Marshallplan scholarship: final report

July 30, 2018

Martin Pontz[a,b]

[a] Department of Mathematics, University of Vienna, Austria
[b] Vienna Graduate School of Population Genetics

Overview

During my stay with Marc Feldman at the Department of Biology of Stanford University, I analyzed two interesting problems from population genetics theory. In particular, I investigated dynamics that are governed by genes on two loci. The first problem concerned the dependence of the invasion rate of a new mutant allele on the recombination rate.

Previous results for deterministic two-locus two-alleles models with and without migration showed that the invasion rate of a new mutant occuring at a locus linked to an already existing polymorphism decreases with increasing recombination rate. However, stochastic versions of the same models show that there can be instances where this monotonic dependence does not hold. We investigated a population that is at a two-locus two-allele polymorphism and faces invasion of a new (third) allele at one of the loci. In particular, the invasion rate of the new (mutant) allele was computed and analysed. To facilitate the derivations, we assume a symmetric viability model generalized to multiple alleles.

In my second project we tried to resolve a long-standing claim for haploid selection models with recombination. In the evolutionary biology literature, it is generally assumed that in deterministic haploid selection models, in the absence of variation-generating mechanisms such as mutation, no polymorphic equilibria can be stable. Thus the genetic variation gets lost. However, analytic results corroborating this claim are scarce and almost always depend upon additional assumptions on the strength of selection with respect to recombination. While we cannot yet prove the claim in general, we establish a necessary condition for the existence of an isolated full polymorphism, i.e., an equilibrium at which all alleles are present at both loci: The number of alleles at each of the two loci must be the same.

In section 1, I present the results for the invasion analysis, while section 2 is concerned with the two-locus haploid dynamics. The two sections are both in itself structured like draft manuscripts. This represents the current state of the research that I started during my stay. The results obtained in section 1 hint on several problems that will be investigated in the future. Work on section 2 is currently continued with new input from Josef Hofbauer and Reinhard Bürger. Marc Feldman has been involved in both projects.

1 Invasion of a mutant allele into a two-locus two-allele polymorphism

Author: Martin Pontz and Marc Feldman

1.1 Introduction

There are two pairs of papers, one from the beginning of multilocus theory, the other rather recent, which are concerned with the possibility of invasion of a single locus polymorphism by a mutant allele. The first pair consists of the papers by Bodmer and Felsenstein (1967) and Ewens(1967). The former analyze the deterministic two-locus two-allele model in great detail. In one section they derive the conditions for the increase of a new allele linked to a polymorphic locus. They find that the rate of this increase is monotone decreasing with the increase of the recombination rate r. In the same year, Ewens investigated the probability of fixation of a mutant at two loci (Ewens 1967). There he applies basically a two-type branching process starting form an equilibrium where one locus is fixed and the other polymorphic in two alleles. The two types in the branching process correspond to the two haplotypes that are formed by the allele at the polymorphic locus with the new allele. Since the equations that yield the exact invasion probability with respect to the two types are transcendental and have thus no closed form solution, he derived an approximation. The average of the two approximations for the two types weighted by the frequency of the two types at equilibrium serves as an approximation to the mean invasion probability. This function gives insight into the mean invasion probability over multiple instances of mutations. For most parameter combinations, this function is always decreasing with r, as in the deterministic case. However, in this stochastic analysis, it becomes clear that in certain parameter regions this function is increasing for small r and decreasing for higher r. It still holds in general that the invasion probability for complete linkage is higher than for no linkage.

The second pair of papers yielded basically the same dichotomy between deterministic and stochastic treatments of a two-locus two-allele model with migration. Akerman and Bürger (2011) provided the deterministic treatment of a two-locus two-allele model based on a continent-island migration scheme. They found that the optimal recombination rate for invasion is 0. This means the invasion rate is also declining with r. The stochastic treatment was done by Aeschbacher and Bürger (2014). Similar to Ewens they applied a two-type branching process and derived an approximate expression for the invasion probability. The approximation was derived by the assumption of a slightly supercritical branching process, which was made rigorous by Athreya (1993). Ewens' approximation was done under basically the same assumption, but it was not yet called the slightly supercritical branching process assumption. As Ewens, Aeschbacher and Bürger (2014) found that for certain parameter values the optimal recombination rate is not zero. In particular they showed that the mean invasion probability is increasing for small r and decreasing for higher r. Outside of the parameter regions where such a behaviour is seen, the mean invasion probability is monotone decreasing with the recombination rate.

Aeschbacher and Bürger provide both a condition for the existence of a non-zero optimal recombination rate and an intuitive explanation based on this. They compute a bound a^* for the fitness of the invading allele, up to which the optimal recombination rate is zero. Above a^*, it is non-zero. As the mean invasion probability is composed of the invasion probabilities of the two types, in general, in the cases of a non-zero optimal recombination rate, one type is beneficial over the other. For the mutant allele this defines a good and a bad background to appear on. Higher recombination rates implies a higher chance for the mutant to recombine out of the initial type. This is detrimental if the mutant starts on the good background but beneficial if it starts on the bad background. The fitness of the mutant type can be seen to be proportional to the time how long it can survive on its own, which implies it can survive longer on the bad background and wait for the next recombination event, if the fitness is higher than a^*. If the mutant fitness is lower than a^*, then it goes extinct on the bad background before it can recombine out. An obvious question is if these phenomena are a general property or specific for the two-locus two-allele case. The next generalisation is clearly a system where one locus has two alleles and the other three alleles. Invasion of the third allele into a two-locus two-allele subsystem is investigated. Whereas the single locus polymorphism is unique, there can be up to seven two-locus two-allele polymorphisms in general. By restricting ourselves to a symmetric fitness scheme, we can reduce this number to three, where their coordinates are computed rather simply.

After introducing the model and determining the equilibria, the deterministic case is investigated followed by the stochastic case.

1.2 Model setup

We want to investigate the invasion rate (deterministic case) and the invasion probability (stochastic case) and their dependence on the recombination rate. We consider a two-locus system with alleles A_1, A_2 and A_3 at locus A and B_1 and B_2 at locus B. We are especially interested in the invasion of two-locus two-allele polymorphisms by a third allele, A_3 say. Such a polymorphism is a boundary equilibrium in the full $6-$dimensional space.

To make the analysis simpler we assume a three parameter symmetric viability scheme generalized to two and three alleles.

Double homozygotes are assigned a fitness of $1-a$, while double heterozygotes have fitness 1. Individuals that are homozygous in the B-locus and heterozygous at the other locus have fitness $1-b$, and those that are heterozygous at the A-locus and homozygous at the B-locus have fitness $1-c$. Note that a, b and c are all between 0 and 1.

The condensed 3×6 fitness scheme is below (since position effects are ignored):

	A_1A_1	A_1A_2	A_2A_2	A_1A_3	A_2A_3	A_3A_3
B_1B_1	$1-a$	$1-b$	$1-a$	$1-b$	$1-b$	$1-a$
B_1B_2	$1-c$	1	$1-c$	1	1	$1-c$
B_2B_2	$1-a$	$1-b$	$1-a$	$1-b$	$1-b$	$1-a$

(1)

In the following, we display the full 6×6 fitness scheme and denote the resulting 6×6 matrix by W, where $W_{11} = 1-a$, etc.

	A_1B_1	A_1B_2	A_2B_1	A_2B_2	A_3B_1	A_3B_2
A_1B_1	$1-a$	$1-c$	$1-b$	1	$1-b$	1
A_1B_2	$1-c$	$1-a$	1	$1-b$	1	$1-b$
A_2B_1	$1-b$	1	$1-a$	$1-c$	$1-b$	1
A_2B_2	1	$1-b$	$1-c$	$1-a$	1	$1-b$
A_3B_1	$1-b$	1	$1-b$	1	$1-a$	$1-c$
A_3B_2	1	$1-b$	1	$1-b$	$1-c$	$1-a$

(2)

As a consequence, we will be able to write the dynamical equations in matrix form, see eq. (3).

We define $\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $\mathbf{D} = \{-D_1, D_1, D_2, -D_2, D_3, -D_3\}$ as the vectors of gamete frequencies (A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B1, A_3B_2) and linkage disequilibria respectivly. For this situation we have three measures of LD where only two are independent ($D_3 = D_1 - D_2$). Then we can write

$$\bar{w}\mathbf{x}_i' = \mathbf{x}_i W \mathbf{x} + r\mathbf{D}_i, \tag{3}$$

for $i = 1, ..., 6$ and $\bar{w} = \mathbf{x}W\mathbf{x}$, the mean fitness. The recombination rate is denoted by $0 \leq r \leq \frac{1}{2}$, where $r = 0$ means that the loci are fully linked and $r = \frac{1}{2}$ corresponds to free recombining loci. Fully linked loci are interpreted as sitting very closely together on the chromosome, whereas free recombination happens when the loci are far apart (eg. on different chromosomes). The D_i are defined in the following way:

$$D_1 = d_1 + d_2, \quad D_2 = d_1 + d_3, \quad D_3 = d_2 - d_3 \tag{4}$$

$$\text{with} \quad d_1 = x_1x_4 - x_2x_3, \quad d_2 = x_1x_6 - x_2x_5, \quad d_3 = x_4x_5 - x_3x_6. \tag{5}$$

No direct biological meaning is known for the d_i.

1.3 Deterministic analysis

1.3.1 Equilibria

For a simpler fitness scheme, Feldman, Lewontin, Franklin and Christiansen (1974) showed existence of many equilibria and categorized them into three classes. However, the exact stability properties of the two-locus two-allele subsystem equilibria that we are interested in, was not investigated.

However from Feldman, Lewontin, Franklin and Christiansen (1974) it is clear that the central point C with $x_i = \frac{1}{6}$ $(\forall i)$ is always an equilibrium.

The classical papers about two-locus two-allele models with this type of symmetric viabilities, inform us that the three polymorphisms that are possible can be viewed as boundary equilibria on the face with $x_5 = x_6 = 0$. Let E_q be the equilibrium satisfying $x_i = \frac{1}{4}$ for $i = 1, 2, 3, 4$ and $x_5 = x_6 = 0$. The two so-called highly complimentarity equilibria with $D := D_1 \neq 0$ are denoted E_+ and E_-. The coordinates of $E_\pm$ are $(\pm v, \mp v, \mp v, \pm v, 0, 0)$, where $\pm v = \frac{1}{4}(1 \pm \sqrt{1 - \frac{4r}{b+c-a}})$. $E_\pm$ are admissible only if $a + 4r < b + c$.

1.3.2 Stability

Bodmer and Felsenstein linearized the dynamics at the single locus polymorphism, which resulted in a decomposition of the characteristic equations into two quadratics. One determines the stability with respect to the alleles present at the equilibrium and the other determines the stability with repect to the missing allele. From this they derived a condition for the increase of the new gamete. In more modern terminology, they used the external eigenvalues to derive this condition. External eigenvalues are the eigenvalues that are transversal to the boundary at an boundary equilibrium. A boundary equilibrium is saturated if the external eigenvalues are smaller than one. If a boundary equilibrium is not saturated, a missing allele can invade into the population. It thus suffices to investigate the local stability of the boundary equilibria by looking at the Jacobian of the full system at this equilibrium.

While invasion of an boundary equilibrium is a local property that does not imply anything about the full flow, it is, nonetheless, of interest how the conditions for invasion of the boundary equilibria affect the stability of the central equilibrium C. It is always admissible and the stability properties are computed in the following paragraph.

Equilibrium C : $\mathbf{\hat{x}_i = \frac{1}{6}}$ $(i = 1, ..., 6)$**.** Let us start with determining the eigenvalues of C:

$$\{0, \frac{2(3 - a - 2b)}{6 - a - 2b - c}, \frac{6 - 2a - b - 2c}{6 - a - 2b - c}, \frac{6 - 2a - b - 6r}{6 - a - 2b - c}\}, \tag{6}$$

where the last two have multiplicity 2. The zero eigenvalue is due to the fact that the state space, the simplex, is actually a five dimensional subspace in the 6 dimensional space and the sixth eigenvalue has no relevance for stability with respect to the state space.

C is locally stable if the following inequalities are fulfilled:

$$c < a + 2b \tag{7a}$$

$$b < a + c \tag{7b}$$

$$\frac{b + c - a}{6} < r \tag{7c}$$

Equilibrium $\mathbf{E_q}$: $\left(\mathbf{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0}\right)$. In the (x_1, x_2, x_3, x_4)-subspace E_q has the following so-called internal eigenvalues:

$$\{\frac{2(2-a-b)}{4-a-b-c}, \frac{2(2-a-c)}{4-a-b-c}, \frac{2(2-a-2r)}{4-a-b-c}\}. \tag{8}$$

Thus it is internally stable if all of the following inequalitites hold

$$c < a+b \tag{9a}$$

$$b < a+c \tag{9b}$$

$$\frac{b+c-a}{4} < r. \tag{9c}$$

The external eigenvalues on the other hand are

$$\{\frac{2(2-b)}{4-a-b-c}, \frac{2(2-b-2r)}{4-a-b-c}\}, \tag{10}$$

where clearly the first is the leading one. It is larger than one if

$$b < a+c. \tag{11}$$

Then invasion by A_3 is possible.

At this point it is easy to see that E_q can never be stable in the 6-dimensional space.

Equilibrium $\mathbf{E_\pm}$: $(\pm\mathbf{v}, \mp\mathbf{v}, \mp\mathbf{v}, \pm\mathbf{v}, \mathbf{0}, \mathbf{0})$. The eigenvalues of E_+ are the same as for E_- because of symmetry. Thus we treat them together as $E_\pm$.

The internal eigenvalues of $E_\pm$ are $\frac{2-b-c+2r}{2-a-2r}$ and the solutions of a quadratic equation. The explicit one is always smaller than one if the equilibrium is admissible, i.e. $r < \frac{b+c-a}{4}$. The other two correspond to a phenomenom called "Ewens gap" and is treated in greater detail later.

The external eigenvalues are

$$\{\frac{2-b}{2-a-2r}, \frac{2-b-2r}{2-a-2r}\}, \tag{12}$$

where the first one is always greater than the second and thus the leading eigenvalue. It is larger than one if

$$b-a < 2r. \tag{13}$$

Pattern of stability: Comparing the conditions (7), (9), (11) and (13) for stability for the different equilibria lead to the following observations:

If E_q is internally stable, then C is stable. Indeed A_3 can then always invade E_q and the ω-limit of this orbit is probably C. As long as E_q is internally stable, $E_\pm$ are not admissible.

If $0 < r < \frac{b+c-a}{4}$, then E_q is internally unstable at least in one direction. However, exactly then $E_\pm$ become admissible and internally stable in at least one direction.

Invasion of A_3 at $E_\pm$ requires $\frac{b-a}{2} < r < \frac{b+c-a}{4}$. Thus $b < a+c$ is necessary for invasion. This implies that also E_q is invaded by A_3 and that C has at least one eigenvalue smaller than one. If additionaly $c < a+b$, then E_q is still unstable, but C may be stable if $\frac{b+c-a}{6} < r$.

If A_3 can not invade $E_\pm$, it could still invade E_q. However, if $b > a+c$, then it cannot invade E_q and C is always unstable.

To sum this up, if $0 < r < \frac{b+c-a}{4}$ and $b < a+c$, then invasion of A_3 at $E_\pm$ depends on the recombination rate. For $0 < r < \frac{b-a}{2}$ $E_\pm$ is not invaded, but E_q is, while C could be stable. For $\frac{b-a}{2} < r < \frac{b+c-a}{4}$ $E_\pm$ is invaded and so is E_q (however not internally stable). C could still be stable. It is never stable, if $r < \frac{b+c-a}{6}$, which is certainly fulfilled before $E_\pm$ get lost ($\frac{b+c-a}{6} < \frac{b+c-a}{4}$).

If $E_\pm$ is invaded by A_3, then always at E_q too. However, this is irrelevant as long as E_q is internally unstable which is always true when $E_\pm$ exist. This comes from the fact that in nature a internally unstable boundary equilibrium will never be observed and thus it does not make sense to talk about invasion of such a point.

Dependence on r An important property of invasion is its dependence on r. The invasion rate at E_q does not depend on r; only one of the internal eigenvalues involves r and it is monotone decreasing in r. This means E_q is internally stabilized by increased r. The same holds for one of the eigenvalues of C. The other eigenvalues of C are independent of r.

The leading external eigenvalue of $E_\pm$ is increasing with r. However when r is too high the equilibrium no longer exists. This means the invasion rate at $E_\pm$ is increasing with decreasing distance to E_q with which they merge at $r = \frac{b+c-a}{4}$. This shows that the invasion rate at $E_\pm$ is always smaller than at E_q.

1.3.3 Lewontin-Kojima case: $b = c$

The stability analysis becomes simpler if we use the well known simplification $b = c$, first analysed by Lewontin and Kojima (1960).

Conditions (7) for C simplify to

$$0 < a + b, \tag{14a}$$

$$0 < a, \tag{14b}$$

$$\frac{2b-a}{6} < r. \tag{14c}$$

The conditions (9) and (11) for internal and external stability of E_q simplify to

$$0 < a, \tag{15a}$$

$$\frac{2b-a}{4} < r \tag{15b}$$

and

$$0 < a. \tag{16}$$

Condition (13) for $E_\pm$ remains unchanged.

The only inequality that does not hold by definition of a and b ($0 < a, b$) for stability of C is

$$r > \frac{2b-a}{6}. \tag{17}$$

As before it is fulfilled if E_q is internally stable, which requires in this case only

$$r > \frac{2b-a}{4}. \tag{18}$$

Invasion of E_q is always possible.

Internal stability of $E_\pm$ is also simpler in this case. The only inequality that is not implied by admissibility is

$$(2b-a)(b-2r) > \sqrt{(a-b)^2(2b-a)(2b-a-4r)}. \tag{19}$$

Admissibility of $E_\pm$ ensures positivity under the root and of the left hand side. Ewens and follow up papers showed that there can be two cases for the range of r in which this inequality holds. Either it holds for $0 < r < \frac{2b-a}{4}$, or there is an interval in r that yields a gap in stability of $E_\pm$. Existence of the gap depends on the ratio between a and b. Karlin and Feldman (1970) state that Ewens gap occurs if and only if

$$\frac{b}{a} < 2 + \sqrt{2}. \tag{20}$$

$E_\pm$ is admissible, if

$$r < \frac{2b-a}{4} \tag{21}$$

holds.

The results can now be summed up in terms of ratios of a and b:

$2b < a$: (17), (18) and (21) imply that C is fully stable, E_q internally stable but never externally and $E_\pm$ do not exist independently of r.

$b < a < 2b$: For $0 < r < \frac{2b-a}{6}$, (17), (18), (19) and (13) imply that C is unstable, E_q is internally unstable, while $E_\pm$ are internally stable but get invaded by A_3. For $\frac{2b-a}{6} < r < \frac{2b-a}{4}$ only the stability of C changes, but nothing else (for the equilibria we study). For $r > \frac{2b-a}{4}$ the qualitative behaviour is the same as in the ($2b < a$) case considered above.

$a < b < 2a$: For $0 < r < \frac{b-a}{2}$ (17), (18) and (19) imply that C and E_q are internally unstable and $E_\pm$ is internally stable, but $E_\pm$ cannot be invaded because of condition (13). For $\frac{b-a}{2} < r$ the behaviour is the same as in the second case ($b < a < 2b$) discussed above, where the lower bound of zero for r is replaced by $\frac{b-a}{2}$.

$2a < b$: Here, the recombination rate threshold above which C is stable, is smaller than that where $E_\pm$ gets invaded. Also, only then can $E_\pm$ become unstable for intermediate recombination rates (Ewens gap).

We did not analyse how the occurence of Ewens gap affects the invasion of A_3. The reason is that it makes no sense biologically to study the invasion of an unstable equilibrium by a mutant haplotype, since the population is never directly at this state, except if it initially starts there.

1.4 Stochastic analysis

Here, we present the results of the stochastic treatment of the invasion properties of a mutant allele invading an equilibrium of the two-lous two-allele subsystem. This will be done by calculating invasion probabilities for a two-type branching process, where the two types are defined by the allele on the B-locus that forms the mutant haplotype with A_3.

For a two-type branching process, we need the mean offspring matrix L, which is also the 2×2 external block of the Jacobian matrix at a boundary equilibrium. This external block contains the entries of the Jacobian $\frac{\partial \dot{x}_i}{\partial x_j}$ such that x_i as well as x_j are not present at the boundary equilibrium. From this matrix the leading eigenvalue is the external leading eigenvalue λ_1 of the full Jacobian. There is a non-negative probability of invasion if it is greater than 1. This is the same condition that was analysed in the previous chapter.

To compute the approximate invasion probabilities by assuming a slightly supercritical branching process, we need the left and right eigenvectors corresponding to λ_1. The left eigenvector u is to one and the right eigenvector v such that $\sum u_i v_i = 1$. With this we can compute B, where

$$B = \sum_i (u_i * \sum_j v_j L_{ij}) + \lambda_1 (1 - \lambda_1) \sum_k u_k v_k^2. \tag{22}$$

Type i has invasion probability π_i. Athreya (1993) showed the following equation for the slightly supercritical branching process:

$$\pi_i = 2(\lambda_1 - 1)\frac{v_i}{B} \tag{23}$$

The average invasion probability $\bar{\pi}$ is the average of the invasion probabilities for the different types weighted by the frequency of the types at the equilibrium:

$$\bar{\pi} = \hat{q}\pi_1 + (1 - \hat{q})\pi_2 \tag{24}$$

$\hat{q}$ is the frequency of B_1 at equilibrium.

The following paragraphs show the computations for the boundary equilibria of interest:

For equilibrium $\mathbf{E_q}$:

$$L = \frac{1}{4-a-b-c}\begin{pmatrix} 2(2-b-r) & 2r \\ 2r & 2(2-b-r) \end{pmatrix} \tag{25a}$$

$$u = (\frac{1}{2}, \frac{1}{2}) \tag{25b}$$

$$v = (1,1) \tag{25c}$$

$$B = \frac{4(2-b)(2-a-c)}{(4-a-b-c)^2}. \tag{25d}$$

This, together with the fact that the frequencies of alleles B_1 and B_2 are the same ($\frac{1}{2}$) for this equilibrium, yields

$$\bar{\pi} = \pi_1 = \pi_2 = \frac{(a-b+c)(4-a-b-c)}{2(2-b)(2-a-c)}, \tag{26}$$

which is independent of r as is the deterministic equivalent.

For equilibrium $\mathbf{E_\pm}$:

$$L = \frac{1}{2-a-2r}\begin{pmatrix} 2-b-r & r \\ r & 2-b-r \end{pmatrix} \tag{27a}$$

$$u = (\frac{1}{2}, \frac{1}{2}) \tag{27b}$$

$$v = (1,1) \tag{27c}$$

$$B = \frac{(2-b)(2-2a+b-4r)}{(2-a-2r)^2}. \tag{27d}$$

This together with the fact that the frequencies of alleles B_1 and B_2 are the same ($\frac{1}{2}$) for this equilibrium, yields

$$\bar{\pi} = \pi_1 = \pi_2 = \frac{2(a-b+2r)(2-a-2r)}{(2-b)(2-2a+b-4r)}. \tag{28}$$

The invasion probability is increasing with r as in the deterministic case.

The invasion probabilities for both types of boundary equilibria coincide exactly when the equilibria coincide.

1.5 Discussion

We have conducted a thorough analysis in both the deterministic setup as well as in a stochastic setup of the invasion rates and the invasion probabilities of a third allele into the population, which is at one of the three admissible two-locus two-allele boundary equilibria, respectively. For the deterministic analysis the classification of the eigenvalues of the Jacobian at boundary equilibria into internal and external ones proved very helpful. The leading external eigenvalue is interpreted as the invasion rate of the missing allele. If it is smaller than unity, then the mutant allele does not invade. Some interesting patterns for the stability of the boundary equilibria emerge. Stability with respect to the boundary of E_q implies stability of C and non-admissibility of $E_{\pm}$. Internal stability of $E_{\pm}$, if the equlibria exist, implies internal instability of E_q. Internal and external stability of $E_{\pm}$ imply that both E_q and C are unstable. This means that the stability of a single two-locus two-allele boundary equilibrium already implies the stability or instability of the central equilibrium C.

Due to the symmetry of the fitness scheme, admissibility and stability of certain two-locus two-allele boundary equilibria is the same as for the equilibria on the other two two-locus two-allele boundaries. This means, if A_3 can invade the boundary equilibrium where A_1 and A_2 are present, if it is initially missing, so can A_1 invade the boundary equilibrium where A_2 and A_3 are present, if it is initially missing.

For the stochastic analysis via a two-type branching process, the derivations were facilitated by the observation that the external block of the Jacobian at an boundary equilibrium is the same as the mean matrix as defined in multi-type branching processe theory. The condition for non-extinction is exactly the same as for invasion in the deterministic setting. The leading eigenvalue of the external block of the Jacobian (i.e. the leading external eigenvalue) has to be larger than unity. The eigenvectors corresponding to this eigenvalue that are needed to perform the approximation with the sligthly supercritical branching process assumption are of simple forms, due to the symmetries of the fitness scheme. The approximate expressions for the invasion probabilities have the same dependences on r as the invasion rates in the deterministic setting.

In general it was shown that the phenomena that occur in the invasion analysis of the two-locus two-allele case do not generalize to a model with two alleles on one locus and three on the other under a symmetric viability model. One point was that the invasion rate of an allele into a single locus polymorphism is monotone decreasing with the recombination rate. Here we have seen that it depends on the exact equilibrium that we investigate and is either independent of r for the equilibrium E_q or monotone increasing for $E_{\pm}$. That is also a major difference that before only a unique boundary equilibrium existed which missed the allele, whereas in the current analysis up to three equilibria can exist, where the same allele is absent and could thus invade.

Another point in the previous analysis was that a stochastic analysis yielded a different invasion pattern. In fact the invasion probability did not depend monotonically on r for some parameter values. Here the stochastic analysis yields qualitatively the same dependence on r as the deterministic analysis.

It is not clear if these phenomena do not generalize to higher numbers of alleles because of the symmetric viability model or because of a more general effect. It could be possible and should be investigated in a follow up analysis thatthe invasion rate, for example, for the $D = 0$ equilibrium E_q is also monotone decreasing in r, if no specific fitness scheme is applied. If the invasion probability will then exhibit a non-zero optimal recombination rate in a certain parameter region is also unclear.

References

Aeschbacher S. and Bürger R. 2014. The effect of linkage on establishment and survival of locally beneficial mutations. Genetics, 197, 317-336.

Athreya K.B. 1993. Rates of decay for the survival probability of a mutant gene II: The multitype case. J. Math. Biol., 32, 45-53.

Bodmer W.F. and Felsenstein F. 1967. Linkage and selection: Theoretical analysis of the deterministic two locus random mating model. Genetics, 57, 237-265.

Bürger R. and Akerman A. 2011. The effects of linkage and gene flow on local adaptation: A two-locus continent-island model. Theoret. Popul. Biol., 80, 272-288.

Ewens W.J. 1967. The probability of fixation of a mutant: The two locus case. Evolution, 21, 532-540.

Feldman M.W., Lewontin R.C., Franklin I.R. and Christiansen F.B. 1974. Selection in complex genetic systems III. An effect of allele multiplicity with two loci. Genetics, 79, 333-347.

Karlin S. and Feldman M.W. 1970. Linkage and selection: Two locus symmetric viability model. Theoret. Popul. Biol., 1, 39-71.

Lewontin R.C. and Kojima K. 1960. The evolutionary dynamics of complex polymorphism. Evolution, 14, 458-472.

2 Two-locus multiallelic haploid selection

Authors: Martin Pontz, Marc Feldman and Josef Hofbauer

2.1 Introduction

A recent paper by Novak and Barton (2017) raises one of the main questions of population genentics right in the title: "When does frequency-independent selection maintain genetic variation?" They note that, while the answer is generally assumed to be "no" for constant selection acting on an idealized haploid population, basically only the cases of no recombination and no selection have been solved and are able to corroborate this claim. Well known results from perturbation theory, of course, expand these results to small parameter values of the respective force. In fact they give a new, more standard, proof for the case of weak selection. The other extreme, strong linkage (low recombination rate), was solved by Kirzhner and Lyubich (1997), where they also arrive at the same conclusion for additive fitnesses and arbitrary linkage. These three results hold for any number of loci and any number of alleles. They all incorporate convergence of the solutions to equilibrium points, via the powerful method of identifying a Lyapunov function.

Besides the general additive case and the trivial one-locus case, only the two-locus two-allele case has recieved attention for intermediate values of recombination and selection. Feldman (1971) was one of the first to rigorously analyze existence and stability of polymorphic equilibria in a two-locus two-allele haploid system with a simple fitness scheme. He showed that whenever a polymorphism exists, it is unique and unstable. A general fitness scheme was considered by Rutschman (1994). He showed convergence of the trajectories to equilibrium points in most fitness parameter regions. However, parameter combinations in which an internal equilibrium was possible, couldn't be treated in the same way. The final answer to the question of loss of genetic variation in the two-locus two-allele case is the paper by Bank, Bürger and Hermisson (2012). They showed that for the fitness parameter combinations not covered by Rutschman, an equilibrium may exist, but it is always unstable. They also used the method of Lyapunov functions to derive this result, which immediately implies that no chaotic or otherwise complicated behaviors such as limit cycles can occur for the two-locus two-allele haploid selection model.

We consider a well mixed haploid population with constant selection on two loci, each with an arbitrary number of alleles. Our fitness scheme is general without any restriction on the epistatic interaction between alleles. For convinience, the dynamics are stated in contiuous time. First, we state and prove that no internal equilibrium exists if the numbers of alleles at the two loci are unequal. This is done by finding a system of linear equations, which has to be solved in order to find an internal equilibrium. For unequal numbers of alleles at the two loci, this system is overdetermined and thus, has no solution by basic linear algebra.

2.2 Model setup and main theorem

In the two locus haploid model considered here, we assume that at one locus there exist alleles $A_1, ..., A_m$, while at the other locus the alleles are $B_1, ..., B_n$. Let p_{ij} and s_{ij} be the frequency and the fitness, respectively, of haplotype A_iB_j and define the matrix $\tilde{S} = (s_{ij})_{m\times n}$. Following Nagylaki (1992) pp. 189-195 and Novak and Barton (2017), we can write the change in frequency over time, $\dot{p}_{ij} = \frac{dp_{ij}}{dt}$, as

$$\dot{p}_{ij} = r(p_i q_j - p_{ij}) + p_{ij}(s_{ij} - \sum_{ij} s_{ij}p_{ij}), \tag{29}$$

where $\sum_{ij} s_{ij}p_{ij}$ is the mean fitness and $p_i = \sum_{j=1}^{n} p_{ij}$ and $q_j = \sum_{i=1}^{m} p_{ij}$ are the marginal frequencies of the alleles. As always, the sum of all haplotype frequencies is one. The quantities $(p_i p_j - p_{ij})$ are the linkage disequilibria (LD).

We want to investigate existence and stability properties of the polymorphic equilibria. At equilibrium, the following equations hold:

$$0 = \dot{p}_{ij} = r(p_i q_j - p_{ij}) + p_{ij}(s_{ij} - \bar{s}). \tag{30}$$

These are mn quadratic equations in mn variables. We order the equations by the subscripts of $\dot{p}$ and represent them as a matrix P, where the ij-th entry is the equation for $\dot{p}_{ij}$ in (30). As usual, the mean fitness is written as $\bar{s}$. For convienience, we use allele frequencies p_i and q_j as helper variables that are linear combinations of the p_{ij}. Thus we have the following $m + n + 2$ additional linear equations:

$$p_i - \sum_{j=1}^{n} p_{ij} = 0, \quad \forall i \tag{31a}$$

$$q_j - \sum_{i=1}^{m} p_{ij} = 0, \quad \forall j \tag{31b}$$

$$\bar{s} - \sum_{ij} s_{ij}p_{ij} = 0, \tag{31c}$$

$$1 - \sum_{ij} p_{ij} = 0. \tag{31d}$$

Throughout the following we assume that the following mn inequalities

$$0 < p_{ij} < 1, \quad \forall i, j \tag{32}$$

hold. The main result is

Theorem 1. *If $m \neq n$, then* (30) *has either no solution for which* (32) *holds or infinitely many. That is, there is no isolated equilibrium with all mn haplotypes present.*

2.3 Weaker version of Theorem 1

During the stay at Stanford a weaker version of Theorem 1 emerged:

Theorem 2. *Let S be a full-rank matrix of size $m \times n$ and r be independent of any s_{ij}. If $m \neq n$, then* (33) *has no solution for which* (32) *holds. That is, there is no equilibrium with all mn haplotypes present.*

The stronger version with a different proof was found later with the help of Prof. Hofbauer in Vienna. Let me first state the proof of Theorem 2:

In this proof we use a slightly different version of (30) to derive the results.

It is easy to see that adding a constant to the entries of the fitness matrix S', doesn't alter (29). We aim to find properties of polymorphic equilibrium ($0 < p_{ij} < 1$, $\forall i, j$); therefore we assume that one exists. The mean fitness at this equilibrium is $\hat{\bar{s}}$ and we can then substract $\hat{\bar{s}}$ from the entries of S' to get S, which rescales $\bar{s}$ at equilibrium to 0. Then (30) becomes

$$0 = \dot{p}_{ij} = rp_iq_j + p_{ij}(s_{ij} - r), \tag{33}$$

because (31c) becomes

$$\bar{s} = \sum_{ij} s_{ij}p_{ij} = 0. \tag{34}$$

To prove Theorem 2, we first need

Lemma 1. *Define $k_i = \frac{s_{i1}-r}{s_{11}-r}$ and $t_i = \frac{p_{i1}}{p_{11}}$ for $1 \leq i \leq m$ and $c_{ij} = \frac{s_{1j}-r}{s_{ij}-r}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Then*

$$p_i = k_it_ip_1, \quad \forall i \tag{35}$$

$$p_{ij} = c_{ij}k_it_ip_{1j}, \quad \forall i, j \tag{36}$$

at equilibrium.

Proof. From (33), we can write

$$-rp_iq_1 = p_{i1}(s_{i1} - r) \quad \text{and} \quad -rp_1q_1 = p_{11}(s_{11} - r). \tag{37a}$$

So,

$$\frac{p_i}{p_1} = \frac{p_{i1}}{p_{11}}\frac{s_{i1} - r}{s_{11} - r} \quad \Leftrightarrow \quad p_i = k_it_ip_1, \quad \forall i, \tag{37b}$$

which proves (35).

Also,

$$-rp_iq_j = p_{ij}(s_{ij} - r) \quad \text{and} \quad -rp_1q_j = p_{1j}(s_{1j} - r). \tag{38a}$$

So,

$$\frac{p_i}{p_1} = k_it_i = \frac{p_{ij}}{p_{1j}}\frac{1}{c_{ij}} \quad \Leftrightarrow \quad p_{ij} = c_{ij}k_it_ip_{1j}, \quad \forall i,\ j, \tag{38b}$$

which proves (36). □

The Lemma tells us that the relevant variables are p_{1j} and p_{i1}, where p_{i1} occurs only as the ratio $\frac{p_{i1}}{p_{11}} = t_i$ for $i \neq 1$.

Proof of Theorem 2. The i-th row sum of the matrix P at equilibrium is:

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \dot{p}_{ij} = \sum_{j=1}^{n} (rp_i p_j + p_{ij}(s_{ij} - r)) && \text{(39a)} \\ &= rp_i \sum_{j=1}^{n} q_j - r \sum_{j=1}^{n} p_{ij} + \sum_{j=1}^{n} s_{ij} p_{ij} && \text{(39b)} \\ &= rp_i - rp_i + \sum_{j=1}^{n} s_{ij} p_{ij}, \text{ by (31d) and (31a)} && \text{(39c)} \\ 0 &= \sum_{j=1}^{n} s_{ij} p_{ij} && \text{(39d)} \\ &\Leftrightarrow \\ p_{in} &= -\frac{1}{s_{in}} \sum_{j=1}^{n-1} s_{ij} p_{ij}. && \text{(39e)} \end{aligned}$$

This means that the n-th column can be written as a sum of the other columns.

The j-th column sum of P is:

$$\begin{aligned} 0 &= \sum_{i=1}^{m} \dot{p}_{ij} = \sum_{i=1}^{m} (rp_i p_j + p_{ij}(s_{ij} - r)) && \text{(40a)} \\ &= rq_j \sum_{i=1}^{m} p_i - r \sum_{i=1}^{m} p_{ij} + \sum_{i=1}^{m} s_{ij} p_{ij} && \text{(40b)} \\ &= rq_j - rq_j + \sum_{i=1}^{m} s_{ij} p_{ij}, \text{ by (31d) and (31b)} && \text{(40c)} \\ &= p_{1j} \sum_{i=1}^{m} s_{ij} c_{ij} k_i t_i, \text{ by (36),} && \text{(40d)} \end{aligned}$$

which we write as

$$0 = \sum_{i=1}^{m} s_{ij} c_{ij} k_i t_i, \text{ by (32).} \tag{40e}$$

Since the n-th column can be written as the sum of the others, as established by (39e), (40e) holds for $1 \leq j \leq n - 1$.

Now, (40e) is a system of $n - 1$ linear equations in $m - 1$ variables t_i ($2 \leq i \leq m$). Since $t_1 = 1$, it is an inhomogenous system which we can write as

$$\sum_{i=2}^{m} c_{ij} s_{ij} k_i t_i = -s_{1j}. \tag{41}$$

For the following, it is more convenient to write (41) as

$$At = b, \tag{42}$$

where A is the $(n-1) \times (m-1)$ matrix with entries $A_{ij} = c_{ij}s_{ij}k_i$, t and b are the vectors $(t_2, ..., t_m)^T$ and $(-s_{11}, ..., -s_{1n})$, respectively.

Without loss of generality we assume $m < n$, which makes (40e),(41) an overdetermined linear system.

For general s_{ij} and r, the rank of the coefficient matrix A and of the augmented matrix $[A|b]$ is maximal. The maximal rank of a matrix is the minimum of the number of its rows and columns, namely $m-1$ for A and m for $[A|b]$, since it has an additional independent column. The Rouché–Capelli theorem states that there is a solution to a linear system of equations if and only if the rank of these two matrices is equal. Therefore, no solution exists for the t_i, as long as the fitness parameters and recombination rate are independent of each other. This implies in turn that no solution exists for (33) with the additional relations (31d), (34), (31a) and (31b). □

2.4 Proof of Theorem 1

The stronger version, Theorem 1, follows from this

Lemma 2. *Define the matrix* $\tilde{\mathbf{S}} = \left(\frac{s_{ij}-\bar{s}}{r+\bar{s}-s_{ij}}\right)_{ij}$ *and the normed vectors* $\mathbf{p} = (p_i)_i$ *and* $\mathbf{q} = (q_j)_j$ *with* $j = 1, ..., n$ *and* $i = 1, ..., m$.

Then (30) *has a solution satisfying* (32) *if and only if there exist* $\mathbf{p} > 0$ *and* $\mathbf{q} > 0$ *normed to one and a number* $\bar{s}$ *such that the following equations and inequalities hold for a positive* r.

$$\tilde{\mathbf{S}}\mathbf{q} = 0 \tag{43a}$$

$$\mathbf{p}^T\tilde{\mathbf{S}} = 0 \tag{43b}$$

$$r + \bar{s} - s_{ij} > 0 \quad \forall i, j. \tag{43c}$$

Proof. (⇒)

At equilibrium, the following identities follow from (30):

$$p_{ij} = \frac{rp_iq_j}{r+\bar{s}-s_{ij}} \quad \forall i, j \tag{44a}$$

$$\dot{p}_i = \sum_j \dot{p}_{ij} = \sum_j p_{ij}(s_{ij}-\bar{s}) = 0 \quad \forall i \tag{44b}$$

$$\dot{q}_j = \sum_i \dot{p}_{ij} = \sum_i p_{ij}(s_{ij}-\bar{s}) = 0 \quad \forall j. \tag{44c}$$

Plug (44a) into (44b) and (44c) to get:

$$\dot{p}_i = rp_i \sum_j \frac{s_{ij}-\bar{s}}{r+\bar{s}-s_{ij}} q_j = 0 \tag{45a}$$

$$\dot{q}_j = rq_j \sum_i \frac{s_{ij}-\bar{s}}{r+\bar{s}-s_{ij}} p_i = 0 \tag{45b}$$

If (32) holds, then it suffices to solve the two system of equations:

$$\sum_j \frac{s_{ij} - \bar{s}}{r + \bar{s} - s_{ij}} q_j = 0 = \tilde{S}q \quad \forall i, \tag{46a}$$

$$\sum_i \frac{s_{ij} - \bar{s}}{r + \bar{s} - s_{ij}} p_i = 0 = p^T \tilde{S} \quad \forall j. \tag{46b}$$

Note that (44a) together with (32) implies $r + \bar{s} - s_{ij} > 0 \quad \forall i, j$.

($\Leftarrow$)

Assume that $\mathbf{p} \in \mathbb{R}^m_+$ and $\mathbf{q} \in \mathbb{R}^n_+$ are normed to one and fulfill (46). Also, assume we have $r > 0$ and $\bar{s} \in \mathbb{R}$ such that $r + \bar{s} - s_{ij} > 0$ holds $\forall i, j$.

The question is, is this then an equilibrium for the haploid selection model? This means, we have to show that the equations of (31) are fulfilled.

Define $p_{ij} = \frac{rp_iq_j}{r+\bar{s}-s_{ij}}$, then (30) is true immediately.

Multiply the equations in (46) with q_j or p_i respectively. Applying the definition of p_{ij} yields

$$\sum_x (s_{ij} - \bar{s}) p_{ij} = 0, \tag{47}$$

where x is either i or j.

Then we sum (30) over x, and get

$$r(\delta_{xj} q_j + \delta_{xi} p_i - \sum_x p_{ij}) + \sum_x p_{ij}(s_{ij} - \bar{s}) = 0. \tag{48}$$

Using (47) simplifies (48) to

$$r(\delta_{xj} q_j + \delta_{xi} p_i - \sum_x p_{ij}) = 0, \tag{49}$$

which satisfies the claims if we take $x = i$ or $x = j$. This in turn implies

$$\sum_{ij} p_{ij} = 1. \tag{50}$$

Summing (30) over all i and j, yields

$$r(1 - \sum_{ij} p_{ij}) + \sum_{ij} p_{ij} s_{ij} - \bar{s} \sum_{ij} p_{ij} = 0 \tag{51a}$$

$$\Rightarrow \sum_{ij} p_{ij} s_{ij} = \bar{s}, \tag{51b}$$

which finishes the proof.

□

With this characterization of the polymorphism at hand, we can prove Theorem 1.

Proof of the Theorem. If $\tilde{\mathbf{S}}$ is such that there exists no vectors and values r and $\bar{s}$ such that all conditions in (43) are fulfilled, then the first part of the theorem is true.

However the following argument shows that if there is a solution that satisfies (43), then there are infinitely many of them provided $m \neq n$.

If we assume that $\mathbf{p}$, $\mathbf{q}$ and $\bar{s}$ exist that fulfill (43) for any $m < n$, then $\text{rk}(\tilde{S}) \leq m-1$. This implies because of the rank-nullity theorem that $\ker(\tilde{S}) \geq n - (m-1) \geq 2$. This means that at least one additional vector q' exists in the kernel of $\tilde{S}$. This solution vector does not necessarily lie in the simplex, however, $\frac{q+\epsilon q'}{1+\epsilon \sum_j q'_j}$ defines a one dimensional manifold that lies in the simplex for $0 < \epsilon < \epsilon^*$ with $\epsilon^* > 0$ sufficiently small and is a solution to (46). □

2.5 Equal number of alleles at the two loci

The Lemma 2 not only gives the means to show the impossibility of isolated polymorphisms for $m \neq n$, but also a simplified problem, where the solution(s) give the polymorphism(s) in the case of $m = n$. However it was not yet possible for me to show that any such solution is unstable. A stable solution that would falsify the claim raised in the introduction is also missing. Quite a large amount of time in Stanford was spent on invastigating this problem. Four possible prove-strategies were identified, but none yielded a answer.

The direct way of computing the eigenvalues of the Jacobian at any internal equilibrium did not help, since they cannot be computed explicitely for large matrices. The usual strategies of using only the trace or the determinant were also not helpful, since the trace has a indefinit sign and the determinant is too complicated to compute for larger matrices.

The more indirect way of using a index theorem of Hofbauer (1992) to infer the stability of internal equilibria from the stability of boundary equilibria, does, unfortunately, also not work. For the three- and three-alellel case for example, a maximum of three vertices can be stable. The work by Bank, Bürger and Hermisson (2012) ensures that in the two- and two-allele subsystems defined by the absence of one of the alleles on each locus, an unstable equilibrium exists. These are boundary equilibria in the full system. There are three of them, which can be saturated or not. If they all are saturated, then the index theorem tells us that at least one internal equilibrium has to exist. It could either be stable or have two positive eigenvalues. Numerical tests have only found equilibria with two positive eigenvalue (unstable).

For the two remaining strategies there exists no recipe and depend thus on luck. One would be to find a Lyapunov function that maximizes at equilibria. The usual candidates for such systems do not work. The second option would be to apply a coordinate transformation in which some derivation are simpler. However, no promising transformation was found.

The problem in the form of Lemma 2 should give a possibility to learn something about the number of equilibria. In fact $\det(\tilde{\mathbf{S}}) = 0$ is necessary for the existence of a solution for (43). In particular, $\det(\tilde{\mathbf{S}})$ depends as a function on $\bar{s}$. If $\det(\tilde{\mathbf{S}}) = 0$ for more than one value of $\bar{s}$, which all have corresponding positive $\mathbf{p}$ and $\mathbf{q}$, then there are more than one internal equilibria. No such example is known until now.

2.6 Discussion

We have conducted a rather elementary mathematical analysis of the haploid two-locus multiallele dynamics under constant selection. The model we use is the standard continuous-time model for haploid selection with recombination.

Two similar theorems are proven that settle the case when the two loci have different numbers of alleles. The proofs are different in the way how one derives the linear systems which determine the polymorphism and are overdetermined for $m \neq n$. The theorem that is proven first (Theorem 2), does not contain a statement about degenerate cases, since the corresponding linear system does not give a direct way of proving it. Theorem 2 does therefore, not in every case, preclude the existence of an isolated polymorphism. That is the reason why Theorem 1 is the stronger one.

For the proof of Theorem 1 a very useful and intuitive characterization of polymorphisms in terms of two linear homogeneous systems of equations is established (Lemma 2). Solvability of both systems in (43) is necessary and sufficient for the existence of internal equilibria. If the number of alleles is unequal among the two loci, then one of the systems is overdetermined and has in general no solution. However, in the degenerate case, where a solution exists, we could show that there is a manifold of them. This means, with more alleles at one of the locus than at the other, there is either no internal equilibrium or infinitely many of them.

If the number of alleles is equal on both loci, there can be at least an isolated equilibrium. We have examples for an unique and unstable polymorphism in the three alleles case. Despite investing a large amount of time, both during the stay with Prof. Feldman and afterwards, no thechnique was found that would enable me to prove the instability of any internal equilibrium as is expected by a huge fraction of the population genetics community. However, with the result presented now, it is clear that genetic variation, if it is maintained in a two locus fashion through haploid selection and recombination, only occurs with the same number (larger or equal to 3) of alleles from both loci present. If it is known that only two alleles occur on one locus, then genetic variation is always los regardless of the number of alleles on the other locus. In the end only a single monomorphism is present.

References

Bank C., Bürger R., Hermisson J. 2012. Limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-island model. Genetics, 191, 845-863.

Feldman M.W. 1971. Equilibrium studies of two locus haploid populations with recombination. Theor. Popul. Biol. 2, 299-318.

Kirzhner V.M. and Lyubich Y. 1997. Multilocus dynamics under haploid selection. J. Math. Biol. 35: 391-408

Nagylaki T. 1992. Introduction to Theoretical Population Genetics. Springer-Verlag

Novak S., Barton N.H. 2017. When does frequency-independent selection maintain genetic variation? Genetics 207: 653-668

Rutschman D. 1994. Dynamics of the two-locus haploid model. Theor. Popul. Biol. 45, 167-176.