Genet. Res., Camb. (1969), **13**, *pp.* 127–141
Printed in Great Britain

Natural selection and gene substitution*

By MOTOO KIMURA AND JAMES F. CROW
National Institute of Genetics, Mishima, Japan, and
University of Wisconsin, Madison, Wisconsin, U.S.A.

(*Received* 21 *May* 1968)

1. INTRODUCTION

A fundamental property of natural selection is that under its operation a more advantageous gene can gradually supplant less advantageous genes in a population without appreciably affecting the total population number.

Because of this property, an advantageous gene combination which was initially very rare or non-existent can finally emerge as the prevailing type in a population whose total number is always restricted by the carrying capacity of the environment. On the other hand, without natural selection, in order to produce even a single individual having advantageous but originally rare genes simultaneously at many loci, the total population number would have to be larger than the entire earth can support.

It is important to note that in each generation in most species a much larger number of young and vastly larger numbers of gametes than adult individuals are produced. Of those young individuals, only a limited number can survive, reaching maturity and serving as parents for the next generation. This strong tendency to increase and the restricted carrying capacity of the environment together with the genetic variation supply the basis for natural selection. In fact, Darwin (1859) in his *Origin of Species* states:

> Owing to this struggle for life, any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species, in its infinitely complex relations to other organic beings and to external nature, will tend to the preservation of that individual, and will generally be inherited by its offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection, in order to mark its relation to man's power of selection.

A realistic treatment of the process of gene frequency change must take into account that the total population number may change very little or not at all during the time that gene frequencies are changing enormously. Feller (1966, 1967) has called attention to a number of interesting paradoxes that arise when the genotypic fitnesses are treated as measures of increase or decrease in actual numbers. In particular he has pointed out that the 'cost of natural selection' may be

* Contribution no. 658 from the National Institute of Genetics, Mishima, Japan. Aided in part by a grant-in-aid from the Ministry of Education, Japan, and also, by a grant from Toyo Rayon Foundation.

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

much less than Haldane (1957) calculated because of Haldane's neglect of the resulting change in population number. We believe, however, that Feller has misinterpreted Haldane's interest. Haldane, we believe, was interested in gene proportions, not numbers, and the cost should be measured in these terms.

In the conventional treatment of the change in (relative) gene frequencies which was extensively applied first by Haldane (1924) and which has been widely used by geneticists, no particular assumption is made regarding the total population number so that it has the advantage of being applicable to expanding, contracting and stationary populations.

The purposes of this paper are twofold. The first is to show that under a variety of models of change in the total population number, equations of the general form

$$dp/dt = sp(1-p),$$

(where p is the *proportion* of a gene or genotype and s is constant, or nearly so) provide a reasonable description of the increase in the proportion of the favoured type. The second purpose is to show that Haldane's (1957) 'cost of natural selection' gives meaningful results, free of the difficulties brought out by Feller, when the formulae deal with the proportions (not numbers) of genes in the process of substitution. We shall also extend Haldane's results to include epistasis and the effect of random drift in small populations.

Throughout this paper we shall consider only haploids, since the basic principles can be discussed without the complexities of diploidy which add only to the mathematical difficulty without revealing any important additional principles.

2. CHANGE OF THE NUMBER OF GENES BY NATURAL SELECTION

Let us consider a pair of alleles A_1 and A_2 and assume that a population consists of two types of haploid individuals A_1 and A_2 whose numbers are n_1 and n_2 respectively. The treatment applies also if A_1 and A_2 represent individuals of two clones making up a population. Thus the total population number denoted by N is the sum of the two numbers, i.e.

$$N = n_1 + n_2.$$

We will denote by p_1 and p_2 the relative frequencies of the two types so that

$$p_1 = n_1/N \quad \text{and} \quad p_2 = n_2/N = 1-p_1.$$

In this section we will assume that the numbers change continuously with time and no stochastic elements are involved.

To investigate the change of gene numbers by natural selection, we will take as our starting-point the following set of equations:

$$\left.\begin{aligned} dn_1/dt &= n_1[\alpha_1 - f_1(n_1, n_2)], \\ dn_2/dt &= n_2[\alpha_2 - f_2(n_1, n_2)]. \end{aligned}\right\} \tag{2.1}$$

In these expressions, α_1 and α_2 ($\alpha_1, \alpha_2 > 0$) stand for the intrinsic growth rates

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

of n_1 and n_2 respectively. They are the rates at which numbers of A_1 and A_2 would grow if the population had unlimited food supply and room for expansion. Parameter t stands for time and changes continuously, but for practical purposes it may conveniently be measured with the length of one generation as the unit. The functions $f_1(n_1, n_2)$ and $f_2(n_1, n_2)$ represent the control mechanism and they depend on various factors such as the food supply, the space available, accumulation of toxic products, territorial behaviour and so on. Following Fisher (1930), we will call $n_1^{-1}dn_1/dt$ and $n_2^{-1}dn_2/dt$ the Malthusian parameters of A_1 and A_2 and denote them by m_1 and m_2 such that

$$\left.\begin{aligned} m_1 &= \frac{1}{n_1}\frac{dn_1}{dt} = \alpha_1 - f_1(n_1, n_2), \\ m_2 &= \frac{1}{n_2}\frac{dn_2}{dt} = \alpha_2 - f_2(n_1, n_2). \end{aligned}\right\} \tag{2.2}$$

To proceed further, we have to assume more concrete forms of f_1 and f_2. So, we will consider several cases that may be useful in treating the process of gene substitution. Similar models have been studied by Egbert Leigh (personal communication).

(2.1) *Model I. The total population number kept constant, gene replacement according to intrinsic growth rates*

In this model, we will assume that

$$f_1(n_1, n_2) = f_2(n_1, n_2) = \bar{\alpha}g(N), \tag{2.3}$$

where $$\bar{\alpha} = (\alpha_1 n_1 + \alpha_2 n_2)/N = p_1\alpha_1 + p_2\alpha_2$$

is the average intrinsic growth rate and $g(N)$ is a positive increasing function of N. With this assumption, equations (2.1) become

$$\frac{dn_1}{dt} = n_1[\alpha_1 - \bar{\alpha}g(N)], \quad \frac{dn_2}{dt} = n_2[\alpha_2 - \bar{\alpha}g(N)]. \tag{2.4}$$

Then $$\frac{dN}{dt} = \frac{dn_1}{dt} + \frac{dn_2}{dt} = N\bar{\alpha} - N\bar{\alpha}g(N),$$

so that $$\frac{1}{N}\frac{dN}{dt} = \bar{\alpha}[1 - g(N)]. \tag{2.5}$$

In the special case of $g(N) = N/K$ in which K is a positive constant, equation (2.5) represents logistic population regulation, i.e.

$$\frac{dN}{dt} = N\bar{\alpha}\left(1 - \frac{N}{K}\right). \tag{2.6}$$

The total population number is kept at the level of $N = K$ at equilibrium and any departure from this state will be reduced roughly at the rate of $\bar{\alpha}$ per unit time.

At equilibrium in which $N = K$, the rates of change of gene numbers are

$$\frac{dn_1}{dt} = \frac{(\alpha_1 - \alpha_2)\, n_1 n_2}{K}, \quad \frac{dn_2}{dt} = -\frac{(\alpha_1 - \alpha_2)\, n_1 n_2}{K}.$$

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

Going back to equations (2.2) and (2.4), we note that

$$\frac{d}{dt}\log\left(\frac{p_1}{p_2}\right) = \frac{d}{dt}\log n_1 - \frac{d}{dt}\log n_2 = m_1 - m_2 = \alpha_1 - \alpha_2,$$

so that

$$\left(\frac{1}{p_1}+\frac{1}{1-p_1}\right)\frac{dp_1}{dt} = \alpha_1-\alpha_2,$$

or

$$\frac{dp_1}{dt} = sp_1(1-p_1), \qquad (2.7)$$

where $s = \alpha_1 - \alpha_2$. Since m_1 and m_2 stand for the Malthusian parameters of A_1 and A_2 as defined in (2.2), s in the above expression represents the selective advantage of A_1 over A_2 measured in Malthusian parameters. This equation is correct whether N is changing or not.

In the somewhat weaker population control given by $g(N) = (\log_e N)/c$, the total population number is kept at the level of $N = e^c$ at equilibrium, but the equation giving the rate of change of gene frequency is exactly the same as (2.7). More generally, if the conditions (2.3) hold, the change of gene frequency is given by (2.7). Note that this model includes the possibility of no population control as a special case of $g(N) \equiv 0$.

(2.2) *Model II. Weaker population control*

In this model we assume that

$$f_1(n_1, n_2) = f_2(n_1, n_2) = g(N), \qquad (2.8)$$

where $g(N)$ is a positive increasing function of N. This yields

$$\frac{dn_1}{dt} = n_1[\alpha_1 - g(N)], \quad \frac{dn_2}{dt} = n_2[\alpha_2 - g(N)]. \qquad (2.9)$$

Some simple examples of this model may be produced by putting $g(N) = cN$, $g(N) = k\log_e N$ and so on. The equation for the rate of change of the relative proportion of gene A_1 is exactly the same as in the first case, because

$$\frac{d}{dt}\log\left(\frac{p_1}{p_2}\right) = \frac{1}{n_1}\frac{dn_1}{dt} - \frac{1}{n_2}\frac{dn_2}{dt} = \alpha_1 - \alpha_2 \equiv s,$$

or

$$\frac{dp_1}{dt} = sp_1(1-p_1). \qquad (2.10)$$

On the other hand, the total population number (N) increases as one gene replaces the other.

For example, if we take

$$g(N) = cN, \qquad (2.11)$$

then, when gene A_1 has replaced gene A_2 (assuming $\alpha_1 > \alpha_2$), $N = \alpha_1/c$. Thus, in this case the replacement of one gene by another results in a population increase

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

proportional to the difference in the intrinsic rates of increase in the two genes. Such a situation is probably not very common in nature. The size of the population is usually determined by factors other than the α's. A replacement of the original gene with a superior one will probably cause only a slight increase in population number or even none at all, so that model I is probably more realistic.

(2.3) *Model III. Selective advantage due to higher resistance to overcrowding*

In this model we assume that

$$\alpha_1 = \alpha_2 \equiv \alpha, \quad f_1(n_1, n_2) = \alpha g_1(N), \quad f_2(n_1, n_2) = \alpha g_2(N), \tag{2.12}$$

where $g_1(N)$ and $g_2(N)$ are both increasing functions of N such that

$$0 < g_1(N) < g_2(N).$$

This leads to the equations

$$\frac{dn_1}{dt} = \alpha n_1 \, [1 - g_1(N)], \quad \frac{dn_2}{dt} = \alpha n_2[1 - g_2(N)]. \tag{2.13}$$

As an example, let us take

$$g_1(N) = N/K_1, \quad g_2(N) = N/K_2$$

in which $$K_1 > K_2.$$

Equations (2.13) are reduced to

$$\frac{dn_1}{dt} = \alpha n_1 \left(1 - \frac{N}{K_1}\right), \quad \frac{dn_2}{dt} = \alpha n_2 \left(1 - \frac{N}{K_2}\right). \tag{2.14}$$

The two genes have the same intrinsic rates of growth, but A_1 has a selective advantage over A_2, because it is more tolerant of overcrowding. The rate of change of logarithmic gene ratio is

$$\frac{d}{dt}\log\left(\frac{p_1}{p_2}\right) = \frac{1}{n_1}\frac{dn_1}{dt} - \frac{1}{n_2}\frac{dn_2}{dt} = \alpha N \left(\frac{K_1 - K_2}{K_1 K_2}\right) \equiv s. \tag{2.15}$$

The total population number changes from roughly K_2 to K_1 as A_1 increases from very low frequency to very high frequency and finally to fixation. If s is small as compared with α, $K_1 K_2$ is roughly equal to N^2 and the right-hand side of (2.15) shows that $(K_1 - K_2)/N \doteqdot s/\alpha$, namely, replacement of one gene by another results in a population increase proportional to the selective advantage expressed as a fraction of the intrinsic growth rate. In this example s is not constant. However, it changes very slowly, especially if N does not change much with the gene substitution. The situation we have in mind is that the favoured allele replaces the other by introducing a greater resistance to overcrowding. For $K_1 = 1{\cdot}01 K_2$, s changes by only 1 % as the frequency changes from 0 to 1; in other words s is nearly constant.

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

The point that we are stressing is that the equation familiar to geneticists,

$$\frac{dp_1}{dt} = sp_1(1-p_1),$$

represents the change of gene frequency by natural selection for a wide variety of realistic situations.

3. THE COST OF A GENE SUBSTITUTION (SUBSTITUTIONAL LOAD)

We will now consider the selection intensity that accompanies the process of substituting A_1 for A_2 by natural selection, assuming that A_1 has selective advantage s over A_2 ($s > 0$). As pointed out in the introduction, in natural populations many more young than adult individuals are usually produced in each generation, but only a fraction survive to maturity and serve as parents for the next generation. The majority of the premature deaths may be non-genetic, that is to say, they strike A_1 and A_2 with equal probability. The remaining deaths are genetic; that is to say, they are caused by A_1 having a selective advantage s over A_2. What we are concerned with here is the latter component of death, for this is the factor which enables A_1 to increase its frequency and leads to its eventual fixation in the population.

Consider a change from time t to $t+dt$. During this short time interval, the amount of genetic death expressed as a fraction of the total population number is

$$p_2(t)sdt = [1-p_1(t)]sdt, \tag{3.1}$$

where $p_2(t) = 1-p_1(t)$ is the relative frequency of the less advantageous gene A_2 in the population at time t. In the discrete generation time model employed by Haldane, the corresponding quantity is $p_2(t)s$, or $d_n = kq_n$ in his terminology (Haldane, 1957), where k is the selection coefficient against less advantageous gene and q_n is the frequency of that gene at the nth generation. Haldane called kq_n 'the fraction of selective deaths in the nth generation'. The expression (3.1) may also be called the load (measured in Malthusian parameters) due to gene substitution during the time interval t to $t+dt$, since, as seen from model I, the average population fitness

$$\bar{m}(t) = p_1(t)m_1+p_2(t)m_2$$

at time t is less by

$$m_1-\bar{m}(t) = (m_1-m_2)p_2(t) \equiv sp_2(t)$$

than the fitness of the favoured genotype A_1. (For the definition of genetic load, see Crow (1958) and Crow & Kimura (1965).)

On the other hand, Feller (1967) takes Haldane's selective death to mean the actual decrement of the number of disadvantageous genes from generation n to generation $n+1$ (i.e. $N'_n - N'_{n+1}$ in Feller's notation) and regards Haldane's expression $d_n = kq_n$ as an approximation. We would like to stress that Feller's expression $(N'_n - N'_{n+1})$ is something quite different from Haldane's. Feller treated μ, the fertility of the favoured genotype, as a constant (the expression,

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

$N'_n - N'_{n+1}$, is that obtained when $\mu = 1$). With constant μ the population does not attain a stable size, unless $\mu = 1$ in which case the final number is the same as the initial number of A genes. Haldane assumed that the population number is relatively stable and counted the number of deaths each generation as a fraction of the population in that generation.

For a realistic treatment of the problem of gene substitution in evolution we must consider a process in which the advantageous genes (A_1) increases from very low frequency to very high frequency and finally to fixation, while the total population number is kept constant or nearly constant throughout the process by population regulating mechanisms such as those discussed in the previous section. The total load or cost due to the gene substitution may then be obtained by summing the quantity (3.1) during the process through which the frequency of A_1 changes from $p_1(0)$ to unity. Namely,

$$L[p_1(0), 1] = \int_0^\infty [1-p_1(t)]s\, dt. \tag{3.2}$$

To simplify the expression, we will write $L(p_1(0))$ for $L\,[p_1(0), 1]$.

If we use the relation

$$\frac{dp_1(t)}{dt} = sp_1(t)\,[1-p_1(t)], \tag{3.3}$$

or

$$dt = dp_1(t)/\{sp_1(t)\,[1-p_1(t)]\},$$

the integral reduces to

$$\int_0^\infty [1-p_1(t)]s\,dt = \int_{p_1(0)}^1 \frac{[1-p_1(t)]s}{sp_1(t)\,[1-p_1(t)]}\,dp_1(t)$$

$$= \int_{p_1(0)}^1 \frac{dp_1(t)}{p_1(t)}, \tag{3.4}$$

so that

$$L(p_1(0)) = -\log_e p_1(0). \tag{3.5}$$

This is the result first obtained by Haldane (1957).

Thus the total load is independent of the selection coefficient s but depends only on the initial frequency $p_1(0)$. The total load becomes larger the lower the initial frequency of A_1. For example, the total load is roughly 6·9 if $p_1(0) = 10^{-3}$, but it is 13·8 if $p_1(0) = 10^{-6}$. Haldane called this load 'the cost of natural selection'. More generally, the total cost due to the frequency of A_1 increasing from $p_1(0)$ to $p_1(t)$ is

$$L[p_1(0), p_1(t)] = \log_e p_1(t) - \log_e p_1(0). \tag{3.6}$$

It should be noted that the selection coefficient s need not be constant throughout the process. As long as s remains positive, formula (3.5) is valid even if s changes from generation to generation. Actually, in the middle integral of (3.4), s in the numerator and in the denominator cancel each other in each infinitesimal time interval and s does not appear in the right side of (3.4).

The Haldane formula gives a measure of the proportion of selective deaths that must occur if gene substitutions are to take place at the specified rate. For example,

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

if the initial frequency $p_1(0)$ is 0·001, the total cost is 6·9; that is, if one gene substitution is to occur on the average every 69 generations there must be 10% selective deaths each generation. The average fertility must be such that 10% of all zygotes can die before maturity and still maintain the population number approximately constant. Of course, this does not include deaths from causes other than gene substitutions, which may be genetic or environmental.

If the population does not have the required average fertility, it does not necessarily mean that it becomes smaller. It may only mean that it cannot make gene substitutions at the calculated rate. Ultimately, however, this may mean that the species loses to a competing species that can evolve faster. Selection is not necessarily acting through death or complete sterility; it may equally, or perhaps more likely, depend on differences in fecundity. The principle is similar, but less simple to state or measure. For selection by death (or complete sterility) the maximum rate of gene substitution, given the original gene frequency, can be calculated directly from the observed rate of premature deaths (or sterility). For selection by fecundity differences the corresponding limit is set by the percentage reduction in the average productivity of the population compared with that of the selectively favoured gene (or genotype). This may be much more difficult to assess.

4. EFFECT OF EPISTASIS ON THE SUBSTITUTIONAL LOAD

If gene substitution is carried out at two or more independent loci, the total load for all of them is the sum of the substitutional load for each locus, provided that there is no epistatic gene interaction in fitness.

In this section we will investigate the effect of epistasis on the substitutional load assuming a haploid population in which gene substitution is carried out simultaneously at two loci. We will denote by $\mathbf{A}_1$ and $\mathbf{A}_2$ a pair of alleles in the first locus with their respective frequencies $p(t)$ and $1-p(t)$ in the population. Similarly, we will denote by $\mathbf{B}_1$ and $\mathbf{B}_2$ a pair of alleles in the second locus with respective frequencies $q(t)$ and $1-q(t)$. Let us assume that the selective advantage of $\mathbf{A}_1$ over $\mathbf{A}_2$ is s_A in combination with $\mathbf{B}_2$ but is $s_A+\epsilon$ in combination with $\mathbf{B}_1$ (See Table 1).

Table 1. *Frequency and fitness of the four haploid genotypes*

Genotype	Frequency	Relative fitness
$\mathbf{A}_1\mathbf{B}_1$	pq	$s_A+s_B+\epsilon$
$\mathbf{A}_2\mathbf{B}_1$	$(1-p)q$	s_B
$\mathbf{A}_1\mathbf{B}_2$	$p(1-q)$	s_A
$\mathbf{A}_2\mathbf{B}_2$	$(1-p)(1-q)$	0

Similarly, assume that the selective advantage of $\mathbf{B}_1$ over $\mathbf{B}_2$ is s_B in combination with $\mathbf{A}_2$ but is $s_B+\epsilon$ in combination with $\mathbf{A}_1$. Throughout this section, we will restrict our consideration to the case where

$$0 \leqslant s_A < s_A+s_B+\epsilon \quad \text{and} \quad 0 \leqslant s_B < s_A+s_B+\epsilon,$$

namely substituting $\mathbf{A}_1$ for $\mathbf{A}_2$ and $\mathbf{B}_1$ for $\mathbf{B}_2$ increases the fitness.

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

Assuming random combination of genes between the two loci, the rates of change of gene frequencies are

$$\frac{dp}{dt} = p(1-p)(s_A+\epsilon q), \quad \frac{dq}{dt} = q(1-q)(s_B+\epsilon p). \tag{4.1}$$

These are approximations that are valid when $|\epsilon|$ is much smaller than the recombination fraction between the two loci, for which case a 'quasi linkage equilibrium' is established in a few generations (cf. Kimura, 1965).

In a population containing $\mathbf{A}_1$ and $\mathbf{B}_1$ with respective frequencies p and q, if we assume that $\mathbf{A}_1\mathbf{B}_1$ has the highest fitness, the load is

$$l(p, q) = s_A(1-p)+s_B(1-q)+\epsilon(1-pq).$$

Thus the total load for substituting $\mathbf{A}_1$ for $\mathbf{A}_2$ and $\mathbf{B}_1$ for $\mathbf{B}_2$ simultaneously is given by

$$L(p(0), q(0)) = \int_0^\infty l(p, q)\, dt, \tag{4.2}$$

where $p = p(t)$ and $q = q(t)$.

In the special case of equal selective advantages $s_A = s_B \equiv s$ and equal initial frequencies $p(0) = q(0) \equiv p_0$, we have $p(t) = q(t)$ throughout the process so that

$$\frac{dp}{dt} = p(1-p)(s+\epsilon p), \tag{4.3}$$

and

$$l(p, q) = 2s(1-p)+\epsilon(1-p^2), \tag{4.4}$$

where $s \geqslant 0$ and $s+\epsilon > 0$. Using (4.3) and (4.4), (4.2) becomes

$$L(p_0, p_0) = \int_{p_0}^1 \frac{2s+\epsilon(1+p)}{p(s+\epsilon p)}\, dp. \tag{4.5}$$

For $s > 0$, the equation reduces to

$$L(p_0, p_0) = (2+\lambda)\log_e\left(\frac{1}{p_0}\right) - (1+\lambda)\log_e\left(\frac{1+\lambda}{1+\lambda p_0}\right), \tag{4.6}$$

where $\lambda = \epsilon/s$. Here the coefficient λ represents the relative magnitude of gene interaction and lies in the range $-1 < \lambda < \infty$. Table 2 lists values of $L(p_0, p_0)$ corresponding to several values of λ, taking $p_0 = 0{\cdot}001$. The value corresponding to $\lambda = \infty$ was obtained from (4.5) by putting $s = 0$, for which case, we have

$$L(p_0, p_0) = \frac{1}{p_0} - 1 + \log_e\left(\frac{1}{p_0}\right). \tag{4.7}$$

The table reveals the interesting fact that as $s/(2s+\epsilon)$ gets small the cost becomes progressively large. This can be understood intuitively as a consequence of the fact that most of the load happens while the favoured gene is rare. When $s = 0$, $\epsilon > 0$, selection is very slow while the double mutant is rare, a situation comparable to a recessive mutant in a diploid. In fact the formula for a recessive gene in diploids is

$$L(p_0) = \frac{1}{p_0} - 1 + \log_e\left(\frac{1}{p_0}\right), \tag{4.8}$$

exactly the same as (4.7.) The situation is mitigated by the fact that the newly

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

favoured mutant probably already exists in mutational equilibrium in the old environment. If so, the initial frequency of the individual mutant gene would be considerably greater, of the order of the square root of the frequency without epistasis, as with a recessive gene, and the cost is thereby reduced. In the terminology of Kimura & Maruyama (1966), the epistatic interaction with respect to the selective advantages of $\mathbf{A}_1$ and $\mathbf{B}_1$ is of 'reinforcing' type if $\lambda > 0$ and of 'diminishing' type if $-1 < \lambda < 0$. Thus, as far as simultaneous substitution of freely recombining genes is concerned, strong epistasis of the reinforcing type is a hindrance to rapid gene substitutions in evolution. Furthermore, in view of 'diminishing returns' and thresholds that are frequently postulated, it might be expected that diminishing type epistasis is more common among advantageous mutant genes than reinforcing type epistasis.

Table 2. *Some values of substitutional load in a haploid population when there is an epistatic interaction in fitness between two loci.*

(In this table, $s_A = s_B = s$, $\lambda = \epsilon/s$ and the substitutional load was computed from equation (4.6) and (4.7) assuming $p_0(= q_0) = 0{\cdot}001$.)

Interaction		Substitutional load
(λ)	$s/(2s+\epsilon)$	$L(p_0, p_0)$
-1	1	6·9
$-0{\cdot}5$	2/3	10·7
0	1/2	13·8
1	1/3	19·3
10	1/12	56·6
100	1/102	248·1
1000	1/1002	699·7
∞	0	1005·9

5. EFFECT OF SMALL POPULATION NUMBER ON THE SUBSTITUTIONAL LOAD

In a small population, change of gene frequencies by natural selection is subject to random genetic drift and this will affect the substitutional load. Furthermore, even in a very large population, if the initial frequency of the advantageous genes is very low, random fluctuation of the number of such genes in the early stage may produce a significant effect on the load.

Since the details of the treatment of this subject will be published elsewhere (Kimura & Maruyama, 1969), we will summarize in this section the result for the simplest case of a haploid population.

Let N_e be the effective population number and let s be the selective advantage of $\mathbf{A}_1$ over $\mathbf{A}_2$. Assuming that the initial frequency of $\mathbf{A}_1$ is p_0, it can be shown that the total load required to substitute $\mathbf{A}_1$ for $\mathbf{A}_2$ in the population is given by

$$L(p_0) = \left\{\frac{1}{u(p_0)} - 1\right\} \int_0^{2Sp_0} \left(\frac{e^y - 1}{y}\right) dy - e^{-2S} \int_{2Sp_0}^{2S} \frac{e^y}{y}\, dy + \log_e \left(\frac{1}{p_0}\right), \qquad (5.1)$$

where $S = N_e s$ and $u(p_0)$ stand for the probability of fixation of A_1, i.e.

$$u(p_0) = \frac{1-e^{-2Sp_0}}{1-e^{-2S}} \tag{5.2}$$

(cf. Kimura, 1957). In the above expression (5.1), if both $S(= N_e s)$ and Sp_0 are very large, only the last term in the right-hand side is significant and (5.1) agrees with (3.5); as expected, it reduces to the value obtained by ignoring random drift.

If, on the other hand, the effective population number is so small that $2N_e s$ is much smaller than unity, (5.1) becomes approximately

$$L(p_0) = 2N_e s \log_e \left(\frac{1}{p_0}\right), \tag{5.3}$$

namely, the load may become much smaller than the standard value $\log_e (1/p_0)$.

Probably, the most important case is the intermediate one in which the population is large enough so that $2N_e s \gg 1$, yet initially the advantageous gene A_1 is so rare that $2N_e sp_0 \ll 1$. In this case, we have roughly

$$L(p_0) = 1+\log_e \left(\frac{1}{p_0}\right), \tag{5.4}$$

namely, the load is larger by about unity as compared with the value derived from deterministic treatment. For example, in a population of $N_e = 10000$, if A_1 has selective advantage $s = 0{\cdot}01$ over A_2 and if the initial frequency of A_1 is $p_0 = 0{\cdot}001$, the total load for the substitution of A_1 is 7·85 from equation (5.1). The corresponding approximate value obtained from (5·4) is 7·91. With the same population size and selective advantage, if the initial frequency is $p_0 = 10^{-4}$, the substitutional load obtained from (5.3) is 10·20, while the approximate value from (5.4) is 10·21. Note that the load for one gene substitution calculated by disregarding random fluctuation is 6·91 for $p_0 = 0{\cdot}001$ and 9·21 for $p_0 = 10^{-4}$. An increase of the load by about unity in the present case is mainly due to the fact that the advantageous gene A_1 becomes fixed (established) only with probability $u(p) \doteqdot 2N_e sp_0$, and, with the remaining probability of $1-u(p)$ it is lost from the population, never contributing to the substitution of A_1 for A_2. In the latter case, the cost (load) is wasted and this inflates the average load per gene substitution.

6. DISCUSSION

In discussing the amount of selective elimination (i.e. 'cost' or 'load') that accompanies the process of substituting one allele for another, it is important to note that the total population number is not controlled to any large extent by the relative proportion of alleles in any particular locus, including the one involved in substitution. Rather, the total population number is determined by the complex interplay between the intrinsic growth rate and the environmental control such as exemplified by Model I in § 2. Because of such population regulating mechanisms,

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

the fitness of a population measured in Malthusian parameters may be zero over a very long period, and yet there is no reason to assume that the population is bound to die out as the application of the standard theory of branching processes may dictate. The population regulating mechanisms (food supply, space available, competing species, density dependent factors of all sorts) are so strong as to overwhelm the ordinary stochastic processes whenever the population gets much too large or too small. A mathematical treatment that regards the population number as deterministic is therefore closer to the true situation than one in which population number is a random variable. With such a deterministic model, the Haldane theory has a reasonable interpretation and is free of the difficulties pointed out by Feller (1967) for unregulated populations.

It has also been claimed that the substitution of a more advantageous allele for a less advantageous one can not be considered a load since the population fitness is thereby increased, and, that the limitation to the rate of evolution set by the magnitude of the substitutional load does not actually exist. This type of argument overlooks the general fact that for each species the environment, both physical and biotic, is constantly deteriorating, while the advantageous genes are always very rare at the start. The substitution load is a measure of the amount of reproductive excess that the population must have in order that there can be enough differential viability and fertility to carry out the gene substitution and maintain the population size. If the excess is not sufficient the rate of gene substitution must be correspondingly less. Thus, as pointed out by Haldane (1957) and also by Van Valen (1965), simultaneous selection at many independent loci can not be carried out without sufficient reproductive excess to permit very intense selection.

If the selection is by differences in fecundity the cost is a measure of the amount by which the favoured type must exceed the fertility of the population average. Thus a new mutant that increases fertility by $2s$ requires the same 'cost' for its ultimate fixation in an infinite population as one that increases it by s; however, it also immediately doubles the capacity of the population for selection, and therefore the population can evolve faster. This is perhaps a good place to reiterate the by now obvious point that the 'load' is not necessarily bad.

In the original calculation of the cost of natural selection, Haldane (1957) assumed a small selection coefficient, but he later (Haldane, 1960) elaborated the cases where the selection coefficient is not small. In the latter paper, Haldane wrote that 'the substitutional load is that part of the "load" or mortality due to unfavourable environment which can be compensated by gene substitutions' and warned the readers not to extend the theory to cover biological situations to which they do not in fact apply. According to him, 'the most important of these appears to be the case where a genotype which is originally unfavourable gradually becomes neutral and then favourable'. The effect of such slowly changing environment on the substitutional load was investigated by Kimura (1967) using a simple model suggested by J. B. S. Haldane (1960, personal communication). For a haploid population, the result may be summarized as follows. Suppose that gene A_1 is originally disadvantageous and is kept in a population in very low frequency by

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

the balance between mutation and selection. Let s be the selection coefficient against A_1 and let u be the mutation rate per generation by which A_1 is produced from its allele A_2 that exists in the population with very high frequency. Assuming that $s \gg u$, the frequency of A_1 at equilibrium is

$$p_0 = u/s. \tag{6.1}$$

Now, suppose that the environment changes suddenly and in one generation, A_1 becomes advantageous such that, thereafter, A_1 has selective advantage s' (> 0) over A_2. Then, as shown in § 3, the total load for substituting A_1 for A_2 is

$$-\log_e p_0 = \log_e s - \log_e u. \tag{6.2}$$

On the other hand, to express the slow change of environment, let us assume that the selective advantage of A_1 over A_2 (measured in Malthusian parameters) is expressed by kt, where k is a small positive constant (say less than 10^{-3}) and t is time measured in generations. Allele A_1 is disadvantageous for $t < 0$ but is neutral at $t = 0$ and becomes advantageous thereafter. Then it can be shown that the frequency of A_1 at $t = 0$ is

$$p_0 = u\sqrt{\frac{\pi}{2k}}. \tag{6.3}$$

The total load for substituting A_1 for A_2 is again expressed by $-\log_e p_0$ but with p_0 given by (6.3). Thus

$$-\log_e p_0 = \tfrac{1}{2}\log_e 2k - \tfrac{1}{2}\log_e \pi - \log_e u. \tag{6.4}$$

For example, assuming $u = 10^{-8}$, the load calculated from (6.2) by taking $s = 10^{-2}$ is about 13·8, while the corresponding value calculated from (6.4) by taking $k = 10^{-6}$ is about 11·3. In other words, as compared with the case in which the selection coefficient changed from $-0{\cdot}01$ to $+0{\cdot}01$ in one generation, if such a change is taken place gradually through 20000 generations, the total load becomes roughly five-sixths as large. The load may be halved if $k = 10^{-10}$, that is if the corresponding change has taken place through two hundred million generations. The actual change of environment on earth, however, must in most cases be much more rapid. In general, we may say that if the gradual change of the selection coefficient as discussed above is common in evolution the value calculated from (6.2) overestimates the load, though the correction to be made is probably much less than 50%.

On the other hand, we have seen in § 4 that a strong epistasis of the reinforcing type may increase the load many fold, though it is unlikely that such a strong epistatic interaction between simultaneously evolving loci is very common.

We have also seen in § 5 that random fluctuation in gene frequency inflates the load roughly by about unity in the realistic case in which the population is large enough so that $2N_e s$ is much larger than unity yet the advantageous allele A_1 is so rare that $2N_e s p_0$ is much smaller than unity.

The cost is strongly dependent on the initial frequency of the favoured gene. A gene starting with a frequency of 10^{-1} instead of 10^{-4} would involve a substitutional load of 2·3 rather than 9·2. If evolution consists more of shifting the

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

frequencies of already common genes than of initially rare genes, as Wright (1931 and later) has suggested, the cost is correspondingly less. For neutral, or nearly neutral, isoalleles the cost becomes very small and the problem becomes not one of the cost, but rather the rate at which gene replacements occur when the major influences are mutation and random drift (Kimura, 1968). Finally, we have discussed the cost for freely recombining genes. If gene substitutions occur in blocks of genes, the cost per gene is less. One possibility is evolution by small duplications.

All those factors modify the load to some extent, but it appears as if the original formula by Haldane (1957), i.e. $-\log_e p_0$ is still useful in estimating the approximate amount of selective elimination that accompanies the process of substituting one allele for another by natural selection.

In the foregoing treatments we have considered the sum total of the load that spreads over many generations. In each species, then, the substitutional load in one generation is given by

$$L_e = \sum_i e_i L_i,$$

where L_i is the total load for one gene substitution in the ith locus and e_i is the rate per generation at which such gene substitution is carried out in that locus.

At the moment, opinion is divided as to how meaningful and useful the concept of the cost or substitutional load is for the study of evolution by natural selection. What it does is to give some insight as to what rate of gene substitution is consistent with a given pattern of genetically determined variability in survival and fertility. However, we have no way of knowing in most organisms how realistic is Haldane's estimate of 10% as the amount of 'substitutional load space' available.

In principle it is easier to measure the variance in fitness or the 'index of opportunity for selection' (Crow, 1958) than it is to measure the difference between the average fitness of the population and that of the type that is being increased by selection. We could then develop a theory analogous to Haldane's, but related to the variance; a beginning attempt has been made by Crow (1968). One difficulty is that the variance approach does not have one of the nicest properties of the Haldane formulation—its independence of s. On the other hand, it is not sensitive to the initial gene frequency. Actually, we believe the two approaches are complementary.

Despite its obvious limitations, the Haldane principle is, we believe, a remarkable, pioneering beginning for a quantitative study of evolutionary rates. We share with Haldane (1957) the belief that 'quantitative arguments of the kind here put forward should play a part in all future discussions of evolution'.

SUMMARY

Using models which describe the change, by natural selection, of the actual numbers of genes rather than their relative frequencies, it is demonstrated that the equation familiar to geneticists, i.e. $dp/dt = sp(1-p)$, is appropriate under a wide range of circumstances. It was pointed out that, for realistic treatment of the

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press

evolutionary process through which gene substitutions are repeated, the models must have the property such that the total population number remains constant or nearly constant throughout the process, and is not appreciably influenced by the genes being substituted.

The load or cost for a gene substitution was studied assuming a haploid population and the effects on the load of such factors as epistatic gene interaction in fitness, finite population number and slow change of environment were investigated. The load may become very large under a strong 'reinforcing' type epistasis between advantageous genes. In a finite population, the load for one gene substitution may be inflated by about unity if the product of the effective population number (N_e) and the selection coefficient (s) is large but $N_e s p_0$ is much smaller than unity, where p_0 is the initial gene frequency. On the other hand, slow change of environment may decrease the load somewhat. It was concluded that despite these and other complicating factors, Haldane's original formula, $-\log_e p_0$, for a haploid population ($-2\log_e p_0$ for the case of a diploid without dominance) is still useful for assessing the approximate amount of selective elimination that accompanies the process of gene substitution in evolution.

REFERENCES

Crow, J. F. (1958). Some possibilities for measuring selection intensities in man. *Hum. Biol.* **30**, 1–13.

Crow, J. F. (1968). The cost of evolution and genetic loads. In *Haldane and Modern Biology*, pp. 165–178. Ed. by K. R. Dronamraju. Baltimore, Md. U.S.A.: Johns Hopkins Press.

Crow, J. F. & Kimura, M. (1965). The theory of genetic loads. *Proc. XI Int. Congr. Genetics* **3**, 495–505.

Darwin, Ch. (1859). *The Origin of Species*. London: John Murray.

Feller, W. (1966). On the influence of natural selection on population size. *Proc. natn. Acad. Sci., U.S.A.* **55**, 733–738.

Feller, W. (1967). On fitness and the cost of natural selection. *Genet. Res.* **9**, 1–15.

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

Haldane, J. B. S. (1924). A mathematical theory of natural and artificial selection. Part I. *Trans. Camb. Phil. Soc.* **23**, 19–41.

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. *J. Genet.* **57**, 351–360.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.* **28**, 882–901.

Kimura, M. (1965). Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. *Genetics* **52**, 875–890.

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

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature, Lond.* **217**, 624–626.

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

Kimura, M. & Maruyama, T. (1969). The substitutional load in a finite population. *Heredity* (in the Press).

van Valen, L. (1965). Selection in natural populations. III. Measurement and estimation. *Evolution* **19**, 514–528.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

https://doi.org/10.1017/S0016672300002846 Published online by Cambridge University Press