THE MAINTENANCE OF ALLELES BY MUTATION—MONTE CARLO RESULTS FOR NORMAL AND SELF-STERILITY POPULATIONS

W. J. EWENS and P. M. EWENS
Australian National University and C.S.I.R.O., Canberra

Received 15.ix.65

1. INTRODUCTION

A POPULATION of fixed size is considered in which mutation at a fixed rate occurs, each new mutant being regarded as an entirely new type. The number of different alleles occurring in any generation is a random variable, and the problem is to determine the mean number of different alleles when the process is in " equilibrium ". An explicit formula is available only in the simplest cases, and the aim of the present paper is to find an approximation for this number by a Monte Carlo experiment, in which the behaviour of the population is simulated on a high-speed computer. Two cases are considered; firstly that of selectively neutral " normal " alleles, and secondly that of a self-sterility population exhibiting a breeding structure like that of *Oenothera organensis*. The former case is of interest because it is the only one for which an explicit formula has been obtained, while the latter is of interest because of the mathematical discussions which have been made to derive an expression for the mutation rate necessary to maintain the large number of alleles observed in this species.

2. MATHEMATICAL THEORY

If we consider a diploid population of size N, then with a mutation rate u to entirely new alleles we expect $2Nu$ new alleles on the average per generation. At equilibrium these new alleles will be balanced by a similar number of " old " alleles being lost from the population by drift and/or mutation. If, in equilibrium, there are $\bar{n}$ different alleles, on the average, in each generation, and if the mean number of generations for which any allele exists in the population is $\bar{t}$, then the relation

$$2Nu = \bar{n}/\bar{t} \qquad \ldots\ldots(1)$$

will express the required balance between new alleles being formed and " old " alleles being lost. Thus once an expression for $\bar{t}$ can be found, the value of $\bar{n}$ follows immediately.

It may be the case that $\bar{t}$ is not a well defined quantity, in the sense that no absolute value of $\bar{t}$ may exist. For example, if selective differences are allowed in the population, then the mean time that any newly formed allele will exist will depend on its selective value, as well as the selective values of alleles currently in the population. This

will effectively be the case for self-sterility populations, and the analysis of such a population must therefore incorporate some approximations and assumptions in the hope of obtaining a reasonably exact result. For selectively neutral populations, on the other hand, once a model has been specified the value of $\bar{t}$ is well defined and applies irrespective of the composition of the population when the new allele appears. In order to fix ideas we now consider briefly this simple case.

The model we use is a particular case of that introduced by Wright (1931). If at any time the number of genes of a particular allele is i, then in the next generation we expect $i(1-u)$ genes of this allele, the decrease $-iu$ being due to mutation. The model is then that the probability p_{ij} that the number of genes of this allele changes from i to j in successive generations is given by the expression

$$p_{ij} = \binom{2N}{j}\left(\frac{i-iu}{2N}\right)^{j}\left(\frac{2N-i+iu}{2N}\right)^{2N-j}. \qquad \ldots\ldots(2)$$

We note that the expression for p_{ij}, apart from the constants u and N, involves only i and j, so that the variable under consideration is Markovian. Since the initial value is necessarily unity, there will exist a definite value $\bar{t}$ for the meantime for the number of such alleles to reach zero, an event which will happen eventually. Further, if we define $\bar{t}_m$ as the mean number of generations for which the number of genes of the allele in question assume the value m, then $\bar{t} = \bar{t}_1+\bar{t}_2+\ldots+\bar{t}_{2N}$ and equation (1) takes the form

$$\bar{n} = 2Nu[\bar{t}_1+\bar{t}_2+\ldots+\bar{t}_{2N}]. \qquad \ldots\ldots(3)$$

The values $(\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_{2N})$ have been called (Ewens (1964*a*) (1965)) the pseudo-transient function of the process under consideration. The exact value of $\bar{t}_i$ is at present unknown for the model under consideration, and the best that can be done is to approximate (3) by the corresponding diffusion expression

$$\bar{n} = 4Nu\int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{4Nu-1}dx, \qquad \ldots\ldots(4)$$

which is valid for all practical purposes when u is at most of order N^{-1}. The expression (4) was first given in Ewens (1964*b*).

For the self-sterility population the analysis is not nearly so easy. The expression for p_{ij} will strictly depend on the frequencies of all genotypes present in the population, and if there are (say) 45 different alleles there will be 990 different possible genotypes. A strict analysis would therefore require joint examination of all 990 variables, which is impossible, and the real problem is to find suitable approximations so that a univariate analysis can be made. Further, if a diffusion approximation such as (4) is to be used, all that is necessary is to find suitable approximations for the mean change $\overline{\Delta x}$ and the variance $\sigma^2_{\Delta x}$ of the change of the frequency x of the allele in question in successive generations.

There has been some discussion (Fisher, (1958), Wright (1960), (1964)) of suitable approximative formulae for Δx. By far the most work on this problem has been carried out by Wright, whose formulae are without doubt extremely satisfactory (see for example, Wright (1960), pp. 66-68). So far as an expression for $\sigma^2_{\Delta x}$ is concerned, the value originally used by Wright (Wright, 1939) was $x(1-x)/2N$. This was altered to $x(1-2x)/2N$ (Wright, 1960) following Fisher (1958), and rechanged to $x(1-x)/2N$ (Wright, 1964). The incorrectness of Fisher's formula was pointed out by Ewens (1964*c*), but so far as the subsequent analysis is concerned, the frequency x is generally sufficiently small to make either expression for $\sigma^2_{\Delta x}$ a reasonably close approximation to the very complicated exact value.

We shall discuss the analysis of Wright (1939, 1960, 1964) and the criticisms of it in more detail in a later section. For the moment we note that the expression for the equilibrium number of alleles maintained will be of the form (3), where the t_i constitute the pseudo-transient function of the process, and that the diffusion approximation to this will be of the form

$$\bar{n} = 2Nu \int_{(2N)^{-1}}^{1} t(x)dx \qquad \ldots\ldots(5)$$

where $t(x)$ is the diffusion approximation to the psuedo-transient function.

3. NUMERICAL RESULTS

It was noted in the previous section that an exact mathematical discussion of the behaviour of the self-sterility population is extremely difficult, and for this reason it was decided to simulate the evolutionary progress of such a population with a high-speed computer. At the same time a programme was run for the selectively neutral case to test the adequacy of the formula (4). Further, for both populations, besides the random mating case, a simulation was made for which a high degree of geographical inbreeding was allowed.

The details are as follows. The population size was 500, allowing 1,000 genes at the locus under consideration. The mutation rate was 10^{-3}, the mutation being deterministic in that one new mutant was formed in every generation (this should cause negligible difference to the stochastic mutation case). This mutation rate is, of course, very high, but more realistic values would give no better values for testing the validity of the formulae under consideration and would require larger populations and longer computer times than could be handled. The process was stochastic in that the genes making up any generation were selected at random (using a random number generator) from those of the previous generation according to the model (2) (or its counterpart in the sterility case). So far as the geographical inbreeding case was concerned, the population was divided up into 25 sub-populations with 20 individuals in each. With probability 0·9 an individual was mated

with an individual from the same sub-population, and with probability 0·1 from the population at large (including the same sub-population), giving a total probability of 0·904 of mating with an individual in the same sub-population. The only exception to this rule was that in the sterility case, if it proved impossible after 10 attempts to find a suitable mate in the sub-population (which would only happen very rarely) a suitable mate was chosen from the population at large.

TABLE 1

Number of generations for which the number of alleles assumed the indicated values in 800 consecutive generations

Neutral population

Number of alleles	5	6	7	8	9	10	11	12	13
Inbred case	2	1	13	30	45	55	71	70	94
Random case	0	0	4	15	38	82	118	139	121
Number of alleles	14	15	16	17	18	19	20	21	Total
Inbred case	110	86	73	69	47	25	7	2	800
Random case	110	80	58	21	11	2	1	0	800

Sterility population

Number of alleles	22	23	24	25	26	27	28	29	30	31	32	33
Inbred case	0	0	0	13	26	20	29	37	53	61	65	115
Random case	2	1	6	5	15	35	62	51	67	65	90	129
Number of alleles	34	35	36	37	38	39	40	41	42	43	Total	
Inbred case	125	91	69	59	23	12	2	0	0	0	800	
Random case	93	68	45	31	11	6	8	6	2	2	800	

In all cases, both neutral and sterility populations were started with twenty different alleles. The results obtained indicated that after 200 generations equilibrium behaviour had been reached. All populations were then run for a further 800 generations. The numbers of generations for which the number of alleles in the population assumed various values in the 800 equilibrium generations are displayed below in table 1.

It is found from table 1 for the random mating case that the mean number of alleles maintained in the normal population is 12·66 and for the sterility population 32·18. For the inbred case the mean numbers are 13·52 and 32·88 respectively. Further, more detailed information about the composition of each generation showed that in the normal case, generally speaking all but two or three alleles occurred quite rarely (less than 10 times in a population of 1,000 genes), whereas in the

sterility case most alleles which were present at any time occurred in reasonable numbers. As an example of this, in the final generation in the random breeding case there were 10 normal genes and 29 sterile genes. Of the 10 normal genes, only three occurred more than ten times, whereas of the 29 sterility genes, twenty-three occurred more than ten times.

So far as verification of equation (4) is concerned, with $N = 500$, and $u = 10^{-3}$ the right-hand side in (4) becomes $2 \log_e 1{,}000 - 2 = 11{\cdot}82$. This is in reasonable agreement with the empirical result 12·66. For the sterility case, Wright's table (Wright, (1964), page 618) for $u = 10^{-3}$, $N = 500$ gives approximately 30 alleles which, considering the number of approximations needed in deriving the table, is in excellent agreement with the empirical value 32·18.

It is also quite clear that rather severe geographical inbreeding has a minor effect on the number of alleles maintained. This is in general agreement with the sort of result obtained from different arguments by Moran (1962, page 178).

4. WRIGHT'S ANALYSIS FOR THE STERILITY POPULATION

The above numerical value for the sterility case agrees with that predicted by Wright's analysis (1939, 1960, 1964). It is therefore of some interest to consider this analysis together with criticisms which have been made of it (Fisher (1958), Moran (1962), Ewens (1964*c*)). So far as Fisher's analysis is concerned, the method of approach is generally similar to that of Wright, and it seems difficult to justify his claim that his results are very different from those of Wright and that his discussion clarifies that of Wright. Fisher's formula for the number of new mutations required per generation to balance losses is

$$\sqrt{N/2\pi}\exp(-2N\alpha^2), \qquad \ldots\ldots(6)$$

where α can be taken as approximately $\bar{n}^{-1}$. In the case $\bar{n} = 32$, $N = 500$, formula (6) gives about 3·36 mutations per generation, compared with the true value of one per generation. Wright's values seem to be far more accurate than Fisher's.

The criticisms of Moran (1962) and Ewens (1964*c*) are more basic and can be considered together. Moran has pointed out the non-Markovian nature of the frequency of any allele and the crude nature of the approximations necessary to find a Markovian variate. This criticism involves formulae used for $\overline{\Delta x}$, and is a valid one, the answer to it being that the numerical example, above, shows that the approximations are far better than could reasonably be expected. Moran's main criticism is the use of a stationary distribution formula for describing this process, a criticism also made by Ewens (1964*c*) and which will be considered later. The first criticism of Ewens (1964*c*) was that an incorrect formula for variance was used by Wright (1960). This criticism is effectively met by the use of the more correct formula by

Wright (1964). It may be useful, in connection with this, to point out that use of the formula $\sigma^2_{\Delta x} = x(1-x)/2\mathcal{N}$ does not imply an " apparently incorrect range " to the variable x (which is restricted to $(0, \frac{1}{2})$) as has been asserted by Wright (1964). This is because the variance formula is not the " usual binomial variance " since the reproductive system is far more complicated than one leading to such a simple formula. The fact that when the number of alleles is large the variance happens to reduce approximately to $x(1-x)/2\mathcal{N}$, the binomial formula, is coincidental. In fact for three alleles the correct formula is $x(1-x)/4\mathcal{N}$.

The second criticism of Ewens was that it is not justifiable to use diffusion approximations (analogous to equation (4)) for self-sterility populations, since the conditions required for their application do not hold for such populations. For details of this point, see Ewens (1964*c*). Despite the validity of this criticism, the answer to it is apparently that although one is not justified in using diffusion formulae, these formulae nevertheless give reasonable answers when applied formally in the case considered. This is possibly fortuitous as in other problems it can be shown that diffusion formulae, when used inappropriately, lead to extremely incorrect results. Thus the belief of Ewens (1964*c*) that " use of diffusion methods has led to very inaccurate results for self-sterility populations", while justifiable on general terms, is not in fact true in this case.

The third criticism concerns the use of stationary distribution formulae, and this turns out to lead to a most remarkable feature of the analysis. Any newly formed mutant allele will in general not last more than a few generations in the population. Using the formula (1) and the numerical results of the previous section, it appears that any newly-formed mutant allele survives, on the average, for about 32 generations. Even this figure is misleading since the detailed numerical results, not presented here, show that the majority of mutant alleles do not survive more than five or ten generations. This makes it clear that it is meaningless to refer to a stationary distribution of any allele. The discussion of Wright (1964, page 611) misses the point of this criticism, since it refers to the eventual dying out of the whole population (which admittedly will take an enormous time) rather than the dying out of the line initiated by a single mutant (which will generally take only a few generations). The relevant transient behaviour of the population is described by the pseudo-transient function.

Even if the mean time until the line initiated by a mutant dies out is very large, it should be pointed out that formal application of the stationary distribution formula in no way describes the transient behaviour of the frequency of the mutant. The best example of this is provided by the model (2) in the case $u = 0$. There are two functions which describe, in different senses, the transient behaviour of the frequency $x = i/2\mathcal{N}$ of the allele in question. The first of these, called the asymptotic conditional distribution, is the conditional distribution

of x, given x is not zero or unity, asymptotically as time increases. This distribution is given, to a close approximation, by

$$f_1(x) = 1, \qquad 0 < x < 1. \qquad \ldots\ldots(7)$$

The second function is the psuedo-transient function $f_2(x)$, having the property that

$$\int_{x_1}^{x_2} f_2(x)dx$$

is the mean time that x assumes a value in the range (x_1, x_2) before being absorbed at zero or unity. This is given to a close approximation by

$$\begin{aligned} f_2(x) &= 2(1-p)/(1-x) && 0 < x \leqq p \\ &= 2p/x && p \leqq x < 1 \end{aligned} \qquad \ldots\ldots(8)$$

where p is the initial value of x. Finally the function $f_3(x)$ derived by formally applying Wright's formula for stationary distributions is

$$f_3(x) = \text{const}/x(1-x), \qquad (2N)^{-1} \leqq x \leqq 1-(2N)^{-1}. \qquad \ldots\ldots(9)$$

It is clear that in no way does (9) resemble (7) or (8), and does not describe, in any sense, the transient behaviour of the process. This is true no matter how large N is, that is how long the mean time to absorption at zero or unity may be. It is this fact which underlies the criticisms of Moran and Ewens concerning formal use of stationary distributions in cases where there is no sort of stationary behaviour whatever.

The remarkable fact referred to above concerning the analogues of (8) and (9) in the context of the self-sterility problem, and which explains the numerical accuracy of Wright's values for the mean number of alleles maintained, is that the analogues of (8) and (9) happen to agree in algebraic form over the part of the range which is really important. In general, if the mean change in x and the variance of this change in successive generations are $\overline{\Delta x}$ and $\sigma^2_{\Delta x}$ respectively, then the general formula for the pseudo-transient function when there is one-way mutation from the allele in question to other alleles, is

$$f_2(x) = \frac{B}{\sigma^2_{\Delta x}} \exp\left[2\int^x \overline{\Delta(y)}/\sigma^2_{\Delta y}\, dy\right] \int_A^x \exp\left[2\int^y \overline{\Delta(z)}/\sigma^2_{\Delta z}\, dz\right] dy$$

$$0 < x \leqq p \qquad \ldots\ldots(10)$$

$$= \frac{C}{\sigma^2_{\Delta x}} \exp\left[2\int^x \Delta(y)/\sigma^2_{\Delta y}\, dy\right] \quad p \; W \; x \leqq 1$$

where p is the initial value of x and A, B and C are constants whose exact values need not concern us. It will be noted that in the range $(p, 1)$, this formula happens to coincide with Wright's formula for stationary distributions, and since in this case $p = (2N)^{-1}$, this agreement in algebraic form extends over practically all the range of the

distribution. It must be emphasised that this agreement in algebraic form does not imply any agreement in interpretation, but does imply that formal use of stationary distribution formulae, in this case, should lead to numerically satisfactory results. It is this remarkable coincidence which justifies use of Wright's table (1964, page 618) for the number of alleles maintained in a given population.

Using this table, and the observed negligible effect of rather close geographical inbreeding, we now confirm properly the conclusion reached by Wright and Fisher, that the explanation for the occurrence of 45 different alleles at one locus in a population of 500, with mutation rate of order 10^{-6}, is a recent reduction in the population size from about 10,000 to 500.

5. SUMMARY

The number of different alleles maintained on the average by a given mutation rate in various populations is considered. For the selectively neutral case, numerical values agree very well with theoretical predictions. The same is true of self-sterility populations, despite criticism of Wright's analysis for the latter case. This analysis is again reviewed and it is found that the main criticism of it, namely the use of stationary distribution formulae in a situation where there is no concept of stationarity, is justified. However it happens that the correct distribution agrees in algebraic form with the inapplicable stationary distribution over most of the range, so that use of the stationary formula, while quite unjustified, happens to lead to satisfactory numerical results.

6. REFERENCES

EWENS, W. J. (1964*a*). "Correcting diffusion approximations in finite genetic models." Stanford University Technical Report 4, Contract NIH GM 10452-01A1 (Department of Mathematics), April 1.

EWENS, W. J. (1964*b*). "The maintenance of alleles by mutation". *Genetics*, *50*, No. 5, 891-898.

EWENS, W. J. (1964*c*). "On the problem of self-sterility alleles." *Genetics*, *50*, No. 6, 1433-1438.

EWENS, W. J. (1965). "Two diffusion distributions in genetics." To appear in *Ann. Hum. Genet.*

FISHER, R. A. (1958). "*The Genetical Theory of Natural Selection.*" Dover, New York.

MORAN, P. A. P. (1962). "*The Statistical Processes of Evolutionary Theory.*" Oxford University Press, London.

WRIGHT, S. (1931). "Evolution in Mendelian populations." *Genetics*, *16*, 97-159.

WRIGHT, S. (1939). "The distribution of self-sterility alleles in populations." *Genetics*, *24*, 538-552.

WRIGHT, S. (1960). "On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in a population of given size: a re-examination." *Biometrics*, *16*, 61-85.

WRIGHT, S. (1964). "The distribution of self-incompatibility alleles in populations." *Evolution*, *18*, 609-619.