Proc. Nat. Acad. Sci. USA
Vol. 72, No. 7, pp. 2761–2764, July 1975
Genetics

Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles*

(population genetics/frequency distribution/electrophoretically detectable alleles)

MOTOO KIMURA AND TOMOKO OHTA

National Institute of Genetics, Mishima, 411 Japan

Contributed by Motoo Kimura, April 25, 1975

ABSTRACT **A formula for the distribution of allelic frequencies in a finite population is derived assuming stepwise production of multiple alleles. Monte Carlo experiments were performed to check the validity of the formula, and excellent agreement was obtained between theoretical distribution and experimental results. The formula should be useful for analyzing genetic variability in natural populations that can be detected by electrophoretic methods.**

For the purpose of analyzing genetic variability that can be detected by electrophoretic methods, we recently proposed a model of stepwise production of alleles (1, 2). In this model, it is assumed that the entire sequence of allelic states can be expressed by integers (. . . , A_{-1}, A_0, A_1, . . .), and that if an allele changes states by mutation it moves either one step in the positive direction or one step in the negative direction in the allele space (Fig. 1). As compared with the conventional model of Kimura and Crow (3) which assumes that every mutation leads to a new, not preexisting allele, this model has a feature that mutations are to some extent recurrent; less frequent alleles in a population tend to be produced repeatedly from mutation of more frequent adjacent alleles. Analytical treatment of this model is much more difficult than that of Kimura and Crow, but we have obtained, using diffusion equation methods, the formula for the effective number of selectively neutral alleles maintained in a finite population (2). Namely, if v is the mutation rate per locus per generation such that the mutational changes toward the positive and the negative directions occur with equal frequencies as shown in Fig. 1, and if N_e is the effective size of the population, then, assuming selective neutrality of mutation, the effective number of alleles at equilibrium is given by

$$n_e = \sqrt{1 + 8N_e v}. \qquad [1]$$

More generally, we have shown that if C_k is the expectation of the product of the frequencies between two alleles that are k steps apart in the allele space, then

$$C_k = H_0 \lambda_1^{\,k}, \qquad [2]$$

where

$$H_0 = 1/\sqrt{1 + 8N_e v} \qquad [3]$$

and

$$\lambda_1 = (1 + 4N_e v - \sqrt{1 + 8N_e v})/(4N_e v). \qquad [4]$$

In particular, H_0 represents the average homozygosity or the expectation of the sum of squares of allelic frequencies. Since the effective number of alleles (n_e) is given by the reciprocal of H_0, formula **1** follows immediately from **3**. Furthermore, the expectation of the product of frequencies of adjacent alleles is

$$C_1 = \frac{1 + 4N_e v - \sqrt{1 + 8N_e v}}{4N_e v\sqrt{1 + 8N_e v}}. \qquad [5]$$

Note that these results are concerned with the second moments of the distribution of allelic frequencies rather than the distribution itself.

However, in order to assess, by detailed analyses of observations, the role of mutation and random drift for the maintenance of protein polymorphisms, knowledge on the actual form of the distribution is required. Although considerable information can be gained on the nature of the distribution through careful Monte Carlo experiments (4), it is much more desirable to derive the distribution analytically. The purpose of the present paper is to show that under a simplifying assumption (as expressed by Eq. **10** below), a distribution can be obtained that satisfies known relations **3** and **5**, and also gives excellent fit to Monte Carlo experiments for $4N_e v$ up to unity.

BASIC THEORY

Let us consider a random mating population of N diploid individuals, and let N_e be the effective size of the population which may be different from the actual size N. We assume that under mutation and random drift a statistical equilibrium is reached with respect to distribution of allelic frequencies. We shall designate this distribution by $\Phi(x)$ such that $\Phi(x)dx$ represents the expected number of alleles whose frequencies in the population are in the range $(x \sim x + dx)$.

Consider an allele (say A_0) whose frequency in the population is x. Let x_{-1} and x_1 be the frequencies of alleles (A_{-1} and A_1) that are adjacent to A_0. Then, under stepwise production of alleles as shown in Fig. 1, the mean and the variance of the change of x per generation are given respectively by

$$M_{\delta x} = -vx + \frac{v}{2} E(x_{-1} + x_1 | x) \qquad [6]$$

and

$$V_{\delta x} = x(1 - x)/(2N_e), \qquad [7]$$

where v is the mutation rate and $E(x_{-1} + x_1|x)$ is the conditional expectation of the sum of the frequencies of the two

* Contribution no. 1057 from the National Institute of Genetics, Mishima, Shizuoka-ken 411 Japan.

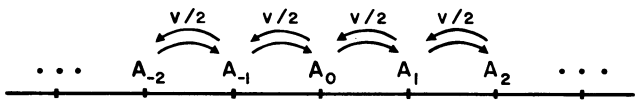

FIG. 1. Diagram showing the model of stepwise production of alleles.

adjacent alleles given that the frequency of the allele in question is x.

If $E(x_{-1} + x_1|x)$ can be expressed as a known function of x, then $\Phi(x)$ may be obtained by applying Wright's (5) formula for the steady-state gene frequency distribution, that is, by

$$\Phi(x) = \frac{C}{V_{\delta x}} \exp\left\{2\int \frac{M_{\delta x}}{V_{\delta x}}\,dx\right\} \qquad [8]$$

where C is a constant determined such that the sum of allelic frequencies in the population is unity:

$$\int_0^1 x\Phi(x)dx = 1. \qquad [9]$$

Note that this condition is different from the one originally used by Wright to determine the constant. The reason for this is that in his case the distribution represents probability density, while in our case $\Phi(x)$ represents density of the expected number of alleles (see also refs. 3 and 6).

Let us now assume that $E(x_{-1} + x_1|x)$ can be expressed, with sufficient accuracy, as a constant fraction of the total frequencies of the remaining alleles so that

$$E(x_{-1} + x_1|x) = b(1 - x), \qquad [10]$$

where b is a constant that may depend on v and N_e but not on x. Substituting Eq. **10** in Eq. **6** we get

$$M_{\delta x} = -vx + \frac{1}{2}vb(1 - x), \qquad [6a]$$

and, this together with Eq. **7**, allows us to obtain the distribution by using formula **8**. Thus, we obtain

$$\Phi(x) = C(1 - x)^{\alpha - 1}x^{\beta - 1}, \qquad [11]$$

where $\alpha = 4N_ev$ and $\beta = 2N_evb$. In this formula, constant C is determined by using condition **9**, and we get

$$C = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha)\Gamma(\beta + 1)}. \qquad [12]$$

In order to determine another constant b, we note that the expectation of the sum of squares of allelic frequencies is $H_0 = 1/\sqrt{1 + 8N_ev}$ as given in Eq. **3**. In terms of the present distribution, we have

$$H_0 = \int_0^1 x^2\Phi(x)dx = (\beta + 1)/(\alpha + \beta + 1), \qquad [13]$$

and therefore, by equating this with $1/\sqrt{1 + 8N_ev}$, we obtain

$$\beta = (\alpha + 1 - \sqrt{1 + 8N_ev})/(\sqrt{1 + 8N_ev} - 1). \qquad [14]$$

Thus, the required distribution may be expressed in the following form:

$$\Phi(x) = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha)\Gamma(\beta + 1)}(1 - x)^{\alpha - 1}x^{\beta - 1}. \qquad [15]$$

where $\alpha = 4N_ev$, $\beta = 2N_evb$ and

$$b = \frac{1 + 4N_ev - \sqrt{1 + 8N_ev}}{2N_ev(\sqrt{1 + 8N_ev} - 1)}. \qquad [16]$$

We shall now show that with this distribution and the above assignment of constant b in Eq. **10** we can derive the correct value for the expectation of the product of frequencies between adjacent alleles. In terms of the present distribution, this is given by

$$C_1 = \frac{1}{2}\int_0^1 xE(x_{-1} + x_1|x)\Phi(x)dx$$

and, noting Eqs. **9, 10, 13,** and **16,** we get

$$\begin{aligned} C_1 &= \frac{b}{2}\int_0^1 x(1 - x)\Phi(x)dx \\ &= \frac{b}{2}(1 - H_0) = \frac{b}{2}\left(1 - \frac{1}{\sqrt{1 + 8N_ev}}\right) \\ &= \frac{1 + 4N_ev - \sqrt{1 + 8N_ev}}{4N_ev\sqrt{1 + 8N_ev}}. \end{aligned}$$

This agrees with Eq. **5**, the result obtained earlier by us (2) using an entirely different method. Furthermore, by letting $f = x_i^n (n \geq 2)$ in $E\{L(f)\} = 0$ in Ohta and Kimura's (ref. 2) formulation, we find that $\Phi(x)$ must satisfy the relation:

$$\begin{aligned} (n - 1)\left\{\int_0^1 x^{n-1}\Phi(x)dx - \int_0^1 x^n\Phi(x)dx\right\} \\ -4N_ev\left\{\int_0^1 x^n\Phi(x)dx \right. \\ \left. - \frac{1}{2}\int_0^1 E(x_{-1} + x_1|x)x^{n-1}\Phi(x)dx\right\} = 0, \end{aligned}$$

and we can show, in fact, this is satisfied by Eq. **15**, assuming Eq. **10**. We can also show that as N_ev gets small, b approaches unity so that $\Phi(x)$ approaches the distribution given by Eq. **11** but with $\alpha = 4N_ev$ and $\beta = 2N_ev$. On the other hand, it is known (see ref. 6) that this form of distribution represents the case in which the entire allelic space consists of three allelic states with stepwise production of mutations. This is reasonable since when N_ev is small, the number of different alleles contained at any moment within a population seldom exceeds three, so that the triallelic space should be sufficient for describing the distribution. Incidentally, we note that b can be expressed in the form:

$$b = \frac{2}{n_e - 1}\lambda_1. \qquad [16a]$$

Since the effective number of alleles excluding the allele under consideration (i.e., A_0) is $n_e - 1$, of which two are in the states adjacent to A_0, and, since $\lambda_1 = C_1/H_0$ represents, in a sense, correlation between adjacent alleles, we think that b given by Eq. **16a** allows us a very natural interpretation.

Table 1. Comparison of n_a between Kimura–Crow and stepwise production models

	N							
	5×10^2		10^3		5×10^3		10^4	
$N_e v$	K–C	Step	K–C	Step	K–C	Step	K–C	Step
0.01	1.274	1.256	1.301	1.280	1.366	1.335	1.394	1.358
0.025	1.675	1.578	1.745	1.627	1.906	1.736	1.975	1.780
0.05	2.324	1.994	2.462	2.068	2.784	2.220	2.923	2.280
0.1	3.557	2.562	3.834	2.649	4.478	2.816	4.755	2.874
0.25	6.908	3.434	7.601	3.501	9.210	3.604	9.903	3.633

Comparison of the average number of alleles (n_a) between the Kimura–Crow model (K–C) and the model of stepwise production of alleles (Step), under various combinations of N and $N_e v$, where N and N_e are respectively the actual and the effective numbers of the population, and v is the mutation rate.

From the frequency distribution **15** thus obtained, we can derive a few formulae that may be useful for analyzing observations on protein polymorphisms. The average number of alleles maintained in a population is given by

$$n_a = \int_{1/2N}^{1} \Phi(x)dx \approx \frac{\alpha + \beta}{\beta}\left\{1 - \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta + 1)}\left(\frac{1}{2N}\right)^{\beta}\right\}. \quad [17]$$

The average number of different alleles in a sample of n individuals is

$$n_a = \int_{1/2N}^{1} \Phi(x)\{1 - (1 - x)^{2n}\}dx = \frac{\alpha + \beta}{\beta}\left\{1 - \prod_{i=0}^{2N-1}\left(\frac{i + \alpha}{i + \alpha + \beta}\right)\right\}. \quad [18]$$

A corresponding expression for the model of Kimura and Crow was earlier given by Ewens (7). The probability that a population is monomorphic can also be obtained by using Eq. **15**. We define that a population is monomorphic if the proportion of the most frequent allele in the population exceeds $1 - q$, where q is a small quantity such as 0.01. Let P_{mono} be the probability that a population is monomorphic. Then, for a small value of q, we have, with good approximation

$$P_{\text{mono}} \approx \int_{1-q}^{1} \Phi(x)dx \approx \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + 1)\Gamma(\beta + 1)} q^{\alpha}, \quad [19]$$

where $\alpha = 4N_e v$ and $\beta = 2N_e vb$. The probability of polymorphism can be computed from $1 - P_{\text{mono}}$.

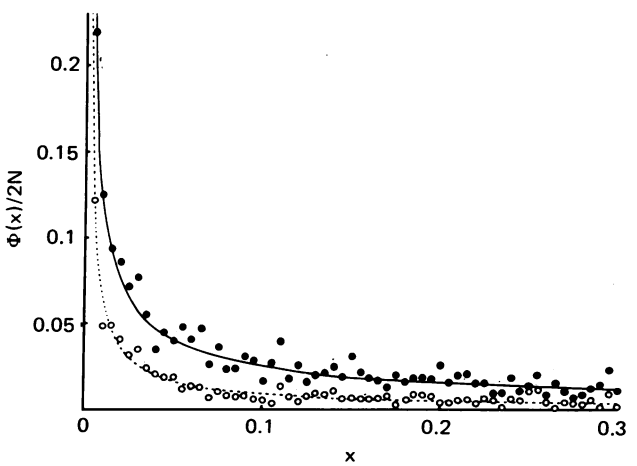

FIG. 2. Comparison between the theoretical distributions and the results of Monte Carlo experiments. The solid curve represents the theoretical distribution and solid circles represent the corresponding experimental results for the case $N_e v = 0.2$. The broken curve represents the theoretical distribution and open circles the experimental results for $N_e v = 0.05$. The abscissa is allelic frequency; the ordinate is the expected number of alleles for each frequency class. For details, see *text*.

MONTE CARLO EXPERIMENTS

In this section, we intend to show that the allelic distribution we obtained gives excellent fit to the results of simulation experiments. The procedure of the experiments follows our previous study (4), namely, the allelic states are arranged on a circle with a sufficiently large number of total states such that the number of segregating alleles within a population at any time is a small fraction of the total states. Each generation of the experiment consists of mutation and sampling. Mutational changes were carried out deterministically. Each experiment, starting from a homogeneous population, was continued until the generation 50,000. The population size assumed was 100. The allelic frequencies were counted every 50 generations. Fig. 2 illustrates the theoretical distribution together with experimental results for the two cases; $N_e v = 0.05$ and 0.2. As seen from the figure, the agreement between the theoretical distribution and experimental results is excellent.

DISCUSSION

The pattern of allelic distribution derived in this paper based on the step model of mutation is somewhat different from the corresponding distribution obtained by Kimura and Crow (3), particularly for a lower frequency range. Namely, the curve describing the distribution rises less steeply toward the origin for the step model than for the Kimura–Crow model, and the difference gets larger as $N_e v$ increases. This is reflected in the difference of the average number of alleles (n_a) for the two models as shown in Table 1. In this table, values of n_a are computed by formula **17** for the step model, while the corresponding values for the Kimura–Crow model are taken from Kimura (6). Note that n_a is always smaller for the step model. The difference of n_a between the two models becomes larger as $N_e v$ gets larger. Furthermore, the difference of n_a is more pronounced than

that of n_e between the two models. This means that the allelic distribution is more uniform for the step model than for the Kimura–Crow model, as we pointed out in one of our previous papers (4). We note also that the probability of polymorphism is only a little influenced under the stepwise production of alleles.

The step model of mutation assumes that one positive and one negative change cancel each other. It is possible, as we pointed out when we first proposed the model (1), that this is an oversimplification for representing the electrophoretically detectable changes. However, deviation from reality is of only minor consequence, and we believe that the step model is more adequate than the conventional Kimura–Crow model for the analysis of genetic variability that can be detected by electrophoretic methods. In fact, the idea of stepwise production of variants has since been used in several studies of molecular evolution and polymorphism (8–13).

1. Kimura, M. & Ohta, T. (1973) *Genetics Suppl.* **73,** 19–35.
2. Ohta, T. & Kimura, M. (1973) *Genet. Res.* **22,** 201–204.
3. Kimura, M. & Crow, J. F. (1964) *Genetics* **49,** 725–738.
4. Ohta, T. & Kimura, M. (1974) *Genetics* **76,** 615–624.
5. Wright, S. (1938) *Proc. Nat. Acad. Sci. USA* **24,** 253–259.
6. Kimura, M. (1968) *Genet. Res.* **11,** 247–269.
7. Ewens, W. J. (1972) *Theor. Pop. Biol.* **3,** 87–112.
8. King, J. L. (1973) *J. Mol. Evol.* **2,** 317–322.
9. King, J. L. (1974) *Genetics* **76,** 607–613.
10. Nei, M. & Chakraborty, R. (1973) *J. Mol. Evol.* **2,** 323–328.
11. Ohta, T. (1974) *Nature* **252,** 351–354.
12. Ohta, T. & Kimura, M. (1975) *Am. Nat.*, in press.
13. King, J. L. & Ohta, T. (1975) *Genetics*, in press.