THE SUBSTITUTIONAL LOAD IN A FINITE POPULATION

JOSEPH FELSENSTEIN
Department of Genetics, University of Washington, Seattle, Washington 98195
Received 26.iii.71

1. Introduction

The substitutional load, or cost of evolution, was first defined by Haldane (1957). He calculated the load for cases in which population size was infinite, so that all changes in gene frequency were deterministic. The substitutional load in a finite population was calculated by Kimura and Maruyama (1969). In this paper I will take issue with their definition of the load, and present calculations based on a definition which I believe corresponds more closely to the spirit of Haldane's original discussion.

In the simplest case, consider a haploid population. Suppose that a gene in the process of substitution has frequency x, and that its selective advantage is s. According to Haldane's definition, the locus contributes $s(1-x)$ during the present generation to the substitutional load. He calculated the load by summing this over all generations as the frequency changed from its initial value to 1. The load per generation was calculated by multiplying by the number of substitutions occurring per generation. I have argued elsewhere that the substitutional load is more meaningful if defined as the reproductive excess necessary to avoid extinction when the gene substitutions occur as responses to periodic unfavourable environmental changes. If the reproductive excess is d, the number of surviving offspring of the present generation will be:

$$(1+d)\prod_i (1-s_i(1-x_i)),$$

where the product is taken over all loci segregating in the offspring generation. The first half of the expression is the fecundity, and the second half is the mean viability of these offspring. The number of offspring surviving after T generations will then be:

$$(1+d)^T \prod_t \prod_i (1-s_i(1-x_i(t))), \tag{1}$$

where s_i is the selection coefficient at the ith locus, and $x_i(t)$ is the frequency of the favourable allele at locus i in generation t.

If the population is not to become extinct, (1) must be greater than one for large T. Taking logarithms, we must have

$$\log_e (1+d) \geqslant -\frac{1}{T}\sum_t \sum_i \log_e (1-s_i(1-x_i(t))). \tag{2}$$

If the s_i are small, which we assume,

$$-\log_e (1-s(1-x)) \simeq s(1-x),$$

so that we must have

$$\log_e (1+d) \geqslant \frac{1}{T}\sum_t \sum_i s_i(1-x_i(t)).$$

If d is small, $\log_e (1+d) \simeq d$, so that we can take as the value of d necessary to avert extinction

$$d = \frac{1}{T} \sum_t \sum_i s_i(1-x_i(t)). \tag{3}$$

Since all of the approximations have the effect of underestimating d, the value given in (3) may be taken as a lower bound in those cases in which the approximation is poor. When we can assume that s is small, but when d is not small, the substitutional load calculated from (3) will measure not the reproductive excess but the natural logarithm of the number of offspring per parent necessary to prevent extinction. Equation (3) is exactly the same as the Haldane definition of the substitutional load. Note that equation (3) gives the substitutional load per generation. For the remainder of this paper the discussion will be in terms of load per substitution. This can be converted to load per generation by multiplying by the number of substitutions per generation.

For the deterministic case in which the gene frequency increases from p to 1, we can use Haldane's formula to obtain the substitutional load:

$$L = \int_0^\infty s(1-x)dt. \tag{4}$$

But when random genetic drift is present, it is not obvious what to do. If we start with a large number of replicate populations, each with initial gene frequency p_0, and allow selection and drift to proceed, some will end up with a gene frequency of 1 and some with a gene frequency of zero. In any population of the first sort, integral (4) will be finite. But if a population loses the favoured allele, it will have a load of s per generation forever, so that its substitutional load, and hence the average substitutional load over all the replicates, will be infinite. Kimura and Maruyama (1969) dealt with this problem by assuming that the substitutional load is $s(1-x)$ as long as x is between zero and 1, but as soon as a population loses the favoured allele, there is no further contribution to the substitutional load. This will prevent the average load from being infinite. They calculate the average load incurred by a population which starts with a gene frequency of p, and then divide this by $U(p)$, the probability of fixation of the favoured allele. Thus, their final result is expressed as the substitutional load per completed substitution.

I have argued elsewhere (Felsenstein, 1971) that the substitutional load is imposed by deterioration of the environment. In our model, each change in the environment depresses the fitness of one allele at a different locus. The unfavourable allele remains in the population for some time after the environmental change. This depression of fitness necessitates a reproductive excess to avoid extinction of the population. Thus, the load results from the lag in evolutionary response to environmental change. On the other hand, if a favourable mutant occurs and is substituted in the absence of any deterioration of the environment, this substitution is not accompanied by any substitutional load. In fact, such a substitution makes the population better able to bear substitutional load, since it either increases fecundity or decreases mortality.

If this view is accepted, then we cannot assume that the load becomes zero as soon as a favourable gene is lost. If the environment deteriorates,

and if the favoured allele is lost as a result of genetic drift, there continues to be a load of s per generation until the favoured allele recurs by mutation and is substituted, thereby finally nullifying the effect of the environmental change on fitness. If the rate of mutation to the favoured allele is zero, the average load imposed by an environmental change will be infinite. A certain fraction of the time there will be no successful evolutionary response to environmental deterioration, and the fitness of the organism will continually decline. To calculate the substitutional load we must, therefore, consider not only the initial gene frequency and the selection coefficient but also the mutation rate. Let us calculate the substitutional load with this in mind.

2. The basic derivation

We will derive a diffusion approximation to the substitutional load. This will be done heuristically, without any pretence at rigor. Consider a haploid population of size N with initial gene frequency p, selection coefficient s and mutation rate u to the favoured allele. Let the unit of time be N generations. We consider a series of these populations, with N getting larger and u and s becoming smaller such that Ns and Nu remain constant. The populations in this series will behave more and more like a diffusion process, the change in gene frequencies in a unit of time becoming more and more nearly continuous. The diffusion equation which describes the limiting process is used to approximate to the behaviour of the initial population in the series.

Let $P(q, p)dq$ be the probability density of the gene frequency (q) in the next generation ($1/N$ of a unit of time later), given that it is p in the present generation. Let $L(p)$ be the function we seek to derive, the expected substitutional load in a population which starts at gene frequency p. Since the total load is the load in the present generation plus that expected to be incurred in future generations, we have

$$L(p) = s(1-p) + \int P(q, p)L(q)dq. \tag{5}$$

If $L(p)$ is analytic, we can expand it in a Taylor series around $q=p$.

Substituting this for $L(q)$ in (5), and ignoring the third and higher powers of $(q-p)$, which we can do since q does not change much from p in one generation,

$$L(p) \simeq s(1-p) + L(p)\int P(q, p)dq + L'(p)\int P(q, p)(q-p)dq + \tfrac{1}{2}L''(p)\int P(q, p)(q-p)^2dq. \tag{6}$$

We note that

$$\int P(q, p)dq = 1, \tag{7}$$

$$\int P(q, p)(q-p)dq = \frac{1}{N}M(p), \tag{8}$$

and

$$\int P(q, p)(q-p)^2dq = \frac{1}{N}V(p). \tag{9}$$

$M(p)$ is the expected change in gene frequency per unit time when the gene

frequency is p. $V(p)$ is the variance of gene frequency per unit time, given a gene frequency of p. We can make use of (7) and eliminate an $L(p)$ from both sides of (6). Substituting (8) and (9), and multiplying both sides by $\mathcal{N}$, we finally obtain the differential equation

$$\tfrac{1}{2}V(p)L''(p)+M(p)L'(p)+\mathcal{N}s(1-p) = 0. \tag{10}$$

In the present case, the variance of gene frequency is due to random genetic drift, and the deterministic change is due to selection as well as mutation to the favoured allele, so that

$$M(p) = \mathcal{N}sp(1-p)+\mathcal{N}u(1-p), \tag{11}$$

and

$$V(p) = p(1-p). \tag{12}$$

Substituting (11) and (12) into (10) and eliminating a factor $(1-p)$, we get

$$\tfrac{1}{2}pL''(p)+[\mathcal{N}sp+\mathcal{N}u]L'(p)+\mathcal{N}s = 0. \tag{13}$$

This is a straightforward second-order differential equation, of a type whose solution is well known. It can be solved if we are given boundary conditions for the function $L(p)$. The conditions used here are

$$L(1) = 0 \tag{14}$$

and

$$L'(0) < \infty. \tag{15}$$

The first is fairly obvious, since there can be no substitutional load if the favoured allele is already fixed at the outset. The finiteness condition on $L'(0)$ is less obvious, but it too can be demonstrated. It seems to give quite usable results. Using (14) and (15), the solution to (13) is

$$L(p) = 2\mathcal{N}s\int_p^1 e^{-2\mathcal{N}sx}x^{-2\mathcal{N}u}\int_0^x e^{2\mathcal{N}sy}y^{2\mathcal{N}u-1}dy\,dx. \tag{16}$$

There is no explicit solution to this integral, but it can be evaluated numerically and approximated for extreme values of $\mathcal{N}s$ and $\mathcal{N}u$.

3. Approximations to the load

(a) *Numerical evaluation of the integral by series approximation*

One of the best, if most tedious, ways of evaluating $L(p)$ is to expand $e^{-2\mathcal{N}sx}$ and $e^{2\mathcal{N}sy}$ in (16) as power series in $\mathcal{N}s$. The body of the double integral can then be replaced by a power series, which can be integrated termwise. The result is

$$L(p) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(-1)^j(2\mathcal{N}s)^{j+k+1}(1-p^{j+k+1})}{j!k!(2\mathcal{N}u+k)(j+k+1)}. \tag{17}$$

This series will converge, but for large $\mathcal{N}s$ convergence may take so long that evaluation by computer is expensive. Table 1 gives values of the load obtained by use of (17). However, the bottom lines of the table, where $2\mathcal{N}s = 100$, are calculated using approximations (*d*) and (*e*), given below.

TABLE 1

Substitutional load calculated from the double integral, for various values of 2Nu, 2Ns *and* p. *The values for* 2Ns = 100 *are calculated using approximations* (d) *and* (e)

		2*Nu*					
2*Ns*	P	10^{-3}	10^{-2}	10^{-1}	1	10	100
10^{-3}	0	0·9995	$9·995 \times 10^{-2}$	$9·995 \times 10^{-3}$	$9·998 \times 10^{-4}$	$1·000 \times 10^{-4}$	$1·000 \times 10^{-5}$
	0·01	0·9895	$9·895 \times 10^{-2}$	$9·895 \times 10^{-3}$	$9·898 \times 10^{-4}$	$9·900 \times 10^{-5}$	$9·900 \times 10^{-6}$
	0·1	0·8995	$8·995 \times 10^{-2}$	$8·996 \times 10^{-3}$	$8·998 \times 10^{-4}$	$9·000 \times 10^{-5}$	$9·000 \times 10^{-6}$
	0·5	0·4996	$4·996 \times 10^{-2}$	$4·997 \times 10^{-3}$	$4·998 \times 10^{-4}$	$5·000 \times 10^{-5}$	$5·000 \times 10^{-6}$
10^{-2}	0	9·950	0·9951	$9·955 \times 10^{-2}$	$9·975 \times 10^{-3}$	$9·995 \times 10^{-4}$	$1·000 \times 10^{-4}$
	0·01	9·850	0·9851	$9·855 \times 10^{-2}$	$9·875 \times 10^{-3}$	$9·895 \times 10^{-4}$	$9·900 \times 10^{-5}$
	0·1	8·951	0·8951	$8·955 \times 10^{-2}$	$8·975 \times 10^{-3}$	$8·996 \times 10^{-4}$	$9·000 \times 10^{-5}$
	0·5	4·963	0·4963	$4·966 \times 10^{-2}$	$4·981 \times 10^{-3}$	$4·997 \times 10^{-4}$	$5·000 \times 10^{-5}$
10^{-1}	0	95·17	9·521	0·9560	$9·755 \times 10^{-2}$	$9·955 \times q0^{-3}$	$9·955 \times 10^{-4}$
	0·01	94·17	9·421	0·9460	$9·655 \times 10^{-2}$	$9·855 \times 10^{-3}$	$9·895 \times 10^{-4}$
	0·1	85·22	8·526	0·8564	$8·758 \times 10^{-2}$	$8·995 \times 10^{-3}$	$8·955 \times 10^{-4}$
	0·5	46·40	4·643	0·4671	$4·817 \times 10^{-2}$	$4·966 \times 10^{-3}$	$4·966 \times 10^{-4}$
1	0	632·4	63·52	6·607	0·7966	$9·569 \times 10^{-2}$	$9·951 \times 10^{-3}$
	0·01	622·5	62·53	6·507	0·7866	$9·469 \times 10^{-2}$	$9·851 \times 10^{-3}$
	0·1	537·3	54·00	5·651	0·6990	$8·574 \times 10^{-2}$	$8·951 \times 10^{-3}$
	0·5	238·9	24·08	2·583	0·3528	$4·680 \times 10^{-2}$	$4·963 \times 10^{-3}$
10	0	1002·7	102·75	12·625	2·880	0·7056	$9·535 \times 10^{-2}$
	0·01	907·6	93·23	11·669	2·782	0·6956	$9·435 \times 10^{-2}$
	0·1	370·3	39·23	6·018	2·083	0·6099	$8·540 \times 10^{-2}$
	0·5	7·527	1·5023	0·8825	0·6920	0·2947	$4·655 \times 10^{-2}$
100	0	1005·18	105·18	15·181	6·104	2·398	0·6931
	0·01	372·27	41·17	8·064	4·713	2·303	0·6832
	0·1	2·348	2·307	2·303	2·301	1·705	0·5978
	0·5	0·6931	0·6931	0·6931	0·6931	0·6061	0·2877

(b) *Approximation when* Ns *is small*

When Ns is small, we can discard all but the first term of (17), the one in which $j = k = 0$, ignoring higher powers of Ns. Then

$$L(p) \simeq \frac{Ns}{Nu}(1-p) = \frac{s}{u}(1-p). \tag{18}$$

The approximation can be motivated as follows: When Ns is very small, selection has almost no influence on the course of events. Suppose that Nu is small, so that we have to wait long periods of time between mutations. A population starting with a gene frequency of p will have a probability $1-p$ of losing the favoured allele due to random genetic drift. If the favoured allele is lost, there will be an average of Nu new mutants occurring per generation, each of which has probability $1/N$ of drifting to fixation rather than loss. The population will then have to wait $1/u$ generations until a new mutant destined to be successful arises. During each of these generations there is a load s. We ignore the reduction of the load by the presence of the favoured allele during the generations preceding its initial loss, as well as the load incurred between the time a successful mutant occurs and the time it is fixed. The total load is then $s(1-p)/u$.

On the other hand, when Nu is large, the changes in gene frequency will by nearly deterministic and mostly the result of mutation pressure. Then we almost never have loss of the favoured alleles initially present. We can use a deterministic treatment:

$$\frac{dp}{dt} = u(1-p)$$

$$L(p) = \int_0^\infty s(1-p)dt = \frac{s}{u}\int_0^\infty \frac{dp}{dt}\,dt = \frac{s}{u}(1-p). \tag{19}$$

This happens to give the same result. It is, therefore, at least reasonable that (18) holds for intermediate values of Nu.

(c) *Approximation when* Nu *is small*

We can apply a similar argument when Ns is large, as long as Nu is so small that most of the load is incurred while waiting for the occurrence of a mutant. The probability that the favoured alleles initially present at frequency p fix is (Kimura, 1962)

$$U(p) = \frac{1-e^{-2Nsp}}{1-e^{-2Ns}}. \tag{20}$$

If fixation occurs, we count the load as being approximately zero. If the favoured alleles are lost, which will happen $1-U(p)$ of the time, we must wait for a successful mutant. There are Nu new mutants per generation, each with probability

$$U\left(\frac{1}{N}\right) \simeq \frac{2s}{1-e^{-2Ns}} \tag{21}$$

of succeeding, so that the total load will be

$$L(p) = \frac{s(1-U(p))}{NuU\left(\frac{1}{N}\right)}$$

$$\simeq \frac{e^{-2Nsp}-e^{-2Ns}}{2Nu}. \tag{22}$$

This equation can also be derived by taking the double power series in (17), and ignoring terms in $(2Nu+1)^{-1}$, $(2Nu+2)^{-1}$, . . . which should be small relative to terms in $(2Nu)^{-1}$ if $2Nu$ is sufficiently small.

(d) *Approximation when both* Ns *and* Nu *are large*

If we divide all terms in equation (13) by Ns, and then let $Ns\to\infty$ assuming that Nu also becomes indefinitely large in such a way that the ratio u/s remains constant, we obtain

$$\left(p+\frac{u}{s}\right)L'(p)+1 = 0. \tag{23}$$

With the boundary condition $L(1) = 0$, we can solve this, obtaining

$$L(p) = \log_e\left(\frac{s+u}{sp+u}\right). \tag{24}$$

This is simply the load in a deterministic model in which forward mutation is present. Note that when $u = 0$, equation (24) reduces to $L(p) = -\log_e p$, Haldane's original result.

(e) *Approximation when* Ns *is very large and* Nu *is small*

When Ns is very large, so that the probability of losing those favoured alleles which are initially present is essentially zero, formula (22) would predict a load of zero. A better approximation can be developed by taking

TABLE 2

Identity and accuracy of the best approximation to the load. Accuracy as number of correct significant figures, to a maximum of four

		$2Nu$					
$2Ns$	P	10^{-3}	10^{-2}	10^{-1}	1	10	100
10^{-3}	0	c 4	c 4	c 4	c, d 3	b, d 4	b, d 4
	0·01	c 4	c 4	c 4	c, d 3	b, d 4	b 4
	0·1	c 4	c 4	c 3	c, d 3	b, d 4	b 4
	0·5	c 4	c 4	c 3	c, d 3	b, d 4	b 4
10^{-2}	0	c 4	c 3	c 3	c, d 2	d 4	b, d 4
	0·01	c 4	c 3	c 3	c, d 2	d 4	b, d 4
	0·1	c 4	c 4	c 3	c, d 2	d 3	b, d 4
	0·5	c 4	c 4	c 3	c, d 2	d 3	b, d 4
10^{-1}	0	c 3	c 3	b 1	d 1	d 3	d 4
	0·01	c 3	c 3	c 2	d 1	d 3	b 3
	0·1	c 3	c 3	c 2	d 1	d 3	d 2
	0·5	c 3	c 3	c 2	d 1	d 3	d 4
1	0	c 3	c 2	c 1	d 0	d 2	d 3
	0·01	c 3	c 2	c 1	d 0	d 2	d 3
	0·1	c 3	c 2	c 1	d 0	d 2	d 4
	0·5	c 3	e 2	e 1	d 0	d 2	d 4
10	0	c 2	e 2	e 1	d 0	d 1	d 3
	0·01	e 4	e 3	e 1	d 0	d 1	d 3
	0·1	e 2	e 2	e 2	d 0	d 1	d 3
	0·5	e 1	e 1	e 0	e 2	d 1	d 3

equation (16) and approximating the inner integral, assuming that Ns is large and Nu small:

$$\int_0^x e^{2Nsy} y^{2Nu-1} dy \simeq x^{2Nu-1} \int_0^x e^{2Nsy} dy + \int_0^x y^{2Nu-1} dy. \tag{25}$$

We substitute this into (16) and finally get

$$L(p) = \int_p^1 \frac{1 - e^{-2Nsx}}{x}\, dx + \frac{e^{-2Nsp} - e^{-2Ns}}{2Nu}. \tag{26}$$

The integral must be evaluated numerically, but this is not difficult.

Table 2 presents, for points corresponding to those in table 1, the identity of the best approximation and a rough indication of its accuracy.

4. Diploidy

All of the above equations are for haploid models. With a diploid model, if the fitnesses of the three genotypes are

$$\begin{array}{ll} AA & 1+2s \\ Aa & 1+s \\ aa & 1, \end{array}$$

and the number of diploid individuals is N, then we can use all of the above equations, except that we must replace $2Ns$ by $4Ns$ and $2Nu$ by $4Nu$.

When the heterozygous genotype is not intermediate, but shows some other level of dominance, we can no longer use equation (16). Such cases will not be dealt with here.

5. Results

Figs. 1 to 6 show calculated values of the substitutional load. These are plotted from table 1, and by using the approximation formulae (*b*) to (*e*) for interpolation and approximation.

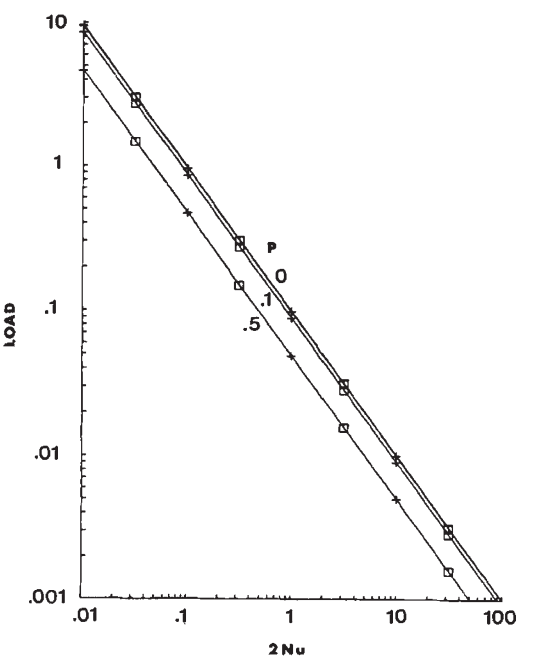

Fig. 1.—Substitutional load as a function of mutation rate and initial gene frequency, with $2Ns = 0{\cdot}1$. Crosses indicate points calculated by numerical evaluation of the double integral, boxes indicate points calculated from the best available approximation formula.

Figs. 1 and 2 show the load plotted against $2Nu$ for small (0·1) and large (10) values of $2Ns$. In both cases the result is as expected: the higher the mutation rate, the lower the substitutional load. When $2Ns$ is small, the

load is nearly proportional to the initial frequency of the disfavoured allele, so that going from an initial gene frequency of zero to a low initial frequency has little effect on the substitutional load. But when $2Ns$ is large and when $2Nu$ is small, initial frequency of the favoured allele has a much larger effect on the load. If even a few copies of the favoured allele exist in the population, the strong selection makes it likely that the favoured allele will rapidly fix. But if no favoured alleles are present initially, the low mutation rate insures a long wait for mutants, with a large load incurred each generation. This interpretation of the form of the curves is bolstered by the fact that when initial gene frequency of the favoured allele is high, the load is not much reduced by an increase in mutation rate, since enough copies of the favoured allele already exist to insure its early fixation irrespective of the mutation rate.

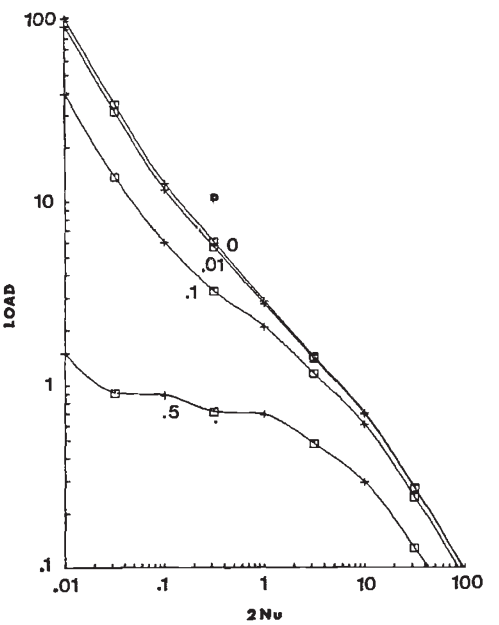

FIG. 2.—Substitutional load as a function of mutation rate and initial gene frequency, with $2Ns = 10$. The wavy nature of the curve for $P = 0{\cdot}5$ may be a result of the inaccuracy of the approximations.

Figs. 3 and 4 show the effect of different values of s on the load, for small (0·1) and large (10) values of $2Nu$. When $2Ns$ is small, selection will have little effect on the dynamics of gene frequency change. Increasing s will increase the impact of selection on the size of the load incurred in each generation, without having much effect on the number of generations which the deleterious alleles persist. As s increases, therefore, the load is approximately proportional to s. In fig. 3, where $2Nu$ is 0·1, as $2Ns$ becomes larger than 1, selection suddenly becomes effective. Increases in s in this range have a much greater effect in increasing the effectiveness of selection than they do in increasing the impact of the deleterious allele on the load. Thus, the load passes through a maximum and begins to decrease. Finally, when $2Ns$ is large, an asymptote is reached at the infinite-population values given

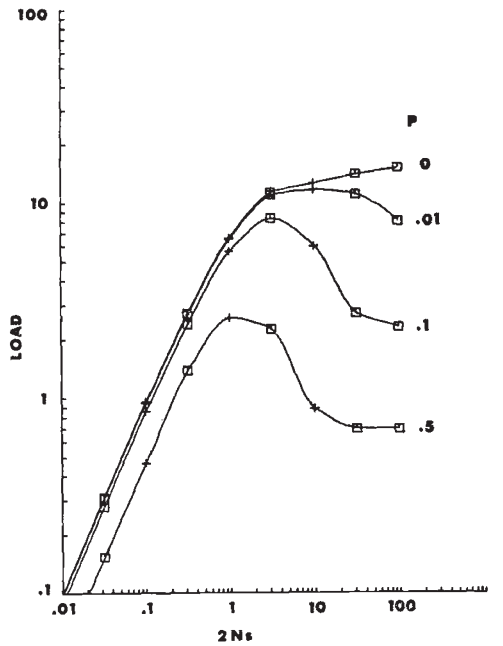

FIG. 3.—Substitutional load as a function of selection coefficient and initial gene frequency, with $2Nu = 0{\cdot}1$.

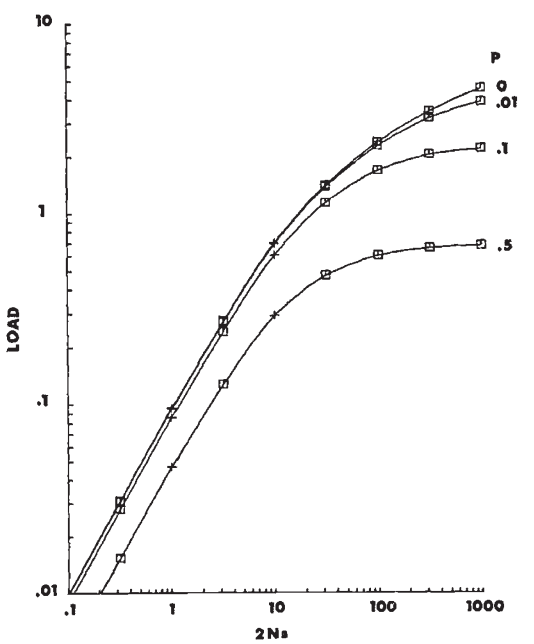

FIG. 4.—Substitutional load as a function of selection coefficient and initial gene frequency, with $2Nu = 10$.

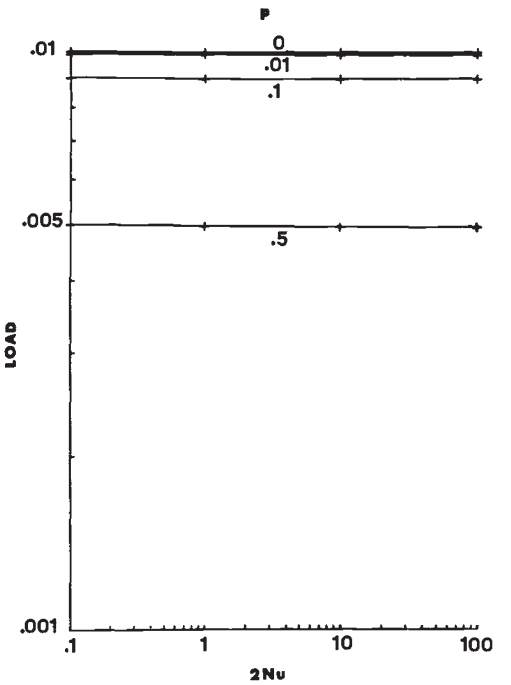

FIG. 5.—Substitutional load as a function of population size and initial gene frequency, with u and s held constant, with $u/s = 100$, so that $2Ns = 0{\cdot}01 \times 2Nu$.

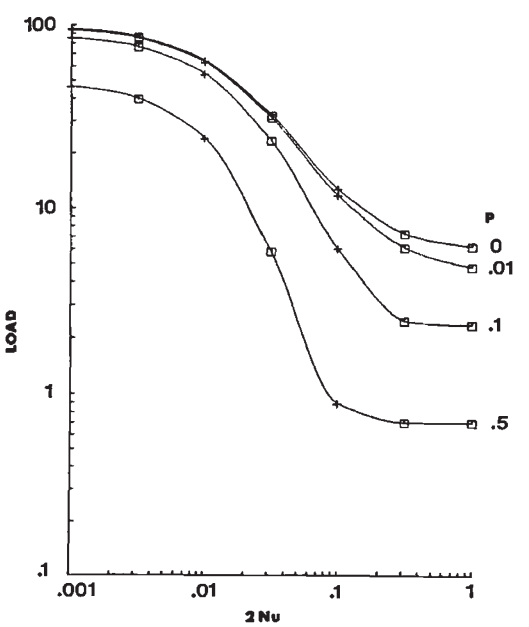

FIG. 6.—Substitutional load as a function of population size and initial gene frequency, with u and s held constant, with $u/s = 0{\cdot}01$, so that $2Ns = 100 \times 2Nu$.

by Haldane (except when $p = 0$). A doubling of the strength of selection will halve the number of generations that the deleterious alleles persist, while doubling the contribution they make in each of those generations to the load. In fig. 4, where $2Nu$ is 10, there are no maxima in the load. The curves pass smoothly from the phase in which load is proportional to s to the phase in which load does not depend on s. It is not clear to me why there should be this difference between figs. 3 and 4. It should be noted that for even smaller values of $2Nu$ the heights of the maximum values of the load curve can be far higher relative to the asymptotic values. Kimura and Maruyama also found maxima in the plot of load against selection coefficient, but those maxima were far less dramatic than the present ones.

Figs. 5 and 6 show the effect of altering N, so that $2Ns$ and $2Nu$ both change but remain in a constant ratio to each other. When s is larger than u (fig. 6) there is a substantial drop in the load with increasing population size. But when u is larger than s, there is almost no effect on the load, since then approximations (*b*), (*c*) and (*d*) are nearly equal.

6. Amino-acid substitution rates

Kimura (1968) has argued that a population carrying out amino acid substitutions at the rate indicated by protein sequence data would incur a large substitutional load. He argues that this load could be circumvented by assuming that most of these substitutions are in fact selectively neutral.

The calculations presented here do not rule this out. For sufficiently small values of s, the load will approach zero. But these results do provide a somewhat different perspective on the problem. Initial gene frequencies are for the most part insignificant in their effects on the load. The mutation rate becomes important. More importantly, the basic definition of the load used here stresses that the load is the result of environmental deterioration. This raises another possible way of avoiding having to assume a high substitutional load. If the mutants are occurring in the absence of environmental change, and are substituted because of their favourable effect on fitness, then there is no substitutional load at all. There need be no contradiction between high rates of substitution and the importance of natural selection in the substitution process, provided that we can envisage a sufficiently high rate of occurrence of favourable mutants.

7. Summary

1. Substitutional load in a finite population is defined. The definition differs from that of Kimura and Maruyama (1969) in that it assumes that the load results from environmental deterioration, and ceases only when the favoured allele has been successfully substituted. Thus the load is a function not only of the selection coefficient, population size, and initial gene frequency, but also of the rate of mutation to the favoured allele.

2. A diffusion approximation to the substitutional load in a finite haploid population can be derived. It turns out to be a double integral which cannot be explicitly integrated. A method of numerical calculation of the integral has been defined. Approximations to the integral were developed for cases in which various parameters of the model take on extreme values.

3. The equations can also be used to calculate the substitutional load in diploid populations, provided that the fitness of the heterozygote is intermediate between the homozygote fitnesses.

4. An increasing rate of mutation to the favoured allele decreases the substitutional load. The only case in which changing the mutation rate has little effect is when the initial frequency of the favoured allele is high, selection is strong, and mutation rates are low.

5. When selection coefficients are small, the substitutional load is also small. As the selection coefficient is increased, the load increases, ultimately reaching an asymptote at the value calculated by Haldane. However, when mutation rates are low, the load will pass through a maximum and approach the asymptote from above. For very low mutation rates, this maximum can be many times higher than the asymptotic value.

6. Increasing the population size, holding mutation rates and selection coefficients constant, will in general result in a decrease in the load. However, this decrease will be very slight if the mutation rate exceeds the selection coefficient.

7. The initial frequency of the favoured allele will have a strong effect on the load only when the mutation rate is low and the selection coefficient is large.

8. Even if rates of gene substitution are as high as is indicated by amino-acid sequence studies on proteins, one need not assume that substitutional load is very large. Either selection coefficients could be very small, as Kimura has suggested, or the substitutions could be the result of favourable mutants occurring in the absence of environmental change.

Acknowledgment.—This research was supported by Task Agreement Number 5 of U.S. Atomic Energy Commission Contract AT(45-1) 2225 with the University of Washington. I wish to thank Professor A. Robertson for helpful suggestions on the presentation of the derivations.

8. References

FELSENSTEIN, J. 1971. On the biological significance of the cost of gene substitution. *American Naturalist, 105,* 1-11.

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

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

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

KIMURA, M., AND MARUYAMA, T. 1969. The substitutional load in a finite population. *Heredity, 24,* 101-114.