Mathematical Biology

Tom Davis
tomrdavis@earthlink.net
http://www.geometer.org/mathcircles
November 26, 1999

Part I

Genes and Chromosomes

Biology is extremely complicated, and it is basically impossible to give an exact mathematical description of many things, but it is possible to give mathematical descriptions of various *models* of biological systems that can be investigated. Almost everything that appears below is a simplification, but for the most part it's true.

Most living things pass much of their "design" to the next generation as information encoded by a chemical called DNA, which is a very long strand composed of two complementary chains of simpler building blocks that can be copied by the cell chemistry. The structure of DNA is often called the "double helix". In higher organisms (organisms more complicated than bacteria), the DNA strands are supported by structures called "chromosomes", but there's basically one strand of DNA arranged along each chromosome.

There are various methods used by cells to take advantage of DNA's ability to duplicate information, but here we're going to be interested primarily in organisms that reproduce sexually. These organisms have two copies of each DNA strand, but the copies may be somewhat different. Thus, at a particular spot on a strand is information that codes for eye color, but each individual has two copies, so one copy may code for brown eyes and the other for blue.

Each chunk of DNA that contains the information for one particular feature (like eye color, ability to roll the tongue, the blood-type, and various diseases like Tay-Sachs and sickle-cell anemia) is called a "gene". These genes can be thought of as arranged like a string of beads along a strand of DNA. Depending on the organism, there are different numbers of strands/chromosomes. In humans, for example, there are 23 pairs, or 46 total chromosomes. Quite often, the spot on a chromosome where a gene is found is called a "locus", and the various possibilities for the genes that can be found at that locus are called "alleles".

Generally, each gene encodes information to construct a protein. Different genes may encode different proteins, or none at all. For example (as a rough approximation), the "brown" gene for eye-color encodes the protein that makes the brown pigment in the eye. If there is no pigment present, the eyes are blue. Since each person has a pair of chromosomes, there are two (possibly different) genes, and if either of them makes the brown pigment protein, the eyes will be brown. If neither gene encodes the brown pigment, the eyes will be blue.

So if B represents the gene encoding the brown pigment, and b the gene encoding no pigment, and if we list the possible pairs of genes that can occur on the pair of chromosomes, we obtain: BB, Bb, bB, and bb. Individuals having any of the first three combinations will have brown eyes because they can make the pigment, and only individuals having bb will have blue eyes, since neither chromosome has instructions to make the brown pigment. If an individual has blue eyes, we know that it has type bb; if it has brown eyes, it may have any of the other combinations. The "phenotype" of an individual is what we can see—in this case, the possible phenotypes are "brown-eyed" and "blue-eyed". The "genotype" is the actual combination of genes present, which may be difficult to determine.

During mating, each parent supplies one gene of the two that are available, so the offspring has one copy from its mother and one from its father. The egg or sperm is called a "gamete", and the combination of a particular egg and sperm is called a "zygote". Normally, the gene provided by each parent is selected at random from the two possibilities, with a 50% probability for each. Continuing with the B-b example above, suppose that both parents have genotype Bb—one of each type. Thus half the gametes from each parent are of type B and half of type b. It's not hard to see that there are four equally-likely possibilities for the offspring: BB, Bb, bB, and bb. Thus, on average, 3 of every 4 offspring will have brown eyes.

Part II

The Hardy-Weinberg Law

In the case of blue versus brown eyes, there is no tremendous difference in the ability of the two phenotypes to survive. The first thing we will show mathematically is known as the "Hardy-Weinberg Law[1]": that if there is no difference in fitness among the phenotypes, then the proportions of the genes will not change (at least in a population that is so large that it can be considered essentially infinite).

One simplifying assumption we make is that mating is random. Assume that in the population, the probability of having genotype AA is P, the probability of having genotype Aa is Q, and the probability of having genotype aa is R. Then the following table displays the outcomes of random mating:

	AA: P	Aa: Q	aa: R
AA: P	AA: P^2	AA: $\frac{1}{2}PQ$ Aa: $\frac{1}{2}PQ$	Aa: PR
Aa: Q	AA: $\frac{1}{2}PQ$ Aa: $\frac{1}{2}PQ$	AA: $\frac{1}{4}Q^2$ Aa: $\frac{1}{2}Q^2$ aa: $\frac{1}{4}Q^2$	Aa: $\frac{1}{2}QR$ aa: $\frac{1}{2}QR$
aa: R	Aa: PR	Aa: $\frac{1}{2}QR$ aa: $\frac{1}{2}QR$	aa: R^2

Before the mating, the probability of an A-type allele is $P + Q/2$ and of a a-type allele is $R + Q/2$; each AA parent is certain to produce an A-type allele, each Aa parent has a 50% probability of doing so, and similarly for the a-type alleles.

After mating, we just need to add up the total number of resulting adults of each type. We use P', Q', and R' to indicate the probabilities of types AA, Aa, and aa after mating, respectively:

$$\begin{aligned}
P' &= P^2 + PQ/2 + PQ/2 + Q^2/4 \\
Q' &= PQ/2 + PR + PQ/2 + Q^2/2 + QR/2 + PR + QR/2 \\
R' &= Q^2/4 + QR/2 + QR/2 + R^2
\end{aligned}$$

The probability of an A-type allele after mating is similarly $P' + Q'/2$ and of an a-type allele is $R' + Q'/2$. Let's evaluate $P' + Q'/2$ (the evaluation of $R' + Q'/2$ is exactly similar). Remember that $P + Q + R = 1$:

$$\begin{aligned}
P' + Q'/2 &= P^2 + PQ/2 + PQ/2 + Q^2/4 \\
&\quad +(PQ/2 + PR + PQ/2 + Q^2/2 + QR/2 + PR + QR/2)/2 \\
&= P^2 + 3PQ/2 + PR + QR/2 + Q^2/2 \\
&= (P^2 + PQ + PR) + (PQ + QR + Q^2)/2 \\
&= P(P + Q + R) + Q(P + Q + R)/2 \\
&= P + Q/2.
\end{aligned}$$

[1]This is also known as the "Hardy-Weinberg Equilibrium"

This proves that if there is random mating and no selective advantage to any of the genotypes, there will be no change in allele frequency as a result of mating.

Part III

Fitness

The biological concept of fitness is very easy to describe, but it is difficult for many people to understand, since it seems somewhat unnatural.

The "fitness" of an individual is simply the expected number of offspring it will leave in the next generation.

If you think of fitness as being strong, or fast, or disease-resistant, you will often be right, but not necessarily—you will be right only if those characteristics help the individual to produce more offspring.

For example, suppose one type of organism can live thorough baths of acid or through a fire, and another cannot. The first type of individual leaves, on average, 2 offspring, but the second type leaves 3. If there are no baths of acid or fires, the second individual is far more fit than the first (in a biological sense).

It is often important to talk about the "relative fitness" of different types of organism which is just the ratio of their fitnesses. For example, if there are individuals of genotype AA, Aa, and aa, and their relative fitnesses are 2, 1, and 1.3, respectively, that means that on average individuals of type AA produce twice as many offspring as those of type Aa, and individuals of type aa produce 1.3 times as many offspring as do those of type Aa.

Part IV

Recessive Lethal Genes

Suppose some gene is so bad that it is certain to kill the individual before that individual has a chance to breed. What will happen to the frequency of that gene? The interesting case is the "recessive lethal", where every individual that has two copies of that allele dies before reproduction. It is easy to imagine a situation like this—suppose the locus codes for some protein that is vital to life, but that an individual can survive just fine with only one functional copy of a gene. In other words, individuals of type AA and Aa do just fine, but individuals of type aa are eliminated from the next generation.

If the initial condition of the population has A-type alleles with probability p and a-type alleles with probability $q = 1 - p$, the resulting distribution of the offspring will have p^2 individuals of type AA, $2pq$ individuals of type Aa, and q^2 individuals of type aa, none of whom survive. So after mating and the death of the unlucky individuals who got the aa combination, what remains of the original population is a relative proportion of p^2 type-AA individuals and $2pq$ type-Aa individuals. Thus the relative proportions of allele A and a are $p^2 + pq$ and pq, respectively.

These are *relative* proportions, however. To get probabilities, we need to divide by $p^2 + 2pq$, giving the probabilities of finding a type-A and type-a allele $p' = (p^2 + pq)/(p^2 + 2pq)$ and $q' = pq/(p^2 + 2pq)$, respectively. (p' and q' are the new probabilities of finding the alleles after breeding.)

So what happens to a population like this over time? Recalling that $p + q = 1$ (so $p = 1 - q$), we obtain:

$$\begin{aligned} q' &= \frac{pq}{p^2 + 2pq} = \frac{q(1-q)}{(1-q)^2 = 2q(1-q)} \\ &= \frac{q - q^2}{1 - 2q + q^2 + 2q - 2q^2} = \frac{q - q^2}{1 - q^2} \end{aligned}$$

$$= \frac{q(1-q)}{(1+q)(1-q)} = \frac{q}{1+q}.$$

Thus the probability of having the lethal gene a goes from q to $q/(1+q)$ with each generation. Since $q > 0$, $(1+q) > 1$, so with each generation, the number of alleles of type a decreases. Natural selection gradually eliminates the lethal gene from the population.

But how fast?

If we denote by the function $f(q) = q/(1+q)$ the result of one generation of selection, then $f(f(q))$ represents the result after two generations, $f(f(f(q)))$ the result after three generations, et cetera. Let's calculate $f(f(q))$ and $f(f(f(q)))$:

$$\begin{aligned} f(f(q)) &= f(q/(1+q)) \\ &= \frac{q/(1+q)}{1+q/(1+q)} \\ &= q/(1+2q). \end{aligned}$$

$$\begin{aligned} f(f(f(q))) &= f(q/(1+2q)) \\ &= \frac{q/(1+2q)}{1+q/(1+2q)} \\ &= q/(1+3q). \end{aligned}$$

If we represent by $f^{(n)}(q)$ the result of n iterations of the function f, it is not hard to prove using induction that:

$$f^{(n)}(q) = q/(1+nq).$$

To get a feeling for this, let's assume that initially (we will call this initial state "generation zero") the population has 50% of each kind of allele. Then as time goes on, here is the proportion of allele a:

Generation	0	1	2	4	8	100	1000
Proportion a	.5	.3333	.25	.1667	.1	.0098	.000998

This is a very slow elimination of the gene—after 1000 generations, still nearly one in a thousand alleles is of the lethal recessive type.

For our next example, consider what happens if the allele a is still lethal when it appears as type-aa, but also is slightly deleterious if it occurs in the form Aa. The fitnesses of the three types AA, Aa, and aa will be $1+s$, 1, and 0, respectively, where s is a small positive number. If $s = .1$, for example, this means that type-AA is 10% more likely to produce offspring as type-Aa.

The proportions after breeding and elimination of the less-fit individuals will be type-AA: $(1+s)p^2$ and type-Aa: $2pq$. Doing exactly the same as above, we obtain:

$$\begin{aligned} q' &= \frac{pq}{(1+s)p^2+2pq} \\ &= \frac{q(1-q)}{(1+s)(1-q)^2+2q(1-q)} \\ &= \frac{q}{(1+s)(1-q)+2q} = \frac{q}{(1+s)+q(1-s)}. \end{aligned}$$

Again, we would like to let $f(q) = q'$ and evaluate the functions $f(f(q))$, $f(f(f(q)))$, et cetera. To simplify the calculation, let's temporarily let $\alpha = (1+s)$ and $\beta = (1-s)$, giving us:

$$q' = f(q) = \frac{q}{\alpha + \beta q}. \tag{1}$$

To figure out what's going on, work out a few of the terms by hand to get:

$$\begin{aligned}
f(q) &= \frac{q}{\alpha+\beta q}\\
f(f(q)) &= \frac{q}{\alpha^2+\beta(1+\alpha)q}\\
f(f(f(q))) &= \frac{q}{\alpha^3+\beta(1+\alpha+\alpha^2)q}\\
f(f(f(f(q)))) &= \frac{q}{\alpha^4+\beta(1+\alpha+\alpha^2+\alpha^3)q}.
\end{aligned}$$

The pattern above is obvious (and easy to prove by induction):

$$f^{(n)}(q) = \frac{q}{\alpha^n+\beta(1+\alpha+\cdots+\alpha^{n-1})q}.$$

If we recall that

$$1+\alpha+\cdots+\alpha^{n-1} = \frac{1-\alpha^n}{1-\alpha},$$

we obtain:

$$f^{(n)}(q) = \frac{q}{\alpha^n+\beta\left(\frac{1-\alpha^n}{1-\alpha}\right)q}.$$

Using this formula, let's check the rate of disappearance of the allele a as a function of time as we did earlier, but this time, we will let $s = .01$, so type-AA is about 1% more likely to survive than type-Aa. We'll again begin with $q = .5$—half the alleles are originally of type-a:

Generation	0	1	2	4	8	100	1000
Proportion a	.5	.3322	.2481	.0964	.1	.0057	.000000472

A quick comparison shows that convergence to zero (the elimination of the detrimental gene) is *much* more rapid if there is even a tiny disadvantage to the heterozygote.

But now let's examine the situation if s is negative—in other words, the heterozygote Aa is slightly *more* fit than the homozygote AA. The formula is exactly the same, but in the table below, $s = -.01$:

Generation	0	1	10	50	500	50000	1000000
Proportion a	.5	.3344	.0872	.0243	.009965	.009901	.009901

The allele a certainly decreases in frequency, but eventually seems to get "stuck" at about 1% of the total. Let's do a similar experiment, but start with $q = .00001$—just the tiniest amount of allele a in the population:

Gen.	0	1	10	50	500	50000	1000000
Prop. a	.0001	.000101	.00011	.00016	.006	.009901	.009901

This time, the frequency of a actually rises until it reaches the same value as previously. What we see above is a special case of the situation known as "heterozygote advantage". In general, if the heterozygote form Aa is more fit than either AA or aa, natural selection will not eliminate either gene, assuming that both are initially present in the population.

Part V

Iterated Functions

One of the features of mathematical genetics is that we frequently use iterated functions—we take the output of a function and plug it back into the original function. This is because the "input" to the next generation is the same as the "output" of the previous generation.

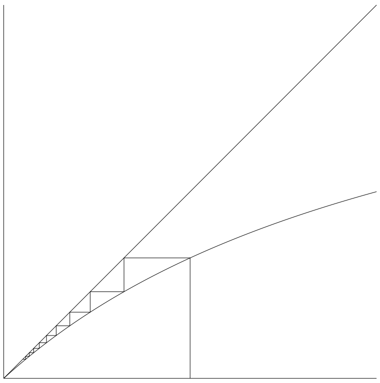

Figure 1: $f(q) = q/(1.1 + .9q)$

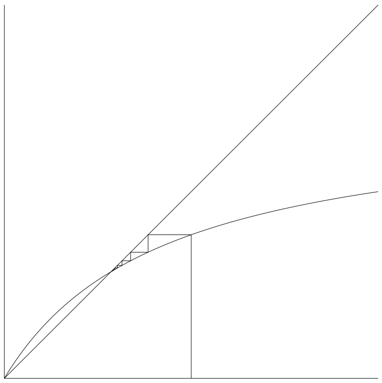

Figure 2: $f(q) = q/(.6 + 1.4q)$

Figure 1 provides a nice example. We have plotted Equation (1) on the same set of axes as the line $y = x$. Then we pick a starting point (in this case, at $x = .5$, the beginning of the iteration, and draw a line up to the curve. The height of the curve is the output value, and if we wish to use that as an input value, we just go horizontally to the line $y = x$, and wherever it meets, that's the new x coordinate for input. Repeat the process, and the staircased line shows how the function iterates—in this case to the fixed point $x = y = 0$.

In fact, wherever our curve crosses the line $y = x$ is a fixed point of the function—the input is equal to the output. Let's consider a different function (just Equation (1) with a different s) that corresponds to a negative s with heterozygote advantage: $f(q) = q/(.6 + 1.4q)$. The result is plotted in Figure 2.

In this case, note that the curve crosses the line $y = x$ at a point other than $(0, 0)$, so it is a possible fixed point. If we trace the staircase, we can see that it converges to the fixed point. As an exercise, trace the staircase starting at a value of x less than the x-coordinate of the fixed point.

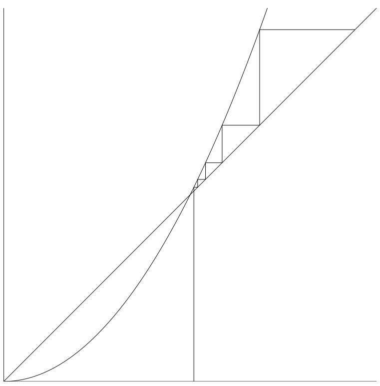

Figure 3: $f(q) = q^2/2$

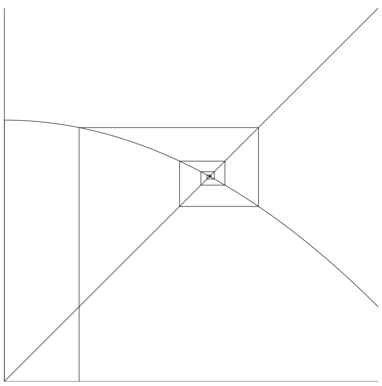

Figure 4: $f(q) = .7 - q^2/2$

Finally, let's look at another type of fixed point as shown in Figure 3. This time the curved line crosses the line $y = x$

from below, and the staircase, even though it began quite near the point of equilibrium, diverged. Try starting just below the point of equilibrium and see what happens.

A final and very interesting example can be seen in Figure 4. There is convergence again, but this time in a spiral. Can you construct a similar example, but where divergence occurs (in a spiral)?

When convergence occurs, notice how easy it is to find the point of convergence. For example, in Figure 2, the iteration will converge to the point where $f(q) = q$. We can solve for that equation as follows:

$$\begin{aligned}
f(q) \ = \ q &= \frac{q}{.6 + 1.4q} \\
(.6 + 1.4q)q &= q \\
.6 + 1.4q &= 1 \\
1.4q &= .4 \\
q &= .4/1.4 \ = \ 2/7.
\end{aligned}$$

Part VI

Fitness Versus Frequency Plots

All the plots that follow have a vertical axis from 0 to 1 that represents the proportion of allele A. The horizontal axis represents generations, from 0 to 100. The 11 curves represent equally-spaced starting proportions of A from 0.01 to 0.99.

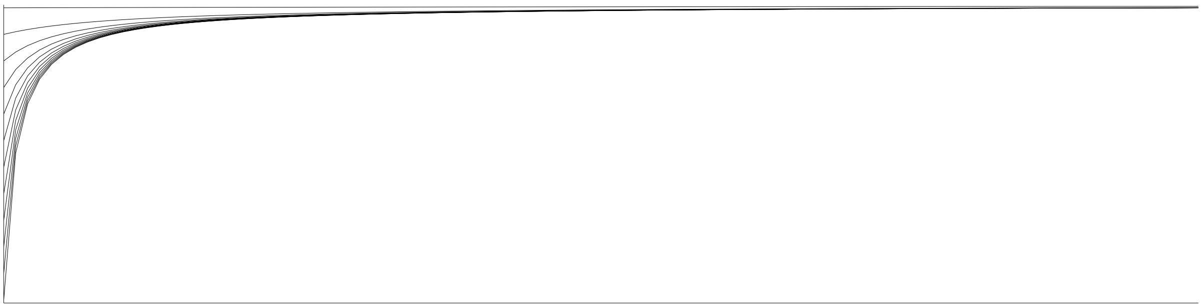

Figure 5: $F_{AA} = 1.0$; $F_{Aa} = 1.0$; $F_{aa} = 0.0$

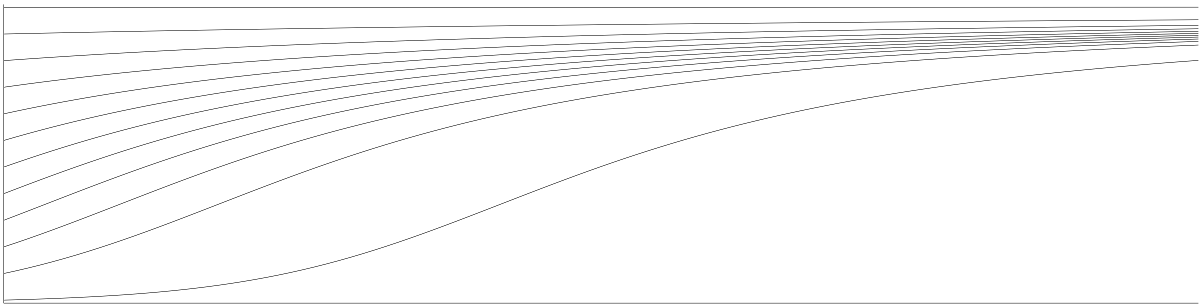

Figure 6: $F_{AA} = 1.0$; $F_{Aa} = 1.0$; $F_{aa} = 0.9$

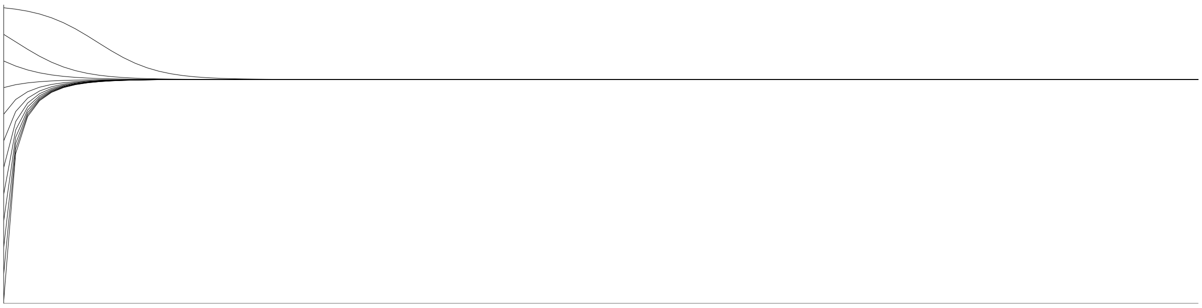

Figure 7: $F_{AA} = 1.0$; $F_{Aa} = 1.5$; $F_{aa} = 0.0$

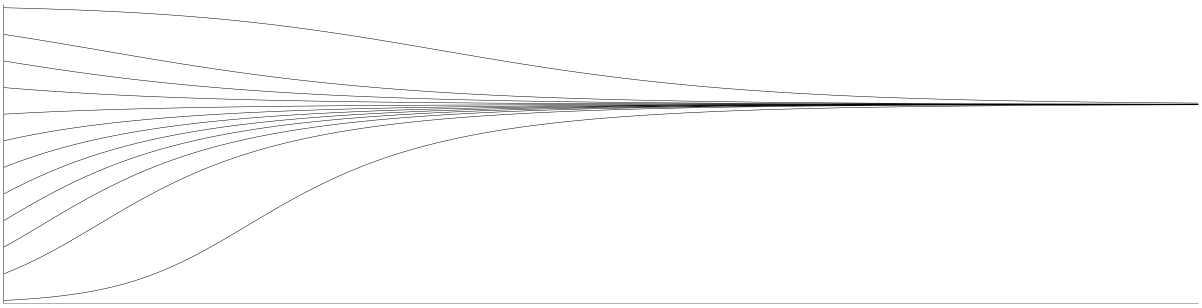

Figure 8: $F_{AA} = 1.0$; $F_{Aa} = 1.1$; $F_{aa} = 0.9$

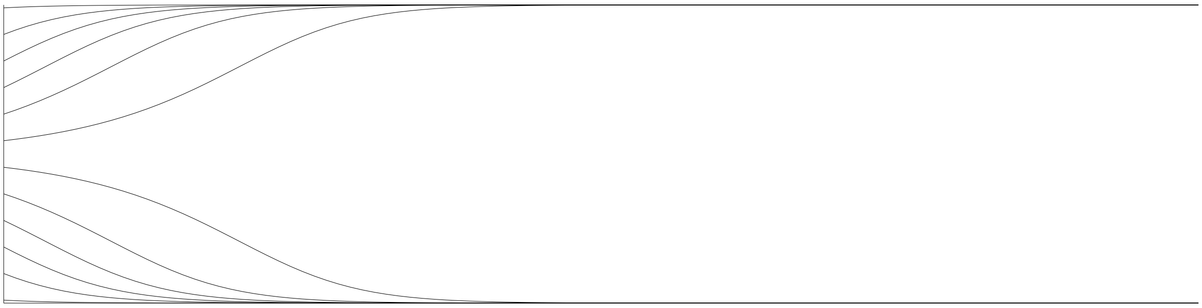

Figure 9: $F_{AA} = 1.0$; $F_{Aa} = 0.8$; $F_{aa} = 1.0$

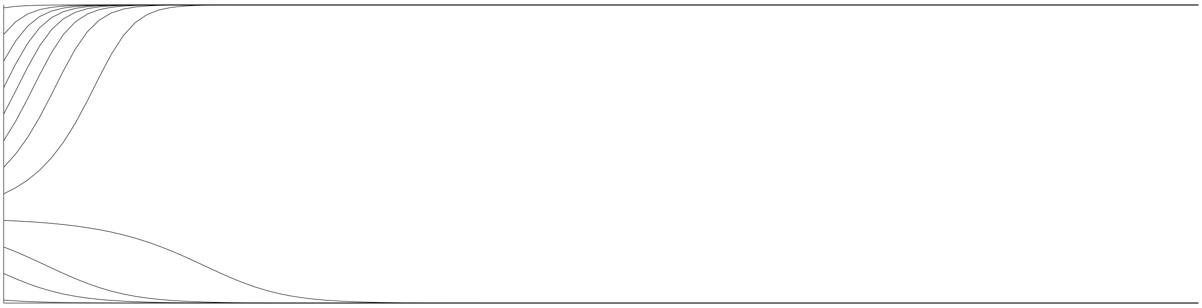

Figure 10: $F_{AA} = 1.0$; $F_{Aa} = 0.5$; $F_{aa} = 0.7$