Genet. Res., Camb. (1968), **11**, *pp.* 247–269
With 7 text-figures
Printed in Great Britain

Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles*

By MOTOO KIMURA
National Institute of Genetics, Mishima, Japan

(*Received* 28 *July* 1967)

1. INTRODUCTION

It is well known that populations of sexually reproducing organisms such as man and *Drosophila* contain a large amount of genetic variability. Ubiquity of lethal and detrimental genes has been demonstrated in various species of *Drosophila*. Inbreeding studies suggest that the same situation is met with also in man and other organisms. The existence of genetic variability in quantitative characters has been amply demonstrated by selection experiments with diverse plants and animals. Moreover, recent studies on enzyme polymorphism in man and *Drosophila* (Harris, 1966; Lewontin & Hubby, 1966) strongly suggest that genetic variability is quite pronounced at the protein level. It is probable that at the level of genetic material, or in terms of nucleotide sequence, variability within a population is still greater.

Since each gene is made up of a sequence of at least hundreds or thousands of nucleotide pairs and since some base substitutions may have very little effect, it is possible, as reasoned by Kimura & Crow (1964), that the wild-type gene is not a single entity, but a set of different isoalleles that are indistinguishable by any ordinary procedure. They investigated the population consequences of such a system, assuming neutral and overdominant mutations. More recently, Wright (1966) discussed the evolutionary implications of such a system under the term 'polyallelic random drift'.

The purpose of the present paper is to present a fuller treatment of this system for neutral mutations. Also, some discussion on the nearly neutral mutations will be presented. The recent findings of 'degeneracy' of DNA code, that is, existence of two or more base triplets coding for the same amino acid, seem to suggest that neutral mutations may not be as rare as previously considered. Furthermore, some amino acid substitution in a polypeptide chain may have very little effect on the biological activity of the protein, still adding to the possibility of neutral or nearly neutral mutations.

* Contribution No. 648 from the National Institute of Genetics, Mishima, Shizuoka-ken, Japan. Aided in part by a Grant-in-Aid from the Ministry of Education, Japan, and also by a Grant from Toyo Rayon Foundation.

2. AVERAGE HOMOZYGOSITY AND THE EFFECTIVE NUMBER OF ALLELES IN A POPULATION

Throughout this paper, I will consider a population of N diploid individuals and designate by N_e the effective population number (cf. Kimura & Crow, 1963), which may be different from the actual number N.

Let us consider a particular locus and assume that there are K possible allelic states $A_1, A_2, \ldots, A_K$ and that each allele mutates with rate $u/(K-1)$ to one of the remaining $(K-1)$ alleles, so that u is the mutation rate per gene per generation and this is equal to all the alleles. Then, if x_i is the frequency of A_i in a population, the amount of change in one generation of x_i denoted by δx_i has mean and variance

$$M(\delta x_i) = \frac{u}{K-1}(1-Kx_i), \tag{2.1}$$

$$V(\delta x_i) = x_i(1-x_i)/(2N_e). \tag{2.2}$$

Thus, it can be shown with rather elementary but exact calculation that, at equilibrium in which the mutation and random sampling of gametes balance each other, the frequency distribution of x_i has the first and the second moments about zero as follows:

$$\mu_1' = 1/K, \tag{2.3}$$

$$\mu_2' = \left\{\frac{4N_eu}{K(K-1)} - \frac{2N_eu^2}{(K-1)^2} + \frac{1}{K}\right\} \Big/ \left\{\frac{4N_euK}{K-1} - \frac{2N_eu^2K^2}{(K-1)^2} + 1\right\}. \tag{2·4}$$

Therefore the average homozygosity, or the expectation of the sum of squares of allelic frequencies is

$$\bar{H}_0 = E\left(\sum_{i=1}^{K} x_i^2\right) = K\mu_2', \tag{2.5}$$

with μ_2' given by (2.4).

The effective number of alleles (n_e) as defined by Kimura & Crow (1964) is the reciprocal of $\bar{H}_0$, so that

$$n_e \equiv 1/\bar{H}_0 = \left\{4N_eu\left(\frac{K}{K-1}\right) - 2N_eu^2\left(\frac{K}{K-1}\right)^2 + 1\right\} \Big/ \left\{\frac{4N_eu}{K-1} - \frac{2N_eu^2K}{(K-1)^2} + 1\right\}. \tag{2.6}$$

If $2N_eu^2$ is much smaller than unity, we have, with good approximation,

$$n_e = \left\{4N_eu\left(\frac{K}{K-1}\right) + 1\right\} \Big/ \left\{4N_eu\left(\frac{1}{K-1}\right) + 1\right\}. \tag{2.7}$$

If, in addition, the number of allelic states is indefinitely large ($K = \infty$), the above reduces to

$$n_e = 1/\bar{H}_0 = 4N_eu + 1, \tag{2.8}$$

a result derived by Kimura & Crow (1964) using a different method. Actually, the formula is valid as long as the number of allelic states K is much larger than $4N_eu + 1$.

On the other hand, if $2N_e u^2$ is not necessarily very small but $K = \infty$, (2.6) reduces to

$$n_e = 4N_e u + 1 - 2N_e u^2. \tag{2.9}$$

Since formula (2.6) is exact and no restrictions are placed on mutation rate u, effective population number N_e and the number of possible allelic states K, it is reassuring to find here that formula (2.8), i.e. $n_e = 4N_e u + 1$ is valid under rather mild restrictions

$$2N_e u^2 \ll 4N_e u + 1 \ll K. \tag{2.10}$$

It is sometimes remarked that a formula like (2.8) is valid only for u up to $1/N_e$, but no such restriction is needed.

3. PROBABILITY DISTRIBUTION OF ALLELIC FREQUENCIES AND THE AVERAGE NUMBER OF ALLELES IN A POPULATION

In this section, we will investigate the distribution of allelic frequencies using the method of diffusing approximation (cf. Kimura, 1964). Let $\phi(p, x; t)$ be the probability density that the frequency of A_i becomes x at the tth generation given that it is p at the zero (initial) generation. In the following, in order to simplify expressions, letter x rather than x_i will be used to represent the frequency of A_i, still assuming that there are K possible allelic states. Since, from (2.1) and (2.2), the mean and the variance of δx per generation are respectively

$$M_{\delta x} = \frac{Ku}{K-1}\left(\frac{1}{K} - x\right) \tag{3.1}$$

and

$$V_{\delta x} = \frac{x(1-x)}{2N_e}, \tag{3.2}$$

$\phi(p, x; t)$ satisfies the following Kolmogorov forward equation

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left\{x(1-x)\phi\right\} - \overline{m}\frac{\partial}{\partial x}\left\{(\bar{x} - x)\phi\right\}, \tag{3.3}$$

where $\overline{m} = Ku/(K-1)$ and $\bar{x} = 1/K$.

The above equation represents a continuous stochastic process in the change of gene frequency x due to linear evolutionary pressures (mutation, migration) and random sampling of gametes. The solution of this process for arbitrary values of p, $\overline{m}$, $\bar{x}$ and N_e was obtained by the present author through the study of the moments of the distribution (cf. Crow & Kimura, 1956). It is given by

$$\phi(p, x; t) = \sum_{i=0}^{\infty} X_i(x) \exp\left\{-i\left(\overline{m} + \frac{i-1}{4N_e}\right)t\right\}, \tag{3.4}$$

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

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

in which $A = 4N_e\overline{m}$, $B = 4N_e\overline{m}\bar{x}$ and $F(.,.,.,.)$ represents the hypergeometric function. For the present case $A = 4N_e uK/(K-1)$ and $B = 4N_e u/(K-1)$.

At the limit $t \to \infty$, the above distribution converges to

$$\phi(p, x; \infty) = X_0(x),$$

which is independent of the initial frequency p. We will denote this distribution by $\phi(x)$. Thus we obtain

$$\phi(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\alpha-1}x^{\beta-1}, \tag{3·5}$$

where $\alpha = A - B = 4N_e u$ and $\beta = B = 4N_e u/(K-1)$.

The first and the second moments about zero of this distribution are

$$\mu_1' = 1/K \tag{3.6}$$

and

$$\mu_2' = \frac{1}{K}\left\{4N_e u\left(\frac{1}{K-1}\right)+1\right\}\Big/\left\{4N_e u\left(\frac{K}{K-1}\right)+1\right\} \tag{3.7}$$

respectively.

The distribution given by (3.5) is the steady-state distribution which is realized when the effects of mutation (or migration) and random sampling of gametes balance each other. It can also be derived by using Wright's formula for the gene frequency distribution at steady state, namely,

$$\phi(x) = \frac{C}{V_{\delta x}}\exp\left(2\int\frac{M_{\delta x}}{V_{\delta x}}dx\right) \qquad \text{(Wright, 1938}a\text{)}, \tag{3.8}$$

in which constant C is determined such that

$$\int_0^1 \phi(x)dx = 1. \tag{3.9}$$

Going back to the general solution (3.4), we note, as pointed out earlier (Crow & Kimura, 1956), that

$$\int_0^1 \phi(p, x; t)dx = 1. \tag{3.10}$$

This means that, for the present case, the procedure (3.9) of determining C is not an arbitrary statistical procedure, but is the one intrinsically determined by the process. Also, it means that with the present formulation no probability mass exists at any time strictly at the boundaries, i.e. at both $x = 0$ and $x = 1$. On the other hand, in an actual population, especially when it is small, we should expect a considerable possibility of an allele being temporarily lost or fixed in the population.

It looks, then, as if the above approach based on the diffusion approximation is inadequate to obtain the probability of temporary loss or fixation. Fortunately, however, it turns out that the required probability mass lies in the intervals

$(0, 1/2N)$ and $(1-1/2N, 1)$. Thus, the probability that allele A_i is temporarily lost from the population may be obtained from

$$f(0) = \int_0^{1/(2N)} \phi(x)dx \tag{3.11}$$

and the probability that it is temporarily fixed in the population from

$$f(1) = \int_{1-1/(2N)}^{1} \phi(x)dx. \tag{3.12}$$

The probability that both A_i and its alleles (collectively denoted by A_i') co-exist in the population is

$$\Omega = \int_{1/(2N)}^{1-1/(2N)} \phi(x)dx. \tag{3.13}$$

If we substitute the distribution formula (3.5) into (3.11), we obtain

$$f(0) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{1/(2N)} (1-x)^{\alpha-1}x^{\beta-1}\,dx \tag{3.14}$$

as the probability that A_i is temporarily lost from the population. Since u and $1/(2N)$ are generally very small, the above reduces, with good approximation, to

$$f(0) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta+1)} \left(\frac{1}{2N}\right)^{\beta}, \tag{3.15}$$

where $\alpha = 4N_e u$ and $\beta = 4N_e u/(K-1)$. Similarly, the probability that A_i is temporarily fixed is

$$f(1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)} \left(\frac{1}{2N}\right)^{\alpha}. \tag{3.16}$$

Now, formula (3.15) may be expressed in the form

$$\tfrac{1}{2}\beta\left(\frac{N}{N_e}\right) f(0) = \frac{1}{2}\left(\frac{N}{N_e}\right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{1}{2N}\right)^{\beta-1} \frac{1}{2N},$$

but, since

$$\phi\left(\frac{1}{2N}\right) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{1}{2N}\right)^{\beta-1}$$

approximately, it may also be expressed as

$$2N\left(\frac{u}{K-1}\right) f(0) = \frac{1}{2}\left\{\phi\left(\frac{1}{2N}\right)\frac{1}{2N}\right\}\left(\frac{N}{N_e}\right). \tag{3.17}$$

The left-hand side of the above equation represents the number of populations which have no A_i genes which move to the class having one or more of them, since $2Nu/(K-1)$ is the expected number of A_i genes produced per generation in a population where A_i is absent. On the other hand, the right-hand side of the equation is half the frequency of the subterminal class ($x = 1/2N$) multiplied by the factor N/N_e and it represents the number of populations containing one or more A_i

genes which move to the class where A_i is absent (cf. Kimura, 1964, p. 12). At statistical equilibrium in which mutational production of an allele is balanced by random extinction of that allele, the above two numbers should be equal and this justifies equation (3.17) and therefore (3.15). A similar argument applies to (3.16).

I will now proceed to derive the effective and the average number of alleles maintained in a population, using the above approach.

The effective number of alleles, as defined in the previous selection, is the reciprocal of the sum of squares of the allelic frequencies. The latter is

$$\bar{H}_0 = E\left(\sum_{i=1}^{K} x_i^2\right) = K\mu_2' = K\int_0^1 x^2\phi(x)dx = \left\{4N_e u\left(\frac{1}{K-1}\right)+1\right\}\Big/\left\{4N_e u\left(\frac{K}{K-1}\right)+1\right\} \tag{3.18}$$

and this gives the effective allele number

$$n_e = 1/\bar{H}_0 = \left\{4N_e u\left(\frac{K}{K-1}\right)+1\right\}\Big/\left\{4N_e u\left(\frac{1}{K-1}\right)+1\right\}, \tag{3.19}$$

which agrees with (2.7) of the previous section. At the limit $K \to \infty$, where the number of possible allelic states is infinite, this reduces to $n_e = 4N_e u + 1$.

The average number of alleles denoted by n_a is equal to the reciprocal of the mean frequency of alleles existing within a population. The mean here is different from the unconditional mean (μ_1') in that temporarily lost alleles are not taken into account. The frequency of a particular allele, say A_i, averaged over all cases in which it is represented at least once in a population is

$$\bar{x}[x \neq 0] = \mu_1'/\{1-f(0)\}. \tag{3.20}$$

In the present model of assuming equal mutation rates, this value is the same for all the alleles and therefore the average number of alleles turns out to be as follows:

$$n_a = K\{1-f(0)\} = \frac{K\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_{1/(2N)}^1 (1-x)^{\alpha-1}x^{\beta-1}\,dx, \tag{3.21}$$

where $\alpha = 4N_e u$ and $\beta = 4N_e u/(K-1)$.

Figure 1 illustrates the relation between n_a and K assuming $4N_e u = 1$ and $N = 10^4$, together with the relation between n_e and K. At the limit $K \to \infty$, the above formula for n_a converges to

$$n_a = 4N_e u\int_{1/(2N)}^1 (1-x)^{4N_e u-1}x^{-1}\,dx. \tag{3.22}$$

This can also be derived immediately by the frequency distribution given by Kimura & Crow (1964),

$$\Phi(x) = 4M(1-x)^{4M-1}x^{-1}, \tag{3.23}$$

where $M = N_e u$, in which u is the mutation rate to new (not pre-existing) alleles. This distribution has a different meaning from the one so far considered, such as (3.5), in that $\Phi(x)dx$ represents the *expected number of alleles* whose frequency is in the range x to $x+dx$ within the population, rather than representing the proba-

bility that *a particular allele* lies in the frequency range x to $x+dx$. Thus, integrating (3.23) from $x = 1/(2N)$ to $x = 1$, we obtain

$$n_a = 4M \int_{1/(2N)}^{1} (1-x)^{4M-1} x^{-1}\, dx, \tag{3.24}$$

which agrees with (3.22). In the special case of $N = N_e$, this reduces a formula given by Ewens (1964, 1966), who derived it by considering the fate of a new allele. Essentially the same formula as that of Ewens was obtained earlier by Wright

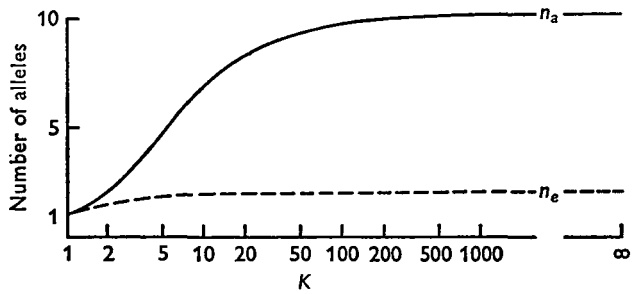

Fig. 1. Relationship between the number of alleles (n_a, n_e) and the number of possible allelic states (K) in a population of $N = 10\,000$, assuming $N_e u = 0{\cdot}25$ (N_e; effective population number, u; mutation rate). The solid line represents the relationship between the actual number of alleles (n_a) and K, while the broken line represents the relationship between the effective number of alleles (n_e) and K.

(1949). It is important to note, however, that in natural as well as controlled populations, the actual population number (N) may be considerably different from the effective population number (N_e).

The effective number of alleles is obtained from the above distribution (3.23) by evaluating the reciprocal of

$$\int_0^1 x^2 \Phi(x)\, dx,$$

giving

$$n_e = 4M + 1 = 4N_e u + 1 \tag{3.25}$$

(Kimura & Crow, 1964). This agrees with formula (2.8) in the previous section. Table 1 lists the average number of alleles obtained by numerical integration of formula (3.24), together with the effective number of alleles derived from (3.25). Also, the relation between the number of alleles and $N_e u$ is illustrated in Fig. 2.

4. SOME SIMULATION STUDIES

In order to check the validity of the foregoing treatments, simulation studies were carried out by using computer IBM 7090. Two programmes (both in Fortran II) were written that differ essentially in the mode of production of mutant genes. In the first program, a pre-determined number of new mutant alleles are intro-

duced into the population in each generation (deterministic mutation). Also, it is so written that all members of the population contribute equally to the gene pool, from which $2N$ gametes are randomly sampled to form the next generation. Thus, the program simulates a monoecious population whose effective number is equal to the actual number, i.e. $N = N_e$. In the second program, mutation to a new allele is induced with a given probability (u) at each step of gamete sampling (random mutation). The program is so written that the population consists of an

Table 1. *The average number (n_a) and the effective number (n_e) of alleles in a population of actual size N and effective size N_e*

N / $N_e u$	5×10^2	10^3	5×10^3	10^4	5×10^4	10^5	5×10^5	10^6	Effective number of alleles n_e
0·001	1·028	1·030	1·037	1·040	1·046	1·049	1·055	1·058	1·004
0·010	1·274	1·301	1·366	1·394	1·458	1·486	1·550	1·578	1·040
0·025	1·675	1·745	1·906	1·975	2·136	2·205	2·367	2·436	1·100
0·050	2·324	2·462	2·784	2·923	3·245	3·384	3·705	3·844	1·200
0·100	3·557	3·834	4·478	4·755	5·399	5·676	6·320	6·597	1·400
0·250	6·908	7·601	9·210	9·903	11·51	12·21	13·82	14·51	2·000
0·500	11·82	13·20	16·42	17·81	21·03	22·41	25·63	27·02	3·000
1·000	20·31	23·08	29·51	32·28	38·72	41·49	47·93	50·70	5·000
2·000	34·58	40·09	52·95	58·50	71·36	76·91	89·78	95·33	9·000
4·000	57·67	68·64	94·30	105·4	131·1	142·2	168·0	179·0	17·000
6·000	76·72	93·08	131·5	148·1	186·7	203·3	242·0	258·6	25·00
8·000	93·18	114·9	166·0	188·1	239·6	261·7	313·2	335·4	33·00
10·000	107·7	134·7	198·5	226·1	290·4	318·1	382·5	410·2	41·00

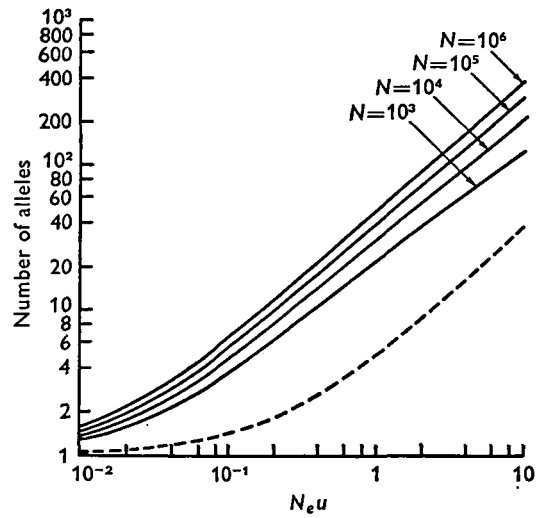

Fig. 2. Graphs showing the relationship between the number of alleles and $N_e u$, where N_e is the effective population number and u is the mutation rate. The solid lines give the average number of alleles (n_a) corresponding to four levels of population number, $N = 10^3$, 10^4, 10^5 and 10^6, while the broken line gives the effective number of alleles (n_e), which is independent of N.

https://doi.org/10.1017/S0016672300011459 Published online by Cambridge University Press

equal number of males and females and that the numbers of breeding males and females may be made smaller than the actual numbers of males and females. Thus it simulates a dioecious population whose effective number may be smaller than the actual number. The effective population number here is given by Wright's formula (Wright, 1938*b*),

$$N_e = 4N_m N_f/(N_m + N_f),$$

where N_m and N_f are the numbers of breeding males and females.

In both programs, sampling of gametes and occurrence of mutation are simulated by generating pseudo-random numbers (using subroutine RAND1). Also, each mutation is treated as a state not pre-existing in the population, so that three formulas, (3.23), (3.24) and (3.25), are relevant in comparing theoretical expectations with computer results. Outputs of both the actual and effective allele numbers were given at pre-assigned intervals. Also, frequency distribution of various alleles within a population was printed out.

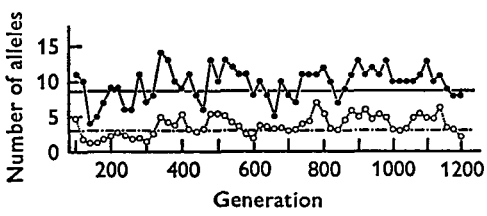

Fig. 3

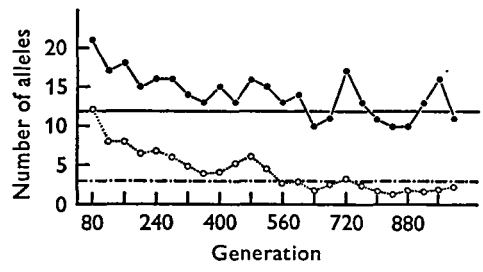

Fig. 4

Figs. 3 and 4. Results of Monte Carlo experiments regarding the number of neutral alleles. In both these experiments, one new mutation is induced in each generation. The average (actual) and the effective number of alleles are plotted respectively by solid and open circles. Horizontal lines, solid and broken, represent corresponding theoretical values derived from the method of diffusion approximation. In Fig. 3, the population consists of 100 monoecious individuals who contribute with equal probability to leaving offspring, namely, $N = N_e = 100$. The mutation rate (u) is 0·005. In Fig. 4, $N = N_e = 500$, $u = 0{\cdot}001$. Neutral alleles deterministic mutation: ●—, average (actual) number, ○ - - -, effective number.

Figures 3–6 illustrate some of the results of Monte Carlo experiments performed by using these two programs. Throughout the experiments, the initial condition was set up such that a population at the zero generation contained $2N$ alleles, that is, all the genes in the initial population were represented by different alleles.

In the experiment shown in Fig. 3, one new mutation was induced in each generation in a population of 100 individuals ($2Nu = 1$, $N = N_e = 100$). Starting with 200 alleles, the balance between mutation and random extinction of alleles has been reached well before generation 100. Actually, a few trials indicated that the majority of the initial 200 alleles are lost within the first 20 generations. Both the average (actual) and the effective numbers of alleles are plotted in the figure at intervals of 20 generations from generation 100 through generation 1200

(56 outputs for each of these two allele numbers). Averaged over these 56 outputs, the average and the effective numbers of alleles turned out to be as follows:

$$n_a = 9{\cdot}68, \quad n_e = 3{\cdot}13 \quad \text{(observed)}.$$

The former was obtained by taking the arithmetic mean of the 56 observed values for the actual allele number, while the latter was obtained by taking the harmonic mean of the 56 observed values for the effective allele number. The corresponding values for n_a and n_e, derived from equations (3.24) and (3.25) by putting $N = 100$ and $N_e u = 0{\cdot}5$, are as follows:

$$n_a = 8{\cdot}61, \quad n_e = 3{\cdot}00 \quad \text{(theoretical)}.$$

These are shown by the horizontal lines in the figure.

Figure 4 illustrates a result of a similar experiment assuming a population of 500 individuals in which one new mutant allele is introduced in each generation. Starting with 1000 different alleles in the zero generation, the balance between mutation and random extinction of alleles is reached well before generation 200. Actually, a majority of the initial 1000 alleles are lost by generation 50. Note that in a very large population the chance of survival of a single neutral gene for t generations is approximately $2/t$ when t is large (Fisher, 1930). Averaged over 21 outputs (from generation 200 through generation 1000 at intervals of 40 generations), the average and the effective numbers of alleles were as follows:

$$n_a = 13{\cdot}43, \quad n_e = 2{\cdot}79 \quad \text{(observed)}.$$

The corresponding values derived from diffusion approximations are

$$n_a = 11{\cdot}82, \quad n_e = 3{\cdot}00 \quad \text{(theoretical)}.$$

Thus, the two experiments assuming deterministic mutation have given results that agree fairly well with theoretical predictions. The diffusion approximation, however, tends to underestimate n_a slightly. The remaining two experiments which are illustrated in Figs. 5 and 6 were carried out by using the second program (random mutation).

In the experiment shown in Fig. 5, the population consists of 50 males and 50 females ($N = 100$), of which only 25 males and 25 females actually participate in breeding ($N_e = 50$). In each generation, 100 male and 100 female gametes are randomly chosen from these 25 breeding males and 25 breeding females to form the next generation. Mutation to a new, not pre-existing, allele is induced in each gamete with probability 0·005 prior to the formation of zygotes ($u = 0{\cdot}005$). The initial population was set up such that it contained 200 different alleles. The balance between mutation and random extinction of alleles was reached well before generation 100. Actually, the majority of the original 200 alleles were lost by generation 20. The figure depicts the course of fluctuation of the average and the effective numbers of alleles in the population at intervals of 40 generations, from generation 120 through generation 2080. The actual computer outputs were

at intervals of 20 generations and gave 100 observed pairs over generations 120–2100, from which the average and the effective numbers of alleles came out as follows:

$$n_a = 6{\cdot}05, \quad n_e = 2{\cdot}07 \quad \text{(observed)}.$$

On the other hand, from equations (3.24) and (3.25), the corresponding values are:

$$n_a = 5{\cdot}30, \quad n_e = 2{\cdot}00 \quad \text{(theoretical)}.$$

In the experiment illustrated in Fig. 6, the actual population number is 100, while the effective population number is 18 (5 breeding males and 45 breeding females). The mutation rate is the same as before. The simulation was carried out until generation 1300 and outputs of both the average and the effective numbers of

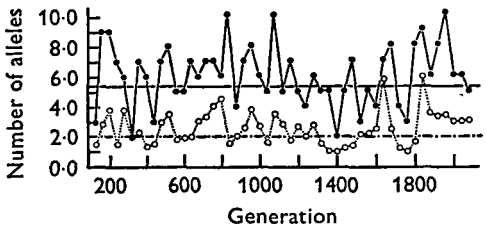

Fig. 5

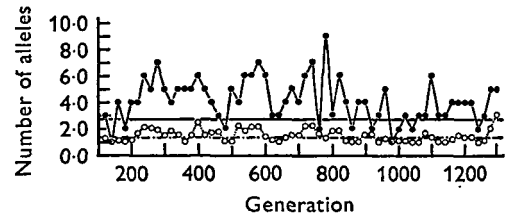

Fig. 6

Figs. 5–6. Results of Monte Carlo experiments as regards the number of neutral alleles. In these 2 experiments, mutation to a new, not pre-existing, allele is induced at each gamete sampling with probability $u = 0{\cdot}005$. The average and the effective numbers of alleles are plotted respectively by solid and open circles. Horizontal lines, solid and broken, represent corresponding theoretical values derived from the diffusion approximations. In Fig. 5, the population consists of 50 males and 50 females, 25 of each sex participating in breeding, so that $N = 100$, $N_e = 50$. In Fig. 6, $N = 100$ but $N_e = 18$. For details, see text. Neutral alleles random mutation: ●—, average (actual) number, ○ - - -, effective number.

alleles are given at intervals of 20 generations starting from generation 120. This yielded 60 pairs of outputs, from which the following values were obtained:

$$n_a = 4.12, \quad n_e = 1{\cdot}38 \quad \text{(observed)}.$$

The corresponding values derived from equations (3.24) and (3.25) by putting $N = 100$, $N_e = 18$, $u = 0{\cdot}005$ are:

$$n_a = 2{\cdot}74, \quad n_e = 1{\cdot}36 \quad \text{(theoretical)}.$$

Additional results of Monte Carlo experiments together with those already mentioned are summarized in Table 2. Despite the smallness of the population number assumed in these experiments, agreement between observed and expected values is fairly good, except that the diffusion approximation tends to underestimate n_a.

These Monte Carlo experiments also gave the frequency distribution of various alleles within a population at equilibrium. An example is given in Fig. 7 in which

observed values are plotted with the squared dots. They were derived from the same experiment from which Fig. 5 was constructed and they were the averages of 100 actual distributions observed from generation 120 through generation

Table 2. *Summary of the results of Monte Carlo experiments regarding the number of neutral alleles in a population*

In the experiments No. 1 and No. 2, mutation is deterministic, but in the remaining experiments, mutation is stochastic. The numbers inside parentheses indicate the numbers of outputs from which n_a and n_e were computed.

Expt. no.	Population size				Mutation rate	Output	Observed means		Diffusion approximations	
	N	Ne	♂	♀			n_a	n_e	n_a	n_e
1 (Fig. 3)	100	100	/	/	0·005	100–1200 (56)	9·68	3·13	8·61	3·00
2 (Fig. 4)	500	500	/	/	0·001	200–1000 (21)	13·43	2·79	11·82	3·00
3 (Fig. 5)	100	50	25	25	0·005	120–2100 (100)	6·05	2·07	5·30	2·00
4	100	100	50	50	0·005	120–2100 (100)	9·34	2·26	8·61	3·00
5 (Fig. 6)	100	18	5	45	0·005	120–1300 (60)	4·12	1·38	2·74	1·36
6	100	100	50	50	0·005	100–1200 (23)	10·91	3·22	8·61	3·00
7	100	50	25	25	0·005	100–1200 (23)	5·52	1·93	5·30	2·00
8	50	50	25	25	0·01	40–400 (19)	9·32	3·67	7·23	3·00
9	100	50	25	25	0·01	50–500 (19)	10·42	3·13	8·61	3·00
10	200	200	100	100	0·01	140–1120 (50)	34·74	10·66	27·30	9·00
11	500	167	50	250	0·001	220–900 (18)	6·78	1·99	5·07	1·67

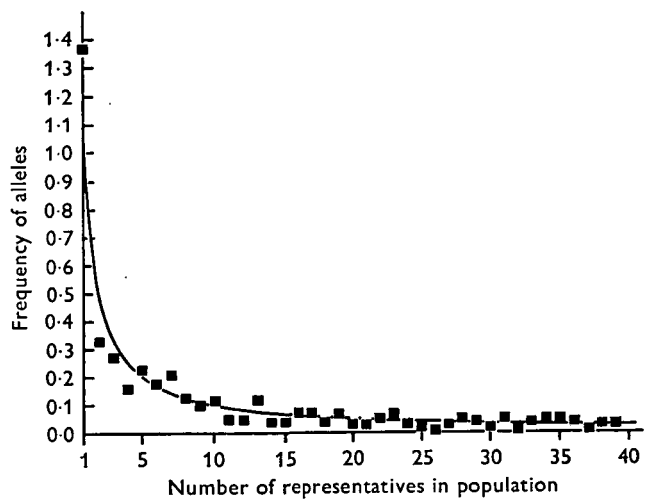

Fig. 7. Frequency distribution of alleles at equilibrium in a population consisting of 50 males and 50 females, of which only 25 males and 25 females actually participate in breeding. The mutation to a new, not pre-existing, allele is induced at each gamete sampling with probability 0·005. Squared dots represent the observed values from a Monte Carlo experiment and the solid curve represents the theoretical distribution obtained from the diffusion approximation. The ordinate stands for the frequency (in absolute number) of alleles having 1, 2, 3, etc., representatives in the population. Frequency distribution of neutral alleles in a population: $N = 100$, $N_e = 50$, $u = 0{\cdot}005$; —, theoretical distributions, ■, observed frequency.

https://doi.org/10.1017/S0016672300011459 Published online by Cambridge University Press

2100 at intervals of 20 generations. The ordinate in the figure stands for the frequency (in absolute numbers) of alleles that have 1, 2, 3, etc., representatives in the population. To make the meaning of the above distribution clearer, let us suppose that a small population consists of 5 individuals, A_1A_1, A_1A_2, A_2A_3, A_3A_4, A_5A_6. In this case, allele A_1 has 3 representatives, alleles A_2 and A_3 have 2 representatives each, and alleles A_4, A_5 and A_6 have a single representative each. Thus the frequencies (in absolute number) of alleles that have 1, 2 and 3 representatives are 3, 2, and 1 respectively. This population contains 6 different alleles but the sum of squares of allelic frequencies (in proportion) is 0·2 and therefore $n_a = 6$, $n_e = 5$.

Going back to the distribution in Fig. 7, the solid curve represents the theoretical distribution derived from $\Phi(x)dx$ by replacing dx by $1/(2N)$. Since, for the present experiment $M = N_e u = 50 \times \cdot 005 = 0{\cdot}25$, formula (3.23) gives

$$\Phi(x)dx = 1/(2Nx),$$

where $2Nx$ is the number of representatives in the population.

The agreement between the observed and the theoretical distributions is fairly good except that the diffusion approximation tends to underestimate the frequency of alleles represented only once in the population. The same tendency was observed also in the experiment performed with $N = 100$, $N_e = 100$, $u = 0{\cdot}005$ (random mutation).

5. DISCUSSION

(1) *Nature of mutant alleles.* The ultimate source of genetic variability in a population is mutation. It is now known that mutation is caused by changes in DNA base arrangements, namely, substitutions, gains and losses. Among them, addition or loss of a single base pair causes a shift of reading frame ('frame shift') and will produce far more drastic effects than single-base alterations. Among the base substitutions, some lead to alteration of amino acids which are quite dissimilar in chemical properties, thus producing marked mutational effects. Especially, changes to the chain termination codons (nonsense codons) would be most damaging. Those leading to substitution of chemically similar amino acids at a position of the polypeptide chain which is different from the active site may produce very little phenotypic effect. Still others cause no alteration of amino acids and their mutational effect in general should be minimal. Sonneborn (1965) called the last category of mutations 'synonymous'. He conjectured that it would not be surprising if 20 % or more of all single-base mutations were synonymous. Since a more complete dictionary of the genetic code is now available, an attempt was made to obtain the probability that a mutation is synonymous, giving due weight to the frequencies of various codons in a haploid chromosome set. The method of calculation is given in detail in Appendix I. The results support Sonneborn's conjecture. That is, the probability is about 0·34 if the base pair replacement is exclusively of transitional type, but is roughly 0·23 if all types of single-base substitution occur with equal frequency.

Sonneborn also suggested a possibility of 'recombinational pseudomutation', that is, production of a codon for a different amino acid by recombination of two synonymous codons. For example, UGC (Cys) would be produced from UCU (Ser) and AGC (Ser) by recombination. This event could take place only in a population which is polymorphic with respect to synonymous codons at a given site on a chromosome. Since the mutation rate per nucleotide pair per replication is estimated to be about 10^{-11} in *Drosophila* and 10^{-12} in man (Kimura, 1967), the mutation rate (u_c) per codon per generation is probably of the order of 10^{-8} in *Drosophila* and 10^{-9} in man even if nearly neutral mutations are included. Thus even for a population of $N_e = 10^4$, $N_e u_c$ is at most of the order of 10^{-4}. This means that at each site of DNA triplet, synonymous polymorphism must be extremely rare and, accordingly, recombinational pseudomutation is probably a very rare phenomenon in nature.

It is important to note here that probably not all synonymous mutations are neutral, even if most of them are nearly so.

Mutations which lead to substitution of somewhat similar amino acids also produce little or no change in biological activity, depending on their position in the polypeptide chain, and such mutations might be called *imperfectly synonymous*. For example, substitution of serine for glycine, in position 47 of the *E. coli* tryptophan synthetase A, leaves the enzymic activity intact (cf. Watson, 1965). Some of the imperfectly synonymous mutations may produce enzymes that have different electrophoretic property, yet differ little in biological activity. Some of the isozyme polymorphisms must be caused by such mutations.

Thus we have a wide mutation spectrum with respect to fitness: the (recessive) lethal mutations damage the developmental processes so drastically that individuals carrying them in homozygous condition cannot survive to maturity. In the second chromosome of *Drosophila melanogaster*, some 500 loci (possibly about one-eighth of the total) are capable of producing lethal mutations and the total rate amounts to about 0·5 %. Mutations causing less deleterious effect, on what Mukai (1964) called viability polygenes, appear to be much more numerous. The mutation rate for such genes is estimated to be about 14 % per second chromosome. This means that the total mutation rate per individual may reach at least 70 %. Probably, the mutational load due to such viability polygenes is reduced by 'reinforcing type' epistasis (Kimura & Maruyama, 1966). Mutations causing still less deleterious effect are difficult to detect, except possibly those found to be isozyme mutations. In addition, a recent analysis of the genetic variation concerning the number of sternopleural bristles in *Drosophila* suggests that the genes responsible for the character are nearly neutral (Robertson, 1967). It is probable that the same situation will be met with when many other quantitative characters are concerned, and, as suggested earlier by Clayton & Robertson (1955), their existing variation could well be maintained by the equilibrium between inbreeding and mutation.

An important problem confronting us now is what the rate is of occurrence of neutral and nearly neutral mutations. According to Robertson (1967), 'apparently

fewer than 30 % of the single amino acid substitutions compatible with the genetic code would cause a change in the electrical charge of the protein molecule' (see also Shaw, 1965). This means that mutations that cannot be detected by electrophoresis occur twice as frequently as those that can be so detected. Substitution of similar amino acids such as leucine with isoleucine will not cause a change in the electrical charge.

In discussing neutral or nearly neutral mutations, the fact that substitution of amino acids in many parts of a polypeptide chain often causes no change of its catalytic activity (cf. Watson, 1965) is probably significant.

If the effects of amino acid substitution on the activity of a polypeptide chain were thoroughly known, it would be possible to assess the frequency of neutral or nearly neutral mutations by estimating the frequencies of synonymous or imperfectly synonymous mutations and the average size of the active site or sites in a chain.

The following is a preliminary (and admittedly crude) attempt along this line. According to Goldberg & Wittes (1966), 20 amino acids may be divided into 8 groups of similar amino acids, namely: {Pro}, {Try, Tyr, Phe, Ileu, Leu, Met, Val}, {Cys}, {Thr, Ser}, {Gly, Ala}, {GluN, AspN}, {Glu, Asp} and {His, Arg, Lys}. For each group, they calculated the probability that a single-base substitution would not lead to a change of group, assuming that all the codons have the same frequency. Using these probabilities but giving due weight to each group of amino acids and their expected frequencies (cf. Table A1, column 3), the average probability (p_s) was calculated. It turned out that p_s is about 0·43, if all types of a single-base substitution occur with equal frequency. If the base substitution is exclusively of transitional type, the corresponding probability is about 0·46. Let a be the fraction of length which the active site or sites occupy on a polypeptide chain. Since the value of a must vary from molecule to molecule, we will take its average value for a. Then, among all mutations due to a single-base substitution with respect to a polypeptide chain, the fraction of mutations that are synonymous or imperfectly synonymous on the non-active site is $(1-a)p_s$. The actual value of a is not known but it is probable that a is at the most 10 %. So, the above fraction should be roughly equal to p_s. Since there are always some (more drastic but less frequent) mutations due to DNA base additions or losses, p_s will impart an upper limit to the fraction of neutral or nearly neutral mutations among all mutations. In conclusion, it seems probable that neutral or nearly neutral mutations might reach some 40 % of all mutations.

If this conclusion turns out to be correct, and, if the remaining 60 % of all mutations are detected as viability polygenes, lethals and semi-lethals, the total mutation rate per gamete in *Drosophila* may reach some 60 % per generation.

(2) *A definition of neutrality*. For any species, there is an upper limit to the total genetic load, or the amount of selective elimination due to genotypic differences. This is because the reproductive capacity of each species is limited. Furthermore, there is always death or sterility due to environmental causes. Thus, as pointed out earlier by Wright (1931) and others, selection intensity per locus depends on

the total number of segregating loci in a population. The former must decrease as the latter increases.

In the last decade, much emphasis has been laid on the possibility of non-existence of neutral mutations. It may be true that the 'visible' mutations, in the sense that they are discernible by the human eye, do have almost always some selective difference. However, the number of polymorphisms due to such mutations must be small as compared with the total number of loci. On the other hand, recent studies of enzyme polymorphisms suggest that a large number of loci are segregating in a population. In Harris's (1966) study, among the 10 arbitrarily chosen enzymes of man, 3 were found to be polymorphic. In the case of *Drosophila pseudoobscura*, Lewontin & Hubby (1966) studied 18 loci responsible for enzymes and other proteins; they found that the average population is polymorphic for about 30% of the loci. They estimated that each individual is heterozygous on the average for 12% of all loci. Since the total gene number in *Drosophila* is estimated to be about 10000 (cf. Muller, 1967), the above findings mean that in this organism each individual is heterozygous for over 1000 loci on the average. The same situation may be met with in man.

This brings us to the problem of natural selection toward holding these polymorphisms. A consideration of the genetic load leads us to conclude that the natural selection acting on the majority of loci at any one time must be small. Kimura & Crow (1964) have shown that if polymorphisms are maintained in thousands of loci by overdominance, each with appreciable selection coefficients, the total load becomes intolerably large.

In considering the effect of selection on each locus, an important quantity is $N_e s$, namely, the product of selection coefficient (s) and effective population number (N_e). A mutant gene may be called *almost neutral* if $|2N_e s|$ is much smaller than unity. Under this definition, neutrality depends not only on s but also on N_e. Thus a gene is almost neutral in a small population but not so in a large one. In this connexion it should be noted that mild overdominance which is efficient enough to maintain a polymorphism in a large population has very little effect in maintaining the polymorphism in a small population.

To see this point more clearly, let us consider a pair of overdominant alleles A_1 and A_2 and assume that the relative fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are $1-s_1$, 1 and $1-s_2$ respectively. Accordingly, in a population of effective size N_e, assuming the most favourable condition $s_1 = s_2 \equiv s$ for maintaining polymorphism, it has been shown by the present author (cf. Robertson, 1962) that the probability of co-existence of both alleles decreases at the rate

$$\lambda_0 = \frac{1}{2N_e}\left\{1 - \frac{2N_e s}{5} + \frac{2^4 . 3}{5^4 . 7}(N_e s)^2 - \frac{2^5}{5^5 . 7}(N_e s)^2 - \quad \ldots\right\} \tag{5.1}$$

per generation. The above power series is valid for $N_e s$ up to about 4.

Since the corresponding rate for a strictly neutral pair of alleles is

$$\lambda_0 = 1/(2N_e),$$

the above formula (5.1) shows that for a small value of $N_e s$, the rate of decay of variance is reduced only by the fraction

$$\tfrac{2}{5}N_e s.$$

Thus for a pair of alleles with 1 % overdominance ($s = 0{\cdot}01$), if an experimental population is kept in a culture bottle with 50 parents ($N_e = 50$) in each generation, the above fraction becomes 0·2. This value will become much less if $s_1 \neq s_2$ and also if N_e is sometimes reduced in the course of breeding. Thus, no appreciable effects of overdominance should be observed under such conditions. On the other hand, in a population of $N_e = 10^4$, the overdominant alleles with $s = 0{\cdot}01$ will be kept in the population almost indefinitely.

(3) *Population structure and migration.* Usually, a species which occupies a wide territory and consists of a large number of individuals does not form a single panmictic unit, but comprises a number of subgroups or 'demes', mating taking place within each deme nearly at random. However, there is always some migration between the subgroups, so that as a whole the species forms a single reproductive community. It has been advocated by Wright (cf. Wright, 1951) that such a subdivided structure is conducive to the maintenance of genetic variability and therefore is favourable to rapid evolutionary progress. He studied the local differentiation of gene frequencies by assuming a continuous population structure.

The problem of local differentiation in gene frequencies was also studied by Kimura & Weiss (1964), who used 'the stepping stone model' of population structure, in which the entire population is subdivided into colonies and migration is restricted to nearby colonies. In this model, if N_e is the effective number of each colony and m_1 is the rate of migration between adjacent colonies, then, assuming mutation as in §2 (K possible allelic states and mutation rates equal in all directions), the gene frequency distribution corresponding to formula (3.5) is approximately

$$\phi(x) = \frac{\Gamma(\bar{A})}{\Gamma(\bar{A}-\bar{B})\Gamma(\bar{B})} x^{\bar{B}-1}(1-x)^{(\bar{A}-\bar{B})-1}, \tag{5·2}$$

where

$$\bar{A} = 4N_e m', \quad \bar{B} = 4N_e m'\bar{x},$$

in which

$$m' = \frac{Ku}{K-1} + m_1(1-r_1), \quad \bar{x} = 1/K.$$

In the above expressions, r_1 is the correlation coefficient of gene frequencies between two adjacent colonies (i.e. colonies 'one step' apart) and its actual value depends on the number of dimensions as well as the rates of migration and mutation (Kimura & Weiss, 1964; Weiss & Kimura, 1965).

From the above distribution (5.2),

$$\bar{H}_0 = K\int_0^1 x^2\phi(x)dx = \left(\frac{4N_e m'}{K}+1\right)\Big/(4N_e m'+1). \tag{5.3}$$

https://doi.org/10.1017/S0016672300011459 Published online by Cambridge University Press

For the one-dimensional stepping-stone model, if $u \ll m_1 \ll 1$, it can be shown that

$$1-r_1 = \sqrt{[\{2Ku/(K-1)\,m_1\}]}.$$

Thus, if $N_e = 10^4$, $u = 10^{-5}$, $K = 10$ and $m_1 = 10^{-1}$, we have $r_1 \approx 0{\cdot}985$, and the average homozygosity is approximately

$$\bar{H}_0 = 0{\cdot}11 \quad \text{(heterozygosity of about 89 \%)}.$$

On the other hand if $m_1 = 0$ but under otherwise the same condition

$$\bar{H}_0 = 0{\cdot}71 \quad \text{(heterozygosity of about 29 \%)}.$$

For the two-dimensional stepping-stone model, if $u \ll m_1 \ll 1$, it can also be shown that

$$1-r_1 = \frac{\pi}{2}\left(\log_e \frac{4}{\sqrt{[2Ku/\{(K-1)\,m_1\}]}}\right)^{-1}.$$

Thus, if $N_e = 10^4$, $u = 10^{-5}$, $K = 10$ and $m_1 = 10^{-1}$, we have $r_1 \approx 0{\cdot}72$ and the average homozygosity is approximately

$$\bar{H}_0 = 0{\cdot}10.$$

On the other hand, if $m_1 = 0$, $\bar{H}_0 = 0{\cdot}71$, as before. These examples show that under subdivided structure and migration, much higher heterozygosity is expected.

When the number of allelic states is infinite, r_k is proportional to the probability that two homologous genes taken one from each of the two colonies k steps apart have the same allelic state. Thus r_k gives the fraction of alleles (in terms of the effective allele number) that are shared by these two colonies.

The advantage of population subdivision in keeping a large number of alleles may best be seen by comparing the average number of alleles maintained in a species under panmixis and under subdivision. For example, in a species of $N = N_e = 10^6$, if $u = 10^{-5}$ and if every mutation is to a new, not pre-existing, allele, then $n_a = 410{\cdot}2$ when the species forms a single random mating unit (see Table 1). On the other hand, $n_a = 1{\cdot}301 \times 1000 \approx 1300$ when the species is subdivided into 1000 completely isolated colonies of size $N = N_e = 10^3$. With a small amount of migration, this number will be reduced but may still be large as compared with the panmictic population.

SUMMARY

1. The average and the effective numbers of alleles maintained in a finite population due to mutational production of neutral isoalleles were studied by mathematical analysis and computer simulation.

2. The exact formula was derived for the effective number (n_e) of alleles maintained in a population of effective size N_e, assuming that there are K possible allelic states and mutation occurs with equal frequency in all directions. If the number of allelic states is so large that every mutation is to a new, not pre-existing,

allele, we have $n_e = 4N_eu + 1 - 2N_eu^2$, where u is the mutation rate. Thus, the approximation formula, $n_e = 4N_eu + 1$, given by Kimura & Crow (1964) is valid as long as $2N_eu^2 \ll 1$.

3. The formula for the average number of alleles (n_a) maintained in a population of actual size N and effective size N_e was derived by using the method of diffusion approximation. If every mutation is to a new, not pre-existing, allele, we obtain

$$n_a = 4M \int_{1/(2N)}^{1} (1-x)^{4M-1} x^{-1}\, dx,$$

where $M = 4N_eu$. The average number of alleles as a function of M and N is listed in Table 1.

4. In order to check the validity of the diffusion approximations, Monte Carlo experiments were carried out using the computer IBM 7090. The experiments showed that the approximations are satisfactory for practical purposes.

5. It is estimated that among the mutations produced by DNA base substitutions, synonymous mutations, that is, those which cause no alterations of amino acids, amount roughly to 0·2–0·3 in vertebrates. *Incompletely synonymous mutations*, that is, those which lead to substitution of chemically similar amino acids at a different position of the polypeptide chain from the active site and therefore produce almost no phenotypic effects, must be very common. Together with synonymous mutations, they might constitute at least some 40% of all mutations. These considerations suggest that neutral and nearly neutral mutations must be more common than previously considered.

I would like to express my thanks to Dr Takeo Maruyama for stimulating discussions in the course of the present work. Thanks are also due to Dr Alan Robertson for reading the manuscript and making valuable suggestions.

APPENDIX I

Probability that a mutation is synonymous

Two or more codons are said to be synonyms (Muller, 1963, cited in Muller, 1967) if they code for the same amino acid. In order to calculate the relative frequency of synonymous mutations among all the mutations produced in an individual by DNA base substitutions, we must know (1) the relative frequencies of various codons in a haploid chromosome set and (2) the frequency of synonymous mutations for each codon. Since the relative frequencies of various codons in the haploid set are not known, we must estimate them either from the frequencies of the four DNA bases, adenine (A), thymine (T), guanine (G) and cytosine (C), or from the frequencies of 16 dinucleotides obtained by the nearest-neighbour analysis (Josse, Kaiser & Kornberg, 1961). Throughout the present calculation, the RNA code dictionary used was the one given by Crick (1966), with UGA included as a nonsense triplet, following Brenner, Barnett, Katz & Crick (1967). We note here that

uracil (U) in RNA code corresponds to thymine (T) in DNA code. Note also that the messenger RNA is complementary to one of the strands of DNA from which the former is 'transcribed'. According to Sueoka (cf. 1965), the G–C content of DNA obtained from various vertebrate species lies within the range 40–44 %. So, in the first calculation, the relative frequencies of A, T, G and C are assumed to be 0·285, 0·285, 0·215 and 0·215, respectively. Actually, the transcription is made from only one of the two strands of DNA and the frequencies of A and G may not necessarily be equal respectively to T and C in this strand. However, we will assume that in

Table A1. *Observed and expected frequencies of various amino acids in the proteins of vertebrates (for details, see text)*

Amino acid	Observed	Relative frequency (%) Expected (1)	Expected (2)
Glu, GluN	10·13	6·51	7·89
Asp, AspN	9·13	7·57	8·28
Gly	7·96	4·91	4·88
Leu	7·92	10·82	12·26
Ser	7·67	9·76	9·10
Ala	7·53	4·91	1·43
Lys	6·70	4·31	5·58
Val	6·68	6·50	9·53
Arg	6·28	8·16	7·34
Thr	5·87	6·50	6·78
Pro	5·21	4·91	4·46
Ileu	4·05	6·77	5·17
Phe	3·81	4·31	5·97
Tyr	3·55	4·31	3·92
Cys	2·77	3·25	2·08
His	2·37	3·25	3·01
Met	1·52	1·85	1·15
Try	0·85	1·40	1·17
Total	100·00	100·00	100·00

higher organisms, A = T, and G = C hold approximately in each strand of DNA. Using the above frequencies of A, T, G, C and assuming independence of base arrangements, relative frequencies of 64 codons may be obtained. For example, the frequency of AAA is $(0{\cdot}285)^3$ or about 0·02315, from which RNA codon UUU is derived. However, three codons, UAA (Ochre), UAG (Amber) and UGA lead to chain termination in polypeptide synthesis. So the relative frequencies of the remaining 61 codons are recalculated after removing these 3 nonsense codons. In order to test the validity of this approach, the relative frequencies of occurrence of various amino acids in proteins were predicted, using the frequencies of those codons. For example, lysine is coded by AAA and AAG. Thus the estimated frequency of this amino acid is the sum of the frequencies of the two codons, which turns out to be 4·312 %. The third column in Table A1 lists the frequencies of

amino acids computed in this way. They should be compared with the corresponding values (second column) actually observed in proteins from vertebrates. The latter values are averages from 61 proteins of vertebrate origin listed in Smith's paper (1966), who compiled the amino acid composition of 80 proteins including those of non-vertebrate origin. Agreement between the observed and the expected frequencies is only fair, but it does indicate that the method is sound as a first approximation in predicting frequencies of various codons. (Prediction is somewhat poor for Ala and Gly but this may be due to some unknown functions of these amino acids.) A similar calculation was carried out using dinucleotide frequencies obtained by Josse *et al.* (1961) for calf thymus DNA (they are as follows: AA 0·089, TT 0·087, CA 0·080, TG 0·076, GA 0·064, TC 0·067, CT 0·067, AG 0·072, GT 0·056, AC 0·052, GG 0·050, CC 0·054, TA 0·053, AT 0·073, CG 0·016, GC 0·044). This second method of calculation assumes that the frequency of codon TTC, for example, is proportional to the product of the frequencies of TT and TC. The last column of Table A1 lists the relative frequencies of amino acids predicted by using frequencies of 61 sense codons thus calculated. The agreement between the

Table A2. *Probability that a mutation is synonymous*

Type of base substitution	Predicted by mononucleotide frequencies	Predicted by dinucleotide frequencies
Transition only, equal in both directions	0·341	0·349
All single-base substitutions with equal frequency	0·233	0·231

observed and the expected frequencies is less satisfactory than in the previous case.

For each codon, the probability that it still codes for the same amino acid after single-base substitution depends on the type and frequency of DNA base replacement. So, two cases were studied. In the first case, only transition ($\frac{G}{C} \rightleftharpoons \frac{A}{T}$) was considered and this was assumed to occur with equal frequency in both directions. Thus, for example, in terms of RNA codon, GCA (Ala) changes to ACA (Thr), GUA (Val) and GCG (Ala) with equal frequency, and therefore the probability is $\frac{1}{3}$ that GCA still codes for the same amino acid after single-base substitution. This probability was calculated for each of the 61 sense codons, and the resulting probabilities were averaged by giving weight to the frequencies of codons predicted by the two different methods mentioned above. The final results are listed in Table A2. From the figures in the upper row of the table, it will be seen that the probability that a mutation is synonymous is about 0·34.

In the second case studied, every one of the four nucleotides is assumed to change to one of the remaining three nucleotides with equal probability. Thus, for example, UUA (Leu) changes to UUU (Phe), UUC (Phe), UUG (Leu), UCA (Ser), UAA (Ochre), UGA (nonsense), CUA (Leu), AUA (Ileu) and GUA (Val) with equal frequency, and therefore the probability is $\frac{2}{9}$ that UUA still codes for the same amino acid after single-base substitution. In this way, the probability was calculated for each of the 61 sense codons and weighted averages were calculated as in the previous case. The results are given in the bottom row of Table A2. They

show that with this type of base substitutions, the probability that a mutation is synonymous is about 0·23.

Summing up, it is estimated that in vertebrate species, the probability that a mutation is synonymous is about 0·34 if the base substitution is exclusively of transition type, but is roughly 0·23 if all types of single-base substitution occur with equal frequency.

REFERENCES

BRENNER, S., BARNETT, L., KATZ, E. R. & CRICK, F. H. C. (1967). UGA: A third nonsense triplet in the genetic code. *Nature, Lond.* **213** (5075), 449–450.

CLAYTON, G. & ROBERTSON, A. (1955). Mutation and quantitative variation. *Am. Nat.* 89, 151–158.

CRICK, F. H. C. (1966). The genetic code: III *Scient. Am.* **215** (4), 55–62.

CROW, J. F. & KIMURA, M. (1956). Some genetic problems in natural populations. *Proc. Third Berkeley Symp. on Math. Stat. and Prob.* **4**, 1–22.

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

EWENS, W. J. & EWENS, P. M. (1966). The maintenance of alleles by mutation—Monte Carlo results for normal and self-sterility populations. *Heredity* **21**, 371–378.

FISHER, R. A. (1930). *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.

GOLDBERG, A. L. & WITTES, R. E. (1966). Genetic code: Aspects of organization. *Science, N.Y.* **153**, 420–424.

HARRIS, H. (1966). Enzyme polymorphism in man. *Proc. Roy. Soc.* B **164**, 298–310.

JOSSE, J., KAISER, A. D. & KORNBERG, A. (1961). Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of nearest neighbour base sequences in deoxyribonucleic acid. *J. Biol. Chem.* **236**, 864–875.

KIMURA, M. (1964). Diffusion models in population genetics. *J. appl. Probability* **1**, 177–232.

KIMURA, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* **9**, 23–34.

KIMURA, M. & CROW, J. F. (1963). The measurement of effective population number. *Evolution* **17**, 279–288.

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

KIMURA, M. & MARUYAMA, T. (1966). The mutational load with epistatic gene interactions in fitness. *Genetics* **54**, 1337–1351.

KIMURA, M. & WEISS, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.

LEWONTIN, R. C. & HUBBY, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

MUKAI, T. (1964). The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics* **50**, 1–19.

MULLER, H. J. (1967). The gene material as the initiator and the organizing basis of life. In *Heritage from Mendel* (ed. R. A. Brink), pp. 419–447. Madison: Univ. of Wisconsin Press.

ROBERTSON, A. (1962). Selection for heterozygotes in small populations. *Genetics* **47**, 1291–1300.

ROBERTSON, A. (1967). The nature of quantitative genetic variation. In *Heritage from Mendel*, pp. 265–280. (ed. R. A. Brink). Madison: Univ. of Wisconsin Press.

SHAW, C. R. (1965). Electrophoretic variation in enzymes. *Science, N.Y.* **149**, 936–943.

SMITH, M. H. (1966). The amino acid composition of proteins. *J. Theoret. Biol.* **13**, 261–282.

SONNEBORN, T. M. (1965). Degeneracy of the genetic code: Extent, nature, and genetic implications. In *Evolving Genes and Proteins*, pp. 377–397 (ed. Bryson, V. and Vogel, H. J.). New York: Academic Press.

SUEOKA, N. (1965). On the evolution of informational macromolecules. In *Evolving Genes and Proteins* (ed. Bryson, V. and Vogel, H. J.), pp. 479–496. New York: Academic Press.

WATSON, J. D. (1965). *Molecular Biology of the Gene.* New York: Benjamin.

WEISS, G. H. & KIMURA, M. (1965). A mathematical analysis of the stepping stone model of genetic correlation. *J. appl. Probability* **2**, 129–149.
WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.
WRIGHT, S. (1938*a*). The distribution of gene frequencies under irreversible mutation *Proc. Natn. Acad. Sci. U.S.A.* **24**, 253–259.
WRIGHT, S. (1938*b*). Size of population and breeding structure in relation to evolution. *Science, N.Y.* **87**, 430–431.
WRIGHT, S. (1949). Genetics of populations. *Encyclopaedia Britannica* **10**, 111–112.
WRIGHT, S. (1951). The genetical structure of populations. *Ann. Eugen.* **15**, 323–354.
WRIGHT, S. (1966). Polyallelic random drift in relation to evolution. *Proc. Natn. Acad. Sci. U.S.A.* **55**, 1074–1081.