THE SUBSTITUTIONAL LOAD IN A FINITE POPULATION*

MOTOO KIMURA and TAKEO MARUYAMA
National Institute of Genetics, Mishima, Japan

Received 11.iii.68

1. INTRODUCTION

THE problem of assessing the genotypic selection intensity that accompanies the process of substituting one allele for another in adaptive evolution was first attacked by Haldane (1957), who used the term " cost of natural selection " in describing the amount of selective elimination in the process. Based on a deterministic treatment, he obtained elegant formulae showing that the sum of the fractions of selective deaths is almost independent of the selection coefficient but depends on the initial frequency of the allele used for the substitution. Later, more exact expressions were derived by Haldane (1960), especially to cope with the cases in which the selection coefficient is not small.

On the basis of Haldane's theory Kimura (1960) developed his theory of the optimum mutation rate, where the term " substitutional load " represents the genotypic selection intensity. Later, the theory was re-examined and also the effect of slowly changing environment on the substitutional load was investigated (Kimura, 1967).

The above treatments are all deterministic in that the random fluctuation of gene frequencies due to random sampling of gametes is disregarded.

However, the actual populations are all finite, and, as will be shown in what follows, random sampling of gametes has a very significant effect on the substitutional load. In the present paper, the senior author (M. K.) is responsible for the theoretical treatments and also for the simulation studies on the haploid population, while the junior author (T. M.) is responsible for the simulation studies on the diploid population.

2. HAPLOID POPULATION

Let us consider a population of haploid organisms and denote by N_e the effective population number. N_e roughly represents the number of breeding individuals and may be different from the actual number of adults. For the difference between actual and effective population number, the reader may refer to Kimura and Crow (1963).

We will assume that the population consists of two types of individuals (or alleles) A_1 and A_2, and denote by x and $(1-x)$ their respective frequencies in the population. We will also assume that A_1 has the selective advantage s (>0) over A_2 such that the mean change of x per generation is $M_{\delta x} = sx(1-x)$, or more exactly, that the expected amount of change in x from time t to $t+dt$ is $sx(1-x)dt$. Before proceeding further, it is important to note, for the treatment to be realistic, that the total population number is not directly

* Contribution No. 682 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 a Grant from Toyo Rayon Foundation.

controlled by the numbers and the relative fitnesses of A_1 and A_2, as assumed by Feller (1967), but is controlled by such environmental factors as food, space, competing species and so on. Such an observation is made on the biological basis of a strong tendency inherent in each organism to increase in number, if unchecked. Namely, in each generation, a large number of young are produced, but only a fraction of them can reach maturity so that the total population number is compatible with the carrying capacity of the environment. Therefore, it is quite realistic to assume that the population number is kept nearly constant by the above population controlling mechanism throughout the process of gene substitution. (For a more detailed discussion, see Kimura and Crow, 1969).

Thus, we will assume that the effective population number N_e is constant, and, in each generation random sampling of gametes (or spores) for the production of next generation takes place in such a way that the variance of the change in gene frequency x per generation is $V_{\delta x} = x(1-x)/N_e$.

The process of change of the frequency (x) of A_1 is now a stochastic process in which x fluctuates from generation to generation and, eventually, either A_1 reaches fixation (*i.e.* x becomes 1) or lost (*i.e.* becomes 0) from the population. We will denote by $u(p)$ the probability of eventual fixation of A_1 when its initial frequency is p. For the gene with selective advantage s in a haploid population of effective size N_e as considered here, the more general formula for $u(p)$ derived by Kimura (1957) reduces to

$$u(p) = (1-e^{-2N_e sp})/(1-e^{-2N_e s}). \tag{2.1}$$

Our aim is to calculate the sum total of the genetic load that accompany the process through which A_1 changes from a low frequency p to a very high frequency and finally to fixation. For this purpose, we will use the method of diffusion equations, especially the one of making use of the Kolmogorov backward equation as developed by Kimura (Kimura, 1957, 1962, 1964).

Let $\phi(p, x; t)$ be the probability density that the frequency of A_1 becomes x at time t (measured one generation as unit) given that it is p at time 0. Then ϕ satisfies the following Kolmogorov backward equation

$$\frac{\partial \phi(p, x; t)}{\partial t} = \frac{p(1-p)}{2N_e} \frac{\partial^2 \phi(p, x; t)}{\partial p^2} + sp(1-p) \frac{\partial \phi(p, x; t)}{\partial p} \tag{2.2}$$

In a particular population containing A_1 and A_2 at the relative frequencies x and $1-x$, the mean fitness of the popultion is less by $s(1-x)$ as compared with the fitness of the optimum genotype A_1, so that the load in this population is

$$l(x) = s(1-x). \tag{2.3}$$

Since the probability is $\phi(p, x; t)dx$ that the frequency of A_1 is x at time t, the expected value of the sum total, denoted by $F(p)$, of the load from time $t = 0$ to time $t = \infty$ is

$$F(p) = \int_0^\infty \left[\int_0^1 l(x)\phi(p, x; t)dx \right] dt, \tag{2.4}$$

in which the integral with respect to x is strictly over the open interval (0, 1) *i.e.* for $0 < x < 1$. It might perhaps be more appropriate to write the limits of integration as $1/2N$ and $1-1/2N$, but, for the sake of simplicity, we write

those limits as 0 and 1. Here we may note that since $s(1-x)dt$ is the amount of selective elimination during a short time interval from t to $t+dt$ in a population containing A_1 and A_2 with relative proportions x and $1-x$, $F(p)$ also represents the expected sum total of the amount of selective elimination that takes place from time 0 to ∞. However, gene A_1 is eventually fixed only with probability $u(p)$ and therefore the load for one gene substitution should be defined by

$$L(p) = F(p)/u(p) \tag{2.5}$$

as pointed out by Maruyama (1967).

We will now proceed to calculate $F(p)$ by using equation (2.2). Multiplying $l(x) = s(1-x)$ on each term of equation (2.2) and integrating the resulting terms first with respect to x on the open interval (0, 1) followed by integrating them with respect to t from $t = 0$ to ∞, we obtain

$$\int_0^\infty \left[\frac{\partial}{\partial t}\int_0^1 s(1-x)\phi(p, x; t)dx\right]dt = \frac{p(1-p)}{2N_e}\left\{\frac{\partial^2 F(p)}{\partial p^2} + 2N_e s\frac{\partial F(p)}{\partial p}\right\}. \tag{2.6}$$

The left hand side of the above equation reduces to

$$\int_0^1 s(1-x)\phi(p, x; \infty)dx - \int_0^1 s(1-x)\phi(p, x; 0)dx = -s(1-p), \tag{2.7}$$

which follows from the facts that (i) the boundaries $x = 0$ and 1 act as absorbing barriers so that

$$\phi(p, x; \infty) = 0$$

for $0 < x < 1$, over which the integral is defined, and, (ii) the initial frequency of A_1 is p so that

$$\phi(p, x; 0) = \delta(x-p),$$

where $\delta(\cdot)$ represents the Dirac delta function.

Combining (2.6) and (2.7), we obtain the following equation for $F(p)$,

$$\frac{d^2F(p)}{dp^2} + 2S\frac{dF(p)}{dp} + \frac{2S}{p} = 0, \tag{2.8}$$

where

$$S = N_e s.$$

The above differential equation may immediately be integrated to give

$$F(p) = C_0 - C_1 e^{-2Sp} + 2S\int_0^p e^{-2S\lambda}\,d\lambda\int_\lambda^1 \frac{e^{2Sx}}{x}\,dx, \tag{2.9}$$

in which two constants C_0 and C_1 may be determined by the boundary conditions,

$$F(0) = F(1) = 0. \tag{2.10}$$

These conditions follow from the fact that, for all x in the interval $0 < x < 1$,

$$\phi(p, x; t) = 0$$

when $p = 0$ or $p = 1$.

Using the boundary conditions (2.10), (2.9) reduces to

$$F(p) = \{1-u(p)\}\int_0^{2Sp} \frac{e^y-1}{y}\,dy - u(p)e^{-2S}\int_{2Sp}^{2S} \frac{e^y}{y}\,dy + u(p)\log_e\left(\frac{1}{p}\right),$$

so that the required formula for the load is

$$L(p) = \frac{F(p)}{u(p)} = \left\{\frac{1}{u(p)} - 1\right\}\int_0^{2Sp} \frac{e^y-1}{y}\,dy - e^{-2S}\int_{2Sp}^{2S} \frac{e^y}{y}\,dy + \log_e\left(\frac{1}{p}\right), \tag{2.11}$$

where $S = N_e s$ and $u(p)$ represents the probability of fixation given by $u(p) = (1-e^{-2Sp})/(1-e^{-2S})$.

The numerical calculation of the load from equation (2.11) is facilitated by noting that the integrals in the right hand side of the equation may be expressed in terms of the exponential integral as follows:

$$\int_0^{2Sp} \frac{e^y-1}{y}\,dy = E_i(2Sp) - \log_e(2Sp) - \gamma$$

$$\int_{2Sp}^{2S} \frac{e^y}{y}\,dy = E_i(2S) - E_i(2Sp)$$

In the above expressions, γ is Euler's constant (0·57721 ...) and $E_i(\cdot)$ is the exponential integral defined by

$$E_i(x) = \int_{-\infty}^{x} \frac{e^t}{t}dt, \qquad (x>0)$$

for which fairly extensive tabulation is available (see for example, Abramowitz and Stegun, 1965). However, for most practical purposes, the following series approximations seem to be sufficient:

For a small value of x (>0)

$$E_i(x) = \gamma + \log_e x + x + \frac{x^2}{4} + \ldots,$$

and, for a large value of x

$$E_i(x) = e^x\left\{\frac{1}{x} + \frac{1}{x^2} + \frac{2!}{x^3}\ldots\right\}.$$

Here we will consider three cases. First, if both $2S(=2N_e s)$ and $2Sp(=2N_e sp)$ are infinitely large,

(2.11) reduces to

$$L(p) = \log_e\left(\frac{1}{p}\right), \tag{2.12}$$

the result first obtained by Haldane (1957). Secondly, if $2S$ is much larger than unity but $2Sp$ is small so that $u(p) = 2Sp$ approximately, we have

$$L(p) = 1 + \log_e\left(\frac{1}{p}\right) \tag{2.13}$$

approximately. For one gene substitution, on the average $(2Sp)^{-1}-1$ equally advantageous genes are lost on the way, each of the latter contributing about $2Sp$ to the load, thus creating extra load of $1-(2Sp)$ or roughly 1 as compared to the first case considered. We believe that this is a new contribution to the concept of substitutional load. Thirdly, if both $2S$ and $2Sp$ are much smaller than unity so that $u(p) = p+Sp(1-p)$ approximately, we have roughly

$$L(p) = 2S\log_e\left(\frac{1}{p}\right),$$

which is much smaller than the load in the first case. Namely, in a small population, a slightly advantageous genes may be substituted with a small load.

Fig. 1 illustrates $L(p)$ as a function of S assuming four different levels of p. The curves in the figure were drawn using values obtained from equation (2.11) by numerical integration.

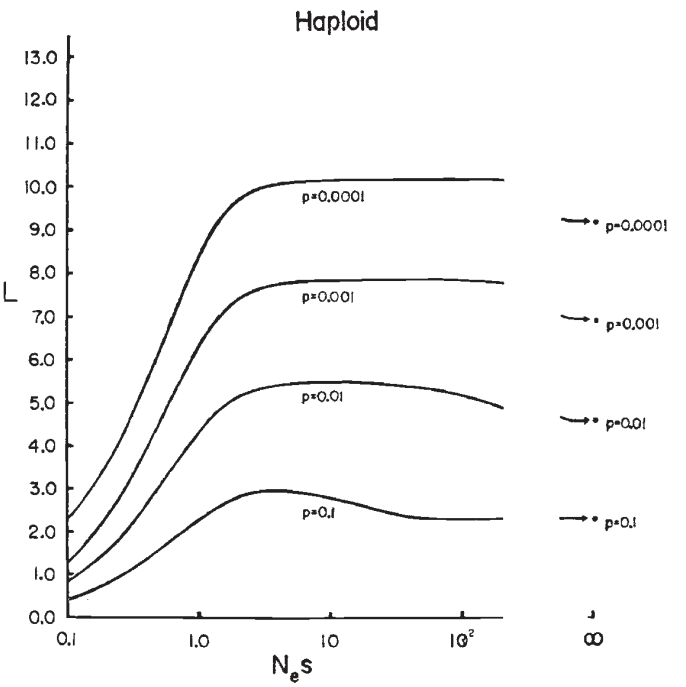

FIG. 1.—The load for one gene substitution (L) in a haploid population as a function of $N_e s$, where N_e is the effective population number and s is the selective advantage of A_1. The relationship between L and $N_e s$ is illustrated for four different initial frequencies, *i.e.* $p = 0{\cdot}1$, $0{\cdot}01$, $0{\cdot}001$ and $0{\cdot}0001$.

3. DIPLOID POPULATION

Let us consider a random mating population consisting of N diploid individuals and having effective population number N_e, and, assume a pair of alleles A_1 and A_2 at an autosomal locus. We will denote by x and $1-x$ the respective frequencies of A_1 and A_2 in the population. We will also denote

by s and sh the selective advantages of A_1A_1 and A_1A_2 over A_2A_2 measured in Malthusian parameters. Then, the mean and the variance of the change in the frequency of A_1 per generation may be given respectively by

$$M_{\delta x} = sx(1-x)\{h+x(1-2h)\}$$

and

$$V_{\delta x} = x(1-x)/(2N_e).$$

Let $\phi(p, x; t)$ be the probability density that the frequency of A_1 becomes x at time t (tth generation) given that it is p at time 0. Then ϕ satisfies the equation

$$\frac{\partial\phi(p,x;t)}{\partial t} = \frac{p(1-p)}{4N_e}\frac{\partial^2\phi(p,x;t)}{\partial p^2} + sp(1-p)\{h+(1-2h)p\}\frac{\partial\phi(p,x;t)}{\partial p}. \quad (3.1)$$

In a particular population in which the frequency of A_1 is x, the average fitness of the population is less by $s-[sx^2+sh2x(1-x)]$ as compared with the optimum genotype A_1A_1 so that the load is

$$l(x) = s(1-x)\{1+(1-2h)x\}. \quad (3.2)$$

As in the haploid case, we are going to calculate the expected value of the sum total of the load that accompanies the process through which x changes from p to unity. For this purpose, we multiply each term of the above differential equation (3.1) by $l(x)$ and integrate the resulting terms first with respect to x over the open interval (0, 1) and then with respect to t from $t = 0$ to ∞. This leads to

$$\int_0^\infty\left[\frac{\partial}{\partial t}\int_0^1 l(x)\phi(p,x;t)dx\right]dt$$

$$= \frac{p(1-p)}{4N_e}\frac{\partial^2F(p)}{\partial p^2} + sp(1-p)\{h+(1-2h)p\}\frac{\partial F(p)}{\partial p}, \quad (3.3)$$

where

$$F(p) = \int_0^\infty\left[\int_0^1 l(x)\phi(p,x;t)dx\right]dt.$$

Since $\phi(p, x; \infty) = 0$ for $0<x<1$, and $\phi(p, x; 0) = \delta(x-p)$, the left hand side of (3.3) becomes

$$\int_0^1 l(x)\phi(p,x;\infty)dx - \int_0^1 l(x)\phi(p,x;0)dx$$

$$= -l(p) = -s(1-p)\{1+(1-2h)p\}.$$

Thus, we have the following differential equation for $F(p)$.

$$\frac{d^2F(p)}{dp^2} + 4S\{h+(1-2h)p\}\frac{dF(p)}{dp} + 4S\left\{\frac{1}{p}+(1-2h)\right\} = 0, \quad (3.4)$$

where $S = N_es$.

The equation can immediately be integrated to give

$$F(p) = C_0 + C_1\int_0^p G(x)dx + 4S\int_0^p G(\lambda)d\lambda\int_\lambda^1\left\{\frac{1}{x}+(1-2h)\right\}G^{-1}(x)dx, \quad (3.5)$$

where C_0 and C_1 are constants and

$$G(x) = e^{-2S[2hx+(1-2h)x^2]}. \tag{3.6}$$

The constants may be determined by the boundary conditions

$$F(0) = F(1) = 1, \tag{3.7}$$

which follows from the fact that when $p = 0$ or $p = 1$, $\phi(p, x; t) = 0$ for $0 < x < 1$. Using these boundary conditions, (3.5) becomes

$$\begin{aligned} F(p) = {} & 4S\{1-u(p)\}\int_0^1 \left\{\frac{1}{x}+(1-2h)\right\} G^{-1}(x)dx \int_0^x G(\lambda)d\lambda \\ & -4S\int_p^1 \left\{\frac{1}{x}+(1-2h)\right\} G^{-1}(x)dx \int_p^x G(\lambda)d\lambda \end{aligned} \tag{3.8}$$

where

$$u(p) = \int_0^p G(x)dx \Big/ \int_0^1 G(x)dx, \tag{3.9}$$

in which $G(x)$ is given by (3.6), is the probability of the ultimate fixation of A_1 (Kimura, 1962). The load for one gene substitution is then given by

$$L(p) = F(p)/u(p). \tag{3.10}$$

In the simplest but important case of " no dominance ", that is, when the mutant gene is semidominant so that $h = 1/2$, we have $G(x) = e^{-2Sx}$,

$$u(p) = (1-e^{-2Sp})/(1-e^{-2S}) \tag{3.11}$$

and the expression for the load is simplified to give

$$L(p) = 2\left\{\frac{1}{u(p)}-1\right\}\int_0^{2Sp} \frac{e^y-1}{y}dy - 2e^{-2S}\int_{2Sp}^{2S} \frac{e^y}{y}dy + 2\log_e\left(\frac{1}{p}\right) \tag{3.12}$$

with $u(p)$ given by (3.11). It is interesting to note that the above expression for $L(p)$ is twice the corresponding expression for the haploid case (2.11). Note, however, that the definition of the selection coefficient s is different for the two cases. Namely, in the haploid case s represents the selective advantage of A_1 over A_2, while in the diploid case with no dominance s represents the selective advantage of A_1A_1 over A_2A_2. Except for such a reservation, discussions given in the previous section for the haploid case apply to the present case.

4. Simulation studies

4.1. *Haploid population.* In order to check the validity of equation (2.11), Monte Carlo experiments were carried out by using computer IBM 7090. The computer program was written in Fortran II to simulate the process of selection in a finite population of haploids, in which the actual number (N) of individuals is equal to the effective number (N_e). The selective values of $1+s$ and 1 were assigned respectively to A_1 and A_2. Starting from Np and $N(1-p)$ individuals of A_1 and A_2, a simulation experiment was continued until one of the alleles became fixed in the population. In each generation, N individuals were sampled to form the next generation in such

a way that in each step of sampling, a pseudo random number R having uniform distribution was generated (using Subroutine RAND 1), and, A_1 was added to the next generation if $R \leqq X$, while A_2 was added if $R > X$, where X is the expected frequency of A_1 after selection. This was continued until N individuals were sampled.

For each experiment, the cumulative total of the load over all the generations was calculated. Then, in order to obtain the substitutional load, a number of such experiments (usually consisting of 200 replicate trials) were carried out with a given set of values of s, p and N, and the average load was computed for those cases in which A_1 was eventually fixed in the population.

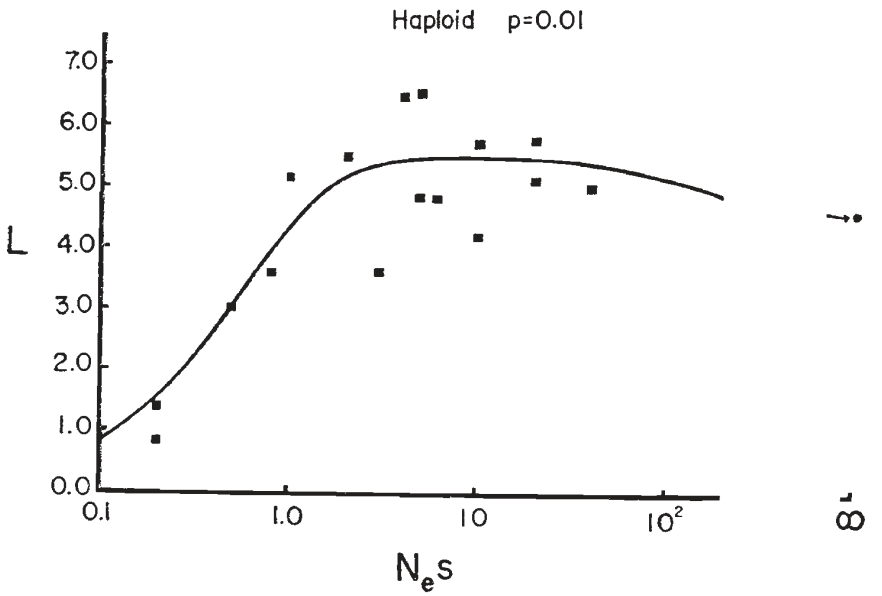

FIG. 2.—Results of Monte Carlo experiments on the substitutional load in a finite population of haploids. The solid line represents theoretical values of L as a function of $N_e s$, where L is the genetic load for one gene substitution, N_e is the effective population number and s is the selection coefficient. Results of the experiments are plotted with square dots. See also table 1.

Fig. 2 illustrates the results of Monte Carlo experiments performed by setting initial gene frequency $p = 0{\cdot}01$. In the figure, the solid line represents the theoretical values of L as a function of $N_e s$. Those values were obtained from formula (2.11) by numerical integration. The value of L at $N_e s = \infty$ is 4·605. The results of the Monte Carlo experiments are plotted with square dots. More detailed results are given in table 1. The agreement between the theoretical predictions and the experimental results appears to be satisfactory.

4.2. *Diploid population.* A similar computer program was written in Fortran IV to simulate the process of selection in a finite population of diploid individuals which are monoecious and among which mating takes place at random. In the simulation process, the effective number (N_e) was set equal to the actual number. The selective values $1+s$, $1+sh$ and 1 were respectively assigned to A_1A_1, A_1A_2 and A_2A_2.

TABLE 1

Results of Monte Carlo experiments on the substitutional load in a haploid population. In all the experiments listed here, the initial frequency of A_1 *was taken as* 0·01 (p=0·01). *The theoretical values were computed from equation* (2.11). N_e *is the effective population number and* s *is the selection coefficient for* A_1.

Case	N_e	s	Number of replicate trials	Number fixed	Substitutional Load (L)	
					From experiment	From theory
1	100	0·002	200	5	0·83	1·55
2	100	0·002	800	11	1·38	1·55
3	100	0·005	200	4	3·01	3·05
4	100	0·008	200	3	3·61	3·95
5	100	0·01	200	3	5·11	4·34
6	100	0·02	200	8	5·50	5·17
7	100	0·03	200	16	3·62	5·37
8	100	0·04	200	10	6·46	5·44
9	100	0·05	200	12	6·52	5·47
10	100	0·05	200	20	4·83	5·47
11	100	0·06	200	22	4·82	5·48
12	100	0·10	200	29	5·73	5·50
13	100	0·10	200	42	4·20	5·50
14	100	0·20	200	54	5·11	5·48
15	200	0·10	100	27	5·78	5·48
16	200	0·20	100	57	5·01	5·40

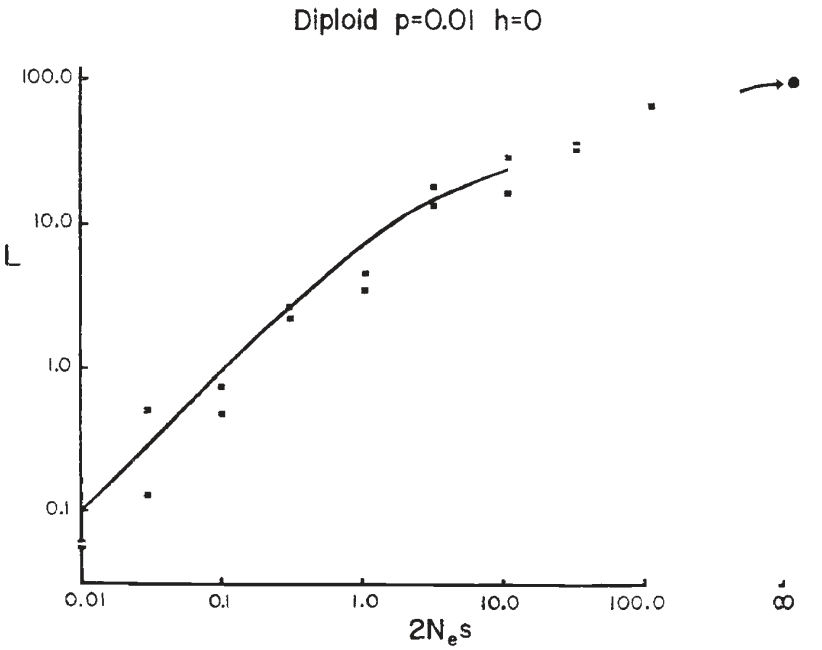

FIG. 3.—Results of Monte Carlo experiments on the substitutional load in a finite population of diploids, assuming that the advantageous mutant gene is completely recessive (h = 0) and its initial frequency p = 0·01. Results of the experiments are plotted with square dots. For details, see text and also table 2.

In Fig. 3, the results of the Monte Carlo experiments, assuming population number $N = N_e = 50$, initial frequency $p = 0{\cdot}01$ and degree of dominance $h = 0$ (A_1 completely recessive) are represented. The square dots in the figure indicate the results of the experiments. These should be compared with the theoretical curve derived from formulae (3.8), (3.9) and

TABLE 2

Results of Monte Carlo experiments on the substitutional load in a finite population of diploid individuals. In all the experiments listed here, the experiments were carried out assuming the effective population number $N_e=50$, *the initial frequency* $p=0{\cdot}01$ *and complete recessiveness of mutant gene* A_1, i.e. $h=0$. *Also, the selective values* $1+s$ *and* 1 *were assigned respectively to the recessive* (A_1A_1) *and dominant* (A_2A_2 *and* A_1A_2) *individuals. The theoretical values were obtained from equations* (3.8), (3.9) *and* (3.10) *by numerical integration.*

	Substitutional Load (L)		
	From Monte Carlo experiments		
$2N_es$	300 replicates	200 replicates	From theory (Diffusion approximation)
0·01	0·05625	0·05766	0·1016
0·02	—	—	0·2026
0·03	0·1328	0·5113	0·3029
0·05	—	—	0·5013
0·08	—	—	0·7939
0·1	0·7545	0·4775	0·9858
0·2	—	—	1·907
0·3	2·255	2·715	2·769
0·5	—	—	4·335
0·8	—	—	6·350
1·0	4·656	3·612	7·509
2·0	—	—	11·81
3·0	18·79	13·49	14·63
5·0	—	—	18·32
8·0	—	—	22·36*
10·0	30·30	17·27	24·45*
30·0	36·98	35·76	—
100·0	—	68·86	—
∞	—	—	103·6

* A value obtained by Gaussian quadrature using the Legendre polynomials with the degree 40.

(3.10) by setting $p = 0{\cdot}01, h = 0$ and applying numerical integration. More detailed values are listed in table 2. Note that in the theoretical treatments based on the diffusion model, L depends only on $2N_es$ when p and h are given. In the present case, the numerical integration presented some difficulty since we had to cope with double integrals. The method used to compute a double integral such as

$$\int_0^1 \left[\frac{1}{x} + (1-2h)\right] G^{-1}(x)dx \int_0^x G(\lambda)d\lambda$$

appearing in equation (3.8) was as follows. The integral was approximated by

$$\sum_{n=1}^{M} I(x_n)(1/M),$$

where

$$x_n = -(1/2M) + n/M$$

and

$$I(x_n) = \left[\frac{1}{x_n} + (1-2h)\right] G^{-1}(x_n) \int_0^{x_n} G(\lambda)d\lambda,$$

in which a value for

$$\int_a^b G(x)dx$$

was computed by the Gaussian quadrature with the degree of the Legendre polynomial used equals to 10 (*cf.* Hildebrand, 1956). The values given in table 2 were obtained with $M = 1000$. The theoretical value at $2N_e s = \infty$, that is 103·6, was computed from $L = (1-p)/p + \log_e(1/p)$ the formula that can be obtained for the completely recessive gene using deterministic theory.

It may be seen from the present study that for the completely recessive genes the value obtained by the deterministic theory can be approached very slowly as the effective population number becomes infinity. Thus, even for $2N_e s = 100$, the load is roughly 65 per cent. as large as the limiting value. Note that $2N_e s = 100$ means $N_e = 5000$ when the selective advantage is 1 per cent.

In addition to the above simulation experiments, numerical studies were also made by multiplying the transition matrix, assuming small population numbers such as $N = 10$, 20 and 30. The results suggest that the theoretical treatment based on the diffusion model gives fairly good approximation already at $N = 30$. For example, with $p = 0{\cdot}05$, $2N_e s = 1$ and $h = 1/2$, we obtain $L = 4{\cdot}066$ from the diffusion model, while $L = 3{\cdot}799$ from matrix multiplication assuming $N = 30$.

5. Discussion

So far, we have considered the cumulative total amount of selective elimination that accompanies the process of substituting one allele for another in a finite population. This total amount is spread over many generations.

Let us now consider a situation in which a large number of loci are available for gene substitution and mutant genes acquire a selective advantage on the average in ν_m of the loci in each generation. We will assume that whenever this happens it takes place in a different locus and that the selective advantage of the mutant gene is s in homozygotes and sh in heterozygotes. In a population consisting of N diploid individuals, the initial frequency p is equal to $1/(2N)$ if the mutant gene is advantageous from the moment of its birth and if every mutant represents a new not pre-existing allele. On the other hand, if the mutant allele is recurrent and initially disadvantageous or neutral but becomes advantageous later due to change of environment, p may sometimes be much larger than $1/(2N)$.

Let us assume then that the above process has proceeded for a large enough number of gene-rations so that the balance is reached between the appearance of advantageous mutations and their random extinction or fixation in the population. Since $\phi(p, x; t)$ is the probability density that the frequency of the mutant gene becomes x after t generations, $\nu_m \phi(p, x; t)dx$

represents the contribution made by the mutant genes that acquired their advantage t generations earlier to the present frequency class having gene frequency $x \sim x+dx$. Noting that the genetic load for a locus with mutant gene frequency x is $l(x) = s(1-x)\{1+(1-2h)x\}$ as given in (3.2), the load in the present generation may be obtained by summing up $\nu_m \phi(p, x; t) l(x) dx$ over all the contributions made by the past generations, *i.e.* $0 \leqslant t \leqslant \infty$, and over all the relevant frequency classes, *i.e.* $0 < x < 1$. Thus if we denote by L_e the substitutional or evolutional load (*cf.* Kimura, 1960), we have, for the present case

$$L_e = \int_0^1 \int_0^\infty \nu_m \phi(p, x; t) l(x) dx dt = \nu_m F(p) = KL(p), \tag{5.1}$$

where $K = \nu_m u(p)$, and, $F(p)$, $u(p)$ and $L(p)$ are respectively given by (3.8), (3.9) and (3.10) in the previous section. Note that $L(p)$ represents the genetic load *for one gene substitution* while L_e represents the substitutional load *at any particular generation* in an equilibrium population in which gene substitution proceeds at the rate K per generation. Note also that the last mentioned figure $K = \nu_m u(p)$ represents the average number of gene substitutions in the population per generation and it may be much smaller than ν_m, the average number of loci in which mutant genes become advantageous per generation, because in each such locus the mutant gene becomes eventually fixed only with probability $u(p)$. For such an equilibrium population, it can be shown that the average number of heterozygous loci per individual due to the advantageous mutant genes is

$$H(p) = 4\left(\frac{\nu_m}{s}\right)[u(p)-p] \tag{5.2}$$

assuming that an infinite number of loci are available for gene substitution and that the mutant gene is semidominant ($h = 1/2$). The derivation of the above formula will be published elsewhere.

In the following, we will consider two cases of special interest, assuming that the mutant genes are semidominant.

First, if the selective advantage is sufficiently large such that $2N_e s \gg 1$ while the initial frequency is very low so that $2N_e sp \ll 1$, we have approximately $u(p) = 2N_e sp$ and $L(p) = 2[1+\log_e(1/p)]$. Namely, the load for one gene substitution is larger by about two as compared with the corresponding value derived by Haldane (1957) who used a deterministic treatment. In a population consisting of N diploid individuals, if ν_m advantageous mutations are produced in each generation in the population and if each mutant represents a new not pre-existing allele in a different locus, $p = 1/(2N)$. In this case, the mutation rate *per gamete* per generation for advantageous mutation is $v = \nu_m/(2N)$. The probability of fixation of each mutant gene is $u = (N_e/N)s$, where s is the selective advantage of the mutant homozygote. The load for one gene substitution is $L = 2[1+\log_e(2N)]$. For example, in a population consisting of 50,000 individuals ($2N = 10^5$) and having the effective population number half as large as the actual number ($N_e = N/2$), if the selective advantage of each mutant gene in single dose is one per cent. ($s/2 = 0{\cdot}01$), the probability of ultimate fixation of each mutant gene is $u = 0{\cdot}01$ and the load for one gene substitution is $L = 25{\cdot}0$. In order that the gene substitution proceeds at the rate of 1 in every 300 generations, the

rate suggested by Haldane (1957) as a representative figure in the ordinary process of evolution, we must have $K = v_m u = 1/300$. Then, the substitutional load in any given generation is $L_e = KL = 0{\cdot}083$. Namely, the amount of selective elimination which is required for the adaptive evolution to proceed at the above rate is 8·3 per cent. per generation. In such a population, the advantageous mutations occur at the rate $v = (1/3) \times 10^{-5}$ per gamete per generation. From equation (5.2) we find that the number of heterozygous loci per individual due to such advantageous mutations is only about 0·7, a very small number.

Secondly, if the mutant gene is almost neutral such that $|2N_e s| \ll 1$, we have approximately $u(p) = p + N_e s p(1-p)$ and $L(p) = 4N_e s \log_e(1/p)$. Namely, as $2N_e s$ approaches zero, the probability of fixation approaches p and the substitutional load may become indefinitely small. For such mutations, there will be no limit to the rate of gene substitution in evolution, provided that mutant genes are produced at correspondingly high rate. Comparative studies of amino acid arrangement of a protein molecule such as hemoglobin or cytochrome *c* among different groups of animals suggest that in mammalian evolution gene substitution had proceeded at the rate of some two nucleotide replacements per generation (Kimura, 1968*b*). This is a surprisingly high rate of gene substitution. It is probable that a majority of such molecular mutations are almost neutral for natural selection (Kimura, 1968*a*) and that the mutation rate for them is very high, amounting to more than one per gamete per generation.

In recent years it has often been claimed that selection coefficients involved in genetic changes of natural populations are in general very large. Certainly, several remarkable cases have been reported including the spread of melanic forms in industrial melanism. However, it might be premature to think that they represent a typical case of gene substitution in evolution, especially when over all genetic loci are considered. In this connection, we should note that a typical mammalian genome could code for some two millions of polypeptide chains each consisting of 500 amino acids and having a size almost five times as large as the mammalian cytochrome *c*. With such a large number of genetic sites, a possibility can not be excluded that an average individual in a large panmictic population is heterozygous at 20 thousands or more of such genetic sites due to steady flux of molecular mutations.

6. Summary

1. In a finite population, the amount of selective elimination that accompanies the process of substituting one allele for another by natural selection (substitutional load) depends not only on the initial gene frequency (p) but also on the product of the effective population number and the selection coefficients. This problem was formulated and solved by the method of diffusion equations.

2. It was found that random sampling of gametes has a significant effect on the substitutional load.

3. In the simplest but important case in which the mutant gene is semidominant, the following results were obtained for a diploid population of effective size N_e and the mutant gene having selective advantage $s/2$ in heterozygotes and s in homozygotes: (i) If the selective advantage is large enough such that $2N_e s \gg 1$, while the initial frequency p of the mutant gene

is so low that $2N_e sp \ll 1$, the load for one gene substitution denoted by $L(p)$ is larger by about two as compared with the corresponding result obtained by Haldane who used a deterministic treatment. (ii) If the mutant gene is almost neutral such that $|2N_e s| \ll 1$, the load $L(p)$ is approximately $4N_e s \log_e(1/p)$. Namely, as $2N_e s$ approaches zero, $L(p)$ may become indefinitely small. For such mutations, there will be no limit to the rate of gene substitution in evolution, provided that mutant genes are produced at a correspondingly high rate.

4. Simulation studies were also performed to check the validity of the formulae derived by analytical treatments.

7. References

ABRAMOWITZ, M., AND STEGUN, I. A. (ed.). 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. U.S. Department of Commerce, Washington, D.C.

FELLER, W. 1967. On fitness and the cost of natural selection. *Genet. Res., 9,* 1-15.

HALDANE, J. B. S. 1957. The cost of natural selection. *J. Genet., 55,* 511-524.

HALDANE, J. B. S. 1960. More precise expressions for the cost of natural selection. *Jour. Genetics, 57,* 351-360.

HILDEBRAND, F. B. 1956. Introduction to Numerical Analysis. McGraw-Hill, New York.

KIMURA, M. 1957. Some problems of stochastic processes in genetics. *Ann. Math. Stat. 28,* 882-901.

KIMURA, M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *Jour. Genet., 57,* 21-34.

KIMURA, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics, 47,* 713-719.

KIMURA, M., AND CROW, J. F. 1963. The measurement of effective population number. *Evolution, 17,* 279-288.

KIMURA, M. 1964. Diffusion models in population genetics. *Jour. Applied Probability, 1,* 177-232.

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

KIMURA, M. 1968*a*. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res. 11,* 247-269.

KIMURA, M. 1968*b*. Evolutionary rate at the molecular level. *Nature, 217,* 624-626.

KIMURA, M., AND CROW, J. F. 1969. Natural selection and gene substitution. *Genet. Res.* (in press).

MARUYAMA, T. 1967. An application of Kimura's formulae to define the evolutionary load in a small population. *Ann. Rep. National Inst. Genetics, Japan No. 17,* pp. 72-73.