

ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN A POPULATION¹

MOTOO KIMURA²

University of Wisconsin, Madison, Wisconsin

Received January 29, 1962

THE success or failure of a mutant gene in a population is dependent not only on selection but also on chance. This fact was first treated quantitatively by FISHER (1922) who later (1930) worked out the probability of ultimate survival of a mutant gene for the case of genic selection (i.e. no dominance). Equivalent results have been obtained by HALDANE (1927) and WRIGHT (1931). Also the probability was estimated for a recessive mutant gene by HALDANE (1927) and WRIGHT (1942).

The present author (KIMURA 1957) extended these results to include any level of dominance. The probability of eventual fixation, $u(p)$, was expressed in terms of the initial frequency, p , the selection coefficients, and the effective population number. This function was used by ROBERTSON (1960) in his theory of selection limits in plant and animal breeding.

The purpose of this note is to present a more general formula for $u(p)$ which encompasses random fluctuations in selection intensity as well as random drift because of small population number. It will also be used to solve a question relating to "quasi-fixation" posed by the author in 1955.

Derivation of the formula: Consider a population in which the frequency of the allele A is p ($0 \leq p \leq 1$). We assume that the population is sufficiently large and the change in p per generation sufficiently small that the change in p through time may be satisfactorily approximated by a continuous stochastic process.

Let $u(p, t)$ be the probability that allele A will be fixed (i.e. its frequency becomes 1) during a time interval t (conveniently measured in generations), given that p is the initial frequency of A . From the nature of Mendelian mechanism of inheritance, the process of change in p is Markovian, i.e. the frequency distribution of A in the next generation is determined by the frequency of A in the present generation and is not dependent on the way in which the present gene frequency was attained. Thus we have

$$u(p, t + \delta t) = \int f(p, p + \delta p; \delta t) u(p + \delta p, t) d(\delta p)$$

where $f(p, p + \delta p; \delta t)$ is the probability density of the change from p to $p + \delta p$ during short time interval δt and the integration is over all possible values of δp .

¹ This paper was prepared for publication under Contract No. DA-11-022-ORD-2059, Mathematics Research Center, United States Army, University of Wisconsin. This also constitutes paper number 391 of the National Institute of Genetics, Mishima-shi, Japan.

² On leave from the National Institute of Genetics, Mishima-shi, Japan.

Here we assume that the probability density depends on p and δt but not on t (i.e. the process is temporary homogeneous).

Expanding $u(p+\delta p, t)$ in the right side of the above equation in terms of δp and putting

$$\lim_{\delta t \rightarrow 0} \frac{1}{\delta t} \int (\delta p) f(p, p+\delta p; \delta t) d(\delta p) = M$$

$$\lim_{\delta t \rightarrow 0} \frac{1}{\delta t} \int (\delta p)^2 f(p, p+\delta p; \delta t) d(\delta p) = V$$

but neglecting higher order terms involving $(\delta p)^3$, $(\delta p)^4$, etc., we have

$$(1) \quad \frac{\partial u(p, t)}{\partial t} = \frac{V}{2} \frac{\partial^2 u(p, t)}{\partial p^2} + M \frac{\partial u(p, t)}{\partial p}$$

Since time is conveniently measured in generations, we substitute $M_{\delta p}$ and $V_{\delta p}$ for M and V , where $M_{\delta p}$ and $V_{\delta p}$ are the mean and the variance of the change of p per generation. The probability $u(p, t)$ may then be obtained by solving the above partial differential equation (known as the Kolmogorov backward equation) with boundary conditions:

$$u(0, t) = 0, \quad u(1, t) = 1.$$

We are especially interested in the present paper in the ultimate probability of fixation defined by

$$u(p) = \lim_{t \rightarrow \infty} u(p, t),$$

for which $\partial u / \partial t = 0$ and which therefore satisfies the equation

$$(2) \quad \frac{V_{\delta p}}{2} \frac{d^2 u(p)}{dp^2} + M_{\delta p} \frac{du(p)}{dp} = 0$$

with boundary conditions:

$$u(0) = 0, \quad u(1) = 1.$$

The equation may readily be solved to give

$$(3) \quad u(p) = \frac{\int_0^p G(x) dx}{\int_0^1 G(x) dx}$$

where

$$(4) \quad G(x) = e^{-\int \frac{2M_{\delta x}}{V_{\delta x}} dx}$$

in which $M_{\delta x}$ and $V_{\delta x}$ are the mean and the variance of the change in gene frequency, x , per generation. Formula (3) gives the probability of fixation of allele A when its initial frequency is p . It has a pleasing simplicity and generality

comparable to WRIGHT's (1938, 1949) well-known formula for the frequency distribution of genes at equilibrium. In my notation, this is

$$(5) \quad \varphi(x) = \frac{C}{V_{\delta x} G(x)}$$

where $\varphi(x)$ is the probability density of the gene frequency x and C is a constant chosen such that

$$\int_0^1 \varphi(x) dx = 1$$

From the derivation it may be seen that formula (3) holds not only for sexually reproducing diploid organisms, but also for haploid or polyploid organisms and asexually reproducing plants.

Finally, the chance of fixation of an individual mutant gene in a randomly mating diploid population is given by

$$(6) \quad u = u\left(\frac{1}{2N}\right),$$

where N is the number of reproducing individuals in the population.

Some applications: The simplest case is genic selection in which A has a constant selective advantage s over its alleles in a randomly mating population of size N . If the frequency of A is x , then the mean and the variance in the rate of change in x per generation are

$$(7) \quad \begin{cases} M_{\delta x} = sx(1-x) \\ V_{\delta x} = x(1-x)/(2N) \end{cases}$$

so that $2M_{\delta x}/V_{\delta x} = 4Ns$, $G(x) = e^{-4Nsx}$ and we obtain from (3)

$$(8) \quad u(p) = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}}$$

For $|2Ns| < \pi$, the right side of the above equation may be expanded in terms of $4Ns$ as follows:

$$(9) \quad \begin{aligned} u(p) &= \sum_{i=1}^{\infty} \frac{\phi_i(p)(-1)^{i-1}}{i!} (4Ns)^{i-1} \\ &= p + 2Nsp(1-p) + \frac{(2Ns)^2}{3} p(p-1)(2p-1) + \dots \end{aligned}$$

where the $\phi_i(p)$'s are Bernoulli polynomials. Thus for a small value of $2Ns$, $u(p) - p$ is $2N$ times $sp(1-p)$. In other words, the total advance is $2N$ times the change in the first generation as pointed out by ROBERTSON (1960).

The probability of fixation of an individual mutant gene is obtained from (8) by putting $p = 1/(2N)$.

$$(10) \quad u = (1 - e^{-2s}) / (1 - e^{-4Ns})$$

If we assume that $|s|$ is small, we obtain

$$(11) \quad u = (2s) / (1 - e^{-4Ns})$$

as a good approximation. This agrees with the result obtained by FISHER (1930) and WRIGHT (1931). This formula is good even for negative s , though u for such a case is very small unless $|Ns|$ is not large. For a positive s and very large N we obtain the known result that the probability of ultimate survival of an advantageous mutant gene is approximately twice the selection coefficient (HALDANE 1927). On the other hand, if we let $s \rightarrow 0$ in (10), we obtain $u = 1/2N$, the result known for a neutral gene.

Next, I shall consider a more general case of zygotic selection under random mating. Let s and sh be, respectively, the selective advantage of mutant homozygote and the heterozygote, then

$$(12) \quad \begin{cases} M_{\delta x} = sx(1-x)[h+(1-2h)x] \\ V_{\delta x} = x(1-x)/(2N) \end{cases}$$

and therefore

$$G(x) = e^{-2cDx(1-x)-2cx}$$

where $c = Ns$ and $D = 2h - 1$. Thus we obtain

$$(13) \quad u(p) = \int_0^p e^{-2cDx(1-x)-2cx} dx / \int_0^1 e^{-2cDx(1-x)-2cx} dx$$

which is the formula given by the present author in a previous paper (KIMURA 1957).

For a completely recessive gene $h = 0$ or $D = -1$ and we have

$$(14) \quad u(p) = \int_0^p e^{-2cx^2} dx / \int_0^1 e^{-2cx^2} dx$$

If s is positive and small but Ns is large, the above formula leads approximately to

$$(15) \quad u = \sqrt{\frac{2s}{\pi N}} \approx 1.128 \sqrt{\frac{s}{2N}}$$

for $p = \frac{1}{2N}$, giving the probability of fixation of individual mutant genes which is advantageous but completely recessive. It is interesting to note that the more exact value given here in (15) lies between $\sqrt{s/N}$, the value obtained by HALDANE (1927) by treating this as a branching process and $\sqrt{s/(2N)}$, obtained by WRIGHT (1942) with his method of integral equations. WRIGHT's numerical approximation, $1.1\sqrt{s/2N}$, is very close to (15).

I shall now investigate the effect of random fluctuation of selection intensity on the fixation of mutant genes.

The simplest situation is again genic selection in which the mutant gene A has a selective advantage s over its alleles, but s is now a random variable with mean \bar{s} and variance V_s . Thus (7) may be replaced by

$$(16) \quad \begin{cases} M_{\delta x} = \bar{s}x(1-x) \\ V_{\delta x} = V_s x^2(1-x)^2 + x(1-x)/(2N) \end{cases}$$

to give

$$(17) \quad G(x) = \left(\frac{\lambda_1 - x}{x - \lambda_2} \right)^{k/\sqrt{1+4r}}$$

where

$$k = 2\bar{s} / V_s$$

and

$$\begin{cases} \lambda_1 = (1 + \sqrt{1+4r}) / 2 \\ \lambda_2 = (1 - \sqrt{1+4r}) / 2 \end{cases}$$

where $r = 1/(2NV_s)$.

There are several interesting special cases. First, if $\bar{s} = 0$, then $G(x) = 1$ and we have

$$u(p) = p$$

a result which might be expected, since the gene is neutral on the average. Secondly, if $\bar{s} \neq 0$, then $k \neq 0$ and I shall consider the following three cases: (1) $r \rightarrow \infty$, (2) $r = 2$ and (3) $r \rightarrow 0$. In all these cases, we assume that $0 < V_s \ll 1$ and therefore $r \gg 1/(2N)$. We will also assume that \bar{s} is positive but small.

Case 1. Consider the situation in which r is very large or equivalently $2NV_s$ is very small. Then, approximately

$$2M_{\delta x} / V_{\delta x} = 4N\bar{s}[1 - 2NV_s x(1 - x)].$$

Assuming further that $N\bar{s}$ is large but $8N^2\bar{s}V_s$ is very small, we have

$$(18) \quad u = 2\bar{s} - V_s,$$

namely the probability of fixation of a favorable gene is reduced by V_s due to random fluctuation in s .

Case 2. Here we consider a special case in which $r = 2$, or equivalently $V_s = 1/(4N)$. For this case

$$G(x) = (2 - x)^{k/3}(1 + x)^{-k/3}$$

If k is very small, u is approximately $1/(2N)$, while if k is very large it is nearly $2\bar{s}$. For an intermediate value of k such as $k = 3$, we obtain

$$u = [N(3 \log 2 - 1)]^{-1}$$

which is roughly $0.9/N$ or $2.4\bar{s}$.

Case 3. We now come to a more interesting case in which r is very small, namely $V_s \gg 1/(2N)$. In this case, as r approaches 0, λ_1 and λ_2 approach 1 and 0, respectively. Here the situation is rather delicate and we restrict our consideration to the behavior of $u(p)$ for an intermediate value of p . For such a value of p which is neither very close to 0 nor 1, the effect of random sampling of gametes in reproduction may be ignored for a large population ($N \rightarrow \infty$), so that we can write

$$G(x) = \left(\frac{1-x}{x}\right)^k$$

keeping in mind that fixation here means strictly "quasifixation" (cf. KIMURA

1954), though in actual cases the prefix "quasi" may be unnecessary, because random sampling of gametes is always at work in the subterminal class. Then we have

$$(19) \quad u(p) = \int_0^p \left(\frac{1-x}{x}\right)^k dx / \int_0^1 \left(\frac{1-x}{x}\right)^k dx$$

It may readily be seen that if $k \geq 1$, both denominator and numerator of the above formula are divergent, which may be interpreted as showing that $u(p) = 1$ for all $p > 0$. This agrees with the obvious fact that once the advantageous mutant gene has reached the stage where it is represented by sufficiently many individuals, it will almost certainly be led to fixation by natural selection. On the other hand, if $0 \leq k < 1$, the integral in the denominator converges and we get

$$\int_0^1 \left(\frac{1-x}{x}\right)^k dx = \frac{\pi K}{\sin \pi K}$$

This shows that if $0 \leq 2\bar{s} < V_s$, then $u(p) < 1$, namely there is a finite chance that gene A will be lost from the population by random fluctuation in selection intensity (plus random sampling of gametes in the neighborhood of $x = 0$) even if the gene is advantageous on the average. This gives an affirmative answer to the question posed by the present author in 1955:

"When \bar{s} and V_s are of the same order of magnitude, or $\bar{s} (> 0)$ is much smaller than V_s , is quasi-fixation still possible on the side of the disadvantageous class?"

For a small value of p (still assuming that $N \rightarrow \infty$ and $1/N \ll p \ll 1$), we have approximately

$$u(p) = \frac{\sin(\pi k) p^{1-k}}{\pi k(1-k)}$$

As an example, if \bar{s} is 1/10 of V_s and if the initial frequency is one percent, the chance of fixation is about three percent.

Application of a similar method for the case of a completely recessive advantageous gene leads to a rather unexpected result that there is no finite chance of its loss if $p \gg 1/N$, where $N \rightarrow \infty$.

SUMMARY

The probability of ultimate fixation of a gene in a population is treated as a continuous stochastic process and the solution is given as a function of the initial frequency and the mean and variance of the gene frequency change per generation. The formula is given by equation (3).

The formula is shown to include previous results as special cases and is applied to solve problems where there is random fluctuation in selection intensity. It is also used to show some circumstances under which, even in a large population, an advantageous gene may be lost because of fluctuating selective values.

ACKNOWLEDGMENTS

The author would like to express his thanks to DR. J. F. CROW who kindly reviewed the draft and made valuable suggestions for improving the manuscript.

LITERATURE CITED

- FISHER, R. A., 1922 On the dominance ratio. *Proc. Roy. Soc. Edinburgh*. **42**: 321-341.
1930 The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh*. **50**: 204-219.
- HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection. Part V: Selection and mutation. *Proc. Cambridge Phil. Soc.* **23**: 838-844.
- KIMURA, M., 1954 Process leading to quasifixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics* **39**: 280-295.
1955 Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia Quant. Biol.* **20**: 33-53.
1957 Some problems of stochastic processes in genetics. *Ann. Math. Stat.* **28**: 882-901.
- ROBERTSON, A., 1960 A theory of limits in artificial selection. *Proc. Roy. Soc. B.* **153**: 234-249.
- WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97-159.
1938 The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. U.S.* **24**: 253-259.
1942 Statistical genetics and evolution. *Bull. Am. Math. Soc.* **48**: 223-246.
1949 Adaptation and selection. pp. 365-389. *Genetics, Paleontology, and Evolution*. Edited by G. L. JEPSON, G. G. SIMPSON, and E. MAYR. Princeton Univ. Press, Princeton, N.J.