nei and Li,

Table 1. Mean and variance of heterozygosity and actual number of alleles from three Monte Carlo experiments and their expectations under two models. For details of formulae used to compute the expectations see text

			Heterozygosity		Actual no. of alleles	
u			Mean	Variance	Mean	Variance
.002	Infinite	Observed	.2724	.0509	2.953	1.872
		Expected	.2857	.0500	2.866	1.704
	Stepwise	Observed	.0869	.0245	1.403	0.326
		Expected	.0871	.0238	1.444	[a]
.008	Infinite	Observed	.6195	.0299	6.645	4.454
		Expected	.6154	.0286	7.184	5.012
	Stepwise	Observed	.2614	.0441	2.198	0.638
		Expected	.2546	.0417	2.271	[a]
.08	Infinite	Observed	.9266	.0003	29.900	14.715
		Expected	.9412	.0003	32.132	17.838
	Stepwise	Observed	.6419	.0106	4.755	1.365
		Expected	.6667	.0110	4.808	[a]

[a] Expectations are not yet obtained theoretically

1976). To get a larger number of replications each run was allowed to proceed for 100 more generations beyond the above mentioned point. Thus, 200 replications of the entire set in fact yielded 400 replications of equilibrium gene (and electromorph) frequency distribution. The equilibrium status was tested by comparing the observed means and variances of the actual number of alleles and heterozygosities per locus with these expected under equilibrium (Ewens, 1972; Kimura and Ohta, 1975; Moran, 1975). The results are shown in Table 1 for the three simulations, each of which involved 400 replications. It is obvious from the table that there is no systematic bias in any of the three simulations. From the above description of the simulation, it is obvious that the present set of experiments is a combination of Ewens and Gillespie's (1974) and Ohta and Kimura's (1974) experiments. In the present case each mutation was given a new identification number and thus tracks were kept for each mutational event at each generation (even if it is lost due to drift from the population).

Distribution of Heterozygosity

It is well known that at steady state the expected heterozygosity under the infinite allele model is given by $M/(1 + M)$ where $M = 4Nu$ (Malecot, 1948; Kimura and Crow, 1964) whereas under the stepwise mutation model it is obtained as $1 - 1/\sqrt{1 + 4M\beta}$ (Ohta and Kimura, 1973). Furthermore, it has been shown recently (Stewart, 1976; Li and Nei, 1975) that the variance of single locus heterozygosities is given by

$$V(h) = \frac{2M}{(1 + M)^2 (2 + M)(3 + M)}$$

at steady state under the infinite allele model. A corresponding variance formula for the stepwise mutation model is obtained by Moran (1975). However, in the general case the theoretical distribution of h is difficult to obtain. In view of this, distributional

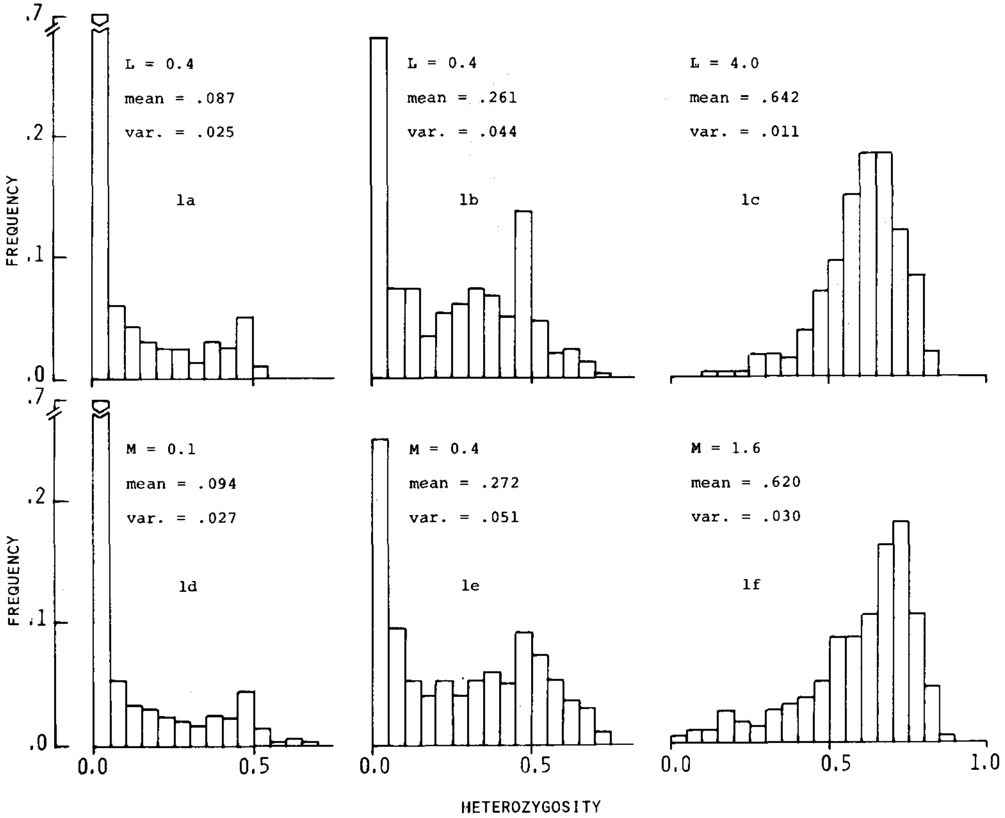

Fig. 1. Distribution of single locus heterozygosities under the stepwise mutation model (Fig. 1a–1c) and the infinite allele model (Fig. 1d–1f) for different values of the parameters as obtained from the Monte Carlo simulations

properties of heterozygosity (h) can now be obtained only empirically. Figure 1 demonstrates the nature of such distributions as seen in the present simulation. The upper histograms (Fig. 1a–c) represents the distribution of h under the stepwise mutation model for three different sets of values of the parameters ($L = 2M$, where $M = 4Nu$, $N = 50$, $\beta = 0.25$, $u = 0.002$, 0.008, and 0.08) whereas the lower diagrams (Fig. 1d–f) show the analogous distributions under the infinite allele model for comparable parametric values. As mentioned before, each diagram is based on 400 replications of the experiment. Figure 1d is from Nei et al. (1976) since comparable sets of parameters were not used in the present set of simulations. From these results it appears that for small values of M the distribution of h is largely L-shaped with a small peak around $h = 0.5$. The bimodality becomes more obvious as M increases. However, for very large M-values the distribution has shifted to one which is largely unimodal. There may additionally be a small peak at lower heterozygosity values which is more conspicuous under the infinite allele model. It is apparent that the two models do not differ significantly until M becomes large enough. In practice, however, the value of M seems to be generally about one or less since the average heterozygosity has been reported as 0.3 or less for all bisexual organisms so far studied (Nei, 1975; Selander, 1971). Therefore, the empirical distribution of Nei et al., obtained from several organisms, seems to be in conformity with the expectations under the stepwise mutation model as well. The bimodality of the distribution of h has also been observed by Ewens and Gillespie (1974) for an infinite allele model, who ascribed it to a mixture of distributions for different values of the actual number of alleles. Using a triallelic model, Stewart (1976) also predicted the existence of such bimodality from a theoretical study. Comparison of Figures 1c and 1f shows that under the stepwise mutation model the major peak shifts towards the left yielding a less prominent smaller peak at lower values of h. This results in a smaller average heterozygosity and a smaller variance as well. Therefore, Ewens and Gillespie's criticism of using h in the estimation of M becomes less severe if the stepwise mutation model is used instead of the infinite allele model.

Actual Number of Electromorphs in a Sample

The preceding section suggests that there is not much difference in the predicted polymorphic pattern between the infinite allele and the step mutation models as long as M is not very large. The biggest difference between the two, however, is in the actual number of alleles (or electromorphs) contained in the sample. Ewens (1972) discussed the sampling behavior of the actual number of alleles for the infinite allele model on the basis of which tests were based on the neutral mutation hypothesis. The mathematical manipulations were greatly facilitated by the Markovian nature of multinomial transition probabilities (Karlin and McGregor, 1967, 1972). Unfortunately, such a sampling theory does not apply to the stepwise mutation model because of the inherent feature that a random fraction of mutations is recurrent in this model, whereas in the infinite allele model every mutation is regarded to result in a new allele not pre-existing in the population. To overcome this difficulty Kimura and Ohta (1975) used the diffusion equation approach and obtained that the expected number of electromorphs in a random sample of s gametes ($s/2$ individuals) from an equilibrium population is given approximately by

$$n_a = \frac{L + L_a}{L_a} \left[1 - \prod_{i=0}^{s-1} \frac{i + L}{i + L + L_a}\right], \qquad (2)$$

where $L = 2M$, $L_a = 2Mb$, and

$$b = \frac{1 + L - \sqrt{1 + 2L}}{L(\sqrt{1 + 2L} - 1)/2} .$$

In the present set of simulations we generated the mean and variance of the actual number of electromorphs in a finite sample for which sampling (with replacement) was performed from distributions of electromorphs (after approximately $10N/(4Nu + 1)$ generations, as mentioned before). The details of the procedure used to obtain the observed number of electromorphs in a sample is given in the Appendix. From the distribution of each replication the average number of electromorphs were computed for several sample sizes and subsequently their mean and variances computed over

Table 2. Sample mean and variance of the number of electromorphs in samples of various sizes from 400 replications of each experiment. s denotes the number of gametes sampled.

	Mean			Variance		
s	$L = 0.1$	$L = 0.4$	$L = 4.0$	$L = 0.1$	$L = 0.4$	$L = 4.0$
10	1.238	1.718	3.261	.141	.298	.410
20	1.293	1.873	3.823	.191	.389	.676
30	1.322	1.951	4.083	.221	.437	.818
40	1.340	2.001	4.238	.241	.470	.912
50	1.352	2.036	4.342	.255	.494	.981
100	1.380	2.124	4.581	.292	.562	1.169
200	1.396	2.176	4.706	.316	.613	1.301
500	1.402	2.197	4.753	.326	.639	1.365

400 replications. The results are given in Table 2. It appears from this table (also from Table 1) that although (2) is an approximate formula for the mean number of electromorphs, it is quite accurate even if L is as large as 4.0. In practice, L seems to be low as mentioned before. Therefore, the approximate mean as obtained by Kimura and Ohta (1975) seems to be reasonable insofar as practical utility is concerned. The variance of this number seems to be high enough to be somewhat disturbing, however, in view of the fact that each electromorph must again be split into a number of alleles at the codon level. Such splitting has its own distribution with a mean and relatively large variance as well (for further discussion see Nei and Chakraborty, 1976; Chakraborty and Nei, 1976). A comparison of Table 2 with Tables 1–2 of Ewens (1972) shows that for larger L and s values the ratio of n_a/n_c (where $n_c = E(K)$), the expected number of alleles under the infinite allele model, (according to Ewens' terminology) decreases substantially. This is so because in such events the number of undetectable alleles is significantly greater for large sample size and higher mutation rate (or larger population size). It should be noted, hoewever, that the standard deviation of the number of alleles is also very large (Chakraborty and Nei, 1976). Thus, for a given value of L and s the number of alleles in an electromorph is expected to vary considerably among samples. In view of this, it is worthwhile to investigate the theoretical distribution of the number of electromorphs since it would elucidate the sampling behavior of electrophoretically silent alleles in a sample of given size.

Is the Oldest Allele in the most Frequent Electromorph?

Ewens and Gillespie (1974) tabulated the empirical probability that the oldest allele is also the most frequent one for various values of M. For the stepwise mutation model similar questions cannot be asked about electromorphs since several alleles may in fact constitute a single electromorph. Instead of devising a measure of the average age of an electromorph, a more relevant question can be framed as: How often does the oldest allele belong to the most frequent electromorphic class? As simple as this might seem, caution must be exercised in answering this question. As mentioned before, all of our simulations initially started with complete monomorphism. An equilibrium heterozygosity is obtained after approximately $10N/(4Nu + 1)$ generations. Insofar as the effect of the initial allele is concerned, however, we know that the loss of alleles

Table 3. Proportion (p) of cases (out of polymorphic ones) in which the oldest allele belonged to the most frequent electromorph (at electrophoretic level) and proportion (p) of events (out of polymorphic ones) in which the oldest allele was most frequent (at codon level) with their standard errors

	Codon level		Electrophoretic level		
u	Number of monomorphic replicates	p	Number of monomorphic replicates	p	Number of replicates
.002	33	.7365 ± .0341	127	.7945 ± .0473	200
.008	0	.5550 ± .0351	40	.7875 ± .0323	200
.08	0	.1983 ± .0163	0	.7633 ± .0174	600

proceeds at a rate of $2u$ per generation (Crow and Kimura, 1956). Therefore, not all replications were useful for a reliable answer to the present question. In the current set of simulations we used the data on generation 500 for $u = 0.002$, 300 for $u = 0.008$, and 50, 100, and 150 for $u = 0.08$. Table 3 exhibits the results based on these replications of $u = 0.002$ ($L = 0.1$), 127 resulted in monomorphism at the electrophoretic level, whereas only 40 replicates were monomorphic for $u = 0.008$ (L = 0.4). The empirical probabilities that the oldest allele belongs to the most frequent electromorph were of the same order for all L-values (heterogeneity χ^2 value is 0.68 with 2 d.f., $P > 0.70$). Furthermore, its appreciable value (nearly 80%) suggests that electromorphs with high population frequency will usually contain older alleles than an electromorph of low population frequency. This would appear to explain Nei and Chakraborty's (1976) findings that more frequently occurring electromorphs contain more alleles than the electromorphs with lower population frequency even if the sample size is the same. It may be argued that the small number of replications used in the present set of experiments is not sufficient to detect differences in the empirical probabilities for the three L-values. To investigate this I computed the empirical probabilities that the oldest allele (at the codon level) was actually the most frequent one as well. These results are also shown in Table 3. It is obvious that the probability sharply decreases as M (= $4Nu$) increases. The values obtained are also in conformity with the values obtained by Ewens and Gillespie (1974).

Discussion

Bulmer's (1971) observation that the most common allele at a locus within a species almost always occurs in the middle of the sequence has already been shown to be in agreement with expectations under the neutral mutation theory of Kimura (1968) when production of alleles is assumed to follow a stepwise model (Maynard-Smith, 1972; Kimura and Ohta, 1973). Furthermore, the present model is more suitable than the infinite allele model for explaining the observed evenness of the allelic frequency distribution (Ohta and Kimura, 1974). The statistic used for this considers the ratio of n_e/n_a, when n_e is the effective number of alleles (= $(1 - h)^{-1}$). I shall not repeat the discussion here since it has been exemplified by Ohta and Kimura's simulations. The only difference between their findings and the present one is in the standard deviation of n_a. From Table 2 we note that the standard deviation of n_a can be larger than one for large sample sizes when L is as large as 4.0. Even for $Nu \lesssim 0.1$ (where

n_a is seldom 3 or more, as seen in Ohta–Kimura's simulation) for large sample sizes n_a can be larger than 3 on an average. However, as far as the n_e/n_a ratio is concerned, the observed values are in the range of 0.45 ~ 0.65 (on average) even when $n_a < 3$ in the whole population. In this range it is in excellent agreement with Yamazaki and Maruyama's (1973) enzyme polymorphism analysis although Ohta and Kimura's predication in this respect was slightly higher. They, however, ascribed the difference to possible non-equlibrium status of populations and/or the presence of a number of slightly deleterious mutations. It may just be noted that bottlenecks in populations also affect the n_e/n_a ratio in a similar manner, although in the current set of simulations no such variation of population size is taken into consideration to illustrate such effects. For $n_a \geq 3$, Yamazaki and Maruyama's (1973) Figure 2 indicates that n_e/n_a is in the range of 0.2 ~ 0.45 which, however, is not in apparent agreement of the step mutation model. This is because in such events a large proportion of alleles are expected to be undetectable by electrophoresis as exemplified by Nei and Chakraborty (1976).

Acknowledgments. The author wishes to express his thanks to Dr. M. Nei and Dr. P. Fuerst for many stimulating discussions and for their comments on an earlier version of the paper.

Appendix

Observed Mean Number of Actual Electromorphs in a Sample of Given Size

In a particular replication the steady state distribution of electromorphs in the population is obtained by simulation as described in the text. To determine the number of distinct electromorphs observed in a sample drawn from this the following procedure is adopted.

Let there be k electromorphs with relative frequencies $p_1, \ldots, p_k$ in that replication. The expected number of different electromorphs chosen in a random sample of n of them from this particular population is given by

$$E_k = \sum_{r=0}^{k-1} \left[(\mathrm{k} - r) \sum_{S_r} \left(\frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{\,x_i} \right) \right] \tag{A1}$$

where, S_r is the set of all k-tuples such that exactly r of the x_i's are all zeros, and $\sum_{i=0}^{k} x_i = n$ ($x_i \geqslant 0$ for $i = 1, \ldots, k$).

From (A1), we readily have

$$E = k - \sum_{r=1}^{k-1} rP_r \tag{A2}$$

when $P_r = \sum_{S_r} (n! \prod_{i=1}^{k} p_i^{\,x_i} / \prod_{i=1}^{k} x_i!)$.

Using combinatorial arguments, the P_r's can be expressed as follows for a given value of k and n.

$$
\begin{aligned}
P_{k\text{-}1} &= Q_1 \\
P_{k\text{-}2} &= Q_2 - \binom{k\text{-}1}{1} Q_1 \\
P_{k\text{-}r} &= Q_r - \binom{k\text{-}r+1}{1} Q_{r\text{-}1} + \genfrac{}{}{0pt}{}{k\text{-}r+2}{2} Q_{r\text{-}2} + \ldots (-1)^{r\text{-}1} \binom{k\text{-}1}{r\text{-}1} Q_1 \\
P_1 &= Q_{k\text{-}1} - \binom{2}{1} Q_{k\text{-}2} + \ldots + (-1)^{k\text{-}2} \binom{k\text{-}1}{1} Q_1
\end{aligned}
$$

where $Q_r = \Sigma\, (p_{i_1} + \ldots + p_{i_r})^n$ in which the summation extends over all $\binom{n}{r}$ combinations of p_i's.

It may be noted that a similar approach may be followed to obtain the actual number of alleles (at the codon level) in an electromorph where p_i's are to be interpreted as the relative allelic frequencies (Nei and Chakraborty, 1976).

References

Avery, P.J. (1975). Genet. Res. **25**, 145–153

Brown, A.H.D., Marshall, D.R., Albrecht, L. (1975). Genet. Res. **25**, 137–143

Bulmer, M.G. (1971). Nature **234**, 410–411

Chakraborty, R., Nei, M. (1976). Genetics **84**, 385–393

Crow, J.F., Kimura, M.: Some genetic problems in natural populations. In: Proceedings of the Third Berkeley Symposium on Mathematics, Statistics, and Probability, Vol. IV, pp. 1–22. Berkeley: University of California Press 1956

Ewens, W.J. (1972). Pop. Biol. **3**, 87–112

Ewens, W.J., Gillespie, J.H. (1974). Theor. Pop. Biol. **6**, 35–57

Karlin, S., McGregor, J.L.: The number of mutant forms maintained in a population. In: Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics, and Probability, Vol. IV, pp. 415–438. Berkeley: University of California Press 1967

Karlin, S., McGregor, J.L. (1972). Theor. Pop. Biol. **3**, 113–116

Kimura, M. (1968). Genet. Res. **11**, 247–269

Kimura, M., Crow, J.F. (1964). Genetics **49**, 725–738

Kimura, M., Ohta, T. (1973). Genetics **73** (April Supplement), 19–35

Kimura, M., Ohta, T. (1975). Proc. Nat. Acad. Sci. **72**, 2761–2764

King, J.L., Ohta, T. (1975). Genetics **79**, 681–691

Li, W-H. (1976). Genet. Res., 28, 119–127

Li, W-H., Nei, M. (1975). Genet. Res. **25**, 229–248

Malecot, G.: Les Mathematiques de l'Heredite. Paris: Masson 1948

Maynard-Smith, J. (1972). Nat. New Biol. **237**, 31

Moran, P.A.P. (1975). Theor. Pop. Biol. **9**, 318–330

Nei, M.: Molecular Population Genetic and Evolution. Amsterdam: North-Holland 1975

Nei, M., Chakraborty, R. (1973) J. Mol. Evol. **2**, 323–328

Nei, M., Chakraborty, R. (1976). J. Mol. Evol. **8**, 381–385

Nei, M., Fuerst, P., Chakraborty, R. (1976). Nature **262**, 491–493

Nei, M., Li, W-H. (1976). Genet. Res., 28, 205–214

Ohta, T., Kimura, M. (1973). Genet. Res. **22**, 201–204

Ohta, T., Kimura, M. (1974). Genetics **76**, 615–624

Selander, R.K.: Stochastic factors in the genetic structure of populations. In: Molecular Study of Biological Evolution, Ayala, F.J. (ed.). Sunderland, Mass: Sinauer 1976

Stewart, F.M. (1976). Theor. Pop. Biol. **9**, 188–201

Wehrhahn, C.F. (1975). Genetics **80**, 375–394

Wright, S. (1948). Genetics of populations. Encyclopaedia Britannica, 14 ed. 10. 111, 111A–D, 112

Yamazaki, T., Maruyama, T. (1973). Nat. New Biol. **245**, 140–141

Received November 13, 1976; Revised Version February 22, 1977