Study of the Meiotic Recombination Hotspot Diffusion Paradox

Jérémy Barbay

Technical Report CS-2006-38
David R. Cheriton School of Computer Science
University of Waterloo, Canada.

Abstract

Zoologists and evolutionists ponder about an apparent paradox in the current model of how sexual reproduction happens, basing their conclusions on extensive simulations. We show through a mathematical analysis that the results of those simulations can be predicted for a larger class of models, and we deduce from this analysis one of the key features of the model which yield the paradox. Based on this analysis, we define another mathematical model which solves this paradox, and we check our results through simulation.

Keywords: recombination hotspot paradox, markov chain, spatiality.

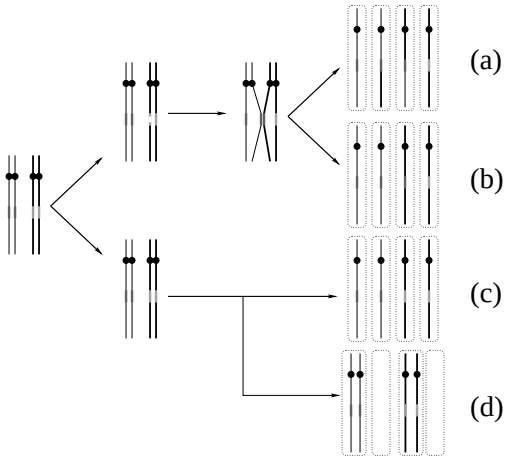

Figure 1: Crossover Mechanism: when a recombination hotspot allele "breaks", it is repaired and replaced by the recombination hotspot allele from the homologous chromosome, initiating a crossover (a), or not (b); when no recombination hotspot allele "breaks" the chromosomes sometime do not segregate properly (d).

1 Introduction

Zoologists and evolutionists ponder about an apparent paradox in the current model of how sexual reproduction happens [1, 10]. Solving this paradox is of practical importance for evolutionary inference about many species as well as for the study of human genetics [6, 7].

Figure 1 illustrates the mechanism on which we focus. Informally, the genetic material of humans, as for all diploid organisms, is composed of $2n$ chromosomes grouped in n pairs such that in the pairs, one chromosome comes from each parent. To reproduce, each parent "shuffles" each pair to produce four haploid gametes of n chromosomes each [5], so that the union of two gametes gives birth to a new individual, with a new genetic package. The "shuffling" procedure happens through specific regions of the chromosomes, called "recombination hotspots" [3, and references therein], which break and are repaired by copying from the corresponding region on the chromosome of the same pair, sometime resulting in an exchange between the two chromosomes: this is called a "crossover". The recombination hotspot alleles can be inactivated by mutations, which change them so that they do not break, and do not initiate a crossover any more. If a recombination is still initiated by the other chromosome from the pair, its recombination hotspot is repaired using the mutated allele, and the gametes will contain only inactivated hotspot alleles.

Boulton, Myers and Redfield [1] and later Pineda-Krch and Redfield [10] performed simulations to show that this model was self-contradicting: if a majority of inactive recombination hotspot alleles are produced each time an active one is activated in conjunction with an inactive one, the active recombination hotspot alleles would disappear from the population faster than they can appear through mutation. Recent studies point out that the degree of activity of recombination hotspots is not boolean, and that it can vary during evolution [2, 8], but the paradox remains the same: if the less active recombination hotspot alleles are the fastest to spread through the population, how comes that we still observe the activity of recombination hotspots, and that the population is not mainly constituted of clones?

Our contribution is twofold:

- We generalize and formalize the models previously studied, and we prove that the active recombination hotspot alleles ineluctably disappear in such models.
- We propose a more realistic model, which takes into account the location of the individuals, and which sustains better the active recombination hotspot alleles.

We summarize in Section 2 the results from Boulton, Myers and Redfield [1] and Pineda-Krch and Redfield [10]. We formalize and study their model in Section 3, producing a theoretical confirmation of

the results of their simulations. We deduce from this analysis that one of the key feature missing from their simulation is *spatiality*, and we introduce a model taking into account this feature in Section 3.5: our simulation results show that the active recombination hotspot alleles resist better the propagation of inactive ones in this model. In Section 5 we criticize our own model and discuss further work.

2 Previous Work

Boulton, Myers and Redfield [1] first identified the potential problem with the model of meiotic recombination hotspots. They illustrated it by the loss of active recombination hotspot alleles in a simple model where each generation begins with a pool of haploid gametes, each containing as its genome a single chromosome with one recombination hotspot and two side alleles. Their results concern the impact of two known positive effects of crossover: the positive effect of the chromosome recombination on their migration during the cell division (without recombination some gametes are sterile with probability .5), and the recovery of side alleles lost to mutations. Their simulations show that none of those effects is strong enough to sustain a positive proportion of active recombination hotspot alleles. Including the presence of active recombination hotspot alleles in the fitness function could maintains this proportion, but they argue that the evolutionary cost of such a mechanism would be too high.

Pineda-Krch and Redfield [10] explored the impact of other positive effects of recombination on the survival of active recombination hotspot alleles in a similar model, where each chromosome contains 10 recombination hotspot locations and 11 side alleles. As Boulton *et al.*'s, their simulation show that the impact of proper segregation and mutation recovery due to the recombination initiated by an active recombination hotspot allele, but also consider the impact of other features, such as

- the fertility selection in a population where most fitness is not yet optimal and can still evolve;
- the back mutation, where an inactive recombination hotspot becomes active due to a mutation;
- different probability of conversion;
- the various levels of activity of active recombination hotspot alleles, in particular depending on other recombination hotspot alleles on the same chromosome;

As Boulton *et al.*, they found that none of those features were strong enough to insure the persistence of active recombination hotspot alleles in the population.

Without describing the details of each models, a key observation is that in all the models studied previously the haploid gametes combined through reproduction are chosen *uniformly at random* in the whole population, after which they pass various viability tests and recombination operations. In real life, it would correspond to constantly shuffling the population, such that all mating pairs occur with the same probability. We show in the next section that under those conditions the active recombination hotspot alleles cannot persist in the population as soon as some inactive recombination hotspot alleles appear.

3 Analysis

3.1 Our theoretical model

We consider a pool of freely interacting haploids. Each haploid consists of a single chromosome, with a single recombination hotspot, and is noted H_i, of fitness f_i, which corresponds to its ability to grab resources. The population size is variable, regulated by the amount N of resources (e.g. food) available. At each step of the process: with probability .5 one haploid H_i is chosen uniformly at random, of fitness f_i, and duplicates with probability $\frac{f_i}{\sum f_i}\frac{N}{2}$, otherwise dies. Otherwise, with the remaining probability .5, two haploids H_i and H_j are chosen uniformly at random, of fitness f_i and f_j. They combine and try to reproduce with probability $\frac{f_i+f_j}{\sum f_i}\frac{N}{4}$, and otherwise die. If both recombination hotspot alleles are inactive, the haploids are unlivable with

probability .5 (because of improper segregation), or duplicates of the original haploids with the remaining probability.

If both recombination hotspot alleles are active, the haploids combine by a crossover at this site, and the four haploids produced have active recombination hotspot alleles. If only one recombination hotspot allele is active, the chromosomes combine by a crossover at this site, but three out of four haploids produced have inactive recombination hotspot alleles, and the one with an active recombination hotspot is a duplicate of the original chromosome whose recombination hotspot allele was active.

Note that for simplicity, we assume that the active recombination hotspot alleles *always* break: in a model with only one recombination hotspot allele this doesn't change the result and simplifies the analysis.

3.2 Evolution of the Population Size

In this model the size S of the population is stable on average. At each step of the process, the probability that the haploids chosen duplicate and increase the population is proportional to the share of resources they get. This share is proportional to their relative fitness, but inversely proportional to the size of the population. Hence the size of the population decreases when it is too large, and increases when it is too small: Lemma 1 proves this formally.

Lemma 1 *The average population size converges to N.*

Proof: First consider the case when a single haploid is chosen for duplication. Let be S the size of the population, ΔS the variation of this size in the first phase of one step, $E(\Delta S|\text{clone})$ the average variation when a single haploid is chosen for cloning, and $E(\Delta S|H_i)$ the average variation when the haploid H_i has been chosen:

$$\begin{aligned} E(\Delta S|H_i) &= +1\left[\frac{f_i}{\sum_l f_l}\frac{N}{2}\right] - 1\left[1 - \frac{f_i}{\sum_l f_l}\frac{N}{2}\right] \\ &= \frac{f_i}{\sum_l f_l}N - 1 \end{aligned} \qquad \begin{aligned} E(\Delta S|\text{clone}) &= \sum_i E(\Delta S|H_i)\frac{1}{S} \\ &= \frac{N}{S} - 1 \end{aligned}$$

Note that ΔS is equal to 1 or -1, and that $E(\Delta S|\text{clone})$ is positive if and only if S is smaller than N.

Now consider the case when two haploids are chosen for breeding. Let be ΔS the average variation of the size of the population in this phase, $E(\Delta S|\text{breed})$ the average variation when a couple of haploids is chosen for breeding, and $\Delta S|H_i, H_j$ the average variation when the haploids H_i and H_j have been chosen:

$$\begin{aligned} E(\Delta S|H_i,H_j) &= +2\left[\frac{f_i+f_j}{\sum_l f_l}\frac{N}{4}\right] \\ &\quad -2\left[1 - \frac{f_i+f_j}{\sum_l f_l}\frac{N}{4}\right] \\ &= \frac{f_i+f_j}{\sum_l f_l}N - 2 \end{aligned} \qquad \begin{aligned} E(\Delta S|\text{breed}) &= \sum_{i,j,i\neq j} E(\Delta S|H_i,H_j)\frac{1}{S(S-1)} \\ &= \frac{\sum_{i,j}(f_i+f_j) - 2\sum_i f_i}{\sum_l f_l}\frac{N}{S(S-1)} - 2 \\ &= (2S-2)\frac{\sum_i f_i}{\sum_l f_l}\frac{N}{S(S-1)} - 2 \\ &= 2\left(\frac{N}{S} - 1\right) \end{aligned}$$

Note that here ΔS is equal to 2 or -2, and $E(\Delta S|\text{breed})$ is positive if and only if S is smaller than N.

Considering the two cases, the average variation of the size of the population is positive if and only if S is smaller than N. On the other hand this variation is always bounded. By analogy with a Markov chain describing ΔS [4] the average size of the population $E(S)$ converges in the stationary distribution to the amount N of resources available. □

3.3 The Effect of Segregation

The segregation produces aneuploids (inviable gametes) with higher probability when no crossover occurs. As this happens more often when there are fewer active recombination hotspot alleles, Boulton *et al.* [1] studied the effects of this property. Their simulations show that this force is not sufficient to maintain a positive proportion of active recombination hotspot alleles. They use a model with a single chromosome per individual, and a single recombination hotspot per chromosome. We give the theoretical analysis of a broader class of models, which shows that the active recombination hotspot alleles indeed disappear in the stationary distribution.

Theorem 1 *Starting from an optimal population, active recombination hotspot alleles disappear in the stationary distribution.*

Proof: As the fitness is uniform, the duplication of haploids doesn't modify on average the repartition of active recombination hotspot alleles, only the sexual reproduction does. Let be p the proportion of active recombination hotspot alleles in the population. As a pair of haploids with inactive recombination hotspot alleles can product 4 haploids or nothing, 3 types of pairs can be formed with 4 possible outcomes:

1. $[AA \to AAAA]$ if the two chosen haploids have active recombination hotspot alleles;
2. $[AN \to ANNN]$ if exactly one chosen haploid has an active recombination hotspot allele;
3. $[NN \to NNNN]$ if none of the chosen haploids has an active recombination hotspot allele, and the gametes properly segregate;
4. $[NN \to \emptyset]$ if none of the chosen haploids have an active recombination hotspot allele, and the gametes does not properly segregate.

In each case we study the variation in the number of active recombination hotspot alleles and in the size of the population, expressed by the variation in the proportion of active recombination hotspot alleles.

1. The first case happens with probability $[p^2]$, and the proportion p then increases by $\frac{4-4p}{S+4}$.
2. The second case happens with probability $[2p(1-p)]$, and the proportion p then increases by $\frac{1-4p}{S+4}$.
3. The third case happens with probability $[(1-p)^2\frac{1}{2}]$, and the proportion p then increases by $\frac{4p}{S+4}$.
4. The fourth case happens with probability $[(1-p)^2\frac{1}{2}]$, and the proportion then does not change.

The expression of the average variation of p is then expressed as the sum of the variations in each case, weighted by their probabilities:

$$\begin{aligned} E(\Delta p) &= \frac{1}{S+2}(2-2p)[p^2] + (-4p)\left[2p(1-p) + (1-p)^2\frac{1}{2}\right] + \frac{1}{S+2}p(1-p)^2 \\ &= \frac{1}{S+2}p(1-p)(-2)(2p+1) + \frac{1}{S+2}p(1-p)^2 \end{aligned}$$

As long as $p \in (0,1)$ this is positive if and only if

$$\begin{aligned} -2(2p+1)(S-2) + (1-p)(S+2) > 0 \qquad &\Leftrightarrow \qquad 6 + 6p - S - 5pS > 0 \\ &\Leftrightarrow \quad p < \tfrac{S-6}{6-5S} \end{aligned}$$

For any value of S larger than 6, the fraction $\frac{S-6}{6-5S}$ is negative. This condition is fulfilled whenever N is sufficiently large (see Lemma 1).

So as long as N is large enough, $E(\Delta p)$ is never positive and p is always decreasing to 0 on average. □

3.4 The Effect of Selection

The active recombination hotspot alleles also have an evolutionary utility: without them, the population evolve only by mutation. A sub-population with active recombination hotspot alleles should have an evolutionary advantage.

Boulton *et al.*'s simulation [1] is starting with a population of optimal individuals, and mutations introduce sub-optimal individuals. In such a model, with a small mutation rate, only a few sub-optimal individuals are generated at each generation, and there is no need of recombination to obtain a better population, as selection just suppresses sub-optimal individuals. On the other hand, with a mutation rate large enough to disrupt the optimality of the population, the possible benefits from recombination are most likely disrupted by a mutation.

Pineda-Krch and Redfield's model [10] starts with a sub-optimal initial population, where recombinations are much more likely to generate better individuals, and the active recombination hotspot alleles triggering these recombinations are more likely to be promoted. To obtain an upper bound on the proportion of active recombination hotspot alleles sustained in such a model, it is sufficient to study a much simpler model where the offsprings obtained by recombination are always better than their parents: this model's unrealism can only increase the proportion of active recombination hotspot alleles. In such a model, the proportion of active recombination hotspot alleles in the stationary distribution is still null: hence even the evolutionary role of the active recombination hotspot alleles does not permit to sustain them.

Theorem 2 *Even in the model where new individuals are always better than their parents, the active recombination hotspot alleles do not persist.*

Proof:

As before, p is the proportion of active recombination hotspot alleles in the population. For a random individual chosen in the population, let be A the event that it has an active recombination hotspot allele, N the event that it has an inactive recombination hotspot allele, and new the event that this individual is different from its parents.

Haploids with a new combination of alleles (event `new`) and active recombination hotspot alleles (event A) can be generated only by breeding two haploids which both have active recombination hotspot alleles. Haploids with a new combination of alleles and inactive recombination hotspot alleles (event N) can be generated only by breeding one haploid, which has an active recombination hotspot allele, with an haploid which has an inactive recombination hotspot allele. In each case, two such haploids are generated.

From the probabilities of those events, we can deduce the probability of generating a new haploid, and the probability that such a haploid has an active recombination hotspot allele:

$$\Pr\{A \wedge \texttt{new}\} = \frac{p^2}{2} \qquad \Pr\{\texttt{new}\} = p(1-\frac{p}{2})$$

$$\Pr\{N \wedge \texttt{new}\} = \frac{2p(1-p)}{2} \qquad \Pr\{A|\texttt{new}\} = \frac{\Pr\{A \wedge \texttt{new}\}}{\Pr\{\texttt{new}\}} = \frac{2}{2-p} - 1$$

This number, which corresponds to the probability that a new haploid has an active recombination hotspot allele, is smaller than p, the proportion of individuals with an active recombination hotspot in the whole population. $\forall p \in [0,1]$:

$$\frac{2}{2-p} - 1 < p \Leftrightarrow 2 < (p+1)(2-p) \Leftrightarrow p(1-p) < 0 \Leftrightarrow p \in [0,1]$$

Hence the proportion of active recombination hotspot alleles among new (and potentially better) solutions is *smaller* than the proportion of active recombination hotspot alleles among the old solutions. So the active recombination hotspot alleles do not persist. □

3.5 Deduction from the analysis

When analyzing the models previously described, one fact is striking: active recombination hotspot alleles do not get a chance to "survive" because the probability for a gamete with an active recombination hotspot allele to mate with a gamete having an inactive recombination hotspot allele is strictly increasing with the number of individuals with inactive recombination hotspot alleles. This is a direct consequence of the uniform choice of individuals at each step of the simulation, and is quite artificial: in practice it would correspond to a constant shuffling of the population, like constantly turning a spoon in the becker containing the population.

In practice, individuals interact mainly with their direct neighbors. Offsprings can move farther from the parents, but often to a limited distance which means that siblings often mate with each other. This slows down evolution, but it also provides the opportunity for sub-populations to appear. Such sub-population would enable individuals with active recombination hotspot alleles to mate with higher probability with each other, to generate offsprings at once better and with active recombination hotspot alleles, providing an advantage over an antagonist sub-population of individuals with inactive recombination hotspot alleles, and unable to generate any new genotypes.

In the following section, we define and simulate a spatial model, which maintains a positive proportion of active recombination hotspot alleles in the population.

4 Simulations

We propose a *spatial* model, similar to *island models* used in zoology for the study of segregation and other evolution phenomenon involving sub-populations. To make the recombination more interesting from the evolution point of view, we consider several recombination hotspot locations per chromosome (as Pineda-Krch *et al.* [10]), a fitness which is exponential in the number of side alleles agreeing with a target, and regular perturbations of the target of the fitness function.

4.1 Control Experiments

We first controlled that in the context of a unique location (i.e. without spatiality), the active recombination hotspot alleles still disappear in our simulation, even when considering the features we added to emphasize the importance of recombination.

Figure 2 shows the results of the simulation with a single recombination hotspot allele, and Figure 3 corresponds to the same simulation but with ten recombination hotspot alleles. Each figure displays the results of five independent runs with an initial population of size 1000, a single location, 100 reproduction steps per generation, where the target of the fitness is randomly renewed every 200 generations. Each recombination hotspot allele in the initial combination is active with probability 0.75, and each side allele is positive with probability inverse to the number of side alleles, so that on average each chromosome of the initial population has one active side allele.

In both case the size of the population is tightly oscillating around its average value, 1000 (it is displayed only once). The fitness of the population converges quite fast to its optimal value (respectively 2^2 and 2^10), but in both cases the number of active recombination hotspot alleles is decreasing fast, and almost reaches zero by generation 1000, for almost all recombination hotspot alleles and in almost all runs.

Those results confirm that, although our population model is slightly different from the ones from Boulton, Myers and Redfield [1] and Pineda-Krch and Redfield [10] because it allows a dynamic population size, which is necessary to avoid introducing global rules in the spatial model, we did not introduce any feature which advantage the active recombination hotspot alleles.

4.2 A basic spatial model

We extend the model described in Section 3.1 by considering a set of islands on a grid. Each individual is associated with a *location*, and can reproduce only with an individual sharing the same location. At each generation, an arbitrary number of individuals and locations are chosen.

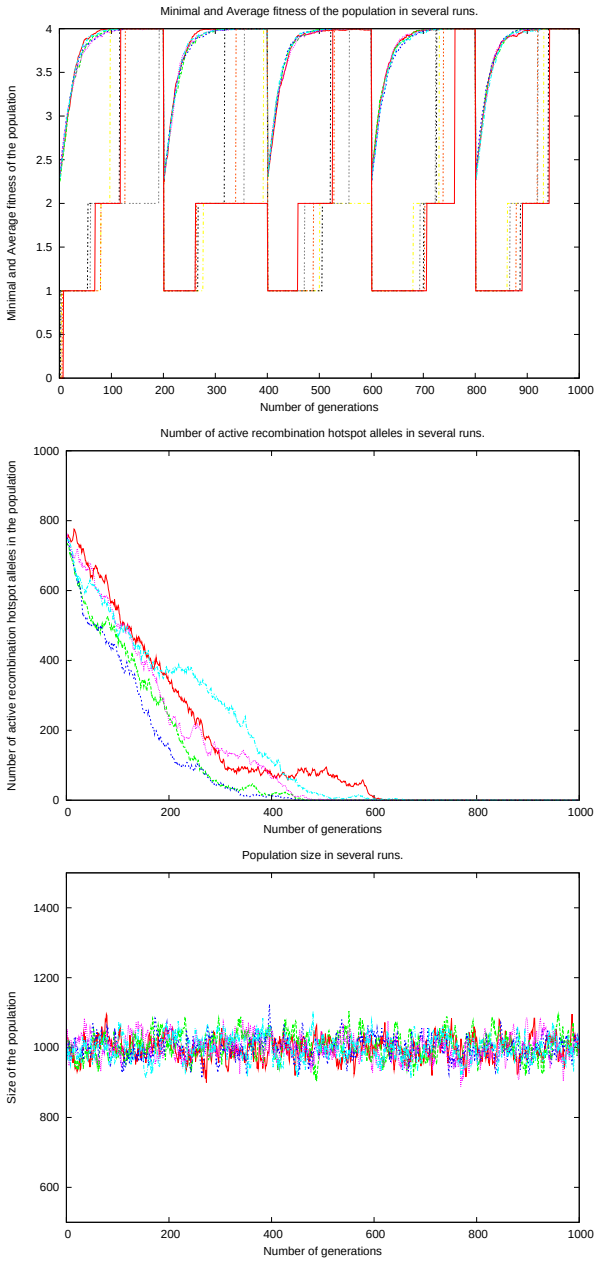

Figure 2: Control experiment with one recombination hotspot location.

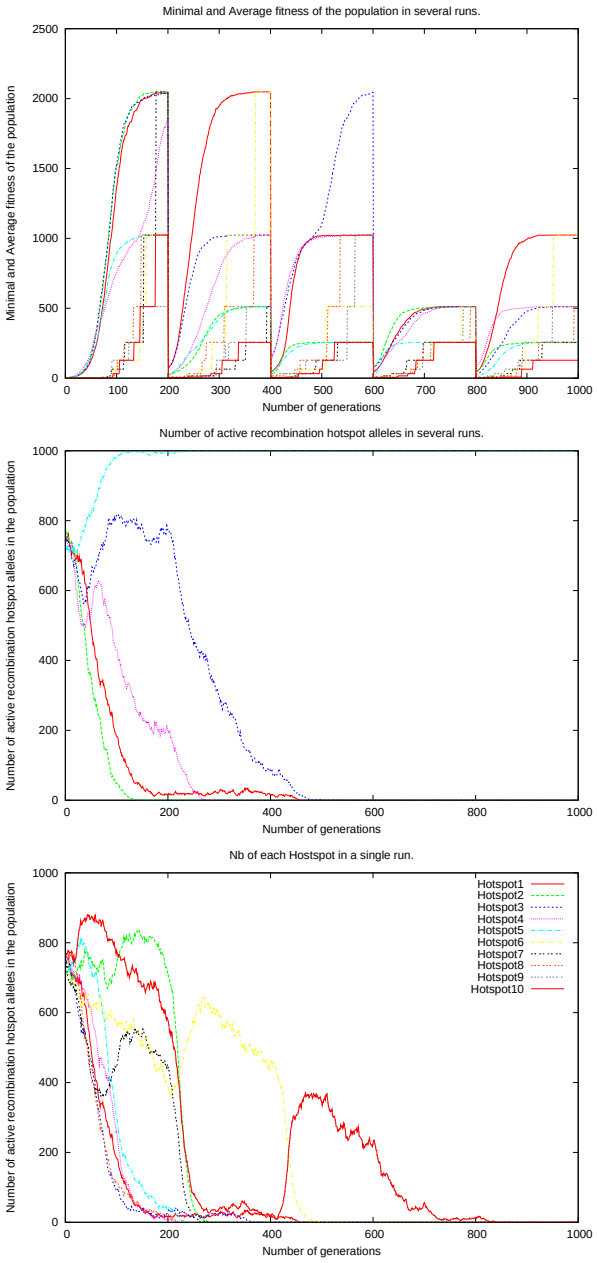

Figure 3: Control experiment with ten recombination hotspot locations.

- For each individual chosen, a second individual is chosen uniformly at random among the individuals sharing the same location, to reproduce. The proportion of resources gathered by the pair of individual, proportional to their fitness, defines the probability that they can combine in a diploid and eventually produce offsprings. The pair of individuals disappear if they fail to reproduce.

- For each location chosen, a random neighboring location is chosen uniformly at random, and the difference g in population size is computed between both locations. A random number between zero and g of randomly chosen individuals is moved from the location with the population of largest size to the other.

Similarly to the control experiments, Figure 4 show the results of the simulation with a single recombination hotspot allele, and Figure 5 correspond to the same simulation but with ten recombination hotspot alleles. Each figure displays the results of five independent runs with an initial population of size 1000, 100 reproduction steps per generation, a grid of 10 but 10 locations, with one migration step per generation.

When each chromosome holds only one recombination hotspot (Figure 4), spatiality does not change much the results, and active recombination hotspot alleles disappear as fast as without spatiality. On the contrary, in the simulation where each chromosome holds ten recombination hotspot locations (Figure 5), the active recombination hotspot alleles survive much longer and sometime invade the whole population. Also, the average fitness of the population is much better than in the simulations without spatiality: because of the large proportion (0.75) of active recombination hotspot alleles in the original population, some inactive recombination hotspot alleles are totally absent from some locations, where a sub-population can optimize the fitness more efficiently and later invade other locations.

Note that active recombination hotspot alleles which do not invade the whole population ultimately disappear, and that in presence of mutations deactivating recombination hotspot alleles they would all disappear: the model is not sustaining indefinitely active recombination hotspot alleles, just supporting them longer.

5 Conclusions

In this paper we provided a theoretical confirmation of some experimental results on a model of the diffusion of inactive recombination hotspot alleles, and we gave an analysis proving that even a more general model does not solve the paradox observed. The models discussed in the literature neglect many properties of natural systems, among which one is spatiality. The models neglecting this property are analogous to the situation where the haploids are continuously shuffled: this happens only rarely in nature, where the haploids can form colonies. Taking into account this spatiality, along with other features which emphasize the importance of recombination to produce individuals of higher fitness, proved to sustain active recombination hotspot alleles longer.

The importance of spatiality is not a surprise, as it proved salutary in other models, in particular for strategies from game theory otherwise believed to disappear: Nowak and May [11] showed that in the Prisoner's Dilemma problem, the strategy always cooperating, usually believed to disappear in competition with defecting, can sustain itself indefinitely. Similar dynamics has been studied by Killigback and Doebeli [9], observing some self-organized criticality behavior.

The paradox is still not fully solved: spatiality does not sustain the active recombination hotspot alleles indefinitely in the simulations, and preliminary experiments regarding the impact of mutations of the recombination hotspot alleles were not conclusive. Indeed, the inability of spatiality to sustain active recombination hotspot alleles indefinitely reflects a slightly different paradox: if active recombination hotspot alleles are useful in populations where they are already dominant, how did such alleles appear to begin with, in populations of organisms practicing cloning rather than sexual reproduction?

Acknowledgments: This work benefited enormously from discussion with Mario Pineda-Krch, Sally Otto, Rosemary J. Redfield and Annie Lee.

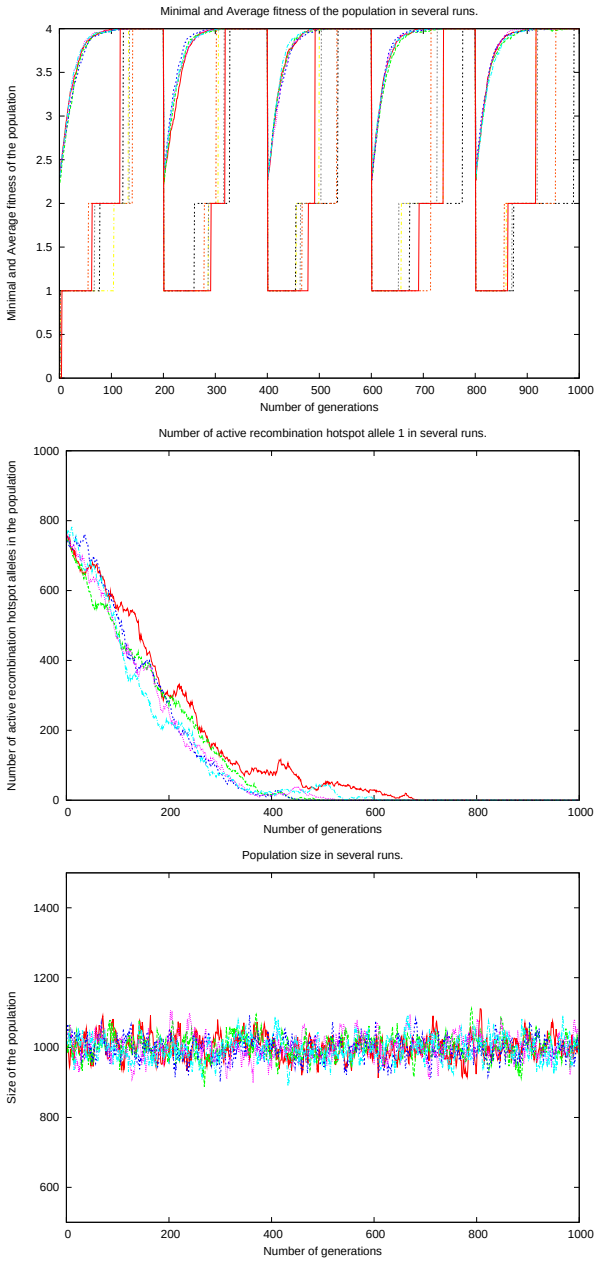

Figure 4: Spatial experiment with one recombination hotspot location.

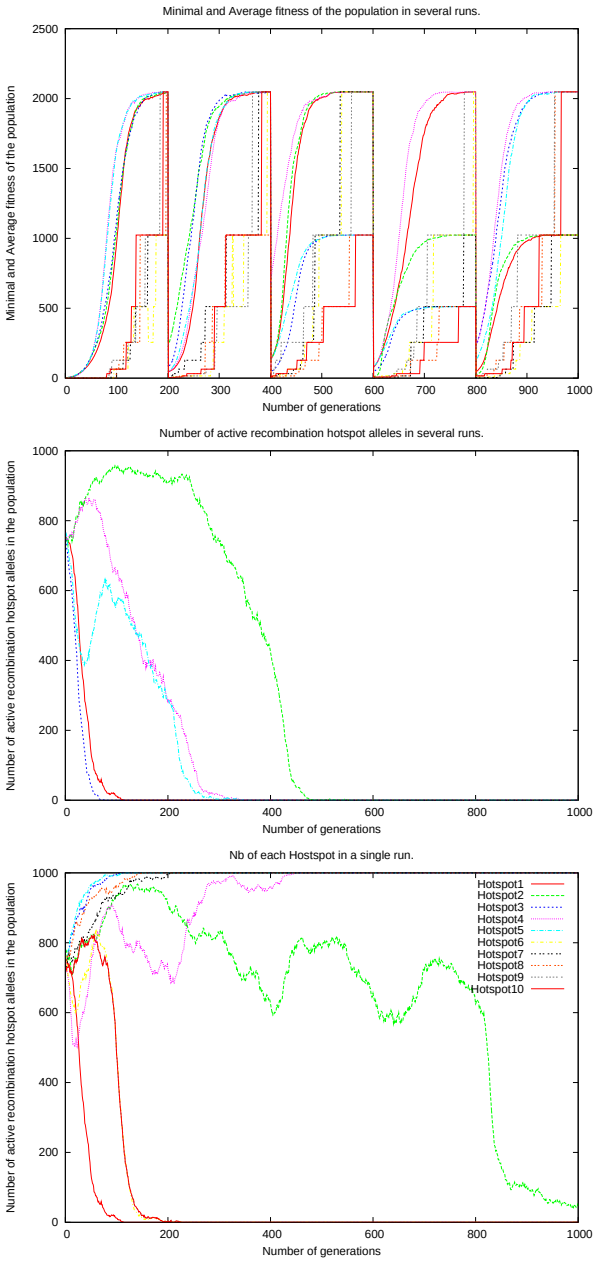

Figure 5: Spatial experiment with ten recombination hotspot locations.

References

[1] A. Boulton, R. S. Myers, and R. J. Redfield. The hotspot conversion paradox and the evolution of meiotic recombination. In *Proc. Natl. Acad. Sci. USA*, volume 94, pages 8058–8063, July 1997.

[2] G. Coop. Can a genome change its (hot)spots? *Trends in Ecology and Evolution*, 20(12):643–645, 2005.

[3] de Massy. Distribution of meiotic recombination sites. *Trends Genet*, 19:514–522, 2003.

[4] G. Fayolle, V. Malyshev, and M. Menshikov. *Topics in the constructive theory of countable Markov chains*. Cambridge [England]; New York, NY : Cambridge University Press, 1995.

[5] A. W. Murray and J. W. Szostak. *Annual Review of Cell Biology*, 1:189–315, 1985.

[6] N. MW. Variations in recombination rate across the genome: Evidence and implications. *Cur Opin Genet Dev*, 12:657–663, 2002.

[7] A. N, C. P, and N. M. Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. *Am J Hum Genet*, 73:5–16, 2003.

[8] R. Neumann and A. J. Jeffreys. Polymorphism in the activity of human crossover hotspots independent of local dna sequence variation. *Human Molecular Genetics*, 15(9):1401–1411, 2006.

[9] M. A. Nowak and R. M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, 1992.

[10] R. R. Pineda-Krch M. Persistence and loss of meiotic recombination hotspots. *Genetics*, pages 169–2319, 2005.

[11] K. T. and D. M. Self-organized criticality in spatial evolutionary game theory. *Theor. Biol.*, 191(3):335–340, Apr 1998.