IMA Journal of Mathematics Applied in Medicine & Biology (1991) **8**, 57–72

Multicomponent Rank Selection as an Alternative to Haldane's Dilemma

F. M. PHELPS IV

290 Cedar Dr., Mt. Pleasant, Michigan 48858, USA

[Received 16 January 1990 and in revised form 10 October 1990]

This paper demonstrates that multicomponent hard selection (hard selection performed sequentially in n independent fitness components of the genome) is incompatible with observed substitution rates. It also shows that multicomponent rank selection permits enough differential viability to account for arbitrarily rapid evolution. This removes a major objection to the acceptance of 'rank' or 'soft' selection as a resolution of 'Haldane's dilemma' and provides firm grounds for rejecting Kimura's genetic load argument for the neutral theory.

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

1. Introduction

IN 1957 Haldane published a paper on the 'cost of natural selection' which seemed to imply an extremely restrictive upper limit to the rate of gene substitution r of about 1 substitution every 300 generations. Haldane himself was bothered by his calculation (Hoyle & Wickramasinghe, 1984), but never published an alternative. The problem became known as 'Haldane's dilemma'.

When protein sequence data became available, Kimura (1968) estimated that the rate of evolution r in higher vertebrates has been of the order of 2 substitutions per generation, which contrasts sharply with Haldane's calculation. Kimura resolved the discrepancy by proposing the neutral theory. This theory avoids Haldane's 'cost', which Kimura called 'the substitutional genetic load', by assuming that there is essentially no selection.

Since 1957, Haldane's calculation has been criticized by several authors. In particular, Haldane and Kimura assumed 'multiplicative fitnesses', which means that each locus contributes to fitness independently of the others. The most popular and influential resolution of Haldane's dilemma (other than Kimura's neutral theory) has been truncation selection (Sved, 1968). Recently Phelps (1991) has generalized truncation selection to the concept of 'rank selection' and shown that there is no theoretical bound to r under any form of rank selection.

However, under rank selection, the entire genome interacts to determine an individual's rank and, therefore, its fitness. Surely this is unrealistic: some components of fitness must be independent of others. This criticism of truncation rank selection was first made by Crow (1970). In particular, if different parts of the genome are expressed at different stages of development, it is impossible for genes which are expressed late in the life cycle to influence viability at an early

stage. Another example of independent contributions to fitness must be genes which detoxify specific poisons or confer resistance to specific diseases.

We seek a model which is intermediate between the two extremes of total independence or total interdependence of fitness effects in the genome. The model we derive will allow fitness to be a function of n independent components. Practically speaking, it is impossible to determine the components or even guess as to the size of n, but the analysis will allow us to judge how important the effect of partial independence is in the discussion of possible rates of gene substitution. This is the primary purpose of this paper. An additional benefit is that the substitutional genetic load in hard selection is found to be as Haldane calculated, with modification for the degree of epistatic interaction, even if alleles have varying detrimental effects.

In Section 2 the definitions of hard and rank selection are reviewed and the model equations are derived. In Section 3, the equations are analysed in the case of multicomponent hard selection. In Section 4, rank selection in n independent fitness components or during n sequential stages of development is discussed. In Section 5, an up-to-date resolution of Haldane's dilemma is proposed and related problems are discussed.

2. The model

2.1 *Hard and Rank Selection*

As there are differing concepts of hard and rank selection in the literature, it is best to state our definitions clearly at the outset.

Let w denote the fitness (probability of survival to adulthood or viability) of an individual and let x denote the number of inferior alleles it carries. If fitness is determined (up to a constant factor) solely by x, then selection is said to be hard.

For each x, let $R(x)$ denote the fraction of individuals in a population having less than or equal to x inferior alleles. Then $R(x)$ is said to be the rank of an individual with x inferior alleles. If fitness is a function of rank, that is, if there is a function G such that $w(x) = G(R(x))$, then the scheme is called rank selection and G is called the viability function.

The multicomponent generalizations of hard and rank selection are straightforward. Let $\boldsymbol{x}$ denote the vector with ith component equal to the number of inferior alleles in the ith component of the genome. Multicomponent hard selection means that w is determined (up to a constant factor) by $\boldsymbol{x}$ alone. Multicomponent rank selection means that rank selection is performed in each of the components in succession. The viability function G may depend on the component.

2.2 *The Multicomponent Equations*

As in essentially all other discussions of the substitutional genetic load, the model will be deterministic. Therefore, every superior mutant will spread so that, in a steady flux, the substitution rate r_i per generation in the ith component must equal the number of loci in the ith component at which a favourable allele

appears each generation. To simplify the treatment, we will assume diploids with no dominance.

Let $\boldsymbol{Z}$ be the vector, with ith component denoted Z_i, equal to the number of inferior alleles in an offspring in the ith component.

At each segregating locus, there is a favourable and an unfavourable allele. A diploid individual can have 0, 1, or 2 copies of a certain unfavourable allele. Assume that (in a given component) all superior alleles have the same frequency when they first arise in the population, as well as the same effect on fitness, so that they spread at the same rate. Then the frequency of a superior allele in the ith component, j generations after it first appears in the population is a fixed quantity, denoted $p_j^{(i)}$. Assuming Mendelian segregation, the probability that an individual receives a superior allele from its mother at a locus which has been undergoing segregation for j generations is $p_j^{(i)}$. The same probability applies to the allele donated paternally. For every nonnegative integer j, there are r_i loci in component i which have been segregating for j generations, giving $2r_i$ chances to receive an inferior allele, each with probability $q_j^{(i)} = 1 - p_j^{(i)}$. Therefore, the (random) number Z_i of inferior alleles in component i than an offspring has before selection is given by

$$Z_i = \sum_{k=1}^{2r_i} \sum_{j=0}^{\infty} B_{j,k}^{(i)},$$

where $B_{j,k}^{(i)}$ is 0 with probability $p_j^{(i)}$ and 1 with probability $q_j^{(i)}$. If we assume no linkage, the collection of random variables is independent. Note that the random variables $B_{j,k}^{(i)}$ $(k = 1,..., 2r_i)$ are independent and identically distributed.

Let $w(\boldsymbol{x})$ be the fitness (viability) of an individual for which $\boldsymbol{Z} = \boldsymbol{x}$. The set of possible values for $\boldsymbol{x}$ is the set $\mathbb{N}^n$, the set of vectors in $\mathbb{R}^n$ with nonnegative integer coordinates. By definition, the average viability of an offspring is

$$V = \sum_{\boldsymbol{x} \in \mathbb{N}^n} w(\boldsymbol{x}) P(\boldsymbol{x}), \tag{1}$$

where $P(\boldsymbol{x})$ denotes the probability that the vector $\boldsymbol{Z}$ of inferior alleles is $\boldsymbol{x}$.

To derive an equation for the selection coefficient, let $V_j^{(i+)}$ be the expected viability given a favourable allele at the locus associated with $B_{j,k}^{(i)}$ (which means $B_{j,k}^{(i)} = 0$), which is independent of k. Let $Q_j^{(i)}(\boldsymbol{x})$ be the probability that $\boldsymbol{Z} - B_{j,1}^{(i)} \boldsymbol{e}_i$ is $\boldsymbol{x}$, where $\boldsymbol{e}_i$ is the unit vector in the ith direction. By definition,

$$V_j^{(i+)} = \sum_{\boldsymbol{x} \in \mathbb{N}^n} w(\boldsymbol{x}) Q_j^{(i)}(\boldsymbol{x}).$$

If $N(t)$ is the population in generation t and B is the per capita birth rate, then

$$N(t)BV = N(t+1).$$

The number of carriers of a superior allele which has been segregating for j generations is $N(t)Bp_j^{(i)}$ before selection and $N(t)Bp_j^{(i)}V_j^{(i+)}$ or $p_{j+1}^{(i)}N(t+1)$ afterwards. Therefore

$$p_j^{(i)} V_j^{(i+)} = p_{j+1}^{(i)} V.$$

With $\Delta p_j^{(i)} = p_{j+1}^{(i)} - p_j^{(i)}$, these equations reduce (after conditioning on $B_{j,k}^{(i)}$ to find the relationship between P and $Q_j^{(i)}$ and much simplification) to

$$\Delta p_j^{(i)} = s_j^{(i)} p_j^{(i)} q_j^{(i)}, \tag{2}$$

where $s_j^{(i)}$ is given by

$$s_j^{(i)} V = \sum_{x \in \mathbf{N}^n} w(x)\, \Delta_i Q_j^{(i)}(x) \tag{3}$$

and $\Delta_i Q_j^{(i)}(x) = Q_j^{(i)}(x) - Q_j^{(i)}(x - e_i)$. This is called the 'selection equation'. The model is fully specified by equations (1), (2), and (3).

3. Approximations

Because we have sectioned the genome into n independent components, the number of inferior alleles an offspring has in a given component is independent of the number in any other component. Thus, we expect the distribution of $\mathbf{Z}$ to be approximately multivariate normal with independent components.

We will not make a mathematically rigorous analysis of these equations, which would include lengthy and technical proofs of the applicability of a generalized version of the central limit theorem (CLT). Instead, we will assume that the CLT is a good approximation if $s_j^{(i)}$ is 'small' for all i and j. Also, the dependence of $s_j^{(i)}$ on j is slight and can safely be ignored in situations where the CLT applies. The approximate selection coefficient will be denoted $s^{(i)}$.

The application of the CLT will be to approximate the selection and viability equations, (3) and (1), by integrals involving the density function for a normally distributed random variable.

The assumption of slow selection also allows us to estimate the mean μ_i and variance σ_i^2 of Z_i. Because the $B_{j,k}^{(i)}$ are independent,

$$\mu_i = \mathsf{E}\Big[\sum_{k=1}^{2r_i} \sum_{j=0}^{\infty} B_{j,k}^{(i)}\Big] = 2r_i \sum_{j=0}^{\infty} \mathsf{E} B_{j,1}^{(i)} = 2r_i \sum_{j=0}^{\infty} q_j^{(i)}.$$

If $s^{(i)}$ is small, then

$$\sum q_j^{(i)} \approx \int_{t=0}^{\infty} q(t)\, \mathrm{d}t,$$

where $q(t) = 1 - p(t)$ satisfies

$$\frac{\mathrm{d}p}{\mathrm{d}t} = s^{(i)} p(t) q(t).$$

The result of changing variables from t to p and integrating is

$$\mu_i \approx \frac{2r_i}{s^{(i)}} \log\,(1/p_0^{(i)}). \tag{4}$$

Similarly,

$$\sigma_i^2 \approx \frac{2r_i}{s^{(i)}} (1 - p_0^{(i)}). \tag{5}$$

We will assume that $p_0^{(i)}$ is independent of i and so will denote it by p_0.

4. Multicomponent hard selection

The goal of this section is to compute the rate of substitution in a large class of multicomponent hard selection models.

Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ denote the vectors with ith component μ_i and σ_i^2, respectively. Let x_i denote the ith component of $\boldsymbol{x}$. Suppose that fitness is given by

$$w(\boldsymbol{x}) = w(\mathbf{0})\mathrm{e}^{-\varepsilon g(\boldsymbol{x})}$$

where $g(\boldsymbol{x})$ is positive. If ε is 0, there is no selection, so weak selection must correspond to small ε. Although $\boldsymbol{Z}$ is a discrete random variable, we approximate it by a continuous multivariate normal random variable. For small ε, the integral approximation of the viability is

$$V \approx \int_{\mathbf{R}_+^n} w(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where $f(\boldsymbol{x})$ is the density function of the multivariate normal distribution

$$(2\pi)^{-n/2} \prod_{i=1}^{n} \left(\frac{1}{\sigma_i}\right) \exp\left[-\tfrac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right].$$

If we make the change of variables $y_i\mu_i = x_i$, the viability integral becomes

$$(2\pi)^{-n/2}\left(\prod_{i=1}^{n} D_i\right) w(\mathbf{0}) \int_{\mathbf{R}_+^n} \exp\left[-\varepsilon g(\mu_1 y_1, \ldots, \mu_n y_n)\right] \prod_{i=1}^{n} \exp\left[-\tfrac{1}{2} D_i^2 (y_i - 1)^2\right] \mathrm{d}\boldsymbol{y},$$

where $D_i = \mu_i/\sigma_i$. For slow selection, D_i is approximately

$$\left(\frac{2r_i}{s^{(i)}(1-p_0)}\right)^{\frac{1}{2}} \log(1/p_0),$$

which is large for fixed r_i and small $s^{(i)}$. As D_i goes to infinity, the function

$$\frac{D_i}{(2\pi)^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2} D_i^2 (y_i - 1)^2\right]$$

approaches $\delta(y_i - 1)$. Thus, the viability is given approximately by

$$V \approx w(\mathbf{0})\mathrm{e}^{-\varepsilon g(\boldsymbol{m})}.$$

The difference $\Delta_i q_j^{(i)}(\boldsymbol{x})$ can be approximated by the derivative $(\partial/\partial x_i) f(\boldsymbol{x})$. The integral approximation to the selection equation is then

$$s^{(i)} V \approx \int_{\mathbf{R}_+^n} w(\boldsymbol{x}) \frac{\partial f}{\partial x_i}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

If the right-hand side is integrated by parts (the boundary terms are negligible for D_i large), we have

$$V s^{(i)} = \varepsilon \int_{\mathbf{R}_+^n} \frac{\partial g}{\partial x_i}(\boldsymbol{x}) w(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

By the same change of variables and delta-function argument,

$$Vs^{(i)} \approx \varepsilon \frac{\partial g}{\partial x_i}(\boldsymbol{\mu}) w(\mathbf{0}) e^{-\varepsilon g(\boldsymbol{\mu})},$$

which simplifies to

$$s^{(i)} \approx \varepsilon \frac{\partial g}{\partial x_i}(\boldsymbol{\mu}).$$

Let $L = \log[w(\mathbf{0})/V]$, so that $L = \varepsilon g(\boldsymbol{\mu})$. The quantity L, the log of the ratio of the fitness of an 'optimal' individual to the population average, is known as the 'substitutional genetic load' in the literature.

Let us now consider a broad class of fitness functions. Suppose g is of the form

$$g(\boldsymbol{x}) = (a_1 x_1^m + \cdots + a_n x_n^m)^{c/m}.$$

Here the vector $\boldsymbol{a} = (a_1, \ldots, a_n)$ must have positive components, m is any positive real number, and c is a positive number known as the 'order of epistatic interaction'.

Let the number A be defined by

$$A = (a_1 \mu_1^m + \cdots + a_n \mu_n^m)^{-1}.$$

The formula for $s^{(i)}$ and some computation shows that

$$\mu_i s^{(i)} \approx L c a_i \mu_i^m A.$$

If we sum this equation over i, use (4), and rearrange, we obtain

$$r \approx \frac{Lc}{2 \log(1/p_0)}, \tag{6}$$

where $r = \sum_i r_i$ is the total substitution rate. This is a very interesting formula in that it agrees with the single component formula (Phelps, 1991) for cth-order epistasis *independently of* $s^{(i)}$, ε, $\boldsymbol{a}$, m, *and* n! Thus, the rate of substitution in hard selection with fitness functions of this quite general form is not affected by the presence of components, differing effects of the alleles, or the variable m. What *is* important is the order of epistatic interaction c, and r is directly proportional to c.

As for Kimura's argument for the neutral theory, the presence of components does not help hard selection explain a rate of substitution of the order of $r = 2$, which cannot be achieved in (6) above for realistic parameter values.

5. Multicomponent rank selection

Suppose that selection operates by performing rank selection on each of the n components in succession. Let G_i denote the viability function for the ith component and let $R_i(m_i)$ denote the rank in the ith component of an individual with m_i inferior alleles in the ith component. The fitness of an individual with

$\boldsymbol{Z} = \boldsymbol{m}$ is then

$$w(\boldsymbol{m}) = \prod_{i=1}^{n} G_i(R_i(m_i)),$$

while the viability is

$$V = \sum_{\boldsymbol{m} \in \mathbf{N}_+^n} \left(\mathsf{P}[\boldsymbol{Z} = \boldsymbol{m}] \prod_{i=1}^{n} G_i(R_i(m_i)) \right).$$

Again we note that the components of $\boldsymbol{Z}$ form an independent collection of random variables. Therefore

$$\mathsf{P}[\boldsymbol{Z} = \boldsymbol{m}] = \prod_{i=1}^{n} \mathsf{P}[Z_i = m_i]. \tag{7}$$

This enables us to factor the equation for V as follows:

$$\begin{aligned} V &= \sum_{\boldsymbol{m} \in \mathbf{N}_+^n} \left(\prod_{i=1}^{n} G_i(R_i(m_i)) \mathsf{P}[Z_i = m_i] \right) \\ &= \prod_{i=1}^{n} \sum_{m_i=0}^{\infty} G_i(R_i(m_i)) \mathsf{P}[Z_i = m_i]. \end{aligned}$$

Let V_i be defined by

$$V_i = \sum_{m_i=0}^{\infty} G_i(R_i(m_i)) \mathsf{P}[Z_i = m_i],$$

so that $V = \prod_{i=1}^{n} V_i$. Because of the independence of the components, the order of the sequential selection does not affect V. In fact, we can consider selection to operate in all components at a single instant according to this formula.

The integral approximation to V_i is

$$V_i = \int_{-\infty}^{\infty} G_i(F(x)) f(x)\, dx,$$

where F and f are the cumulative distribution and density functions of an N(0, 1) random variable.

The equation for the ith selection coefficient is

$$s_j^{(i)} V \approx \sum_{\boldsymbol{m} \in \mathbf{N}_+^n} \left(\prod_{l=1}^{n} G_l(R_l(m_l)) \right) \Delta_i Q_j^{(i)}(\boldsymbol{m}).$$

If we write out $\Delta_i Q_j^{(i)}(\boldsymbol{m})$ in terms of probabilities, and use independence to factor those probabilities as in (7) above, we have a product summed over n indices. If we then pull all factors independent of m_1 out of the sum over m_1, we find that what is left sums to V_1. Continue this procedure to arrive at

$$s_j^{(i)} V_i = \sum_{m_i=0}^{\infty} G_i(R_i(m_i)) \left(\mathsf{P}[Z_i - B_{j,1}^{(i)} = m_i] - \mathsf{P}[Z_i - B_{j,1}^{(i)} = m_i - 1] \right).$$

We note that Z_i is distributed approximately $N(\mu_i, \sigma_i^2)$ so that $R_i(m_i)$ is nearly $F\big((m_i - \mu_i)/\sigma_i\big)$. The difference can be approximated by a derivative:

$$\begin{aligned}
&P[Z_i - B_{\cdot,1}^{(i)} = m_i] - P[Z_i - B_{\cdot,1}^{(i)} = m_i - 1] \\
&\qquad \approx \frac{1}{\sigma_i} f\left(\frac{m_i - \mu_i + B_{\cdot,1}^{(i)}}{\sigma_i}\right) - \frac{1}{\sigma_i} f\left(\frac{m_i - \mu_i + B_{\cdot,1}^{(i)}}{\sigma_i} - \frac{1}{\sigma_i}\right) \\
&\qquad \approx \frac{1}{\sigma_i^2} f'\left(\frac{m_i - \mu_i}{\sigma_i} + \frac{B_{\cdot,1}^{(i)}}{\sigma_i}\right) \\
&\qquad \approx \frac{1}{\sigma_i^2} f'\left(\frac{m_i - \mu_i}{\sigma_i}\right)
\end{aligned}$$

for large σ_i. These approximations lead to

$$s_j^{(i)} V_i \sigma_i^2 \approx \sum_{m_i = -\infty}^{\infty} G_i\left(F\left(\frac{m_i - \mu_i}{\sigma_i}\right)\right) f'\left(\frac{m_i - \mu_i}{\sigma_i}\right).$$

Because $s_j^{(i)}$ is nearly independent of j, it will be denoted simply by $s^{(i)}$.

Let

$$x_{m_i} = (m_i - \mu_i)/\sigma_i \quad \text{and} \quad \Delta x = 1/\sigma_i.$$

Then

$$s^{(i)} V_i \sigma_i \approx \sum_{m_i = -\infty}^{\infty} G_i\big(F(x_{m_i})\big) f'(x_{m_i})\, \Delta x.$$

The integral approximation is

$$s^{(i)} V_i \sigma_i \approx \int_{-\infty}^{\infty} G_i\big(F(x)\big) f'(x)\, dx. \tag{8}$$

The estimates for μ_i and σ_i are as before. Let K_i be the ratio of the two integrals

$$K_i = \frac{\displaystyle\int_{-\infty}^{\infty} G_i\big(F(x)\big) f'(x)\, dx}{\displaystyle\int_{-\infty}^{\infty} G_i\big(F(x)\big) f(x)\, dx} = \frac{\displaystyle\int_{-\infty}^{\infty} G_i\big(F(x)\big) f'(x)\, dx}{V_i}.$$

Then, from (8), we have

$$(s^{(i)})^2 \sigma_i^2 \approx K_i^2.$$

Using (5) above and solving for r_i, this becomes

$$r_i \approx \frac{K_i^2}{2(1 - p_0) s^{(i)}}.$$

To find r, we sum over i:

$$r \approx \sum_{i=1}^{n} \frac{K_i^2}{2(1 - p_0) s^{(i)}}.$$

This is a fundamental equation for the rate of substitution in a rank selection model with components. Recall that, in hard selection, r_i was found to be independent of $s^{(i)}$, but here r_i is inversely proportional to $s^{(i)}$.

Note that G_i uniquely determines the value of K_i. Therefore, in order to determine r, we must know each of the n selection coefficients $s^{(i)}$ and each of the n viability functions G_i. Of course, there is no hope of obtaining such data (or even identifying the components) in the near future, but this entire investigation is an attempt to determine what substitution rates are theoretically *possible*. In the following subsections, we will consider *plausible* choices of the unknowns.

5.1 *Sequential Threshold Selection*

Threshold or 'truncation' selection has a long history in the literature of population genetics and has applications to artificial selection in breeding programmes. The viability function for truncation selection is

$$G_i(y) = \begin{cases} 1 & \text{if } y < V_i, \\ 0 & \text{if } y \geq V_i. \end{cases}$$

From the definition of K_i,

$$K_i = \frac{f(F^{-1}(V_i))}{V_i}.$$

Thus, we can think of K_i as a function of the parameter V_i. Let the function H be defined on $(0, 1)$ by

$$H(x) = K_i^2(x) = \left(\frac{f(F^{-1}(x))}{x}\right)^2.$$

Let us assume the selection coefficients $s^{(i)}$ are all equal to s.

A question of major interest is how the n-component model compares to the one-component model. In the one-component model,

$$r = \frac{1}{2(1-p_0)s} H(V),$$

while, in the n-component model,

$$r = \frac{1}{2(1-p_0)s} \sum_{i=1}^{n} H(V_i).$$

In order to make a fair comparison, both models should have the same viability, so that

$$V = \prod_{i=1}^{n} V_i.$$

The easiest case to analyse is the one where $V_i = V^{1/n}$ for every i. In this case, let

$r_n(V)$ be the rate of substitution in an n-component model. Then

$$r_n(V) = \frac{n}{2(1-p_0)s} H(V^{1/n}).$$

It is important to observe that, even in multicomponent selection, the substitution rate is inversely proportional to the selection coefficient. This means that *any* rate of substitution can be achieved if s is small enough. The quantity $r_n(V)$ with $s = 0{\cdot}01$ and p_0 small is plotted for various values of n in Fig. 1. Note that, even with $n = 100$, it is possible to achieve Kimura's estimated rate $r = 2$ with roughly 50% mortality. Multicomponent threshold selection is seen to be less efficient than single-component threshold selection, but $r = 2$ is still attainable for $s = 0{\cdot}01$ and fairly realistic viability. Furthermore, the selection coefficient may be much less than $0{\cdot}01$, which makes $r = 2$ even less of a problem.

Let $\rho_n(V)$ be the ratio r_n/r_1. For sequential threshold selection,

$$\rho_n(V) = \frac{nH(V^{1/n})}{H(V)}.$$

This function is plotted in Figs. 2 and 3. It is interesting to note that, for $n = 100$ and a wide range of viabilities, sequential threshold selection is about 5–10% as efficient as single-component threshold selection.

It is important to know how $\rho_n(x)$ behaves for large n. This is partially answered in the next two lemmas.

LEMMA 1 *For all* $x \in (0, 1)$, $\lim_{n\to\infty} n\rho_n(x) = \infty$.

Proof. Let

$$\alpha_n(x) = \frac{f(F^{-1}(x))}{x}[n\rho_n(x)]^{\frac{1}{2}}.$$

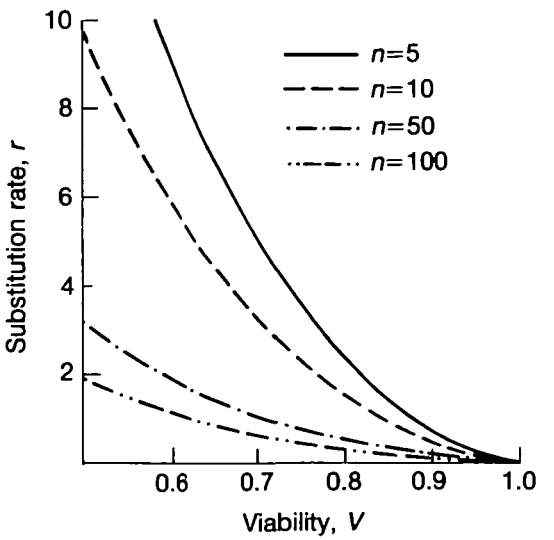

FIG. 1. Substitution rate in n-component sequential truncation selection with $s = 0{\cdot}01$ and p_0 small.

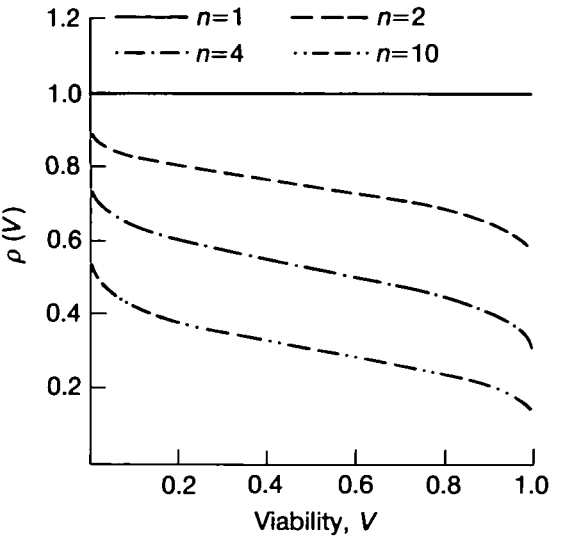

FIG. 2. Efficiency of n-component sequential truncation selection relative to simple truncation selection for small n.

Then

$$\alpha_n(x) \to \infty \quad \Leftrightarrow \quad n\rho_n(x) \to \infty.$$

Now

$$\lim_{n\to\infty} \alpha_n(x) = \left[\lim_{n\to\infty} \left(\frac{f(F^{-1}(x^{1/n}))}{1/n}\right)\right]\left[\lim_{n\to\infty} x^{-1/n}\right].$$

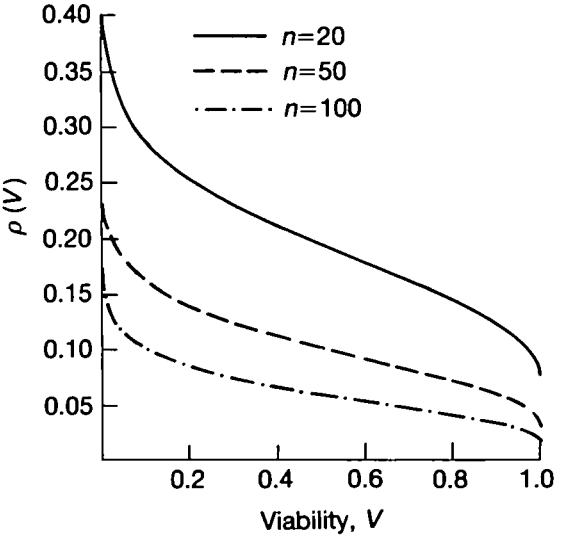

FIG. 3. Efficiency of n-component sequential truncation selection relative to simple truncation selection for large n.

The second factor goes to one and so can be ignored. If we apply l'Hôpital's rule and the properties of the functions F and f, we find that

$$\lim_{n\to\infty} \alpha_n(x) = \lim_{n\to\infty} \frac{f'(F^{-1}(x^{1/n}))(\log x)x^{1/n}}{f(F^{-1}(x^{1/n}))} = -(\log x) \lim_{n\to\infty} F^{-1}(x^{1/n}).$$

Since $\log x$ is negative and $x^{1/n} \to 1$, the right-hand side goes to $+\infty$. □

This result is useful in proving the next lemma.

LEMMA 2 *If $c < 1$, then, for all $x \in (0, 1)$, $\lim_{n\to\infty} n^c \rho_n(x) = 0$.*

Proof. Let

$$\alpha_n(x) = \frac{f(F^{-1}(x))}{x} [n^c \rho_n(x)]^{\frac{1}{2}}.$$

Then

$$\alpha_n(x) \to 0 \quad \Leftrightarrow \quad n^c \rho_n(x) \to 0.$$

Now

$$\lim_{n\to\infty} \alpha_n(x) = \left[\lim_{n\to\infty} \frac{f(F^{-1}(x^{1/n}))}{n^{-\beta}}\right]\left[\lim_{n\to\infty} x^{-1/n}\right],$$

where $\beta = \frac{1}{2}(1 + c) < 1$. As in the previous lemma, the second factor goes to one and so can be ignored. By l'Hôpital's rule,

$$\lim_{n\to\infty} \alpha_n(x) = -(\log x) \lim_{n\to\infty} \frac{F^{-1}(x^{1/n})}{\beta n^{1-\beta}}.$$

Both the numerator and the denominator go to ∞, so we apply l'Hôpital's rule again:

$$\begin{aligned}\lim_{n\to\infty} \alpha_n(x) &= \frac{(\log x)^2}{\beta(1-\beta)} \left(\lim_{n\to\infty} \frac{1}{n^{1-\beta}}\right)\left(\lim_{n\to\infty} \frac{1}{nf(F^{-1}(x^{1/n}))}\right) \\ &= \frac{x}{f(F^{-1}(x))} \frac{(\log x)^2}{\beta(1-\beta)} \left(\lim_{n\to\infty} \frac{1}{n^{1-\beta}}\right)\left(\lim_{n\to\infty} \frac{1}{x^{1/n}[n\rho_n(x)]^{\frac{1}{2}}}\right).\end{aligned}$$

Since $\beta < 1$, the first limit on the right-hand side goes to zero. The second limit goes to zero by the previous lemma. □

These lemmas show that, for large n, the substitution rate decays slightly less rapidly than the function $1/n$.

5.2 *Sequential Linear Rank Selection*

It may be objected that truncation selection is too extreme to be put forth as a model of natural selection. It is therefore necessary to investigate the behaviour of linear rank selection models with n components. Consider the general linear viability function

$$G_i(y) = b_i - m_i y.$$

For a fixed viability V_i, substitution is most rapid when the absolute value of the slope m_i is maximized, which is when $b_i = 1$. (It should be intuitively clear that, when $b < 1$, there is a $1 - b_i$ probability that a given individual will not be viable *in addition to* the differential viability due to the slope $-m_i$, so that not all of the deaths contribute to directional selection.) Therefore, for a fair comparison, let us assume $b_i = 1$ which implies $m_i = 2(1 - V_i)$. A simple calculation gives

$$r_i = \frac{1}{2\pi(1-p_0)s^{(i)}}\left(\frac{1-V_i}{V_i}\right)^2.$$

Suppose that $s^{(i)} = s$ for all i. Then

$$r = \frac{1}{2\pi(1-p_0)s}\sum_{i=1}^{n}\left(\frac{1-V_i}{V_i}\right)^2.$$

Again, let us assume that $V_i = V^{1/n}$ for all i. If $r_n(V)$ and $\rho_n(V)$ are defined as in the previous section, then

$$r_n(V) = \frac{n}{2\pi s(1-p_0)}\left(\frac{1-V^{1/n}}{V^{1/n}}\right)^2$$

and

$$\rho_n(V) = n\left(\frac{1-V^{1/n}}{V^{1/n}}\right)^2\left(\frac{V}{1-V}\right)^2.$$

The substitution rate with $s = 0{\cdot}01$ and p_0 small is plotted in Figs. 4 and 5. In Fig. 4, we see that the existence of components greatly reduces the efficiency of the substitution process. If n is larger, Fig. 5 indicates that the reduction is quite

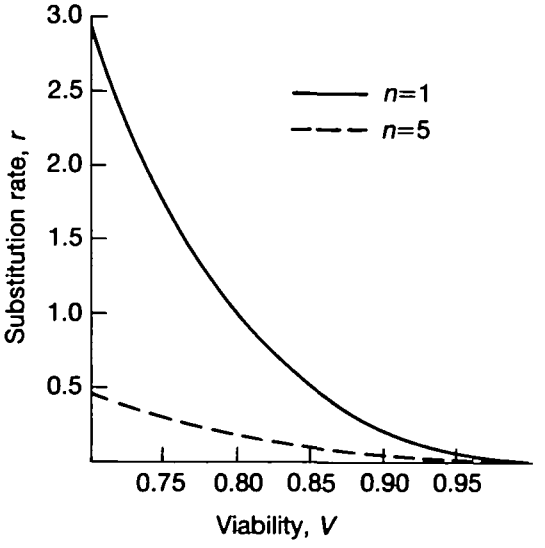

FIG. 4. Comparison of substitution rates in sequential linear rank selection with $s = 0{\cdot}01$ and p_0 small for $n = 1$ and $n = 5$.

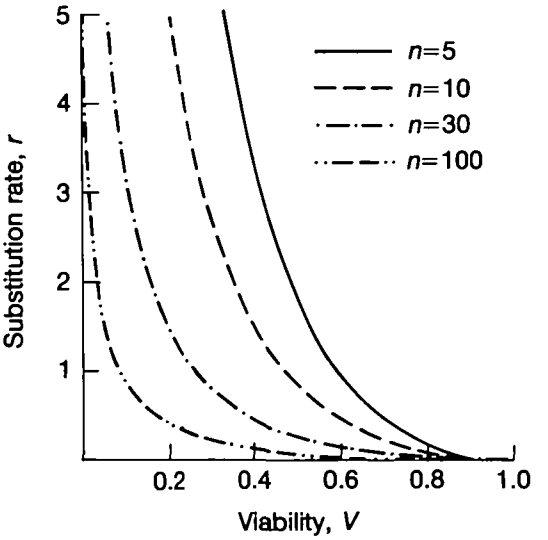

FIG. 5. Substitution rate in n-component sequential linear rank selection with $s = 0{\cdot}01$ and p_0 small.

severe. If $n = 100$ and $s = 0{\cdot}01$, it is quite unlikely that $r = 2$ because that would require well over 90% mortality due to directional (as opposed to purifying or balancing) selection each generation. This is the first indication that $r = 2$ may be difficult to achieve in a rank selection model and suggests that the 'average value' of s is unlikely to be as large as 0·01 in nature.

In Fig. 6, $\rho_n(x)$ is plotted. The behaviour is quite different from the truncation

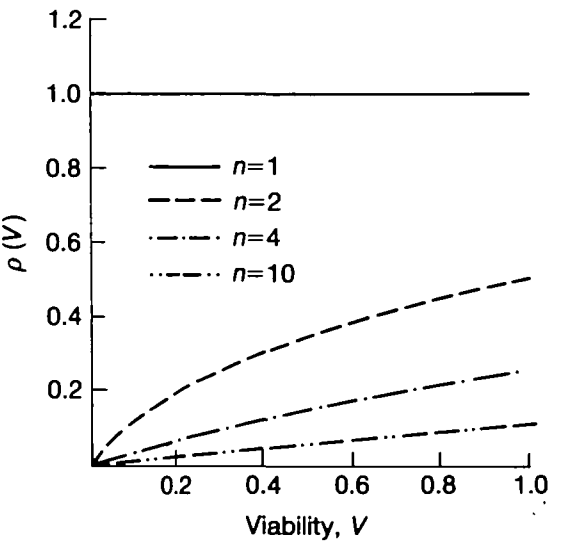

FIG. 6. Efficiency of sequential linear rank selection with n components relative to simple linear rank selection.

selection case. It is also true that the decay in efficiency is more rapid than in the threshold case. This is the content of the next lemma.

LEMMA 3 *For every* $V \in (0,1)$,

$$\lim_{n\to\infty} n\rho_n(V) = \left(\frac{V \log V}{1-V}\right)^2.$$

Proof. As in the previous section, let $\alpha_n = (n\rho_n)^{\frac{1}{2}}$. Then

$$\lim_{n\to\infty} \alpha_n(V) = \frac{V}{1-V} \lim_{n\to\infty} \left(\frac{1-V^{1/n}}{n^{-1}}\right) \lim_{n\to\infty} V^{-1/n}.$$

Apply l'Hôpital's rule to obtain

$$\lim_{n\to\infty} \alpha_n(V) = \frac{-V \log V}{1-V},$$

which implies that

$$\lim_{n\to\infty} n\rho_n(V) = \left(\frac{V \log V}{1-V}\right)^2. \qquad \square$$

It is worth stressing the point that the behaviour of sequential linear rank selection is quite different from sequential threshold selection. As this lemma shows, the differences are qualitative, and not just quantitative. Thus it is inappropriate to base a theory on even the qualitative behaviour of threshold selection models, since they are certainly too extreme to correspond to nature.

6. Haldane's dilemma and the neutral theory

'Haldane's dilemma' has a special place in the history of evolutionary theory as it was one of the primary arguments leading to the neutral theory proposed by Kimura (1968). Although this argument and the multiplicative fitnesses model on which it is based have been in disfavour for quite some time, the most widely quoted alternative, simple truncation selection, is unrealistic in two respects. First, the notion of 'truncation' is too extreme, and, secondly, truncation selection assumes that no two (or more) components of fitness are independent. Recently (Phelps, 1991) has relaxed truncation to rank selection, thereby removing the former objection. The present paper has shown that *any* multicomponent rank selection model is consistent with arbitrarily rapid rates of evolution because the rate r is inversely proportional to the selection coefficient. Rank selection models are more realistic than hard selection models because viability depends upon the quality of the competition in nature. Multicomponent rank selection is more realistic than single-component rank selection because it does not require the complete interdependence of the entire genome in determining fitness. Since such realistic models are consistent with arbitrarily rapid rates of substitution, Haldane's dilemma has been resolved. This is not an argument against the neutral theory, but merely a rejection of Kimura's original reason for proposing it.

A related issue is the intolerable segregational genetic load calculated by

Lewontin & Hubby (1966). Although details of the problems differ, multicomponent rank selection models surely will allow large amounts of heterozygosity to be maintained by heterosis. Wills (1978) has already applied single-component rank selection to this problem.

On the other hand, the application of multicomponent rank selection models to the mutational genetic load is not as straightforward and must be approached with caution. This is because rank selection may not apply to the weeding out of harmful mutations.

There remain two areas for further research. One is rigorously establishing the integral approximations used throughout this work. This was begun in Phelps (1989). A more important question from the applied side is how the model behaves if stochastic effects are included.

Acknowledgements

I wish to thank J. Keener and J. Murray for their support and W. Ewens for his helpful suggestions on the manuscript. This work was supported by NSF grant DMS-8801446 as well as by the Science and Engineering Research Council of England through a grant to the Centre for Mathematical Biology, Oxford.

References

Crow, J. F. 1970 Genetic loads and the cost of natural selection. In: *Mathematical Topics in Population Genetics* (K. Kojima, ed.). Berlin: Springer, p. 168.

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

Hoyle, F., & Wickramasinghe, C. 1984 *From Grains to Bacteria.* Cardiff, UK: University College Cardiff Press, pp. 194–5.

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

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

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.

Phelps, F. M. 1989 The substitutional genetic load and the neutral theory of molecular evolution. Ph.D. thesis, University of Utah.

Phelps, F. M. 1991 A unifying model for the substitutional genetic load. *IMA J. Math. Appl. Med. Biol.*, this issue.

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.