Genet. Res., Camb. (1994), **63**, *pp.* 63–78 *With 9 text-figures*

Error thresholds and stationary mutant distributions in multi-locus diploid genetics models

PAUL G. HIGGS
University of Sheffield, Department of Physics, Hounsfield Road, Sheffield S3 7RH, U.K.

(*Received 2 September 1993 and in revised form 18 October 1993*)

Summary

We study multi-locus models for the accumulation of disadvantagenous mutant alleles in diploid populations. The theory used is closely related to the quasi-species theory of molecular evolution. The stationary mutant distribution may either be localized close to a peak in the fitness landscape or delocalized throughout sequence space. In some cases there is a sharp transition between these two cases known as an error threshold. We study a multiplicative fitness landscape where the fitness of an individual with j homozygous mutant loci and k heterozygous loci is $w_{jk} = (1-s)^j (1-hs)^k$. For a sexual population in this landscape there are two types of solution separated by an error threshold. For a parthenogenetic population there may be three types of solution and two error thresholds for some values of h. For a population reproducing by selfing the solution is independent of h, since the frequency of heterozygous individuals is negligible. The mean fitnesses of the populations depend on the reproductive method even for the multiplicative landscape. The sexual may have a higher or lower fitness than the parthenogen, depending on the values of h and u/s. Selfing leads to a higher mean fitness than either sexual reproduction or parthenogenesis. We also study a fitness landscape with epistatic interactions with $w_{jk} = \exp(-s(2j+k)^\alpha)$. The sexual population has a higher fitness than the parthenogen when $\alpha > 1$. This confirms previous theories that sexual reproduction is advantageous in cases of synergistic epistasis. The mean fitness of a selfing population was found to be higher than both the sexual and the parthenogen over the range of parameter values studied. We discuss these results in relation to the theory of the evolution of sex. The fitness of the stationary distribution in cases where unfavourable mutations accumulation is one factor which could explain the observed prevalence of sexual reproduction in natural populations, although other factors may be more important in many cases.

1. Introduction

Many problems in population genetics may be viewed as a competition between selection and mutation. Selection acts to increase the numbers of individuals with the fittest genes, and mutation continually produces new genes which are often of lower fitness. In many cases a balance arises between these effects leading to a stationary distribution of disadvantageous mutations. If a mutant gene has a large disadvantageous effect then selection will act strongly against it, and its frequency in the population will remain low. If a mutant gene is only slightly disadvantageous then selection is less effective against it, and it may occur with a significant frequency. Kimura (1983) has argued that many naturally occurring mutations are close to being neutral in effect. Gillespie (1991) has also discussed the evidence for and against the neutral theory. If slightly disadvantageous mutations occur at many places on the genome the overall effect on the fitness of the individual may be severe, even if each mutation alone is almost neutral. In this article we will consider the balance of selection and mutation for multi-locus models with slightly disadvantageous mutations.

Theoretical biologists have usually studied diploid models with one or two loci (Wright, 1969; Crow & Kimura, 1970). Such mocels can easily be solved in a wide variety of cases. Generalization to multi-locus models is possible for some simple cases, and often requires the use of numerical methods (Kimura & Maruyama, 1966; Haigh, 1978; Kondrashov, 1982; Charlesworth, 1990). Another approach is the 'molecular quasi-species' theory (Eigen *et al.* 1989; Swetina & Schuster, 1982), which is intended to describe populations of self-replicating macromolecules. In the

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

language of population genetics these would be haploid multi-locus models. It is the object of this article to show how ideas from the quasi-species theory are relevant to diploid multi-locus models.

In the quasi-species theory for molecular evolution each macromolecular sequence has a 'fitness' associated with it which represents its rate of replication. In the absence of mutation the sequence with the highest replication rate would grow to dominate the population, and all surviving sequences would become identical. If mutation occurs then new sequences are continually being formed which are not of optimal fitness. In this case the concentrations of the different sequences converge to a stationary distribution in which sequences which differ from the optimal sequence are present in non-negligible amounts. This stationary distribution has been termed the quasi-species.

The stationary distribution which arises will depend on the shape of the fitness peak, i.e. on the way in which fitness decreases as successive mutations are made to the optimal sequence. The stationary distribution can be obtained analytically for simple choices of fitness landscape. In general the distribution becomes broader as the mutation rate is increased. For large mutation rates selection becomes ineffective, and the stationary distribution is spread over the whole of sequence space. In some cases there is a well defined transition between a localized and a delocalized distribution as the mutation rate is increased. This transition is called the error threshold, and has analogies with phase transitions in statistical physics (Leuthäuser, 1986; Tarazona, 1992).

In this article we will ask how the stationary mutant distribution depends on the method of reproduction, and on the shape of the fitness peak. The plan of the article is as follows. Section 2 discusses haploid models and points out the relationship to the quasi-species theory of Eigen *et al.* (1989). Section 3 presents new results on multi-locus diploid models. The stationary mutant distributions are obtained for sexual, parthenogenetic and selfing species, in the simplest case where the effect of the different loci is multiplicative. Section 4 looks at the effect of epistatic interactions which alter the shape of the fitness peak, and hence of the stationary distribution. Analytical solutions can be obtained in certain cases, and these are compared to previous approximate solutions and to numerical results. Section 5 discusses the relative fitness of sexual and asexual populations in different landscapes, with reference to the widespread recent debate on the evolution and costs of sex.

2. Haploid Models

The simplest possible case is a single locus haploid model. We consider one single gene in each individual, which may either exist as an optimal fitness allele with relative fitness $w_0 = 1$ or as a mutant allele having a lower relative fitness $w_1 = 1-s$. Usually s is called the selection coefficient (Crow & Kimura, 1970). Let x be the frequency of the optimal allele, and $y = 1-x$ be the combined frequency of the mutant alleles. Let there be a probability u per generation that the optimal allele mutates into one of the other alleles. Subsequent mutations occurring on already mutant genes are very unlikely to recreate the optimal allele, thus we may neglect 'back mutations'. If the allele frequencies at generation t are $x(t)$ and $y(t)$ then at generation $t+1$ we have $x(t+1) = (1-u)\,x(t)/(x(t)+(1-s)\,y(t))$. Upon repeated iteration of this equation $x(t)$ will converge to $x = 1-u/s$, if $u \leqslant s$, or to $x = 0$, if $u > s$. Although this argument is trivial it contains the essence of the error threshold idea. In cases where the back mutation rate may be neglected selection can only counteract the effects of mutation if the mutation rate is small enough. In this case the error threshold is at $u = s$. If $u > s$ the optimal allele disappears.

Now consider a haploid genome sequence of L loci. At each locus there may either be an optimal allele or a mutant one. Let w_k be the fitness of a genome sequence containing k mutant alleles. It will be assumed that fitness simply depends on the number of mutant genes and not on their precise positions in the sequence. Hence the fitness landscape is a single peak with maximal fitness $w_0 = 1$. At generation t let $C_k(t)$ be the combined concentration of all sequences having k mutations. The C_k are normalized so that the sum of the concentrations is equal to 1. In the subsequent generation

$$C_j(t+1) = \frac{1}{W}\sum_k M_{jk}\, w_k\, C_k(t), \tag{2.1}$$

where W is the constant needed to maintain the normalization condition. In fact, W is the mean fitness of the popultion.

$$W = \sum_k w_k\, C_k(t). \tag{2.2}$$

In equation (2.1) the mutation matrix M_{jk} is the probability that a sequence with k mutant genes mutates into a sequence with j mutant genes in one generation. Since we are neglecting back mutations $j \geqslant k$.

$$M_{jk} = \binom{L-k}{j-k} u^{j-k}(1-u)^{L-j} \quad (k \leqslant j \leqslant L). \tag{2.3}$$

A simplification of the mutation matrix is possible if $L \gg 1$, and the mutation rate per locus $u \ll 1$. In this case the total mutation rate per genome is $U = uL$, and the binomial distribution of equation (2.3) becomes a Poisson distribution.

$$M_{jk} = \exp(-U)\frac{U^{j-k}}{(j-k)!} \quad (k \leqslant j). \tag{2.4}$$

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

It is possible to write a mutation matrix which includes back mutations at a finite rate. This would be necessary in a binary sequence space, such as the one considered in Eigen (1989), or in a DNA sequence space with four possible nucleotides at each site. However, in the limit $L \gg 1$ we still end up with equation (2.4), which is independent of the back mutation rate. This is because of combinatorial factors which occur for large L. Thus we are always entitled to neglect back mutations for long sequences.

If we are interested in a numerical solution then we have merely to iterate equation (2.1). The fitnesses w_k may be set according to any model which we choose, and starting from any initial choice of concentration distribution, the concentrations will eventually converge to the stationary distribution for the given w_k. We will now look at some particular choices of w_k for which the stationary distribution may be obtained analytically, at least approximately.

One landscape which has been widely studied is the 'master sequence' landscape or 'isolated peak' landscape (Swetina & Schuster, 1982). In this case the optimal sequence (called the master sequence) has a high fitness ($w_0 = 1$) and all the other sequences have a lower fitness which is the same for all of them ($w_k = 1-s$ for $k \neq 0$). If $L \gg 1$ the stationary distribution C_k satisfies

$$C_j = \frac{e^{-U}}{W} \sum_{k=0}^{j} \frac{U^{j-k}}{(j-k)!} w_k C_k. \tag{2.5}$$

Setting $j = 0$ gives straight away that $W = e^{-U}$. In fact this result is independent of the choice of the w_k, hence for a haploid asexual population the mean fitness is independent of the shape of the fitness landscape. This result is at first sight surprising, but has been known for some time (Kimura & Maruyama, 1966; Kondrashov, 1982). If the selection strength is increased the stationary distribution will become more strongly concentrated about the fitness peak, but the mean fitness is not affected.

Setting $j = 1$ gives $C_1 = (U/s) C_0$. We may obtain an approximate solution for higher j if we suppose that $s \ll 1$ and $U \ll 1$, and work to first order in these small parameters. In this case

$$C_j = \left(\frac{U}{s}\right)^j C_0 \tag{2.6}$$

and using the normalization condition we have

$$C_0 = 1 - U/s \quad (U \leqslant s). \tag{2.7}$$

This is the same as the single locus result except that u has been replaced by the whole genome mutation rate U. The error threshold is thus at $U = s$ in the master sequence landscape.

Another simple landscape is the multiplicative fitness landscape. Each mutant gene reduces the fitness of the sequence by a factor of $1-s$ independently of the others, so that $w_k = (1-s)^k$. We will use the form of M_{jk} for finite L.

$$C_j = \frac{1}{W} \sum_{k=0}^{j} \binom{L-k}{j-k} u^{j-k} (1-u)^{L-j} (1-s)^k C_k. \tag{2.8}$$

As before, if we work to first order in s and in u then we obtain

$$W = (1-u)^L, \tag{2.9}$$

$$C_j = \binom{L}{j} (u/s)^j (1-u/s)^{L-j}. \tag{2.10}$$

This result means that the loci are behaving independently of each other in the multiplicative landscape. There is a fraction $x = 1-u/s$ of optimal alleles and a fraction $y = u/s$ of mutant alleles at each locus just as in the single locus problem. The mutant genes are distributed randomly along the sequence giving rise directly to the binomial distribution in (2.10). The error threshold thus remains at its value for the single locus model: $u = s$. In fact, if L is large then the threshold has little significance since the fraction of the optimal sequence is $C_0 = (1-u/s)^L$, which becomes very small even for u much less than s. If we take the limit $L \gg 1$ then U becomes the appropriate variable instead of u, and (2.10) becomes

$$C_j = e^{-U/s} \frac{(U/s)^j}{j!}. \tag{2.11}$$

Thus the error threshold disappears altogether in this limit and we simply have a gradual delocalization of the population from the fitness peak as U is increased. This solution has previously been given by Haigh (1978).

3. Diploid models with multiplicative fitness landscapes

In diploid models each locus may either be homozygous for the optimal gene or heterozygous optimal/mutant, or may have two mutant genes. We will usually call this latter state 'homozygous' for mutant genes, although it should be remembered that the two mutants will usually be different since there are very many possible mutant alleles. The concentration of individuals with j loci which are homozygous for mutant genes, and k loci which are heterozygous optimal/mutant will be denoted C_{jk}. If there are L loci then the number of loci which are homozygous for optimal genes is $L-j-k$. The concentrations are normalized so that their sum is equal to 1.

In this section we consider models where the fitness contributions from different loci are multiplicative. Each homozygous mutation contributes a factor $1-s$ to the fitness, and each heterozygous mutation contributes a factor $1-hs$, where h is known as the

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

dominance coefficient. The fitness of individuals of type jk is

$$w_{jk} = (1-s)^j (1-hs)^k. \tag{3.1}$$

This landscape has been studied for both finite and infinite population sizes (Charlesworth *et al.* 1992, 1993). Here we will give a general solution for all values of the number of loci L. We consider three types of breeding system.

(i) *Parthenogen*

By a parthenogen we mean a diploid organism reproducing asexually by straightforward copying of the whole of its diploid genome. More specifically, this is known as apomixis (Maynard Smith, 1978). The offspring are therefore identical to the parent apart from any new mutations which have occurred. By analogy with equation (2.5) we may write

$$C_{jk}(t+1) = \frac{1}{W} \sum_{n=0}^{L} \sum_{m=0}^{L-n} M_{jknm}\, w_{nm}\, C_{nm}(t), \tag{3.2}$$

where

$$W = \sum_{n=0}^{L} \sum_{m=0}^{L-n} w_{nm}\, C_{nm}(t), \tag{3.3}$$

and M_{jknm} is the probability that an individual of type nm mutates to an individual of type jk.

$$M_{jknm} = u^{j-n}(1-u)^{m-j+n}(2u)^{j-n+k-m}(1-2u)^{L-j-k} \binom{m}{j-n}\binom{L-n-m}{j-n+k-m}. \tag{3.4}$$

To arrive at j homozygous mutations it is necessary to make $j-n$ mutations from among the m sites which were initially heterozygous. To make up the number of heterozygous sites to k requires $j-n+k-m$ mutations from among the $L-n-m$ sites which were originally homozygous for the optimal allele. These mutations occur at rate $2u$ since there are two possible genes at which mutation may occur. It has been assumed that simultaneous mutations of the two genes at the same locus do not occur. In other words if more than one mutation occurs in a generation then they will occur at different loci. This is entirely reasonable if the number of loci is large. To include the possibility of simultaneous mutations at a single locus would require the introduction of an extra summation variable in (3.4) and would complicate the subsequent analysis. Also we are interested principally in the limit $u \ll 1$ and $s \ll 1$ as before. The additional terms would be of higher order in u and would not contribute to the answer in this limit.

Numerical iteration of equation (3.2) will lead to the stationary distribution for any choice of w_{jk}. For the multiplicative landscape it is possible to find an analytical solution. One way to do this is to assume a solution of the form

$$C_{jk} = \frac{L!}{j!k!(L-j-k)!} a^j b^k c^{L-j-k}. \tag{3.5}$$

Here a is the fraction of loci which are homozygous for mutations, b is the fraction of heterozygous loci, and $c = 1-a-b$ is the fraction of homozygous optimal loci. The factorials are simply the combinatorial factor appropriate to random distribution of mutations throughout the sequence. Substitution of this into (3.2) with the multiplicative landscape (3.1) leads to a solution for a, b and c. In the usual limit of $u \ll 1$ and $s \ll 1$ everything becomes a function of u/s as before and we find

$$a = \frac{2(u/s)^2}{h+(1-2h)u/s}, \quad b = \frac{2u/s(1-2u/s)}{h+(1-2h)u/s},$$
$$c = \frac{(h-u/s)(1-2u/s)}{h+(1-2h)u/s}. \tag{3.6}$$

The fact that a consistent solution can be found by the substitution procedure shows that (3.5) was the correct form of solution to begin with. The fractions a, b and c do not depend on L, i.e. the fraction of mutations at each site in the L loci model is the same as for a single locus model ($L = 1$). Thus we have proved that the different loci are acting independently. Recall that the same thing happened for the haploid model in equation (2.10). This is a special property of the multiplicative landscape, and it is not true for the epistatic landscapes considered below.

It is worth examining this solution in more detail, since it has several features which can be interpreted in terms of the error threshold idea. If $u/s < h$ and $u/s < \frac{1}{2}$ then a, b and c are all non-zero. We will call this range of the parameter space phase *I*. In phase *I* all the different genotypes occur with a non-zero probability. As the mutation rate is increased the fraction of homozygous optimal loci c first becomes zero when $u/s = \frac{1}{2}$ or when $u/s = h$, depending on whether h is greater than or less than $\frac{1}{2}$. If $h \geqslant \frac{1}{2}$ then there is an error threshold at $u/s = \frac{1}{2}$. For $u/s > \frac{1}{2}$ we have $a = 1$, $C_{L0} = 1$, and all other concentrations are zero. We will call this phase *II*. In phase *II* only mutant genes remain. If $h < \frac{1}{2}$, on the other hand, c becomes equal to zero when $u/s = h$, and at this point b is still non-zero. Thus for $u/s > h$, another form of trial solution is appropriate (phase *III*).

$$C_{j,L-j} = \frac{L!}{j!(L-j)!} a^j b^{L-j}. \tag{3.7}$$

It is found by substitution into (3.2) that $a = (u/s)/(1-h)$, and $b = 1-a$. In phase *III* there is at least one mutant gene at every locus, i.e. $C_{jk} = 0$ if $k \neq L-j$, but there is still a non-zero fraction of the optimal alleles. The highest fitness genotype which remains is heterozygous at all loci. This solution

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

applies in the range $h < u/s < 1-h$, since when $u/s = 1-h$ we have $a = 1$, and we enter phase *II*. There are thus two error thresholds in this model when $h < \frac{1}{2}$. All the transitions are continuous (second order). The fraction of optimal alleles is $b/2+c$. This is shown as a function of u/s in figure 1 for various values of h.

The mean fitness of the population is

$$\begin{aligned} W &= (1-u)^{2L} \approx (1-2u)^L, && \text{phase } I, \\ W &= (1-s)^L, && \text{phase } II, \\ W &= (1-u)^L(1-hs)^L, && \text{phase } III. \end{aligned} \tag{3.8}$$

The result for phase *I* is independent of the fitness landscape for the same reasons as in the haploid case. If $j = k = 0$ in (3.2) and $C_{00} \neq 0$ then we have immediately that $W = M_{0000}$. A similar argument in phase *III* shows that W is only dependent on the totally heterozygous fitness w_{0L} and not on the rest of the landscape.

(ii) *Sexual*

We will now consider a sexual population with random mating, and free recombination between loci. In addition to the concentrations C_{jk} defined before, it will be necessary to calculate the gamete concentrations X_n, defined as the fraction of haploid gametes having n mutant and $L-n$ optimal genes. It is necessary to account for three processes occurring in the reproductive cycle: the random fusion of gametes to give a new diploid population, the production of new gametes by individuals according to their fitnesses, and the possibility of mutations occurring. The general equations for the stationary distribution are

$$C_{jk} = \sum_n \sum_m X_n X_m P(jk;nm), \tag{3.9}$$

$$X'_m = \frac{1}{W} \sum_j \sum_k C_{jk} w_{jk} g(m;jk) \tag{3.10}$$

$$X_n \sum_m M_{nm} X'_m. \tag{3.11}$$

Equation (3.9) represents the random fusion of gametes. $P(jk;nm)$ is the probability that two gametes with n and m randomly positioned mutations fuse to give an individual of type jk. Combinatorial arguments give

$$P(jk;nm) = \frac{\binom{n}{j}\binom{L-n}{m-j}}{\binom{L}{m}} = \frac{\binom{n}{j}\binom{L-n}{k+j-n}}{\binom{L}{k+2j-n}}. \tag{3.12}$$

Note that only three of the four variables $j\,k\,n$ and m are independent, since they are related by $k = n+m-2j$, hence the two alternative forms given in (3.12). In equation (3.10) X'_m is the fraction of gametes of type m which would be formed if there were no mutation, whilst X_n in (3.11) is the gamete concentration after taking account of mutations. In (3.10) $g(m;jk)$ is the probability that a gamete produced by a type jk individual will be of type m.

$$g(m;jk) = \binom{k}{m-j}\frac{1}{2^k} \quad (j \leqslant m \leqslant j+k). \tag{3.13}$$

The mutation matrix M_{nm} in (3.11) is the usual haploid mutation matrix of equation (2.3). It is simpler to take account of mutation after the gamete production in equations (3.10) and (3.11). This does not mean that events happened in that order. The mutation could have occurred in the line of cells which eventually divided to form the gamete. This does not change the equations. It would be possible to combine (3.9), (3.10) and (3.11) into one single equation for C_{jk}. We have not done this since the result is very unwieldy and the meaning of the different terms is less easy to see.

Let us assume that there is a fraction x of optimal alleles and a fraction $y = 1-x$ of mutant alleles at each locus, and that the concentrations at each locus are independent (or in other words linkage disequilibrium is very weak). This implies that

$$X_n = \binom{L}{n} y^n x^{L-n}. \tag{3.14}$$

If a solution of this form can be found which satisfies the equations then this justifies the assumption of independent loci, and shows that linkage disequilibrium can be neglected in the stationary state. Substituting into (3.9) gives the diploid concentrations.

$$C_{jk} = \frac{L!}{j!k!(L-j-k)!} y^{2j}(2xy)^k x^{2(L-j-k)}. \tag{3.15}$$

This is of exactly the same form as (3.5) with $a = y^2$, $b = 2xy, c = x^2$. These proportions are just the Hardy–Weinberg equilibrium values in single locus models. If we now substitute this into (3.10) and (3.11) we find that the X_n which emerges is indeed of the form (3.14) provided y satisfies equation

$$y^2(1-2h) + hy - u/s = 0. \tag{3.16}$$

It has again been assumed that $u \ll 1$ and $s \ll 1$. The general solution is of course

$$y = \frac{-h + \sqrt{[h^2 + 4(1-2h)\,u/s]}}{2(1-2h)} \tag{3.17}$$

and $y = 2u/s$ in the particular case of $h = \frac{1}{2}$. The error threshold occurs when $y = 1$, and from (3.16) it can be seen that this will occur when $u/s = 1-h$. However, if $h > \frac{1}{2}$ there is potentially a problem with (3.17) since the term in the square root becomes negative if $u/s > h^2/4(2h-1)$. As long as $h < \frac{2}{3}$ then the error threshold occurs before we reach this point, in other words x decreases smoothly with increasing u/s and goes to zero at $u/s = 1-h$. On the other hand if $h > \frac{2}{3}$ then a problem occurs at $u/s = h^2/4(2h-1)$. In fact x goes discontinuously to zero at this value. There is always a trivial solution with $x = 0$, and this must be the

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

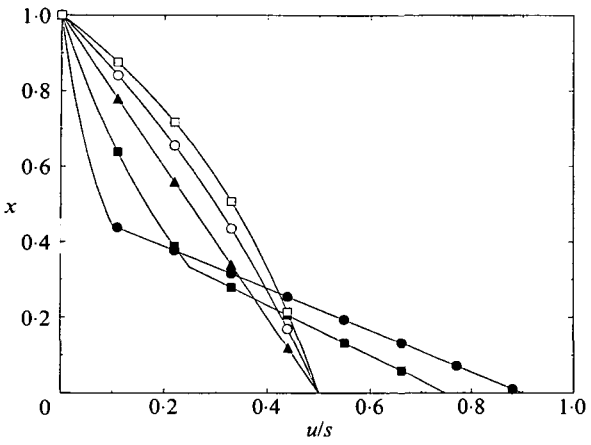

Fig. 1. The fraction x of optimal alleles as a function of u/s for the parthenogenetic population in the single locus problem (or equivalently in the multi-locus problem with multiplicative fitnesses). The curves for $h < \frac{1}{2}$ show a kink at the first error threshold $u/s = h$, and decrease to zero at the second threshold, $u/s = 1-h$. The curves for $h \geqslant \frac{1}{2}$ have a single threshold at $u/s = \frac{1}{2}$: (●), $h = 0{\cdot}1$; (■), $h = 0{\cdot}25$; (▲), $h = 0{\cdot}5$; (○), $h = 0{\cdot}75$; (□), $h = 1{\cdot}0$.

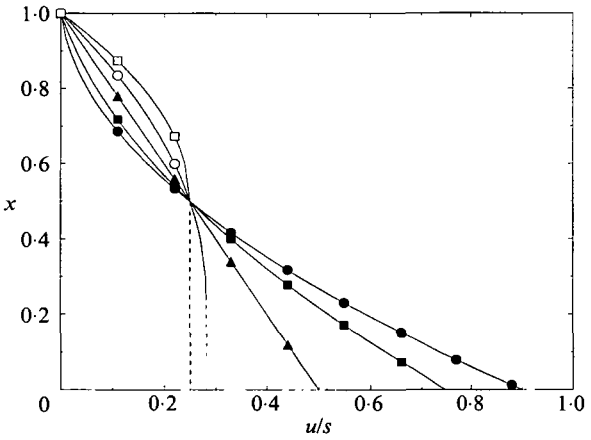

Fig. 2. As Fig. 1, but for a sexual population. The curves for $h < \frac{2}{3}$ decrease continuously to zero at $u/s = 1-h$. The curves for $h > \frac{2}{3}$ jump discontinuously to zero at $u/s = h^2/4(2h-1)$. This is shown by a dotted line. Symbols as in Fig. 1.

correct solution when (3.17) becomes a complex number. Thus there is a discontinuous error threshold (first order phase transition) in this model if $h > \frac{2}{3}$. There is a clear difference between the sexual and parthenogenetic cases. Figure 2 shows the fraction x of optimal alleles as a function of u/s in the sexual population, which may be compared with Fig. 1 for the parthenogen.

The mean fitness of the population is $(1-u-hsy)^L$ which is to be compared with $(1-2u)^L$ for the parthenogen. Thus sexual reproduction may lead either to an advantage or a disadvantage in terms of the mean fitness, according to the values of h and u/s. The parthenogenetic population has a higher mean fitness if $u < hsy$, which is true if $h > \frac{1}{2}$ and $u/s < \frac{1}{2}$. The sexual population has higher mean fitness if $h < \frac{1}{2}$ and $u/s < 1-h$. If $u/s > \frac{1}{2}$ and $h > 1-u/s$ then both

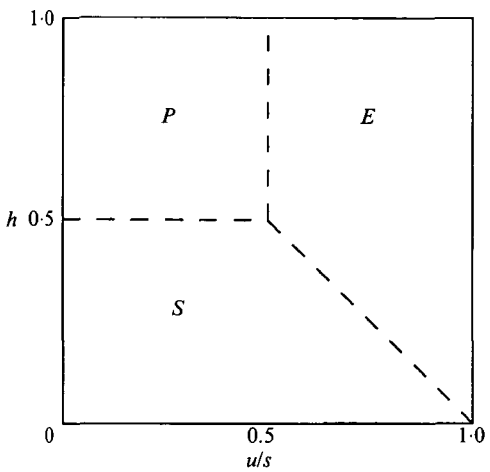

Fig. 3. Regions of the parameter space h versus u/s in which the parthenogen has higher mean fitness (P) or the sexual has higher mean fitness (S). In the region E both populations have passed the error threshold and therefore have equal fitnesses. Fitnesses are also equal along the line $h = \frac{1}{2}$ and when $u/s \ll 1$. This diagram applies for the single locus problem and the multi-locus problem with multiplicative fitnesses.

populations have passed the error threshold and therefore have equal fitness $W = (1-s)^L$. The ranges of relative advantage and disadvantage for the sexual and parthenogen are shown in Fig. 3. The relative difference in these two fitnesses may be considerable if L is large and u is of similar magnitude to s.

Although the results for sexual and parthenogen are in general different, there are certain special cases when they become equal. If $h = \frac{1}{2}$ then the concentrations and the mean fitnesses are identical for the two cases. Also in the limit $u \ll s$ then $y \approx u/sh$ from equation (3.16), and therefore $W \approx (1-2u)^L$. The mean fitnesses are thus identical in this limit.

The solution for y in the multiplicative landscape is the same as in the standard treatment of single locus models. Equation (3.16) appears in Crow & Kimura (1970) (section 6.2) and there is an equivalent treatment in Wright (1969) (chapter 3). They were both mainly interested in the limit $u \ll s$. In taking this limit all the interesting behaviour of the model is lost. Also it leads to the conclusion that mean fitness is independent of the reproduction method, and we stress that this is not true even for single locus models, although the relative difference of the two fitnesses is extremely small for $L = 1$ (it is of order u).

Figure 4 summarizes the phase behaviour for the parthenogen, and for the sexual population. The two cases are completely different. There is no equivalent of phase *III* for the sexual case. Figures 5 and 6 show the shape of the stationary mutant distribution for the multiplicative landscape with $L = 12$, and $h = 0{\cdot}25$ at several different values of u/s. The horizontal axis shows the numnber j of homozygous mutations and the vertical axis shows the number k of heterozygous mutations. The grey-scale indicates the concentration, with darker shades indicating higher concentrations.

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

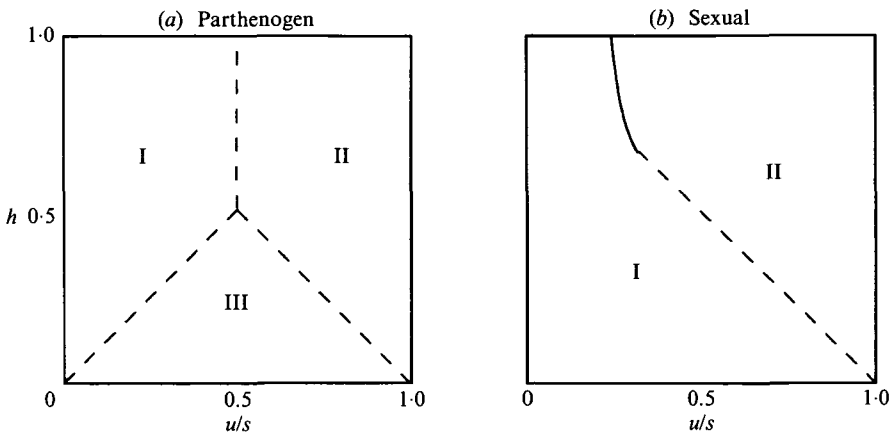

Fig. 4. Phase diagrams for the multi-locus problem with multiplicative fitnesses for (*a*) parthenogenic and (*b*) sexual populations. In phase *I* all the genotypes are present with non-zero concentration. In phase *II* only the completely homozygous mutant genotype is present. In phase *III* (which occurs only for the parthenogen) there is at least one mutant gene at every locus, and the best surviving genotype is heterozygous at every locus. Dotted lines indicate continuous (second order) transitions, and the solid line in (*b*) is a discontinuous (first order) transition.

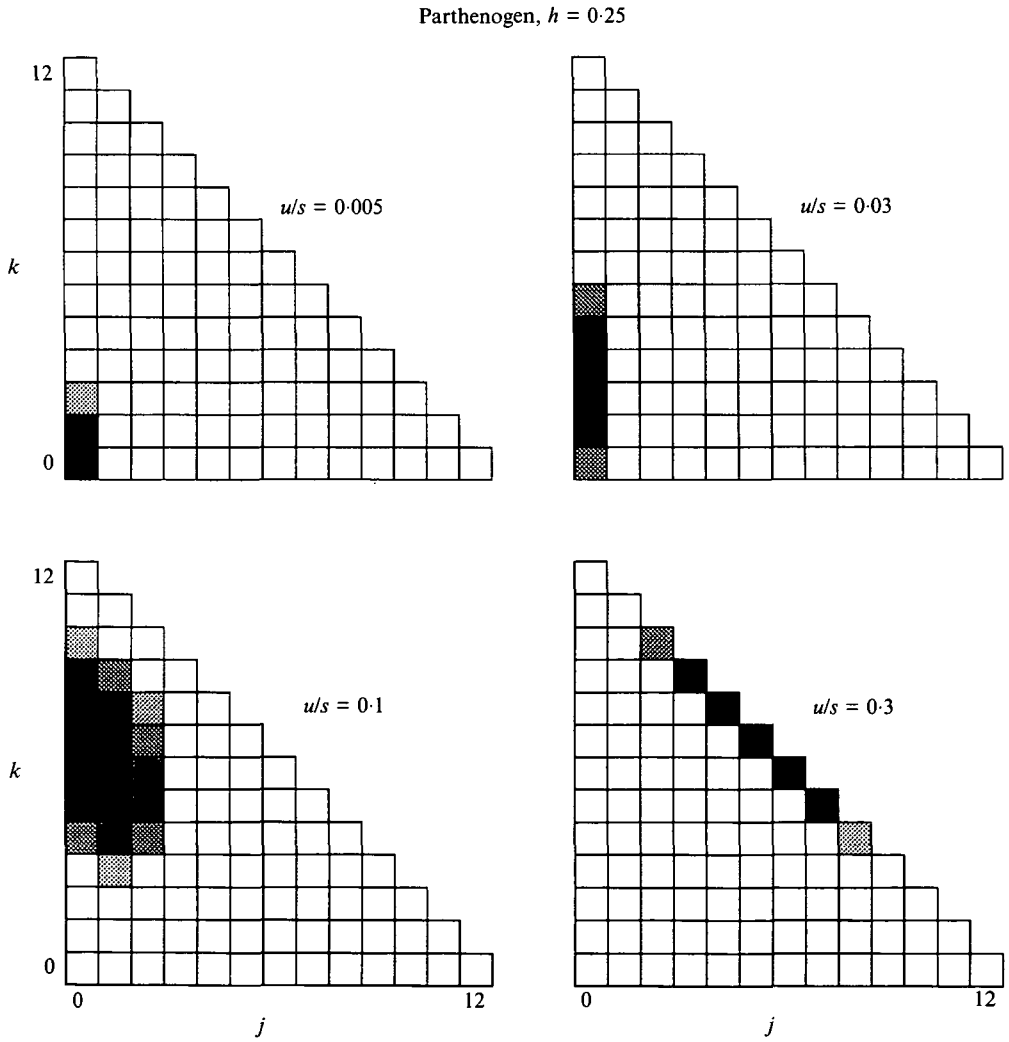

Fig. 5. Representation of the stationary quasi-species distribution for the parthenogen with 12 loci and multiplicative fitnesses. Four different values of u/s are shown with $h = 0{\cdot}25$. The number of homozygous mutations j is shown on the horizontal axis, and the number of heterozygous mutations k is shown on the vertical axis. The element of the C_{jk} matrix which is largest has been coloured black, and other elements of the matrix have been coloured with one of 10 different shades of grey according to their concentration relative to the element with the highest concentration. For small u/s the population is localized close to the origin. As u/s increases the number of heterozygous mutations increases and the quasispecies spreads vertically on the diagram. When $u/s = 0{\cdot}3$ the population is in phase *III* and the distribution lies entirely along the diagonal of the matrix.

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

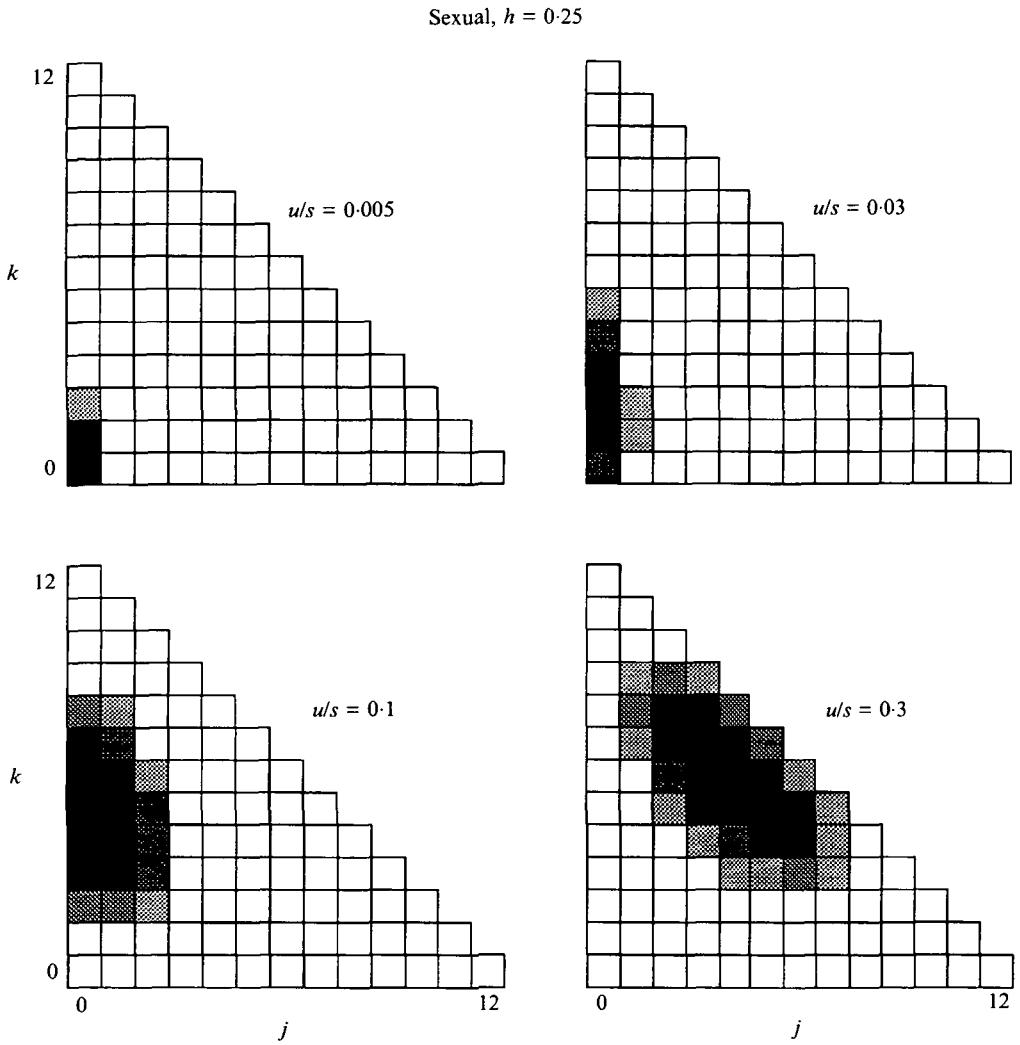

Fig. 6. Representation of the quasi-species distribution for the sexual population which may be compared to the parthenogen in Fig. 5. The figures are very similar for small u/s but become increasingly different for larger u/s. There is no phase *III* in the sexual case.

The highest concentration element of the matrix has been coloured black, and the other elements have been assigned to one of 10 different grey levels according to their concentration relative to the highest concentration. In the case of the parthenogen we see that for small mutation rates (see figure 5 with $u/s = 0{\cdot}005$) the distribution is localized close to the origin. As the mutation rate increases the distribution begins to move away from the origin, mostly in the vertical direction (figure 5 with $u/s = 0{\cdot}03$). This shows that for fairly small mutation rates most of the mutations occur in heterozygous form. Further increase of the mutation rate leads to a significant number of homozygous mutations as well ($u/s = 0{\cdot}1$). Since $h = 0{\cdot}25$ in this example the transition to phase *III* occurs at $u/s = 0{\cdot}25$. Only elements on the diagonal of the matrix ($j+k = L$) are non-zero in this phase. The figure with $u/s = 0{\cdot}3$ shows the stationary distribution along the diagonal elements. Further increase of u/s causes the distribution to move towards the bottom right hand corner ($j = L, k = 0$), and phase *II* will be reached at $u/s = 0{\cdot}75$ in this example. The sexual case is illustrated in figure 6 for the same parameter values as for the parthenogen in figure 5. Figures 5 and 6 are very similar for small u/s but become different for larger u/s since there is no phase *III* in the sexual case.

The centre of the stationary distribution is at the point $(\langle j\rangle, \langle k\rangle)$ within the triangular diagrams, where $\langle j\rangle$ and $\langle k\rangle$ are the mean values of the numbers of mutations calculated from the distributions (3.5) and (3.7). In phase *I* and in phase *III* $\langle j\rangle/L = a$, and $\langle k\rangle/L = b$. Figure 7 shows the path traced by the centre of the quasispecies for the parthenogen as the mutation rate increases. The path is different for different values of h. In each case the centre of the distribution leaves the origin in a vertical direction, indicating that only heterozygous mutations occur for small u/s. All the paths eventually end up in the bottom right hand corner for sufficiently large u/s. The shape of these paths is independent of L.

In the sexual case the centre of the quasi-species is at $\langle j\rangle/L = y^2$ and $\langle k\rangle/L = 2y(1-y)$. The path traced out by the centre of the quasi-species is determined by the parameter y, and its shape is therefore independent

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

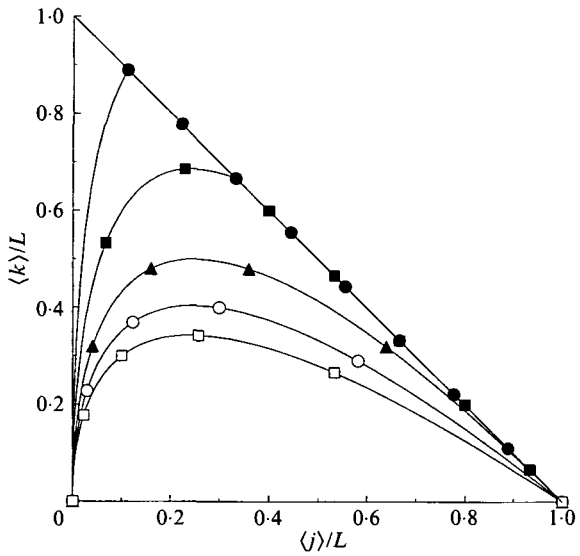

Fig. 7. Plotting $\langle k\rangle/L$ versus $\langle j\rangle/L$ shows the path traced out by the centre of the quasispecies as the mutation rate is increased. The paths are shown for the parthenogen in the multiplicative fitness landscape for several values of h (symbols as in Fig. 1). All paths leave the origin in a vertical direction, indicating that only heterozygous mutations occur for $u/s \ll 1$. In phase III the paths lie along the diagonal. In the sexual case the quasispecies follows the path indicated by triangles for all values of h, but the position on this path for a given value of u/s depends on h.

of h. It is the same path as shown in Fig. 7 for the parthenogen with $h = \frac{1}{2}$. For a given value of u/s the position of the quasi-species on this path does depend on h, however.

(iii) *Selfing*

In this case the parent produces gametes of both sexes which subsequently fuse to give a diploid offspring. The offspring is descended from only one parent but is not identical to its parent. It is convenient to define C'_{jk} as the concentrations of the offspring before accounting for mutations, and C_{jk} as the concentrations after mutation. Again we assume free recombination. The resulting equations are

$$C'_{nm} = \frac{1}{W}\sum_j\sum_k S_{nmjk}\, w_{jk}\, C_{jk}, \tag{3.18}$$

$$C_{jk} = \sum_n\sum_m M_{jknm}\, C'_{nm}. \tag{3.19}$$

The mutation matrix M_{jknm} is as in equation (3.4). S_{nmjk} is the probability that an individual of type jk produces an offspring of type nm by selfing, not accounting for mutations. The heterozygous loci in the parent segregate in the ratio $\frac{1}{4}:\frac{1}{2}:\frac{1}{4}$, therefore:

$$S_{nmjk} = \frac{k!}{(n-j)!\,m!(k+j-n-m)!}\left(\frac{1}{4}\right)^{n-j}\left(\frac{1}{2}\right)^{m}\left(\frac{1}{4}\right)^{k+j-n-m}. \tag{3.20}$$

Once again the loci may be shown to be independent. The fraction of loci homozygous for the mutation is u/s, and the fraction homozygous for the optimal alleles is $1-u/s$. Heterozygotes have negligible concentration if $u \ll 1$ and $s \ll 1$. The result is

$$C_{j0} = \binom{L}{j}(u/s)^j(1-u/s)^{L-j}, \tag{3.21}$$

which can be checked by substitution into (3.18) and (3.19). The interpretation of this is that heterozygotes are being formed at rate $2u$ from the optimal homozygotes, but the heterozygotes segregate rapidly to give the two types of homozygote in equal concentration. There is thus a net mutation rate of u from the homozygous optimal genotype to the homozygous mutant. The solution (3.21) is independent of h. The error threshold is at $u = s$, which is a larger value than for either of the other two reproductive systems. The mean fitness is $W = (1-u)^L$. This is greater than either the sexual or the parthenogen over the whole of the range of h and u/s. It is the fact that selfing eliminates the heterozygotes which leads to a higher fitness in this type of fitness landscape. We have assumed that the fitness of the heterozygote is intermediate between the fitnesses of the two homozygotes. If the heterozygote had an advantage over both types of homozygote then selfing would clearly lead to a lower fitness than sexual reproduction and parthenogenesis precisely because it eliminates the heterozygotes. We will not consider models of this type here.

(iv) *The limit $L \gg 1$*

The results of sections (i) (ii) and (iii) are valid for general L, and therefore apply if $L \to \infty$ with fixed u. However, for large L it is more natural to consider the overall mutation rate per genome U instead of the individual gene mutation rate u. For diploid models $U = 2uL$. If we take the limit $L \to \infty$ with U fixed, then in the case of the parthenogen we obtain from (3.5)

$$C_{0k} = \exp(-U/hs)\frac{(U/hs)^k}{k!}, \quad W = \exp(-U). \tag{3.22}$$

For the sexual case, taking the limit of (3.15) gives exactly the same result. Thus in this special limit the sexual and parthenogen have exactly the same stationary distribution and mean fitness. Only heterozygous mutations occur in this limit ($C_{jk} = 0$ if $j \neq 0$). The error thresholds have disappeared in this limit, just as they did for the haploid model in (2.11). Most previous work on multi-locus models has assumed this limit from the outset (Kimura & Maruyama, 1966; Kondrashov, 1982; Charlesworth, 1990), hence the general understanding that fitness does not depend on the reproductive method in multiplicative landscapes. In biological terms this limit is probably fairly reasonable since the per gene mutation rates u are

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

exceedingly small, even if the per genome rates U are sometimes appreciable (see for example Houle *et al.* 1992). However, in taking the $L \gg 1$ limit and dealing with U/s we are implicitly assuming that $u/s \ll 1$, and in view of the evidence that some mutations may be very nearly neutral we should be careful of making this assumption.

As a final point, if we take the same limit for the selfing population, we obtain from (3.21)

$$C_{j0} = \exp(-U/2s)\frac{(U/2s)^j}{j!}, \quad W = \exp(-U/2). \tag{3.23}$$

This is entirely different from (3.22) since only homozygotes are involved, instead of only heterozygotes. The fitness in this case may be substantially higher than for the sexual and the parethenogen if U is reasonably large.

In summary, it has been shown in Section 3 that even for the simplest possible diploid model with a multiplicative fitness landscape the mean fitness of the population depends on the breeding system. This is true even for a single locus, since we may just put $L = 1$ into the above formulae. Selfing gives an advantage over both sexual reproduction and parthenogenesis for all values of h and of u/s in this model. Sexual reproduction may be either advantageous or disadvantageous relative to parthenogenesis, depending on the parameter values.

4. Diploid models with epistatic interactions

Genes may interact with each other in such a way that the fitness contribution of any one gene depends on the other genes which are present in the sequence, i.e. the fitness contributions from the different loci are not multiplicative. A situation where the disadvantageous effect of a new mutant gene increases with the number of mutations already present is called synergistic epistasis, whilst a situation where the disadvantageous effect diminishes with the number of mutations already present is called diminishing epistasis. Consider a series of fitness peaks governed by the parameter α, such that the fitness of an individual with n mutant genes is

$$w(n) = \exp(-sn^\alpha). \tag{4.1}$$

If $\alpha > 1$ then the fitness decreases more rapidly than exponentially, and the interactions are of synergistic form. If $\alpha < 1$ then the interactions are of diminishing form. If $\alpha = 1$ then (4.1) is a special case of the multiplicative landscape already considered above. Several approximate solutions and numerical solutions have been given for landscapes very similar to this (Kimura & Maruyama, 1966; Kondrashov, 1982; Charlesworth, 1990). We will obtain analytical solutions as far as possible, and only use numerical solutions for verification. We will establish in what cases the previous approximate solutions are valid. In this section we will deal only with the large L limit where U is the appropriate variable for the mutation rate. In this case only heterozygous mutations appear for both the sexual and parthenogenetic populations, and therefore $w_{0k} = w(k)$ in terms of the previous notation.

(i) *Parthenogen*

The stationary distribution satisfies the equation

$$C_{0k} = \frac{1}{W}\sum_{m=0}^{k}\frac{U^{k-m}}{(k-m)!}\exp(-U)\exp(-sm^\alpha)C_{0m}. \tag{4.2}$$

The equation for $k = 0$ gives $W = \exp(-U)$ providing $C_{00} \neq 0$, which shows once again that W is independent of the fitness landscape for the parthenogen. Working only to first order in s and U we may obtain a convenient closed form approximation for the stationary distribution valid for $s \ll 1$ and $U \ll 1$. In this limit $\exp(-sm^\alpha) \approx 1 - sm^\alpha$. Setting $k = 1$ gives $C_{01} = (U/s)C_{00}$, and equations for the other k values give

$$C_{0k} = \frac{(U/s)^k}{(k!)^\alpha}C_{00}, \tag{4.3}$$

and C_{00} can then be obtained using the normalization condition. Since the sum

$$\sum_{k=0}^{\infty}\frac{(U/s)^k}{(k!)^\alpha}$$

converges for all positive values of α there is no error threshold with this type of landscape, and there is a finite concentration C_{00} for all finite U/s. If $\alpha = 1$ the different loci are independent and therefore (4.3) is a Poisson distribution as in (3.22). For other values of α the result (4.3) implies that there are correlations between the positions of mutant genes at different loci. The fraction of mutant genes is not the same as for the single locus problem, as it was for multiplicative landscapes.

Charlesworth (1990) has considered quadratic landscapes ($\alpha = 2$) in detail and finds that the stationary distribution can be approximated by a Gaussian distribution. Approximate analytical expressions for the mean and variance can be found. The method appears to give a good approximation over a wide range of parameter values; however there seems to be no way of generalizing this to non-quadratic landscapes. The approximation of (4.3) is only accurate at small U and small s, but is applicable for all values of α.

(ii) *Sexual*

Kimura & Maruyama, 1966, studied quadratic landscapes ($\alpha = 2$) with sexual populations, and obtained an approximate analytical solution. They assumed

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

that the stationary distribution was a Poisson distribution with mean λ and then found an approximate value for λ. The result was found to be reasonably close to the exact numerical solutions. This result is unsatisfactory, however, because no proof was given that the distributin was a Poisson distribution, and no suggestion was given as to how the approximation might be improved. In fact we can show that the exact solution is not a Poisson distribution in general, but that it tends to a Poisson distribution in the limit $s \ll 1$ and $U \ll 1$. Here we will obtain the solution to the quadratic landscape by more rigorous means, then we will show how the approximate treatment of Kimura & Maruyama (1966) can be used for general values of α.

We will try a solution for the gamete concentrations of the form

$$X_n = \frac{1}{A}\frac{e^{-\lambda/2}}{n!}\left(\frac{\lambda}{2}\right)^n (1-\epsilon n^2+\ldots). \tag{4.4}$$

It will be assumed that ϵ is a small parameter of order s, and work as usual only to first order in U and s. If ϵ were zero then this would be a Poisson distribution. It will be shown that the first correction to the Poisson distribution is in fact the term ϵn^2. The constant required to normalize the distribution is $A = 1-\epsilon(\lambda^2/4+\lambda/2)$. From equation (3.12) with $L \gg 1$ we find that $P(0k;nm) = 1$ if $n+m=k$ and is otherwise zero. Therefore from (3.9)

$$C_{0k} = \sum_{n=0}^{k} X_n X_{k-n} = \frac{1}{A^2}\frac{e^{-\lambda}}{k!}\lambda^k\left(1-\epsilon\left(\frac{k}{2}+\frac{k^2}{2}\right)\right). \tag{4.5}$$

$$W = \frac{1}{A^2}\sum_{k=0}^{\infty}\frac{e^{-\lambda}}{k!}\lambda^k\left(1-\frac{\epsilon}{2}(k+k^2)\right)(1-sk^2+\ldots) = 1-s(\lambda+\lambda^2)+\ldots. \tag{4.6}$$

The appropriate form for the mutation matrix in (3.11) is

$$M_{nm} = \exp(-U/2)\frac{(U/2)^{n-m}}{(n-m)!},$$

since the mutation rate per gamete is $uL = U/2$. From (3.10) and (3.11) a result for X_n can be found which has to be compared to (4.4). A solution of this form is possible only if $\epsilon = 2s$, and λ is the root of

$$2\lambda^2+\lambda-U/s = 0. \tag{4.7}$$

This is the same equation as was found by Kimura & Maruyama, 1966 (equation 1.10 of their paper). The advantage of the present method is that we have proved that the trial solution (4.4) satisfies the equations (3.9)–(3.11), whereas the previous approximate method involved no proof. The approximate method did not include the correction term ϵn^2 in (4.4), and we note that if this term is not included then it is impossible to find a consistent solution for λ by the substitution method above. The present method also shows that the approximation is valid if $U \ll 1$ and $s \ll 1$, but U/s may be either large or small. The approximation could be improved by including higher order correction terms if it were wished. If we compare the mean fitness obtained in (3.6) with the parthenogen result $W = e^{-U} \approx 1-U$, then we find that the sexual has a higher fitness than the parthenogen in the quadratic landscape for all values of U/s.

The physical meaning of λ is the mean number of mutations per diploid genome. This number may be large if U/s is large; however we have already implicitly assumed when we took the limit $L \gg 1$ that the fraction of mutations per locus, λ/L, is small. Because there is free recombination between the different alleles in the sexual case the mutant genes are constantly reshuffled and the stationary distribution is kept close to a Poisson distribution. This is likely to happen whatever the shape of the fitness peak. In the parthenogen there is no such reshuffling and the stationary distribution (3.3) is not close to a Poisson in the quadratic landscape. The fact that both sexual and parthenogen should give a Poisson distribution solution in the multiplicative landscape (equation (3.22)) shows that the multiplicative landscape is by no means a general case.

Having established that the method of Kimura & Maruyama gives a good approximation in the quadratic landscape for small U and small s we will now use a generalized form of their argument for the landscape with general α. Thus it will be assumed that

$$C_{0k} \approx \exp(-\lambda)\frac{\lambda^k}{k!}$$

for some λ which is to be determined. The mean fitness is therefore

$$W = \sum_k \exp(-\lambda)\frac{\lambda^k}{k!}\exp(-sk^\alpha) \approx 1-s\sum_k \exp(-\lambda)\frac{\lambda^k}{k!}k^\alpha. \tag{4.8}$$

The strength of selection against a new mutant gene depends on the number of mutations already present. Kimura & Maruyama define the 'mean selection coefficient' s^* by

$$s^* = \frac{1}{W}\sum_k C_{0k}(w_{0k}-w_{0(k+1)}) \approx s\sum_k \exp(-\lambda)\frac{\lambda^k}{k!}((k+1)^\alpha-k^\alpha). \tag{4.9}$$

The result for s^* has only been given up to first order in s since this argument turns out only to be valid in that limit anyway. In the stationary state mutation and selection pressures must balance, therefore $U = s^*\lambda$. This gives a closed equation for λ.

$$U/s = \lambda\sum_k \exp(-\lambda)\frac{\lambda^k}{k!}((k+1)^\alpha-k^\alpha). \tag{4.10}$$

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

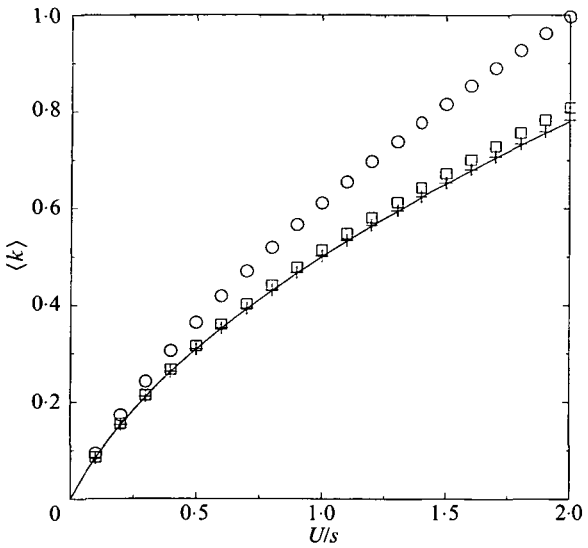

Fig. 8. Mean number of mutations per genome as a function of U/s in the quadratic landscape. Solid line shows analytical theory, which is exact for small U and small s. Symbols show numerically calculated values: (○), $s = 0{\cdot}1$; (□), $s = 0{\cdot}01$; +, $s = 0{\cdot}001$.

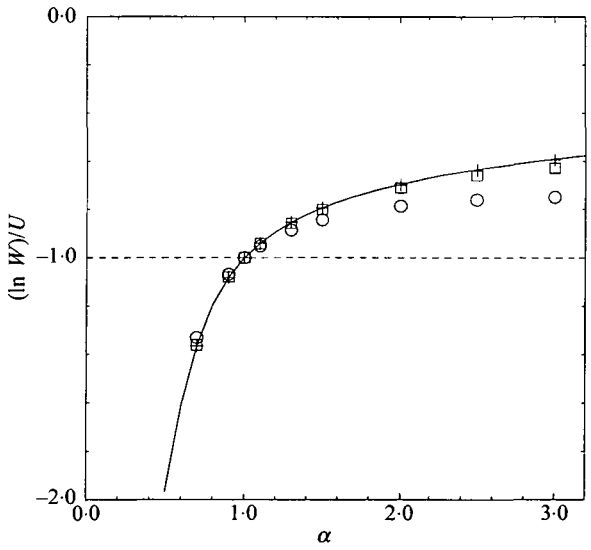

Fig. 9. Mean fitness of sexual population as a function of the exponent α which controls the epistatic interactions. Solid line shows theory for small U and small s. Symbols show numerical values for different values of s (as in Fig. 8). For each curve $U/s = 2$. For parthenogenic populations $(\ln W)/U = -1$ for all α, (dotted line). Sexual reproduction leads to a higher fitness for $\alpha > 1$.

When $\alpha = 2$ this is the same as (4.7). If α is an integer this gives a polynomial for λ of order α. A numerical solution for λ is possible from (4.10) for any value of α (even non-integers).

Figures 8 and 9 show numerical solutions for the stationary distribution which are compared to the analytical approximations. Figure 8 gives the mean number of mutations per genome as a function of U/s in the quadratic landscape. For small U and small s this converges to the value λ which is the solution of (4.7). Figure 9 shows the way the mean fitness varies as the shape of the landscape is changed. For the parthenogen $(\ln W)/U = -1$ in all cases. The curves for the sexual population are higher than -1 for $\alpha > 1$ indicating that sexual reproduction leads to an advantage when there is synergistic epistasis, and a disadvantage when there is diminishing epistasis (cf. Kondrashov, 1982). The solid line is the theoretical fitness from (4.8) with the value of λ obtained from (4.10). The theory appears to be exact in the limit of small U and small s for all values of α.

Another type of approximate analytical solution has been given by Charlesworth (1990) for the quadratic landscape with sexual reproduction. This involves replacing the discrete distribution V_{0k} by a Gaussian distribution and treating the number of mutations k as a continuous variable. This method appears to give a rather good approximation in the examples given by Charlesworth, although numerical analysis is required to solve the equations for the mean and variance of the distribution. The method given above is exact in the limit considered (small U and small s) whereas the Gaussian method is at best an approximation. Also it is not possible to generalize the Gaussian method to non-quadratic landscapes.

(iii) *Selfing*

We will now look at a population reproducing by selfing in the epistatic landscape of (4.1). For an individual with k heterozygous and j homozygous mutations the total number of mutant genes is $2j+k$, and the fitness is $w_{jk} = \exp(-s(2j+k)^{\alpha})$. Writing out equations (3.15) and (3.16) explicitly in this case gives

$$C'_{nm} = \frac{1}{W}\sum_{j=0}^{n}\sum_{k=n+m-j}^{\infty} \frac{k!}{(n-j)!\,m!(k+j-n-m)!}\,\frac{\exp(-s(2j+k)^{\alpha})\,C_{jk}}{2^m 4^{k-m}}, \quad (4.11)$$

$$C_{jk} = \sum_{m=0}^{k}\frac{U^{k-m}}{(k-m)!}\exp(-U)\,C'_{jm}. \quad (4.12)$$

The limit $L \gg 1$ has been taken, which greatly simplifies the mutation matrix in (3.16) and (3.4). In (3.4) mutations at previously homozygous optimal loci occur at rate $2u$, and these are accompanied by a factor of L from the binomial coefficients when $L \gg 1$. Hence $U = 2uL$ appears in (4.12). The mutations occurring at already heterozygous loci occur at rate u in (3.4), but these are negligible in the large L limit. Thus for the selfing population new heterozygous loci are created by mutation (4.12) and these segregate to produce homozygous loci in (4.11). The rate of formation of homozygous mutant loci by mutation alone is negligible if $L \gg 1$.

We have not found a general analytical solution of these equations, however an approximation is possible if $U \ll 1$ and $s \ll 1$. In this case the segregation rate is much quicker than the mutation rate, and the result is

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

to give an effective mutation rate of $U/2$ from homozygous optimal loci to homozygous mutant loci. The fraction of heterozygous loci is negligible. Thus $C_{jk} = 0$ if $k \neq 0$, and (4.11) and (4.12) can be combined to give

$$C_{j0} = \frac{1}{W}\sum_{n=0}^{j}\frac{(U/2)^{j-n}}{(j-n)!}\exp(-U/2)\exp(-s(2j)^{\alpha})\,C_{n0}. \tag{4.13}$$

This equation is very similar to (4.2) and can be solved in the same way.

$$C_{j0} = \left(\frac{U}{2^{\alpha+1}s}\right)^{j}\frac{C_{00}}{(j!)^{\alpha}} \tag{4.14}$$

This implies that $W = \exp(-U/2)$ independent of the fitness landscape. We have verified numerically that the distribution converges to (4.14) for small s and small U for several values of α. The mean fitness was measured numerically, together with the fraction of the population containing at least one heterozygous locus (i.e. the sum of the concentrations with $k \neq 0$). With $\alpha = 2$ it was found that with $s = 0{\cdot}001$ and $U = 0{\cdot}002$ only 0·4% of the population contained heterozygous loci and $(\ln W)/U = -0{\cdot}501$, whilst for $s = 0{\cdot}01$ and $U = 0{\cdot}02$ these values were 4% and $-0{\cdot}51$. These values are very close to the theoretical limit. For $s = 0{\cdot}1$ and $U = 0{\cdot}2$ there was a substantial difference from the limit: $(\ln W)/U = -0{\cdot}6$ and 30% of the population contained at least one heterozygous locus. If the mutation rate is too large then segregation does not remove the heterozygous loci quickly enough, and the approximation (4.13) is not valid. For each of these three sets of s and U values the mean fitness was measured as a function of α, as was done in Figure 9 for the sexual case. With selfing the mean fitnesses were almost constant over the range $0{\cdot}5 \leqslant \alpha \leqslant 3{\cdot}0$. This is in contrast to the sexual case where the fitness varies strongly with α. For these values of s and U the mean fitness for the selfing population was found to be higher than both the sexual and the parthenogen over the whole of this range of α. We also looked at the numerical solutions in the quadratic landscape with larger values of U where the analytical approximations are not valid. For a moderate degree of selection ($s = 0{\cdot}05$) numerical solutions were obtained for U varying between 0·01 and 10. It was found that the selfing population had a higher fitness than the sexual population over the whole of this range, and that both were higher than the value for the parthenogen ($W = e^{-U}$).

5. Discussion

We have seen that error thresholds can arise in two ways. These are seen in the two examples of the multiplicative landscape and the master sequence landscape discussed in Section 2. The multiplicative landscape behaves like a single locus model. There is an error threshold when $u/s \sim 1$ (the precise value depends on h for diploid models). In this case the single gene mutation rate becomes so high that the optimal allele at each locus can no longer be maintained by selection. In the master sequence landscape there is an error threshold when $U/s \sim 1$. In this case the overall genome mutation rate becomes too high, and the fittest genome sequence can no longer be maintained by selection. This second type of threshold can be seen in models where we take the $L \gg 1$ limit and work in terms of U, whereas the first type requires finite L and analysis in terms of u. The presence or absence of the second type of threshold depends on the shape of the fitness peak for large numbers of mutations. If the fitness continues to decrease smoothly as we move further from the optimal sequence (as with the function $w(n) = \exp(-sn^{\alpha})$) then there is no threshold. If the fitness remains constant when the number of mutations is large (as in the master sequence landscape) then there is a threshold. There is a clear physical analogy: a particle in an infinitely deep potential well is always bound, but a particle in a well of finite depth may be either bound or free depending on its energy. It follows that if the landscape decreases in fitness up to a certain number of mutations n_0 and is thereafter constant then there will still be a threshold whatever the value of n_0. In more realistic landscapes, of course, if we move too far from the original fitness peak we will encounter other peaks. We must then ask how likely it is for the population to shift from one peak to another (Barton & Rouhani, 1987). Stationary distributions may be of little relevance in complex landscapes, since equilibrium may never be reached.

The first type of threshold requires single gene mutation rate of order s. Rates per gene are known to be very small in cases where they can be measured (Gillespie, 1991; Houle *et al.* 1992), and it may be that the error threshold would not be reached in real organisms except for alleles which are almost exactly neutral. The second type of threshold is more likely to occur naturally, since per genome mutation rates may be quite large. Since this type of error threshold depends crucially on the landscape it would be of interest to know about the shape of the fitness landscape for real gene systems.

Whether or not there is an error threshold may be of more interest to mathematicians than biologists. The calculation of the mean fitness of the population is of direct relevance to biology, however, since it relates to the question of the relative advantages and disadvantages of different reproductive methods, and the reason for the evolution of sexual reproduction. The topic is very broad and we can only make a few very brief points here. For more details see Maynard Smith, 1978; Lewis, 1987; Stearns, 1987; Kondrashov, 1988; Hamilton *et al.* 1990; Charlesworth, 1990.

The basic problem is the two-fold 'cost of males' associated with sexual reproduction. An all female

https://doi.org/10.1017/S0016672300032092 Published online by Cambridge University Press

parthenogenetic strain arising in a sexual population should in principle reproduce twice as fast as the sexual individuals, half of which are male. This means that sexual strains would die out very rapidly if there were not some other large advantage associated with sex. Even though the realised cost of sex may be less than a factor of two in some cases (Lewis, 1987), the cost is probably substantial in many cases, and there must therefore be a rather large advantage of sex to counter this cost. Many suggestions have been made for the origin of this advantage.

It has been shown above that even in the simplest possible fitness landscapes the mean fitness of the population in the stationary state depends on the reproductive system. Kondrashov (1988) has argued that this is a significant factor in the evolution of sexual reproduction. It has been shown, and we have confirmed above, that when epistatic interactions are of synergistic form ($\alpha > 1$ in our model) there is an advantage of sexual reproduction relative to parthenogenesis. We have also shown that even in a multiplicative landscape there may be significant differences in fitness between the two types of population. The quantity $1-W$ measures the difference between the mean fitness and the maximum possible fitness, and is known as the mutational load (Kimura & Maruyama, 1966). When the mutation rate U is small the mutational load is small and the populations have mean fitnesses close to 1 irrespective of the reproductive system. For larger U the mutational load will be higher and the fitnesses of sexual and parthenogen may differ by a factor of order 2. Thus in principle models of this type can explain the advantage of sexual reproductive relative to parthenogenesis in situations where the mutational load is high. Kondrashov (1985) has also shown that sex can also given an advantage to selfing for the case of truncation selection (which is similar to $\alpha \gg 1$ in our models). On the contrary, for all the examples investigated numerically in section 4(iii) it was found that selfing gave the highest fitness of any of the three reproductive systems considered, although we have by no means attempted a full exploration of all the parameter space. In the models discussed here the heterozygote has a fitness intermediate between the two homozygotes. If the heterozygote had a fitness higher than both the homozygotes then there would be a clear disadvantage to selfing, since selfing removes the heterozygotes.

A sexual organism has both haploid and diploid stages in its life cycle. In principle the mutation rate is twice as high in the diploid phase, and the mean fitness in the stationary state should therefore be lower. Kondrashov & Crow (1991) have noted that there is a twofold cost of diploidy rather like the twofold cost of sex, and have discussed the reasons why diploidy may lead to an advantage, despite this cost.

A problem for the theory based on stationary mutant distributions is that it applies only in the limit of very large populations. For an asexual population of finite size there is no stationary mutant distribution because Muller's ratchet will operate, i.e. if the highest fitness genotypes are lost due to random fluctuations then there is no way to get them back again in the absence of back mutations. The mean fitness of the population thus continues to decrease indefinitely, with a rate dependent on the population size. In sexual populations the ratchet is countered by recombination even when there is no back mutation. This gives a substantial advantage to sex. Occasionally unfavourable mutations will become fixed in a finite sexual population due to random drift. Thus there may be a slow decrease in fitness for a sexual population even if the ratchet does not occur. It may be that to explain the advantage of sex we should not think in terms of stationary fitness values, but rather in terms of the rate of accumulation of mutations. In diploid organisms the effect of recombination rate and selfing rate on Muller's ratchet has been investigated in a large number of simulations by Charlesworth *et al.* 1992, 1993. Muller's ratchet in haploid organisms has been studied by several authors: Haigh, 1978; Nowak & Schuster, 1989; Stefan *et al.* 1993.

A further feature which may be important for the evolution of sex is that the fitness of any given gene sequence for a given species may change in time, either because of changes in the environment or because of the evolution of other species which interact with the species in question. It has been argued that the principal advantage of sex and recombination is to allow rapid response to changing fitness landscapes. One of the main features determining the fitness of a species maybe its ability to resist predators, parasites and disease-causing organisms (Hamilton *et al.* 1990). These organisms are themselves evolving and are likely to change in such a way as to increase their ability to prey on (or to infect) the original species. Sex and recombination are thought to provide a significant advantage in situations of host/parasite or predator/prey co-evolution (Hamilton *et al.* 1990; Stearns, 1987).

The models discussed above all have a single peak landscape with an optimum fitness genotype surrounded by lower fitness mutations. This is an extreme oversimplification. It may be more realistic to think of a rugged fitness landscape where there are many high fitness gene combinations which are peaks and ridges in the multi-dimensional sequence space. Evolution of populations in rugged landscapes has much in common with the statistical physics of disordered systems and with other types of complex optimization problems (Perelson & Kauffman, 1991). In fact, even in flat fitness landscapes (neutral evolution) many concepts of statistical physics, such as overlap distributions of genome sequences within a population, can be applied to problems of evolution (Derrida & Peliti, 1991; Higgs & Derrida, 1991, 1992). Whilst much is known about stochastic properties of finite popu-

lations evolving in neutral landscapes (Donnelly & Tavaré, 1987), very little is known about finite populations in even the simplest non-neutral landscapes. More recently the behaviour of quasispecies has been studied in rugged fitness landscapes with many local optima (Tarazona, 1992), and in landscapes with fitnesses related directly to RNA sequences (Fontana *et al.* 1989). RNA molecules have both rugged fitness landscapes which determine their evolution and rugged energy landscapes which determine their folding behaviour (Bonhoeffer *et al.* 1993; Higgs, 1993; Huynen & Hogeweg, 1993; Fontana *et al.* 1993). An important aim is to develop a theory of evolution at the molecular level on realistic landscapes such as these.

Returning to the models discussed in this paper, it would appear that considerations of mean fitnesses in the stationary distribution provide one possible explanation for the usefulness of sex and recombination, but that other factors such as fluctuations in finite popuations and time dependence of the fitness landscape also play a very important part. These other factors are in general rather difficult to treat analytically, and any attempt to do so must rely on a full understanding of the time-independent, infinite population case, which we have studied in this paper.

6. Conclusions

The object of this paper has been to draw a link between the quasi-species theory of molecular evolution and multi-locus diploid models used in population genetics. The ideas of the stationary mutant distribution (or quasi-species) and the error threshold are relevant in both cases. Sexual, Parthenogenetic, and Selfing populations have been studied in several types of landscape. Even in the simplest cases the mean fitness of the population depends on the reproductive system.

In the multiplicative landscape with fitness $w_{jk} = (1-s)^j(1-hs)^k$ we have given a solution for a general model with L loci. Several types of solution are possible which are separated by error thresholds. The sexual may have a higher or lower fitness than the parthenogen, depending on the values of h and u/s. The two have equal fitnesses in the limit u tends to zero and when $h = \frac{1}{2}$. Selfing leads to a higher mean fitness than either sexual reproduction or parthenogenesis.

A fitness landscape with epistatic interactions has also been studied with $w_{jk} = \exp(-s(2j+k)^\alpha)$. For the sexual population the stationary distribution is close to a Poisson distribution for $U \ll 1$ and $s \ll 1$. This confirms that a previous approximation of Kimura & Maruyama (1966) is in fact exact in the small U limit, but is not valid for general U. We have also extended the approximate treatment to general α values, and shown that it is exact for $U \ll 1$. Comparison of the mean fitness of sexual and parthenogen shows that the sexual population has a higher fitness when $\alpha > 1$. This confirms previous theories that sexual reproduction is advantageous in cases of synergistic epistasis. The mean fitness of a selfing population was found to be higher than both the sexual and the parthenogen over the range of parameter values studied.

The fitnesses of the statinary distributions are one factor which may explain the evolution of sexual reproduction and its prevalence over other systems. The stationary distribution may be of little importance, however, unless the population is of an extremely large size, since Muller's ratchet occurs in finite populations. Also if the fitness landscape is changing continuously then a stationary distribution may never be reached. In view of these other factors it is unlikely that the stationary distributions alone can explain the observed prevalence of sexual reproduction. A thorough understanding of the stationary case is nevertheless required before a theory of more complicated time-dependent, finite size population models can be developed.

I wish to thank the Royal Society and the University of Sheffield for my recent appointment as Sorby Research Fellow.

References

Barton, N. H. & Rouhani, S. (1987). The frequency of shifts between alternative equilibria. *Journal of Theoretical Biology* **125**, 397–418.

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. & Schuster, P. (1993). RNA multi-structure landscapes. *European Biophysical Journal* **22**, 13–24.

Charlesworth, B. (1990). Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genetical Research* **55**, 199–221.

Charlesworth, D., Morgan, M. T. & Charlesworth, B. (1992). The effect of linkage and population size on inbreeding depression due to mutational load. *Genetical Research* **59**, 49–61.

Charlesworth, D., Morgan, M. T. & Charlesworth, B. (1993). Mutation accumulation in finite outbreeding and inbreeding populations. *Genetical Research* **61**, 39–56.

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

Derrida, B. & Peliti, L. (1991). Evolution in a flat fitness landscape. *Bulletin of Mathematical Biology* **53**, 355–382.

Donnelly, P. & Tavaré, S. (1987). The population genealogy of the infinitely many neutral alleles model. *Journal of Mathematical Biology* **25**, 381–391.

Eigen, M., McCaskill, J. & Schuster, P. (1989). The molecular quasi-species. *Advances in Chemical Physics* **75**, 149–263. Also in abridged form in *Journal of Physical Chemistry* (*1988*) **92**, 6881–6891.

Fontana, W., Schnabl, W. & Schuster, P. (1989). Physical aspects of evolutionary optimization and adaptation. *Physical Review A*. **40**, 3301–3321.

Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D. & Schuster, P. (1993). RNA folding and combinatory landscapes. *Physical Review E* **47**, 2083–99.

Gillespie, J. H. (1991). *The causes of Molecular Evolution.* Oxford University Press.
Haigh, J. (1978). The accumulation of deleterious genes in a population – Muller's Ratchet. *Theoretical Population Biology* **14**, 251–267.
Hamilton, W. D., Axelrod, R. & Tanese, R. (1990). Sexual reproduction as an adaptation to resist parasites. *Proceedings of the National Academy of Sciences (USA)* **87**, 3566–3569.
Higgs, P. G. (1993). RNA secondary structure: a comparison of real and random sequences. *Journal de Physique (France) I* **3**, 43–59.
Higgs, P. G. & Derrida, B. (1991). Stochastic models for species formation in evolving populations. *Journal of Physics A (Mathematical and General)* **24**, L985–L991.
Higgs, P. G. & Derrida, B. (1992). Genetic distance and species formation in evolving populations. *Journal of Molecular Evolution* **35**, 454–465.
Houle, D., Hoffmaster, D. K., Assimacopoulos, S. & Charlesworth, B. (1992). The genomic mutation rate for fitness in Drosophila. *Nature* **359**, 58–60.
Huynen, M. A. & Hogeweg, P. (1993). Evolutionary dynamics and the relationship between RNA structure and RNA landscapes. Proceedings of the European Conference on Artificial Life, Brussels, May 1993.
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**, 1337–1351.
Kondrashov, A. S. (1982). Selection against harmful mutations in large sexual and asexual populations. *Genetical Research* **40**, 325–332.
Kondrashov, A. S. (1984). Deleterious mutations as an evolutionary factor. I. The advantage of recombination. *Genetical Research* **44**, 199–217.
Kondrashov, A. S. (1985). Deleterious mutations as an evolutionary factor. II. Facultative apomixis and selfing *Genetics* **111**, 635–653.
Kondrashov, A. S. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440.
Kondrashov, A. S. & Crow, J. F. (1991). Haploidy or diploidy: which is better? *Nature* **351**, 314–5.
Leuthäuser, I. (1986). An exact correspondence between Eigen's evolution model and a two dimensional Ising system. *Journal of Chemical Physics* **84**, 1884–1885.
Lewis, W. M. Jr. (1987). The costs of sex. In *The Evolution of Sex and its Consequences.* (ed. S. C. Stearns). Basel: Birkhäuser Verlag.
Maynard Smith, J. (1978). *The Evolution of Sex.* Cambridge University Press.
Nowak, M. & Schuster, P. (1989). Error thresholds of replication in finite populations, mutation frequencies, and the onset of Muller's ratchet. *Journal of Theoretical Biology* **137**, 375–395.
Perelson, A. S. & Kauffman, S. A. (Eds.) (1991). Molecular evolution on rugged landscapes: Proteins, RNA, and the immune system. (Santa Fe Institute). Addison-Wesley, California.
Stearns, S. C. (1987). Why sex evolved and the difference it makes. In *The Evolution of Sex and its Consequences.* (ed. S. C. Stearns). Basel: Birkhäuser Verlag.
Stephan, W., Chao, L. & Smale, J. G. (1993). The advance of Muller's ratchet in a haploid asexual population: approximate solutions based on diffusion theory. *Genetical Research* **61**, 225–231.
Swetina, J. & Schuster, P. (1982). Self-replication with errors – A model for polynucleotide replication. *Biophysical Chemistry* **16**, 329–345.
Tarazona, P. (1992). Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. *Physical Review A.* **45**, 6038–6050.
Wright, S. (1969). *Evolution and the Genetics of Populations (Vol. 2)* University of Chicago Press.