2. Elements of population genetics

Reinhard Bürger

Institut für Mathematik

universität wien

2.1 Introduction

The foundations of modern population genetics

were laid by the work of

Ronald A. Fisher, J.B.S. Haldane, and Sewall Wright,

who reconciled Mendelism with Darwinism during the second and third decades of the twentieth century.

They demonstrated that the theory of evolution by natural selection, proposed by Charles Darwin (1859), can be justified on the basis of genetics as governed by Mendel's laws.

Mendel's (1866) prime achievement

was the recognition of the particulate nature of the hereditary determinants, now called genes.

A gene may have different forms, called alleles. In diploid organisms each autosomal cell carries two copies of each gene. The allelic types may be different or the same.

Mendel recognized that in diploid organisms each reproductive cell (gamete – egg or sperm) contains only one of the two alleles, and that each gamete is equally likely to contain one or the other.

The separation of paired alleles is called segregation and occurs during meiosis. At mating, two reproductive cells fuse and form a zygote.

The allelic composition is called the genotype, and the set of observable properties derived from the genotype is called the phenotype.

Terminology

In diploid organisms, if there are two alleles $\mathcal{A}_1$ and $\mathcal{A}_2$, then there are three possible genotypes, $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$.

They are called homozygous if the two alleles are identical, otherwise they are heterozygous. In general, the genotypes $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_1$ cannot be distinguished.

When the phenotype of the heterozygote $\mathcal{A}_1\mathcal{A}_2$ is the same as $\mathcal{A}_1\mathcal{A}_1$, allele $\mathcal{A}_1$ is called dominant and $\mathcal{A}_2$ is called recessive.

Today, it is known that the genetic material is deoxyribonucleic acid (DNA), and the code has been deciphered.

In modern language, genes are functional units of DNA. The position of a gene on the DNA is called its locus.

2.2 The Hardy-Weinberg Law

The Hardy-Weinberg Law

We consider a single diploid locus with I alleles $\mathcal{A}_i$ in an (infinitely) large population with discrete, nonoverlapping generations.

The frequency of the ordered genotype $\mathcal{A}_i\mathcal{A}_j$ is P_{ij}. Then the frequency of allele $\mathcal{A}_i$ is

$$p_i = \sum\nolimits_j P_{ij} \, .$$

The Hardy-Weinberg law states that after one generation of random mating, the zygotic frequencies satisfy

$$P'_{ij} = p_i p_j \quad \text{for every } i \text{ and } j \, . \tag{2.1}$$

If gametes combine randomly, this holds by definition. If diploid individuals mate, it is less obvious but can be proved using elementary probability theory.

Important consequences of the HW law

The allele frequencies satisfy

$$p_i' = p_i \quad \text{for every } i\,. \tag{2.2}$$

In other words, in a (sufficiently large) randomly mating population reproduction does not change allele frequencies.

Therefore, no genetic variability is lost by reproduction if mating is random!

A population is said to be in Hardy–Weinberg equilibrium if

$$P_{ij} = p_i p_j\,. \tag{2.3}$$

It follows that gamete frequencies are sufficient to describe such a population!

2.3 The discrete-time selection model

The model

We shall be concerned with the evolutionary consequences of selection caused by differential viabilities (i.e., the probability that an offspring survives to reproductive age).

Throughout, we consider an (infinitely) large population with discrete, nonoverlapping generations. Therefore, we can ignore random genetic drift. Unless stated otherwise, our population is diploid and mates at random. The sexes are indistinguishable.

We suppose that at an autosomal locus the alleles $\mathcal{A}_1, \ldots, \mathcal{A}_I$ occur. We count individuals at the zygote stage and denote the (relative) frequency of the ordered genotype $\mathcal{A}_i\mathcal{A}_j$ by $P_{ij} (= P_{ji})$.

As above, frequency of allele $\mathcal{A}_i$ is $p_i = \sum_j P_{ij}$.

Since mating is at random, the genotype frequencies P_{ij} are in Hardy-Weinberg proportions, i.e., $P_{ij} = p_i p_j$.

The model

We denote the fitness (viability) of $\mathcal{A}_i\mathcal{A}_j$ individuals by $w_{ij} \geq 0$. We assume that the fitnesses satisfy $w_{ij} = w_{ji}$.

Then the frequency of $\mathcal{A}_i\mathcal{A}_j$ genotypes among adults that have survived selection is

$$P^*_{ij} = \frac{w_{ij}P_{ij}}{\bar{w}} = \frac{w_{ij}p_ip_j}{\bar{w}}\,, \tag{2.4}$$

where

$$\bar{w} = \sum_{ij} w_{ij}P_{ij} = \sum_{ij} w_{ij}p_ip_j = \sum_i w_ip_i \tag{2.5}$$

is the mean fitness of the population and

$$w_i = \sum_j w_{ij}p_j \tag{2.6}$$

is the marginal fitness of allele $\mathcal{A}_i$.

The selection equation

Therefore, the frequency of $\mathcal{A}_i$ after selection is

$$p_i^* = \sum_j P_{ij}^* = p_i \frac{w_i}{\bar{w}} \,. \tag{2.7}$$

Because of random mating, the allele frequency p_i' among zygotes of the next generation is also p_i^*. Therefore, the allele frequencies evolve according to the selection equation

$$p_i' = p_i \frac{w_i}{\bar{w}} \,, \quad \text{for every } i \,. \tag{2.8}$$

The selection equation

This recursion equation preserves the relation

$$\sum_i p_i = 1$$

and describes the evolution of allele frequencies at a single autosomal locus in a diploid population.

We view the selection equation (2.8) as a dynamical system on the simplex

$$\Delta_I = \left\{ p = (p_1, \ldots, p_I)^T \in \mathbb{R}^I : p_i \geq 0 \; \forall i \, , \sum_i p_i = 1 \right\} .$$

Multiplicative fitnesses, or the haploid case

We say fitnesses are multiplicative if there exist constants v_i such that

$$w_{ij} = v_i v_j \tag{2.9}$$

holds for every i, j. Then $w_i = v_i \bar{v}$, where $\bar{v} = \sum_i v_i p_i$, and $\bar{w} = \bar{v}^2$. Therefore, the selection dynamics simplifies to

$$p_i' = p_i \frac{v_i}{\bar{v}}\,, \quad \text{for every } i\,, \tag{2.10}$$

which also describes the dynamics of an asexual haploid population under selection.

(2.10) has the explicit solution (prove this!)

$$p_i(t) = \frac{p_i(0) v_i^t}{\sum_j p_j(0) v_j^t}\,. \tag{2.11}$$

In haploids, the best allele becomes fixed

Assume (2.10). If one allele, say $\mathcal{A}_1$ has higher fitness than all others ($v_1 > v_i$ for every $i \neq 1$), then $(v_j/v_1)^t \to 0$ for $j \neq 1$ as $t \to \infty$.

Therefore, (2.11) shows (prove this!) that $p_1(t) \to 1$ as $t \to \infty$, i.e., in the long run the best allele becomes fixed.

As we shall see, this is not necessarily so in diploids.

Selection is very efficient

Consider two alleles, $\mathcal{A}_1$ and $\mathcal{A}_2$. If $\mathcal{A}_1$ is the wild type and we assume that $\mathcal{A}_2$ is a new beneficial allele, we may set (without loss of generality!) $v_1 = 1$ and $v_2 = 1 + s$. Then we obtain from above:

$$\frac{p_2(t)}{p_1(t)} = \frac{p_2(0)}{p_1(0)} \left(\frac{v_2}{v_1} \right)^t = \frac{p_2(0)}{p_1(0)} (1+s)^t \,. \tag{2.12}$$

Thus, $\mathcal{A}_2$ increases exponentially relative to $\mathcal{A}_1$.

If $s = 0.5$, then after 10 generations the frequency of $\mathcal{A}_2$ has increased by a factor of $(1+s)^t = 1.5^{10} \approx 57.7$ relative to $\mathcal{A}_2$.
If $s = 0.05$ and $t = 100$, this factor is $(1+s)^t = 1.05^{100} \approx 131.5$.

Slight fitness differences may have a big long-term effect!

Mean fitness is increasing

A fundamental property of the selection dynamics (2.8) is that mean fitness is increasing along non-constant trajectories, i.e.,

$$\bar{w}' = \bar{w}(p') \geq \bar{w}(p) = \bar{w}\,, \tag{2.13}$$

and equality holds if and only if p is an equilibrium (proof is not easy!).

This statement is a special case of Fisher's Fundamental Theorem of Natural Selection (see below). It implies that the dynamics is gradient like and trajectories always converge to an equilibrium.

The equilibria are precisely the solutions of

$$p_i(w_i - \bar{w}) = 0 \quad \text{for every } i\,. \tag{2.14}$$

In general, it is difficult to determine the equilibria. There can be up $2^I - 1$ equilibria.

Evolutionary dynamics for two alleles

Now we specialize to two alleles, $\mathcal{A}_1$ and $\mathcal{A}_2$, and write p and $1-p$ instead of p_1 and p_2. We use relative fitnesses and assume

$$w_{11} = 1\,,\ w_{12} = 1 - hs\,,\ w_{22} = 1 - s\,, \tag{2.15}$$

where s is called the selection coefficient and h describes the degree of dominance. We assume $s > 0$.

The allele $\mathcal{A}_1$ is called dominant if $h = 0$, partially dominant if $0 < h < \frac{1}{2}$, recessive if $h = 1$, and partially recessive if $\frac{1}{2} < h < 1$.

If $h = \frac{1}{2}$, there is no dominance or additivity (of alleles).
If $h < 0$, there is overdominance or heterozygote advantage.
If $h > 1$, there is underdominance or heterozygote inferiority.

Evolutionary dynamics for two alleles

We obtain

$$w_1 = 1 - hs + hsp \quad \text{and} \quad w_2 = 1 - s + s(1-h)p\,, \tag{2.16a}$$

and, from (2.5), the mean fitness is

$$\bar{w} = 1 - s + 2s(1-h)p - s(1-2h)p^2\,. \tag{2.16b}$$

It is easily verified that the allele-frequency change from one generation to the next can be written as

$$\Delta p = p' - p = \frac{p(1-p)}{2\bar{w}}\,\frac{\mathrm{d}\bar{w}}{\mathrm{d}p} \tag{2.17a}$$

$$= \frac{p(1-p)s}{\bar{w}}\,[1 - h - (1-2h)p]\,. \tag{2.17b}$$

Evolutionary dynamics for two alleles

Therefore, there exists a polymorphic equilibrium, i.e., both alleles are present, if and only if $h < 0$ or $h > 1$. It is given by

$$\hat{p} = \frac{1-h}{1-2h}\,. \tag{2.18}$$

This equilibrium is globally asymptotically stable if there is overdominance. It is unstable if there is underdominance.

In the latter case, the evolutionary outcome depends on the initial condition, and the allele with higher (homozygous) fitness may be lost.

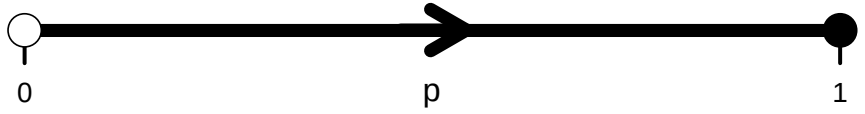
For intermediate dominance ($0 \leq h \leq 1$), $\hat{p} = 1$ is globally asymptotically stable, i.e., the fitter allele is established by selection.

Convergence patterns for selection with two alleles

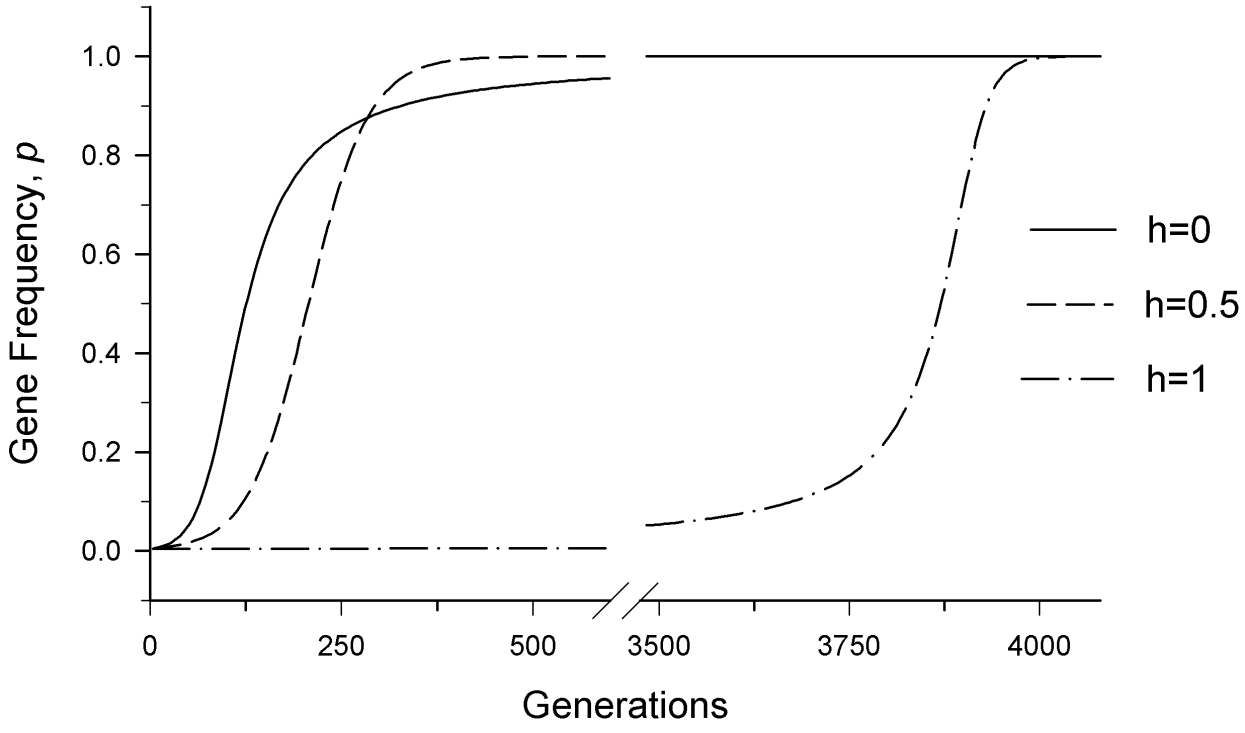
$0 < s < 1,\ h > 1$

$0 < s < 1,\ h < 0$

$0 < s < 1,\ 0 \le h \le 1$



Dominance and the rate of adaptation



2.4 The continuous-time selection model

Derivation from the discrete-time model

Most higher animal species have overlapping generations because birth and death occur continuously in time. Because a rigorous derivation from biological principles (invoking age structure etc.) is a formidable task, we view the continuous-time model as an approximation to the discrete dynamics if selection is weak.

To this end, we set

$$w_{ij} = 1 + sm_{ij} \quad \text{for every } i, j\,, \tag{2.19}$$

where $s > 0$ is assumed to be small. We rescale time according to $t = \lfloor \tau/s \rfloor$, where $\lfloor\ \rfloor$ denotes the closest smaller integer. Then s may be interpreted as generation length and we write $\pi_i(\tau) = p_i(t)$, where for $p_i(t)$ satisfies the difference equation (2.8).

Derivation

Then we obtain formally

$$\begin{aligned}
\frac{\mathrm{d}\pi_i}{\mathrm{d}\tau}(\tau) &= \lim_{s\downarrow 0} \frac{1}{s}\left[\pi_i(\tau+s) - \pi_i(\tau)\right] \\
&= \lim_{s\downarrow 0} \frac{1}{s}\left[p_i(t+1) - p_i(t)\right] \\
&= \lim_{s\downarrow 0} \frac{1}{s} \frac{s p_i(t)(m_i(t) - \overline{m}(t))}{1 + s\overline{m}(t)} \qquad \text{by (2.8) and (2.19)} \\
&= \pi_i(\tau)(m_i(\tau) - \overline{m}(\tau)\,.
\end{aligned}$$

where

$$m_i = \sum_j m_{ij} p_j \quad \text{and} \quad \overline{m} = \sum_i m_i p_i = \sum_{i,j} m_{ij} p_i p_j\,, \tag{2.20}$$

are the marginal (Malthusian) fitness of allele A_i and the mean (Malthusian) fitness of the population, respectively.

The continuous-time selection equation

Therefore, the dynamics of allele frequencies becomes

$$\dot{p}_i = p_i(m_i - \bar{m}) \quad \text{for every } i\,, \tag{2.21}$$

and

$$\Delta p_i \approx s\dot{\pi}_i = sp_i(m_i - \bar{m})\,. \tag{2.22}$$

The continuous-time selection equation

Therefore, the dynamics of allele frequencies becomes

$$\dot{p}_i = p_i(m_i - \bar{m}) \quad \text{for every } i\,, \tag{2.21}$$

and

$$\Delta p_i \approx s\dot{\pi}_i = sp_i(m_i - \bar{m})\,. \tag{2.22}$$

Note that (2.8) is essentially the Euler scheme for (2.21).

The above defined fitness parameters m_{ij} can be interpreted as the (Malthusian) fitness m_{ij} of genotype $\mathcal{A}_i\mathcal{A}_j$, i.e., as its birth rate minus its death rate.

Selection in continuous time

The equilibria of (2.21) are obtained by solving the non-linear system

$$p_i(m_i - \overline{m}) = 0 \text{ for every } i\,. \tag{2.23}$$

Therefore, the discrete-time and the continuous-time selection equation have the same equilibria. The equilibria have also the same stability properties.

Selection in continuous time

The equilibria of (2.21) are obtained by solving the non-linear system

$$p_i(m_i - \overline{m}) = 0 \text{ for every } i\,. \tag{2.23}$$

Therefore, the discrete-time and the continuous-time selection equation have the same equilibria. The equilibria have also the same stability properties.

Exercise 1. Validate the above statement about stability by deriving the Jacobian in each case.
Exercise 2. By constructing an example, prove that the selection equation with I alleles can have up to $2^I - 1$ equilibria (provided equilibria are isolated).

Two alleles

For two alleles, (2.21) simplifies considerably because it is sufficient to track the allele frequency $p = p_1$. In addition, we write $q = 1 - p$.

Scaling the Malthusian parameters in the following way

$$\begin{array}{ccc} \mathcal{A}_1\mathcal{A}_1 & \mathcal{A}_1\mathcal{A}_2 & \mathcal{A}_2\mathcal{A}_2 \\ 0 & -hs & -s \end{array}, \tag{2.24}$$

we obtain the simple representations

$$\dot{p} = \tfrac{1}{2}spq \quad \text{if } h = \tfrac{1}{2} \text{ (no dominance)} \tag{2.25}$$

and

$$\dot{p} = spq^2 \quad \text{if } h = 0 \ (\mathcal{A}_1 \text{ is dominant}) . \tag{2.26}$$

Equation (2.25) is also obtained for a haploid population in which $\mathcal{A}_2$ has a selective disadvantage of $\frac{1}{2}s$ relative to $\mathcal{A}_1$.

The Fundamental Theorem of Natural Selection

In his Fundamental Theorem of Natural Selection, Fisher (1930) stated that the rate of change of mean fitness is equal to the additive genetic variance in fitness,

$$\sigma_A^2 = 2\sum_i p_i(w_i - \bar{w})^2 \,. \tag{2.27}$$

In general, $\sigma_A^2 < \sigma_G^2 = \sum_{i,j} p_i p_j (w_{ij} - \bar{w})^2$, where σ_G^2 is the total genetic variance.

The classical interpretation of Fisher's Fundamental Theorem has been that

$$\Delta\bar{w} = \sigma_A^2/\bar{w} \,, \tag{2.28}$$

at least approximately. Unless there is no dominance, (2.28) does generally not hold exactly.

Mean fitness increase in continuous time

In continuous time, we have $\sigma_A^2 = 2\sum_i p_i(m_i - \overline{m})^2$ and obtain easily

$$\begin{aligned}
\dot{\overline{m}} &= \frac{d}{dt}\sum_{i,j} m_{ij}p_ip_j \\
&= 2\sum_i m_i\dot{p}_i \\
&= 2\sum_i (m_i - \overline{m})\dot{p}_i \quad \text{(because } \overline{m}\sum_i \dot{p}_i = 0\text{)} \\
&= 2\sum_i p_i(m_i - \overline{m})^2 = \sigma_A^2 \geq 0\,.
\end{aligned}$$

Therefore, mean fitness is a Lyapunov function! It strictly increases along solutions except at equilibria (when it remains constant).