

# The Effects of Recombination on a Haploid Quasispecies Evolving on a Single-peak Fitness Landscape

Lionel Barnett

Centre for the Study of Evolution  
Centre for Computational Neuroscience and Robotics  
School of Cognitive and Computing Sciences  
University of Sussex  
Brighton BN1 9QH, UK

lionelb@cogs.susx.ac.uk

January 24, 2000

## Abstract

In this paper we investigate the effects of (uniform) recombination on the evolution of a haploid quasispecies on a single-peak “spike” fitness landscape. It is found that at low mutation rates the long-term behaviour of the quasispecies exhibits one stable and one unstable equilibrium. At a critical mutation rate which we identify as an *error threshold*, the stable and unstable equilibria coalesce. Beyond the error threshold there is no (biologically meaningful) equilibrium; errors accumulate and the optimum genotype is inevitably lost. In contrast to the asexual case where the error threshold is discontinuous, the sexual error threshold represents a 2nd order (continuous) phase transition. It is mooted that the unstable equilibrium could act as a barrier to the fixation of a favourable mutant. We derive analytic approximations for the equilibrium concentrations of the optimum genotype and for the error threshold. We also analyse the stability of the equilibria.

## 1 Introduction

Manfred Eigen, in his quasispecies formalism (Eigen, 1971; Eigen and Schuster, 1979; Eigen et al., 1989), developed an approach to analysing the evolution of large populations of information-encoding sequences based on (deterministic) flow-reactor kinetics, whereby concentrations of sequence types change according to differential rates of replication, destruction and, via mutation, transformation to different types. This formalism led to the concept of a *quasispecies* as a distribution of sequences localised in sequence space and clustered around the most frequent sequence variety. The existence of an *error threshold* of mutation (or replication fidelity) was established, beyond which the fittest sequence type would inevitably be lost from the population. The implication is that if the mutation rate is too high a favourable mutant can never become established in an evolving population. Furthermore, the error threshold typically decreases with increasing sequence length, so that there is effectively (for a given per-allele mutation rate) a limit to the sequence length beyond which an evolving population of sequences can maintain sufficiently high fitness to be viable. This observation leads to the so-called “error catastrophe”; in nature, the genomes of organisms have comparatively low effective per-allele mutation rates due to the existence of *error correction* mechanisms in effect during replication. However, these error correction mechanisms must themselves be coded for in the organism’s genome - they are functionally non-trivial and are likely to require lengthy coding sequences, greater than the “raw” per-allele mutation rate would permit due to the error threshold. How, then, could

these error correction mechanisms (and hence complex organisms requiring long genomes) have evolved?

There has been a persistent and recurrent idea that an answer to this conundrum may lie with sex and recombination (Maynard Smith, 1978; Kimura and Maruyama, 1966; Kondrashov, 1982; Charlesworth, 1990). Thus it has been suggested that, under certain circumstances, recombination can act as a kind of error repair mechanism. It is, therefore, of great interest to examine the effects of recombination on the dynamics of a quasispecies and on error thresholds in particular. In attempting to extend the “classical” quasispecies formalism to include recombination we immediately come up against two problems. The first is that in the asexual case analysis of the quasispecies dynamics is greatly abetted by the (near) *linearity* of the system; recombination introduces a quadratic non-linearity. Secondly, in the asexual case (and particularly if sequence lengths are long) we are generally entitled to ignore “back-mutation” of less fit sequences to the fittest sequence type. This simplifies the analysis considerably, enabling us to separate out the dynamics of the concentration of the fittest sequence variety. When recombination is present we may still neglect back-mutation, but we *cannot* ignore “back-recombination” (this is in a sense the essence of the error-correction potential of recombination) so that the dynamics of the fittest sequence type are inextricably linked to the concentrations of types nearby in sequence space. Our approach then is to develop approximations that reflect at least qualitatively the dynamics of the sexual quasispecies.

The basic quasispecies model employed in this paper is as follows: we consider a large (effectively infinite) population of haploid genotypes, considered as binary sequences of fixed length  $N$  evolving under selection, mutation and recombination. There is a single “optimal” genotype<sup>1</sup> and the fitness of any sequence depends only on the number of errors; i.e. the Hamming distance of that sequence from the optimal sequence.

We shall be interested mainly in the *long sequence length limit*  $N \rightarrow \infty$ ; all analytical results are strictly valid only in this limit. Numerical simulations are of necessity performed with finite sequence length, although care was taken to use the longest sequence lengths compatible with clarity and feasible within the constraints of computational resources. In what follows (unless otherwise stated) all Latin indices  $i, j, \dots$ , Greek indices  $\alpha, \beta, \dots$  and summations run from 0 to  $N$  (where  $N$  may be  $\infty$ ).

Let  $w_i$  denote the fitness of a sequence with  $i$  errors. We now specialise to a “spike” fitness landscape defined by:

$$w_i = 1 + \delta_{i0} \sigma = \begin{cases} 1 + \sigma & \text{if } i = 0 \\ 1 & \text{if } i > 0 \end{cases} \quad (1)$$

where  $\sigma > 0$  is the *selection coefficient*<sup>2</sup> of the optimum sequence. It must be remarked that while this fitness landscape arguably lacks biological relevance, it has the advantage of simplicity and allows for direct comparison with known results from asexual quasispecies theory. We shall return to this topic in Section 6 below.

We use  $x_i(t)$  to denote the proportion (or *concentration*) of sequences with  $i$  errors at generation  $t$ , so that  $\sum_i x_i(t) = 1$ .  $(x_i)_{i=1, \dots, N}$  represents the *quasispecies distribution* of the population. We will use the generating functions  $g_t(z)$  for the  $x_i(t)$  defined by:

$$g_t(z) \equiv \sum_k x_k(t) (1-z)^k \quad (2)$$

Note that  $g_t(0) = 1$  and (by convention)  $g_t(1) = x_0$ . We also define:

$$\theta(t) \equiv \sum_k k x_k(t) \quad (3)$$

---

<sup>1</sup> Also known in the literature as the “wild-type” or “master sequence”.

<sup>2</sup> It is commonplace in the population genetics literature to take the optimum fitness as 1 and that of other genotypes as  $1 - s$ . Since we shall only consider fitness-proportional selection, there is no essential difference;  $\sigma$  and  $s$  are related by  $1 + \sigma = \frac{1}{1-s}$ .

the *mean number of errors per sequence*. In terms of the generating functions  $g_t(z)$  we have:

$$\theta(t) = -g'_t(0) \quad (4)$$

where the prime denotes differentiation with respect to  $z$ . If the concentrations  $x_i(t)$  are time-independent we drop the argument  $t$ .

The remainder of the paper is organised as follows: Section 2 reviews the pertinent features of the model in the absence of recombination. Section 3 introduces recombination to the model while Section 4 presents the approximations used to analyse the sexual quasispecies. Section 5 addresses stability issues and Section 6 discusses some biological implications of the results.

## 2 The Asexual Quasispecies

Let us suppose that evolution of the quasispecies operates as follows: generations are non-overlapping. At each generation sequences are selected for reproduction proportional to their fitness. Each allele of a selected sequence then mutates (i.e. the binary allele flips) independently with probability  $0 < u < \frac{1}{2}$ . We also set  $U \equiv Nu = \text{mean number of mutations per sequence}$ . We then have:

$$x_i(t+1) = \frac{1}{W(t)} \sum_j m_{ij} w_j x_j(t) \quad (5)$$

where we have set:

$$m_{ij} \equiv \mathbf{P}(\text{a sequence with } j \text{ errors mutates to a sequence with } i \text{ errors}) \quad (6)$$

(note the order of indices) and  $W(t)$  is simply the population *mean fitness*:

$$W(t) \equiv \sum_k w_k x_k(t) = \sigma x_0(t) + 1 = \sigma g_t(1) + 1 \quad (7)$$

Equation (5) may be viewed as defining a (discrete)  $N$ -dimensional dynamical system. A straightforward calculation gives, for the mutation probabilities  $m_{ij}$ :

$$m_{ij} = \sum_{\alpha, \beta} \delta_{i, j - \alpha + \beta} \binom{j}{\alpha} \binom{N-j}{\beta} u^{\alpha+\beta} (1-u)^{N-(\alpha+\beta)} \quad (8)$$

In terms of the generating function  $g_t(z)$  we note the following: if  $(x_i)$  is the quasispecies distribution at a given generation and  $g(z)$  its generating function (2) then selection transforms  $g(z)$  according to:

$$g(z) \mapsto \frac{1}{W} [\sigma x_0 + g(z)] = \frac{\sigma g(1) + g(z)}{\sigma g(1) + 1} \quad (9)$$

In the long sequence length limit  $N \rightarrow \infty$  the action of mutation on the generating function is (see Appendix A):

$$g(z) \mapsto e^{-Uz} g(z) \quad (10)$$

Note that it follows that in the long sequence length limit  $m_{ij} = 0$  for  $i < j$ ; i.e. back-mutation becomes negligible. We may write (5) in terms of the generating function as:

$$g_{t+1}(z) = e^{-Uz} \frac{\sigma g_t(1) + g_t(z)}{\sigma g_t(1) + 1} \quad (11)$$

If the population is in dynamic equilibrium,  $x_i(t) = x_i$  for all  $i$  and  $t$ , then (11) becomes:

$$g(z) = e^{-Uz} \frac{\sigma g(1) + g(z)}{\sigma g(1) + 1} \quad (12)$$

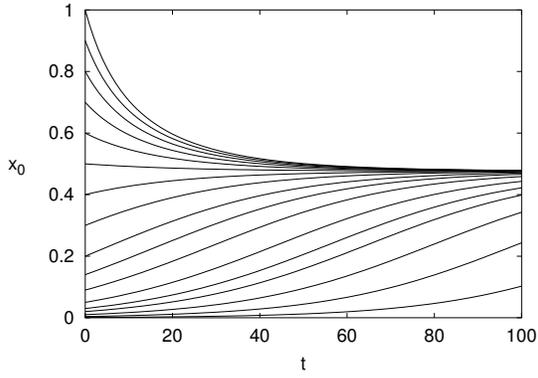


Figure 1: Genotype concentrations  $x_0(t)$  for the asexual quasispecies (5) plotted against time. Sequence length  $N = 20$ , selection coefficient  $\sigma = 0.1$ , per-genome mutation rate  $U = 0.05$ . We note that for this value of  $\sigma$ ,  $U_a \approx 0.0953$ .

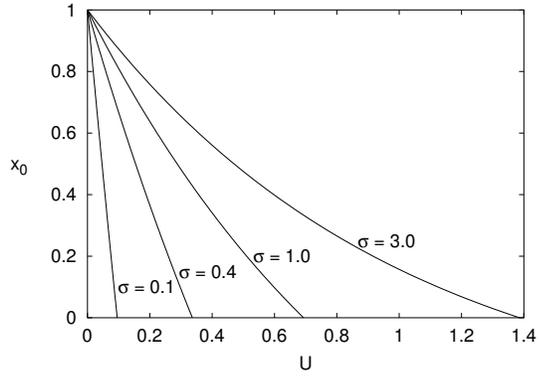


Figure 2: Equilibria of (5) plotted against per-genome mutation rate.

which may be solved directly for  $g(z)$ . We find in particular, setting  $z = 1$ , that the optimum genotype concentration is given by either  $x_0 = g(1) = 0$  or:

$$x_0 = g(1) = \frac{1}{\sigma} [e^{-U}(\sigma + 1) - 1] \quad (13)$$

Now  $x_0$  must be non-negative. From examination of (13) we see that, given a selection coefficient  $\sigma$ , there can only be an equilibrium solution with a non-vanishing concentration of the optimum genotype if  $U$  is less than a certain critical value  $U_a$  given by:

$$U_a = \log_e(1 + \sigma) \quad (14)$$

This critical mutation rate has been termed an *error threshold*. The behaviour of the model is illustrated in Figs. 1 and 2. In Fig. 1 the optimum genotype concentration  $x_0(t)$  as calculated from (5) is plotted against time for  $U < U_a$ . We see that there is a single stable equilibrium. As the mutation rate is increased to the critical rate  $U_a$  the equilibrium approaches zero discontinuously. In Fig. 2 the equilibrium optimum genotype concentrations are plotted against per-genome mutation rate for a few selection coefficients. The transition in the equilibrium behaviour of the quasispecies as the parameter  $U$  crosses the error threshold  $U_a$  is of a form that would be recognised by physicists as a *1st order* (or *discontinuous*) *phase transition*.

### 3 The Sexual Quasispecies

We now add recombination to the above model as follows: at each generation sequences are selected for reproduction proportional to their fitness. Selected sequences pair off at random; each pair produces an offspring with *uniform crossover* (Syswerda, 1989); i.e. each allele in the offspring sequence is chosen independently from one of its two parents with probability  $\frac{1}{2}$ . Each allele of the offspring then mutates as before. This model is similar to the model of retrovirus replication with superinfection presented in (Boerlijst et al., 1996). Equation (5) now becomes:

$$x_i(t+1) = \frac{1}{W(t)^2} \sum_{j,k,l} m_{ij} r_{jkl} w_k w_l x_k(t) x_l(t) \quad (15)$$

where we have set:

$$r_{jkl} \equiv \mathbf{P} \left( \begin{array}{l} \text{the offspring of a sequence with } k \text{ errors recombined} \\ \text{with a sequence with } l \text{ errors has } j \text{ errors} \end{array} \right) \quad (16)$$

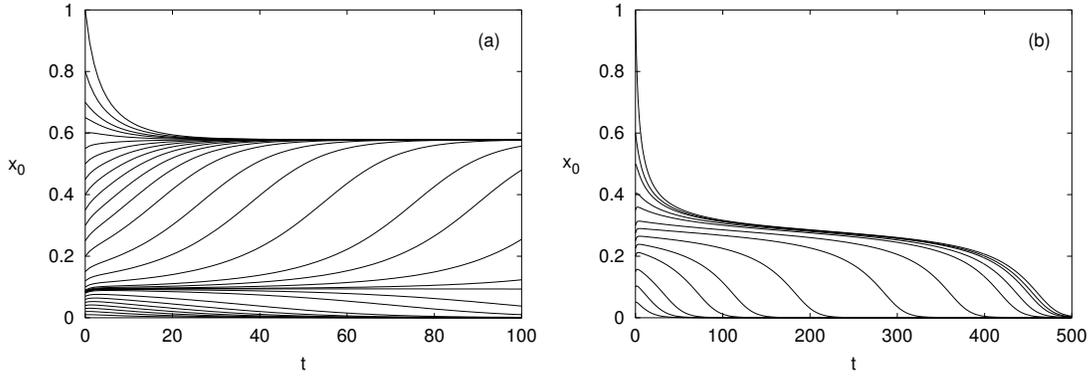


Figure 3: Genotype concentrations  $x_0(t)$  for the quasispecies with recombination (Eq. 15) plotted against time. Sequence length  $N = 20$ , selection coefficient  $\sigma = 0.4$  and (a) per-genome mutation rate  $U = 0.11$ , (b)  $U = 0.15$ .

Note that for (16) to make sense we have to suppose that recombination probabilities do not depend on the full distribution of all sequence types within the population, rather only on the quasispecies distribution. This will be the case if the population is in *linkage equilibrium*, which we explicitly assume. Now it is well-known that linkage equilibrium is reinforced by recombination and mutation but tends to be destroyed by finite population drift (Crow and Kimura, 1970; Maynard Smith, 1998). To test the assumption Monte Carlo simulations were run with finite populations of binary sequences for population sizes in the range of 100 – 10,000 genotypes, under the evolutionary algorithm described above. The results indicated that for a wide range of parameter values the population remains close to linkage equilibrium and that, furthermore, the infinite-population model provides a good approximation to the finite-population (stochastic) dynamics (but see also Section 5 below). Assuming linkage equilibrium it is not difficult to show that:

$$r_{jkl} = \sum_{\alpha} \binom{k}{\alpha} \binom{N-k}{l-\alpha} \binom{N}{l}^{-1} \binom{k+l-2\alpha}{j-\alpha} \left(\frac{1}{2}\right)^{k+l-2\alpha} \quad (17)$$

(note that this is actually symmetric in  $k, l$ ).

Analogous to (10), in the long sequence length limit  $N \rightarrow \infty$  the action of recombination on the generating function is given by (see Appendix B):

$$g(z) \mapsto g\left(\frac{1}{2}z\right)^2 \quad (18)$$

We may thus write (15) in terms of the generating function as:

$$g_{t+1}(z) = e^{-Uz} \left( \frac{\sigma g_t(1) + g_t\left(\frac{1}{2}z\right)}{\sigma g_t(1) + 1} \right)^2 \quad (19)$$

At equilibrium (19) becomes:

$$g(z) = e^{-Uz} \left( \frac{\sigma g(1) + g\left(\frac{1}{2}z\right)}{\sigma g(1) + 1} \right)^2 \quad (20)$$

Unlike (12) we cannot solve this equation explicitly for  $g(z)$  or indeed for  $x_0 = g(1)$ . We can, however, simulate (15) numerically. The results are illustrated in Fig. 3. Here the optimum genotype concentration  $x_0(t)$  as calculated from (15) is plotted against time. For the initial conditions binomial quasispecies distributions were chosen (see Section 4 below for justification). We see that at the lower mutation rate the dynamical system (15) apparently has a stable equilibrium (at  $x_0 \approx 0.6$ ) and an unstable equilibrium (at  $x_0 \approx 0.1$ ). There is also apparently a stable equilibrium at  $x_0 \approx 0$ , but we shall argue (see Section 6) that this equilibrium cannot be

biologically meaningful. At the higher mutation rate only the (unrealistic)  $x_0 \approx 0$  equilibrium remains. At a critical per-genome mutation rate  $U_s$  between these values the system bifurcates, the unstable and stable equilibria coalescing and vanishing. We identify this critical mutation rate as an error threshold since beyond this value the optimum genotype concentration inevitably falls to (nearly) zero. Again, a physicist would recognise this transition as a *2nd order* (or *continuous*) *phase transition*.

## 4 Approximations for the Sexual Quasispecies

Simulation of the sexual quasispecies model indicates that, due to the “shuffling” effect of recombination, the quasispecies distribution rapidly attains (from any initial conditions) a distribution close to a *binomial distribution*, which, in the long sequence length limit approaches a *Poisson* distribution. We thus proceed as follows: taking at generation  $t$  the Poisson distribution:

$$x_k(t) = e^{-\theta(t)} \frac{\theta(t)^k}{k!} \quad (21)$$

with generating function:

$$g_t(z) = e^{-\theta(t)z} \quad (22)$$

the evolutionary equation (15) yields for the next generation a distribution which will be “nearly Poisson”. We approximate this distribution by another Poisson distribution, choosing  $\theta(t+1)$  judiciously. This we shall do in two ways, according as the selection coefficient  $\sigma$  is small or large; in either case we effectively reduce the evolution of the quasispecies from an  $N$ -dimensional to a 1-dimensional dynamical system.

### 4.1 Small- $\sigma$ Approximation

If  $\sigma$  is small, the evolution of the quasispecies from one generation to the next was found empirically to be dominated by the mean number of errors  $\theta(t)$ . For the long sequence length limit we thus choose  $\theta(t+1)$  to be the mean number of errors one generation on, starting with a Poisson distribution (21) at generation  $t$ . Substituting  $g_t(z)$  from (22) in the right hand side of (19) then using the relation (4) we find immediately:

$$\theta(t+1) = U + \frac{\theta(t)}{\sigma e^{-\theta(t)} + 1} \quad (23)$$

The equilibrium condition  $\theta(t) = \theta(t+1) = \dots = \theta$  yields, after rearranging terms:

$$e^{-\theta} = \frac{U}{\sigma \theta - U} \quad (24)$$

which may be solved numerically for  $x_0 = e^{-\theta}$ . Equation (24) is observed to have two solutions for  $U$  smaller than a threshold value  $\widehat{U}_s$  which approximates the error threshold  $U_s$  of the exact model (15) for small  $\sigma$ .

We can calculate the approximate error threshold  $\widehat{U}_s$  as follows: the two solutions for  $\theta$  of (24) correspond to the points where the curves  $f(\theta) = e^\theta$  and  $g(\theta) = \frac{\sigma}{U}(\theta - U)$  intersect. At the approximate error threshold  $U = \widehat{U}_s$  these curves are tangential; i.e.  $f(\theta) = g(\theta)$  and  $f'(\theta) = g'(\theta)$ . Solving these equations we find that  $\widehat{U}_s$  is the (unique) solution of:

$$U e^{U+1} = \sigma \quad (25)$$

which may be solved numerically for  $\widehat{U}_s$  in terms of  $\sigma$ . We note that for small  $\sigma$ ,  $\widehat{U}_s$  is of the same order as  $\sigma$  and we have:

$$\widehat{U}_s = \frac{\sigma}{e} + O(\sigma^2) \quad (26)$$

This may be compared with  $U_a = \sigma + O(\sigma^2)$  for the asexual case (14). It is also not difficult to show that at the error threshold:

$$x_0 = \frac{1}{e} + O(\sigma) \quad (27)$$

## 4.2 Large- $\sigma$ Approximation

If  $\sigma$  is large, the evolution of the quasispecies was found to be dominated by the optimum genotype concentration  $x_0(t)$ . We proceed as for the small- $\sigma$  case, except that we now choose  $\theta(t+1)$  such that  $x_0(t+1) = e^{-\theta(t+1)}$  is the optimum genotype concentration in the next generation, again starting with the Poisson distribution (21) at generation  $t$ . Substituting  $g_t(z)$  from (22) in the right hand side of (19), setting  $z = 1$  and noting that  $x_0(t) = e^{-\theta(t)}$  we find:

$$x_0(t+1) = e^{-U} \left( \frac{\sigma x_0(t) + \sqrt{x_0(t)}}{\sigma x_0(t) + 1} \right)^2 \quad (28)$$

At equilibrium,  $x_0(t) = x_0(t+1) = \dots = x_0$ , we find (assuming  $x_0 > 0$  and taking square roots of both sides):

$$\sigma x_0 + 1 = e^{-\frac{1}{2}U} (\sigma \sqrt{x_0} + 1) \quad (29)$$

This is a quadratic equation for  $\sqrt{x_0}$  which may be solved explicitly, yielding two values for  $x_0$  so long as  $U$  is less than a critical value  $\tilde{U}_s$  which approximates the error threshold  $U_s$  of the exact model (15) for large  $\sigma$ .  $\tilde{U}_s$  is easily found to be:

$$\tilde{U}_s = -2 \log_e \left( \frac{2}{\sigma} (\sqrt{1+\sigma} - 1) \right) \quad (30)$$

For large  $\sigma$  we see that  $\tilde{U}_s$  scales as:

$$\tilde{U}_s = \log_e \frac{\sigma}{4} + O\left(\sigma^{-\frac{1}{2}}\right) \quad (31)$$

so that  $U_a - \tilde{U}_s = \log_e 4 + O\left(\sigma^{-\frac{1}{2}}\right) \approx 1.3863$  for large  $\sigma$ . We also find that at the error threshold:

$$x_0 = \frac{1}{\sigma^2} (\sigma - 2\sqrt{1+\sigma}) \quad (32)$$

which, for large  $\sigma$ , scales as:

$$x_0 = \frac{1}{\sigma} + O\left(\sigma^{-\frac{3}{2}}\right) \quad (33)$$

In Fig. 4 we plot optimum genotypes concentration  $x_0$  for the equilibria of (15) with  $N = 60$ , against per-genome mutation rate  $U$  for several values of the selection coefficient  $\sigma$ . The small- and large- $\sigma$  approximations (24), (29) for  $x_0$  are plotted on the same graph. In this figure the upper branches of the curves represent the stable and the lower branches the unstable equilibria. It was also found that for any  $\sigma, U$  the optimum genotype concentration  $x_0$  at equilibrium is always smaller with recombination than without.

Fig. 5 plots the error threshold  $U_s$  computed from numerical simulation of (15) with sequence length  $N = 80$  as well as the small- and large- $\sigma$  approximations  $\hat{U}_s$  and  $\tilde{U}_s$  against  $\sigma$ . The asexual error threshold  $U_a$  is also plotted for comparison.

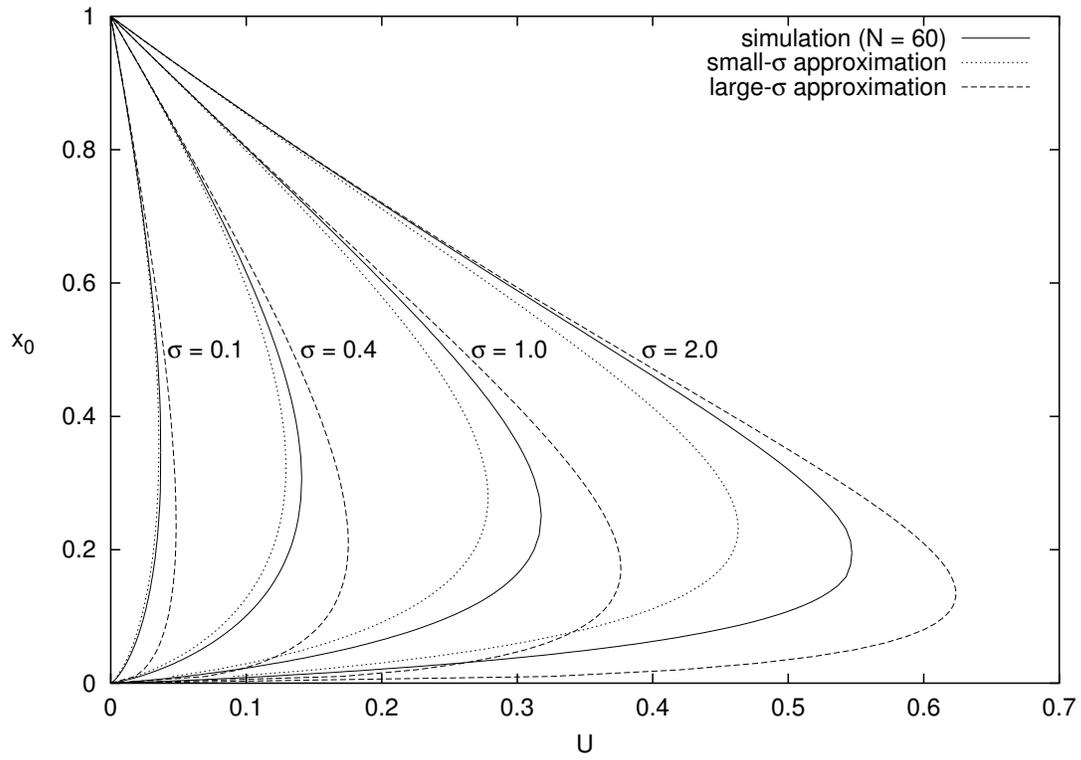


Figure 4: Equilibria of (15) and approximations (24), (29) plotted against per-genome mutation rate.

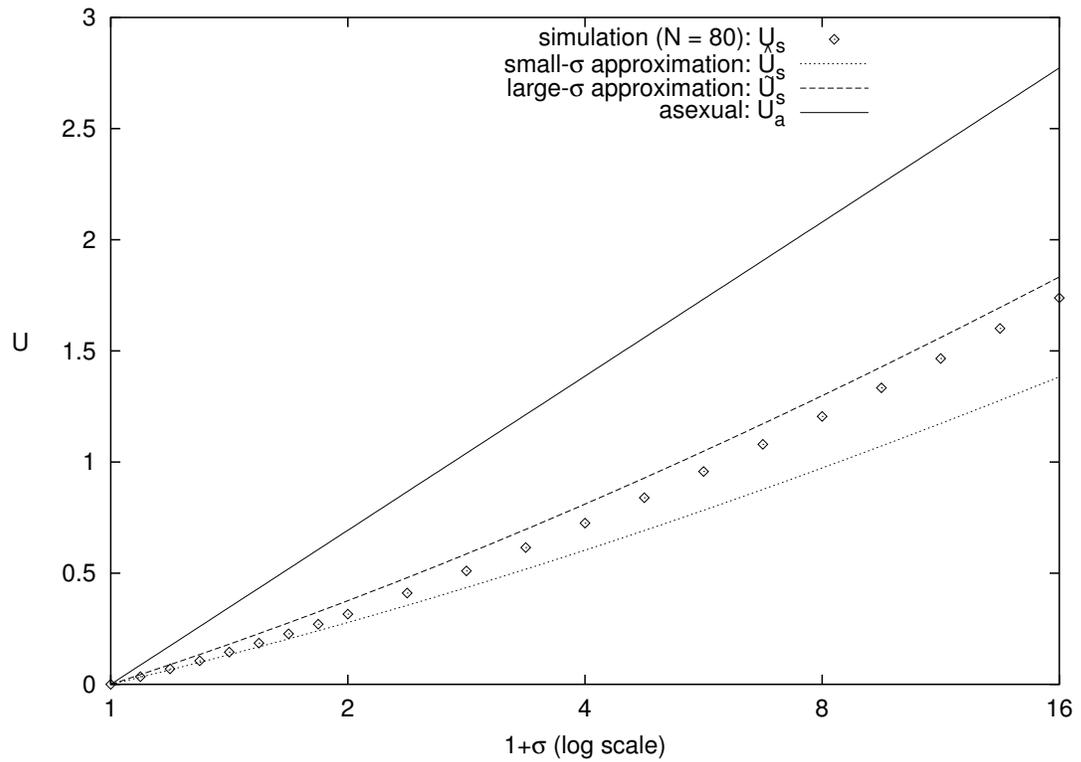


Figure 5: Error thresholds  $U_s$ ,  $\hat{U}_s$ ,  $\tilde{U}_s$  and  $U_a$  plotted against  $1 + \sigma$  (note logarithmic scale).

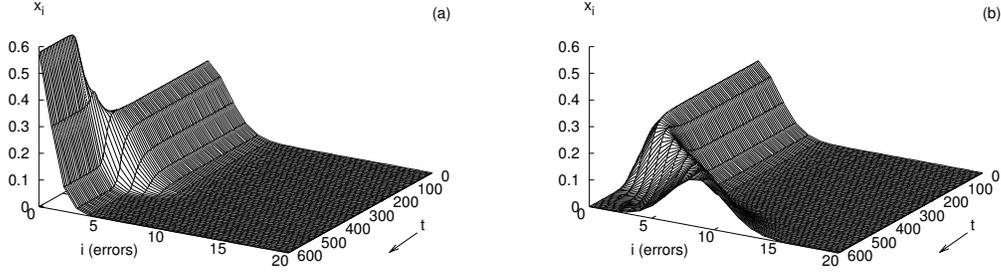


Figure 6: Behaviour of the sexual quasiespecies near the unstable equilibrium. In both cases  $N = 20$ ,  $\sigma = 0.4$ ,  $U = 0.11$  and the population was initialised with a binomial distribution (a) just above and (b) just below the unstable equilibrium.

## 5 Stability of Equilibria

We wish to investigate the *stability* of the equilibrium solutions to (15). This is of particular importance to analysis of finite-population models for which (15) may be an approximation, since stochastic fluctuations will occur in the concentrations  $x_i(t)$  which might destabilise a deterministic equilibrium. Furthermore, we note that, particularly for small  $\sigma$ , the system may persist in a state apparently close to the unstable equilibrium for a considerable time before destabilising (Fig. 6); we should like to elucidate the mechanism by which these “nearly stable” quasiespecies destabilise.

Consider a discrete dynamical system:

$$\mathbf{x}(t+1) = \mathbf{F}(\mathbf{x}(t)) \quad (34)$$

where  $\mathbf{x}$  is a real vector ( $x_i$ ) and  $\mathbf{F}(\mathbf{x})$  a (smooth) vector-valued function with component functions  $F_i(\mathbf{x})$ . Suppose further that  $\boldsymbol{\xi}$  is a *fixed-point* of (34); i.e.:

$$\boldsymbol{\xi} = \mathbf{F}(\boldsymbol{\xi}) \quad (35)$$

Suppose now that at time  $t$ ,  $\mathbf{x}(t)$  is close to  $\boldsymbol{\xi}$ ; i.e.  $\delta \equiv |\mathbf{x}(t) - \boldsymbol{\xi}|$  is small. We find then from (34) and (35) that:

$$\mathbf{x}(t+1) - \boldsymbol{\xi} = \nabla \mathbf{F}(\boldsymbol{\xi}) \cdot (\mathbf{x}(t) - \boldsymbol{\xi}) + o(\delta) \quad (36)$$

where  $\nabla \mathbf{F}(\boldsymbol{\xi})$  is the matrix with components  $\left. \frac{\partial F_i}{\partial x_j} \right|_{\mathbf{x}=\boldsymbol{\xi}}$  and (36) is the *linearisation* of the dynamical system (34) about the fixed-point  $\boldsymbol{\xi}$ . It represents the linear transformation mapping points in the vicinity of a fixed-point to their positions in the next generation. Now the *principal eigenvalue* of a linear transformation indicates the degree of “stretching” in the direction of greatest stretching; a fixed-point of a dynamical system (34) will be stable iff  $|\lambda_0| < 1$  where  $\lambda_0$  is the principal eigenvalue of  $\nabla \mathbf{F}$  at that fixed-point. Our evolutionary equations (15) are of the form (34) with  $\mathbf{F}$  given by:

$$F_i(\mathbf{x}) = \frac{1}{W(\mathbf{x})^2} \sum_{j,k,l} m_{ij} r_{jkl} w_k w_l x_k x_l \quad (37)$$

with the added constraint  $\sum_i x_i = 1$ . We find that at a fixed-point  $\boldsymbol{\xi}$ :

$$[\nabla \mathbf{F}(\boldsymbol{\xi})]_{ij} = \frac{2w_j}{W(\boldsymbol{\xi})} \left\{ -\xi_i + \frac{1}{W(\boldsymbol{\xi})} \sum_{k,l} m_{il} r_{ljk} w_k \xi_k \right\} \quad (38)$$

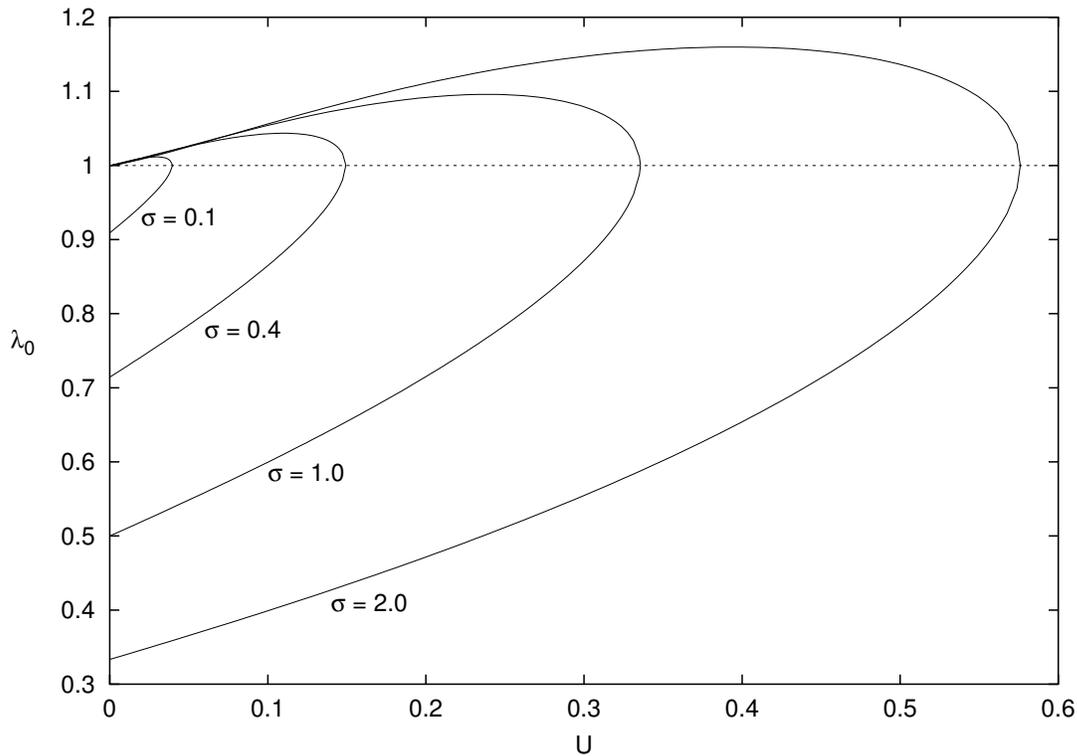


Figure 7: Principal eigenvalues of  $\nabla \mathbf{F}(\xi)$  for the stable (lower branches) and unstable (upper branches) equilibria of (15). Sequence length  $N = 60$ .

To analyse the linear transformation  $\nabla \mathbf{F}(\xi)$  given by (38) we calculated its eigenvalues  $\lambda_0 > \lambda_1 > \lambda_2 > \dots > \lambda_N = 0$  for the stable and unstable equilibria<sup>3</sup>. Fig. 7 plots the principal eigenvalues  $\lambda_0$  for a range of mutation rates and a few  $\sigma$  values, for the stable (lower branches) and unstable (upper branches) equilibria. It was also found empirically that the remaining eigenvalues fall off roughly exponentially; i.e. for fixed  $\sigma$  and  $U$  there is a constant  $c \approx \frac{1}{2}$  such that for  $k = 1, 2, \dots$  we have  $\lambda_k \approx c^k \lambda_0$ . It was certainly the case that for the stable equilibrium  $|\lambda_0| < 1$  (confirming stability) while for the unstable equilibrium  $|\lambda_0| > 1$  (confirming instability) and that in both cases  $|\lambda_k| < 1$  for  $k > 0$ . This latter implies in particular that the unstable equilibrium of (15) is only unstable along a single dimension - we might think of it as a narrow steep-walled saddle-shaped gully with a shallow curvature in the direction of the principal eigenvector of  $\nabla \mathbf{F}(\xi)$ . For small  $\sigma$  (see Fig. 7 and analysis below) we see that  $\lambda_0$  is only slightly larger than 1. This explains the comparative stability of the unstable equilibrium (Fig. 6). It is also interesting to note that for a given selection coefficient  $\sigma$  there is a critical mutation rate at which the instability of the unstable equilibrium is greatest. For higher mutation rates the unstable equilibrium becomes less unstable as the error threshold is approached.

To approximate the principal eigenvalues, we proceed as follows: in Section 4 we approximated the  $N$ -dimensional system (15) by the 1-dimensional systems (23) and (28). Consider the general situation where there is a vector function  $\phi(y) = (\phi_i(y))$  of a new variable  $y$  and a scalar function  $f(y)$  satisfying the relation  $\phi(f(y)) = \mathbf{F}(\phi(y)) \forall y$  or, in functional notation:

$$\phi \circ f = \mathbf{F} \circ \phi \quad (39)$$

This equation<sup>4</sup> formalises the notion of “reducing the dimension” of the dynamical system (34)

<sup>3</sup>It was found (although not proven analytically) that all eigenvalues were non-negative. We note that  $\sum_i x_i = 1$  implies that  $\nabla \mathbf{F}(\xi)$  is a *projection*, so that there must exist at least one zero eigenvalue.

<sup>4</sup>In mathematical “Category Theory” Equation (39) would define  $\phi$  as an *endomorphism* within the category of (discrete) dynamical systems.

to the new 1-dimensional dynamical system  $y(t + 1) = f(y(t))$ . We then have:

$$\phi'_i(f(y))f'(y) = \sum_j F_{i,j}(\phi(y))\phi'_j(y) \quad \forall y \quad (40)$$

where primes denote differentiation, so that if  $\eta$  is a fixed-point of  $f$  then  $f'(\eta)$  is an eigenvalue of  $\nabla \mathbf{F}(\boldsymbol{\xi})$  for  $\boldsymbol{\xi} = \boldsymbol{\phi}(\eta)$ , with eigenvector  $\boldsymbol{\phi}'(\eta)$ .

The small- $\sigma$  approximation of Section 4.1 is an approximation to just such a reduction of dimension (in the sense that the relation (39) is “almost” satisfied) if we identify  $y$  with  $\theta$ .  $\boldsymbol{\phi}(\theta)$  is then specified by (21) and  $f(\theta)$  by (23). The eigenvalue  $\hat{\lambda}_0 \equiv f'(\theta)$  at the stable (resp. unstable) fixed-point  $\theta$  is found to be:

$$\hat{\lambda}_0 = (1 + U) \left( 1 - \frac{U}{\theta} \right) \quad (41)$$

where  $\theta$  represents the stable (resp. unstable) solution of the equilibrium equation (24).

For the large- $\sigma$  approximation of Section 4.2 we identify  $y$  with  $x_0$ ;  $\boldsymbol{\phi}(x_0)$  is then specified by (21) and  $f(x_0)$  by (28). The eigenvalue  $\tilde{\lambda}_0 \equiv f'(x_0)$  at the stable (resp. unstable) fixed-point  $x_0$  is found to be:

$$\tilde{\lambda}_0 = \frac{2 - e^{-\frac{1}{2}U}}{\sigma x_0 + 1} \quad (42)$$

where  $x_0$  represents the stable (resp. unstable) solution of the equilibrium equation (29).

Numerical computation of  $\hat{\lambda}_0$  and  $\tilde{\lambda}_0$  showed them to be reasonable approximations to the principal eigenvalue  $\lambda_0$  of  $\nabla \mathbf{F}(\boldsymbol{\xi})$  (for both stable and unstable equilibria) for small and large values of  $\sigma$  respectively. We may also conclude that for small  $\sigma$  the unstable equilibrium is most sensitive to perturbations of  $\theta$ , the mean number of errors per sequence, while for large  $\sigma$  it is more sensitive to perturbations of  $x_0$ .

Finally, we return to a remark made in the Section 3, namely that the infinite-population model (15) is generally a good approximation to the corresponding finite-population (stochastic) model. This is not entirely true near the unstable equilibrium; unsurprisingly stochastic fluctuations will tend to dislodge the population from the vicinity of the unstable equilibrium, whence the population will either converge to the stable equilibria or errors will accumulate with loss of the optimum genotype (Fig. 8).

## 6 Discussion

Before discussing the implications of the results presented here, some remarks are in order about the fitness landscape. Several charges of biological implausibility might be leveled against our “spike” landscape. One is that it seems unlikely that there should be just a single optimum genotype and that all genotypes just a few mutations away from that optimum should have *equal* lower fitness; conventional wisdom has it that as mutations accumulate fitness should drop off according to some scheme (often taken to be multiplicative).

Perhaps a more serious charge is that it seems highly unlikely that *any number* of mutations could accumulate without impacting fitness. This will happen when the quasispecies distribution destabilises, in particular when the mutation rate is higher than the error threshold. This may be observed in Fig. 6b where after destabilisation the quasispecies distribution moves off in a “wave” towards a higher number of errors<sup>5</sup>. It is for this reason that we must reject the “near zero” optimum concentration in the finite sequence length case as a plausible equilibrium. Of course in the infinite sequence length idealisation there is no equilibrium at all - the “wave” simply keeps on moving. It should be noticed, furthermore, that in our derivation of the formulae for mutation and recombination probabilities in the infinite sequence length limit (Appendices A, B), the number of errors is held fixed during passage to the limit; for long (but finite) sequence

<sup>5</sup>This phenomenon was correctly predicted by John Maynard Smith (private communication).

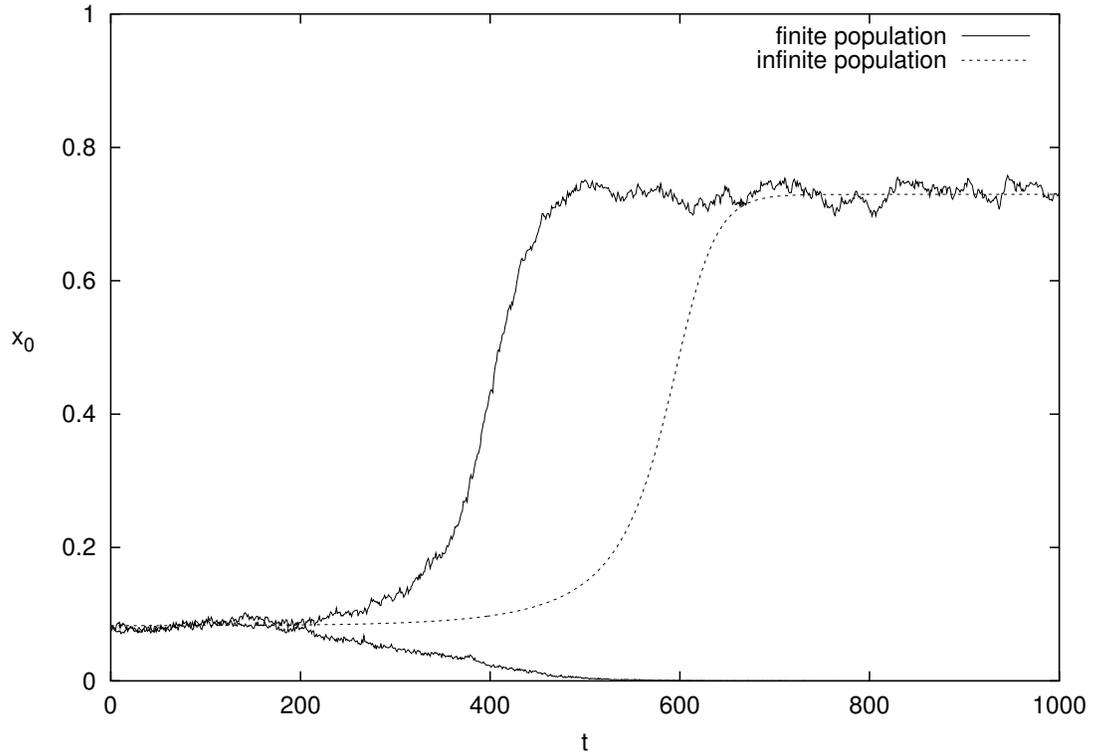


Figure 8: Optimum genotype concentration plotted against time for two typical simulations of a finite population (stochastic) sexual quasispecies initialised near the unstable equilibrium, alongside the corresponding infinite-population model (15). Sequence length is  $N = 80$ , selection coefficient  $\sigma = 0.4$ , per-genome mutation rate  $U = 0.1$  and population size for the finite population runs is 10,000.

lengths we should thus not expect our analytical model to furnish a good approximation if there are appreciable concentrations of genotypes with accumulations of errors comparable to the sequence length.

To test whether this last objection could perhaps be dispensed with we tried, for long sequence lengths, setting the fitness of any genotype with errors at more than e.g. 5% of the number of loci to zero; effectively any mutation beyond a fixed number would be lethal. It was found that except for very small selection coefficients and *provided the quasispecies doesn't destabilise* the effect of this alteration was surprisingly slight - perhaps less surprising if we recall that, according to our analysis, the concentration of genotypes falls off roughly exponentially with increasing number of errors, so that genotypes with significant accumulations of errors are not in any case contributing much to the behaviour of the model. Allied to this, we might also note that, in line with (45) the contribution of recombination falls off roughly exponentially with the number of errors in either parent sequence.

Comparing the behaviour of the sexual with the asexual quasispecies there are several striking features. In particular it seems clear that on the spike fitness landscape recombination is a distinct disadvantage for several reasons:

- The error threshold is lower with recombination than without.
- Even below the error threshold the optimum genotype concentration  $x_0$  (and hence the population mean fitness  $W$ ) is lower with recombination than without.
- Suppose that in a finite population our optimum genotype has been recently discovered by mutation/recombination. Even if any copies of the optimum genotype survived elimination by random drift, the concentration of the optimum genotype would have to rise above the level of the unstable equilibrium before selection could begin to “pull” it towards fixation - in the meantime mutation and recombination conspire to *reduce* the concentration. In particular, in a large population it is difficult to see how the optimum could ever fixate.

We should note that these conclusions do not contradict the theory propounded by Kondrashov and others (Kimura and Maruyama, 1966; Kondrashov, 1982; Charlesworth, 1990) that recombination can lead to a lower mutational load, since this is only claimed in the presence of *synergistic epistasis*; our spike landscape represents almost the opposite end of the spectrum to synergistic epistasis. Although by setting the fitness of the “error tail” to zero as described above our fitness landscape appears to acquire synergistic epistasis (i.e. it becomes technically sub-multiplicative), as long as the population doesn't destabilise it essentially only “sees” the non-synergistic portion of the landscape so that Kondrashov's principle is not violated.

Another striking difference is the following: in the asexual case, if the quasispecies is in equilibrium just within the error threshold we would expect to see a low concentration  $x_0$  of the optimal genotype (Equation (13) and Fig. 2). With recombination, we would expect to see a substantial concentration of the optimal genotype (Fig. 4), particularly if the selection coefficient  $\sigma$  is small when from (27) we have  $x_0 \approx 1/e \approx 0.3679$ . Thus if we observed a sexual population in equilibrium to have a reasonably high concentration of the optimum genotype we could not infer, as we might in the asexual case, that the mutation rate was well within the error threshold; in effect, a small change in mutation rate or selection pressure could push a seemingly stable sexual population catastrophically over the error threshold.

Finally, it was remarked in Section 3 that our model is similar to that in (Boerlijst et al., 1996). The principal difference is that in their model recombination occurs only with a given probability. They also consider fitness landscapes with a “plateau” of higher fitness around the optimum genotype as well as an isolated fitness spike. They found that in certain parameter regimes their model exhibited *bistability* (i.e. two stable equilibria). It seems, however, that in those cases their lower-fitness equilibrium corresponds to an unstable quasispecies which we have argued here should be rejected as a plausible equilibrium; further research is required. In particular, we hope to extend the analysis presented here to variable recombination rates and to landscapes featuring a plateau-like optimum - it is expected that results might differ substantially, since “back-recombination” would be expected to exert a more marked effect. We also intend to investigate more fully the implications of the model for finite populations.

## Acknowledgements

The author would like to thank Inman Harvey and Prof. John Maynard Smith for helpful discussions.

## A Transformation of $g(z)$ by Mutation

From (8) we have for any  $z$  and fixed  $U \equiv Nu$ :

$$\begin{aligned}
 \sum_i m_{ij}(1-z)^i &= \sum_{\alpha, \beta} \binom{j}{\alpha} \binom{N-j}{\beta} u^{\alpha+\beta} (1-u)^{N-(\alpha+\beta)} (1-z)^{j-\alpha+\beta} \\
 &= \sum_{\alpha} \binom{j}{\alpha} u^{\alpha} (1-u)^{j-\alpha} (1-z)^{j-\alpha} \\
 &\quad \times \sum_{\beta} \binom{N-j}{\beta} u^{\beta} (1-z)^{\beta} (1-u)^{N-j-\beta} \\
 &= (1-z+uz)^j (1-uz)^{N-j} \\
 &= \left(1 - \frac{1}{N}Uz\right)^N \left(\frac{1-z + \frac{1}{N}Uz}{1 - \frac{1}{N}Uz}\right)^j \\
 &\rightarrow e^{-Uz} (1-z)^j \quad \text{as } N \rightarrow \infty
 \end{aligned}$$

holding  $j$  fixed, where in the last step we have used  $\left(1 - \frac{1}{N}Uz\right)^N \rightarrow e^{-Uz}$  as  $N \rightarrow \infty$ . The result follows immediately.

## B Transformation of $g(z)$ by Recombination

Let us set:

$$c_{jk, \alpha} \equiv \binom{j}{\alpha} \binom{N-j}{k-\alpha} \binom{N}{k}^{-1} \quad (43)$$

Note that  $c_{jk, \alpha}$  is symmetric in  $j, k$ . Then from (17) we have:

$$r_{ijk} = \sum_{\alpha} c_{jk, \alpha} \binom{j+k-2\alpha}{i-\alpha} \left(\frac{1}{2}\right)^{j+k-2\alpha} \quad (44)$$

Now using Stirling's formula (Stirzaker, 1994) it is not difficult to show that, holding  $j, k$  and  $\alpha$  fixed we have  $\lim_{N \rightarrow \infty} c_{jk, \alpha} = \delta_{\alpha 0}$ . Thus, holding  $i, j$  and  $k$  fixed, we have:

$$\lim_{N \rightarrow \infty} r_{ijk} = \binom{j+k}{i} \left(\frac{1}{2}\right)^{j+k} \quad (45)$$

[c.f. Kimura and Maruyama (1966) - this is equivalent to neglecting the probability of *homozygous* mutant alleles occurring at any locus during recombination]. In the limit:

$$\sum_i r_{ijk} (1-z)^i = (1 - \frac{1}{2}z)^{j+k} \quad (46)$$

and the result follows.

## References

- Boerlijst, M. C., Bonhoeffer, S., and Nowak, M. A. (1996). Viral quasi-species and recombination. *Proc. R. Soc. Lond. B.*, 263:1577–1584.
- Charlesworth, B. (1990). Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.*, 55:199–221.
- Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, New York.
- Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 10:465–523.
- Eigen, M., McCaskill, J., and Schuster, P. (1989). The molecular quasispecies. *Adv. Chem. Phys.*, 75:149–263.
- Eigen, M. and Schuster, P. (1979). *The Hypercycle - a Principle of Natural Selforganization*. Springer, New York.
- Kimura, M. and 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. *Genet. Res.*, 40:325–332.
- Maynard Smith, J. (1978). *The Evolution of Sex*. Cambridge University Press, Cambridge, UK.
- Maynard Smith, J. (1998). *Evolutionary Genetics*. Oxford University Press, Oxford.
- Stirzaker, D. (1994). *Elementary Probability*. Cambridge University Press.
- Syswerda, G. (1989). Uniform crossover in genetic algorithms. In Schaffer, J. D., editor, *Proceedings of the Third International Conference on Genetic Algorithms and their Applications*, pages 2–9. Morgan Kauffman.