IMA Journal of Mathematics Applied in Medicine & Biology (1995) **12**, 1–11

A model for the evolution of the genome: The effect of stochasticity on genetic loads

F. M. PHELPS IV

Aksai 3-A dom 66 kv 10, 480031 Almaty, Kazakhstan

[Received 22 September 1994 and in revised form 29 March 1995]

In this paper a general model is given for the evolution of the genome incorporating stochastic factors. The model is applied to the substitutional genetic load problem. All of the major hard selection load formulae in the literature are extended and, where necessary, corrected (for stochasticity). Turning to rank selection, formulae for stochastic factors are also corrected and harmful mutations included. A simple formula for the selection coefficient as a function of the nonneutral substitution rate and the mutation profile is obtained. Further, it is noted that the formulae derived also apply (for different parameter values) to the mutation load, thus unifying the two loads under a single theory. A general formula for the mutation load under hard selection is given, extending previous results. Finally, the author derives a formula showing how many harmful mutations can be effectively eliminated by rank selection and discusses its relevance to the question of the possible buildup of harmful mutations in the human gene pool due to long-term exposure to low-level radiation.

Keywords: cost of natural selection; genetic loads; neutral theory; rate of evolution; hard selection; rank selection.

1. Overview

It is relatively easy to understand how natural selection can increase the frequency of a favoured allele at one locus, but understanding how selection can work on all favoured loci in the genome at the same time is a complicated problem. The simplest models assume multiplicative fitnesses or hard selection and run into the problem of intolerable genetic loads.

Three times (Mueller, 1950; Lewontin & Hubby, 1966; Kimura, 1968) in the history of population genetics, the concept of 'genetic load' has been associated with major developments. Yet derivations of expressions for genetic load have only seldom (see Kimura & Maruyama, 1966, 1969) included stochastic factors (such as the deaths wasted selecting for favourable alleles which are ultimately lost to drift), so it may be said that calculations of various genetic loads have never been done correctly, except in a few special cases. In all three loads—mutational (Crow & Kimura, 1978), segregational (Milkman, 1967; Wills, 1978), and substitutional (Sved, 1968)—'soft' selection models have been shown to eliminate problems associated with prohibitively large loads, and yet, except for Wills (1978), none of these treatments has included stochastic factors. This work was motivated by an attempt to incorporate stochastic factors into the general model for the substitutional genetic load (SGL) developed by Phelps (1991).

In Section 2 we develop the model, combining the diffusion equation method with the formula for selection coefficient as a function of some character. In Section 3, following Kimura (1969), we write down the solution to the diffusion equation. In Section 4.1 we briefly review the history of 'Haldane's dilemma' and the SGL, and its relevance to the neutral theory of molecular evolution. In Section 4.2 we give new results for the SGL in hard selection. In Section 4.3 we show that incorporating stochastic factors and harmful as well as beneficial mutations does not limit the rate at which evolution can proceed, but does restrict possible selection coefficients. In Section 4.4 we note that the formulae derived for the SGL also apply to the mutational load (ML), leading to an extension of previously published results. The incorporation of stochasticity does not alter classical formulae for the ML, which is found to be very large in hard selection. We note that, as with the SGL, rank selection can essentially eliminate the ML. We derive an expression showing just how many harmful mutations can be eliminated under rank selection.

2. General theory

2.1 *The diffusion equation*

One of the most productive tools for analysing the effects of stochasticity in population genetics models has been the diffusion equation method (see Crow & Kimura, 1970), which we take as our starting point. Let us assume that a pair of alleles A_1 (the new mutant) and A_2 (the original) are segregating in a population. The Kolmogorov backward diffusion equation for the probability density of the frequency of A_1 is

$$\frac{\partial\phi(p, x; t)}{\partial t} = \tfrac{1}{2}V_{\delta p}\frac{\partial^2\phi(p, x; t)}{\partial t^2} + M_{\delta p}\frac{\partial\phi(p, x; t)}{\partial p}, \tag{2.1}$$

where $\phi(p, x; t)$ is the probability density that the frequency of A_1 is x at time t (measured in generations), given that it is p at time $t = 0$, and the mean and the variance of the amount of change in allele frequency p during a short time interval from t to $t + \delta t$ are $M_{\delta p}\delta t$ and $V_{\delta p}\delta t$ respectively. For simplicity, we will treat the haploid case where $V_{\delta p} = p(1 - p)/N_e$ and $M_{\delta p} = sp(1 - p)$, where N_e is the effective population number and s is the selective advantage of A_1, meaning that an individual with the new allele has a $1 + s : 1$ fitness advantage over an otherwise genetically identical individual carrying A_2.

2.2 *The selection coefficient*

In many papers, s for a given allele has unnecessarily been assumed not to depend upon population composition (hard selection), which leads to intolerable 'genetic loads' (discussed below). To avoid this restriction, let us consider the definition of selective advantage s. Suppose fitness w is a function of some variable X which can be calculated for any individual from its genome. In this paper we will assume that X is 'additive', that is, for every polymorphic locus (indexed by i) there is a contribution X_i (depending on the allele at the ith locus) such that $X = \sum X_i$.

To define X we assume that as long as the ith locus is polymorphic, the original allele contributes a units to X ($X_i = a$) in every individual carrying it, while the new allele contributes 0 ($X_i = 0$) so that the 'average effect' of the mutant is $-a$. Of course a depends upon the locus i, but we will usually not write out this dependence. Once a mutant is fixed or lost, the locus is no longer polymorphic and the fixed allele contributes nothing to X.

Let us assume that new mutations of effect $-a$ and initial frequency p are occurring (and have been since the distant past) in the population at rate (two-dimensional probability density function) $v(a, p)$ per generation. Theoretically, a mutation can occur at any locus in any individual, but, to simplify the treatment, we assume all mutations occur at monomorphic loci.

Let the probability density function of X be denoted $\tilde{f}(\chi)$. Let $\bar{w}$ denote the average population fitness (survival to adulthood). By definition,

$$\bar{w} = \int_{-\infty}^{\infty} w(\chi)\tilde{f}(\chi)\,d\chi. \tag{2.2}$$

Let w_1 and w_2 denote the conditional fitnesses given A_1 and A_2 respectively. Then $s = (w_1 - w_2)/\bar{w}$. Assuming a is small,

$$s = \frac{1}{\bar{w}}\int_{-\infty}^{\infty} w(\chi)[\tilde{f}(\chi + a(1-x)) - \tilde{f}(\chi - ax)]\,d\chi \approx \frac{a}{\bar{w}}\int_{-\infty}^{\infty} w(\chi)\tilde{f}'(\chi)\,d\chi. \tag{2.3}$$

Additivity over an enormous number of mostly independently evolving polymorphic loci suggests (by various central limit theorems) that X is normally distributed. Departures from this assumption due to finite population size, linkage, or domination of X by the contributions of a few loci are assumed to be second-order effects. Thus we take $\tilde{f}(\chi) = (1/\sigma)f((\chi - \mu)/\sigma)$, where $f(\xi) = e^{-\xi^2/2}/(2\pi)^{1/2}$ and μ and σ^2 are the mean and variance of X respectively. Thus, to first order, (2.3) and (2.2) become

$$s = \frac{a}{\sigma^2\bar{w}}\int_{-\infty}^{\infty} w(\chi)f'\left(\frac{\chi - \mu}{\sigma}\right)d\chi \tag{2.4}$$

and

$$\bar{w} = \frac{1}{\sigma}\int_{-\infty}^{\infty} w(\chi)f\left(\frac{\chi - \mu}{\sigma}\right)d\chi. \tag{2.5}$$

2.3 *Hard, soft, and rank selection*

In order to use (2.4) or (2.5) we need to specify $w(\chi)$. This brings us to a discussion of various models for natural selection. We consider two fairly general cases, hard and rank selection.

Selection is said to be *hard selection* if, in a given environment, the viability of an individual for which $X = \chi$ is determined up to a density dependent factor by χ without reference to other individuals in the population. For hard selection, fitness has the form

$$w(\chi) = w(0)e^{-\varepsilon g(\chi)} \tag{2.6}$$

for some function $g(\chi)$ and positive constant ε. In the case of beneficial mutations of effect $-a$, we assume $a > 0$ and take $g(\chi)$ to be an increasing positive function. Of course, ε could be absorbed into the function g, but we retain it (and assume it is small) to stress that our analysis applies in the slow selection limit. Here $w(0) = w(0, N)$ is allowed to depend on population size N (the density-dependent factor mentioned above), but we will not write out this dependence. If selection is not hard, it is said to be *soft selection.*

The rank $\mathscr{R}(\chi)$ of an individual with $X = \chi$ in a population is defined to be the fraction of individuals for which $X < \chi$. Note that, by this definition,

$$\mathscr{R}(\chi) = \int_{-\infty}^{\chi} \tilde{f}(\zeta)\,\mathrm{d}\zeta = \int_{-\infty}^{\chi} \frac{1}{\sigma} f\left(\frac{\zeta - \mu}{\sigma}\right)\mathrm{d}\zeta = F\left(\frac{\chi - \mu}{\sigma}\right),$$

where F is the cumulative distribution function for f, defined by $F(\zeta) = \int_{-\infty}^{\zeta} f(\xi)\,\mathrm{d}\xi$. If fitness is a function of rank, that is, if there is a function G such that $w(\chi) = G(\mathscr{R}(\chi))$, so that

$$w(\chi) = G\left(F\left(\frac{\chi - \mu}{\sigma}\right)\right), \tag{2.7}$$

we say that the selection scheme is *rank selection.* In such a case G is called the *viability function.* In this paper we assume that G, and so fitness, is a decreasing function of X and rank 0 is optimal. Rank selection is a special form of soft selection. The classical example of rank selection is the threshold or truncation selection model where the most fit fraction $\bar{w}$ of the population survive and the least fit fraction $1 - \bar{w}$ die.

3. Determining μ and σ^2

Following Kimura (1969), we consider the functional $I_\gamma(p)$ defined by

$$I_\gamma(p) = v(a, p) \int_{t=0}^{t=\infty} \int_{x=0^+}^{x=1^-} \gamma(x)\phi(p, x, t)\,\mathrm{d}x\,\mathrm{d}t.$$

Clearly, if a 'mutation' has initial frequency $p = 0$ or $p = 1$ the locus remains monomorphic so that $\phi(0, x, t) = \phi(1, x, t) = 0$ for all x on the interval $(0^+, 1^-)$. Thus $I_\gamma(0) = I_\gamma(1) = 0$.

If $\gamma(x) = a(1 - x)$, then $I_\gamma(p)$ is $\hat{\mu}(a, p)$, the contribution to the mean μ of all the polymorphic loci for which the mutant arose t generations ago, has effect $-a$, and has initial frequency p, integrated over all time. So, we have

$$\mu = \int_{a=-\infty}^{a=\infty} \int_{p=0^+}^{p=1^-} \hat{\mu}(a, p)\,\mathrm{d}p\,\mathrm{d}a. \tag{3.1}$$

If $\gamma(x) = a^2x(1 - x)$, then $I_\gamma(p)$ is $\hat{\sigma}^2(a, p)$, the contribution to the variance σ^2 of all the polymorphic loci for which the mutant arose t generations ago, has effect $-a$,

and has initial frequency p, integrated over all time. In this case,

$$\sigma^2 = \int_{a=-\infty}^{a=\infty} \int_{p=0^+}^{p=1^-} \hat{\sigma}^2(a, p)\, dp\, da. \tag{3.2}$$

We seek a differential equation for I_γ. Multiply (2.1) by $v(a, p)\gamma(x)$, integrate with respect to x on $(0^+, 1^-)$ and then with respect to t on $[0, \infty)$. After using the definition of I_γ and boundary conditions $\phi(p, x, \infty) = 0$ and $\phi(p, x, 0) = \delta(x - p)$, we arrive at

$$\tfrac{1}{2} V_{\delta p} I_\gamma''(p) + M_{\delta p} I_\gamma'(p) + v(a, p)\gamma(p) = 0,$$

with boundary conditions $I_\gamma(0) = I_\gamma(1) = 0$.

Kimura (1969) gives the solution to this equation (in our notation) as

$$I_\gamma(p) = [1 - u(p)] \int_0^p \psi_\gamma(\xi) u(\xi)\, d\xi + u(p) \int_p^1 \psi_\gamma(\xi)[1 - u(\xi)]\, d\xi, \tag{3.3}$$

where $u(p)$ is the probability of ultimate fixation (see Crow & Kimura, 1970) of the mutant allele given by $u(p) = \int_0^p \Gamma(\xi)\, d\xi / \int_0^1 \Gamma(\xi)\, d\xi$ with $\Gamma(\zeta) = \exp(-2 \int_0^\zeta (M_{\delta p}/V_{\delta p})\, dp)$, and $\psi_\gamma(\xi) = 2v(a, p)\gamma(\xi)/V_{\delta\xi} u'(\xi)$. In the haploid case, $\Gamma(\xi) = e^{-2N_e s\xi}$ and

$$u(p) = \frac{1 - e^{-2N_e sp}}{1 - e^{-2N_e s}}. \tag{3.4}$$

For $\gamma(x) = a(1 - x)$ and using (3.1), the solution (3.3) reduces to

$$\mu = \int_{a=-\infty}^{a=\infty} \int_{p=0^+}^{p=1^-} \frac{a}{s} v(a, p) \mathscr{F}(p)\, dp\, da. \tag{3.5}$$

where

$$\mathscr{F}(p) = [1 - u(p)] \int_0^{2N_e sp} \frac{e^{\xi - 1}}{\xi}\, d\xi - u(p) e^{-2N_e s} \int_{2N_e sp}^{2N_e s} \frac{e^\xi}{\xi}\, d\xi + u(p) \ln(1/p).$$

For $\gamma(x) = a^2 x(1 - x)$ and using (3.2), (3.3) reduces to

$$\sigma^2 = \int_{a=-\infty}^{a=\infty} \int_{p=0^+}^{p=1^-} \frac{a^2}{s} v(a, p)[u(p) - p]\, dp\, da. \tag{3.6}$$

4. Applications to genetic loads

4.1 *Overview: Substitutional genetic load*

In 1957 Haldane attempted to compute a theoretical upper limit to the rate of evolution based on what he called the 'cost of natural selection' (also known as the substitutional genetic load (SGL) defined below). His observation was that, in order for a new allele to replace an old allele in a population, all carriers of the old allele must eventually be weeded out. Now, if rapid evolution is proceeding simultaneously at many loci, there is a limit to how many individuals can be wiped out each generation if the population is to survive. By making what amounted to an

assumption of multiplicative fitnesses (mathematically the easiest to handle), Haldane estimated that evolution could not proceed at a rate of much more than one substitution every 300 generations.

Haldane's argument and its extremely restrictive conclusion (known as 'Haldane's dilemma') generated a lengthy controversy (see review in Phelps, 1991), which received renewed attention when Kimura (1968) used the discrepancy between the rate of evolution estimated from protein sequence data (two substitutions per generation) and Haldane's calculation to propose the neutral theory (for which there is no SGL) of molecular evolution. See Kimura (1983) for a historical summary of the neutral theory. One of the most important objections to Kimura's SGL argument was due to Sved (1968), who used a deterministic rank selection model to show that there is no theoretical limit to the rate of molecular evolution. Phelps (1991) created a general deterministic model which encompasses each previous model as a special case. In this section we modify that model to include stochastic factors.

4.2 *Hard selection and the SGL*

Let us assume selection is hard and take $w(\chi)$ to be given by (2.6) above where g is positive and increasing. Our assumptions that selective differences are small, X is essentially normally distributed, and $X > 0$ imply that $D = \mu/\sigma \gg 1$. Note that the scaling on a is arbitrary—we can use the free parameter ε to ensure that selective differences are small, even if the average effect of new alleles is large.

Put $y = \chi/\mu$ in (2.5) and use the definition of f to arrive at

$$\bar{w} = \frac{w(0)\mu}{(2\pi)^{1/2}\sigma}\int_0^\infty e^{-\varepsilon g(\mu y)} e^{-D^2(y-1)^2/2}\,dy.$$

For large D, the function $[D/(2\pi)^{1/2}]e^{-D^2(y-1)^2/2}$ acts like $\delta(y-1)$ and so $\bar{w} \approx w(0)e^{-\varepsilon g(\mu)}$. The quantity $L = \ln[w(0)/\bar{w}]$ is known as the substitutional genetic load (SGL) in the literature. Thus, $L = \varepsilon g(\mu)$.

A similar argument gives $s = a\varepsilon g'(\mu)$, so that $s(a) = Lg'(\mu)/g(\mu)$. Using (3.5), this becomes

$$L = \frac{g(\mu)}{g'(\mu)\mu}\int_{a=0}^{a=\infty}\int_{p=0+}^{p=1} v(a,p)\mathscr{F}(s(a),p)\,dp\,da. \tag{4.1}$$

The density function for the substitution rate $r(a, p)$ is given by $r(a, p) = u(a, p)v(a, p)$, which follows from the fact that $u(a, p)$ is the probability of ultimate fixation for a mutant allele. Thus L is given by

$$L = \frac{g(\mu)}{g'(\mu)\mu}\int_{a=0}^{a=\infty}\int_{p=0+}^{p=1-} r(a,p)\mathscr{L}(s(a),p)\,dp\,da, \tag{4.2}$$

where $\mathscr{L}(s(a), p) = \mathscr{F}(s(a), p)/u(s(a), p)$ is the 'load formula' for a single substitution, denoted $L(p)$ in Kimura & Maruyama (1969).

This result (4.2) generalizes and improves upon what has been written on the SGL. We discuss several cases below. Since we are considering substitution of new slightly

favourable mutants, we usually assume $p = 1/N$, although we will often retain p dependence in the formulae.

As pointed out by Kimura & Maruyama, there are three cases in which $\mathscr{L}(s, p)$ reduces to a simple expression.

Case I. If $2N_e sp \gg 1$, then, using (3.4), $\mathscr{L}(s, p)$ reduces to

$$\mathscr{L}(s, p) = \ln(1/p). \tag{4.3}$$

This is the *nonstochastic* case. It does not apply for $p = 1/N$, as weak selection ($s \ll 1$) implies that $2N_e s/N$ is not large. Note that it is independent of s.

Case II. If $2N_e sp \ll 1$ but $2N_e s \gg 1$, so that $u(s, p) \approx 2N_e sp$, then

$$\mathscr{L}(s, p) = \ln(1/p) + 1. \tag{4.4}$$

This is the *selectionist* case and is of major importance in evolutionary theory. Note that it is also independent of s.

Case III. If $2N_e s \ll 1$, then

$$\mathscr{L}(s, p) = 2N_e s \ln(1/p) \ll 1.$$

This is the neutral case, in which many genetic loads essentially disappear.

Now we use the selectionist expression to compute the SGL as a function of the nonneutral substitution rate R for the genome, assuming hard selection with 'nth-order epistasis' or $g(x) = x^n$. Assuming that all mutations have $p = 1/N$, we have $r(a, p) = r(a)\delta(p - 1/N)$ for some density function $r(a)$ for which $\int_{a=0}^{a=\infty} r(a)\, da = R$. In this case (4.2) becomes

$$L = \frac{1}{n} \int_{a=0^+}^{a=\infty} r(a)\mathscr{L}(s(a), 1/N)\, dp\, da.$$

Now, in the selectionist limit (4.4), $\mathscr{L}$ is independent of a, so that this reduces to

$$L = \frac{R}{n}(\ln N + 1) \tag{4.5}$$

independently of $r(a)$. Thus, the result is not affected by the distribution of mutation effects. Phelps (1991) showed that nth-order epistasis reduces the SGL by a factor of n in hard selection, but that treatment was in the nonstochastic case, which does not apply for reasonable parameter values. Here we see that in the stochastic case a similar reduction occurs.

The formula obtained by Phelps (1991) differs from the above in that the 1 is absent. Thus, the effect of stochasticity (which can also be seen by comparing (4.3) to (4.4) above) is the increase of the SGL by the inclusion of the 1 in (4.5). The simplest stochastic case—the case of multiplicative fitnesses ($n = 1$) with fixed effect a_0—was fully treated by Kimura & Maruyama (1969), so that (4.5) extends their work.

To investigate the consequences of this formula, let us take as a reasonable guess $n = 2$ and, following Haldane (1957), $N = 10^5$. Then, even if $w(0) = 1$, its maximum possible value, $\bar{w} = e^{-6R}$. The nonneutral mutation rate cannot be even a substantial

fraction of the total substitution rate, estimated by Kimura (1968) to be two substitutions per generation, without implying an intolerably small average population fitness, Thus we must conclude that most substitutions are essentially neutral or that selection is mostly soft.

4.3 *Rank selection as a solution to Haldane's dilemma*

Let us now consider the relationship between the SGL and the substitution rate in a rank selection model. Note that both the fitness of the optimal individual $G(0)$ and the average population fitness $\int_0^1 G(\xi)\,d\xi$ are fixed functions of G. Thus the SGL, being the log of the ratio of these quantities, is a fixed function of G alone, unrelated to the substitution rate. This effectively eliminates Haldane's dilemma, even if we include stochastic factors. One possible objection is that, for large R, the selection coefficients may become so small as to be effectively neutral. We consider this problem below.

Given a rank selection fitness rule, we can allow a to be negative so as to handle both beneficial and harmful mutations in a single formula. Let us define the constant K, which depends only upon the viability function G, by

$$K = \frac{\int_{-\infty}^{\infty} G(F(\zeta))f'(\zeta)\,d\zeta}{\int_{-\infty}^{\infty} G(F(\zeta))f(\zeta)\,d\zeta}.$$

In this notation, using (2.4), (2.5), and (2.7), we arrive at

$$s(a) = aK/\sigma. \tag{4.6}$$

We now seek to determine σ. From (3.6) and (4.6) and assuming p is fixed at a single value, we have

$$\sigma = \frac{1}{K}\int_{a=-\infty}^{a=\infty} av(a)[u(p) - p]\,da. \tag{4.7}$$

In the nonstochastic case, $u(p) \quad p \approx 1 \cdot\cdot p$. In the selectionist case, $u(p) - p \approx 2N_e sp$. In the case of a harmful mutation where $2N_e s \ll 0$, $u(p) \approx 0$. Near-neutral mutations contribute essentially nothing to σ. Let us make the approximation that the selectionist case pertains on $(0^+, \infty)$ and the harmful mutation formula holds on $(-\infty, 0^-)$. Then, from (4.7) and applying (4.6), we arrive at the following quadratic equation for σ:

$$\sigma^2 - \frac{I_1 p}{K}\sigma - 2N_e p I_2 = 0,$$

where $I_1 = -\int_{-\infty}^{0} av(a)\,da$ and $I_2 = \int_{0+}^{\infty} a^2 v(a)\,da$. Now $R = 2N_e p \int_{0+}^{\infty} sv(a)\,da$, so that, using (4.6) again and solving for σ, we obtain

$$\sigma = \frac{1}{K}(I_1 p + RI_2/I_3),$$

where $I_3 = \int_{0+}^{\infty} av(a)\,da$. Using (4.6) to solve for s, we arrive at

$$s(a) = \frac{aK^2}{I_1 p + RI_2/I_3}. \tag{4.8}$$

Formula (4.8) generalizes and corrects the formulae existing in the literature. For example, in the simplest case, where we neglect harmful mutants and all beneficial mutants have the same effect a_0, $I_1 = 0$ and (4.8) reduces to

$$s = K^2/R, \tag{4.9}$$

correcting the result $R = K^2/[s(1-p)]$ given in Phelps (1991), which ignored stochastic factors. But ignoring harmful mutants is obviously unwarranted. The power of (4.8) is that it expresses s as a simple function of the nonneutral substitution rate R and *any* mutation profile $v(a)$, which may include harmful mutations.

A more realistic simple case is to assume that mutations are one of three types: neutral (which do not enter the formulae), harmful with effect a_- and mutation rate v_{1-} per individual per generation, or beneficial with effect $-a_+$ and mutation rate v_{1+} per individual per generation. Let $A = a_-/a_+$. Then

$$s(a_+) = \frac{K^2}{R + Av_{1-}}. \tag{4.10}$$

Assuming harmful mutants are often more harmful than beneficial mutations are helpful, we suggest $1 < A < 5$. Following Maynard Smith (1989), we assume the mutation rate per base pair per cell division is in the range of 10^{-9} to 10^{-10} and a generation in higher animals represents 30 cell divisions. Assuming a genome consisting of 3×10^9 base pairs, we obtain a mutation rate in the range of 10 to 100 sites per individual per generation. Certainly, at least half of the mutations occurring in the 2% of the sites which are translated should be deleterious so that $0{\cdot}1 < v_{1-} < 100$. From Kimura (1968) we have $0 < R < 2$. As for K, we take $0.1 < K < 1$ as suggested by the linear rank selection model $G(y) = m(1-y)$, for which $K^2 = 1/\pi$. With these numbers $10^{-5} < s < 10$. Admittedly, this is a broad range, but in principle there is no trouble keeping $2N_e s \gg 1$ so that the selectionist case pertains. If, as is likely, most mutations in nontranslated loci are neutral, s will be safely larger than the lower bound (10^{-5}) given above.

4.4 *Applications to the mutation load*

Both the hard and rank selection formulae derived above contain the variable p, which we have taken to be the initial frequency of a mutant favourable allele, usually $1/N$. But a new *harmful* mutation with initial frequency $1/N$ creates a polymorphic locus with the frequency of the favourable allele equal to $(N-1)/N$. *Thus we can consider a harmful mutant as a favourable mutant with initial frequency* $p = (N-1)/N$.

This trick enables us to use the same formulae to evaluate the mutational load (ML), which is a measure of the loss of population fitness due to the weeding out of harmful mutations. Understanding how harmful mutations are eliminated is of vital

long-term concern to the human race. In particular, will there be a buildup of unfavourable mutations in the human gene pool due to long-term exposure to low-level radiation?

From (4.2) above, assuming nth-order epistasis, i.e., $g(\zeta) = \zeta^n$, and that $v(a, p) = v(a)\delta(p - (N-1)/N)$, we have

$$L = \frac{1}{n}\int_{a=0}^{a=\infty} v(a)\mathscr{F}(s(a), (N-1)/N)\,da.$$

Now, with $p = (N-1)/N$, the stochastic case does not arise because $2N_e s \gg 1$ implies $2N_e sp \gg 1$ as well. This suggests that ignoring stochasticity does not alter formulae for the ML. Thus there are only two cases: the nonstochastic (because fixation of the favourable allele, i.e. elimination of the unfavourable allele, is essentially certain) limit ($2N_e s \gg 1$) and the neutral limit ($2N_e s \ll 1$). A simple calculation shows that the ML essentially disappears in the neutral case, so we will treat it no further.

In the nonstochastic case (4.4), $u = 1$ (from (3.5)) and $\mathscr{F}(s, (N-1)/N) = \ln[N/(N-1)] = 1/N$. Thus

$$L = v_1/n, \tag{4.11}$$

where $v_1 = (1/N)\int v(a)\,da$ is the rate of harmful mutations per individual per generation. Remarkably, this formula is independent of the effects of the mutations. Note that, just as in the SGL, nth-order epistasis reduces the ML by a factor of n. This extends the work of Kimura & Maruyama (1966), who established the formula for $n = 2$. Note also that if most mutations are harmful, so that $10 < v_1 < 100$, then the mutation load is intolerably large if selection is hard. Again, there are two alternatives: either most mutations are neutral or selection is soft.

Just as with the SGL, rank selection models eliminate any problem with mutation loads. Equation (4.10) above applies just as well to this case.

Let us consider the question of how many harmful mutations can be eliminated by rank selection before the s becomes so small that the selectionist case no longer pertains and harmful mutations begin to be fixed by drift. Put $R = 2N_e s v_{1+}$ in (4.10) and solve for s. Using the selectionist criterion, $2N_e s \gg 1$, we obtain

$$v_{1-} \ll \frac{2N_e K^2 - v_{1+}}{A}. \tag{4.12}$$

Unless the effective population size becomes very small, no population experiencing rank selection should be unable to withstand a harmful mutation rate even as high as $v_{1-} = 100$.

As to the question posed above about the accumulation of harmful mutations in the human gene pool, these calculations make it clear that if selection is predominantly hard there will be a heavy price to pay in mean population fitness, while if selection is approximated by rank selection models the population will remain healthy as long as (4.12) holds.

5. Discussion

Incorporating stochasticity into our model does not alter the conclusion that we must reject Kimura's original hard selection load argument for the neutral theory. Including stochasticity strengthens the case for the rank selection resolution of 'Haldane's dilemma'. As argued in Phelps (1991), rank selection, with all its shortcomings, is likely to be a much better model for adaptive natural selection than hard selection. For a given rank selection model, (4.8) gives the selection coefficient as a function of the rate of nonneutral substitution and the mutation profile. As a general model for the evolution of a genome, it is much superior to the commonly utilized multiplicative fitnesses model and may prove to be of use in other contexts.

The ability of a population to withstand a high mutation rate is closely related to the mode (hard or rank) of selection eliminating the harmful mutations. The importance of this question seems to imply that experiments should be performed to determine the nature of selection.

References

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

Crow, J. F., & Kimura, M., 1978. Efficiency of truncation selection. *Proc. Natl. Acad. Sci. USA* **76**, 396–9.

Haldane, J. B. S., 1957. The cost of natural selection. *J. Genet.* **55**, 511–24.

Kimura, M., 1968. Evolutionary rate at the molecular level. *Nature* **217**, 624–6.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.

Kimura, M., 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press.

Kimura, M., & Maruyama, T., 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* **54**, 1337–51.

Kimura, M., & Maruyama, T., 1969. The substitutional load in a finite population. *Heredity* **24**, 101–14.

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

Maynard Smith, J., 1989. *Evolutionary Genetics.* Oxford University Press.

Milkman, R., 1967. Heterosis as a major cause of heterozygosity in nature. *Genetics* **55**, 493–95.

Mueller, H. J., 1950. Our load of mutations. *Am. J. Human Genet.* **2**, 111–76.

Phelps, F. M., 1991. A unifying model for the substitutional genetic load. *IMA J. Math. Appl. Med. Biol.* **8**, 31–56.

Sved, J. A., 1968. Possible rates of gene substitution in evolution. *Am. Nat.* **102**, 283–93.

Wills, C., 1978. Rank order selection is capable of maintaining all genetic polymorphisms. *Genetics* **89**, 403–17.