Genet. Res., Camb. (1973), **21**, *pp.* 115–120
Printed in Great Britain

Natural selection with varying selection coefficients – a haploid model

By J. H. GILLESPIE
Department of Biology, University of Pennsylvania, Philadelphia, Pa. 19104

(*Received* 5 *June* 1972)

SUMMARY

In this paper an exact treatment is given for the stochastic behaviour of the frequency of haploid genotypes in an infinite population when the absolute fitnesses of the two genotypes vary at random over generations. The main qualitative result from this treatment is that natural selection will favour that allele with the largest geometric mean fitness. A diffusion equation is derived whose solution is identical to the exact solution. The drift coefficient for this equation is of the form $-\mu p(1-p)+\sigma^2(\frac{1}{2}-p)p(1-p)$. This differs from the drift coefficient used in previous treatments of this problem and reduces the rate of quasi-fixation. Various waiting time problems are solved using this diffusion equation.

1. INTRODUCTION

The behaviour of haploid alleles in uncorrelated environments has been investigated by Kimura (1954) and Dempster (1955), and recently reviewed by Crow & Kimura (1970). The mathematical techniques employed in the first two papers are different, though the underlying model is the same (Crow & Kimura, 1970). As far as I have been able to discover, no one has carried Dempster's approach to the point of displaying the probability density function for the process. When this is done, it differs in rather important ways from the density obtained by Kimura using a diffusion approximation. As will be shown in this paper, a better diffusion approximation is possible, and its simple relationship to the Brownian motion process allows various waiting-time problems to be readily solved. Before arriving at this approximation I will redescribe Dempster's model in a way which will emphasize its biologically important properties. In particular it will be shown that the mean fitness of a population can decrease through the action of natural selection in a stochastic environment.

2. THE STOCHASTIC MODEL

Consider a discrete-generation haploid population of two genotypes, A_1 and A_2, whose absolute fitnesses in the nth generation are $1+U_n$ and $1+V_n$, respectively. If the frequency of A_1 in the nth generation is X_n, the difference equation describing the trajectory of X_n is given by

$$\Delta X = \frac{X_n(1-X_n)(U_n-V_n)}{1+X_nU_n+(1-X_n)V_n}, \tag{1}$$

whose solution is

$$X_n = \left\{1 + \frac{1-x_0}{x_0} \exp\left[\sum_{i=0}^{n-1} \ln\left(\frac{1+V_i}{1+U_i}\right)\right]\right\}^{-1}.$$

If U_n and V_n are random variables, not necessarily independent, and if the density of the sum

$$Y_n = \sum_{i=0}^{n-1} \ln\left(\frac{1+V_i}{1+U_i}\right)$$

is $f_n(Y_n)$, then the density of X_n is

$$\frac{f_n\left[\ln\left(\frac{x_0}{1-x_0}\frac{1-x_n}{x_n}\right)\right]}{x_n(1-x_n)}. \tag{2}$$

In the special case where the random vector (U_n, V_n) is independent of and identically distributed with (U_{n+j}, V_{n+j}), $f_n(Y_n)$ approaches a normal distribution with moments

$$\left.\begin{aligned} EY_n &= nE\ln\left(\frac{1+V_i}{1+U_i}\right) = n\mu, \\ \operatorname{var} Y_n &= n \operatorname{var} \ln\left(\frac{1+V_i}{1+U_i}\right) = n\sigma^2. \end{aligned}\right\} \tag{3}$$

The central limit theorem implies that the density of X_n becomes, asymptotically,

$$\phi_n(x) = \frac{\exp\left\{\frac{-\frac{1}{2}\left[\ln\left(\frac{x_0}{1-x_0}\cdot\frac{1-x}{x}\right) - n\mu\right]^2}{n\sigma^2}\right\}}{\sqrt{(2\pi n\sigma^2)}x(1-x)}. \tag{4}$$

This density is exact, for all n, if $1+U_n$ and $1+V_n$ are lognormally distributed with moments as above. Otherwise, the rapidity with which the approximation (4) approaches the exact solution (2) depends on the density of (U_n, V_n). Note that this model has the quasi-fixation property:

$$\lim_{n\to\infty} Pr\{X_n \in (\delta, 1-\delta)\} = 0.$$

From a biological point of view, the most important property of this model rests with the fate of alleles as a function of the first- and second-order moments of (U_n, V_n). This information may be obtained by examining the probability mass of $\phi_n(x)$ in the interval $(0, \alpha)$:

$$p_n(\alpha) = \int_0^\alpha \phi_n(x)\,dx.$$

The change of variable

$$y = \ln(1-x)/x$$

shows that $p_n(\alpha)$ is equal to the integral of the standardized normal over the interval

$$\left(\frac{-n\mu + \ln\frac{1-x_0}{x_0}\frac{1-\alpha}{\alpha}}{\sqrt{(n\sigma^2)}}, \infty\right).$$

https://doi.org/10.1017/S001667230001329X Published online by Cambridge University Press

As $n \to \infty$ the asymptotic value of this interval is determined solely by the sign of μ. Using (3) we can conclude that

$$p_n(\alpha) \to 0 \quad \text{iff} \quad \mu < 0,$$
$$p_n(\alpha) \to 1 \quad \text{iff} \quad \mu > 0,$$
$$p_n(\alpha) \to \tfrac{1}{2} \quad \text{iff} \quad \mu = 0.$$

Since the geometric mean of a random variable is just the expectation of the logarithm raised to the power e, the above implies that the allele with the largest geometric mean fitness is favoured by natural selection. In the case where both alleles have equal geometric mean fitnesses, both alleles have (asymptotically) equal probabilities of being found in the population.

If the stochastic effects are small, μ can be approximated by

$$\mu \simeq \overline{U} - \overline{V} - \tfrac{1}{2}(EV^2 - EU^2)$$

which illustrates the role of the second-order moments of fitness in determining the fate of an allele. Note that the arithmetic mean tends to overestimate the true effect of fitness on gene frequency changes. This admits the possibility of selection favouring an allele which actually lowers the mean fitness of the population. For example, if $\overline{U} < \overline{V}$, but (5) holds, a population consisting almost entirely of allele A_2 with mean fitness $1 + \overline{V}$ will be replaced by a population of A_1 individuals with mean fitness $1 + \overline{U}$, resulting in a drop in mean fitness of $\overline{U} - \overline{V}$.

The covariance between U_n and V_n plays no role in condition (5), but does affect the rate of quasi-fixation, this being defined as the rate of increase in the variance of

$$\ln \frac{1 - X_n}{X_n}$$

(Gillespie, 1972). For the model under consideration the rate of quasi-fixation is obviously

$$\sigma^2 \simeq \sigma_U^2 + \sigma_V^2 - 2\sigma_{UV}.$$

Of particular interest here is the role of σ_{UV}. If the covariance of the fitnesses of the two alleles is negative, the quasi-fixation process can proceed very rapidly. It is minimal, for fixed σ_U^2 and σ_V^2, when $\sigma_{UV} = \sigma_U \sigma_V$. The bounds on σ^2 are

$$(\sigma_U - \sigma_V)^2 \leqslant \sigma^2 \leqslant (\sigma_U + \sigma_V)^2$$

for fixed σ_U^2, σ_V^2.

3. THE DIFFUSION APPROXIMATION

In order to arrive at a diffusion approximation for the process defined in the preceding section we can begin by noting that Y_n is a simple random walk and thus can be approximated by a Brownian motion process with drift and diffusion coefficients of μ and σ^2. The resulting density for the random function $X(t)$ is

$$\phi(x,t) = \frac{\exp\left\{\dfrac{-\frac{1}{2}\left[\ln\dfrac{1-x}{x}\dfrac{x_0}{1-x_0}-\mu t\right]^2}{t\sigma^2}\right\}}{x(1-x)\sqrt{(2\pi\sigma^2 t)}}. \tag{5}$$

To discover the diffusion equation satisfied by (5) we need only examine

$$M(x) = \lim_{\tau\to 0}\frac{E[X(t+\tau)|x(t)]-x(t)}{\tau},$$

$$V(x) = \lim_{\tau\to 0}\frac{\mathrm{var}\,[X(t+\tau)|x(t)]}{\tau},$$

which are the drift and diffusion coefficients of the process $X(t)$. To evaluate $M(x)$, use

$$E(X(t+\tau)) = \frac{1}{\sqrt{(2\pi\sigma^2\tau)}}\int_{-\infty}^{\infty}\left[\frac{1}{1+\dfrac{1-x(t)}{x(t)}e^{y}}\right]\exp\left\{-\frac{1}{2}\frac{(y-\mu\tau)^2}{\sigma^2\tau}\right\}dy.$$

The bracketed expression under the integral may be approximated near the origin by its Taylor series:

$$x(t)-x(t)\,(1-x(t))y+x(t)\,(1-x(t)\,(\tfrac{1}{2}-x(t))y^2.$$

Using this approximation it is easily verified that

$$M(x) = x(1-x)\,[-\mu+\sigma^2(\tfrac{1}{2}-x)].$$

Similarly $$V(x) = \sigma^2x^2(1-x)^2.$$

That (5) does, in fact, satisfy

$$\frac{\partial\phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}\{x(1-x)\,[-\mu+\sigma^2(\tfrac{1}{2}-x)]\,\phi\}+\frac{1}{2}\frac{\partial^2}{\partial x^2}\{\sigma^2x^2(1-x)^2\phi\}$$

may be shown by substitution. Warren Ewens (pers. com.) has shown me a derivation of $M(x)$ directly from (1). This will be published elsewhere.

In his general treatment of this problem that includes the possibility of a non-zero drift coefficient Kimura (1955) uses a drift coefficient or the form

$$\bar{s}x(1-x),$$

where $\bar{s}$ is interpreted by Kimura as the difference in the Arithmetic mean fitnesses of the two genotypes. This coefficient can be compared directly to the one derived in

this paper by noting that the Brownian motion approximation to the discrete random walk is obtained by shrinking the mean and variance of U_n and V_n to zero at the same rate, while assuming all higher moments shrink faster. In terms of these first two moments the drift coefficient is exactly

$$x(1-x)\left[\overline{U}-\overline{V}+\tfrac{1}{2}(\sigma_V^2-\sigma_U^2)+\sigma^2(\tfrac{1}{2}-x)\right].$$

This suggests that Kimura's coefficient should be viewed as a first-order approximation which can be considerably improved by the addition of the term

$$x(1-x)\left[\tfrac{1}{2}(\sigma_V^2-\sigma_U^2)+\sigma^2(\tfrac{1}{2}-x)\right].$$

Kimura's symmetric case (his $M(x)=0$) considered in his 1954 paper obviously applies to the situation where the two genotypes have equal geometric mean fitnesses so the drift coefficient should be

$$M(x)=\sigma^2 x(1-x)\,(\tfrac{1}{2}-x).$$

This coefficient will cause the quasi-fixation process to proceed considerably slower than Kimura's description of the process would indicate. This can be seen directly by comparing the density (5) with $\mu=0$ to the density in Kimura's (1954) paper.

The use of the Brownian approximation of Y_n points out the simple relationship between this process and the genetic process. In fact

$$Y(t)=\ln\left(\frac{1-x(t)}{x(t)}\,\frac{x(0)}{1-x(0)}\right).$$

Many of the genetic properties may be arrived at as a consequence of this transformation. Consider, for example, the various waiting-time problems associated with $X(t)$. The distribution of the waiting-time for an allele with an initial frequency $x(0)$ to leave the interval (a,b) is the same as the time required for the Brownian motion process with initial value zero to leave the interval

$$\left(\ln\frac{x(0)}{1-x(0)}\,\frac{b}{1-b},\quad \ln\frac{x(0)}{1-x(0)}\,\frac{a}{1-a}\right).$$

This distribution is well known and we shall simply note that the mean value of the distribution is

$$ET=-\frac{1}{\mu}\left\{\ln\frac{x(0)}{1-x(0)}\,\frac{a}{1-a}+\ln\frac{b}{1-b}\,\frac{1-a}{a}\left[\frac{\left(\dfrac{x(0)}{1-x(0)}\,\dfrac{a}{1-a}\right)^{2\mu/\sigma^2}-1}{\left(\dfrac{a(1-b)}{b(1-a)}\right)^{2\mu/\sigma^2}-1}\right]\right\}.$$

In particular, when allele A_1 has a mean selective advantage over A_2 ($\mu<0$), the moments of the time to reach a final frequency of x are:

$$ET=\frac{\ln\left(\dfrac{x(0)}{1-x(0)}\,\dfrac{1-x}{x}\right)}{-\mu}$$

$$\operatorname{var}T=-\frac{\sigma^2}{2\mu^3}\ln\left(\frac{x(0)}{1-x(0)}\,\frac{1-x}{x}\right).$$

https://doi.org/10.1017/S001667230001329X Published online by Cambridge University Press

Remarkably, the mean time is the same as the associated deterministic process ($\sigma^2 = 0$).

In an analogous fashion we can examine the probability that $X(t)$ leaves the interval (a, b) for the first time on the left side. From the Brownian motion theory this is given by

$$p = \frac{\left(\frac{x(0)}{1-x(0)}\frac{1-b}{b}\right)^{2\mu/\sigma^2} - 1}{\left(\frac{a}{1-a}\frac{1-b}{b}\right)^{2\mu/\sigma^2} - 1}.$$

One use which can be made of this involves the probability of A_1, when advantages, reaching a frequency of ϵ (quasi-lost) before attaining the frequency $1-\epsilon$ (quasi-fixed). If ϵ and $x(0)$ are small, and if $\epsilon < X(0)$, then

$$p \simeq (x(0)/\epsilon)^{2\mu/\sigma^2}.$$

REFERENCES

CROW, J. F. & KIMURA, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

DEMPSTER, E. (1955). Maintenance of genetic heterogeneity. *Cold Spring Harbor Symposia on Quantitative Biology* 20: 25–32.

GILLESPIE, J. H. (1972). The effect of stochastic environments on allele frequencies in natural populations. *Journal of Theoretical and Population Biology* (in the Press).

KIMURA, M. (1954). Processes leading to quasi-fixation of genes in natural populations due to random fluctuations in selection intensities. *Genetics* **39**, 280–295.

KIMURA, M. (1955). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* **20**, 33–53.