Error catastrophe in populations under similarity-essential recombination

Marcus A.M. de Aguiar[1*], David M. Schneider[1], Eduardo do Carmo[2], Paulo R. A. Campos[3], Ayana B. Martins[1,4]

[1]*Instituto de Física 'Gleb Wataghin', Universidade Estadual de Campinas, 13083-970, Campinas, SP, Brazil*

[2]*Universidade Federal da Integração Latino Americana 85867-970, Foz do Iguaçu, PR, Brazil*

[3]*Departamento de Física, Universidade Federal de Pernambuco, 50670-901, Recife, PE, Brazil*

[4]*Instituto de Biociências, Universidade de São Paulo, 05508-090, São Paulo, SP, Brazil*

Abstract

Organisms are often more likely to exchange genetic information with others that are similar to themselves. One of the most widely accepted mechanisms of RNA virus recombination requires substantial sequence similarity between the parental RNAs and is termed similarity-essential recombination. This mechanism may be considered analogous to assortative mating, an important form of non-random mating that can be found in animals and plants. Here we study the dynamics of haplotype frequencies in populations evolving under similarity-essential recombination. Haplotypes are represented by a genome of B biallelic loci and the Hamming distance between individuals is used as a criterion for recombination. We derive the evolution equations for the haplotype frequencies assuming that recombination does not occur if the genetic distance is larger than a critical value G and that mutation occurs at a rate μ per locus. Additionally, uniform crossover is considered. Although no fitness is directly associated to the haplotypes, we show that frequency-dependent selection emerges dynamically and governs the haplotype distribution. A critical mutation rate μ_c can be identified as the error threshold transition, beyond which this selective information cannot be stored. For $\mu < \mu_c$ the distribution consists of a dominant sequence surrounded by a cloud of closely related sequences, characterizing a quasispecies. For $\mu > \mu_c$ the distribution becomes uniform, with all haplotypes having the same frequency. In the case of extreme assortativeness, where individuals only recombine with others identical to themselves ($G = 0$), the error threshold results $\mu_c = 1/4$, independently of the genome size. For weak assortativity ($G = B - 1$) $\mu_c = 2^{-(B+1)}$ and for the case of no assortativity ($G = B$) $\mu_c = 0$. We compute the mutation threshold for $0 < G < B$ and show that, for large B, it depends only on the ratio G/B. We discuss the consequences of these results for recombination in viruses and for speciation.

*Corresponding author: aguiar@ifi.unicamp.br

Keywords: haplotype frequencies, frequency-dependent selection, quasispecies theory, neutral models.

I. INTRODUCTION

Mate choice by phenotypic similarity, or assortative mating, is a form of non-random mating that plays important roles in evolution and speciation. The mechanism has been explored in several mathematical and computational models and is often implemented as occurring in a Mendelian trait determined by a single locus [1–3] or in a single quantitative trait determined by multiple loci with additive effects [4–8].

Mate choice, however, often relies on multiple cues that are determined genetically [9–12], such that the effect of the state of these traits on phenotypic similarity is additive [13–15]. In the past twenty years different models have described this type of assortative mating [16–22], which may also be interpreted as a multilocus generalization of the Bateson-Dobzhansky-Muller model of intrinsic genetic incompatibilities [23].

Interestingly, a form of assortativity also occurs in virus populations. Contrary to what was initially thought, recombination is now considered to be a general phenomenon in RNA viruses and might play a major role as a driving force in virus variability and evolution [24]. Although the mechanisms of viral RNA recombination are only now beginning to be elucidated, in the most widely accepted mechanism of viral recombination the enzyme responsible for replication switches from one sequence to another during the synthesis generating a recombinant genome [25, 26]. This sequence switch is known to be dependent of the extent of similarity between the recombining genomes [27–29] and referred to as similarity-essential recombination [30]. Although there is strong evidence that the genetic exchange promoted by recombination can offer advantages, random recombination destroys more good alleles than it creates, leading to a selective pressure towards close similarity in the process [31].

In this paper we develop a theory for the evolution of recombinant haplotypes subjected to point mutations and similarity-essential recombination, but no other selective pressures. Each sequence will be represented by a binary string of length B and we will assume that sequences differing in more than G loci do not recombine. This mode of recombination is analogous to the assortative mating that often appears in models of population genetics, but may also represent replication of RNA viruses as described above. We will write the evolution equations for the haplotype frequencies and find their equilibrium solutions. For the case of zero mutation probability we will show that the population evolves to an equilibrium quasispecies where G loci are polymorphic and all the remaining $B - G$ loci get fixed. This

equilibrium configuration is equivalent to that of a population with only G loci evolving under unconstrained recombination. However, which loci will remain polymorphic depends on the initial conditions. For non-zero mutation probability the scenario is more subtle. Our results show that, at least qualitatively, the population under similarity-essential recombination behaves in a way similar to the original quasispecies model for replicating macromolecules [32].

The quasispecies theory was originally developed by Eigen and Schuster to study the evolution of prebiotic RNA molecules exploring the consequences of the mutation-selection dynamics in near-infinite populations. Mutation rates are thought to have been much higher in the early history of life. As a result, equilibrium populations can be described as a a distribution of related genotypes known as quasispecies [32, 33]. Later the theory has successfully been applied to the study of viral evolution, especially RNA viruses [34–36].

One of the main results of the quasispecies theory is the existence of a mutation rate above which selection cannot overcome the mutation load (*i.e.* the error threshold). The amount of information that can be encoded in such evolutionary systems is limited by the genome length, since longer sequences suffer from mutations more than shorter ones. This leads to a logical enigma called Eigen's paradox [37, 38]: given the mutations rates of this prebiotic scenario, these early genomes would not be long enough to encode the enzymes required to increase replication accuracy [32, 33]. Different mechanisms have been proposed to overcome or alleviate the genome size constraint imposed by the error threshold and warrant stable integration of information contained in the self-replicative units, like the theory of hypercycles [32], group selection models [39] and models incorporating recombination [40–42], and/or more complex genotype-phenotype mapping [43–45].

Here we show that, as in the quasispecies theory, recombinant haplotypes evolving under similarity-essential recombination exhibit two equilibrium regimes separated by a critical mutation rate μ_c. In the first regime, which takes place for $\mu < \mu_c$, a dominant haplotype coexists with a cloud of closely related haplotypes. In the second regime, there is an information crisis and an uniform distribution of haplotypes is obtained. Depending on the degree of assortativity, described by the parameter G, the error threshold can be as high as $\mu_c = 1/4$ and independent of the genome size, or as low as $\mu_c = 2^{-(B+1)}$, in contrast with the $1/B$ behavior obtained in the original quasispecies model. We compute the mutation threshold for all values of G and show that, for large B, it depends only on the ratio G/B.

II. MATERIALS AND METHODS

We consider a population of haploid individuals with B biallelic loci. The genome of each individual is represented by a string of B binary digits

$$i = i_1 i_2 \dots i_B \tag{1}$$

where the alleles i_k are either 0 or 1. We introduce the genotypic distance between two haplotypes as the number of different alleles between them:

$$d(i,j) = \sum_{k=1}^{B} |i_k - j_k|. \tag{2}$$

Similarity-essential recombination corresponds to forbidding mating if $d(i,j) > G$, where $G \leq B$.

A. Unconstrained recombination ($G = B$)

Using the compact notation p_i for the frequency of haplotype $i = i_1 i_2 \dots i_B$, the equation determining the frequency p_i^{t+1} in terms of the frequencies at time t assumes the form

$$p_i^{t+1} = \sum_{j,k} c_\mu(j,k;i) p_j^t \, p_k^t \tag{3}$$

where $c_\mu(j,k;i)$ is the probability that individuals with haplotypes j and k produce a recombinant haplotype i if the mutation rate is μ, whereas $p_j^t \, p_k^t$ is the probability of an encounter of haplotypes j and k at time t. To determine these coefficients we assume independent segregation (uniform crossover) and look at one locus at a time. For a given allele i_n there are three possibilities for j_n and k_n:

(a) $j_n = k_n \neq i_n$.
In this case the allele transmitted to the recombinant sequence is $(1 - i_n)$ and it contributes to p_i only if it mutates to i_n. Therefore it contributes a factor μ to the probability. We call α the number of loci satisfying this condition:

$$\alpha = \sum_{n=1}^{B} [1 - |j_n - k_n|] \; |i_n - j_n|. \tag{4}$$

(b) $j_n = k_n = i_n$.

The allele transmitted is i_n if it does not mutate. It contributes a factor $(1-\mu)$ and the number of loci in this case is β:

$$\beta = \sum_{n=1}^{B} [1 - |j_n - k_n|]\ [1 - |i_n - j_n|] \tag{5}$$

(c) $j_n \neq k_n$.

The allele transmitted is either i_n or $1 - i_n$. It contributes a factor $\frac{1}{2}(1-\mu) + \frac{1}{2}\mu = \frac{1}{2}$. The number of loci of this type is

$$\gamma = \sum_{n=1}^{B} |j_n - k_n| = d(j,k). \tag{6}$$

It can be checked that $\alpha + \beta + \gamma = B$ and that

$$\alpha = \frac{d(i,j) + d(i,k) - d(j,k)}{2}. \tag{7}$$

With these considerations equation (3) becomes

$$p_i^{t+1} = \sum_{j,k} p_j^t\ p_k^t\ (1-\mu)^{B-\alpha-\gamma} \mu^\alpha \left(\frac{1}{2}\right)^\gamma. \tag{8}$$

It is interesting to report the two limiting cases $\mu = 0$ and $\mu = 1/2$. In the first case only haplotypes with $\alpha = 0$ contribute to offspring and $d(j,k) = d(i,j) + d(i,k)$, showing that the sum of the genetic distances from the recombinant sequence to each original sequence is the genetic distance between the parents. If $\mu = 1/2$, corresponding to maximum randomness, each pair of parental genomes contribute equally with weight 2^{-B}.

The normalization condition is

$$\sum_i c_\mu(j,k;i) = \sum_i (1-\mu)^{B-\alpha-\gamma} \mu^\alpha \left(\frac{1}{2}\right)^\gamma = 1 \tag{9}$$

and can be easily verified explicitly (see Electronic Supplementary Material, section I).

B. Similarity-essential recombination ($G \neq B$)

When genomes whose alleles differ in more than G loci are considered incompatible for recombination, equations (8) have to be modified. In this case the sums over j and k on the right hand side have to be restricted to parental sequences with $\gamma = d(j,k) \leq G$ (see eq.(2))

and several terms are removed from the sum. Consequently, the equation has to be modified in order to satisfy the normalization condition $\sum_i p_i = 1$. Normalization is ensured with the introduction of an auxiliary function Φ such that

$$p_i^{t+1} = \sum_{j,k,\gamma \leq G} p_j^t \; p_k^t \; c_\mu(j,k;i) - p_i^t(\Phi - 1). \tag{10}$$

Summing over i on both sides and using that $\sum_i c_\mu(j,k;i) = \sum_i p_i = 1$ we find

$$\Phi = \sum_{j,k,\gamma \leq G} p_j^t \; p_k^t. \tag{11}$$

C. Analogy with the quasispecies theory and the error catastrophe

The quasispecies theory is originally a theory of molecular evolution [32]. In the Eigen model molecules are represented by binary sequences of length L and the concentrations x_i of each type follow the equation

$$\frac{dx_i}{dt} = \sum_j x_j f_j q_{ji} - x_i \phi, \tag{12}$$

which assumes that the molecules replicate by cloning with mutations. In Eq. (12) f_j refers to the replication rate (hereafter referred to as fitness) and the element of the mutation matrix $q_{ji} = (1-\mu)^{B-d(i,j)}\mu^{d(i,j)}$ gives the probability that a molecule of type i produces a molecule of type j if the mutation probability per digit is μ. If the master string 00...0 has fitness $f_0 > 1$ and all the remaining ones have fitness $f_i = 1$, it can be shown that a cloud of mutant sequences surrounding and including the fittest master sequence (wild type) settles in the population if

$$\mu < \frac{\log f_0}{B} \equiv \mu_c. \tag{13}$$

Above the mutation threshold μ_c the population can no longer equilibrate in a mutation-selection balance and the selection information is lost (error catastrophe).

Equation (10), which assumes similarity-essential recombination and discrete time, can also be written in a similar form as the discrete time version of the Eigen's equation

$$p_i^{t+1} - p_i^t \equiv \sum_j p_j^t \; F_j \; Q_{ji} - p_i^t \Phi \tag{14}$$

with

$$Q_{ji} = \frac{\sum\limits_{d(k,j)\leq G} p_k^t c_\mu(j,k;i)}{\sum\limits_{d(k,j)\leq G} p_k^t} \tag{15}$$

and

$$F_i = \sum_{d(i,j)\leq G} p_j. \tag{16}$$

Q_{ji} is the average probability that a haplotype j produces i by recombining with all compatible haplotypes k. At this point the definition of F_i as the fitness of individuals of type i comes about naturally, and is readily interpreted as the fraction of compatible individuals in the population. Note that $\Phi = \sum\limits_{j,k,\gamma\leq G} p_j^t\, p_k^t = \sum_j p_j F_j$ is, therefore, the average fitness of the population.

It is important to point out that the model assumes that all individuals are selectively equivalent regardless of their identities (neutral model). However, equation (14) demonstrates that frequency dependent selection arises from the similarity-essential recombination and can be quantified by equation (16). The fitness of an individual depends not on its specific haplotype but on the population composition. More importantly it is large if the individual is amongst compatible pairs (with which recombination is possible) and low if it is surrounded by incompatible mates. This idea concurs with the proper definition of quasispecies, at which natural selection is no longer directed toward a single variant but instead acts on the whole haplotype distribution [32].

III. RESULTS

A. Unconstrained recombination

When $G = B$ recombination is possible between every pair of individuals. In this case, since the alleles in each locus are segregated independently and there are no correlations between them, the result is that, for $\mu = 0$ the allele frequencies remain constant from the first generation and the haplotype frequencies asymptotically reach the linkage equilibrium. For $\mu \neq 0$ the haplotypes converge to the uniform distribution, and thereby all frequencies are equal to $p_i = 2^{-B}$. These results are well known [46, 47] and are demonstrated in the present context in the Electronic Supplementary Material, section II.

B. Extreme assortativeness

In the case of extreme assortativeness, $G = 0$, individuals only recombine with others identical to themselves. In this case $\gamma = d(j,k) = 0$, $\alpha = d(i,j)$, $F_j = p_j$ and $Q_{ij} = q_{ij}$. Equation (14) becomes

$$p_i^{t+1} - p_i^t \equiv \sum_j \left(p_j^t\right)^2 \; q_{ji} - p_i^t \Phi \tag{17}$$

which is identical to (12) with $F_i = p_i$. For $\mu = 0$ the equation simplifies to

$$p_i^{t+1} - p_i^t \equiv \left(p_i^t\right)^2 - p_i^t \Phi \tag{18}$$

and the only stationary solutions are:
(a) the single haplotype solution $p_{i_0} = 1$ and $p_i = 0$ for $i \neq i_0$ and;
(b) the uniform solution $p_i = 1/2^B$ for all i.

For small mutations a cloud of haplotypes similar to i_0 is generated. Which haplotype survives, along with its mutant cloud, is determined by the initial population [47–49]. As the mutation rate increases the cloud spreads and, at $\mu = \mu_c$, the uniform solution becomes stable. The mutation threshold can be calculated and results $\mu_c = 1/4$, independent of the size of the genome B (see Electronic Supplementary Material, section VI). This should be compared with equation (13), where the threshold becomes small as the genome size B increases. If we define an effective fitness F_0 as the fitness of the corresponding quasispecies, whose value is constant, that will result in the same error threshold, we find that $1/4 = \log(F_0)/B$ and so

$$F_0 = e^{B/4}. \tag{19}$$

The upper panel in figure 1 shows the frequencies of the haplotypes as a function of the mutation probability μ. The scenario displayed in the plot is essentially the pattern exhibited by the quasispecies model: a dominant haplotype surrounded by a cloud of closely related haplotypes.

Interestingly, for $G = 1$ both expressions (12) and (15) work, even though $Q_{ij} \neq q_{ij}$. The reason for this coincidence is that for $G = 1$ reproduction occurs effectively with a single locus and is equivalent to cloning one of the original haplotypes with equal probability. Recombination affects only genetic exchanges between individuals with $d(j,k) \geq 2$ [47]. Indeed, for $G \geq 2$ only (14) is true.

A proof that the uniform distribution is always a solution for any value of G is presented in the Electronic Supplementary Material, section V.

C. Error threshold for arbitrary G and B

Explicit stationary solutions of equations (10) or (14) are not known, except for $B = 2$ [47]. Because of the mating restriction imposed by the condition $d(i, j) \leq G$, the loci are not independent and the dynamics of allele and haplotype frequencies are more complex and richer. For zero mutation probability the haplotype frequencies converge to a distribution where $B - G$ loci are monomorphic (fixed in either 0 or 1) and the remaining G loci are polymorphic in linkage equilibrium. Which loci become polymorphic depends on the initial conditions and there are many possibilities. Indeed, for a population with this type of haplotype distribution all individuals have maximum fitness $F_i = 1$, since the genetic distance between any pair satisfies $d(i, j) \leq G$ (see equation (16)). Moreover, $\Phi = 1$ and equations (10) become identical to (8) with $B \to G$.

However, the introduction of a small mutation rate $\mu > 0$ generates mutants that decrease the fitness of all resident types, resulting in further dynamics that converges to a single dominant type plus a set of low frequency mutants. As μ increases the distribution widens and the uniform distribution, where $p_i = 2^{-B}$ for all haplotypes, eventually takes over. The mutation threshold for small values of G and for the limit case $G = B - 1$ are:

$$\begin{aligned}
\mu_c(B, G = 0) &= \frac{1}{4} \\
\mu_c(B, G = 1) &= \frac{(B-1)}{4B} \\
\mu_c(B, G = 2) &= \frac{(B-1)(B-2)}{4(B^2 - B + 2)} \\
\mu_c(B, G = 3) &= \frac{(B-1)(B-2)(B-3)}{4(B^3 - 3B^2 + 8B)} \\
\mu_c(B, G = 4) &= \frac{(B-1)(B-2)(B-3)(B-4)}{4(B^4 - 6B^3 + 23B^2 - 18B + 24)}
\end{aligned} \tag{20}$$

and

$$\mu_c(B, G = B - 1) = 2^{-(B+1)}. \tag{21}$$

A detailed discussion is presented the Electronic Supplementary Material, section VI. These results are in qualitative agreement with previous numerical simulations in similar systems [50, 51].

Figure 2 shows μ_c as a function of G/B for several values of B as calculated with the procedure indicated in the SI. For a fixed number of loci B the region under the corresponding curve indicates where a single haplotype or a set of closely related haplotypes dominates the population. It is interesting to note that this region is roughly independent of B and that it shrinks fast for $G/B > 1/2$. The more restrictive is the criterion for recombination (smaller G/B) the larger the interval of mutations leading to non-uniform distribution of alleles. This means that similarity-essential recombination turns the population less susceptible to the error-prone replication, i.e., the selective information can be kept for a broader range of values of mutation probabilities. This result is in contrast with the effects of standard recombination, which tends to lower the value of the critical mutation probability [49]. Figure 1 also shows how the distribution of haplotypes changes with μ. It is quite noticeable the shift of the error threshold to lower values as the assortativity G is reduced.

IV. DISCUSSION

Here we studied the evolution of haplotype frequencies in an infinite population with similarity-essential recombination. We assumed that recombination is not possible if the genetic distance between two sequences is greater than a certain threshold G. In viral populations recombination often occurs when the replication enzyme switches from one molecule to another, and reducing G is equivalent to increasing the extent of similarity required for template switching. Depending on the similarity threshold the recombination is classified as precise or imprecise [24] and its value may be interpreted as a by-product of physical-chemical properties of the molecules involved in nucleic acid replication.

We derived the evolution equations for the haplotype frequencies assuming that each locus segregates independently (uniform crossover, as in [40, 41, 50]). This assumption is known not to be realistic. However, it is the simplest case to consider in order to highlight the effects of assortativity in the process of recombination. The correspondence between our equations and the quasispecies model allowed us to quantitatively determine the contribution of this type of recombination constraint to fitness, which we have shown to be equal to the

fraction of all compatible sequences in the population. It is important to highlight that, aside from the differences in fitness resulting from this constraint, our model is essentially neutral, since all individuals are assumed to be selectively identical. In spite of this neutrality at the individual level, natural selection arises as an outcome of the internal dynamics, which favors the selection of common haplotypes.

In the case of RNA viruses, most experimental studies are performed under strong selective pressures, so that only the higher fitness types are detected [24]. Adding intrinsic fitness to the haplotypes would be a natural, though non-trivial, extension of our work. Depending on the initial conditions and on the strength of selection, the dynamics could give rise to a competition between the higher fitness types and those starting with large fractions of compatible individuals, leading to interesting properties of the haplotype distributions.

One of our main results is the observation and calculation of the error catastrophe in a population under similarity-based recombination. We demonstrated that for small G/B the threshold tends to $1/4$, which is very large and independent of the genome size. At the other extreme, where $G = B - 1$ the error threshold decreases exponentially as $2^{-(B+1)}$. Both these behaviors are in contrast with the $1/B$ formula of the original Eigen model. The main consequence of the information crisis is the prediction of a maximum length for the sequence size beyond which the selective information is lost. In the context of prebiotic evolution, this limits our understanding about how complex molecular structures can emerge from the prebiotic scenario. Here we have shown that replicating units subjected to similarity-essential recombination are able to safely transmit information at higher mutations through the emergence of a stable distribution of closely related haplotypes. These haplotypes naturally arise from the dynamics without defining a priori the set with large fitness. Because the error threshold is large and independent of the genome size for strong assortativity there is no restriction on the amount of information that can be stored in the system. The framework developed here provides exact results on the mutation thresholds for any values of genome size and assortativeness that could be applied to these populations.

Taken in a broader sense, the quasispecies framework may describe the evolution of any population of reproducing organisms [40]. In this case, similarity-based recombination is equivalent to assortative mating. This form of non-random mating plays an important role in the reproductive behavior of many populations. In mathematical models it can be introduced by prohibiting mating between individuals whose phenotypes are too dissimilar.

Depending on how genotypes are related to phenotypes, mate choice may be translated into a rule to be applied directly to genotypes. For haploid individuals with B biallelic loci where each locus represents a trait, assortative mating can be implemented by preventing mating between individuals whose haplotypes differ in more than G loci [4, 17, 18, 20–22]. These models assume that there are many characteristics controlling mating preference. In birds, for example, important traits are the color of plumage, patterns on plumage, song length, song complexity, beak size, body size, etc. Modeling each trait by a binary label results in 2^B different phenotypes, where B is the genome size. For example, color is blue or red, beak is small or large, song is short or long, etc. Similar associations between haplotype and phenotype have also been used to study the branching of languages [52].

Our results also have implications for neutral models of speciation where reproductive isolation results from incompatibilities between individuals at the boundary between species [21, 22]. Individuals from each species have high fitness when amongst their own kind, but lower fitness at the boundary with other species, where the fraction of compatible mates drops. This feature keeps the populations isolated and prevents mixing in the absence of environmental selection.

ACKNOWLEDGMENTS

This work was partly supported by FAPESP (MAMA, DMS and ABM) and CNPq (MAMA, PRAC).

[1] H. S. Jennings, "The numerical results of diverse systems of breeding," *Genetics*, vol. 1, pp. 53–89, Jan. 1916.

[2] S. Wright, "Systems of mating. III. assortative mating based on somatic resemblance," *Genetics*, vol. 6, pp. 144–161, Mar. 1921.

[3] S. P. Otto, M. R. Servedio, and S. L. Nuismer, "Frequency-dependent selection and the evolution of assortative mating," *Genetics*, vol. 179, pp. 2091–2112, Aug. 2008.

[4] A. S. Kondrashov and M. Shpak, "On the origin of species by means of assortative mating.," *Proceedings of the Royal Society B: Biological Sciences*, vol. 265, pp. 2273–2278, Dec. 1998.

[5] M. Higashi, G. Takimoto, and N. Yamamura, "Sympatric speciation by sexual selection," *Nature*, vol. 402, pp. 523–526, Dec. 1999.

[6] U. Dieckmann and M. Doebeli, "On the origin of species by sympatric speciation," *Nature*, vol. 400, pp. 354–357, July 1999.

[7] D. I. Bolnick, "Waiting for sympatric speciation," *Evolution*, vol. 58, pp. 895–899, Apr. 2004. ArticleType: research-article / Full publication date: Apr., 2004 / Copyright 2004 Society for the Study of Evolution.

[8] V. Schwmmle, K. Luz-Burgoa, J. S. S. Martins, and S. M. de Oliveira, "Phase transition in a mean-field model for sympatric speciation," *Physica A: Statistical Mechanics and its Applications*, vol. 369, pp. 612–618, Sept. 2006.

[9] N. Burley, "Mate choice by multiple criteria in a monogamous species," *The American Naturalist*, vol. 117, pp. 515–528, Apr. 1981. ArticleType: research-article / Full publication date: Apr., 1981 / Copyright 1981 The University of Chicago.

[10] U. Candolin, "The use of multiple cues in mate choice," *Biological Reviews*, vol. 78, no. 4, pp. 575–595, 2003.

[11] S. F. Chenoweth and M. W. Blows, "Dissecting the complex genetic basis of mate choice," *Nature Reviews Genetics*, vol. 7, pp. 681–692, Sept. 2006.

[12] P. A. Hohenlohe and S. J. Arnold, "Dimensionality of mate choice, sexual isolation, and speciation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 16583–16588, Sept. 2010.

[13] J. M. Jawor, S. U. Linville, S. M. Beall, and R. Breitwisch, "Assortative mating by multiple

ornaments in northern cardinals (cardinalis cardinalis)," *Behavioral Ecology*, vol. 14, pp. 515–520, July 2003.

[14] J. Blais, M. Plenderleith, C. Rico, M. I. Taylor, O. Seehausen, C. v. Oosterhout, and G. F. Turner, "Assortative mating among lake malawi cichlid fish populations is not simply predictable from male nuptial colour," *BMC Evolutionary Biology*, vol. 9, p. 53, Mar. 2009.

[15] G. L. Conte and D. Schluter, "Experimental confirmation that body size determines mate preference via phenotype matching in a stickleback species pair," *Evolution*, vol. 67, no. 5, pp. 1477–1484, 2013.

[16] M. Serva and L. Peliti, "A statistical model of an evolving population with sexual reproduction," *Journal of Physics A: Mathematical and General*, vol. 24, p. L705, July 1991.

[17] P. G. Higgs and B. Derrida, "Stochastic models for species formation in evolving populations," *Journal of Physics A: Mathematical and General*, vol. 24, p. L985, Sept. 1991.

[18] P. G. Higgs and B. Derrida, "Genetic distance and species formation in evolving populations," *Journal of Molecular Evolution*, vol. 35, pp. 454–465, Nov. 1992.

[19] F. Manzo and L. Peliti, "Geographic speciation in the derrida-higgs model of species formation," *Journal of Physics A: Mathematical and General*, vol. 27, p. 7079, Nov. 1994.

[20] S. Gavrilets, H. Li, and M. D. Vose, "Rapid parapatric speciation on holey adaptive landscapes.," *Proceedings of the Royal Society B: Biological Sciences*, vol. 265, pp. 1483–1489, Aug. 1998.

[21] M. A. M. de Aguiar, M. Baranger, E. M. Baptestini, L. Kaufman, and Y. Bar-Yam, "Global patterns of speciation and diversity," *Nature*, vol. 460, pp. 384–387, July 2009.

[22] A. B. Martins, M. A. M. d. Aguiar, and Y. Bar-Yam, "Evolution and stability of ring species," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 5080–5084, Mar. 2013.

[23] S. Gavrilets, *Fitness Landscapes and the Origin of Species (MPB-41)*. Princeton University Press, July 2004.

[24] R. Aaziz and M. Tepfer, "Recombination in RNA viruses and in virus-resistant transgenic plants," *Journal of General Virology*, vol. 80, pp.1339–1346, 1999.

[25] K. Delviks-Frankenberry, A. Galli, O. Nikolaitchik, H. Mens, V. K. Pathak, and W.-S. Hu, "Mechanisms and factors that influence high frequency retroviral recombination," *Viruses*, vol. 3, pp. 1650–1680, Sept. 2011.

[26] E. Simon-Loriere and E. C. Holmes, "Why do RNA viruses recombine?," *Nature Reviews*

Microbiology, vol. 9, pp. 617–626, Aug. 2011.

[27] J. Zhang and H. M. Temin, "Retrovirus recombination depends on the length of sequence identity and is not error prone.," *Journal of Virology*, vol. 68, pp. 2409–2414, Apr. 1994.

[28] H. A. Baird, R. Galetto, Y. Gao, E. Simon-Loriere, M. Abreha, J. Archer, J. Fan, D. L. Robertson, E. J. Arts, and M. Negroni, "Sequence determinants of breakpoint location during HIV-1 intersubtype recombination," *Nucleic Acids Research*, vol. 34, pp. 5203–5216, Oct. 2006.

[29] E. Petterson, M. Stormoen, O. Evensen, A. B. Mikalsen, and O. Haugland, "Natural infection of atlantic salmon (salmo salar l.) with salmonid alphavirus 3 generates numerous viral deletion mutants," *Journal of General Virology*, vol. 94, pp. 1945–1954, Sept. 2013.

[30] P. D. Nagy and A. E. Simon, "New insights into the mechanisms of RNA recombination," *Virology*, vol. 235, pp. 1–9, Aug. 1997.

[31] M. Worobey and E. C. Holmes, "Evolutionary aspects of recombination in RNA viruses," *Journal of General Virology*, vol. 80, pp. 2535–2543, 1999.

[32] M. Eigen and P. Schuster, "A principle of natural self-organization," *Naturwissenschaften*, vol. 64, pp. 541–565, Nov. 1977.

[33] M. Eigen, "Selforganization of matter and the evolution of biological macromolecules," *Naturwissenschaften*, vol. 58, pp. 465–523, Oct. 1971.

[34] E. Domingo, D. Sabo, T. Taniguchi, and C. Weissmann, "Nucleotide sequence heterogeneity of an RNA phage population," *Cell*, vol. 13, pp. 735–744, Apr. 1978.

[35] D. A. Steinhauer, J. C. d. l. Torre, E. Meier, and J. J. Holland, "Extreme heterogeneity in populations of vesicular stomatitis virus.," *Journal of Virology*, vol. 63, pp. 2072–2080, May 1989.

[36] E. Domingo, J. Sheldon, and C. Perales, "Viral quasispecies evolution," *Microbiology and Molecular Biology Reviews*, vol. 76, pp. 159–216, June 2012.

[37] J. M. Smith and J. F. Y. Brookfield, "Models of evolution [and discussion]," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 219, pp. 315–325, Oct. 1983.

[38] E. Szathmary, "The integration of the earliest genetic information," *Trends in Ecology & Evolution*, vol. 4, pp. 200–204, July 1989.

[39] E. Szathmary and L. Demeter, "Group selection of early replicators and the origin of life," *Journal of Theoretical Biology*, vol. 128, pp. 463–486, Oct. 1987.

[40] M. C. Boerlijst, S. Bonhoeffer, and M. A. Nowak, "Viral quasi-species and recombination," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 263, pp. 1577–1584, Nov. 1996.

[41] M. N. Jacobi and M. Nordahl, "Quasispecies and recombination," *Theoretical Population Biology*, vol. 70, pp. 479–485, Dec. 2006.

[42] J.-M. Park and M. W. Deem, "Phase diagrams of quasispecies theory with recombination and horizontal gene transfer," *Physical Review Letters*, vol. 98, p. 058101, Jan. 2007.

[43] F. K. d. Boer and P. Hogeweg, "Eco-evolutionary dynamics, coding structure and the information threshold," *BMC Evolutionary Biology*, vol. 10, p. 361, Nov. 2010.

[44] J. R. Peck and D. Waxman, "Is life impossible? information, sex, and the origin of complex organisms," *Evolution*, vol. 64, pp. 3300–3309, Nov. 2010.

[45] E. S. Colizzi and P. Hogeweg, "Evolution of functional diversification within quasispecies," *Genome Biology and Evolution*, vol. 6, pp. 1990–2007, Aug. 2014.

[46] W.J. Ewens. *Mathematical Population Genetics. I. Theoretical Introduction.* Series: Biomathematics, Vol. 9 (New York: Springer Verlag, 1979).

[47] D. M. Schneider, E. do Carmo, A. B. Martins, and M. A. M. de Aguiar, "Toward a theory of topopatric speciation: The role of genetic assortative mating," *Physica A: Statistical Mechanics and its Applications*, vol. 409, pp. 35–47, Sept. 2014.

[48] D. M. Schneider, E. do Carmo, and M. A. M. de Aguiar, "A dynamical analysis of allele frequencies in populations evolving under assortative mating and mutations," *Physica A: Statistical Mechanics and its Applications*, vol. 421, pp. 54–68, 2015.

[49] M. N. Jacobi, and M. Nordahl,"Quasispecies and recombination," *Theor. Popul. Biol.*, vol. 70(4), pp. 479–485, 2006.

[50] G. Ochoa and K. Jaffe, "Assortative mating drastically alters the magnitude of error thresholds," in *Parallel Problem Solving from Nature - PPSN IX* (T. P. Runarsson, H.-G. Beyer, E. Burke, J. J. Merelo-Guervs, L. D. Whitley, and X. Yao, eds.), no. 4193 in Lecture Notes in Computer Science, pp. 890–899, Springer Berlin Heidelberg, Jan. 2006.

[51] E. Tannenbaum and J. F. Fontanari, "A quasispecies approach to the evolution of sexual replication in unicellular organisms," *Theory in Biosciences*, vol. 127, pp. 53–65, Mar. 2008.

[52] V. Schwmmle and P. M. C. de Oliveira, "A simple branching model that reproduces language family and language population distributions," *Physica A: Statistical Mechanics and*

its Applications, vol. 388, pp. 2874–2879, July 2009.

FIGURE CAPTIONS

Figure 1: (Online version in color.) Equilibrium haplotype frequencies for $B = 3$ and $G = 0$, $G = 1$ and $G = 2$ as a function of the mutation rate. Lines correspond, from top to bottom, to: 000 (black); 100, 010 and 001 (blue); 110, 101 and 011 (red); 111 (green). The initial condition is $p_{000} = 1$ and the other frequencies zero.

Figure 2: (Online version in color.) Critical mutation as a function of G/B. The uniform distribution of haplotype frequencies becomes stable for $\mu > \mu_c$. The area below the curves correspond to a stable distribution of closely related haplotypes.

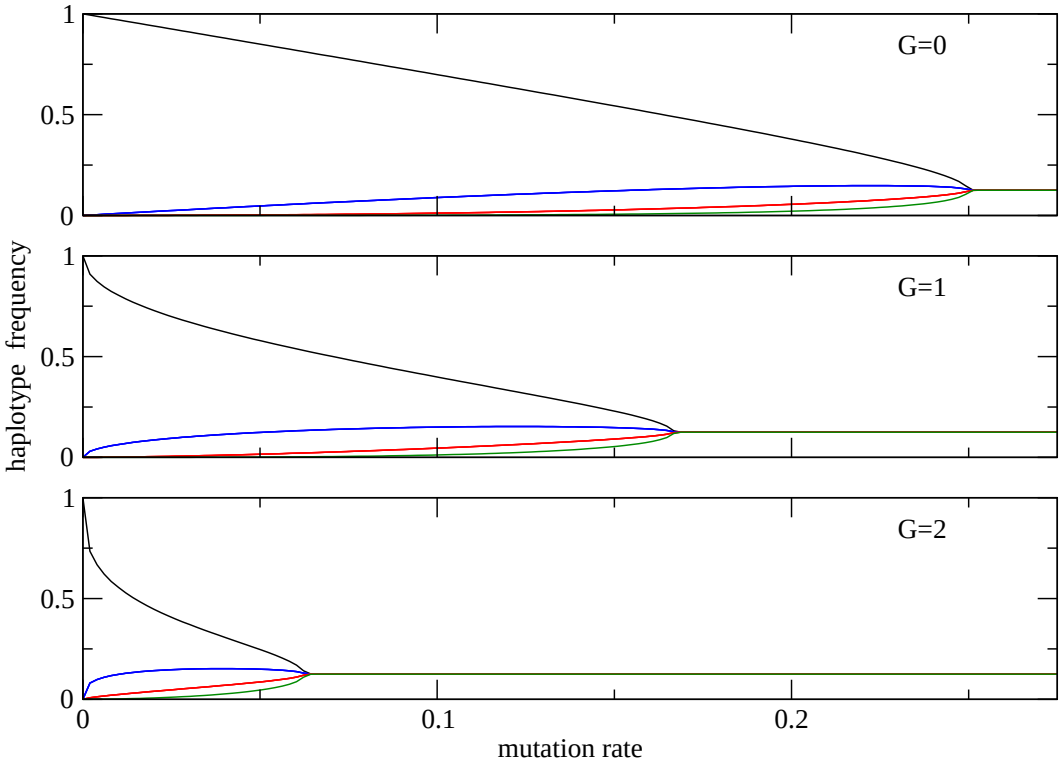

Figure 1.

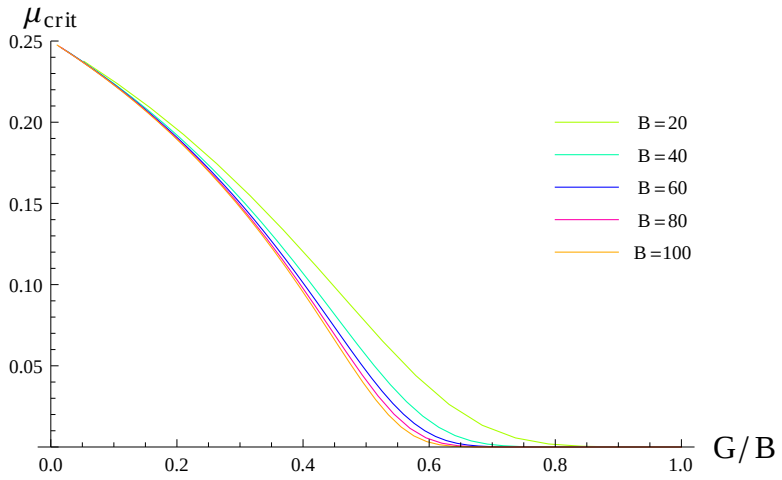
μ_{crit}
0.25
0.20
0.15
0.10
0.05
0.0
0.2
0.4
0.6
0.8
1.0
G/B
B = 20
B = 40
B = 60
B = 80
B = 100

Figure 2.