Genet. Res., Camb. (1974), **23**, *pp.* 23–35
With 2 text-figures
Printed in Great Britain

The hitch-hiking effect of a favourable gene

By JOHN MAYNARD SMITH and JOHN HAIGH
University of Sussex, Falmer, Brighton BN1 9QH

(*Received* 22 *May* 1973)

SUMMARY

When a selectively favourable gene substitution occurs in a population, changes in gene frequencies will occur at closely linked loci. In the case of a neutral polymorphism, average heterozygosity will be reduced to an extent which varies with distance from the substituted locus. The aggregate effect of substitution on neutral polymorphism is estimated; in populations of total size 10^6 or more (and perhaps of 10^4 or more), this effect will be more important than that of random fixation. This may explain why the extent of polymorphism in natural populations does not vary as much as one would expect from a consideration of the equilibrium between mutation and random fixation in populations of different sizes. For a selectively maintained polymorphism at a linked locus, this process will only be important in the long run if it leads to complete fixation. If the selective coefficients at the linked locus are small compared to those at the substituted locus, it is shown that the probability of complete fixation at the linked locus is approximately $\exp(-Nc)$, where c is the recombinant fraction and N the population size. It follows that in a large population a selective substitution can occur in a cistron without eliminating a selectively maintained polymorphism in the same cistron.

1. INTRODUCTION

When a selectively favourable mutation occurs in a population and is subsequently fixed in that population, this process will alter the frequencies of alleles at closely linked loci. Alleles present on the chromosome on which the original mutation occurred will tend to increase in frequency, and other alleles will thus decrease in frequency. We refer to this as the 'hitch-hiking effect', because an allele can get a lift in frequency from selection acting on a neighbouring allele. The aim of this paper is to consider the importance of the hitch-hiking effect on natural populations.

A previous attack on this problem was made by Kojima & Scheffer (1967). Unfortunately, their conclusions are wrong by several orders of magnitude; one reason for this is that they assume that each genotype has a fixed probability of replication, irrespective of the frequencies of the other genotypes in the population.

In this paper we are concerned with two problems. The first is the effects of selective substitution on genetic polymorphism for selectively neutral alleles; in § 2 we consider the effect of substitution at one locus on the frequencies of neutral alleles at a linked locus, and in §§ 3 and 4 we attempt to estimate the importance of this process for average heterozygosity. The relevance of this to controversies

about neutral polymorphism is considered in § 6. The second problem, considered in § 5, concerns the effect of substitution on a linked locus at which there is a selectively maintained polymorphism; here we are concerned only with complete fixation of one allele at the linked locus, because, if complete fixation is not achieved when substitution at the first locus is completed, then gene frequencies at the linked locus will tend to return to their equilibrium values.

2. THE EFFECT OF A SINGLE GENE SUBSTITUTION

Consider first the haploid case, where an allele b becomes substituted by a favourable allele B. At a neighbouring locus, a pair of neutral alleles a, A are segregating. Thus, in generation n, the possible genotypes, their frequencies and fitnesses are

Genotype ...	AB	aB	Ab	ab
Frequency	$p_n Q_n$	$p_n(1-Q_n)$	$(1-p_n)R_n$	$(1-p_n)(1-R_n)$
Fitness	$1+s$	$1+s$	1	1

Here $s > 0$, p_n is the frequency of B and Q_n, R_n are the proportions of A in those chromosomes containing B, b respectively.

In order to see what effect the favourable allele B has on the relative proportions of A, a, we assume that all B individuals are descended, without further mutation, from a single mutant aB individual in generation 0, and that the recombination fraction between the two loci is c. Thus $Q_0 = 0$, and we are considering here the deterministic situation.

By examining the ten possible types of mating in generation n, we obtain the following equation for the frequency of AB in generation $n+1$:

$$\begin{aligned}(1+p_n s)^2 p_{n+1} Q_{n+1} = {} & (1+s)^2 p_n^2 Q_n^2 + (1+s)^2 p_n^2 Q_n(1-Q_n) + (1+s) p_n(1-p_n) Q_n R_n \\ & + (1+s) p_n(1-p_n) Q_n(1-R_n)(1-c) \\ & + (1+s) p_n(1-p_n)(1-Q_n) R_n c.\end{aligned}$$

This simplifies to

$$(1+p_n s)^2 p_{n+1} Q_{n+1} = p_n(1+s)[Q_n(1+p_n s) + c(1-p_n)(R_n - Q_n)]. \tag{1}$$

Similarly, the equation for B is

$$(1+p_n s)^2 p_{n+1} = p_n(1+s)(1+p_n s), \tag{2}$$

so (1) simplifies to

$$(1+p_n s) Q_{n+1} = (1+p_n s) Q_n + c(1-p_n)(R_n - Q_n). \tag{3}$$

Considering Ab, we find that

$$(1+p_n s) R_{n+1} = (1+p_n s) R_n + c(1+s) p_n(Q_n - R_n). \tag{4}$$

Subtracting (4) from (3),

$$Q_{n+1} - R_{n+1} = (Q_n - R_n)(1-c)$$

and so, since $Q_0 = 0$, we have

$$Q_n - R_n = -R_0(1-c)^n. \tag{5}$$

Equation (2) can, of course, be solved explicitly to give

$$p_n = p_0(1+s)^n/\{1-p_0+p_0(1+s)^n\}, \tag{6}$$

which, with (5) and (3), yields the recurrence relation

$$Q_{n+1} = Q_n + cR_0(1-p_0)(1-c)^n/\{1-p_0+p_0(1+s)^{n+1}\}. \tag{7}$$

Thus
$$Q_\infty = cR_0(1-p_0)\sum_{n=0}^{\infty}(1-c)^n/\{1-p_0+p_0(1+s)^{n+1}\}, \tag{8}$$

which is the (deterministic) final proportion of AB in the populations when B has replaced b. The sum in (8) appears not to simplify, but, since $1-p_0+p_0(1+s)^{n+1} \geqslant 1$, we can deduce the simple (but not very useful) inequality

$$Q_\infty \leqslant R_0(1-p_0). \tag{9}$$

This calculation for a deterministic model is not a good representation of reality at the times when the frequency of B is either very small or very large. Even when B is favourable, most new occurrences of B will be eliminated by chance, and so, in the cases where B is fixed, the initial increase of B will be faster than in the deterministic model. Near fixation, we can expect B to be fixed faster than in the deterministic model, because of chance fluctuations, as reference to tables 6.1 and 6.2 of Ewens (1969), pp. 61, 62, confirms. There will therefore be fewer opportunities for recombination than the deterministic model implies, and thus the effect on the frequency of A will be correspondingly greater. Hence the deterministic model *underestimates* (but probably only by a very small amount) the drop in frequency of allele A, provided we realize that we are looking only at the cases where B is not eliminated.

Further, if $0 < c \ll s \ll 1$, and $p_0 \ll 1$, we can approximate (8) to obtain, eventually, (14), but a derivation of (14) using differential equations to approximate (2), (3) and (4) will be given. These equations are

$$\dot{p} = sp(1-p)/(1+sp), \tag{10}$$

$$\dot{Q} = c(1-p)(R-Q)/(1+sp), \tag{11}$$

$$\dot{R} = -cp(1+s)(R-Q)/(1+sp), \tag{12}$$

where a dot denotes differentiation with respect to time. Subtracting (12) from (11) and integrating, we find $Q-R = -R_0\exp(-ct)$ (compare (5)). Dividing out (10) and (11), and using the value of $Q-R$, we have

$$\frac{\mathrm{d}Q}{\mathrm{d}p} = \frac{cR_0}{sp}\mathrm{e}^{-ct}. \tag{13}$$

If c is so small that, over the major time that p increases from p_0 to 1, e^{-ct} remains effectively at 1, integration of (13) gives

$$\frac{Q_\infty}{R_0} \simeq \frac{c}{s}\log\frac{1}{p_0}. \tag{14}$$

For a range of values of c, s and p_0, Table 1 shows the values of Q_∞/R_0 calculated from (8), and compares them with the approximation (14). When Q_∞/R_0 is less than

Table 1. *Q_∞/R_0 is a measure of the proportional reduction in frequency of a neutral allele A after fixation of a linked favourable allele B*

(The table compares the exact values of Q_∞/R_0, calculated from (8), with the approximation $(c/s)\log N$, for various values of the recombination fraction c, selective advantage s and population size N.)

N	s	c	$(c/s)\log N$	Q_∞/R_0	N	s	c	$(c/s)\log N$	Q_∞/R_0
10^2	0·01	3×10^{-5}	0·0138	0·0138	10^2	0·01	3×10^{-4}	0·1382	0·1283
10^2	0·1	3×10^{-4}	0·0138	0·0142	10^2	0·1	3×10^{-3}	0·1382	0·1325
10^6	0·01	10^{-5}	0·0138	0·0138	10^6	0·01	10^{-4}	0·1382	0·1295
10^6	0·1	10^{-4}	0·0138	0·0143	10^6	0·1	10^{-3}	0·1382	0·1344
10^2	0·01	6×10^{-5}	0·0276	0·0273	10^2	0·01	6×10^{-4}	0·2763	0·2379
10^2	0·1	6×10^{-4}	0·0276	0·0282	10^2	0·1	6×10^{-3}	0·2763	0·2454
10^6	0·01	2×10^{-5}	0·0276	0·0274	10^6	0·01	2×10^{-4}	0·2763	0·2419
10^6	0·1	2×10^{-4}	0·0276	0·0285	10^6	0·1	2×10^{-3}	0·2763	0·2506
10^2	0·01	$1·5\times10^{-4}$	0·0691	0·0667	10^2	0·01	$1·5\times10^{-3}$	0·6908	0·4822
10^2	0·1	$1·5\times10^{-3}$	0·0691	0·0689	10^2	0·1	$1·5\times10^{-2}$	0·6908	0·4951
10^6	0·01	5×10^{-5}	0·0691	0·0670	10^6	0·01	5×10^{-4}	0·6908	0·4984
10^6	0·1	5×10^{-4}	0·0691	0·0697	10^6	0·1	5×10^{-3}	0·6908	0·5130

0·1, (14) is a good approximation to (8), but when Q_∞/R_0 has reached 0·5, (14) is a substantial overestimate. Fig. 1 shows Q_∞/R_0 as a function of c – the shape of the curve is similar for other values of p_0 and s.

In the case of a diploid organism, we again suppose that A, a are selectively neutral, and thus the relative fitnesses depend on the allele at the (B, b) locus. The fitnesses of BB, Bb and bb are taken as $1+s$, $1+hs$ and 1, and a similar deterministic argument to the haploid case gives

$$p_{n+1}-p_n = sp_n(1-p_n)\,(h+p_n(1-2h))/\{1+sp_n(2h+p_n(1-2h))\}, \tag{15}$$

$$Q_{n+1}-Q_n = c(1+hs)\,(1-p_n)\,(R_n-Q_n)/\{1+s(h+p_n(1-h))\}, \tag{16}$$

$$R_{n+1}-R_n = c(1+hs)\,p_n(Q_n-R_n)/\{1+shp_n\}. \tag{17}$$

This time, when Q_n-R_n is obtained by recursion from (16) and (17), a simple result like (5) does not arise, but we can, of course, compute Q_∞ to any required degree of accuracy for given values of c, h, s and p_0.

The corresponding differential equations are

$$\dot{p} = sp(1-p)\,(h+p(1-2h))/\{1+sp(2h+p(1-2h))\}, \tag{18}$$

$$\dot{Q} = c(1+hs)\,(1-p)\,(R-Q)/\{1+s(h+p(1-h))\}, \tag{19}$$

$$\dot{R} = c(1+hs)\,p(Q-R)/(1+shp). \tag{20}$$

Suppose c is so small that R is effectively constant, and that Q remains much less than R. We find

$$\frac{1}{R_0}\frac{\mathrm{d}Q}{\mathrm{d}p} \simeq \frac{c(1+hs)}{s}\,\frac{1+sp(2h+p(1-2h))}{p(h+p(1-2h))\{1+s(h+p(1-h))\}}. \tag{21}$$

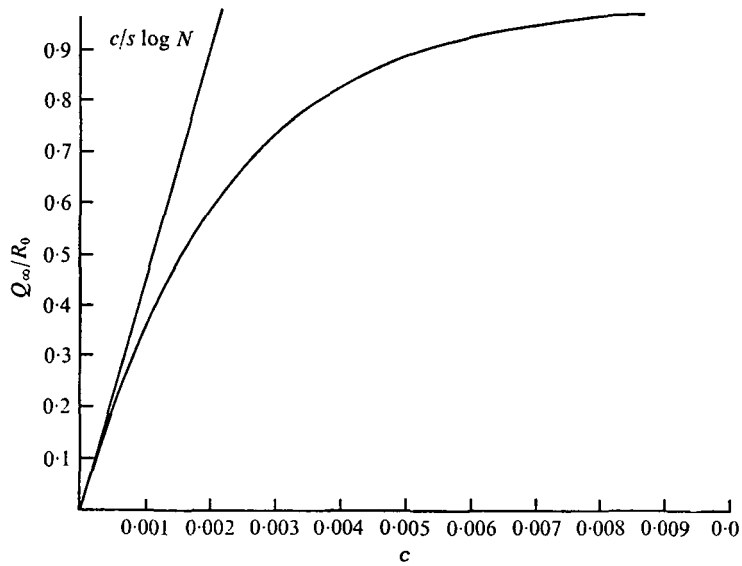

Fig. 1. Q_∞/R_0 is a measure of the proportional reduction in frequency of a neutral allele A after fixation of a linked favourable allele B. Its exact value, calculated from (8) with $N = 10^4$ and $s = 0{\cdot}02$, as a function of c, is shown here, along with the approximation (14).

Equation (21) exhibits qualitatively different behaviour in the three cases of special interest, namely $h = 0, \frac{1}{2}, 1$. For the recessive case $h = 0$ the principal term is

$$\frac{Q_\infty}{R_0} \sim \frac{c}{s}\frac{1}{p_0}. \tag{22}$$

In the additive case, $h = \frac{1}{2}$, we find

$$\frac{Q_\infty}{R_0} \sim \frac{2c}{s}\log\frac{1}{p_0}, \tag{23}$$

while in the dominant case $h = 1$ the integral diverges at $p = 1$, but if we integrate from $p = p_0$ to $p = 1 - p_0$, we again obtain

$$\frac{Q_\infty}{R_0} \sim \frac{2c}{s}\log\frac{1}{p_0}. \tag{24}$$

(In a population of size N, p_0 will usually be $1/(2N)$, so when p reaches $1 - p_0$, all individuals will have at least one B.)

For other values of h in the range $(\epsilon, 1-\epsilon)$ $(\epsilon > 0$, small), we find

$$\frac{Q_\infty}{R_0} \simeq \frac{c}{hs}\log\frac{1}{p_0}. \tag{25}$$

To get an idea of what these imply, suppose $s = 10\,\%$ and $p_0 = 10^{-6}$, and ask for what values of c will Q_∞/R_0 be less than $0{\cdot}1$, i.e. when will the polymorphic allele A be reduced to at most 10 % of its initial value? For additive or dominant inheritance, c must be no more than $3{\cdot}6 \times 10^{-4}$ and, if s or p_0 is smaller, c must be smaller still. For recessive inheritance, with the same parameters, c must be at most 10^{-8}.

Equations (22)–(25) give estimates of the extent to which the fixation of a favourable allele will reduce neutral genetic polymorphism in its neighbourhood. The effect of the favourable allele is much greater if it is wholly or partly dominant, but even for a dominant allele the effect is very local. It will later be shown that the effect can be important in aggregate.

3. THE EFFECT OF SUBSTITUTION ON AVERAGE HETEROZYGOSITY

The heterozygosity H at a locus is the probability that the locus will be heterozygous. If there are just two alleles at the locus, of frequencies R, $1-R$, then $H = 2R(1-R)$. Kimura (1964) showed that, in the absence of selection and mutation,

$$H_n = H_0 \exp[-n/(2N_e)], \tag{26}$$

where H_0 is the initial heterozygosity, H_n that after n generations, and N_e is the effective population size. Equation (26) gives the rate at which heterozygosity decays due to random drift in a finite population, and we shall compare this with the rate of decay due to hitch-hiking effects.

Let H_0 denote the initial average heterozygosity over the genome, and let H_∞ denote the mean heterozygosity after fixation. Then

$$\frac{H_\infty}{H_0} = \frac{Q_\infty(1-Q_\infty)}{R_0(1-R_0)}.$$

See Fig. 2. Assuming additive fitness, so that diploid and haploid cases are effectively the same, when $Q_\infty \ll R_0$, it follows from (23) that

$$\frac{H_\infty}{H_0} \simeq \frac{2c}{s(1-R_0)} \log \frac{1}{p_0} = \beta c, \tag{27}$$

where $\beta = \dfrac{2 \log 1/p_0}{s(1-R_0)}$ is the slope of the line in Fig. 2.

For $c = 0$, we have complete homozygosity, and the shaded area represents the amount of heterozygosity lost. The replacement of the actual curve by the line, and thus the shaded area by the triangle AOB will underestimate the effect of hitch-hiking. AOB has area $1/\beta$, so this underestimate is equivalent to making completely homozygous a total length of chromosome of $1/\beta$. Suppose that the ratio of the shaded area to AOB is α; the total equivalent length of chromosome in map units (% recombination) made homozygous is thus

$$L \simeq \frac{50\alpha s(1-R_0)}{\log(1/p_0)}. \tag{28}$$

α depends on R_0, s and p_0, but calculations for various values of these parameters all give values of α reasonably close to 2.

It may be objected that to make a length L completely homozygous is not equivalent to making some greater length less heterozygous. This would be a valid objection for a selectively maintained polymorphism, because a locus rendered completely homozygous would remain so until a further mutation, whereas if any

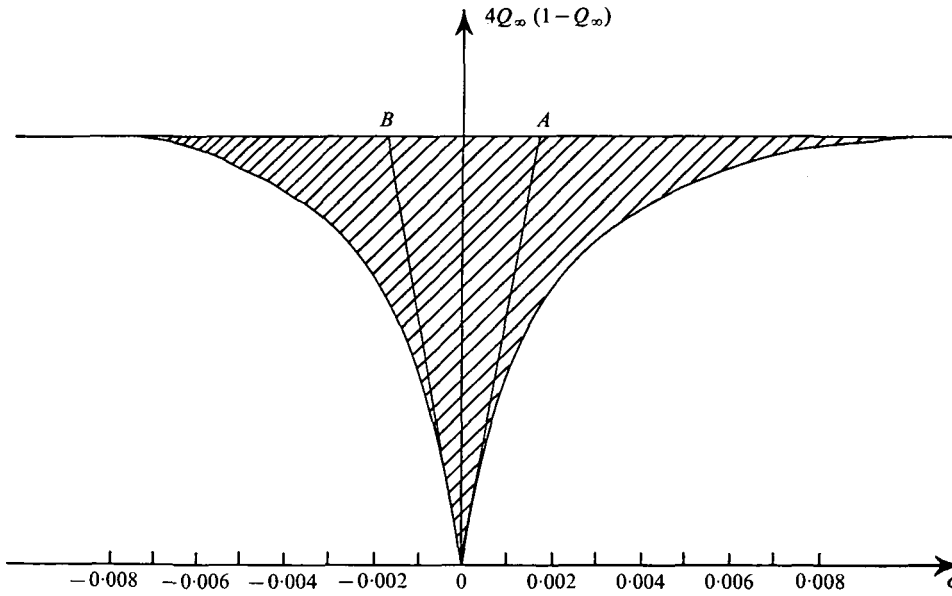

Fig. 2. $4Q_\infty(1-Q_\infty)$ is the final amount of heterozygosity at a locus, when initial frequencies of a, A are 0·5. The graph here, with $N = 10^6$ and $s = 0{\cdot}01$, is calculated from (8).

heterozygosity remained, the gene frequencies would return towards their equilibrium frequencies. But we are assuming a *neutral* polymorphism, for which (26) states that the expected heterozygosity at some future time will be a fixed fraction of what it is now.

Since, in (28), L depends on R_0, we can calculate the mean value of L given any initial distribution of frequencies at the polymorphic locus. If the initial distribution of frequencies is uniform over [0, 1], let the initial frequencies of A, a at the polymorphic locus be $1-x, x$. The probability that, when B arises it is linked to a is x, and the resulting value of L is $100sx/\log(1/p_0)$, (taking $\alpha = 2$). Hence

$$E(L) = 2\int_0^1 100sx/\log(1/p_0)\,x\,\mathrm{d}x$$
$$= \frac{200s}{3\log 1/p_0}. \tag{29}$$

(29) is the expected equivalent length of chromosome, in map units, made homozygous by the substitution of a single favourable mutation.

Thus the effect is greatest if there is a large selective advantage s per locus, and if the population is small. For a selective advantage of 0·1, and a population of size 10^6, the equivalent length of chromosome made homozygous would be 0·48 map units.

4. THE AGGREGATE EFFECT OF GENE SUBSTITUTION

To estimate the aggregate effect of gene substitution on neutral polymorphism, we need an estimate of the rate of substitution of favourable alleles. This is the most uncertain part of our investigation, with two possible lines of approach. One is to extrapolate from the known rates of evolution of well-sequenced proteins. For example, assuming that in vertebrates there are 10^5 genes evolving at the same rate as the genes for haemoglobin – i.e. approximately 8×10^6 years per amino acid substitution (Dickerson, 1971) – and taking an average of 5 years per generation, then we get $k \simeq 16$, where k is the number of generations per completed substitution. This estimate suffers from the drawbacks that we do not know how many genes there are, nor can we justify the assumption that all or most of the substitutions in haemoglobin are selectively favourable.

An alternative approach is to use the concept of the cost of natural selection (Haldane, 1957). Using this, Felsenstein (1971) has shown that in a stationary haploid population,

$$k \geqslant \log(1/p_0)/\log(1+d), \tag{30}$$

where p_0 is again the initial frequency of the favourable allele, and $1+d$ is the mean number of offspring produced by the fittest genotypes in the population. (30) also holds for a diploid with additive fitness.

There are a number of difficulties in applying (30). First, its derivation assumed that genes at different loci act independently, so that fitnesses are multiplicative; Sved (1968) and Maynard Smith (1968) have shown that if this assumption is dropped, a given intensity of selection can produce a much higher rate of gene substitution. Also (30) is an inequality, not necessarily sharp, and $1+d$ must be estimated. For man, $1+d$ can hardly exceed 5, for then some genotypes would average more than 10 surviving children in a stationary population; in a species of high fecundity, $1+d$ might be much higher, but need not be so if deaths are independent of genotype. Further, we want to know that part of $(1+d)$ which arises from the additive effects on fitness, and which therefore contributes to gene substitution. Fortunately, $\log(1+d)$ varies much more slowly than $1+d$, but we shall consider the effect of a range of values of $1+d$ in Table 2.

Taking (30) as an equality; if l is the equivalent length of chromosome (in map units) made homozygous per generation, from (29)

$$l = E(L)/k \sim \frac{200s\log(1+d)}{3(\log 1/p_0)^2}. \tag{31}$$

Thus the effect of hitch-hiking on reducing heterozygosity is greatest when the selective advantage of the substituted locus is high, the fittest genotype is much fitter than average, and, taking $p_0 = 1/(2N)$, the population is small.

To measure the importance of hitch-hiking, we will compare the expected half-life of a neutral polymorphism towards fixation by drift with the expected half-life towards fixation by hitch-hiking. The former, from (26), is $(2\log_e 2)N_e$ generations, which will be about the same quantity as N, the population size. For the hitch-hiking

half-life, suppose the total length of the genome is m map units, and that favourable mutations are distributed at random uniformly along the chromosome. Neglecting end effects, the probability that a neutral polymorphism will not be fixed after n generations is $(1-l/m)^n \simeq \exp[-ln/m]$. Thus the half-life $T \simeq m(\log 2)/l$, which from (31) gives

$$T \simeq \frac{3m(\log 2)[\log(1/p_0)]^2}{200s\log(1+d)}. \tag{32}$$

Some values of this expression are given in Table 2, taking $p_0 = 1/(2N)$. We also give the values of k, the number of generations per substitution, which correspond, via (30), to the reference value of $1+d$.

Table 2. *The half-life in generations, T, of a neutral polymorphism against fixation by hitch-hiking, for various values of the population size N and the intensity of selection $(1+d)$*

(The four values of T are for selective advantage s = 0·01, 0·02, 0·05 and 0·1 respectively. k is the number of generations per completed substitution (taking (30) as an equality). The values of T should be compared with the half-life of a neutral polymorphism against fixation by drift, which is approximately equal to N generations.)

$1+d$ ...	1·25		2		5		10	
N	T	k	T	k	T	k	T	k
10^2	65500	24	21100	8	9080	3·3	6340	2·3
	32800		10500		4540		3170	
	13100		4220		1820		1270	
	6550		2110		908		634	
10^4	229000	44	75000	14	31700	6	22200	4·3
	114000		37500		15800		11100	
	45800		15000		6330		4430	
	22900		7500		3170		2220	
10^6	491000	65	158000	21	68000	9	47500	6·3
	246000		79000		34000		23800	
	98200		31600		13600		9500	
	49100		15800		6800		4750	

Table 2 shows that for large populations, hitch-hiking may well be much more important in reducing neutral polymorphism than fixation by random drift. The uncertainties are in the use of (30) as an equality for k, and in the value of the intensity of selection, $1+d$; even for small values of $1+d$, hitch-hiking is more important than drift if the population size is 10^6 or more.

5. THE PROBABILITY OF COMPLETE FIXATION

We now consider the effect of hitch-hiking on a selectively maintained polymorphism. We assume that the selective forces maintaining the polymorphic locus (A, a) are small compared with those responsible for the substitution of b by B, for if the reverse were true, linkage would prevent the substitution until recombination had occurred, and fixation at the (A, a) locus would be impossible. As in § 2, we

suppose that when B arises it is initially linked to a. We are interested only in the probability that, when B is fixed, A will be completely eliminated; otherwise it will gradually attain its equilibrium frequency again.

Consider a diploid population of size N with the relative fitnesses of bb, bB and BB being $1:1+s/2:1+s$. In any one generation, suppose the gamete frequencies are

AB	aB	Ab	ab
0	p	$(1-p)R$	$(1-p)(1-R)$

In the next generation, the probability of an AB zygote arising is $cp(1-p)R$ (selection can be ignored here when c is small). If an AB zygote does arise by recombination, it has a probability x, say, of being established, which is the same as the probability that a new mutant arising now gets established. x is related to the form of the offspring distribution, but in particular to the selective advantage k of the new mutant. With a Poisson offspring distribution, and k small, $x \simeq 2k$, but this will often be an overestimate. To preserve flexibility, we take $x = \theta k$. Now,

$$k = \frac{\text{mean fitness of zygotes with } AB \text{ chromosome}}{\text{mean fitness of population}} - 1$$

$$= \frac{1+sp+s(1-p)/2}{(1+s)p^2+2(1+s/2)p(1-p)+(1-p)^2} - 1,$$

i.e.
$$k \simeq \frac{s(1-p)}{2(1+sp)}. \tag{33}$$

Thus the probability g that a gamete chosen at random in the next generation is an AB gamete which eventually becomes established is

$$g = cp(1-p)R\frac{\theta s(1-p)}{2(1+sp)},$$

i.e.
$$g = \frac{\theta csRp(1-p)^2}{2(1+sp)}. \tag{34}$$

Thus, with $2N$ gametes, the probability that no AB arises in this generation and eventually becomes established is $(1-g)^{2N} \sim \exp[-2Ng]$. Hence the probability that no established AB ever arises, which is the probability that A is eliminated, is

$$P_0 \simeq \exp[-2N\Sigma g], \tag{35}$$

where Σg denotes the summation of the expression (34) over the generations during which B substitutes b.

Taking R as constant during the period,

$$\log P_0 \simeq -sN\theta cR\int_0^T \frac{p(1-p)^2}{1+sp}\,\mathrm{d}t,$$

where T is the time to fixation. Now, from (10) (in which we replace s by $s/2$), $\mathrm{d}p/\mathrm{d}t = sp(1-p)/(2+sp)$. Hence

$$\log P_0 \simeq -N\theta cR\int_{p_0}^1 (1-p)\frac{(2+sp)}{1+sp}\,\mathrm{d}p.$$

Since s and p_0 are small, we find

$$P_0 \simeq \exp(-\theta NcR). \tag{36}$$

Note that (36) does not contain s; the intuitive explanation is that, as s increases, so the time to fixation decreases, so an AB chromosome has less chance of arising – but if it does arise, its chance of establishment increases, and these forces tend to cancel out. Since θ will usually be between 1 and 2, and R will be about 0·5, (36) shows that for a reasonable chance of fixation c must be not much larger than $1/N$. Thus, in a population of size 1000, there is a distance of about 0·1 map units on either side of the substituted locus at which the polymorphic locus has a reasonable chance of being fixed.

The derivation of (36) involved many approximations and assumptions, so it was thought prudent to put (36) to the test of simulation. Consider a haploid population, of fixed size N, with initial numbers and fitnesses

	AB	aB	Ab	ab
Initial number	0	1	NR	$N-NR-1$
Fitness	$1+s$	$1+s$	1	1

Then, given numbers m_1, m_2, m_3 and m_4 of these gametes in generation n, the probabilities that a randomly selected member of the next generation has each one of these genotypes can be calculated, and thus generation $(n+1)$ obtained by drawing N individuals according to these probabilities. Successive generations are then produced until either B dies out by random chance, or is eventually fixed by selection. The results of these simulations are given in Table 3.

Table 3. *A simulation test of the estimate (36) of P_0, the probability that allele A will be eliminated when a linked favourable allele B is fixed*

(For given values of selective advantage s and recombination fraction c, a population of 200 haploids, initially 0:1:100:99 of $AB:aB:Ab:ab$ was taken, and successive generations produced until either B was eliminated by chance (column 5) or fixed by selective drift (column 6). When B was fixed, the presence or absence of A was noted (columns 7, 8). The number of eliminations should be compared with the theoretical expectations in the last column. Note that (38), the probability of fixation of a favourable allele, is confirmed by columns 5 and 6.)

						When B was fixed		
s	θ	c	e^{-500c}	Times B eliminated	Times B fixed	A present	A eliminated	Expected times A eliminated
0·1	1·65	0·01	0·438	70	17	10	7	7·45
0·1	1·65	0·02	0·192	71	14	10	4	2·69
0·1	1·65	0·03	0·084	89	13	11	2	1·09
0·1	1·65	0·04	0·037	57	15	14	1	0·56
0·1	1·65	0·05	0·016	65	13	13	0	0·21
0·2	1·39	0·01	0·499	57	26	11	15	12·97
0·2	1·39	0·02	0·249	46	24	20	4	5·98
0·2	1·39	0·03	0·125	59	25	21	4	3·13
0·2	1·39	0·04	0·062	44	25	22	3	1·55
0·2	1·39	0·05	0·031	66	27	26	1	0·81
0·3	1·18	0·01	0·554	64	35	19	16	19·39
0·3	1·18	0·02	0·307	53	37	28	9	11·36
0·3	1·18	0·03	0·170	65	33	29	4	5·61
0·3	1·18	0·04	0·094	65	39	36	3	3·67
0·3	1·18	0·05	0·052	42	37	35	2	1·92

The difference between a haploid and an additive diploid model is only in the population size, so (36) becomes

$$P_0 \simeq \exp[-\theta NcR/2]. \tag{37}$$

The method described above of obtaining successive generations corresponds to assuming a Poisson offspring distribution, and, in order to obtain a reasonable number of fixations in a relatively short computing time, s was taken to be between 0·1 and 0·3. Thus, rather than take $\theta = 2$, we must use standard branching process methods (see, for example, Feller (1966), p. 296), which give the probability of fixation x as the relevant solution of

$$1-x = \exp[-(1+s)x].$$

This solution is

$$x \simeq \frac{2s}{(1+s)^2}, \tag{38}$$

so, for $s = 0{\cdot}1$, we take $\theta = 1{\cdot}65$. In each simulation, $N = 200$, $R = 0{\cdot}5$, so

$$P_0 \simeq \exp[-50\theta c].$$

A comparison of the last two columns of Table 3 by a standard χ^2 goodness-of-fit test confirms the impression of agreement of theory and simulation. Thus (36) is validated.

6. DISCUSSION

It is clear from Table 2 that in a large population the hitch-hiking effect is likely to be more important than random drift in determining the level of heterozygosity for neutral alleles; the effect is overriding in populations of size 10^6, or more, and may be predominant in populations as small as 10^4 individuals.

This conclusion is important for the following reason. It is gradually emerging (Lewontin, 1973) that the extent of enzyme polymorphism is surprisingly constant between species. If the polymorphism is neutral, this is very difficult to explain. Thus, at equilibrium between mutation and random fixation, the mean heterozygosity

$$H = 4N_e u/(1+4N_e u), \tag{39}$$

where N_e is the effective population size and u the neutral mutation rate. If H lies between 0·1 and 0·5, then N_e lies between $0{\cdot}028u^{-1}$ and $0{\cdot}25u^{-1}$, and it is not plausible that the effective population sizes of all species lie within such narrow limits.

It can be objected that species have not had time to reach their equilibrium values, but we know that H will be some function ϕ of past numbers. The exact form of ϕ is not known, but its value is likely to be as sensitive to changes in N_e as is the equilibrium formula (39). (See, for example, Haigh & Maynard Smith (1972) for the effect of bottleneck on numbers.) Thus it seems that the uniformity of H between species is powerful evidence against the view that the observed polymorphisms are in the main selectively neutral.

The investigation in §§ 2–4 can therefore be regarded as a last ditch attempt to save the neutral mutation theory by showing that there is another process which can account for the uniformity of H between species. The attempt has been partially

successful. If the only force opposing neutral mutation were random fixation, a change in population size from 10^4 to 10^8 would alter the half-life of a neutral polymorphism by a factor of 10^4; if we allow for the hitch-hiking effect, the half-life will change only by a factor of about 4.

The effect of hitch-hiking on selectively maintained polymorphisms is very local. The region of chromosome (in map units) which will be made homozygous by a single substitution is of order N^{-1}. In a large population, say 10^6 individuals, a selectively maintained polymorphism could survive a substitution in the same cistron, although one or other of the alleles would usually be reduced to a low frequency.

We thank Professor R. C. Lewontin for provoking us into solving this problem, and the Science Research Council for financial support.

REFERENCES

Dickerson, R. E. (1971). The structure of cytochrome *c* and the rates of molecular evolution. *Journal of Molecular Evolution* **1**, 26–45.

Ewens, W. J. (1969). *Population Genetics*. London: Methuen.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, 3rd ed., vol. 1. Wiley.

Felsenstein, J. (1971). On the biological significance of the cost of gene substitutions. *American Naturalist* **105**, 1–11.

Haigh, J. & Maynard Smith, J. (1972). Population size and protein variation in man. *Genetical Research* **19**, 73–89.

Haldane, J. B. S. (1957). The cost of natural selection. *Journal of Genetics* **55**, 511–522.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232.

Kojima, K. I. & Scheffer, H. E. (1967). Survival process of linked mutant genes. *Evolution* **21**, 518–531.

Lewontin, R. C. (1973). *The Genetic Basis of Evolutionary Change*. (In the Press.)

Maynard Smith, J. (1968). 'Haldane's dilemma' and the rate of evolution. *Nature* **219**, 1114–1116.

Sved, J. A. (1968). Possible rates of gene substitution in evolution. *American Naturalist* **102**, 283–292.