Multiple loci at linkage equilibrium: deterministic case

One locus, two alleles

In the very simplest case, assume haploids with two alleles (P, Q), frequencies $p + q = 1$, and fitnesses W_P, W_Q. Then:

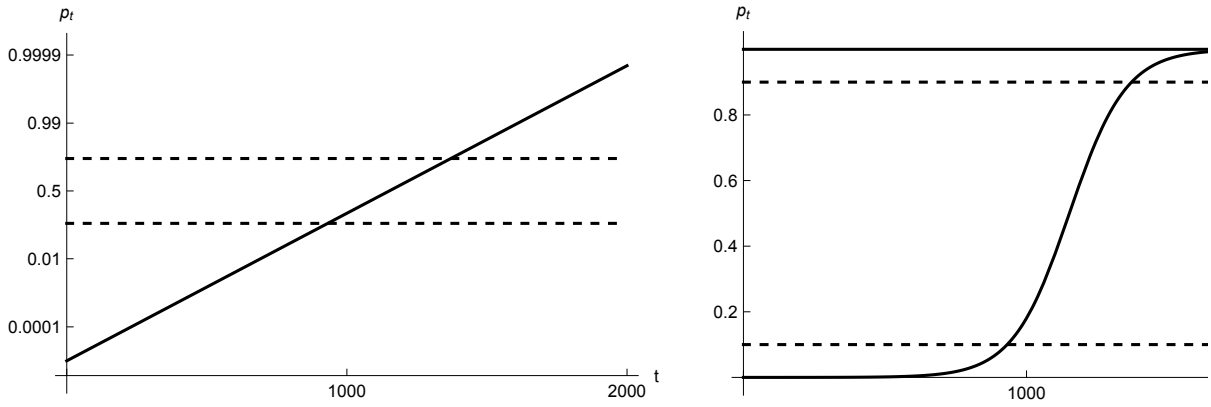
$$p_t = p_{t-1}\frac{W_P}{\bar{W}} \quad \text{where } \bar{W} = qW_Q + pW_P$$

It is often simpler to follow the ratio of the two types, which changes geometrically if relative fitnesses are constant:

$$\frac{p_t}{q_t} = \frac{p_{t-1}}{q_{t-1}}\frac{W_P}{W_Q} = \frac{p_0}{q_0}\left(\frac{W_P}{W_Q}\right)^t$$

Plotting p_t/q_t on a log scale (a *logit* transform) gives a straight line, with slope equal to the selection coefficient. The example below shows selection $s = \log\left(\frac{W_P}{W_Q}\right) = 0.01$, $p_0 = 10^{-5}$. The allele takes $\sim\log(1/p_0)/s$ generations to rise to appreciable frequency, but spends a much shorter time, $\sim 1/s$, polymorphic. Dashed lines indicate $p = 0.1, 0.9$.

Out[14]=



Selection gradients

The change in allele frequency is proportional to the gradient of log mean fitness with respect to allele frequency:

$$\Delta p = pq\frac{\partial \log(\bar{W})}{\partial p} \quad \text{where } \bar{W} = qW_Q + pW_P,\ p + q = 1$$

Provided that the population is at linkage equilibrium, and fitnesses are constant, this generalises to multiple loci:

$$\Delta p_i = p_i q_i \frac{\partial \log(\bar{W})}{\partial p_i} \quad \text{where } \bar{W} = (q_1\, q_2 \ldots)W_{00\ldots} + \ldots + (p_1\, p_2 \ldots)W_{11\ldots} \quad \text{(haploids)}$$

$$\Delta p_i = \frac{p_i q_i}{2}\frac{\partial \log(\bar{W})}{\partial p_i} \quad \text{where } \bar{W} = \left(q_1^2\, q_2^2 \ldots\right)W_{00\ldots} + \ldots + \left(p_1^2\, p_2^2 \ldots\right)W_{22\ldots} \quad \text{(diploids)}$$

where diploid genotypes {QQ, PQ, PP} are labelled 0, 1, 2.

WIth haploids and constant fitnesses, there may be multiple equilibria, but these are always fixed ($p_i = 0$ or 1). With diploids, selection favouring heterozygotes can maintain stable polymorphisms with $0 < p_i < 1$.

Chastain et al. (2014) show that selection on allele frequencies is equivalent to the *multiplicative weights update algorithm* (MWUA) which gives a remarkably efficient solution to the "experts problem".

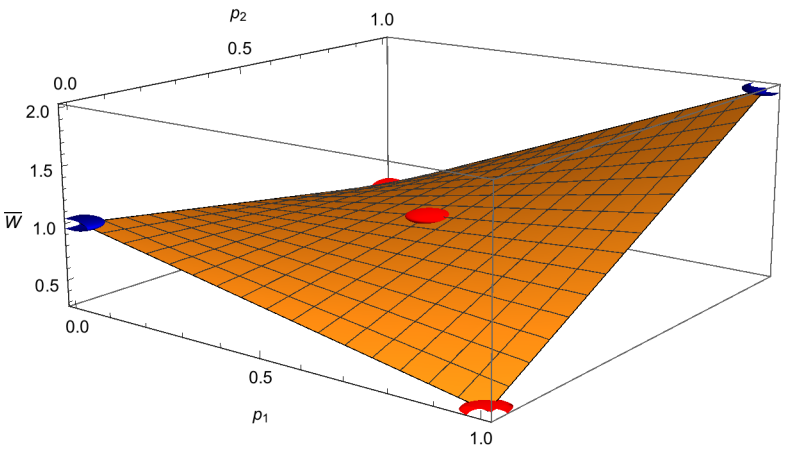
Adaptive landscapes

This can be visualised by plotting mean fitness against allele frequencies. For a haploid population with two loci, and fitnesses:

$$\begin{pmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.3 & 1 \end{pmatrix} \qquad \overline{W} = q_1\, q_2\, W_{00} + q_1\, p_2\, W_{01} + p_1\, q_2\, W_{10} + p_1\, p_2\, W_{11}$$

The figure shows the adaptive landscape" - a graph of mean fitness $\overline{W}$ against p_1, p_2. Red dots are unstable equilibria, blue dots are locally stable.

Out[36]=



Stabilising selection on a quantitative trait

Define $z = \sum_i \gamma_i X_i$; $X_i = 0$ or 1 so $p_i = \mathbb{E}[X_i]$. Assume haploidy and linkage equilibrium. Then, $\overline{z} = \sum_i \gamma_i p_i$, $V_g = \mathrm{var}(z) = \sum_i \gamma_i^2\, p_i\, q_i$.

Fitness is $W = 1 - \frac{1}{2\,V_s}(z - z_{\mathrm{opt}})^2$, which we assume is close to 1. Then, $\overline{W} = 1 - \frac{1}{2\,V_s}\left((\overline{z} - z_{\mathrm{opt}})^2 + V_g\right)$, and $\log(\overline{W}) \sim -\frac{1}{2\,V_s}\left((\overline{z} - z_{\mathrm{opt}})^2 + V_g\right)$. We assume symmetrical mutation at rate μ, so that the change in allele frequency is $\mu\,(1 - 2\,p_i)$. The change in allele frequency is small and can be approximated as a continuous rate of change:

$$\frac{dp_i}{dt} = \mu\,(1 - 2\,p_i) + p_i\,q_i\left(\frac{\partial \log(\overline{W})}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial p_i} + \frac{\partial \log(\overline{W})}{\partial V_g}\frac{\partial V_g}{\partial p_i}\right)$$

$$= \mu\,(1 - 2\,p_i) + \frac{p_i\,q_i}{2\,V_s}\left(-2\,(z - z_{\mathrm{opt}})\,\gamma_i - (1 - 2\,p_i)\,\gamma_i^2\right)$$

We are interested mainly in the rate of change of the mean and the variance. The mean changes as:

$$\frac{d\,\overline{z}}{dt} = \sum_i \gamma_i \frac{dp_i}{dt} = \sum_i \mu\,\gamma_i\,(1 - 2\,p_i) + \frac{1}{2\,V_s}\sum_i p_i\,q_i\left(-2\,(z - z_{\mathrm{opt}})\,\gamma_i^2 - (1 - 2\,p_i)\,\gamma_i^3\right)$$

$$= \mu\,(z_{\max} - 2\,\overline{z}) - \frac{V_g}{V_s}(z - z_{\mathrm{opt}}) - \frac{M_3}{2\,V_s}$$

where $z_{\max} = \sum_i \gamma_i$ is the maximum possible z, and $M_3 = \sum_i p_i\,q_i(1 - 2\,p_i)\,\gamma_i^3$ is the third moment of z. The first term is likely to be small, since μ is the per-locus mutation rate. The last term is also likely to be small, both because $(1 - 2\,p_i)$ fluctuates in sign, and because it is smaller in magnitude than $\sum_i p_i\,q_i\,\gamma_i^3 = |\gamma|\,V_g$, and the allelic effects γ are assumed small. Therefore,

$\frac{d\bar{z}}{dt} \approx -\frac{V_g}{V_s}(z - z_{opt})$, which does not depend on the genetic details.

The rate of change of variance is:

$$\frac{dV_g}{dt} = \sum_i \gamma_i^2 (1 - 2p_i)\frac{dp_i}{dt}$$

$$= \mu \sum_i \gamma_i^2 (1 - 2p_i)^2 + \frac{1}{2V_s}\sum_i p_i q_i (1 - 2p_i)\left(-2(z - z_{opt})\gamma_i^3 - (1 - 2p_i)\gamma_i^4\right)$$

$$= \mu \sum_i \gamma_i^2 (1 - 4p_i q_i) + \frac{1}{2V_s}\sum_i p_i q_i \left(-2(1 - 2p_i)(z - z_{opt})\gamma_i^3 - (1 - 4p_i q_i)\gamma_i^4\right)$$

$$= (V_m - 4\mu V_g) - \frac{(z - z_{opt})}{V_s} M_3 - \frac{1}{2V_s}\sum_i p_i q_i (1 - 4p_i q_i)\gamma_i^4$$

where $V_m = \mu \sum_i \gamma_i^2$ is the rate of increase of variance in a completely inbred population; necessarily, $V_m > 4\mu V_g$. (Note that $(1 - 2p_i)^2 = (1 - 4p_i q_i)$). The first term is no larger than V_m; the second term is $\sim|\gamma|\frac{V_g}{V_s}(z - z_{opt})$ (as argued above); and the last term is smaller than $\sim -\frac{V_g|\gamma^2|}{2V_s}$. Thus, the variance will change much more slowly than the mean if $|\gamma^2|$ is small, justifying the *infinitesimal model*.

We can easily find the equilibrium allele frequencies if the mean has reached the optimum: $p_i q_i = 2\mu V_s / \gamma_i^2$ (assuming that γ_i^2 is large enough that this is smaller than the maximum possible $pq = 1/4$). The genetic variance is $V_g = \sum_i \gamma_i^2 p_i q_i = 2UV_s$, where $U = \sum_i \mu$ is the genomic mutation rate. The loss of mean fitness due to genetic variation (the *mutation load*) is $\frac{V_g}{2V_s} = U$ - as predicted by Haldane's Principle.

Multiple loci at linkage equilibrium: random genetic drift

Populations near fixation

If $4N\mu << 1$, populations will typically be near fixation. Then, evolution consists of jumps from one genotype to another, when mutations are occasionally fixed. The probability of fixation of an allele with advantage s, initially at p_0, is:

$$P = \frac{1 - \exp(-2Nsp_0)}{1 - \exp(-2Ns)}$$

foe haploids; for diploids, the 2 are replaced by 4. If the mutation is initially in one copy, $p_0 = 1/N$ in a haploid population, so:

$$P = \frac{2s}{1 - \exp(-2Ns)}$$

For diploids, we have $P = 2s/(1 - \exp(-4Ns))$. Note that for $s < 0$:

$$P = \frac{2|s|}{\exp(2N|s|) - 1}$$

Note that the ratio of fixation probabilities in the two directions is just e^{2Ns}, for haploids. The stationary state is remarkably simple: the probability that the population will fix a genotype with fitness $W(X)$ is just proportional to W^{2N}.

Wright's formula

This is a special case of a much more general formula for the stationary distribution of allele frequencies, which applies for high as well as low mutation rates. The mutation rate from Q_i to P_i is μ_i, and is v_i in the opposite direction. Then, the probability density of the allele frequencies is:

$$\left(\prod_i p_i^{2\,N\mu_i-1}\, q_i^{2\,N\nu_i-1}\right) \bar{w}^{2\,N}$$

for haploids; for diploids, the $2N\mu$, $2N\nu$ are replaced by $4N\mu$, $4N\nu$, but the exponent in $\bar{w}^{2N}$ remains the same.

Multiple loci with linkage disequilibrium

A general method

(See Barton and Turelli, *Genetics* 1991)

Assume selection on haploids. Let $\zeta_i = X_i - p_i$, and define $D_U = \mathbb{E}[\zeta_U]$, where ζ_U is shorthand for $\prod_{i\in U}\zeta_i$. Necessarily, $D_i = 0$; D_{ij} is the standard pairwise LD. Now, define fitness as:

$$\frac{w}{\bar{w}} = 1 + \sum_U a_U\,(\zeta_U - D_U)$$

where the sum is over all sets U that affect fitness. Then:

$$D_S^* = D_S + \sum_U a_U\,(D_{SU} - D_S\,D_U)$$

Note that $D_i = 0$, and $D_i^* = \Delta p_i$. This gives a simple formula for the effects of selection on associations amongst arbitrary sets of loci. The effect of recombination can be written as:

$$D_S^{**} = D_S^* + \sum_{AB=S} r_{A,B}\,D_A^*\,D_B^*$$

Here, D_A^* are associations amongst the set of genes A from male gametes, after selection, and D_B^* amongst the set of genes B from female gametes, assumed here to come together at random.

The equations for selection are not closed: the change in D_S depends on higher order associations D_U. However, with two alelles per locus, we can use the following relation to close the equations:

$$\zeta_i^2 = p_i\,q_i + (1-2\,p_i)\,\zeta_i$$
$$D_{iiU} = \mathbb{E}\left[\zeta_i^2\,\zeta_U\right] = p_i\,q_i\,D_U + (1-2\,p_i)\,D_{iU}$$

with the convention $D_\emptyset = 1$.

These associations are defined with respect to the initial allele frequencies. A final step should be to chenge the *reference point* so that the *D's* are defined relative to the new allele frequencies. However, in he example below such corrections are second order, and can be ignored for weak selection.

Hitch-hiking: derivation using the general method

See Maynard Smith and Haigh (1974).

Suppose that a favourable mutation, with advantage s arises in a haploid population, at initial frequency $p_{1,0} = \frac{1}{N}$; it is linked to a neutral locus at a second locus, with recombination rate c. Allele frequencies are denoted p_1, p_2, and linkage disequilibrium $D_{1,2}$. In the notation just defined, $a_1 = s$ is the only selection coefficient. Selection has effect:

$$\begin{aligned}\Delta p_1 &= D_1^* = D_1 + a_1\,D_{1,1} = s p_1\,q_1\\ \Delta p_2 &= D_2^* = D_2 + a_1\,D_{1,2} = s D_{1,2}\\ D_{1,2}^* &= D_{1,2} + a_1\,D_{1,1,2} = D_{1,2} + s\,(1-2\,p_1)\,D_{1,2}\end{aligned} \tag{1}$$

Using the general method (above), recombination does not change allele frequencies, and changes $D_{1,2}^*$ to a sum over the four possible origins of the two alleles:

$$D_{1,2}^{**} = \frac{(1-c)}{2}\,(D_{1,2}^*\,D_\emptyset^* + D_\emptyset^*\,D_{1,2}^*) + \frac{c}{2}\,(D_1^*\,D_2^* + D_2^*\,D_1^*) = (1-c)\,D_{1,2}^* + c\,D_1^*\,D_2^* \tag{2}$$

using the fact that the *D's* are the same in the two sexes, and that $D_\emptyset = 1$. Now, $D_i^* = \Delta p_i$, and so

the second term above can be neglected; the effect of changing allele frequency on the definition of D can also be neglected (see above)

Hitch-hiking: the dynamics

Approximating to continuous time:

$$\frac{dp_1}{dt} = s p_1 q_1 \qquad \frac{dp_2}{dt} = s D_{1,2} \qquad \frac{dD_{1,2}}{dt} = -c D_{1,2} + s (1 - 2 p_1) D_{1,2} \tag{3}$$

This cna be simplified by letting $\delta = D_{1,2} / (p_1 q_1)$, which is just the diffewrence in frequency of the neutral allele between the two genetic backgrounds. Selection does not affect this, and so it decays by recombination alone:

$$\frac{d\delta}{dt} = -c\delta \quad \therefore \quad \delta = \delta_0 e^{-ct} \tag{4}$$

Initially, $\delta_0 = 1 - p_{2,0} = +q_{2,0}$ with probability $p_{2,0}$, and $-p_{2,0}$ with probability $q_{2,0}$. The total change of the neutral allele can be written:

$$\Delta p_2 = \int_0^\infty \frac{dp_2}{dt} dt = \int_0^\infty s D_{1,2}\, dt = \int_0^\infty s p_1 q_1 \delta_0 e^{-ct}\, dt = \int_0^\infty \frac{dp_1}{dt} \delta_0 e^{-ct}\, dt \tag{5}$$

Now, $\frac{dp_1}{dt}$ peaks as p_1 sweeps from low to high frequency, around time $T = \frac{1}{s} \log(1/p_{1,0})$. To a good approximation, we can take e^{-ct} outside the integral to give:

$$\Delta p_2 \approx \delta_0 e^{-cT} \int_0^\infty \frac{dp_1}{dt} dt \approx \delta_0 e^{-cT} = \delta_0 p_{1,0}^{c/s} \tag{6}$$

With probability $p_{2,0}$ the mutation arises in coupling with P_2, and so $\delta_0 = q_{2,0}$, and p_2 increases by $q_{2,0} p_{1,0}^{c/s}$. With probability $q_{2,0}$, the mutation arises in repulsion, and so $\delta_0 = -p_{2,0}$, and p_2 decreases by $p_{2,0} p_{1,0}^{c/s}$.

The expected change in allele frequency is proportional to $\mathbb{E}[\delta_0] = 0$. The variance of fluctuations is $\mathbb{E}[\delta_0^2] p_0^{2c/s} = p_{2,0} q_{2,0} e^{-2(c/s)\log(N)}$ if the initial frequency of the favourable mutation is $p_{1,0} = 1/N$. This process, in which occasional substitutions cause sudden jumps in allele frequency, has been called *genetic draft* by Gillespie.

The same formula can be derived easily by a coalescent argument, as the chance that neither of two lineages will escape by recombination in time $T = (1/s) \log(1/p_{1,0})$, e^{-2cT}.