A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. By Mr J. B. S. HALDANE, Trinity College.

[*Received* 21 May, *read* 25 July, 1927.]

New factors arise in a species by the process of mutation. The frequency of mutation is generally small, but it seems probable that it can sometimes be increased by changes in the environment (1, 2). On the whole mutants recessive to the normal type occur more commonly than dominants. The frequency of a given type of mutation varies, but for some factors in Drosophila it must be less than 10^{-6}, and is much less in some human cases. We shall first consider initial conditions, when only a few of the new type exist as the result of a single mutation; and then the course of events in a population where the new factor is present in such numbers as to be in no danger of extinction by mere bad luck. In the first section the treatment of Fisher (3) is followed.

In a large population let p_r be the chance that a factor present in a zygote at a given stage in the life-cycle will appear in r of its children in the next generation. If the individual considered is homozygous, this is the chance of leaving r children, if mutation is neglected. Let $\sum\limits_{r=0}^{\infty} p_r x^r = f(x)$. Therefore $f(1) = 1$, $f(0) = p_0$, the probability of the factor disappearing, while $f'(1) = \sum\limits_{r=0}^{\infty} rp_r$, *i.e.*, the probable number of individuals possessing the factor in the next generation. The probability of m individuals bearing one each of the factors considered leaving r descendants is clearly the coefficient of x^r in $[f(x)]^m$, if we neglect the possibility of a mating between two such individuals, which we may legitimately do if m is small compared with the total number of the population. If then the probability of the factor being present in r zygotes of the nth generation be the coefficient of x^r in $F(x)$, the corresponding probability in the $(n+1)$th generation is the same coefficient in $F[f(x)]$. Hence if a single factor appears in one zygote, the probability of its presence in r zygotes after n generations is the coefficient of x^r in $S_f^n(x)$, *i.e.* $f(f(f...f(x)...))$, the operation being repeated n times. The probability of its disappearance is therefore $\operatorname*{Lt}\limits_{n\to\infty} S_f^n(0)$. By Koenigs' theorem (4) this is the root of $x = f(x)$ in the neighbourhood of zero.

https://doi.org/10.1017/S0305004100015644 Published online by Cambridge University Press

Now in the case of a dominant factor appearing in a population in equilibrium, and conferring an advantage measured by k, as in Part I (5), $f'(1) = 1 + k$. Since $f'(x)$ and $f''(x)$ are positive when x is positive, and $f(0)$ is positive, $x = f(x)$ has two and only two real positive roots, one equal to unity, the other lying between 0 and 1, but near the latter value if k be small. Hence any advantageous dominant factor which has once appeared has a finite chance of survival, however large the total population may be.

If a large number of offspring is possible, as in most organisms, the series p_n approximates to a Poisson series, provided that adult organisms are counted, and since $f'(1) = 1 + k, f(x) = e^{(1+k)(x-1)}$. Hence the probability of extinction $1 - y$ is given by

$$1 - y = e^{-(1+k)y}.$$

Hence $$(1 + k)y = -\log(1 - y) \text{..............} (1{\cdot}0),$$

and $$k = \frac{y}{2} + \frac{y^2}{3} + \frac{y^3}{4} + \dots,$$

and if k be small, $y = 2k$ approximately. Hence an advantageous dominant gene has a probability $2k$ of survival after only a single appearance in an adult zygote, and if in the whole history of a species it appears more than $\frac{\log_e 2}{2k}$ times it will probably spread through the species. But, however large k may be, the factor may be extinguished after a single appearance. Thus if $k = 1$, so that the new type probably leaves twice as many offspring as the normal, the probability of its extinction is still ·203. If in any generation there are m dominant individuals the probability of extinction is reduced to y^m, where y is the smaller positive root of $x = f(x)$. When k is small this reduces to $(1 - 2k)^m$. Hence if in any generation more than $\frac{\log_e 2}{2k}$ adult dominants exist, the factor will probably spread through the whole population.

On the other hand a recessive factor whose phenotype is advantageous has a quite negligible advantage in a random mating population provided that the number of its bearers is small compared with the square root of the total population. This is best seen by considering the case of a hermaphrodite: in a dioecious organism the argument, though similar, is more complicated. Let N be the fixed number of the population, and z_n the number of heterozygotes plus double the number of recessives for the factor A in the nth generation. It therefore produces gametes in the ratio $(2N - z_n) A : z_n a$. If now the recessives have a small advantage measured by k, the probabilities of production of each genotype in the next generation are

$$(2N - z_n)^2 AA : 2z_n(2N - z_n) Aa : (1 + k) z_n^2 aa.$$

https://doi.org/10.1017/S0305004100015644 Published online by Cambridge University Press

Hence if, as above, $f(x)$ be the function defining the probable number of offspring of a dominant, so that $f'(1) = 1$, the probability of r heterozygotes in the $(n + 1)$th generation is the coefficient of r in

$$[f(x)]^{\frac{2Nz_n(2N - z_n)}{4N^2 + kz_n^2}},$$

that of r recessives the same coefficient in

$$[f(x)]^{\frac{N(1+k)z_n^2}{4N^2 + kz_n^2}}.$$

Hence the probability of z_{n+1} in the next generation is the coefficient of $x^{z_{n+1}}$ in

$$[f(x)]^{\frac{2Nz_n(2N + kz_n)}{4N^2 + kz_n^2}},$$

or, approximately, if z_n be small compared with N, in

$$[f(x)]^{z_n\left(1 + \frac{kz_n}{2N}\right)}.$$

The corresponding expression for a dominant factor is

$$[f(x)]^{z_n(1+k)}.$$

Hence provided that z_n is small the probability of escaping extinction is much smaller than k. I have been unable to evaluate it exactly, but it seems from a comparison with the case of a dominant factor, that the value of z_n such that the factor is as likely to survive as to be extinguished, is of the order of $\left(\frac{N}{k}\right)^{\frac{1}{2}}$, *i.e.* generally $> N^{\frac{1}{2}}$. So if N is sufficiently large the probability of a single mutation leading to the establishment of a recessive factor is negligible.

When the population is wholly self-fertilized or inbred by brother-sister mating, on the other hand, a recessive factor has almost as good a chance of survival as a dominant. With partial self-fertilization or inbreeding it can be shown by methods similar to those of Part II (6) that an advantageous recessive factor has a finite chance of establishment after one appearance, however large be the population.

If mutation occurs with a finite frequency any advantageous or not too disadvantageous factor will certainly be established. Consider a random mating population in which, in each generation, a proportion p of the A genes mutate to a, a proportion q of the a genes to A, and the coefficient of selection is k. Let u_n be the gametic ratio of the nth generation. But for mutation we should have

$$u_{n+1} = \frac{u_n(u_n + 1)}{u_n + 1 - k};$$

allowing for mutation

$$u_{n+1} = \frac{(1-p)(u_n^2+u_n)+q(u_n+1-k)}{(1-q)(u_n+1-k)+p(u_n^2+u_n)}.$$

Hence $$\Delta u_n = \frac{ku_n^2 - pu_n(u_n+1)^2 + q(u_n+1)(u_n+1-k)}{u_n+1-k+pu_n(u_n+1)-q(u_n+1-k)} \quad \ldots(2{\cdot}0)$$

$$= \frac{ku_n}{u_n+1} - pu_n(u_n+1) + q(u_n+1) \quad \ldots\ldots\ldots\ldots\ldots(2{\cdot}1)$$

approximately, if p, q, and k are small, as is generally the case. It is clear that u_n must lie between $\frac{1-p}{p}$ and $\frac{q}{1-q}$, *i.e.* between $\frac{1}{p}$ and q approximately, and that when near these values it alters rapidly. But as p and q may be less than 10^{-6} these limits are very wide. The population is in equilibrium when

$$pu^3 + (2p-q)u^2 + (p-2q-k+kq)u - q + kq = 0.$$

There is always one real positive root since p and q are positive and less than unity. If k be positive there is only one such root, defining a stable condition towards which the population tends when dominants have the advantage. If k or q be large compared with p this root approximates to $\left(\frac{k}{p}\right)^{\frac{1}{2}}$ or $\frac{q}{p}$ as the case may be, *i.e.* recessives nearly disappear. If p be of the same order of magnitude as the larger of k and q, u has a moderate value and the population is dimorphic. If p be much larger than k or q, u is small and approximates to $\frac{q}{p}$, *i.e.* dominants are rare.

If k be negative all the roots are positive if they are real, provided $q > 2p$ and $-k(1-q) > 2q-p$. They are real if $\frac{\Delta}{3k}$, *i.e.*

$$4(p+q)^3 + [-27p^2 + 18p(p+q)(1-q) + (p+q)^2(1-q)^2]k + 4p(1-q)^3k^2$$

is positive, that is to say, when q is small, if

$$4pk^2 + (-8p^2 + 20pq + q^2)k + 4(p+q)^2$$

is positive. All these three conditions can rarely be fulfilled, but such cases may presumably occur. Thus if $p = {\cdot}000{,}001$, $q = {\cdot}0004$, $k = -{\cdot}008$; $u^3 - 398u^2 + 7197{\cdot}8u - 403{\cdot}2 = 0$. Therefore $u = {\cdot}057$, $18{\cdot}93$, or $379{\cdot}0$, giving $89{\cdot}5\ {}^\circ/_\circ$, $0{\cdot}252\ {}^\circ/_\circ$, or ${\cdot}000{,}693\ {}^\circ/_\circ$ of recessives. In such a case the middle root defines an unstable equilibrium, the other two equilibria being stable. Thus the above considered population would be stable with only about seven recessives per million, the small tendency of dominant genes to mutate to recessive being balanced by reverse mutation. But if a group containing more than one recessive gene in twenty were isolated from it, selection would be effective, and it would pass into a condition where only 10·5 °/₀ were dominants, this number being kept up by mutation.

Usually when k is negative there is only one real root. If p or $-k$ be large compared with q, it is small and approximates to $\frac{q}{p}$ or $\frac{-q}{k}$ as the case may be, so that dominants are rare. If q be of the same order of magnitude as the larger of p and $-k$, the root has a moderate value and the population is dimorphic. If q be larger than p or $-k$, u is large and approximates to $\frac{q}{p}$, so that recessives are few.

The rate of approach to equilibrium is given by

$$\frac{du_n}{dn} = \frac{ku_n}{u_n+1} - pu_n(u_n+1) + q(u_n+1) \text{ ......(2·2)},$$

provided that the constants are small. The exact expression for n in terms of u_n depends on the nature of the roots and the side from which an equilibrium is being approached, but it always contains logarithmic terms. Hence the numbers of the rarer type of the population in succeeding generations always lie between two geometric series until equilibrium is nearly reached. That is to say, the march of events is comparatively rapid.

In a self-fertilizing population we can similarly show that

$$\Delta u_n = ku_n - pu_n(u_n+1) + q(u_n+1) \text{ ....(2·3)}.$$

Only one equilibrium is possible, and the course of events can readily be calculated in any given case. Similarly for a sex-linked factor

$$\Delta u_n = \frac{ku_n(u_n+3)}{3(u_n+1)} - pu_n(u_n+1) + q(u_n+1) \text{......(2·4)}.$$

In this case if k be negative, three equilibria are sometimes found, and selection is more effective than in the autosomal case when recessives are rare.

To sum up, if selection acts against mutation, it is ineffective provided that the rate of mutation is greater than the coefficient of selection. Moreover, mutation is quite effective where selection is not, namely in causing an increase of recessives where these are rare. It is also more effective than selection in weeding out rare recessives provided that it is not balanced by back mutation of dominants. Mutation therefore determines the course of evolution as regards factors of negligible advantage or disadvantage to the species. It can only lead to results of importance when its frequency becomes large.

https://doi.org/10.1017/S0305004100015644 Published online by Cambridge University Press

Addendum. Equilibrium and selection in Sciara and similar animals.

In Part I of this series all the then known types of single-factor Mendelian inheritance were discussed. Since then Metz (7) has discovered a new type in Sciara which is here treated on the lines of Part I. Gametogenesis is normal in the female, but spermatozoa are formed from maternal chromatin only. Hence there are two types of heterozygous male, which may be symbolized by $A(a)$ and $(A)a$ according as the A is received from the mother or father. They yield A and a spermatozoa respectively, the other genotypes behaving normally.

In the absence of selection let eggs and spermatozoa be produced by the mth generation in the proportions $u_m A : 1a$ and $v_m A : 1a$, respectively. The next generation is therefore:

$$♀\ u_m v_m AA : (u_m + v_m)Aa : 1aa.$$

$$♂\ u_m v_m AA : u_m A(a) : v_m(A)a : 1aa.$$

Hence
$$\left.\begin{aligned} u_{m+1} &= \frac{2u_m v_m + u_m + v_m}{u_m + v_m + 2} \\ v_{m+1} &= u_m \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots(3{\cdot}0),$$

which is the same as equation (6·0) of Part I (5). Hence, as in the above equation, we find, if y_m be the proportion of recessives in the mth generation,

$$y_m = y_\infty - \left(\frac{-1}{2}\right)^m c^{\frac{1}{2}} y_\infty^{\frac{1}{2}} + \left(\frac{-1}{2}\right)^{2m-1} c \quad \ldots\ldots(3{\cdot}1),$$

where c is a constant depending on the initial conditions. Hence equilibrium is rapidly approached, the values in successive generations being alternately greater and less than the final value.

If selection occurs with a coefficient k in ♀s, l in ♂s, then

$$\left.\begin{aligned} u_{n+1} &= \frac{2u_n v_n + u_n + v_n}{u_n + v_n + 2 - 2k} \\ v_{n+1} &= \frac{u_n(v_n + 1)}{v_n + 1 - l} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(3{\cdot}2).$$

If the population is nearly in equilibrium apart from selection and k and l are small, so that u_n and v_n are nearly equal,

$$\Delta u_n = \frac{v_n - u_n}{2} + \frac{ku_n}{u_n + 1},$$

$$\Delta v_n = u_n - v_n + \frac{lu_n}{u_n + 1}, \text{ both approximately.}$$

https://doi.org/10.1017/S0305004100015644 Published online by Cambridge University Press

Hence
$$\Delta u_n = \frac{2k+l}{3}\frac{u_n}{u_n+1},$$

and
$$\frac{2k+l}{3}n = u_n - u_0 + \log_e\left(\frac{u_n}{u_0}\right) \quad \text{.........(3·3)},$$

approximately. Selection therefore occurs much as with a normally inherited autosomal factor.

REFERENCES.

1. Harrison and Garrett. *Proc. Roy. Soc.* B, 99, p. 241, 1926.
2. Gager and Blakeslee. *Proc. Nat. Ac. Sci.* 13, p. 75, 1927.
3. Fisher. *Proc. Roy. Soc. Edin.* 42, p. 321, 1922.
4. Koenigs. *Darb. Bull.* (2) 7, p. 340, 1883.
5. Haldane. *Trans. Camb. Phil. Soc.* 23, p. 19, 1924.
6. Haldane. *Biol. Proc. Camb. Phil. Soc.* 1, p. 158, 1924.
7. Metz. *Proc. Nat. Ac. Sci.* 12, p. 690, 1926.

https://doi.org/10.1017/S0305004100015644 Published online by Cambridge University Press