THE ELIMINATION OF DOUBLE DOMINANTS IN LARGE RANDOM MATING POPULATIONS

By J. B. S. HALDANE and S. D. JAYAKAR
Genetics and Biometry Laboratory, Government of Orissa, Bhubaneswar

Introduction

The mathematical theory of natural and artificial selection is likely to become as complicated as that of gravitational astronomy. However a few problems are fairly simple. The deterministic theory of artificial, or complete, selection based on the elimination of one phenotype is simple when the phenotypes are genetically determined by a single factor, or a pair of allelomorphic genes, in a very large population. A fully or partially dominant gene can be eliminated in one generation. The effects of mutation are neglected in this article. In a moderately large population they must be treated stochastically. In a very large one they can be treated deterministically, the number of mutants in each generation becoming nearly constant. Punnett (1917) gave the expression for the rate at which recessives are eliminated when they are prevented from breeding in a random mating population segregating for a pair of autosomal allelomorphs. If both homozygotes are eliminated segregation continues indefinitely. If heterozygotes are eliminated then if the ratio of **AA** to **aa** in generation n is u_n,

$$u_{n+1} = u_n^2,$$
$$\text{so} \quad u_n = u_o^{2^n}$$
$$\text{or} \log u_n = 2^n \log u_o. \tag{1.1}$$

So there is a very unstable equilibrium when $u=1$, the two genes being equally frequent; and in general one or other disappears very rapidly. However in a finite population the theory becomes stochastic, and even the simplest problems have not yet been fully solved.

When two pairs of allelomorphs, or factors, are concerned, then unless selection is complete in one generation, the mathematics are at least as difficult as those raised by slow selection at a single locus, and in the case here solved, are formally like them. Consider a large population segregating for two pairs of autosomal allelomorphs. Let the n-th generation be formed by random mating, the gametic pool being

$$p_n\mathbf{AB} + q_n\mathbf{Ab} + r_n\mathbf{aB} + s_n\mathbf{ab},$$

where $p_n+q_n+r_n+s_n=1$. If the population includes double heterozygotes the values will not be the same in the two sexes if the loci concerned are linked. We may however neglect this complication if double heterozygotes are not permitted to breed. The next generation consists of the genotypes shown in Fig. 1.

If **A** and **B** are dominant, there are four phenotypes, indicated by the four areas in Fig. 1. They may be designated **A- B-, A- bb, aa B-,** and **aa bb.** Any one, two

or three may be eliminated as breeders of the next generation. There are therefore ten types of complete selection.

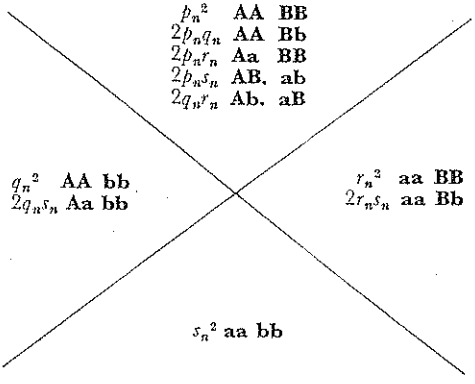

Fig. 1. Genotype frequencies in F_n

(1) **A- B-** eliminated. This is the type here considered.
(2) **A- bb** *or* **aa B-** eliminated.
(3) **aa bb** eliminated.
(4) **A- B-** and **A- bb** *or* **A- B-** and **aa B-** eliminated.
(5) **A- B-** and **aa bb** eliminated.
(6) **A- bb** and **aa B-** eliminated.
(7) **A- bb** and **aa bb** *or* **aa B-** and **aa bb** eliminated.
(8) All but **A- B-** eliminated.
(9) All but **A- bb** *or* **aa B-** eliminated.
(10) All but **aa bb** eliminated.

Type (10) leads to a homogeneous **aa bb** population in one generation. Type (4) causes the elimination of **A** or **B** as the case may be in one generation. After this the population is in stable equilibrium, segregating for **B** and **b** or for **A** and **a**. Type (9) also leads to a simple result. If **A- bb** is selected, **B** is eliminated in one generation, and after this **a** is slowly eliminated according to Punnett's calculation, the final genotype being **AA bb**. Type (5) is very artificial, and unlikely to occur as the result of natural or human selection.

In the other five types of selection, namely (2), (3), (6), (7), and (8), double heterozygotes of both types breed. So though their treatment is not very difficult when **A** and **B** are unlinked, the algebra becomes complicated when they belong to the same linkage group. It is often, however, easy to see what the final result will be; for example if double dominants **A- B-** are selected, the final population will be all **AA BB**, though we have not found explicit expressions for its composition after n generations. In the next section we give a partial solution for type (3).

The final results are as follows. In cases (2), (4), (7), and (9) the first alternative only is considered, e.g. in case (2) **A- bb** is eliminated.

(1) The rarer of **A** and **B** disappears.

(2) **A** *or* **b** disappears.
(3) The rarer of **a** and **b** disappears.
(4) **A** disappears.
(5) **A** and **b** *or* **a** and **B** disappear.
(6) **A** and **B** *or* **a** and **b** disappear.
(7) **b** disappears.
(8) **a** and **b** disappear.
(9) **a** and **B** disappear.
(10) **A** and **B** disappear.

Where only one gene disappears, the other locus gives a Hardy-Weinberg equilibrium.

THE CALCULATION OF EQUILIBRIA. THE FICTION OF NON-VIOLENCE.

In some cases the elimination of one gene is necessarily accompanied by that of another of known type. Thus if **aa bb** is eliminated, for every **b** gene removed, one and only one **a** gene is also removed, so the process will continue till all the **a** genes have gone, if they are less frequent than **b.** It will then cease. This is true for any mating system which does not lead to complete homozygosis. In such cases it is sometimes useful to adopt the following fiction. We start with N individuals, formed from $2N$ gametes. This number is constant in each generation. But instead of the discarded individuals being killed or castrated, they are subjected to *apartheid*, and form a new population with equal opportunities for breeding. Thus in type (3) selection there were $2p_1N$ **AB**, $2q_1N$ **Ab**, $2r_1N$ **aB**, $2s_1N$ **ab** gametes forming the first generation. Suppose $q_1 > r_1$. That is to say there were more **b** than **a** gametes. In each generation some **a** gametes are removed, and an equal number of **b.** Finally we have in the part of the population removed $2(r_1+s_1)N$ **ab** gametes, or $(r_1+s_1)N$ **aa bb** zygotes. The remainder contains $2(q_1+s_1)N-2(r_1+s_1)N$, i.e., $2(q_1-r_1)N$ **b** gametes, along with $2(p_1+r_1)N$ **B** gametes. The final composition is thus $2(p_1+r_1)N$ **AB**, $2(q_1-r_1)N$ **Ab**, and (fictionally) $2(r_1+s_1)N$ **ab.** The frequency of **AB** gametes in the selected population is thus

$$\frac{p_1+r_1}{p_1+q_1}, \text{ that of } \mathbf{Ab} \text{ gametes } \frac{q_1-r_1}{p_1+q_1}; \text{ or}$$

$$p_\infty = \frac{p_1+r_1}{p_1+q_1}, \quad q_\infty = \frac{q_1-r_1}{p_1+q_1}, \quad r_\infty = s_\infty = 0. \tag{1.1}$$

Applying the same treatment where **A- B-** is eliminated we find that after the first generation the zygotes eliminated are all **AB. ab** and **Ab. aB.** Thus one **A** is eliminated for every **B.** And if $q_1 > r_1$, we are left with

$$q_\infty = \frac{q_1-r_1}{q_1-r_1+s_1}, \quad s_\infty = \frac{s_1}{q_1-r_1+s_1}, \quad p_\infty = r_\infty = 0. \tag{1.2}$$

If, on the other hand **A- bb** individuals are eliminated (type 2), it is not so simple to calculate what will happen. $q_\infty = 0$, and either p_∞ or $s_\infty = 0$, but the dynamics must, we think, be investigated before one can decide between these alternatives, at least for some sets of values of p_1, q_1, r_1, and s_1.

Elimination of Double Dominants. The Trajectories.

Any population can be represented by a point in a space of $g-1$ dimensions where there are g diploid genotypes. If however mating is at random we only need space of one dimension less than the number of haploid genotypes. A glance at Fig. 1 shows that the population of diploids could be represented by a point in a simplex in 9-dimensional space. However its gametes could be represented by a point in a tetrahedron in three dimensions. Within such a simplex the points representing successive generations will lie on a line, straight or otherwise, which may be called a trajectory. It is often possible to calculate these trajectories when one cannot calculate where the representative point will be after n generations without the use of new mathematical functions, or of an electronic computer.

If double dominants are eliminated, p_n is zero after the first generation, so $q_n+r_n+s_n=1$, and we can represent our population by a point in a triangle, or on one or two of its edges (Fig. 2). Thus if p_0 is positive, $p_1=0$, and our calculation can

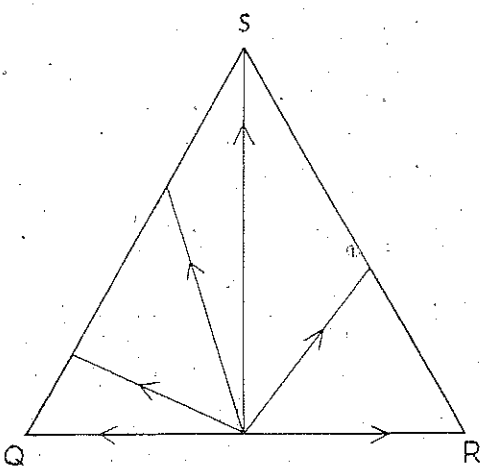

Fig. 2. Trajectories of the point (q_n, r_n, s_n) representing the gametes of F_n

conveniently start with $n=1$. In each generation a fraction $2q_nr_n$ is eliminated, and in the next generation

$$\left.\begin{aligned} q_{n+1} &= \frac{q_n(1-r_n)}{1-2q_nr_n} \\ r_{n+1} &= \frac{r_n(1-q_n)}{1-2q_nr_n} \\ s_{n+1} &= \frac{s_n}{1-2q_nr_n}. \end{aligned}\right\} \qquad (2.1)$$

First let us find the equilibrium. s_{n+1} can only equal s_n, if $s_n=0$, $q_n=0$, or $r_n=0$.

So one or two of these three must vanish at any equilibrium, which is therefore represented by a point on an edge of the triangle or by one of its corners. Unless $q_n=0$, or $r_n=0$, $s_{n+1}>s_n$ from (2.1). So s_∞ cannot be zero unless $s_1=0$. If $s_1=0$, then, as will be shown later, unless $q_1=r_1=\frac{1}{2}$, one will vanish, so $q_\infty=1$ and $r_\infty=1$ are stable equilibria, and $q_\infty=r_\infty=\frac{1}{2}$ an unstable one.

If $s_1=0$, then if q_1 or $r_1=0$ the population is already in equilibrium. I suppose that none of q_1, r_1, or s_1 is zero.

$$\triangle q_n = \frac{q_n r_n (2q_n-1)}{1-2q_n r_n}.$$

So if $q_n>\frac{1}{2}$, $q_{n+1}>q_n$. Thus q_n increases with n if $q_1>\frac{1}{2}$, remains constant if $q_1=\frac{1}{2}$, and diminishes if $q_1<\frac{1}{2}$. Thus $q=\frac{1}{2}$ is a trajectory. It follows from (2.1) that

$$\frac{q_{n+1}-r_{n+1}}{s_{n+1}} = \frac{q_n-r_n}{s_n}. \tag{2.2}$$

If $q_1>r_1$ this is positive, and we immediately obtain the expressions (1.2) for the stable equilibria. The trajectories are given by

$$\frac{q_n-r_n}{q_1-r_1} - \frac{s_n}{s_1} = 0, \tag{2.3}$$

which are represented by straight lines in Fig. 2. Thus all trajectories pass through the foot of the perpendicular from S on QR, where $q=r=\frac{1}{2}$, $s=0$. There are three special trajectories, two which run along QR, or $s=0$, and one which runs along its perpendicular bisector, or $q-r=0$. This has a special interest, as it would represent the state of affairs in F_n from a cross between two pure lines. In a real and therefore finite population the representative point would never follow one of these lines except $s=0$ exactly, but would remain close to one in a large population.

In a population where all of p_n, q_n, r_n, and s_n are non-zero this simple type of argument will not permit the calculation of the trajectory through a given point; but it will permit the calculation of a plane or other surface in which it must lie.

DYNAMICS IN THE GENERAL CASE

From (2.3), $$\frac{q_n-r_n}{1-q_n-r_n} = \frac{q_1-r_1}{s_1} = \frac{1-s_\infty}{s_\infty} = \frac{1-S}{S},$$

where $S=s_\infty=\dfrac{s_1}{q_1-r_1+s_1}$, if $q_1>r_1$, in which case $r_\infty=0$.

$$q_n=1-S+(2S-1)\,r_n,\ s_n=S\,(1-2r_n).$$

Hence from (2.1)

$$r_{n+1} = \frac{r_n[S+(1-2S)r_n]}{1+2(S-1)r_n+2(1-2S)r_n^2}. \tag{3.1}$$

From this equation successive values of r_n may be calculated. Unless $1-S$ is small, it tends rapidly to zero. For example if $q_1=0{\cdot}6$, $r_1=0{\cdot}3$, $s_1=0{\cdot}1$, $S=0{\cdot}25$,

$$r_{n+1} = \frac{r_n(1+2r_n)}{2(2-3r_n+2r_n^2)}, \text{ so}$$

$$r_2=\frac{3}{16}=\cdot1875,\ r_3=\frac{33}{386}=\cdot0854922,\ r_4=\frac{3729}{130978}=\cdot02847021, \text{ etc.}$$

Hence $q_4=\cdot735765$, $s_4=\cdot235765$.

(3.1) is only soluble in finite terms if $q_1=\frac{1}{2}$, whence $S=\frac{1}{2}$, when

$$r_{n+1} = \frac{r_n}{2(1-r_n)}, \text{ so that}$$

$$r_n = \frac{r_1}{2^n(1-2r_1)+2r_1}. \tag{3.2}$$

When $S<\frac{1}{2}$, we can calculate successive values of r_n without much difficulty, as in the example already given. But when S is near unity, r_n only approaches zero slowly. To calculate its value in this case, we use the method of Haldane (1963), and put

$$x_n=\frac{(1-S)\sqrt{2S-1}\ r_n}{S(1-S)+(2S^2-1)r_n}, \text{ whence } r_n=\frac{S(1-S)x_n}{(1-S)\sqrt{2S-1}-(2S^2-1)x_n}.$$

$$\text{Then}\qquad x_{n+1}=Sx_n\left(1-\frac{x_n^2}{1-ax_n}\right)^{-1}$$

$$\text{where}\qquad a=\frac{3S-2}{(1-S)\sqrt{2S-1}},$$

that is to say the values of x_n approximate much more closely to a geometric progression than do those of r_n.

$$\text{Now let } (n+C)\ \ln S=\ln x_n+\sum_{r=2}^{\infty} b_r x_n^r,$$

where C and b_r are constants to be determined. It follows that

$$(n+1+C)\ \ln S=\ln x_{n+1}+\sum_{r=2}^{\infty} b_r x_{n+1}^r.$$

On subtracting we find the identity

$$\sum_{r=2}^{\infty} b_r x_n^r\left[1-S^r\left(1-\frac{x_n^2}{1-ax_n}\right)^{-r}\right]\equiv-\ln\left(1-\frac{x_n^2}{1-ax_n}\right).$$

On equating the coefficients of powers of x_n we find

$$(n+C)\ln S=\ln x_n+\sum_{r=2}^{\infty}\frac{a^{r-2}x_n^2}{1-S^r}+\frac{x_n^4}{1-S^4}\left(\frac{2}{1-S^2}-\frac{3}{2}\right)+$$

$$\frac{ax_n^5}{1-S^5}\left(\frac{2}{1-S^2}+\frac{3}{1-S^3}-4\right)+\ldots \tag{3.3}$$

Provided x_n and ax_n are numerically small, and $1-S$ is not very small, this is close to

$$n+C=\frac{\log_{10}x_n}{\log_{10}S}, \tag{3.4}$$

the error being approximately $\frac{3\cdot04x_n^2}{7(1-S^2)\log_{10}S}$.

If $S=\frac{2}{3}$, $a=0$, and (3·3) becomes

$$(n+C)\ln S=\ln x_n+\frac{x_n^2}{1-S^2}+\left(\frac{2}{1-S^2}-\frac{3}{2}\right)\frac{x_n^4}{1-S^4}+\left[\frac{8}{(1-S^2)(1-S^4)}-\frac{6}{1-S^4}+\frac{3}{1-S^2}-\frac{14}{3}\right]\frac{x_n^6}{1-S^6}+\ldots. \tag{3.5}$$

Once C is calculated, we can derive from (3.3)

$$S^{n+C}=x_n+\sum_{r=2}^{\infty}\left[\frac{a^{r-2}x_n^{r+1}}{1-S^r}\right]+\left[\frac{3}{1-S^2}-2\right]\frac{x_n^5}{1-S^4}+\left[\frac{3}{1-S^2}+\frac{4}{1-S^3}-5\right]\frac{ax_n^6}{1-S^5}+\ldots \tag{3.6}$$

And if $t=S^{n+C}$,

$$x_n=t-\sum_{r=2}^{\infty}\left[\frac{a^{r-2}t^{r+1}}{1-S^r}\right]+\left[\frac{3}{1-S^2}-1\right]\frac{t^5}{1-S^4}+\left[\frac{4}{1-S^2}-2\right]\frac{at^6}{1-S^5}+\ldots \tag{3.7}$$

In practice (3.4) is generally accurate enough. For example let $q_1=\cdot244$, $r_1=\cdot1$, $s_1=\cdot656$, whence

$$S=\cdot82,\quad a=\frac{115}{36}=3\cdot1944,\quad q_\infty=\cdot18,\quad r_\infty=0,\quad s_\infty=\cdot82.$$

Also $r_{n+1}=\frac{r_n(41-32r_n)}{50-18r_n-64r_n^2}$, $x_n=\frac{360r_n}{369+862r_n}$, $r_n=\frac{369x_n}{360-862x_n}$.

Hence $r_2=\cdot079479$, the progress to zero being rather slow.

$$x_1=\frac{45}{569}=\cdot0794451,\text{ so from (3.4)}$$

$$1+C=12\cdot7851.$$

The first two corrections given by (3·3) are $-\cdot0962$ and $-\cdot0178$. The correct value of C is thus about 11·7. But the simple formula (3·4) is sufficient, and its error becomes quite negligible when x_n is small. Taking $C=11\cdot7$, we can answer the following questions. How many generations are needed to reduce r_n to ·001? $x_n=\cdot00097334$. So $n+C=34\cdot945$, $n=23\cdot2$. What is the value of r_n when $n=30$?

$$\log_{10}x_n=4\cdot40604,\quad x_n=2\cdot547\times10^{-4},\quad r_n=2\cdot609\times10^{-4}.$$

THE CASE WHEN GENE FREQUENCIES ARE EQUAL

If two pure lines have been crossed, say **AA bb** and **aa BB**, the frequencies of **a** and

b will be the same, that is to say $q_n = r_n$, so $q_\infty = r_\infty = 0$, $s_\infty = 1$.

From (2·1) we have

$$q_{n+1} = \frac{q_n(1-q_n)}{1-2q_n^2} \tag{4.1}$$

If the F_1 is **AB.ab,** and c and c' are the recombination frequencies in the two sexes, then in F_2 after culling the double dominants,

$$q_2 = \frac{c+c'}{2(1+c+c'-cc')} . \tag{4.2}$$

If the F_1 is **Ab.aB**, $q_2 = \dfrac{2-c-c'}{2(2-cc')}$.

Thus $\frac{1}{2} > q_n > 0$, and in the absence of linkage $q_2 = \frac{2}{7} = \cdot 2857$, $q_3 = \cdot 2439$, $q_4 = \cdot 2093$, $q_5 = \cdot 1655$, etc. The values diminish rather slowly, and ultimately approximate to a harmonic progression. To calculate them when n is large, we put

$$n + C = \frac{1}{q_n} - \ln q_n - \tfrac{1}{2}\ln(1-q_n) + \tfrac{3}{2}\ln(1-2q_n) + \sum_{r=2}^{\infty} b_r q_n^r .$$

$$\text{So } n+1+C = \frac{1}{q_{n+1}} - \ln q_{n+1} - \tfrac{1}{2}\ln(1-q_{n+1}) + \tfrac{3}{2}\ln(1-2q_{n+1}) + \sum_{r=2}^{\infty} b_r q_{n+1}^r$$

On subtracting we have the identity :—

$$\sum_{r=2}^{\infty} b_r q_n^r \left[1 - \left(\frac{1-q_n}{1-2q_n^2}\right)^r\right] \equiv -q_n(1-q_n)^{-1} - \tfrac{1}{2}\ln(1-q_n) - \tfrac{1}{2}\ln(1-q_n-q_n^2).$$

Whence $n + C = q_n^{-1} - \ln q_n - \frac{1}{2}\ln(1-q_n) + \frac{3}{2}\ln(1-2q_n) - \frac{1}{12}q_n^2(1 + \frac{5}{3}q_n + \frac{23}{20}q_n^2 - \frac{19}{15}\frac{7}{0}q_n^3$

$$- \tfrac{7}{1}\tfrac{4}{2}\tfrac{3}{6}q_n^4 \ \ldots\ldots) \tag{4.3}$$

The series probably diverges, but even when $q_2 = \frac{2}{7}$ the last term is only ·0003, and it is best merely to sum to the numerically smallest term, whence

$$C = 1 \cdot 640.$$

Hence we can, for example, calculate that if $q_n = \cdot 01$, $n = 102 \cdot 94$. The value of C of course depends on that of q_2. It must however be emphasized that in any but a very large population the frequencies of **a** and **b** will become unequal as the result of "drift," and there will almost always be some selection with a similar effect. Thus the final population will almost certainly contain some **A** or some **B** genes.

Discussion

While such calculations as this are inevitably artificial because populations are always finite, and genes never or rarely either selectively neutral or immutable, they have the same kind of value as those of "classical" physics or the dynamics of bodies

conceived of as frictionless, perfectly elastic, perfectly rigid, and so on. It is notable that such calculations, which were sometimes stated to be only of academic interest, turned out during the present century to be applicable to particles of atomic and lesser size. Many fairly simple genetical problems of this type are beyond the scope of existing mathematics, but could easily be solved by an electronic computer. We believed that an international committee could usefully draw up a list of such problems, with a view to such computation.

REFERENCES

HALDANE, J. B. S. (1963). (*in press*).

PUNNETT, R. C. (1917). Eliminating feeblemindedness. *Journ. Hered.*, **8**, 464.