Genet. Res., Camb. (1976), **28**, *pp.* 205–214
With 3 text-figures
Printed in Great Britain

The transient distribution of allele frequencies under mutation pressure

BY MASATOSHI NEI AND WEN-HSIUNG LI

Center for Demographic and Population Genetics, University of Texas at Houston, Texas 77030

(*Received* 28 *November* 1975)

SUMMARY

The transient distribution of allele frequencies in a finite population is derived under the assumption that there are k possible allelic states at a locus and mutation occurs in all directions. At steady state this distribution becomes identical with the distribution obtained by Wright, Kimura and Crow when $k = \infty$. The rate of approach to the steady state distribution is generally very slow, the asymptotic rate being $2v + 1/(2N)$, where v and N are the mutation rate and effective population size, respectively. Using this distribution it is shown that when population size is suddenly increased, the expected number of alleles increases more rapidly than the expected heterozygosity. Implications of the present study on testing hypotheses for the maintenance of genetic variability in populations are discussed.

Wright (1949) and Kimura & Crow (1964) independently derived a formula for the expected number of neutral alleles with a given gene frequency under the assumptions that (1) there are an infinite number of possible alleles at a locus, and (2) the effects of mutation and random genetic drift are balanced. The first of these assumptions seems to be roughly correct if allele differences are studied at the nucleotide or codon level. The second assumption, however, does not always hold, since the size of a species or Mendelian population often changes drastically in the evolutionary process, and once the equilibrium is disturbed, it takes a long time for the new equilibrium to be attained.

In the present paper we shall remove the second assumption and derive a formula for the transient distribution of allele frequencies. This formula seems to be important in the study of the mechanism of maintenance of protein polymorphism and molecular evolution.

Distribution of allele frequencies

We intend to derive a formula for the expected number of selectively neutral alleles in transient states whose frequencies are from x to $x + \mathrm{d}x$. Following Wright (1949) and Kimura (1968), we assume that there are k possible alleles at a locus and each allele mutates to one of the $k-1$ remaining alleles with a frequency of $v/(k-1)$ per generation. Therefore, the mutation rate per gene per generation is v. We consider a randomly mating population of effective size N. Let x be the frequency of a particular allele and $\phi(p, x; t)$ be the distribution of gene frequency x

at time t, given the initial gene frequency p. The mean ($M_{\delta x}$) and variance ($V_{\delta x}$) of the gene frequency change per generation are given by

$$M_{\delta x} = -vx + v_1(1-x), \tag{1}$$

$$V_{\delta x} = x(1-x)/(2N), \tag{2}$$

where $v_1 = v/(k-1)$. Kimura (Crow & Kimura, 1956) studied the frequency distribution of an allele under mutation pressure [$\phi(p, x; t)$], and in the present case it is given by

$$\phi(p, x; t) = \sum_{i=0}^{\infty} X_i(x)\, e^{-\lambda_i t}, \tag{3}$$

where $\lambda_i = i(A+i-1)/(4N)$, and

$$X_i(x) = x^{B-1}(1-x)^{A-B-1}F(A+i-1, -i, A-B, 1-x)$$

$$\times F(A+i-1, -i, A-B, 1-p)\frac{\Gamma(A-B+i)\Gamma(A+2i)\Gamma(A+i-1)}{i!\Gamma^2(A-B)\Gamma(B+i)\Gamma(A+2i-1)}, \tag{4}$$

in which $M = 4Nv$, $A = kM/(k-1)$, and $B = M/(k-1)$.

We have assumed that there are k possible alleles, so that the expected number of alleles whose frequencies are from x to $x+dx$ is given by

$$\Phi_k(x, t)dx = \sum_{j=1}^{k} \phi(p_j, x; t)dx, \tag{5}$$

where p_j is the initial gene frequency of the jth allele. As mentioned earlier, the number of possible alleles at a locus is almost infinite at the nucleotide or codon level. Therefore, in most cases we may put $k \to \infty$. We denote $\Phi_\infty(x, t)$ by $\Phi(x, t)$. In the following we consider two specific cases in detail.

Case 1: The first case is that where there is no genetic variability in the initial population. In this case $p = 1$ for a particular allele and $p = 0$ for all other alleles. Therefore, the distribution of allele frequencies is given by

$$\Phi(x, t) = \lim_{k\to\infty} [(k-1)\phi(0, x; t) + \phi(1, x; t)]. \tag{6}$$

To evaluate the first term in (6), we have to know $\lim_{k\to\infty}(k-1)X_i(x)$ for $p = 0$. Namely,

$$\lim_{k\to\infty}(k-1)X_i(x) = \lim_{k\to\infty}(k-1)x^{B-1}(1-x)^{A-B-1}F(A+i-1, -i, A-B, 1-x)$$

$$\times F(A+i-1, -i, A-B, 1)\frac{\Gamma(A-B+i)\Gamma(A+2i)\Gamma(A+i-1)}{i!\Gamma^2(A-B)\Gamma(B+i)\Gamma(A+2i-1)}.$$

If we note $(k-1)B = M$, $A-B = M$, $\lim_{k\to\infty}(k-1)/\Gamma(B) = M$, and

$$F(A+i-1, -i, A-B, 1) = (-1)^i \frac{\Gamma(B+i)\Gamma(A-B)}{\Gamma(B)\Gamma(A-B+i)},$$

we obtain

$$\lim_{k\to\infty}(k-1)X_i(x) = (-1)^i M x^{-1}(1-x)^{M-1}F(M+i-1, -i, M, 1-x)$$

$$\times \frac{(M+2i-1)\Gamma(M+i-1)}{i!\,\Gamma(M)}.$$

https://doi.org/10.1017/S0016672300016906 Published online by Cambridge University Press

On the other hand, to evaluate the second term in (6), we have to know $\lim_{k\to\infty} X_i(x)$ for $p = 1$. For $i = 0$,

$$\lim_{k\to\infty} X_0(x) = x^{-1}(1-x)^{M-1}F(M-1, 0, M, 1-x) \times F(M-1, 0, M, 0)\lim_{k\to\infty}\frac{\Gamma(A)}{\Gamma(M)\Gamma(B)} = 0.$$

For $i \geqslant 1$,

$$\lim_{k\to\infty} X_i(x) = x^{-1}(1-x)^{M-1}F(M+i-1, -i, M, 1-x) \times \frac{(M+2i-1)\Gamma(M+i-1)\Gamma(M+i)}{i!\,\Gamma^2(M)\Gamma(i)}.$$

Therefore, $\Phi(x, t)$ is given by

$$\Phi(x;t) = Mx^{-1}(1-x)^{M-1} + \sum_{i=1}^{\infty} x^{-1}(1-x)^{M-1}F(M+i-1,-i, M, 1-x) \times \frac{(M+2i-1)\Gamma(M+i-1)}{i!\,\Gamma(M)}\left[(-1)^i M + \frac{\Gamma(M+i)}{\Gamma(i)\Gamma(M)}\right]e^{-\lambda_i t}. \quad (7)$$

It is clear that at $t = \infty$, $\Phi(x; t)$ becomes identical with Kimura & Crow's formula $\Phi(x) = Mx^{-1}(1-x)^{M-1}$.

In general, the asymptotic rate of approach to the steady state distribution is given by the smallest eigenvalue, i.e. λ_1. In the present case, however, the coefficient of $e^{-\lambda_1 t}$ becomes 0. Therefore, the asymptotic rate is given by

$$\lambda_2 = 2v + 1/(2N).$$

The reason why λ_1 drops out is that we are considering the configuration of allele frequencies rather than a particular allele frequency. This result agrees with that obtained by Ewens & Kirby (1975) and Karlin & Avni (1975) (see also Ewens & Gillespie, 1974) in their studies of the eigenvalues of the configuration process with the discrete time model.

In passing we note that $\phi(1, x; t)$ for $k \to \infty$, which is given by

$$\phi(1, x; t) = \sum_{i=1}^{\infty} x^{-1}(1-x)^{M-1}F(M+i-1, -i, M, 1-x) \times \frac{(M+2i-1)\Gamma(M+i-1)\Gamma(M+i)}{i!\,\Gamma^2(M)\Gamma(i)} e^{-\lambda_i t}, \quad (8)$$

is identical with formula (2) in Nei & Li (1975) (see also Crow & Kimura, 1970). This identity can be easily shown if we note the relationship

$$F(M+i-1, -i, M, 1-x) = xF(-i+1, M+i, M, 1-x).$$

The above formula gives the distribution of the allele that was present in the original population. Note that the rate of steady decay for this distribution is given by $\lambda_1 = v$ rather than λ_2. This is because we are considering a particular allele in this case. Ewens & Gillespie (1974) have called λ_1 a labelling eigenvalue.

In the derivation of formula (7) no consideration was made about the fact that the sum of frequencies over all alleles in a population is unity. This is because the initial condition for the gene frequency is sufficient to determine equation (3)

uniquely (cf. Crow & Kimura, 1970, p. 441). Therefore, if we consider the initial conditions for all possible alleles, the sum of gene frequencies should be 1 at any generation. In fact, it can be shown that $\int_0^1 x\Phi(x, t)\,dx$ is always 1.

Case 2: Let us now consider the case in which the initial population is in equilibrium with a given value of $4Nv$, i.e. M_0, and then because of the change in population size or mutation rate $4Nv$ becomes M. In practice, of course, population size would change gradually rather than suddenly in a single generation. However, the change in population size is generally much quicker than that of genetic variability, so that the assumption of sudden change in population size would not affect our final result appreciably. Let $\Phi_k(p)$ be the stationary distribution of allele frequencies for $4Nv = M_0$ when the number of possible alleles is k. Then,

$$\Phi_k(p) = \frac{k\Gamma(A_0)}{\Gamma(A_0-B_0)\Gamma(B_0)}(1-p)^{A_0-B_0-1}p^{B_0-1}, \tag{9}$$

where $A_0 = kM_0/(k-1)$ and $B_0 = M_0/(k-1)$ (Kimura, 1968). Therefore, for a given value of k, the distribution of allele frequencies in the tth generation is given by

$$\Phi_k(x, t) = \int_0^1 \phi(p, x; t)\,\Phi_k(p)\,dp. \tag{10}$$

If we note

$$\int_0^1 F(A+i-1, -i, A-B, 1-p)_k(p)\,dp\,\Phi = k\sum_{n=0}^{i}\frac{(A+i-1)_n(-i)_n}{(A-B)_n n!}\frac{(A_0-B_0)_n}{(A_0)_n},$$

where $(a)_n = a(a+1), \ldots, (a+n-1)$, then $\Phi_k(x, t)$ is given by

$$\Phi_k(x, t) = k\sum_{i=0}^{\infty} x^{B-1}(1-x)^{A-B-1}F(A+i-1, -i, A-B, 1-x)$$
$$\times \sum_{n=0}^{i}\frac{(A+i-1)_n(A_0-B_0)_n(-i)_n}{(A-B)_n(A_0)_n n!}\frac{\Gamma(A-B+i)\Gamma(A+2i)\Gamma(A+i-1)}{i!\,\Gamma^2(A-B)\Gamma(B+i)\Gamma(A+2i-1)}e^{-\lambda_i t}. \tag{11}$$

It can be shown that if $A = A_0$ and $B = B_0$, then $\Phi_k(x, t) = \Phi_k(x)$, as expected.

The distribution $\Phi(x, t)$ can be obtained by putting $k \to \infty$ in $\Phi_k(x, t)$. The first term in (11), corresponding to $i = 0$, becomes $Mx^{-1}(1-x)^{M-1}$, as expected. For $i \geqslant 1$, we note that

$$\lim_{k\to\infty} k\sum_{n=0}^{i}\frac{(A+i-1)_n(A_0-B_0)_n(-i)_n}{(A-B)_n(A_0)_n n!} = \sum_{n=1}^{i}\frac{(-i)_n(M+i-1)_n}{(M)_n n!}$$
$$\times \sum_{j=0}^{n-1}\left(\frac{M}{M+i-1+j} - \frac{M_0}{M_0+j}\right),$$

which is 0 if $i = 1$. Therefore, we have

$$\Phi(x, t) = Mx^{-1}(1-x)^{M-1} + \sum_{i=2}^{\infty} x^{-1}(1-x)^{M-1}F(M+i-1, -i, M, 1-x)$$
$$\times \sum_{n=1}^{i}\frac{(-i)_n(M+i-1)_n}{(M)_n n!}\sum_{j=0}^{n-1}\left(\frac{M}{M+i-1+j} - \frac{M_0}{M_0+j}\right)$$
$$\times \frac{(M+2i-1)\Gamma(M+i)\Gamma(M+i-1)}{i!\,(i-1)!\,\Gamma^2(M)}e^{-\lambda_i t}. \tag{12}$$

https://doi.org/10.1017/S0016672300016906 Published online by Cambridge University Press

It can be shown that for $M_0 = M$ or for $t = \infty$, $\Phi(x, t) = Mx^{-1}(1-x)^{M-1}$, as expected. Note that the asymptotic rate of approach to equilibrium is again given by λ_2, and $\int_0^1 x\Phi(x, t)\mathrm{d}x = 1$ for all t's.

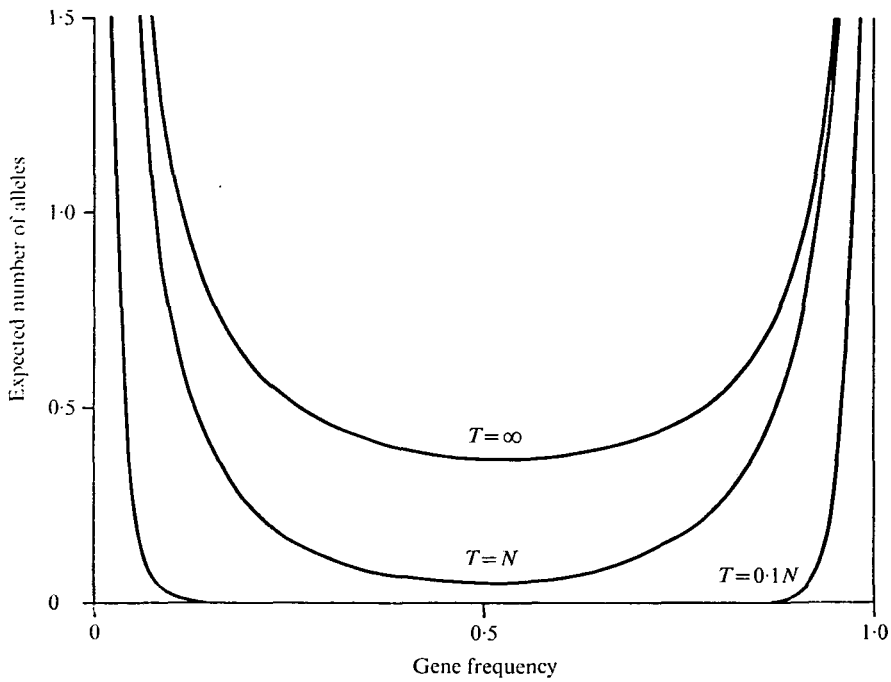

Fig. 1. Transient distributions of allele frequencies with $4Nv = 0{\cdot}1$. The initial population was assumed to be monomorphic for a particular allele. T: time in the unit of $2N$ generations.

Some of the numerical results obtained by using (7) and (12) are given in Figs 1–3. Fig. 1 refers to the case where the initial population was completely homozygous and $M = 0{\cdot}1$. (We note that in many natural populations the value of M is about 0·1; Nei, 1975.) In the early generations ($t = 0{\cdot}1N$) allele frequencies are concentrated either near 0 or near 1. This is because most new mutant alleles exist in low frequency, while the frequency of the original allele remains high. The number of alleles whose frequency is close to 0·5, however, increases gradually, and by generation $t = N$ it becomes about one-sixth of the value at steady state. At $t = 10N$ generations the distribution of allele frequencies becomes almost indistinguishable from that of the steady state in this case. Fig. 2 refers to the case where the initial population was in equilibrium with $M_0 = 0{\cdot}1$ and later the value of M has increased 10 times. In this case, the expected number of alleles whose frequency is close to 0 gradually increases due to accumulation of new mutations, whereas the expected number of alleles whose frequency is close to 1 declines. With $M = 1$, the steady state distribution of allele frequencies becomes inverse-J shaped. If a locus is defined as monomorphic when the frequency of the most

https://doi.org/10.1017/S0016672300016906 Published online by Cambridge University Press

common allele is 0·99 or more, then the probability of monomorphism is only 0·01 at $t = \infty$ (cf. Kimura, 1971). Fig. 3 refers to the case where $M_0 = 0{\cdot}05$ and $M = 5$. In this case, because of the large value of M, new mutations are rapidly accumulated and the probability of monomorphism quickly declines. At $t = 0{\cdot}1N$ there arises a peak in the distribution around the gene frequency equal to 0·9. This is because the originally monomorphic alleles still have a high gene frequency. In the present case the distribution of allele frequencies becomes almost indistinguishable from that of the steady state by $t = 2N$ generations.

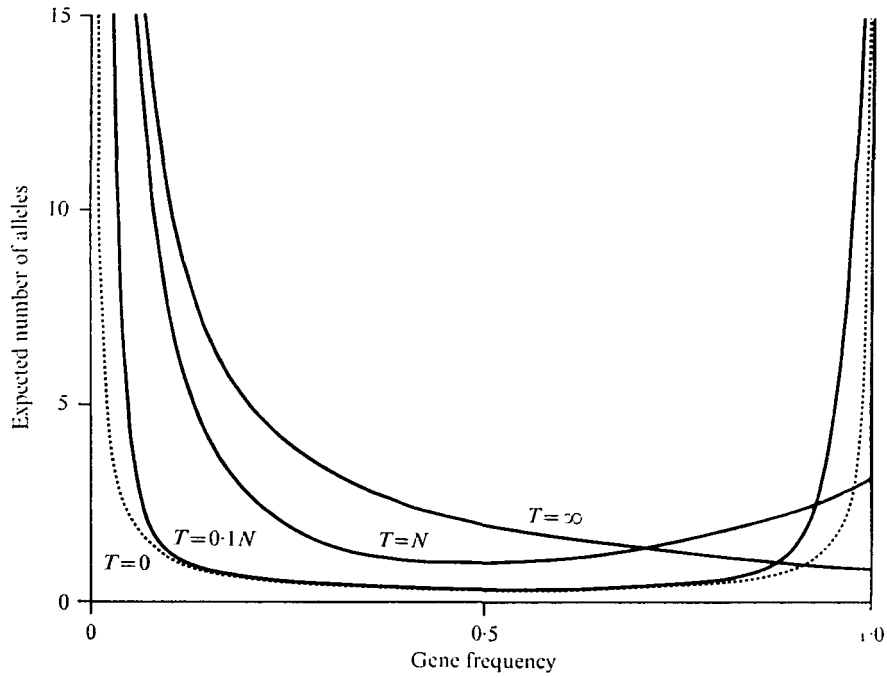

Fig. 2. Transient distributions of allele frequencies. It is assumed that the initial population was in equilibrium with $4Nv = 0{\cdot}1$, but because of the increase of population size, $4Nv$ was raised to 1·0. T: time in the unit of $2N$ generations.

Using amino acid substitution data in evolution, Kimura & Ohta (1971) have estimated the mutation rate for protein loci to be of the order of 10^{-7} per year under the assumption of neutral mutations. If we use this estimate, the population size becomes $1{\cdot}25 \times 10^7$ for $M = 5$ in an organism whose generation time is one year. Thus, $2N$ generations, which are required for reaching steady state, correspond to about 25 million years. This is an extremely long time.

Heterozygosity and average number of alleles per locus

Average or expected heterozygosity is an important measure of genetic variability of a population. The expected heterozygosity in the tth generation is defined as

$$H_t = 1 - \int_0^1 x^2 \Phi(x, t)\, dx.$$

https://doi.org/10.1017/S0016672300016906 Published online by Cambridge University Press

Thus, this can be evaluated by using either (7) or (12), depending on the initial condition. In both cases it can be shown to be

$$H_t = \frac{M}{1+M} + \left(H_0 - \frac{M}{1+M}\right) e^{-(M+1/2N)t}. \quad (13)$$

As expected, this agrees with the result obtained by Nei & Feldman (1972) and Li & Nei (1975) (see also Malécot, 1948).

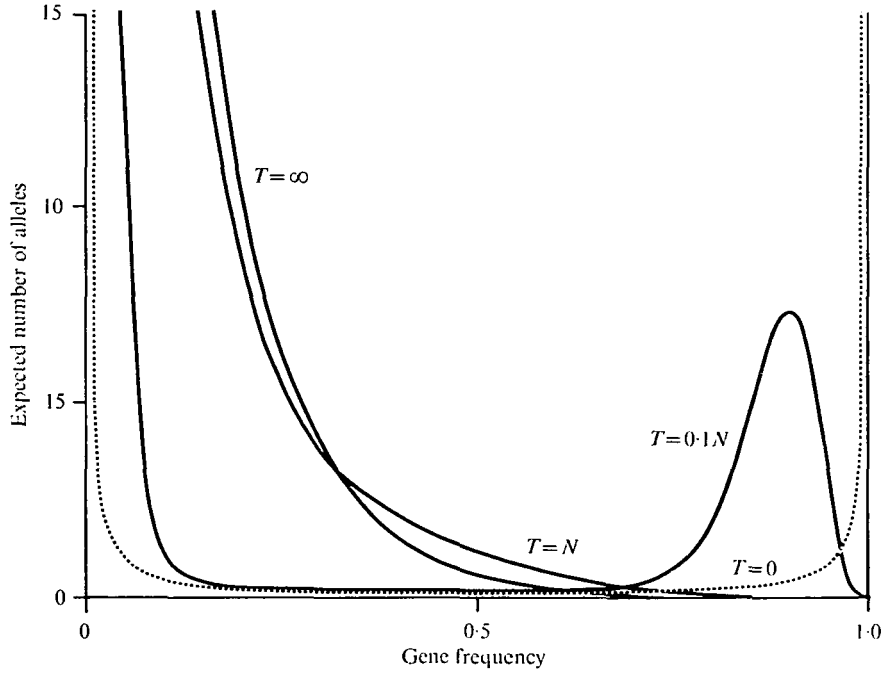

Fig. 3. Transient distributions of allele frequencies. It is assumed that the initial population was in equilibrium with $4Nv = 0{\cdot}05$, but because of the increase of population size, $4Nv$ was raised to 5·0. T: time in the unit of $2N$ generations.

Table 1. *Expected number of alleles and heterozygosity in transient states*

(Case 1: $M = 1$ and $N = 500$ with the initial population being completely monomorphic. Case 2: $M = 1$ and $N = 500$ with the initial population being in equilibrium with $M_0 = 0{\cdot}1$ and $N = 50$)

Time in generations ...	0	$0{\cdot}1N$	N	$10N$	∞
Case 1: n_a	1	4·1	6·1	7·1	7·1
H	0	0·05	0·32	0·5	0·5
Case 2: n_a	1·4	4·4	6·2	7·1	7·1
H	0·09	0·13	0·34	0·5	0·5

Another parameter which is of interest in population genetics studies is the expected number of alleles that are existing in a population (Wright, 1949; Ewens, 1964; Kimura, 1968). This number is obtained by

$$n_a = \int_{1/2N}^{1} \Phi(x, t)\,dx.$$

https://doi.org/10.1017/S0016672300016906 Published online by Cambridge University Press

This value was computed for two different cases. In Case 1, $M = 1$ and $N = 500$ were assumed with the initial population being completely monomorphic. In Case 2, $M = 1$ and $N = 500$ were again assumed but the initial population being in equilibrium with $M_0 = 0{\cdot}1$ and $N = 50$. We assumed a high mutation rate simply because it makes the numerical computation easier. The results obtained are given in Table 1, together with the values of expected heterozygosity. It is clear that the expected number of alleles rapidly increases in the early generations, whereas the increase in expected heterozygosity is rather slow. This is of course expected, since most of the new mutant alleles in the early generations exist in low frequency, so that they do not contribute to heterozygosity very much.

DISCUSSION

In the last decade the formula $\Phi(x) = Mx^{-1}(1-x)^{M-1}$ has been used extensively in the study of the mechanism of maintenance of protein polymorphism. Thus, Ewens (1972) and Maruyama & Yamazaki (1974) developed tests of selective neutrality of genes, based on this distribution. In practice, however, the assumption of steady state on which this formula is based does not always hold. Therefore, we must be cautious about the conclusion obtained by these methods. This is also true with the test proposed by Yamazaki & Maruyama (1972). Recently, Latter (1975) applied this technique and showed that the observed amount of heterozygosity for low gene frequency classes is much higher than the expected. He took this as evidence for his hypothesis of optimum model selection for protein activity. However, his results can also be explained by the hypothesis that the populations he studied (species of the *Drosophila willistoni* group) experienced a bottleneck recently, as indicated by Nei (1976). In fact, the present study shows that if population size increases, many low frequency alleles are accumulated in the population in the early generations (Figs 1 and 2). Thus, the relative amounts of heterozygosity for low gene frequency classes are expected to be higher than those for other classes. It is therefore difficult to distinguish between the two competing hypotheses.

A similar difficulty arises in the study of the distribution of allele frequencies. Ohta (1975) compared the observed distributions of allele frequencies in *Drosophila* and man with the expected distributions which were obtained under the assumption of steady state. She found an excess of low frequency alleles in both *Drosophila* and man. This excess is expected to occur if her hypothesis of slightly deleterious mutations is correct. However, it can also be explained by the hypothesis of recent bottleneck or population expansion, since in this case the number of low frequency alleles first increases (see, for example, Fig. 2). In this connexion it is noted that there is another hypothesis for explaining the excess of low frequency alleles. Namely, if mutation rate varies from locus to locus, the excess of low frequency alleles is again expected to occur even if the population is in equilibrium. This is because the expected number of low frequency alleles increases rapidly with increasing value of $4Nv$, whereas the number of intermediate or high

frequency alleles is less sensitive to the variation of $4Nv$. This can be seen from comparison of the steady state distributions for the different values of $4Nv$ in Figs 1–3.

As mentioned earlier, the present study is based on Kimura & Crow's infinite allele model. In practice, however, protein polymorphism is usually studied by electrophoresis. Recently, Kimura & Ohta (1975) derived an approximate formula for the steady state distribution of allele frequencies appropriate to electrophoretic data. Their study suggests that as long as the value of $4Nv$ remains small compared with unity, the infinite allele model applies approximately to electrophoretic data. In practice, $4Nv$ for protein loci seems to be generally small compared with unity, since the observed value of average heterozygosity is almost always smaller than 0·3 (Nei, 1975).

This study was supported by U.S. Public Health Service Research Grant GM 20293.

REFERENCES

Crow, J. F. & Kimura, M. (1956). Some genetic problems in natural populations. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4**, 1–22.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

Ewens, W. J. (1964). The maintenance of alleles by mutation. *Genetics* **50**, 891–898.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Ewens, W. J. & Gillespie, J. H. (1974). Some simulation results for the neutral allele model, with interpretations. *Theoretical Population Biology* **6**, 35–57.

Ewens, W. J. & Kirby, K. (1975). The eigenvalues of the neutral alleles process. *Theoretical Population Biology* **7**, 212–220.

Karlin, S. & Avni, H. (1975). Derivation of the eigenvalues of the configuration process induced by a labeled direct product branching process. *Theoretical Population Biology* **7**, 221–228.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology* **2**, 174–208.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kimura, M. & Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469.

Kimura, M. & Ohta, T. (1975). Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences U.S.A.* **72**, 2761–2764.

Latter, B. D. H. (1975). Enzyme polymorphisms: gene frequency distributions with mutation and selection for optimal activity. *Genetics* **79**, 325–331.

Li, W.-H. & Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genetical Research* **25**, 229–247.

Malécot, G. (1948). *Les Mathématiques de l'hérédité*. Paris: Masson et Cie.

Maruyama, T. & Yamazaki, T. (1974). Analysis of heterozygosity in regard to the neutrality theory of protein polymorphism. *Journal of Molecular Evolution* **4**, 195–199.

Nei, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam and New York: North Holland.

Nei, M. (1976). Comments on B. D. H. Latter's paper. In *Population Genetics and Ecology* (ed. S. Karlin and E. Nevo), p. 408. New York: Academic Press.

NEI, M. & FELDMAN, M. W. (1972). Identity of genes by descent within and between populations under mutation and migration pressures. *Theoretical Population Biology* **3**, 460–465.

NEI, M. & LI, W.-H. (1975). Probability of identical monomorphism in related species. *Genetical Research* **26**, 31–43.

OHTA, T. (1975). Statistical analyses of *Drosophila* and human protein polymorphisms. *Proceedings of the National Academy of Sciences U.S.A.* **72**, 3194–3196.

WRIGHT, S. (1949). Genetics of populations. *Encyclopedia Britannica* **10**, 111, 111A–D, 112.

YAMAZAKI, T. & MARUYAMA, T. (1972). Evidence for the neutral hypothesis of protein polymorphism. *Science* **178**, 56–58.