NATURAL SELECTION AND THE FUNDAMENTAL THEOREM

Dr. Bob Gardner

Department of Mathematical and Statistical Sciences

(DoMaSS)

East Tennessee State University

presented in

Independent Study - Mathematical Biology

Summer 2003

Genotypes

Note. "Higher" organisms ("eukaryotes") have cells with organelles which have a nucleus that contains the genetic material (DNA). The DNA is in chromosomes. A cell with a single copy of each chromosome is *haploid* and a cell with two copies of each chromosome is *diploid*. We will concentrate on on diploid organisms and consider different versions ("genotypes") of the "homologous" pairs of chromosomes.

Note. A location on a chromosome is called a *locus*. At each locus is an *allele* (or gene). We denote the alleles as $A_1, A_2, \ldots, A_n$. If, at a given locus, one chromosome has allele A_i and the other chromosome has allele A_j where $i \neq j$, then the organism is a *heterozygote* at that locus. If $i = j$ then the organism is a *homozygote* at that locus. The *genotype* of the organism is the pair of alleles $A_i A_j$ present at the locus. The way these alleles are expressed determine the *phenotype* of the organism.

Note. In some cases, one allele is *dominant* and another is *recessive*. If A is the dominant allele and a is the recessive allele, then there are three genotypes and two phenotypes:

genotype	phenotype
AA	*A–*
Aa	*A–*
aa	*aa*

The Hardy-Weinberg Law

Note. We denote the *frequency* of allele A_i as x_i where $i = 1, 2, \ldots, n$. For the gene pair (A_i, A_j), where allele A_i is inherited from the father and A_j is inherited from the mother, we denote the frequency as x_{ij}.

Theorem. The Hardy-Weinberg Law

If alleles are randomly mixed from one generation to the next, then:

(a) the gene frequencies remain unchanged from generation to generation, and

(b) from the first daughter generation onward, the frequency of genotype (A_i, A_i) is x_i^2 and the frequency of genotypes (A_i, A_j) and (A_j, A_i) <u>together</u> is $2x_i x_j$ where $i \neq j$.

Proof. If we choose an allele at random from all the genotypes $\{(A_i, A_j) \mid 1 \leq i \leq n, 1 \leq j \leq n\}$, then we get allele A_i with probability

$$
\begin{aligned}
x_i &= \frac{1}{2} \times (\text{the frequency of genotype } (A_i, A_j)) \\
&\quad + \frac{1}{2} \times (\text{the frequency of genotype } (A_j, A_i)) \\
&= \frac{1}{2} \sum_{j=1}^{n} x_{ij} + \frac{1}{2} \sum_{j=1}^{n} x_{ji}.
\end{aligned}
$$

Now let x_i' and x_{ij}' denote the frequencies in the next generation.

With random mating,

$$x'_{ij} = x_i x_j \text{ and } x'_{ji} = x_j x_i.$$

Therefore

$$
\begin{aligned}
x'_i &= \frac{1}{2} \sum_{j=1}^{n} x'_{ij} + \frac{1}{2} \sum_{j=1}^{n} x'_{ji} \\
&= \frac{1}{2} \left(\sum_{j=1}^{n} x_i x_j \right) + \frac{1}{2} \left(\sum_{j=1}^{n} x_j x_i \right) \\
&= \sum_{j=1}^{n} x_i x_j = x_i \left(\sum_{j=1}^{n} x_j \right) = x_i.
\end{aligned}
$$

Next, the frequency of (A_i, A_j) in the next generation is $x'_{ij} = x_i x_j = x'_i x'_j$ and claim **(b)** holds. ∎

Note. We get the Hardy-Weinberg Law by assuming random mating (or the random union of gametes). In biology, the assumptions are:

1. There is no mutation at the given locus.

2. There is no migration (emigration or immigration).

3. The population is large (infinite, in fact).

4. There is no selection pressure at the locus.

5. The locus has autosomal alleles (i.e. the locus is not sex linked).

6. Reproduction is at random with respect to the locus.

7. Generations are non-overlapping.

Note. If we only consider two alleles A and a with frequencies p and q, respectively, then we have $p + q = 1$ and:

genotype	frequency
AA	p^2
Aa	$2pq = 2p(1-p)$
aa	$q^2 = (1-p)^2$

We can therefore represent the state of such a population as a point $p \in [0, 1]$.

Note. In a sense, the Hardy-Weinberg Law describes a sampling problem. Imagine an urn containing an infinite number of marbles, some of them red ($p \times 100\%$) and some of them blue ($q \times 100\%$). Create a "new population" by taking an infinite sample. As we will see, if the population is finite, then the sampling process may violate the Hardy-Weinberg Law and the population may undergo *genetic drift*.

Note. We can represent the "state" of a population which has three possible alleles at a given locus using a *DeFinetti Diagram*. In such a diagram, a point is plotted in an equilateral triangle and the perpendicular distance from the point to each edge of the triangle represents an allele frequency:
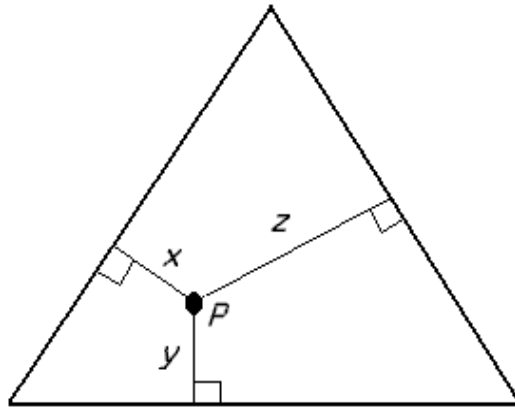
We can also use a DeFinetti diagram to represent the genotypic frequencies AA, Aa, and aa.

The Selection Model

Note. We now associate a *selective value* (or *fitness*) with each genotype A_iA_j of w_{ij}. We take $w_{ij} \geq 0$ for all i, j. Since the gene pairs (A_i, A_j) and (A_j, A_i) are genotypically identical, then we have $w_{ij} = w_{ji}$.

Note. If N individuals are conceived in a new generation, then the number of individuals with gene pair (A_i, A_j) is $x_i x_j N$ and the number that survive to maturity is $w_{ij} x_i x_j N$ (so here we take $w_{ij} \in [0, 1]$). The total number of individuals reaching maturity in the next generation is then

$$\sum_{r,s=1}^{n} w_{rs} x_r x_s N.$$

Then the frequency of gene pairs (A_i, A_j) in the next generation is

$$x'_{ij} = \frac{w_{ij} x_i x_j N}{\sum_{r,s=1}^{n} w_{rs} x_r x_s N}.$$

Note. Since $x'_i = \dfrac{1}{2} \sum_{j=1}^{n} x'_{ij} + \dfrac{1}{2} \sum_{j=1}^{n} x'_{ji}$ then

$$x'_i = \frac{1}{2} \sum_{j=1}^{n} \left(\frac{w_{ij} x_i x_j N}{\sum_{r,s=1}^{n} w_{rs} x_r x_s N} \right) \times 2$$

(the factor of 2 comes from the fact that $w_{ij} = w_{ji}$), or

$$x'_i = x_i \left(\frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{r,s=1}^{n} w_{rs} x_r x_s} \right).$$

This equation describes the evolution of the gene frequencies under selection.

Theorem. The Fundamental Theorem of Natural Selection

The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time. Symbolically:

$$\sum \alpha \, dp = dt \sum \sum {}'(2pa\alpha) = W \, dt$$

where

- a is the rate of increase of bearers of a particular gene <u>above the average</u>,

- α is the average effect upon an individual of introducing the gene in question,

- p is the gene frequency,

- t is time,

- $W \, dt$ is the change in fitness, and

- $\sum'(2pa\alpha)$ is the contribution of each factor to the genetic variance in fitness.

This was stated by Ronald A. Fisher in his 1930 *The Genetical Theory of Natural Selection.*

The Increase in Average Fitness

Definition. We create the *fitness matrix* $W = [w_{ij}]$. The *average fitness* is then

$$\overline{w} = \sum_{r,s=1}^{n} w_{rs} x_r x_s = \vec{x} \cdot (W\vec{x}).$$

Note. In order to prove our "big result," we need the following:

Theorem. Jensen's Inequality. If f is a strictly convex (i.e. concave up) function defined on some interval I, then

$$f\left(\sum p_i x_i\right) \leq \sum p_i f(x_i)$$

for all $x_1, x_2, \ldots, x_n \in I$ and constants $p_1, p_2, \ldots, p_n$, with equality if and only if all x_i coincide.

Note. The text describes the following as a "consequence of the Fundamental Theorem of Natural Selection." However, the following is often taken as the Fundamental Theorem.

Theorem. For the dynamical system $\vec{x} \to \vec{x}'$ given by

$$x_i' = x_i \frac{\sum_j w_{ij} x_j}{\sum_{r,s} w_{rs} x_r x_s} = x_i \frac{(W\vec{x})_i}{\vec{x} \cdot (W\vec{x})},$$

the average fitness $\overline{w}(\vec{x}) = \vec{x} \cdot (W\vec{x})$ increases along every orbit in the sense that $\overline{w}(\vec{x}') \geq \overline{w}(\vec{x})$ with equality if and only if $\vec{x}$ is a fixed point.

Proof. We assume $\overline{w}(\vec{x}) = \vec{x} \cdot W\vec{x} > 0$, and will show that

$$(\overline{w}(\vec{x}))^2(\overline{w}(\vec{x}')) = (\vec{x} \cdot W\vec{x})^2(\vec{x}' \cdot W\vec{x}') \geq (\vec{x} \cdot W\vec{x})^3 = (\overline{w}(\vec{x}))^3.$$

First,

$$
\begin{aligned}
(\vec{x} \cdot W\vec{x})^2(\vec{x}' \cdot W\vec{x}') &= (\vec{x} \cdot W\vec{x})^2 \left(\sum_{i=1}^{n} x_i'(W\vec{x})_i \right) \\
&= (\vec{x} \cdot W\vec{x})^2 \sum_{i=1}^{n} \left(x_i \frac{(W\vec{x})_i}{\vec{x} \cdot W\vec{x}} \right) \left(\sum_{j=1}^{n} w_{ij}x_j' \right) \\
&= (\vec{x} \cdot W\vec{x})^2 \sum_{i=1}^{n} \left(x_i \frac{(W\vec{x})_i}{\vec{x} \cdot W\vec{x}} \right) \left(\sum_{j=1}^{n} w_{ij}x_j \frac{(W\vec{x})_j}{\vec{x} \cdot W\vec{x}} \right) \\
&= \sum_{i=1}^{n} \left(x_i \left[\sum_{k=1}^{n} w_{ik}x_k \right] \right) \sum_{j=1}^{n} w_{ij}x_j(W\vec{x})_j \\
&= \sum_{i,j,k}^{n} x_i w_{ik}x_k w_{ij}x_j(W\vec{x})_j \equiv s(2). \qquad (*)
\end{aligned}
$$

Now if we swap i and j in $(*)$ we get

$$\sum_{i,j,k}^{n} x_i w_{ij}x_j w_{ik}x_k(W\vec{x})_k \equiv s(1).$$

Now

$$s(1) = s(2) = \frac{s(1) + s(2)}{2} \geq \sqrt{s(1)s(2)}$$

and so

$$s(1) = s(2) = \frac{1}{2} \sum_{i,j,k} x_i w_{ij} x_j w_{ik} x_k [(W\vec{x})_j + (W\vec{x})_k]$$

$$\geq \sum_{i,j,k}^{n} x_i w_{ij} x_j w_{ik} x_k (W\vec{x})_j^{1/2} (W\vec{x})_k^{1/2}$$

$$= \sum_{i=1}^{n} x_i \sum_{j=1}^{n} w_{ij} x_j (W\vec{x})_j^{1/2} \sum_{k} w_{ik} x_k (W\vec{x})_k^{1/2}$$

$$= \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} w_{ij} x_j (W\vec{x})_j^{1/2} \right)^2 .$$

Now Jensen's Inequality applied to $f(x) = x^2$ implies

$$\left(\sum x_i y_i \right)^2 \leq \sum x_i y_i^2$$

and so the last expression is

$$\geq \left(\sum_{i=1}^{n} x_i \sum_{j=1}^{n} w_{ij} x_j (W\vec{x})_j^{1/2} \right)^2$$

$$= \left(\sum_{j=1}^{n} x_j (W\vec{x})_j^{1/2} \sum_{i=1}^{n} x_i w_{ij} \right)^2$$

$$= \left(\sum_{j=1}^{n} x_j (W\vec{x})_j^{1/2} \sum x_i w_{ji} \right)^2 \quad \text{since } w_{ij} = w_{ji}$$

$$= \left(\sum_{j=1}^{n} x_j (W\vec{x})_j^{1/2} (W\vec{x})_j \right)^2$$

$$= \left[\sum_{j=1}^{n} x_j (W\vec{x})_j^{3/2} \right]^2$$

$$\geq \left[\left(\sum_{j=1}^{n} x_j (W\vec{x})_j \right)^{3/2} \right]^2 \quad \text{by Jensen's Inequality applied to } f(x) = x^{3/2}$$

$$= \left(\sum_{j=1}^{n} x_j (W\vec{x})_j \right)^3 . \quad (**)$$

Now by $(*)$

$$(\vec{x} \cdot W\vec{x})^2(x' \cdot W\vec{x}') = s(1) = s(2)$$
$$\geq \left(\sum_{j=1}^{n} x_j(W\vec{x})_j\right)^3 \quad \text{by } (**)$$
$$= (\vec{x} \cdot W\vec{x})^3.$$

Hofbauer and Sigmund (p. 239) state: if $\overline{w}(\vec{x}) = \overline{w}(\vec{x}')$, there must exist a value c such that $(W\vec{x})_j = c$ for all j with $x_j > 0$. This means that $\vec{x}$ is a rest point. ∎

Corollary. Let $T : \vec{x} \to \vec{x}'$ be defined as

$$x_i' = x_i \frac{(W\vec{x})_i}{\vec{x} \cdot W\vec{x}}.$$

Then every orbit $T^k\vec{x}$ converges to a rest point as $k \to \infty$. In fact, the convergence is monotone in the sense that $\overline{w}(\vec{x}') \geq \overline{w}(\vec{x})$. Alternatively, every accumulation point of an orbit is a rest point of T.

The Case of Two Alleles

Note. In this section, we consider a single locus with two alleles, A and a, present. This is similar to a previous analysis, but this time we use the Fundamental Theorem of Natural Selection and consider a discrete dynamical system. We denote the frequency of A as p and the frequency of a as $q = 1 - p$. We associate the fitness values with genotypes:

genotype	fitness
AA	w_{11}
Aa	$w_{12} = w_{21}$
aa	w_{22}

From above ("The Selection Model" section), we see that

$$
\begin{aligned}
p' &= p\frac{pw_{11} + qw_{12}}{p^2 w_{11} + pqw_{12} + pqw_{21} + q^2 w_{22}} \\
&= \frac{p(pw_{11} + w_{12}(1-p))}{p(w_{11}p + w_{12}(1-p)) + (1-p)(w_{21}p + w_{22}(1-p))} \qquad (*) \\
&\equiv \frac{a_1}{a_1 + a_2} \text{ where } a_1 = p(w_{11}p + w_{12}(1-p)) \text{ and} \\
&\qquad\qquad a_2 = (1-p)(w_{12}p + w_{22}(1-p)).
\end{aligned}
$$

The average fitness is

$$
\begin{aligned}
\overline{w}(p) &= p^2 w_{11} + 2p(1-p)w_{12} + (1-p)^2 w_{22} \\
&= p^2 w_{11} + 2pw_{12} - 2p^2 w_{12} + w_{22} - 2pw_{22} + p^2 w_{22} \\
&= p^2[(w_{11} - w_{12}) + (w_{22} - w_{12})] - 2p(w_{22} - w_{12}) + w_{22} \\
&= a_1 + a_2.
\end{aligned}
$$

Now from the recurrence relation $(*)$ we get the *difference equation*:

$$
\begin{aligned}
p' - p &= \frac{p(pw_{11} + w_{12}(1-p))}{\overline{w}(p)} - p \\
&= \frac{p^2 w_{11} + p(1-p)w_{12} - p\overline{w}(p)}{\overline{w}(p)} \\
&= \frac{p^2 w_{11} + p(1-p)w_{12} - p(p^2[w_{11} - 2w_{12} + w_{22}] - 2p(w_{22} - w_{12}) + w_{22})}{\overline{w}(p)} \\
&= \frac{p^2 w_{11} + p(1-p)w_{12} - p^3(w_{11} - 2w_{12} + w_{22}) + 2p^2(w_{22} - w_{12}) - pw_{22}}{\overline{w}(p)}. \quad (**)
\end{aligned}
$$

Now

$$
\frac{d}{dp}[\overline{w}(p)] = 2p[w_{11} - 2w_{12} + w_{22}] - 2(w_{22} - w_{12})
$$

and

$$
\begin{aligned}
p(1-p)\frac{d}{dp}[\overline{w}(p)] &= p(1-p)(2p[w_{11} - 2w_{12} + w_{22}] - 2(w_{22} - w_{12})] \\
&= 2\{p^2(w_{11} - 2w_{12} + w_{22}) - p(w_{22} - w_{12}) \\
&\quad -p^3(w_{11} - 2w_{12} + w_{22}) + p^2(w_{22} - w_{12})\} \\
&= 2\{p^2 w_{11} + p(1-p)w_{12} - p^3(w_{11} - 2w_{12} + w_{22}) \\
&\quad -p^2 w_{12} + p^2 w_{22} - pw_{22} + p^2 w_{22} - p^2 w_{12}\} \\
&= 2\{p^2 w_{11} + p(1-p)w_{12} - p^3(w_{11} - 2w_{12} + w_{22}) \\
&\quad +2p^2(w_{22} - w_{12}) - pw_{22}\}.
\end{aligned}
$$

From $(**)$

$$
p' - p = \frac{p(1-p)}{2\overline{w}(p)}\frac{d}{dp}[\overline{w}(p)].
$$

Therefore, we have a fixed point in the recurrence relation when either $p = 0$, $p = 1$, or $\frac{d}{dp}[\overline{w}(p)] = 0$.

15

Note. We consider cases:

(1) If $w_{12} = \dfrac{1}{2}(w_{11} + w_{22})$ then

$$\overline{w}(p) = -2p(w_{22} - w_{12}) + w_{22}$$

and $\dfrac{d}{dp}[\overline{w}(p)] \neq 0$ (unless $w_{22} = w_{12}$ and hence $w_{11} = w_{22} = w_{12}$ — in this event, $\overline{w}(p)$ is a constant and all points are rest points.

Now

$$\frac{d}{dp}[\overline{w}(p)] = -2(w_{22} - w_{12}) = -2\left(w_{22} - \frac{1}{2}(w_{11} + w_{22})\right)$$

$$= -2\left(-\frac{1}{2}w_{11} + \frac{1}{2}w_{22}\right) = w_{11} - w_{22}$$

and so $p' - p > 0$ if $w_{11} > w_{22}$ and $p \to 1$ in this case (where w_{11} is the largest fitness value — and the fitness of AA).

If $w_{11} < w_{22}$ then $p \to 0$ (where w_{22} is the largest fitness value — and the fitness of aa). So in the first case, the most fit homozygotic form goes to fixation.

(2) If $w_{12} \neq \dfrac{1}{2}(w_{11} + w_{22})$ then $\overline{w}(p)$ is a parabola with critical point

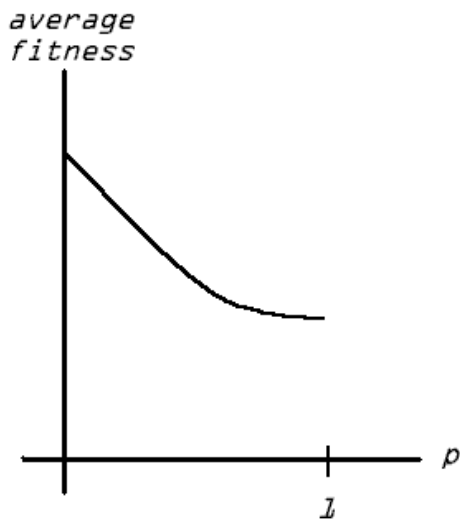$$\overline{p} = \frac{w_{22} - w_{12}}{(w_{11} - w_{12}) + (w_{22} - w_{12})}.$$

In this case we have:

      (a) if w_{12} is between w_{11} and w_{22} then $\overline{w}$ has a MAX at either $p = 0$ (if $w_{22} > w_{12} > w_{11}$) or at $p = 1$ (if $w_{11} > w_{12} > w_{22}$).
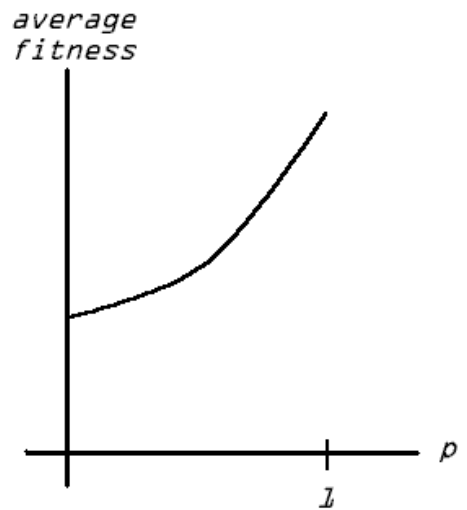
      (b) if $w_{12} > w_{11}$ and $w_{12} > w_{22}$ then this is the *heterozygote advantage* and $\overline{w}(p)$ has a critical point in $(0, 1)$ where $\overline{w}$ has a local MAX.

(c) if $w_{12} < w_{11}$ and $w_{12} < w_{22}$ then $\overline{w}(p)$ has a critical point in $(0,1)$ where $\overline{w}$ has a critical point in $(0,1)$ where $\overline{w}$ has a local MIN.
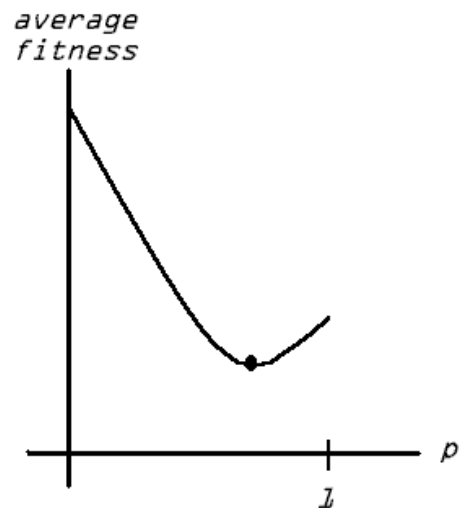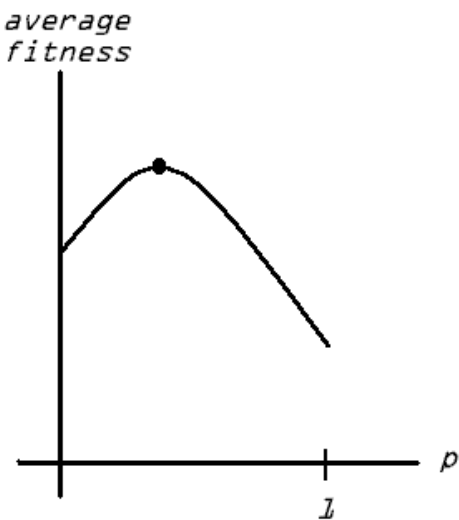
Note. Graphically, we have:

$$w_{22} > w_{12} > w_{11}$$

$$w_{11} > w_{12} > w_{22}$$

Note. Biologically, in this case (one locus, two alleles), polymorphism is maintained in the population only under the heterozygote advantage model.

Note. In the event of one locus and three alleles, we can use a DeFinetti diagram to represent the state of the population and we can watch the population evolve with time as the point moves around. We can also plot the mean fitness $\overline{w}$ as a surface above the DeFinetti diagram, called an *"adaptive topography."*