Genet. Res., Camb. (1973), **22**, *pp.* 201–204
With 2 text figures
Printed in Great Britain

SHORT PAPERS

A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*

BY TOMOKO OHTA AND MOTOO KIMURA
National Institute of Genetics, Mishima, Japan

(*Received* 20 *February* 1973)

SUMMARY

A new model of mutational production of alleles was proposed which may be appropriate to estimate the number of electrophoretically detectable alleles maintained in a finite population. The model assumes that the entire allelic states are expressed by integers $(\ldots, A_{-1}, A_0, A_1, \ldots)$ and that if an allele changes state by mutation the change occurs in such a way that it moves either one step in the positive direction or one step in the negative direction (see also Fig. 1). It was shown that for this model the 'effective' number of selectively neutral alleles maintained in a population of the effective size N_e under mutation rate v per generation is given by

$$n_e = \sqrt{(1+8N_e v)}.$$

When $4N_e v$ is small, this differs little from the conventional formula by Kimura & Crow, i.e. $n_e = 1+4N_e v$, but it gives a much smaller estimate than this when $4N_e v$ is large.

Since a model of isoalleles with infinite states was proposed by Kimura & Crow (1964) it has been used extensively to estimate the number of selectively neutral isoalleles that can be maintained in a finite population under a given mutation rate. In this model it is assumed that the number of possible allelic states at a locus is so large that whenever mutation occurs it represents a new, not pre-existing allele. This model could be applied directly to actual situations if individual variants were identified at the level of nucleotide or amino acid sites. At present, however, our experimental analyses of the genetic variability of natural populations are at much cruder level of identifying electrophoretically detectable variants. In other words, a gene mutation can be detected only when it leads to a replacement of amino acid which causes change in electric charge of the molecule. Not only such variants occupy a relatively small fraction of the entire variants at the molecular level, but also they are identified only as a discrete spectrum of broad bands on the electrophoresis gels. This means that the electrophoretic method does not have the resolving power which the model of Kimura & Crow presupposes.

The purpose of the present note is to propose a model which may be more appropriate to estimate the number of electrophoretically detectable alleles, allowing us to compare theoretical predictions with actual observations. Let us assume that the entire sequence of allelic states are expressed by integers as shown in Fig. 1, and that if an allele changes its state by a single step mutation, the change occurs in such a way that it moves either one step in the positive direction or one step in the negative direction. In other words, it can

* Contribution no. 922 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411, Japan.

mutate only to one of the two adjacent states. In this model, one positive and one negative changes (in charge) cancel each other, leading the allele back to the original state. An actual example of this type of change is afforded by mutant proteins A 11 and A 46 of tryptophan synthetase of *E. coli*. According to Henning & Yanofsky (1963), A 11 moves toward the negative direction after electrophoresis on cellulose acetate while A 46 moves toward the positive direction. The mobility of the double mutant A 11–46 protein was found to be identical with that of the wild-type A protein.

Consider a diploid population with the effective size N_e, and let v be the mutation rate per locus per generation. To simplify the treatment, we shall assume that under mutation, changes toward the positive and the negative directions occur with equal frequencies (i.e. each with $\frac{1}{2}v$ as shown in Fig. 1). Let x_i be the frequency of the ith allele A_i (i = integer), and also let

$$C_k = E\{\sum_i x_i x_{i+k}\} \quad (k = 0, 1, \ldots), \tag{1}$$

where E stands for the operator for taking expectation. The summation is over all relevant alleles in the population. Note that C_0 is the expected value of the sum of squares of allelic frequencies, so that it gives the average homozygosity under random mating. The correlation between frequencies of alleles that are k steps apart may be given by C_k/C_0.

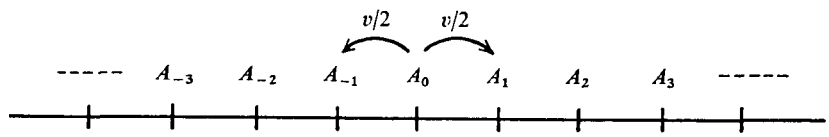

Fig. 1. Diagram illustrating the model of production of electrophoretically detectable alleles.

In order to obtain a set of equations giving the rate of change in C_k, we use the basic equation for generating the moments (Ohta & Kimura, 1971). This equation takes the following form:

$$(d/dt)\,E(f) = E\{L(f)\}, \tag{2}$$

where L is the differential operator of the Kolmogorov backward equation (see equation A 5 of Ohta & Kimura, 1971) and f is an arbitrary continuous function of x_i's. Assuming that the alleles are selectively neutral, the mean (M), the variance (V) and the covariance (W) of gene frequency changes per generation are

$$M_{\delta x_i} = \tfrac{1}{2}v(x_{i-1} + x_{i+1}) - vx_i,$$
$$V_{\delta x_i} = \frac{x_i(1-x_i)}{2N_e} \tag{3}$$

and

$$W_{\delta x_i \delta x_j} = -\frac{x_i x_j}{2N_e} \quad (i \neq j).$$

Let $f = x_i^2$ in formula (2), then

$$\frac{d}{dt}E(x_i^2) = E\left\{v(x_{i-1} + x_{i+1})\,x_i - 2vx_i^2 + \frac{x_i(1-x_i)}{2N_e}\right\}$$

or

$$\frac{d}{dT}E(x_i^2) = E\{2N_e v(x_{i-1}x_i + x_i x_{i+1}) - (4N_e v + 1)\,x_i^2 + x_i\},$$

where $T\,(= t/(2N_e))$ is time measured in the unit of $2N_e$ generations. By summing up the above equation for all i, and noting (1), we obtain

$$dC_0/dT = -(1 + 4N_e v)\,C_0 + 4N_e v C_1 + 1, \tag{4}$$

https://doi.org/10.1017/S0016672300012994 Published online by Cambridge University Press

since $\sum_i x_i = 1$. Similarly, by letting $f = x_i x_{i+k}$, we obtain the following equation for C_k:

$$\frac{dC_k}{dT} = 2N_e v C_{k-1} - (1+4N_e v) C_k + 2N_e v C_{k+1} \quad (k \geqslant 1). \tag{5}$$

At equilibrium, we have $dC_k/dt = 0$ for all k, and the appropriate equilibrium solution for the set of equations (4) and (5) which vanishes at $k = \infty$ is given by

$$C_k = H_0 \lambda_1^k, \tag{6}$$

where

$$H_0 = 1/\sqrt{(1+8N_e v)} \tag{7}$$

and

$$\lambda_1 = \frac{1+4N_e v - \sqrt{(1+8N_e v)}}{4N_e v}. \tag{8}$$

To check this solution, we considered a finite (n) set of allelic states arranged on a circle, and derived a set of equations transforming C_k's ($k = 0, 1, \ldots, n$) from one generation to the

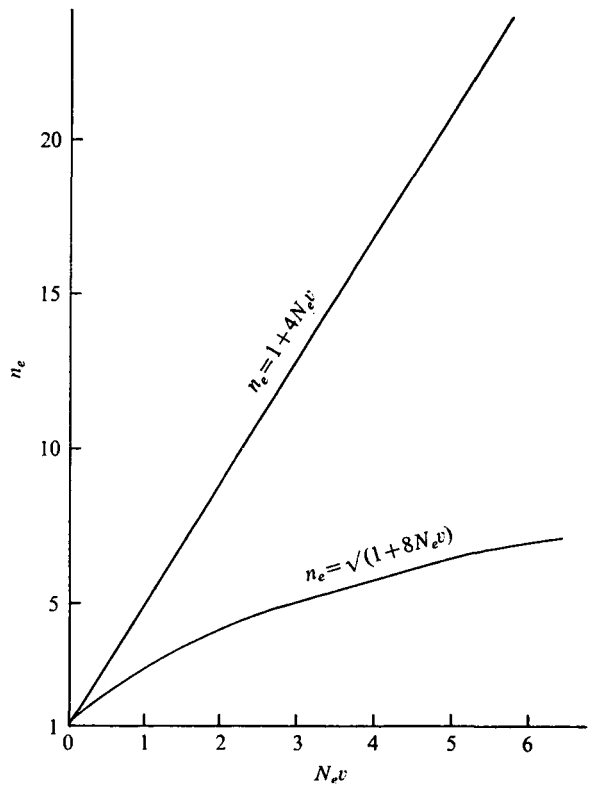

Fig. 2. Relationship between the effective number of alleles (n_e) and $N_e v$ under the two models (the model of Kimura & Crow and the present model).

next. Then we multiplied the matrix corresponding to the finite set of transformations a large number of times by a computer assuming $n = 10, 20$ and 40 allelic states and $N_e v = 0{\cdot}5$ and $2{\cdot}0$. It was found that formula (6) gives good approximation to the equilibrium values of C_0, C_1, etc., obtained by the matrix multiplication.

The 'effective' number of alleles is given by the reciprocal of the average homozygosity H_0. Thus, we obtain

$$n_e = \sqrt{(1+8N_e v)}. \tag{9}$$

https://doi.org/10.1017/S0016672300012994 Published online by Cambridge University Press

Monte Carlo experiments were performed to check this formula for various combinations of values of N_e and v and the results were satisfactory. The above formula for n_e should be compared with the corresponding formula,

$$n_e = 1 + 4N_e v, \tag{10}$$

obtained by Kimura & Crow (1964). Fig. 2 illustrates the relationship between $N_e v$ and n_e for these two models.

It may be seen from the figure that for a small value of $N_e v$ these two formulae differ rather little. For example, if $4N_e v = 0{\cdot}2$, we have $n_e = 1{\cdot}18$ from (9) but $n_e = 1{\cdot}20$ from (10). The former gives the average heterozygosity of 15·5 % while the latter gives 16·7 %. However, these two formulae give very different estimates for n_e when $N_e v$ is large. For example, if $N_e v = 100$, the present model (9) gives $n_e = 28{\cdot}3$, while the conventional model (10) gives $n_e = 401$. Recently, Ayala *et al.* (1972) in criticizing our neutral theory of protein polymorphism (Kimura & Ohta, 1971), point out that in *D. willistoni*, which is estimated to have 10^9 breeding flies per generation, if we take a lower estimate of 10^{-7} for mutation rate, we should have $4N_e v = 400$, while the observed average heterozygosity per locus is about 18 %. In other words, the theoretical value of n_e obtained from Kimura & Crow's formula overestimates the true value by the factor of some 400. Such a marked discrepancy is reduced, however, if we use the present formula, although the theoretical value is still some 20 times greater. In addition, there are two possibilities which have to be taken into account and both of which reduce our theoretical estimate based on the neutral theory. First, as pointed out by Kimura & Ohta, the rate $v = 10^{-7}$ may be appropriate for *the neutral mutation rate per year*, but not per generation. This means we should take much smaller value for the neutral mutation rate per generation for fruit flies that breed all the year round in tropical forests. Secondly, there is possibility that $N_e = 10^9$ is an overestimate, not as the number of breeding flies per generation, but as the number applicable to the formula of effective allele number. Not only this number is controlled by the minimum population size when population number fluctuates from generation to generation, but also it takes the length of time in the order of the population size for the equilibrium state in the distribution of allelic frequencies to be established. (Treatment of this subject will be published elsewhere.) It is quite likely that the effect of small population number during the last glaciation still remains in the genetic composition of tropical fruit flies. In this connexion we would like to point out that neutral alleles behave quite differently from lethal genes (having short life-span) and inversion polymorphisms (many of which may be subject to 'balancing selection').

REFERENCES

AYALA, F. J., POWELL, J. R., TRACEY, M. L., MOURÃO, C. A. & PÉREZ-SALAS, S. (1972). Enzyme variability in the *Drosophila willistoni* group. IV. Genetic variation in natural populations of *Drosophila willistoni*. *Genetics* **70**, 113–139.

HENNING, U. & YANOFSKY, C. (1963). An electrophoretic study of mutationally altered A proteins of the tryptophan synthetase of *Escherichia coli*. *J. Molecular Biology* **6**, 16–21.

KIMURA, M. & CROW, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

KIMURA, M. & OHTA, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469.

OHTA, T. & KIMURA, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**, 571–580.