

8 Additions: Time till absorption

Assume that both $x = 0$ and $x = 1$ are absorbing boundaries (e.g., corresponding to a loss or fixation of an allele). Let $\phi(t, p)$ be the density function of the time until absorption occurs given the initial value of x is p . Under diffusion approximation, $\phi(t, p)$ satisfies the equation

$$\frac{\partial \phi(t, p)}{\partial t} = a(p) \frac{\partial \phi(t, p)}{\partial p} + \frac{1}{2} b(p) \frac{\partial^2 \phi(t, p)}{\partial p^2}.$$

Let

$$\bar{t}(p) = \int_0^\infty t \phi(t; p) dt$$

be the average time till absorption. Note that $\bar{t}(0) = \bar{t}(1) = 0$ and that the first and second derivatives of $\bar{t}(p)$ are

$$\frac{d\bar{t}(p)}{dp} = \int_0^\infty t \frac{\partial \phi(t; p)}{\partial p} dt$$

and

$$\frac{d^2 \bar{t}(p)}{dp^2} = \int_0^\infty t \frac{\partial^2 \phi(t; p)}{\partial p^2} dt.$$

Following Ewens (1979,2004) and using integration by parts,

$$\begin{aligned} -1 &= - \int_0^\infty \phi(t; p) dt = \\ &= -[t\phi(t, p)]_0^\infty + \int_0^\infty t \frac{\partial \phi}{\partial t} dt \\ &= 0 + \int_0^\infty t \left[a(p) \frac{\partial \phi}{\partial p} + \frac{1}{2} b(p) \frac{\partial^2 \phi}{\partial p^2} \right] dt, \end{aligned}$$

so that

$$1 + a(p) \frac{d\bar{t}(p)}{dp} + \frac{1}{2} b(p) \frac{d^2 \bar{t}(p)}{dp^2} = 0. \quad (46a)$$

Thus, the average time till absorption can be found by solving the second order ordinary differential equation (1a) subject to boundary conditions

$$\bar{t}(0) = \bar{t}(1) = 0. \quad (46b)$$

9 Additions: Evolution of holey landscapes

Let x_i be the frequency of the i -th genotype and w_i be its fitness (that is, the expected number of offspring). Then after selection

$$x'_i = \frac{w_i}{\bar{w}} x_i, \quad (47a)$$

where

$$\bar{w} = \sum_i w_i x_i \quad (47b)$$

is the average fitness of the population. Assume that individuals surviving selection produce offspring which may differ from its parents as a result of mutations. The genotype frequencies in offspring (that is, at the beginning of a new generation) are

$$x''_i = (1 - \mu_i) x'_i + \sum_{j \neq i} \mu_{ji} x'_j, \quad (48)$$

where μ_{ji} is the probability that genotype j mutates to genotype i and $\mu_i = \sum_{j \neq i} \mu_{ij}$ is the overall probability of mutation for genotype i . In equation (48), the first term in the right-hand side gives the frequency of genotypes i that did not mutate and the second term gives the frequency of genotypes i formed by mutation of other genotypes.

9.1 Derivations of Error threshold

Let us consider an asexual population of haploid individuals. I will assume that the fitness landscape has a percolating neutral network of genotypes with fitness 1 and that all other genotypes have a reduced fitness $1 - s$, where $s > 0$. Let ν be the overall probability of mutation per genotype per generation. Asymptotically, the population is expected to reach a mutation-selection balance. Let $\hat{\mathcal{N}}$ be the average number of high-fitness one-step neighbors per high-fitness individual in the population. I will refer to $\hat{\mathcal{N}}$ as the *population neutrality*.

Let variable x stand for the overall frequency of genotypes with fitness 1, and let variable y stand for the overall frequency of all other genotypes, with $x + y = 1$. Then, after selection,

$$x' = \frac{1}{\bar{w}} x, \quad (49a)$$

where $\bar{w} = 1 - sy$ is the average fitness of the population. Assume that mutations at different loci have equal probabilities and disregard the probability of more than one mutation per genotype.

Then, the overall frequency of high-fitness genotypes after mutation is

$$x'' = x' - \nu \left(1 - \frac{\hat{\mathcal{N}}}{\mathcal{D}}\right) x' \quad (49b)$$

where we neglected mutations into the high-fitness network from the low-fitness genotypes. This assumption is justified because the frequency of low fitness genotypes is expected to be low.

Equations (49) control the dynamics of genotype frequencies in time. It is easy to see that if

$$\nu < \nu_c = \frac{s}{1 - \hat{\mathcal{N}}/\mathcal{D}}, \quad (50)$$

that is, if mutation rate is sufficiently small, then the equilibrium frequency of high-fitness genotypes is positive and can be written as

$$x^* = 1 - \frac{\nu(1 - \hat{\mathcal{N}}/\mathcal{D})}{s}. \quad (51)$$

If condition (50) is not satisfied, then the only equilibrium state is the one with no high-fitness genotypes present: $x^* = 0$. This shows that the optimum sequence can be maintained in the population only if ν is not too big (i.e., if

there is minimal replication accuracy). If the mutation rate per locus μ or the sequence length \mathcal{L} is too large, the optimum sequence will be lost. This transition between two different regimes is known as the *error threshold*. Both ν_c and ϕ^* increase with $\hat{\mathcal{N}}$. Thus, with neutrality, genotypes with the highest fitness can be maintained at larger mutation rates and at higher frequencies than in the case of no neutral mutations. In other words, the existence of neutral mutations increases the error threshold.

Using equation (51), one also finds the average fitness of the population at equilibrium:

$$\bar{w} = 1 - \nu \left(1 - \frac{\hat{\mathcal{N}}}{\mathcal{D}}\right). \quad (52)$$

The last equation shows that the average fitness of the population increases with increasing population neutrality $\hat{\mathcal{N}}$. Note that we reached these conclusions even without the knowledge of the exact value of the population neutrality $\hat{\mathcal{N}}$. The value of $\hat{\mathcal{N}}$ will be found in the next section, where we consider genetic canalization.

9.2 Derivations of equation 4.14 (Genetic canalization)

Let us consider the same model as in the previous section, but now I will concentrate on the frequencies of different high-fitness genotypes, which will be denoted as x_i , rather than on just the overall frequency $\phi = \sum \phi_i$ of such genotypes. Let $x^* = (x_1^*, x_2^*, \dots)^T$ be a column vector of equilibrium genotype frequencies. Let us also define a matrix M with the element M_{ij} equal to one if the genotypes i and j are one-step neighbors and equal to zero otherwise. In the terminology of graph theory, M is the *adjacency matrix* for the network of viable genotypes.

The frequency of the i -th viable genotype after selection is

$$x'_i = \frac{1}{\bar{w}} x_i, \quad (53a)$$

where \bar{w} is the average fitness of the population. With the same assumptions regarding mutation as in the previous subsection, the mutation rate per gamete is $\mu_i = \nu$ and the probability of allele j mutating to allele i is $\mu_{ji} = \nu/\mathcal{D}$ if the genotypes j and i are one-step neighbors and is zero otherwise. Then after mutation

$$x''_i = (1 - \nu)x'_i + \sum_{j \in V_i} \frac{\nu}{\mathcal{D}} x'_j, \quad (53b)$$

where V_i is the set of high-fitness one-step neighbors of a high-fitness genotype i . Combining equations (53), one can see that the genotype frequencies x_i^* at mutation-selection balance must satisfy a system of algebraic equations

$$\frac{\mathcal{D}(\bar{w} - 1 + \nu)}{\nu} x_i^* = \sum_{j \in V_i} x_j^*. \quad (54)$$

At this state, the average fitness of the population is given by equation (52). Using this equation, it is easy to see that the ratio in the left-hand side of the above equation is equal to the population neutrality $\hat{\mathcal{N}}$. This allows one to rewrite equations (54) in matrix notation as

$$\hat{\mathcal{N}} x^* = M x^*. \quad (55)$$

9.3 Additions: Molecular clock

Zuckerandl and Pauling (1962): the total number of amino acid substitutions in a particular locus over a period of t years has a Poisson distribution with mean λt :

$$Prob(n = i) = \frac{\exp(-\lambda t)(\lambda t)^i}{i!},$$

where λ is the mean rate of substitutions per locus per year.

Deviations: (a) lineage effect, (b) non-Poisson distribution.

Index of dispersion as a measure of departure from a Poisson process:

$$I = \frac{var(n)}{\bar{n}}$$

For a Poisson distribution, $I = 1$.

Kimura and Ohta (1971): statistical analysis of the constance of molecular clock. α -globin, β -globin and cytochrome c : the variation is larger than expected by chance.

Kimura (1983): α -globin ($I = 1.3$), β -globin ($I = 3.1$), myoglobin ($I = 1.7$), cytochrome c ($I = 3.3$), ribonuclease ($I = 2.4$).

Gillespie (1991): many more proteins; I ranges between 0.2 and 34.1; molecular evolution as an “episodic” process in which a burst of amino acid substitutions is followed by a static phase.

Ohta: can be explained by variations in population sizes and nearly neutral mutations

Gillespie: can be explained by fluctuating selection

10 Multilocus models

10.1 Two loci: notations

Here, we will consider not one but two different loci simultaneously. We consider diallelic loci with alleles **A**, **a** at the first locus, and alleles **B**, **b** at the second locus. Each individual has 4 genes (with two possible alleles each) resulting in that 16 different genotypes are possible. These genotypes are formed by four different gametes: **AB**, **Ab**, **aB**, **ab**, which frequencies we will denote x_1, x_2, x_3 and x_4 , respectively ($\sum x_i = 1$). The frequencies of alleles can be computed from the frequencies of gametes: the frequency of **A** is $p_1 = x_1 + x_2$, the frequency of **a** is $q_1 = x_3 + x_4$; the frequency of **B** is $p_2 = x_1 + x_3$, and the frequency of **b** is $q_2 = x_2 + x_4$.

If alleles were combined into gametes *randomly*, the following equalities would be true: $x_1 = p_1 p_2, x_2 = p_1 q_2$ etc. *Linkage disequilibrium*

$$D = x_1 - p_1 p_2$$

measures the deviation from randomness. Equivalently, linkage disequilibrium can be defined as

$$D = x_1 x_4 - x_2 x_3.$$

Homework: prove the equivalence of these two definitions.

Yet another way to define D is by using *indicator variables* l_1 and l_2 defined for a gamete: $l_1 = 1$ if the allele at the first locus is **A**, $l_1 = 0$ if the allele at the first locus is **a**, $l_2 = 1$ if the allele at the second locus is **B**, and $l_2 = 0$ if the allele at the second locus is **b**. Note $l_i^2 = l_i$. Calculating the mean values (over the population)

$$\begin{aligned} E\{l_1\} &= 1 \times p_1 + 0 \times q_1 = p_1, \\ E\{l_2\} &= 1 \times p_2 + 0 \times q_2 = p_2, \\ E\{l_1 l_2\} &= 1 \times x_1 + 0 \times x_2 + 0 \times x_3 + 0 \times x_4 = x_1. \end{aligned}$$

Thus,

$$D = x_1 - p_1 p_2 = E\{l_1 l_2\} - E\{l_1\} E\{l_2\} = \text{cov}(l_1, l_2).$$

Linkage disequilibrium D measures covariance of l_1 and l_2 . Note that $\text{var}\{l_i\} = E\{(l_i - p_i)^2\} = \dots = p_i q_i$ is a measure of genetic variation at a locus. Gamete frequencies can be represented in terms of allele frequencies and linkage disequilibrium:

$$\begin{aligned} x_1 &= p_1 p_2 + D, \\ x_2 &= p_1 - x_1 = p_1 q_2 - D, \\ x_3 &= p_2 - x_1 = q_1 p_2 - D, \\ x_4 &= q_1 q_2 + D. \end{aligned}$$

In general, a set of gamete frequencies (x_1, x_2, x_3, x_4) and a set of allele frequencies and linkage disequilibrium (p_1, p_2, D) provide two alternative ways to characterize genetic structure of populations.

Under random mating, the genotype frequencies are equal to the products of the corresponding gamete frequencies:

$$X_{ij} = x_i x_j.$$

That is zygotic Hardy-Weinberg proportions are attained in a single generation.

10.2 Crossing-over and recombination

Crossing-over is an exchange of portions of chromatids between homologous chromosomes. Recombination is the formation of a non-parental gamete from the maternal alleles at a set of loci and the paternal alleles at the remaining loci. Recombination is a result of crossing-over. Recombination changes gamete frequencies but not allele frequencies.

Double heterozygotes can produce 4 different gametes. Double heterozygotes **AB/ab** will produce gametes **AB** and **ab** if there is no recombination, and gametes **Ab** and **aB** if there is recombination. In a similar way, double heterozygotes **Ab/aB** will produce gametes **Ab** and **aB** if there is no recombination, and gametes **AB** and **ab** if there is recombination. Let r be the probability of recombination (defined as the proportion of recombinant gametes; $0 \leq r \leq 1/2$). If $r = 1/2$, the genes are *unlinked* (they are on different chromosomes). If $r < 1/2$, the genes are *linked* (they are on the same chromosome). For linked loci, the recombination rate depends on the physical distance between the loci: the closer the loci, the smaller the rate of recombination.

Let x_1, p_1 and p_2 be the frequencies of gamete **AB** and alleles **A** and **B** in this generation. In the next generation

$$x'_1 = (1 - r)x_1 + rp_1p_2,$$

where the first term in the right-hand side is the frequency of gametes **AB** that were not destroyed by recombination, and the second term is the frequency of gametes **AB** that were created by recombination. Thus, $x'_1 - p'_1p'_2 = (1 - r)(x_1 - p_1p_2)$, which can be rewritten as $D' = (1 - r)D$. This shows that linkage disequilibrium D decays exponentially:

$$D_t = (1 - r)^t D_0$$

With no other factors the population approaches a state of *linkage equilibrium* at that $D = 0$. If $D = 0$, then the gamete frequencies are in *Robbins* (1918) proportions: $x_1 = p_1p_2, x_2 = p_1q_2, x_3 = q_1p_2, x_4 = q_1q_2$. These equalities mean that the alleles are combined into gametes randomly. The time scale for attaining linkage equilibrium between two loci is $1/r$ generations where r is the rate of recombination between the loci. The genetic structure of populations at linkage equilibrium can be completely characterized in terms of allele frequencies.

10.2.1 Major points

In randomly mating diploid populations

- allele frequencies do not change
- genotype frequencies attain Hardy-Weinberg proportions in one generation and do not change after that;
- linkage equilibrium is attained asymptotically;
- the time scale for attaining linkage equilibrium between two loci is $1/r$ generations where r is the rate of recombination between the loci.

11 Selection in Multilocus models

Let x_i be the frequency of gamete i . With random mating the frequency of genotype i/j is $z_{ij} = x_i x_j$. Let w_{ij} be fitness (viability) of genotype i/j . Then in the next generation

$$x'_k = \frac{\sum_{i,j} x_i x_j w_{ij} R(i, j \rightarrow k)}{\bar{w}}, \quad (56)$$

where $R(i, j \rightarrow k)$ is the probability that genotype i/j produces gamete k (as a result of segregation/recombination), and $\bar{w} = \sum_{i,j} w_{ij} x_i x_j$ is the mean fitness of the population. This is a very general equation valid for any number of loci and any number of alleles. However, this equation is not easy to use.

11.1 Two-locus two-allele models: viability selection

Here we consider a simpler two-locus two-allele case. There are four gametes: **AB**, **Ab**, **aB**, **ab**. Let x_1, x_2, x_3 , and x_4 be their frequencies ($\sum x_i = 1$). There are $4 \times 4 = 16$ different genotypes. Let z_{ij} be the frequency of a genotype formed by gametes i and j . After random mating, $z_{ij} = x_i x_j$. Selection can be specified by a viability matrix

	AB	Ab	aB	ab
AB	w_{11}	w_{12}	w_{13}	w_{14}
Ab	w_{21}	w_{22}	w_{23}	w_{24}
aB	w_{31}	w_{32}	w_{33}	w_{34}
ab	w_{41}	w_{42}	w_{43}	w_{44}

There are some natural symmetries in this matrix. In particular, if there are no maternal/paternal effects, then $w_{ij} = w_{ji}$, and if there are no cis-trans effects, then $w_{14} = w_{23}$. These symmetries allow one to represent the viability matrix in an alternative way:

	AA	Aa	aa
BB	v_{11}	v_{12}	v_{13}
Bb	v_{21}	v_{22}	v_{23}
bb	v_{31}	v_{23}	v_{33}

where $v_{11} = w_{11}, v_{12} = w_{12}$ etc.

To derive the dynamic equations for gamete frequencies one needs to consider all possible matings and resulting offspring (see Table 1). In doing so, one has to keep in mind that because of recombination, double heterozygotes will produce recombinant gametes. For example, genotype **AB/ab** produces non-recombinant gametes **AB** and **ab** with probability $1 - r$, and recombinant gametes **aB** and **Ab** with probability r , where r is the rate of recombination.

Table 1. Gamete production table.

Genotype	Frequency	Viability	Gametes produced			
			AB	Ab	aB	ab
AB/AB	x_1^2	w_{11}	1	0	0	0
AB/Ab	$2x_1x_2$	w_{12}	1/2	1/2	0	0
AB/aB	$2x_1x_3$	w_{13}	1/2	0	1/2	0
AB/ab	$2x_1x_4$	w_{14}	$(1-r)/2$	$r/2$	$r/2$	$(1-r)/2$
...
Ab/aB	$2x_2x_3$	w_{23}	$r/2$	$(1-r)/2$	$(1-r)/2$	$r/2$
...

Using Table 1, one finds that in the next generation:

$$\begin{aligned}
 x'_1 &= \frac{w_{11}x_1^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + (1-r)w_{14}x_1x_4 + rw_{23}x_2x_3}{\bar{w}} \\
 &= \frac{w_{11}x_1^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + w_{14}x_1x_4 - rw_{14}x_1x_4 + rw_{23}x_2x_3}{\bar{w}} \\
 &= \frac{x_1(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4) - rw_{14}(x_1x_4 - x_2x_3)}{\bar{w}} \\
 &= \frac{x_1w_1 - rw_{14}D}{\bar{w}},
 \end{aligned}$$

where $w_1 = \sum w_{1i}x_i$ is the induced fitness of gamete **AB**, and $D = x_1x_4 - x_2x_3$ is linkage disequilibrium. $\bar{w} = \sum w_{ij}x_i x_j$ is the mean fitness of the populations.

The general dynamic equations for two-locus two-allele systems can be written as

$$x'_i = \frac{w_i x_i}{\bar{w}} \pm \frac{r w_{14} D}{\bar{w}}, \quad (57)$$

where the sign is + for $i = 2, 3$ and is - for $i = 1, 4$ (Lewontin and Kojima, 1960).

Note that if there is no recombination (if $r = 0$), the dynamics are the same as that of an one-locus four-allele system. We also know that if there is no selection (if $w_{ij} = const$), then linkage disequilibrium $D \rightarrow 0$ exponentially.

Conditions for stability of monomorphic equilibria can be found easily (use *Maple*). For example,

AB/AB is stable if $w_{11} > w_{12}, w_{13}, (1-r)w_{14}$ that is if $v_{11} > v_{12}, v_{21}, (1-r)v_{22}$, and

ab/ab is stable if $w_{44} > w_{42}, w_{43}, (1-r)w_{41}$ that is if $v_{33} > v_{32}, v_{23}, (1-r)v_{22}$; etc

Note that increasing recombination rate r makes monomorphic equilibria “more stable”. If a monomorphic equilibrium is stable with no recombination, it is stable with any positive r .

If all four monomorphic equilibria are unstable, genetic variation is maintained in at least one locus.

11.1.1 Additive fitnesses: $v_{ij} = \alpha_i + \beta_j$.

Here we assume that the loci contribute additively to fitness. Note that this assumption does not necessarily imply the additivity of allele contributions. The viability matrix can be represented as

	AA	Aa	aa
BB	$\alpha_1 + \beta_1$	$\alpha_2 + \beta_1$	$\alpha_3 + \beta_1$
Bb	$\alpha_1 + \beta_2$	$\alpha_2 + \beta_2$	$\alpha_3 + \beta_2$
bb	$\alpha_1 + \beta_3$	$\alpha_2 + \beta_3$	$\alpha_3 + \beta_3$

At equilibrium, $x_i = x'_i$ and

$$\begin{aligned}\bar{w}x_1 &= w_1x_1 - rw_{14}D, & | \times (1/x_1), \\ \bar{w}x_2 &= w_2x_2 + rw_{14}D, & | \times (-1/x_2), \\ \bar{w}x_3 &= w_3x_3 + rw_{14}D, & | \times (-1/x_3), \\ \bar{w}x_4 &= w_4x_4 - rw_{14}D, & | \times (1/x_4).\end{aligned}$$

Multiplying each of these equations by the term indicated and summing up the results one finds that an equality

$$0 = w_1 - w_2 - w_3 + w_4 - rw_{14}D\left(\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}\right)$$

must be true. But

$$w_1 - w_2 - w_3 + w_4 = (\text{use Maple}) = 0,$$

thus, at equilibrium $D = 0$ (the system is at linkage equilibrium).

Let p_1 and p_2 be the frequencies of **A** and **B**. One can show that the mean fitness of the population is

$$\bar{w} = (\alpha_1p_1^2 + 2\alpha_2p_1q_1 + \alpha_3q_1^2) + (\beta_1p_2^2 + 2\beta_2p_2q_2 + \beta_3q_2^2).$$

The mean fitness does not depend on linkage disequilibrium D .

Equations (57) can be used to derive the dynamic equations for allele frequencies. Because $p_1 = x_1 + x_2$ and $p_2 = x_1 + x_3$ the equations for allele frequencies are independent of r . In particular, the allele frequencies will change in exactly the same way as if recombination was absent ($r = 0$). Therefore, the dynamics of the mean fitness \bar{w} does not depend on r . It follows that the mean fitness in non-decreasing and $\Delta\bar{w} = 0$ only at equilibria (because this is how \bar{w} behaves if $r = 0$).

It is straightforward to show that genetic variation in a locus is maintained only if the locus is overdominant. In particular, the doubly polymorphic equilibrium with

$$p_1^* = \frac{\alpha_2 - \alpha_3}{2\alpha_2 - \alpha_1 - \alpha_3}, \quad p_2^* = \frac{\beta_2 - \beta_3}{2\beta_2 - \beta_1 - \beta_3}$$

exists and is globally stable if

$$\alpha_2 > \alpha_1, \alpha_3, \quad \beta_2 > \beta_1, \beta_3$$

that is if there is overdominance in the both loci.

11.1.2 Additive fitnesses with n loci

In the multilocus case, additive fitnesses can be described as

$$w = \sum_{i=1}^n w_i,$$

where $w_i = \alpha_{1i}, \alpha_{2i}$ and α_{3i} if the genotype at the i -th locus is A_iA_i, A_ia_i and a_ia_i , respectively.

At equilibrium, gamete frequencies are products of the corresponding allele frequencies, e.g. $freq(A_iA_ja_k) = p_i^*p_j^*q_k^*$. Genetic variability will be maintained only in the overdominant loci, that is at the loci with $\alpha_{i2} > \alpha_{1i}, \alpha_{3i}$.

11.1.3 Multiplicative fitnesses: $v_{ij} = \alpha_i\beta_j$.

Here we assume that the loci contribute multiplicatively to fitness. Note that this assumption does not necessarily imply the multiplicativity of allele contributions. The viability matrix can be represented as

	AA	Aa	aa
BB	$\alpha_1\beta_1$	$\alpha_2\beta_1$	$\alpha_3\beta_1$
Bb	$\alpha_1\beta_2$	$\alpha_2\beta_2$	$\alpha_3\beta_2$
bb	$\alpha_1\beta_3$	$\alpha_2\beta_3$	$\alpha_3\beta_3$

In the multiplicative case, the overdominance in a locus is necessary and sufficient for the maintenance of genetic variation in this locus.

To illustrate other features of the multiplicative selection model, we consider a special case of within-locus overdominance with $\alpha_2 = \beta_2 = 1, \alpha_1 = \alpha_3 = \beta_1 = \beta_3 = 1 - s, s > 0$.

	AA	Aa	aa
BB	$(1 - s)^2$	$1 - s$	$(1 - s)^2$
Bb	$1 - s$	1	$1 - s$
bb	$(1 - s)^2$	$1 - s$	$(1 - s)^2$

Here, the monomorphic equilibria are unstable for any $s > 0$ and any r (why?), and, thus, genetic variation is protected.

From the symmetry, one expects that at equilibrium

$$p_1^* = p_2^* = 1/2.$$

At such equilibria the gamete frequencies are $x_1 = 1/4 - D, x_2 = 1/4 + D, x_3 = 1/4 + D, x_4 = 1/4 - D$, where D is the corresponding linkage disequilibrium. Note that $x_1 = x_4$ and $x_2 = x_3$ that is the complementary gametes have the same frequency.

One can show that there are 3 possible D values and, thus, three possible equilibria:

- An equilibrium with $D = 0$ and gamete frequencies

$$x_1 = x_2 = x_3 = x_4 = 1/4.$$

This equilibrium exists always (*Maple*). It is locally stable if (*Maple*)

$$r > s^2/4$$

that is if recombination rate is sufficiently high.

- A pair of equilibria with

$$D^* = \pm \frac{1}{4} \sqrt{1 - 4 \frac{r}{s^2}}.$$

At these equilibria, the gamete frequencies are as follows. At equilibrium D^+ ,

$$x_1 = x_4 = \frac{1}{4} + \frac{1}{4} \sqrt{1 - 4 \frac{r}{s^2}}$$

$$x_2 = x_3 = \frac{1}{4} - \frac{1}{4} \sqrt{1 - 4 \frac{r}{s^2}}$$

At equilibrium D^- ,

$$x_1 = x_4 = \frac{1}{4} - \frac{1}{4} \sqrt{1 - 4 \frac{r}{s^2}}$$

$$x_2 = x_3 = \frac{1}{4} + \frac{1}{4} \sqrt{1 - 4 \frac{r}{s^2}}$$

These equilibria exist and are stable if (*Maple*)

$$r < s^2/4.$$

that is if linkage is sufficiently tight.

Bifurcation diagram.

Example. With $s = .1$, equilibria D^+ and D^- exist and are stable if $r < .0025$ (tight linkage). The equilibrium gamete frequencies at D^+ for different recombination rates are

r	$x_1(\mathbf{AB})$	$x_2(\mathbf{Ab})$	$x_3(\mathbf{aB})$	$x_4(\mathbf{ab})$
.002	.36	.14	.14	.36
.0005	.47	.03	.03	.47
.0001	.49	.01	.01	.49

Recall that the genotype frequencies are $z_{ij} = x_i x_j$. Thus, with $r = .0005$ the most common genotypes will be **AA/BB**, **Aa/Bb** and **ab/ab**. Although only three genotypes out of ten possible will be observed in the population, there is a lot of hidden genetic variability. If selection is relaxed, recombination will quickly recreate all possible genotypes.

In the symmetric case (with $\alpha_1 = \alpha_3, \beta_1 = \beta_3$), at a stable doubly polymorphic equilibrium $D = 0$ if $r > r_c$ and $D \neq 0$ if $r < r_c$.

In the asymmetric case: "Ewens gap" (Ewens, 1968) and "overlap" (Franklin and Feldman, 1977; Karlin and Feldman, 1978; Hastings, 1981).

11.1.4 Multiplicative fitnesses with n loci

In the multilocus case, multiplicative fitnesses can be described as

$$w = \prod_{i=1}^n w_i,$$

where $w_i = \alpha_{1i}, \alpha_{2i}$ and α_{3i} if the genotype at the i -th locus is $A_i A_i, A_i a_i$ and $a_i a_i$, respectively. Genetic variability will be maintained only in the overdominant loci, that is at the loci with $\alpha_{i2} > \alpha_{1i}, \alpha_{3i}$.

11.1.5 Symmetric fitnesses:

	AA	Aa	aa
BB	$1 - \delta$	$1 - \beta$	$1 - \alpha$
Bb	$1 - \gamma$	1	$1 - \gamma$
bb	$1 - \alpha$	$1 - \beta$	$1 - \delta$

Symmetric fitnesses model exhibits *epistasis* in fitness. *Epistasis* usually means interactive effects between loci. Sometimes it is defined as non-additivity of the loci effects on the trait under consideration (e.g., fitness). Alternatively, one says there is epistasis if the contribution of a locus to a trait depends on genetic background.

Symmetric fitness model has two types of doubly polymorphic equilibria. The *symmetric equilibria* are of the form

$$x_1 = x_4 = 1/4 + D, x_2 = x_3 = 1/4 - D$$

($p_1 = p_2 = 1/2$), where the values of D are solutions of the cubic equation

$$64lD^3 - 16(\delta - \alpha)D^2 - 4(l - 8r)D + (\delta - \alpha) = 0,$$

where $l = 2(\beta + \gamma) - (\alpha + \delta)$.

In addition to these equilibria, there may exist up to four *unsymmetric polymorphic equilibria* with $x_1 \neq x_4$ and/or $x_2 \neq x_3$.

Results on *stable* equilibria:

- there can be four monomorphic equilibria plus one polymorphic equilibrium or 4 monomorphic equilibria plus 2 polymorphic equilibria (Feldman and Liberman, 1979);
- there can be four polymorphic equilibria (Hastings, 1985). If $\alpha > r$, $\alpha - r \ll 1$, and $|\beta - \gamma| > \alpha$. Example: $\alpha = .1, \beta = .2, \gamma = .4, r = .09$. Allele frequencies and linkage disequilibrium:

p_1	p_2	D
.125	.169	.074
.125	.831	-.074
.875	.169	-.074
.875	.831	.074

- equilibria with $D = 0$ and $D \neq 0$ can be stable simultaneously

11.1.6 Stabilizing selection on an additive quantitative trait.

Quantitative traits are phenotypic characters that exhibit *continuous* variation (e.g. size, weight etc.). Genetic variation in these characters is usually based on many loci of small effect.

We assume that the genes contribute additively to a certain (quantitative) trait z . Let a_1 and $-a_1$ be the contributions of alleles **A** and **a**, and a_2 and $-a_2$ be the contributions of alleles **B** and **b**. The values of trait values z are defined by matrix

	AA	Aa	aa
BB	$a_1 + a_2$	a_1	$a_1 - a_2$
Bb	a_2	0	$-a_2$
bb	$-a_1 + a_2$	$-a_1$	$-a_1 - a_2$

We assume that fitness depends of *phenotype* z rather than on genotype, that is $w = w(z)$. We consider the case of *stabilizing* selection which we describe using a quadratic fitness function

$$w(z) = 1 - sz^2,$$

where parameters s characterizes the strength of selection, and it is assumed that the optimum phenotype is zero. The notion of stabilizing selection reflects the fact that many quantitative traits have an intermediate optimum with natural selection acting against extreme phenotypes.

Quadratic stabilizing selection on an additive trait results in symmetric fitness scheme with

$$\alpha = s(a_1 - a_2)^2, \quad \beta = sa_1^2, \quad \gamma = sa_2^2, \quad \delta = s(a_1 + a_2)^2.$$

Conditions for existence and stability of different equilibria in this model are shown in Fig. 1.

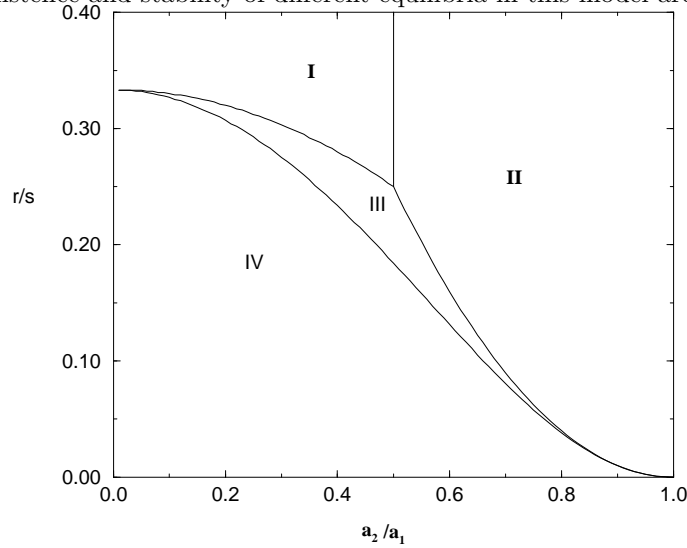


Figure 4: **Regions of stability of different equilibria on the $(a_2/a_1, r/s)$ plane. The roman numerals denote the regions where the following equilibria exist and are stable: I - singly polymorphic equilibria $(x_1, 0, x_3, 0)$ and $(0, x_2, 0, x_4)$; II - monomorphic equilibria $(0, 1, 0, 0)$ and $(0, 0, 1, 0)$; III - a pair of “unsymmetric” doubly polymorphic equilibria; IV - “symmetric” doubly polymorphic equilibrium.**

A surprising conclusion is that strong stabilizing selection (small r/s) can maintain genetic variation in both loci whereas this is not possible with weak selection (large r/s).

11.1.7 General fitnesses

- induced overdominance: counterexamples (Hastings 1982),
- cycles are possible (Akin 1982, Hastings 1981),
- limit on D : $r|D| < (\Delta w)_{max}/10$ (Hastings 1986).
- numerical results with random fitnesses (Turelli and Ginzburg 1983):

Only 6.4% of fitness matrices result in stable two-locus polymorphism (with unlinked loci); in 75% of these cases $\bar{w}_2 > \bar{w}_1 > \bar{w}_0$ where \bar{w}_i is the average fitness of genotypes with i heterozygous loci; in 97%, $\bar{w}_2 > \bar{w}_0$.

Increasing linkage (decreasing r) increases the probability of polymorphism:

r	%
.5	6.4
.1	11.7
.01	15.75

In three-locus two-allele models only .6% of fitness matrices have resulted in stable equilibria with all 8 gametes present.

- In general, with diallelic loci

# of loci	# of polymorphic equilibria
1	1
2	7
3	193
4	63,775
5	4,294,321,153

11.2 Weak selection (linkage equilibrium) approximation

Let us consider a *haploid* population with individuals different with respect to n diallelic loci. We will use indicator variables

$$l_i = \begin{cases} 1 & \text{if } A_i, \\ 0 & \text{if } a_i. \end{cases} \quad (58)$$

Recall that $l_i^2 = l_i$ and that $E\{l_i\} = p_i$ (the frequency of allele \mathbf{A}_i in the population). Any fitness configuration can be represented as

$$w = \mu + \sum_i a_i l_i + \sum_{i,j} b_{ij} l_i l_j + \sum_{i,j,k} c_{ijk} l_i l_j l_k + \cdots + d l_1 l_2 \dots l_n.$$

Here a_i are additive effects, b_{ij} are *pairwise epistatic* effects etc. Collecting terms that include l_i and those that do not:

$$w = l_i A(l_1, \dots, l_{i-1}, l_{i+1}, \dots, l_n) + B(l_1, \dots, l_{i-1}, l_{i+1}, \dots, l_n). \quad (59)$$

The general equation for the change in allele frequency, which is valid for any strength of selection, is

$$\Delta p_i = \frac{w_{A_i} - \bar{w}}{\bar{w}},$$

where w_{A_i} is the induced fitness of allele \mathbf{A}_i (that is the average fitness of genotypes having this allele). If selection is weak relative to selection, the system approaches the state of linkage equilibrium. Assuming linkage equilibrium (that is the independence of indicator variables: $E\{l_i l_j\} = E\{l_i\}E\{l_j\}$ etc.) and taking the expectation of both sides of equation (59)

$$E\{w\} \equiv \bar{w} = E\{l_i\}E\{A\} + E\{B\} \equiv p_i \bar{A} + \bar{B}.$$

Note that

$$\frac{\partial \bar{w}}{\partial p_i} = \bar{A}.$$

The induced fitness of allele \mathbf{A}_i can be represented as

$$w_{A_i} = E\{w|l_i = 1\} = 1 \times \bar{A} + \bar{B} = \bar{A} + \bar{B},$$

and, thus,

$$w_{A_i} - \bar{w} = (\bar{A} + \bar{B}) - (p_i \bar{A} + \bar{B}) = q_i \bar{A} = p_i \frac{\partial \bar{w}}{\partial p_i}.$$

Summarizing,

$$\Delta p_i = p_i q_i \frac{\partial \ln \bar{w}}{\partial p_i}.$$

In the diploid case, a similar procedure leads to

$$\Delta p_i = \frac{p_i q_i}{2} \frac{\partial \ln \bar{w}}{\partial p_i}$$

(Wright, 1935). Gradient-type dynamics (because $\Delta \bar{w} \geq 0$): evolution towards equilibrium, no cycles, no chaos.

Example. Diploid population with n loci. Two sets of indicator variables: l_i (for maternal genes) and l'_i (for paternal genes).

A general class of fitness functions that include additive (a), dominant (b) and pairwise additive-by-additive epistatic (c) effects:

$$w = \mu + \sum_i [a(l_i + l'_i) + 2bl_i l'_i] + \sum_{i \neq j} c(l_i + l'_i)(l_j + l'_j).$$

The mean fitness of the population (assuming linkage equilibrium)

$$\bar{w} = \mu + \sum_i [2ap_i + 2bp_i^2] + \sum_{i \neq j} 4cp_i p_j.$$

Thus,

$$\frac{\partial \bar{w}}{\partial p_i} = 2a + 4bp_i + 8c \sum_{j \neq i} p_j.$$

The dynamics of allele frequencies are described by

$$\dot{p}_i = p_i q_i (a + \sum_j S_{ij} p_j),$$

where

$$S_{ii} = 2b, \quad S_{ij} = 4c, \quad i \neq j.$$

A unique completely polymorphic equilibrium with

$$p_i = p^* \equiv \frac{a}{4c(1-n) - 2b}$$

exists if $0 < p^* < 1$. Stability matrix

$$S = \begin{pmatrix} 2b & 4c & 4c & \dots & 4c & 4c \\ 4c & 2b & 4c & \dots & 4c & 4c \\ 4c & 4c & 2b & \dots & 4c & 4c \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 4c & 4c & 4c & \dots & 4c & 2b \end{pmatrix}.$$

has eigenvalues

$$\lambda_1 = \lambda_2 = \dots = \lambda_{n-1} = 2b - 4c, \quad \lambda_n = 2b + 4c(n-1).$$

The completely polymorphic equilibrium exists and is stable if the following conditions are satisfied:

$$a > 0, \quad b < 0, \quad -\frac{|b|}{2} < c < \frac{2|b| - a}{4(n-1)}.$$

Model 1. Quadratic stabilizing selection on an additive trait. Trait value:

$$z = \sum_i a(l_i + l'_i).$$

Fitness function:

$$w = w_{stab} = 1 - sz^2.$$

Model 2. “Corridor model”. Let there be a set of n diallelic genes with pleiotropic effects on two quantitative characters. (“Pleiotropy” means that the same gene affects multiple traits). One trait is additive ($z_1 = \sum_i a_1(l_i + l'_i)$) and is under quadratic stabilizing selection ($w_{stab} = 1 - sz^2$). There is some dominance in the second trait,

$$z_2 = \sum [a_2(l_i + l'_i) + 2b_2l_i l'_i]$$

which is under linear directional selection:

$$w_{dir} = 1 + tz_2.$$

Assuming that both forms of selection are weak ($s, t \ll 1$), the overall fitness is

$$w = w_{stab} w_{dir} \approx 1 - sz_1^2 + tz_2.$$

Model 3. Pleiotropic overdominance. The number of heterozygous loci, h , can be represented as

$$h = \sum_i (l_i + l'_i - 2l_i l'_i)$$

(why?). Assume that an additive trait ($z = \sum_i a(l_i + l'_i)$) is under quadratic stabilizing selection ($w_{stab} = 1 - sz^2$) and, in addition, each heterozygous locus increases fitness by value t . The overall fitness is

$$w = w_{stab} + t h.$$

Homework. Can genetic variation be maintained in these three models? If yes, under what conditions?