Recombination and Error Thresholds in Finite Populations

Gabriela Ochoa and Inman Harvey
Centre for the Study of Evolution
Centre for Computational Neuroscience and Robotics
School of Cognitive and Computing Sciences
The University of Sussex
Falmer, Brighton BN1 9QH, UK
E-mail: {gabro, inmanh}@cogs.susx.ac.uk

Abstract

This paper introduces the notions of 'quasi-species' and 'error threshold' from molecular evolutionary biology. The error threshold is a critical mutation rate beyond which the effect of selection on the population changes drastically. We reproduce, using GAs — and hence finite populations — some interesting results obtained with an analytical model — using infinite populations — from the evolutionary biology literature. A reformulation of a previous analytical expression , which explicitly indicates the extent of the reduction in the error threshold as we move from infinite to finite populations, is derived. Error thresholds are shown to be lower for finite populations. Moreover, as in the infinite case, for low mutation rates recombination can reduce the diversity of the population and enhance overall fitness. For high mutation rates, however, recombination can push the population over the error threshold, and thereby cause a loss of genetic information. These results may be relevant to optimizing the exploration-exploitation balance in GAs. Choices for critical GA parameters such as population size, mutation and recombination rates, should be reconsidered in the light of this new knowledge.

1 Introduction

One of the major issues in genetic algorithms (GAs) is the relative importance of two genetic operators: mutation and recombination (crossover) [Spe93]. Although there exists a large body of conventional wisdom concerning the roles of recombination and mutation, these roles have not been completely characterized. Furthermore, recombination is the primary operator distinguishing GAs from other stochastic search methods, much theoretical work in GAs is aimed at depicting the role of recombination in genetic search but precise knowledge is still lacking [MFH92, MH93, Mit96].

We are not alone. Questions on the evolution of sex and the role of sexual reproduction in nature have been among the major unsolved issues in evolutionary biology over many decades. Several hypotheses and models have been proposed to explain why sexual reproduction is maintained in most organisms in spite of the high cost associated with it [Wil75, MS78, ML88]. The GA theory community is beginning to pay attention to results coming from evolutionary biology [Boo93, Lev91, Lev95], and in particular from the field of population genetics. There is much more to be learned that is of potential interest to GA theory.

In this paper we reproduce, using GAs — and hence finite populations — some interesting results obtained with an analytical model — using infinite populations — from the evolutionary biology literature. These results concern the interaction between recombination and mutation on evolving populations of the so called 'quasi-species' (explained below), and show some unexpected effects of recombination on population evolution and the magnitude of the 'error threshold' . The error threshold is a critical mutation rate at which the population evolutionary dynamics change radically. There exists a phase transition between and "ordered" (selection-dominated) regime and a "disordered" (mutation-dominated) one. Mutation rates above this critical value cause a loss of the genetic information gained so far. The notion of error threshold, then, seems to be related with the idea of an optimal balance between exploitation and exploration in genetic search. Molecular biology research reveals that real virus populations — which are very efficiently evolving entities — have mutation rates very close but below their theoretically expected error threshold. In consequence, we argue that the notion of error threshold may well be related to the notion of optimal mutation rates in GAs.

Section 2 introduces the theory of quasi-species, and the notion of error-threshold. Section 3 reformulates an expression introduced by Nowak and Shuster [NS89]. This new reformulation explicitly indicates the reduction in the error threshold as we move from infinite to finite populations. Section 3 also discusses the relevance of error threshold to optimal mutation rates. Section 4 gives a detailed recapitulation of an analytical model dealing with recombination in populations of quasi-species. Section 5 describes the GA implementation translating this analytical model. Section 6 presents the main results, and ,finally, section 7 discusses results and their relevance to the theory of GAs.

2 Quasi-Species

The concept of a 'quasi-species' was developed in the context of polynucleotide replication, and in particular studies of early RNA evolution [Eig71, ES79, EMS88]. A protein space, [MS70] or more generally a sequence space, can be modeled as the space of all possible sequences of length ν drawn from a finite alphabet of size A. Each sequence has a fitness value which specifies its replication rate, or expected number of offspring per unit time. The fitnesses of all A^ν possible sequences define a 'fitness landscape'. When $A = 2$, a binary alphabet, the fitness landscape is equivalent to specifying fitness values at each vertex of a ν-dimensional hypercube; with some mathematical imagination (and some caution ...)

this can be pictured as spread out over a geographical landscape where fitness is analogous to height, and the dynamics of evolution of a population correspond to movement of the population over such a landscape.

Given an infinite population, and a specified mutation rate governing errors in (asexual) replication, one can determine the stationary sequence distribution reached after any transients from some original distribution have died away [EMS88]. Unless the mutation rate is too large or differences in fitnesses too small, the population will typically cluster around the fittest sequence(s), forming a concentrated cloud; the average Hamming distance between two members of such a distribution drawn at random will be relatively small. Such a clustered distribution is called a 'quasi-species'.

With a large finite population on the same fitness landscape the sequence distribution after many generations will typically be similar to that of the infinite case. The distribution will be noisy due to stochastic effects of the finite population size N. N is commonly far less than the number of possible sequences A^{ν}. With finite populations we can relax the requirement of waiting for a stationary final distribution; a finite population will cluster very earlier on in an evolutionary run. Even in the absence of selective pressures the convergence time for a population of size N is of order N generations in asexual populations and approximately order $N\left(ln(\nu)\right)^{1.1}$ generations in a sexual population with uniform crossover [AM94].

With finite populations we can also relax the requirement for a significant variation of fitnesses across the landscape. Even on a completely flat fitness landscape, where all sequences have the same fitness, a finite population will drift around in a quasi-species cluster or family of clusters [DP91].

These general results are relevant to GAs, with mutation and crossover in a finite population. Populations will genetically converge to a cluster or quasi-species after a limited number of generations, the 'width' of such clusters being largely determined by the balance between selection (inward) and mutation (outward) pressures, modified by the effects of genetic drift. For long term evolution within a GA, almost all of the run will be with such a clustered population. This genetic convergence is not a bar to further exploration and increase in fitness, and GAs can be modified to behave appropriately, as in Species Adaptation Genetic Algorithms [Har92, Har93].

2.1 The Error Threshold

When there are variations in fitness — the landscape is more or less rugged — and a low mutation rate, then the stationary sequence distribution of an infinite population will be focused around the point(s) of highest fitness. The same can be seen with a finite population centered around point(s) that are locally of highest fitness. As the mutation rate is increased, the local distribution widens and ultimately loses its hold on the local optima. Genetic Algorithm search can be considered as a balance between exploration (of the new) and exploitation (of what has been previously found to be fit). When mutation rates are too high then the search process can no longer exploit its history and it becomes random search.

This can be seen at its clearest in an extreme form of a fitness landscape which contains a single peak of fitness $\sigma > 1$, all other sequences having a fitness of 1. With an infinite population there is a phase transition at a particular error rate p, the mutation rate at each of the ν loci in a sequence. Following [ES79], we can determine analytically this critical error rate, which is defined as the rate above which the proportion of the infinite population on the peak drops to chance levels. The characteristic population distribution above and below this phase transition can be observed in figure $3a$.

Let $q = 1 - p$ be the per-locus replication accuracy. Then at the phase transition the probability of accurate replication of the 'master sequence' on the peak needs to be balanced by its superior replication rate, so as to equate with the replication of all the other sequences (we are, following Eigen and Shuster model [ES79], ignoring back-mutations from these to the master sequence).

$$\sigma q^{\nu} = 1 \tag{1}$$

$$(1-p)^{\nu} = \frac{1}{\sigma} \tag{2}$$

$$\left[(1-p)^{\frac{1}{p}}\right]^{\nu p} = \frac{1}{\sigma} \tag{3}$$

p is very small, so we can approximate the contents of the square brackets by e^{-1}. Hence

$$e^{-\nu p} = \frac{1}{\sigma} \tag{4}$$

$$p = \frac{ln(\sigma)}{\nu} \tag{5}$$

For mutation rates lower than this critical value, the error threshold, then the proportion of master sequences in the population will build up, giving the quasi-species centered around the peak.

The error threshold is of significance for GAs because it determines a critical balance between exploration and exploitation. In general, to maximize exploration the mutation rate should be as high as possible, but it should not be above the error threshold. Thus, the optimal balance between exploration and exploitation in GAs is assumed to be found with a mutation rate close to the error threshold but below it.

The single peak landscape abstraction employed here, is analogous to an scenario more familiar to the GA community, that is, a single block in the Royal Road Landscape [MFH92]. In the final stage of the search in this landscape, the final block(s) need to be completed without losing those that have been completed already. These spiky landscapes are one extreme of a continuum, with any less rugged landscape the error threshold would be smaller and the phase transition less sharp.

The above calculations are for infinite populations. For practical applications in GAs we must consider how the picture changes for finite populations.

3 The Error Threshold in Finite Populations

In [NS89] the calculations of an error threshold for *infinite* asexually replicating populations (which we will now call p_{∞})are extended to *finite* populations (where we shall call the critical rate p_N for a population of size N). In this latter case it is easier for a population to lose its grip on the solitary spike of superior fitness in a single peak landscape because of the added hazard of natural fluctuations in a finite population. The main result is presented as:

> The error threshold can be expanded in a power series of the reciprocal square root of the population size, and this increases with $1/\sqrt{N}$ in sufficiently large populations.

More precisely, the reciprocal square root factor applies to the *difference* between the critical *replication accuracy* in an infinite population $q_{min}(\infty)$, and the equivalent q_N in a population size N. The reference is to the second term in the following expansion, on the assumption that the third and subsequent terms are relatively insignificant and can be ignored [NS89].

$$q_N = q_\infty \left(1 + \frac{2\sqrt{\sigma - 1}}{\nu\sqrt{N}} + \frac{2(\sigma - 1)}{\nu N} + \frac{(\sigma - 1)^{3/2}}{\nu N^{3/2}} + \ldots\right) \tag{6}$$

where as before ν is the genotype length, and σ is the selection strength or superiority parameter of the master sequence. Since in many practical circumstances σ may lie between 1 and 5, then this implies that for values of $N \geq 100$ and of $\nu \geq 10$ then q_N should differ from q_∞ by only of the order of 1% or less. However, error thresholds are usually reckoned in terms of critical error rates $p = 1 - q$; and it turns out that the proportionate changes in critical values of p are much more significant in finite populations than the changes in q. Equation 6 was introduced by Nowak and Shuster [NS89]. Here we derive a reformulation of this equation, which makes explicit the reduction in the critical mutation rate as we move from an infinite population to one of size N. In other words, instead of calculating the critical *replication accuracy* (q_N), we wish to calculate the critical *error rate* (p_N):

$$\begin{aligned} \frac{p_\infty - p_N}{p_\infty} &= \frac{q_N - q_\infty}{p_\infty} && (7) \\ &= \frac{2\sqrt{\sigma - 1}(1 - p_\infty)}{\nu\sqrt{N}p_\infty} && (8) \end{aligned}$$

ignoring further terms in the expansion. Using (5) to substitute for p_∞, we have as the proportionate reduction in the error threshold:

$$\begin{aligned} \frac{p_\infty - p_N}{p_\infty} &= \frac{2\sqrt{\sigma - 1}}{\nu\sqrt{N}} \frac{(1 - \frac{1}{\nu}ln(\sigma))}{\frac{1}{\nu}ln(\sigma)} && (9) \\ &= \frac{2\sqrt{\sigma - 1}}{\sqrt{N}} \left(\frac{1}{ln(\sigma)} - \frac{1}{\nu}\right) && (10) \end{aligned}$$

For large values of ν the second term in the bracket is relatively insignificant and we have

$$\frac{p_\infty - p_N}{p_\infty} \simeq \frac{2\sqrt{\sigma - 1}}{ln(\sigma)\sqrt{N}} \tag{11}$$

Alternatively, we can present equation (10) as:

$$p_N = \frac{ln(\sigma)}{\nu} - \frac{2\sqrt{\sigma - 1}}{\nu\sqrt{N}} + \frac{2ln(\sigma)\sqrt{\sigma - 1}}{\nu^2\sqrt{N}} \tag{12}$$

In the experiments to be discussed below $\sigma = 5/3.5$; the genotype length is small at $\nu = 15$, making it less easy to ignore in equation (10). We can calculate in this case $p_\infty = 0.023778$, $p_{1000} = 0.021084$ a reduction of 11.33% and $p_{100} = 0.015257$ a reduction of 35.84% (using the less accurate equation (11) would give $p_{1000} \simeq 0.021018$ and $p_{100} \simeq 0.015050$).

3.1 Relevance of Error Threshold to Optimal Mutation Rates

On a single spike fitness landscape, the error threshold specifies a maximum mutation rate above which the population or quasi-species will lose any presence that it may have on the peak. What relevance does this have to GA problems, where researchers are generally more interested in the different problem of finding the peak in the first case? GA researchers may be more concerned to find an optimal mutation rate than the unfamiliar concept of an error threshold. In this section we shall briefly and selectively survey work on optimal mutation rates, and then relate these to the concerns of this paper.

Optimal mutation rates can be thought of as those which maintain an ideal balance between exploration and exploitation. Too low a mutation rate implies too little exploration — in the limit of zero mutation, successive generations of selection remove all variety from the population, and once the population has converged to a single point in genotype space all further exploration ceases. On the other hand, clearly mutation rates can be too excessive; in the limit where mutation places a randomly chosen allele at every locus on an offspring genotype, then the evolutionary process has degenerated into random search with no exploitation of the information gleaned in preceding generations.

Any optimal mutation rate must lie between these two extremes, but its precise position will depend on a number of factors including, in particular, the form of the fitness landscape under consideration. In conventional GAs, choice of mutation rates tends to be a low figure, typically 0.01 or 0.001 per bit as a background operator. However, the work on quasi-species and error thresholds (section 2) suggest that evolution works efficiently when mutation rates are directly below the threshold value above which information is destroyed. This idea is supported by T. Bäck [Bäc91] who suggests that "an optimal mutation rate for a GA is relatively large and turns mutation into an additional search operator".

Moreover, in conventional GAs, mutation rates are usually decided upon without regard to the genotype length. This despite suggestions from experimentation in [SCED89] that optimal rates $m_{opt} \simeq 1.829/(N^{1.0732}l^{0.4867}$ (where N is population size and l is genotype length); in [HM91] that earlier higher values should decrease exponentially towards $m_{opt} = \alpha'/(N\sqrt{l})$, for some constant α; and in [DeJ75] quoted in [HM91] as recommending $m_{opt} = 1/l$. The notion of error threshold confirms that the choice of an optimal mutation rate should consider the genotype length. Thus, such formulae point towards the right direction in this respect. However, they cannot be generally applicable, because they ignore at least two factors which are relevant:

1. Selection pressure.

2. 'Junk' or redundant loci.

One can propose simple thought experiments that support these ideas. Considering the first point, selective pressure, a generational GA can be seen as alternate applications of selection and then genetic operators. Any given selective pressure S_H can be emulated by two successive applications of a lower selective pressure (S_L where $S_L S_L = S_H$) *without* genetic operators. If the only genetic operator under consideration is mutation, where application of the optimal mutation rate is symbolized by M, then one generation of selection (at the original selective pressure) followed by mutation can be expressed as $S_H M = S_L S_L M \neq S_L M S_L M$. Clearly optimal mutation rates cannot be identical under these different selective pressures, so any general formula must take account of this.

Turning to the second point, let us consider a different thought experiment where for genotypes of length l the optimal mutation rate under selective pressure S is m mutations per genotype. We could in principle add an arbitrary number, for instance l, of 'junk' or redundant loci to the genotypes, without affecting the evolutionary dynamics as long as the expected number of mutations in the l relevant loci remained at M. However, direct application of the formula 'm mutations per genotype' (calculated on the new genotype length of $2l$) gives a revised, sub-optimal, mutation rate applied to the non-redundant loci. So once again, any general formula expressed in terms of mutations-per-genotype must take account of redundancy.

It can be seen from this discussion that general formulae for optimal mutation rates cannot be along the lines cited above. It can, however, be proposed that a mutation rate just below the error threshold is an optimal mutation rate for one extreme form of landscape under specific conditions. The single peak landscape studied above represents this extreme case, where the limiting behavior of the population as mutation rates increase gives rise to a phase transition at the error threshold. Here there is explicitly no redundancy, and the error threshold is indeed calculated in terms of selective pressure.

Suppose one starts with a population genetically converged at some point on the landscape other than the peak, and is seeking to (a) first *find* and (b) then *maintain* a presence on the peak. Then (in the absence of elitism [1]) we need a mutation rate which (a) maximizes the rate of (in this context) random search and yet (b) lies below the error threshold.

So error thresholds can be seen to be related to optimal mutation rates in this one extreme special case. It should be mentioned here that some authorities suggest that, in the natural world, effective mutation rates per genotype are generally maintained close to the error threshold at something of the order of one per genotype where genotype lengths vary from 4500 (bacterial viruses) to 6.10^9 (humans) [ES79].

4 Viral Quasi-Species and Recombination

Most mathematical models describing quasi-species focus on point mutations as the principal source of variation. However, Boerlijst et al. [BBN96] propose a mathematical model of quasi-species dynamics which incorporates both mutation and recombination. In particular they study virus populations. Viruses are infectious agents found in all life forms (plants,

[1] We are here assuming that circumstances do not allow elitism to be a viable option — for instance there may be some noise in fitness evaluations.

animals, fungi and bacteria). A virus particle consist of a core of nucleic acid, which may be DNA or RNA, surrounded by a protein coat. Certain viruses named 'retro-viruses' (e.g. HIV) can recombine their genetic material. They carry two copies of their genetic material in every virus particle, thus, recombination may occur when two distinct strains of the same virus simultaneously infect a single cell. Virus populations are quasi-species. The model of Boerlijst and co-workers specifically deals with retro-virus recombination. They first consider viral quasi-species dynamics without recombination. Distinct viral strains are represented by bitstrings of length L. A set of differential equations (see Appendix) describe the change in uninfected cells x, infected cells y_i and free viruses v_i. In this model a parameter, p, stands for the mutation rate per bit. Next paragraph describes the analytical results obtained in this case.

From now on, following Boerlijst et al. notation, we will use p to represent a variable mutation rate and p_c to indicate the critical mutation rate (or error threshold).

Without mutation ($p = 0$), the strain with the largest reproductive ratio R_i will outcompete all other strains. With mutation ($p > 0$) there is a critical error rate, p_c (the error threshold discussed in section 2), beyond which the strain with the highest R_i fails to be selected. Boerlijst et al. consider a single peak fitness landscape, where a strain F has the highest reproductive ratio, R_F, and all other strains have the same but lower reproductive ratio R. If $p < p_c$ the quasi-species will be centered around the fittest strain F, which will be the most abundant. If $p > p_c$ the fittest strain F will not be selected and each virus strain will have essentially the same relative abundance.

4.1 Bitstring recombination model

In [BBN96] the mathematical model is then adapted to include recombination (see Appendix), and here we will summarize their results. Variables for double infected cells, and for viruses produced by these cells are incorporated. Double infected cells Y_{ij}, are infected with strain i and superinfected by strain j. v_{ij} represents the free virus produced by these super-infected cells, of which 25% will be homozygous type i, 25% will be homozygous type j, and 50% will be heterozygous. Due to this characteristic of the model, the recombination rate, r, has a maximum at $r = 0.5$, because only heterozygous virus particles can (effectively) recombine. To model recombination itself 'uniform crossover' [Sys89] is employed.

The steady state structure of the new set of equations including recombination is studied. Bitstrings have length 15. The recombination rate has a maximum at $r = 0.5$, for the reasons exposed above. Two abstract fitness landscapes, (a) Isolated peak landscape, and (b) Plateau landscape, are considered.

(a) Isolated peak landscape First, the case where only one strain F has an increased R_i value, a so-called 'isolated peak' landscape, is studied. This single bit string has fitness $R_F = 5$, all other strings (designed as mutants) have fitness $R_i = 3.5$. The steady state mutant distribution for this landscape, produces the following results. For an error rate of $p = 0.07$, the recombinant population (compared against the population without recombination) is in some sense more compact: there are less rare mutants, but there is also less of strain F. This distribution is qualitatively similar to that obtained experimentally for finite populations (in this paper), thus figure $1a$ illustrates this distribution, although for a distinct mutation rate.

On the other hand, for a slightly increased error rate of $p = 0.08$, recombination drives the population beyond the error threshold, resulting in an almost uniform distribution of mutants (see figure 1*b*, again for a qualitatively similar distribution). Thus, for an isolated peak landscape, recombination is always disadvantageous for the virus, because it decreases the abundance of F and shifts the error threshold towards lower mutation rates.

(b) Plateau landscape In this scenario, the fitness of mutants close by the fittest strain F is increased to $R_{H_1} = 4.8$, and $R_{H_2} = 4.6$. Where H_1 is the set of all mutants with a Hamming distance of 1 from the fittest string F, and H_2 the set of all mutants with a Hamming distance of 2 from F. In this case the steady state distribution of mutants shows that, before the error threshold at $p = 0.011$, the recombinant population is again more compact, and it has more of it mass in the middle of the fitness plateau (figure 2*a* mirrors these results, although for a distinct mutation rate). If the error rate is increased, at a certain point (around $p = 0.015$), and fairly suddenly, recombination can no longer keep the population in the middle of the fitness plateau (see figures 2*b* and 4*d*). On the other hand, the transition around the error threshold with no recombination is very smooth, and the magnitude of the error threshold itself is larger, acting around $p = 0.05$ (figure 4*c* qualitatively mirrors this behavior). Thus, in this situation, recombination is advantageous to the virus for small mutation rates.

4.1.1 Main conclusions for infinite recombinant populations

To summarize, Boerlijst et al. main conclusions are:

- For small mutation rates (i.e. below the error threshold), recombination can focus the quasi-species around a fitness optimum.
- Recombination shifts the error threshold to lower mutation rates, and make the transition sharper.
- Recombination is advantageous (in the sense that it increases average population fitness) if fitness is more correlated —as in the plateau landscape (b) — and if the mutation rate is sufficiently small.

Finally, the authors claim that they have extensively tested the diploid bit-string model for other fitness distributions such as 'smooth' fitness peaks, multiple peaks and random distributions; that they have looked to alternatives to uniform crossover, such as one-point and multi-point crossover; and that in all this cases the main conclusion holds: recombination shifts the error threshold towards lower mutation rates and makes the transition sharper.

5 Methods

Now we have described in detail the analytical model of Boerlijst and co-workers, we can move to the discrete world of computer simulations. Results obtained using infinite population models, can not be expected to automatically apply to the more realistic case of finite populations. Thus, we endeavored to develop a genetic algorithm simulation model to study similar scenarios in the latter case. Moreover, Boerlijst et al. study deals with a particular type of recombination in viruses. Our study employs a more general scheme of recombination — that used in GAs. For the GA implementation the following choices were made. A generational GA with fitness proportionate selection is employed. The genetic operators utilized are bit mutation and uniform crossover. Chromosomes have length 15. For both abstract fitness landscapes modeled, the isolated peak and plateau landscapes, the fittest string F is considered to be the string of all zeros — 000000000000000 — with no loss of generality. Any other bitstring or strain is referred to as a 'mutant', and belongs to one of the Hamming distance classes H_i, where i is the Hamming distance to F — in this case the number of ones in the bitstring. To run the experiments, the populations are initialized as follows. For the single peak landscape, around 50% of the population is set on the peak and the rest is randomly generated. For the plateau landscape 25% is set on the peak, 25% on the H_1 compartment, 25% on the H_2 compartment, and the rest is randomly generated. The fitness values, for both landscapes, are those employed by Boerlijst et al. (see section 4.1 above). Population sizes are set to 100 for one group of experiments and to 1000 for another. This is intended to study the effect of population size on the magnitude of the error threshold. To be able to compare the results with those of Boerlijst et al., the crossover rate is set to 0.5 in all experiments for sexual populations. The per bit mutation rate p is the subject of study, thus it is varied from $p = 0.005$ to $p = 0.04$, with a step size of 0.005. The number of generations per GA run is set to 250. This value was empirically selected, the distribution of mutants is fairly stable by this point in all cases. In order to cope with stochastic noise, each GA run is repeated 50 times and the results are averaged. GA parameters are summarized in the following table:

Chromosome length	15
Population size	100 or 1000
Crossover rate	0.0 or 0.5
Mutation rate	0.005 to 0.04, Step = 0.005
Generations	250
Trials per GA run	50

Table 1: GA parameters

6 Results

The experimental results obtained with the GA model described above mirrored qualitatively those produced by Boerlijst et al. (section 4.1). However, the error-threshold magnitudes differ considerably. In fact, the error threshold for finite populations is, in all scenarios, significantly smaller than for the infinite case.

Before further discussing the results obtained with the GA model, let us consider "a basic principle of recombination", as exposed by Boerlijst et. al. ([BBN96]). This principle holds for **any** type of recombination, and turns out to be an important element for understanding the effects of recombination in population evolution, and the stable distribution of mutants.

> Consider two sequences i and j with a genetic distance d_{ij} (for a bitstring model d_{ij} is the Hamming distance). Assume that these sequences recombine to produce an offspring k. If recombination is the only source of variation, we have
>
> $d_{ik} + d_{jk} = d_{ij}$.

> The genetic difference between the parents equals the sum of the genetic difference between offspring and each of the parents. This relation is important for our understanding of recombination. It shows that in sequence space recombination is always inwards pointing. ([BBN96], p. 1578)

> Next subsections discuss in detail the results obtained with the GA model.

6.1 Single peak landscape

Figure 1 reflects the distribution of mutants, above and below the error threshold for the recombinant population in an isolated peak fitness landscape. These plots, using logarithmic scale, are almost mirror images of those shown in ([BBN96], p. 1579).

Figure 1*a* shows mutant distribution for an error rate of $p = 0.01$, with or without recombination. The recombinant population turns out to be more compact — less diverse — in some sense: there are fewer rare mutants, there is also fewer of strain F. This effect of recombination can be understood as follows ([BBN96], p. 1579). Most of the population is of strain F. If strain F recombines with e.g. a strain in H_8, then, according to the principle of recombination discussed above, the offspring lies anywhere between F and H_8.

On the other hand, figure 1*b* shows that for a slightly increased error rate ($p = 0.015$) recombination drives the population beyond the error threshold, resulting in an almost uniform distribution of mutants. As it can be seen, the bulk of the recombinant population is in the H_7 and H_8 compartments, because these contain the most strains. The explanation suggested by Boerlijst and co-workers is as follows. Where recombination acts as a converging operation when F is involved, it acts as a diverging operation in other cases. If for instance two mutants in H_4 recombine, the product lies everywhere between F and H_8.

These distributions are observed at error rates of $p = 0.07$, and $p = 0.08$ respectively, for infinite populations [BBN96].

6.2 Plateau landscape

In the isolated peak landscape, recombination seems to be disadvantageous for the population, because it decreases the abundance of F and shifts the error threshold towards lower mutation rates. However, recombination can be advantageous for more correlated fitness landscapes, as for instance the plateau landscape (see section 4.1). Figure 2*a* shows the distribution of mutants in a plateau landscape for an error rate $p = 0.02$ — now with a linear scale. It can be seen that the bulk of the population is in the H_2 compartment. Recombination between two H_2 strains generates offspring anywhere between F and H_4. Recombination thus shifts part of the population back to the middle of the fitness plateau. However, for a slightly increased error rate, $p = 0.025$, recombination drives, again, the population beyond the error threshold (see figure 2*b*).

6.3 Population size and the magnitude of the error threshold

Figures 3 and 4 show graphically the critical mutation rate in the distinct scenarios for two population sizes — 100 and 1000.

It should be mentioned that whereas for infinite populations on a single peak landscape the definition of the error threshold is straight forward (there is a clear phase transition), this is not the case for finite populations (where the transition is less sharp). Moreover, if fitness is more correlated —as in the plateau landscape, the transition is even less noticeable. Nevertheless the error threshold can be identified visually for finite populations (see figures 3 and 4) with some degree of uncertainty.

Table 2 summarize the error thresholds values for finite populations in the single peak and plateau landscapes, as observed experimentally to the nearest step size of 0.005. Whereas table 3 show the error thresholds for infinite populations, as calculated in section 3 for an asexual population in a single peak landscape, and as reported by Boerlijst et al. [BBN96] for the other cases.

It should be noticed that the error thresholds observed experimentally for finite asexual populations — sizes 100 and 1000 — in the single peak fitness landscape, coincide very accurately with the values for these critical error as calculated in section 3 ($p_{1000} = 0.021084$ and $p_{100} = 0.015257$).

	Single Peak		Plateau	
	100	**1000**	**100**	**1000**
Asexual	0.015	0.020	0.030	0.035
Sexual	0.010	0.010	0.020	0.020

Table 2: Error thresholds for finite populations —sizes 100 and 1000.

	Single Peak	Plateau
Asexual	0.023778	$\simeq 0.05$
Sexual	$\simeq 0.075$	$\simeq 0.02$

Table 3: Error thresholds for infinite populations

It can be observed that:

- Error thresholds for finite populations are lower in most situations than for the infinite case.
- The error threshold for an asexually replicating population is in all scenarios smaller than for a sexually replicating one.
- For asexual replication, the error threshold is smaller the smaller the population size. This reduction can not be seen for sexually replicating populations, however, this may not be conclusive as experiments with a smaller mutation rate step should be realized.
- Error thresholds are higher in all situations for the more correlated fitness landscape studied — the plateau landscape.
- The transition in mutant distribution around the error threshold, is sharper in the case of sexually replicating populations compared to the asexual ones. Moreover, the transition seems to be sharper the smaller the population in all cases.

7 Discussion

For finite populations and in both abstract fitness landscapes studied, the stable mutant distribution was seen to be qualitatively similar to that for infinite populations. Thus, the main conclusions of Boerlijst and co-workers, summarized in section 4.1.1 above, hold in this case. However, the error thresholds are smaller in most scenarios for finite populations. Moreover, for asexually replicating populations, the smaller the population, the smaller the magnitude of the error threshold (or the greater the extent of the reduction from) compared to the infinite case. In the single peak landscape, the experimental results for asexually replicating populations were accurately predicted by the analytic expression derived in section 3.

The relevance of these results to the theory of GAs is twofold. First, in the study of optimal mutation rates, if as mentioned in section 3.1 the notion of error threshold turns out to be relevant in this respect. Secondly, in understanding both th role of recombination, and the interaction between recombination and mutation in GAs operation.

Although we have studied simple fitness landscapes, the isolated fitness landscape is an extreme case in the continuum of less rugged landscapes. The plateau landscape is a less extreme case that also showed distinct behaviors below and above a critical mutation rate. Further experiments are currently being designed to asses the correlation between error thresholds and optimal mutation rates in distinct scenarios. Particularly, the 'Royal Road' fitness landscape [MFH92] will be employed as it allows for easily representing landscapes with distinct degrees of fitness correlation and neutrality. If, as we expect, optimal mutation rates are closely related to error thresholds; higher values for mutation rates should generally be used in GAs for practical applications. Moreover, the following general suggestions, could be made:

- Given that error thresholds are inversely proportional to chromosome length; the mutation rate should be smaller, the longer the chromosome.
- Given that error thresholds were shown to be lower for small-sized populations; the mutation rate should be smaller, the smaller the population size.
- Given that recombination shifts the error threshold to lower mutation rates, the mutation rate should be smaller when recombination is used.
- Given that recombination was shown to increase the population average fitness in more correlated landscapes; the more correlated the fitness landscape is, the more the advantages of using recombination.

These suggestion should be tested using more realistic fitness functions. However, simple abstract fitness landscapes turn out to be very useful tools to explore evolutionary dynamics, and to test hypotheses regarding the roles of genetic operations in population evolution.

Finally a computational 'microanalytical' or 'agent-based' model — in this case the GA — could offer some advantages over an analytical model for evolutionary biology studies. In particular, there is the possibility of modifying the general assumption of random mating, allowing instead more biologically inspired patterns of sexual selection. Preliminary studies show that assortative mating can enlarge considerably the critical error rate. This allows, in consequence, the use of higher mutation rates without losing genetic information — which may have implications for the exploration aspect of GAs.

Acknowledgements

Many thanks to H. Buxton and A. Meier for helpful comments and discussion regarding this work.

References

[AM94] H. Asoh and H. Mühlenbein. On the mean convergence time of evolutionary algorithms without selection and mutation. In Y. Davidor, H-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature: PPSN III*, pages 88–97, Berlin, 1994. Springer–Verlag.

[Bäc91] T. Bäck. Self-adaptation in genetic algorithms. In F. J. Varela and P. Bourgine, editors, *Proceedings of the First European Conference on Artificial Life. Toward a Practice of Autonomous Systems*, pages 263–271, Paris, France, 11-13 December 1991. MIT Press, Cambridge, MA.

[BBN96] M. C. Boerlijst, S. Bonhoeffer, and M. A. Nowak. Viral quasi-species and recombination. *Proc. R. Soc. London. B*, 263:1577–1584, 1996.

[Boo93] Lashon B. Booker. Recombination distributions for genetic algorithms. In L. Darrell Whitley, editor, *Proceedings of the Second Workshop on Foundations of Genetic Algorithms*, pages 29–44, San Mateo, July 26– 29 1993. Morgan Kaufmann.

[DeJ75] K. A. DeJong. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. PhD thesis, University of Michigan, Ann Arbor, MI, 1975. Dissertation Abstracts International 36(10), 5140B, University Microfilms Number 76-9381.

[DP91] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53(3):355–382, 1991.

[Eig71] M. Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

[EMS88] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *J. Phys. Chem.*, 92:6881–6891, 1988.

[ES79] M. Eigen and P. Schuster. *The Hypercycle: A Principle of Natural Self-Organization*. Springer-Verlag, 1979.

[Har92] I. Harvey. Species adaptation genetic algorithms: The basis for a continuing SAGA. In F. J. Varela and P. Bourgine, editors, *Toward a Practice of Autonomous Systems: Proc. First Eur. Conf. on Artificial Life*, pages 346–354. MIT Press/Bradford Books, Cambridge, MA, 1992.

[Har93] I. Harvey. Evolutionary robotics and SAGA: the case for hill crawling and tournament selection. In C. Langton, editor, *Artificial Life III, Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVI*, pages 299–326. Addison Wesley, 1993.

[HM91] J. Hesser and R. Männer. Towards an optimal mutation probability for genetic algorithms. In H.-P. Schwefel and R. Männer, editors, *Parallel Problem Solving from Nature*. Springer-Verlag, Lecture Notes in Computer Science Vol. 496, 1991.

[Lev91] James R. Levenick. Inserting introns improves genetic algorithm success rate: Taking a cue from biology. In Rick Belew and Lashon Booker, editors, *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 123–127, San Mateo, CA, 1991. Morgan Kaufman.

[Lev95] James R. Levenick. Metabits: Generic endogenous crossover control. In Larry Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 88–95, San Francisco, CA, 1995. Morgan Kaufmann.

[MFH92] M. Mitchell, S. Forrest, and J. H. Holland. The Royal Road for genetic algorithms: fitness landscapes and GA performance. In F. J. Varela and P. Bourgine, editors, *Proceedings of the First European Conference on Artificial Life. Toward a Practice of Autonomous Systems*, pages 245–254, Paris, France, 11-13 December 1992. MIT Press, Cambridge, MA.

[MH93] Melanie Mitchell and John H. Holland. When will a genetic algorithm outperform hill climbing? In Stephanie Forrest, editor, *Proceedings of the 5th International Conference on Genetic Algorithms*, pages 647–647, San Mateo, CA, USA, July 1993. Morgan Kaufmann.

[Mit96] M. Mitchell. *An Introduction to Genetic Algorithms.* MIT Press, 1996.

[ML88] R.E. Michod and B.R. Levin, editors. *The Evolution of Sex: An examination of Current Ideas.* Sinauer Associates, MA, 1988.

[MS70] J. Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[MS78] J. Maynard Smith. *The Evolution of Sex.* Cambridge University Press, Cambridge, 1978.

[NS89] M. Nowak and P. Schuster. Error thresholds of replication in finite populations: Mutation frequencies and the onset of muller's ratchet. *J. Theor. Biol.*, 137:375–395, 1989.

[SCED89] J.D. Schaffer, R.A. Caruana, L.J. Eshelman, and R. Das. A study of control parameters affecting online performance of genetic algorithms for function optimization. In J. D. Schaffer, editor, *Proceedings of the 3rd ICGA*, San Mateo CA, 1989. Morgan Kaufmann.

[Spe93] William M. Spears. Crossover or mutation? In L. Darrell Whitley, editor, *Proceedings of the Second Workshop on Foundations of Genetic Algorithms*, pages 221–238, San Mateo, July 26– 29 1993. Morgan Kaufmann.

[Sys89] Gilbert Syswerda. Uniform crossover in genetic algorithms. In J. David Schaffer, editor, *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 2–9, George Mason University, June 1989. Morgan Kaufmann.

[Wil75] G. C. Williams. *Sex and Evolution.* Princeton University Press, Princeton, 1975.

Appendix

Boerlijst et al. model [BBN96] describes the change in uninfected cells x, infected cells y_i, and free virus v_i without recombination:

$$\frac{dx}{dt} = \lambda - \delta x - x \sum_i \beta_i v_i \tag{1}$$

$$\frac{dy_i}{dt} = x \sum_j Q_{ij} \beta_j v_j - a_i y_i \tag{2}$$

$$\frac{dv_i}{dt} = k_i y_i - u_i v_i \tag{3}$$

In this model λ is the influx rate of uninfected cells; δ, a_i and u_i are the death rates of, respectively, uninfected cells, infected cells, and free virus; β_i is the infection rate; k_i the production rate of new free virus; and Q_{ij} is the probability of strain j mutating to strain i. The mutation matrix is given by:

$$Q_{ij} = p^{H_{ij}} (1-p)^{L-H_{ij}} \tag{4}$$

Here p is the mutation rate per bit, L is the bitstring length, and H_{ij} is the Hamming distance between strings i and j. Error free replication is given by $Q_{ij} = (1-p)^L$.

Equations (1)-(3) are adapted to include recombination. Double infected cells Y_{ij} are added, which are infected with strain i and superinfected with strain j. v_{ij} is the free virus produced by this super-infected cells, of which 25% will be homozygous type i, 25% will be homozygous type j, and 50% will be heterozygous. The new set of equations becomes:

$$\frac{dx}{dt} = \lambda - \delta x - xV \tag{5}$$

$$\frac{dy_i}{dt} = xV_i - a_i y_i - s y_i V \tag{6}$$

$$\frac{dy_{ij}}{dt} = s y_i V_j - a_{ij} y_{ij} \tag{7}$$

$$\frac{dv_i}{dt} = k_i y_i - u_i v_i \tag{8}$$

$$\frac{dv_{ij}}{dt} = k_{ij} y_{ij} - u_{ij} v_{ij} \tag{9}$$

Here s is the rate of super-infection, $V = \sum_i \beta_i v_i + \sum_{ij} \beta_{ij} v_{ij}$ is the sum of all infectious virus and $V_i = \sum_j Q_{ij} \beta_j v_j + \sum_j Q_{ij} \sum_{kl} M_{jkl} \beta_{kl} v_{kl}$ is the sum of infectious virus of type i, after mutation and recombination, with M_{jkl} being the probability of strain k and l recombining to strain j. All other variables and parameters are as described in equations (1)-(3).

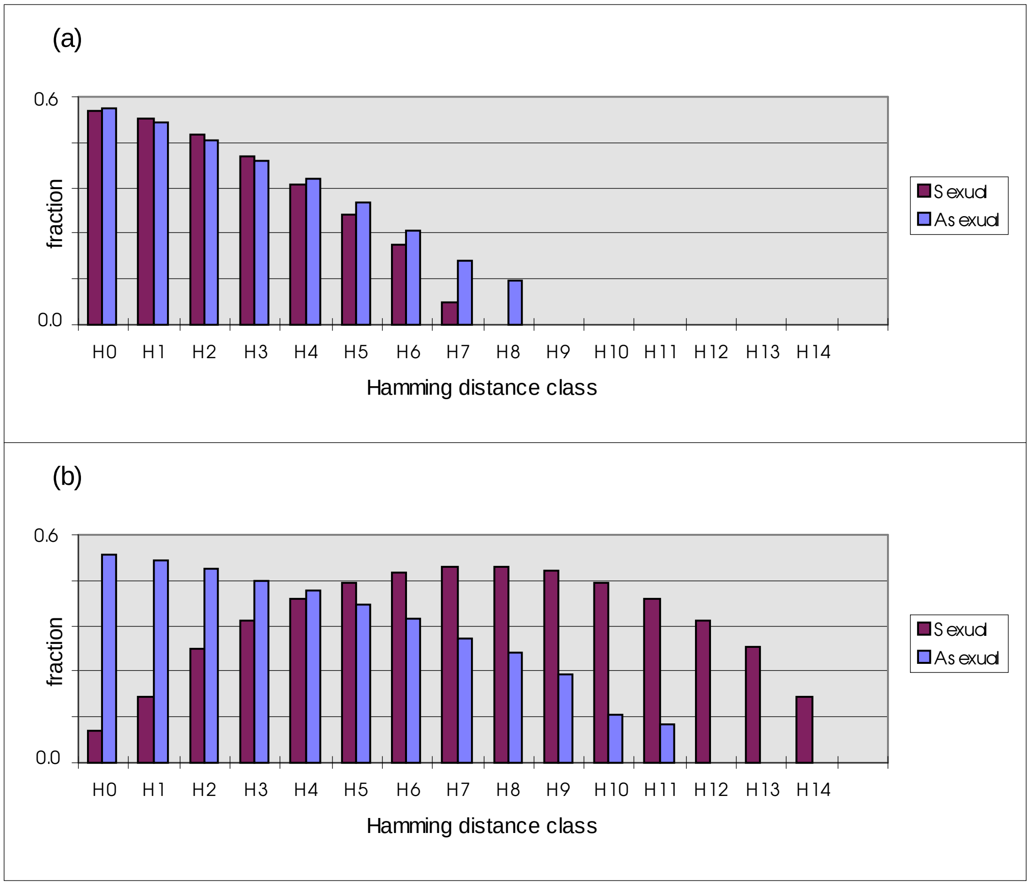

Figure 1: The effect of recombination on mutant distribution in a single peak fitness landscape for a population size of 1000 — logarithmic scale. (a) Below the error threshold ($p = 0.01$) the recombinant population is more compact. (b) For a slightly increased mutation rate $p = 0.015$, recombination can push the population over the error threshold. H_i denotes the sum of all mutants with a Hamming distance i to F (the fittest string)

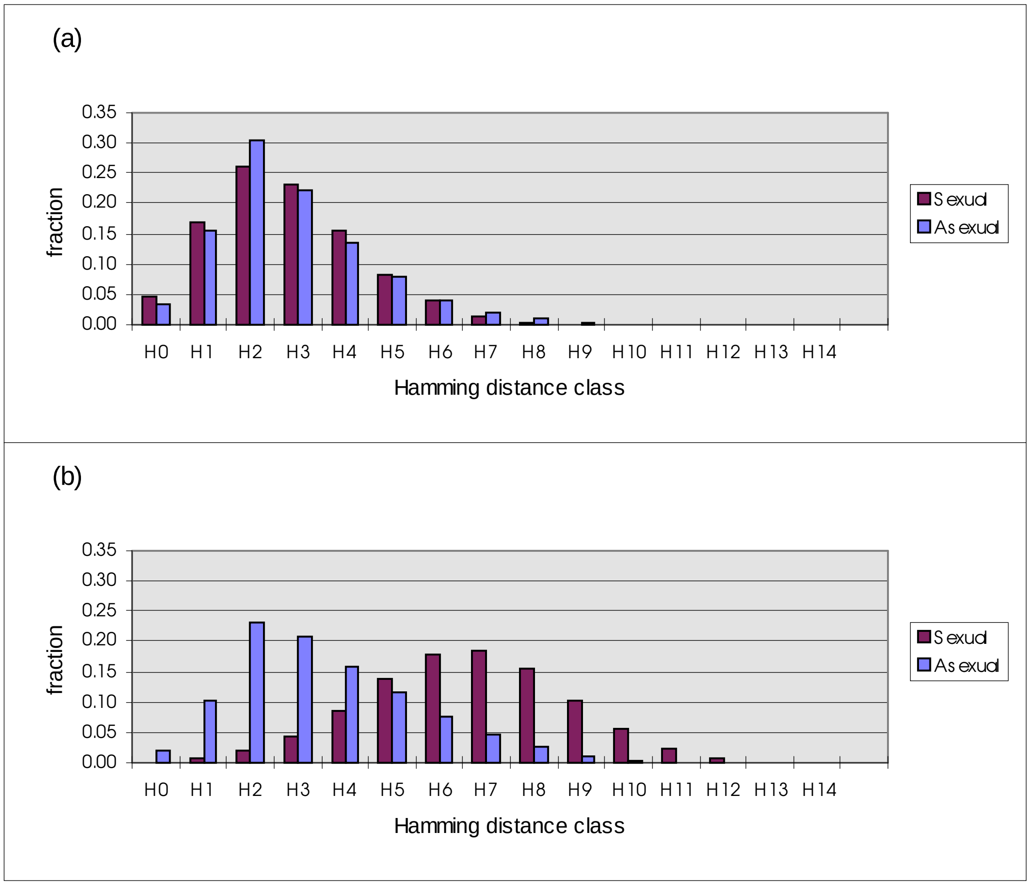

Figure 2: The effect of recombination on mutant distribution in a plateau fitness landscape for a population size of 1000 — linear scale. (a) Below the error threshold ($p = 0.02$) the recombinant population is again more compact, and it has more of its mass in the middle of the fitness plateau. (b) For a slightly increased mutation rate $p = 0.025$, recombination can no longer keep the population in the middle of the fitness plateau.