Printed in Great Britain

Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome*

By NAOYUKI TAKAHATA

National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan

(*Received 24 March 1981 and in revised form 11 August 1981*)

SUMMARY

A general model of linked genes or a part of a genome is proposed which enables us to study various problems in molecular population genetics in a unified way. Several formulae with special reference to the linkage disequilibrium and genetic distance are derived for neutral mutations in finite populations, based on the method of diffusion equations. It is argued that the model and formulae are useful particularly when observations are made in terms of DNA sequence.

1. INTRODUCTION

To understand the genetic construction of Mendelian populations, we have to know the linkage disequilibrium or nonrandom association of genes between different loci so that the study of linkage disequilibrium is one of the major subjects in current population genetics. A number of papers have been published since the early report of Haldane (1931). Among them, one line of study incorporating the effect of random genetic drift on disequilibrium was begun by Robertson (1961), Hill (1968) and Hill & Robertson (1968). Subsequently, Ohta & Kimura (1969*a*, *b*; 1971) studied the problem of the two locus two allele model taking into account mutation, and Hill (1974*a*, *b*) developed methods for computating linkage disequilibrium among several linked neutral loci and obtained the variances and covariances of disequilibria. He also investigated the total squared linkage disequilibrium among multiple neutral alleles at two loci (Hill, 1975) where the infinite allele model is assumed (Kimura & Crow, 1964). Note that attention in the above studies, except that of Ohta & Kimura (1971), was paid mainly to linkage disequilibrium between different loci.

However, recent developments in molecular biology, in particular the technique of rapid DNA sequencing by Sanger, Nicklen & Coulson (1977) and Maxam & Gilbert (1977), enable us to study the genetic make-up at the nucleotide level. In

* Contribution no. 1374 from the National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan.

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

terms of population genetics, it is expected that we will soon know the genetic structure of populations in terms of DNA sequence. Although many models have been proposed to account for such data, each has its own range or condition by which the application is limited (see, for example, Ewens 1979 for a review). Therefore, it is desirable to develop a model reflecting the real structure of a gene or a chromosome. Such a general model can be constructed by assuming that there are multiple sites linearly arranged on a chromosome, and that each site can take multiple states. For example, the model can be applied to a gene consisting of r nucleotide sites each of which can be occupied by four bases.

Based on the above model, linkage disequilibrium between two sites is re-examined in relation to the work of Ohta & Kimura (1969*b*) and Hill (1975). Assuming more than two sites and complete linkage between them, we derive several mathematical formulae with special reference to genetic distance (Nei, 1972) and evolutionary distance (Jukes & Cantor, 1969; Upholt, 1977; Nei & Li, 1979). In particular, we demonstrate how to estimate the evolutionary distance when populations are polymorphic with respect to DNA sequence or restriction enzyme maps in a particular region of the genome. The formulae are compared with those obtained mainly from the infinite allele model.

2. MODEL AND ANALYSIS

Let us consider a random mating population of diploid organisms with effective size N_e and r 'sites' linked on a chromosome in which the term 'site' may be a nucleotide site, codon, cistron and so forth. We assume that the pth site consists of K_p states, and that neutral mutation occurs with equal likelihood between those states at a rate v_p per site per generation. We also assume that recombination occurs randomly at any one of $r-1$ joints of r linked sites at a rate c per generation. Each chromosome with r such sites is specified by a vector $i = (i_1, i_2, \ldots, i_r)$ and designated by A_i where the element i_p is an integer which takes the values $1, 2, \ldots, K_p$. Let us denote by $x_i(t)$ the frequency of the ith chromosome within a population at the tth generation. For convenience, we introduce the following marginal functions of $x_i(t)$ defined by

$$\left.\begin{aligned} x_{i,p}(t) &= \sum_{i_p=1}^{K_p} x_i(t) = \sum_{i_p=1}^{K_p} x_{(i_1, i_2, \ldots, i_r)}(t), \\ x^*_{i,p}(t) &= \sum_{i_{p+1}=1}^{K_{p+1}} \cdots \sum_{i_r=1}^{K_r} x_i(t) \quad \text{and} \quad x^{**}_{i,p}(t) = \sum_{i_1=1}^{K_1} \cdots \sum_{i_p=1}^{K_p} x_i(t). \end{aligned}\right\} \tag{1}$$

Neglecting the higher order terms, the frequency $x_i(t+1)$ of A_i chromosomes due to mutation is given by

$$x_i(t+1) = \left(1 - \sum_{p=1}^{r} v_p\right) x_i(t) + \sum_{p=1}^{r} \sum_{\substack{i'_p=1 \\ i'_p \neq i_p}}^{K_p} \frac{v_p}{K_p - 1} x_{(i_1 \ldots i'_p \ldots i_r)}(t), \tag{2}$$

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

in which the first term of the right hand side corresponds to the probability of no change and the second to the contribution coming from all chromosomes that can produce A_i chromosome by a single step mutation. The amount of change in one generation of $x_i(t)$, denoted by $\Delta x_i(t)$, is

$$\Delta x_i(t) = \sum_{p=1}^{r} \frac{v_p}{K_p-1} x_{i,p}(t) - \left(\sum_{p=1}^{r} \frac{K_p v_p}{K_p-1}\right) x_i(t) \tag{3}$$

using the notation in (1).

Recombination also changes the frequency A_i and we have

$$\Delta x_i(t) = c \sum_{p=1}^{r-1} x^{*}_{i,p} x^{**}_{i,p} - c(r-1)\, x_i, \tag{4}$$

where the first and second terms in the right hand side correspond to the amount of addition and loss of x_i due to recombination, respectively.

In the following, we describe the stochastic process of $x_i(t)$ due to random sampling of gametes by a diffusion approximation (Kimura, 1964). Combining (3) and (4), the mean and covariance are as follows;

$$\left.\begin{aligned} M(\delta x_i) &= \sum_{p=1}^{r} \frac{v_p}{K_p-1} x_{i,p} + c\sum_{p=1}^{r-1} x^{*}_{i,p} x^{**}_{i,p} - \left\{\sum_{p=1}^{r} \frac{K_p v_p}{K_p-1} + c(r-1)\right\} x_i, \\ V(\delta x_i\, \delta x_{i'}) &= \frac{1}{2N_e} x_i(\delta_{ii'} - x_{i'}), \end{aligned}\right\} \tag{5}$$

where $\delta_{ii'} = 1$ only if $i_p = i'_p$ for all $p = 1, 2, \ldots, r$, and is zero otherwise. From (5), we get the multi-dimensional diffusion operator L_r for the Kolmogorov backward equation multiplied by $4N_e$,

$$L_r = \sum_i \left\{\sum_{p=1}^{r} \theta_p x_{i,p} + C\sum_{p=1}^{r-1} x^{*}_{i,p} x^{**}_{i,p} - \left(\sum_{p=1}^{r} K_p \theta_p + C(r-1)\right) x_i\right\} \frac{\partial}{\partial x_i} + \sum_i \sum_{i'} x_i(\delta_{ii'} - x_{i'}) \frac{\partial^2}{\partial x_i\, \partial x_{i'}}, \tag{6}$$

where $\theta_p = 4N_e v_p/(K_p-1)$ and $C = 4N_e c$. The operator (6) governs the process of change in chromosomal frequencies, and gives the moment equation

$$\frac{dE\{f\}}{dT} = E\{L_r f\}, \tag{7}$$

in which $E\{\quad\}$ stands for taking expectation, f is a continuous function of random variables and T is measured in units of $4N_e$ generations. However, it is not easy to treat (7) in the general case so that we shall consider two special cases to study the amount of polymorphism, linkage disequilibrium and genetic distance. One case is $c \neq 0$ but $r = 2$, and the other is $r > 2$ but $c = 0$.

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

(i) *Two locus multiple allele model* ($r = 2$, $c \neq 0$)

Linkage disequilibrium in the two locus two allele model was investigated by Ohta & Kimura (1969*b*) incorporating the different rates between forward and backward mutations at both loci. Hill (1975) studied the same problem assuming that all mutant alleles differ from pre-existing ones, i.e. the infinite allele model at both loci. Their results are rather similar when not more than two mutant alleles are segregating. In this section, we shall see the relationship between them keeping the number of states per site, K_p, finite. Any moment equation is easily obtained by substituting a function of random variables x_i for f in (7) with $r = 2$. For simplicity, we consider the first two loci in (7) and denote $\Sigma_{i_p=1}^{K_p} x_i$ by x_{i_1} for $p = 2$ and by x_{i_2} for $p = 1$.

At equilibrium we readily obtain the moments of lower order

$$\left.\begin{aligned} E\{x_i\} &= E\{x_{i_1} x_{i_2}\} = \frac{1}{K_1 K_2}, \\ E\{x_{i_1}^2\} &= \frac{1}{K_1}\frac{1+\theta_1}{1+K_1\theta_1}, \quad E\{x_{i_2}^2\} = \frac{1}{K_2}\frac{1+\theta_2}{1+K_2\theta_2}, \\ E\{x_i x_{i_1}\} &= E\{x_{i_1}^2 x_{i_2}\} = \frac{1}{K_2} E\{x_{i_1}^2\}, \\ E\{x_i x_{i_2}\} &= E\{x_{i_1} x_{i_2}^2\} = \frac{1}{K_1} E\{x_{i_2}^2\}. \end{aligned}\right\} \tag{8}$$

Here and subsequently we suppress a symbol to indicate the equilibrium state. To get the moments of higher order, let us represent the relevant equation in matrix form

$$\begin{pmatrix} \alpha+C & -C & 0 \\ -2 & \beta+C & -C \\ 0 & -4 & \gamma \end{pmatrix} \begin{pmatrix} E\{x_i^2\} \\ E\{x_i x_{i_1} x_{i_2}\} \\ E\{x_{i_1}^2 x_{i_2}^2\} \end{pmatrix} = \begin{pmatrix} a \\ 2b \\ b \end{pmatrix}. \tag{9}$$

In the above equation, the constants in the matrix and vector are, respectively,

$$\left.\begin{aligned} a &= \frac{1}{K_1 K_2}[1+\theta_1 K_2 E\{x_{i_2}^2\} + \theta_2 K_1 E\{x_{i_1}^2\}], \\ b &= \frac{1}{K_1 K_2}[(1+\theta_1) K_2 E\{x_{i_2}^2\} + (1+\theta_2) K_1 E\{x_{i_1}^2\}], \\ \alpha &= 1 + K_1\theta_1 + K_2\theta_2, \\ \beta &= 6 + 2K_1\theta_1 + 2K_2\theta_2, \\ \gamma &= 6 + K_1\theta_1 + K_2\theta_2. \end{aligned}\right\} \tag{10}$$

Accordingly, we have

$$\left.\begin{aligned} E\{x_i^2\} &= \frac{1}{|M|}\{a\beta\gamma + C(bC + (2b+a)\gamma - 4a)\}, \\ E\{x_i x_{i_1} x_{i_2}\} &= \frac{1}{|M|}\{2(b\alpha + a)\gamma + bC(C+\alpha+2\gamma)\}, \\ E\{x_{i_1}^2 x_{i_2}^2\} &= \frac{1}{|M|}\{8a + b\alpha(\beta+8) + bC(C+\alpha+\beta+6)\}, \end{aligned}\right\} \tag{11}$$

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

where $|M| = \alpha\beta\gamma + C\{(\gamma-4)C + (\alpha+\beta-2)\gamma - 4\alpha\}$. Instead of (11), Ohta & Kimura (1969*b*) and Hill (1974*a*, *b*, 1975) formulated the following three quantities directly;

$$\left.\begin{aligned} X_i &= x_{i_1}x_{i_2}(1-x_{i_1})(1-x_{i_2}),\\ Y_i &= D_i(1-2x_{i_1})(1-2x_{i_2}),\\ Z_i &= D_i^2, \end{aligned}\right\} \tag{12}$$

where $D_i = x_i - x_{i_1}x_{i_2}$ in the present notation. Using (8) to (11), we can obtain the equilibrium values of those quantities in a straightforward manner. However, as the calculation is rather tedious, we substitute (12) for (7) directly, and get

$$\left.\begin{aligned} E\{(2+K_1\theta_1+K_2\theta_2)X_i - Y_i\} &= A,\\ E\{(10+C+2K_1\theta_1+2K_2\theta_2)Y_i - 8Z_i\} &= 0,\\ E\{X_i+Y_i-(3+C+K_1\theta_1+K_2\theta_2)Z_i\} &= 0, \end{aligned}\right\} \tag{13}$$

where

$$A = \frac{1}{K_1K_2}\left\{\frac{(K_1-1)(K_2\theta_2-\theta_1)\theta_1}{1+K_1\theta_1} + \frac{(K_2-1)(K_1\theta_1-\theta_2)\theta_2}{1+K_2\theta_2}\right\}. \tag{14}$$

By solving (13), we have

$$\left.\begin{aligned} E\{X_i\} &= \frac{A(\beta'\gamma'-8)}{|M'|},\\ E\{Y_i\} &= \frac{8A}{|M'|},\\ E\{Z_i\} &= \frac{A\beta'}{|M'|}, \end{aligned}\right\} \tag{15}$$

in which

$$\begin{aligned} \alpha' &= 2+K_1\theta_1+K_2\theta_2,\\ \beta' &= 10+C+2K_1\theta_1+2K_2\theta_2,\\ \gamma' &= 3+C+K_1\theta_1+K_2\theta_2 \end{aligned}$$

and

$$|M'| = \alpha'\beta'\gamma' - 8(1+\alpha').$$

The standard linkage disequilibrium, $\sigma_d^2 = E\{Z_i\}/E\{X_i\}$, becomes

$$\sigma_d^2 = 1/[(3+C+K_1\theta_1+K_2\theta_2) - 8/(10+C+2K_1\theta_1+2K_2\theta_2)]. \tag{16}$$

All quantities treated above are concerned with a particular type of chromosome. To obtain the expectation over all types of chromosomes, it is necessary to sum over all i. However, the states of a site are all mutually equivalent so that the procedure is simply to multiply by the number of states. Let us consider two sites p and q which are $|p-q|$ steps apart in terms of site units. For instance, the expected homozygosity at the pth site, F_p, is

$$F_p = \sum_{i_p=1}^{K_p} E\{x_{i_p}^2\} = \frac{1+\theta_p}{1+K_p\theta_p} \tag{17}$$

and the expectation of the product of the homozygosities,

$$F_{p,q} = \sum_{i_p=1}^{K_p} \sum_{i_q=1}^{K_q} E\{x_{i_p}^2 x_{i_q}^2\},$$

is
$$F_{p,q} = \{8a'+b'\alpha(\beta+8)+b'C'(C'+\alpha+\beta+6)\}/|\mathrm{M}| \tag{18}$$

from (10) to (12), where

$$a' = 1+\theta_q F_p+\theta_p F_q, \quad b' = (1+\theta_q)F_p+(1+\theta_p)F_q, \quad C' = C|p-q|,$$
$$\alpha = 1+K_p\theta_p+K_q\theta_q, \quad \beta = 2(3+K_p\theta_p+K_q\theta_q), \quad \gamma = 6+K_p\theta_p+K_q\theta_q$$

and
$$|M| = \alpha\beta\gamma+C'\{(\gamma-4)C'+(\alpha+b-2)\gamma-4\alpha\}.$$

Table 1. *Dependence of* $\Sigma_i E\{X_i\}$, $\Sigma_i E\{Y_i\}$, $\Sigma_i E\{Z_i\}$ *and* σ_d^2 *on the number of possible states per site,* K, *obtained by using* (15) *and* (16), *assuming* $v_1 = v_2 = v$ *and* $K_1 = K_2 = K$

(The last column denoted by R represents the ratio of Σ_i E{Z_i} to the identity excess.)

K	$4N_e c$	$\Sigma_i E\{X_i\}$	$\Sigma_i E\{Y_i\}$	$\Sigma_i E\{Z_i\}$	σ_d^2	R
			$4N_e v = 0{\cdot}02$			
2	0·0	0·4430 × 10^{-3}	0·1521 × 10^{-3}	0·1932 × 10^{-3}	0·4362	0·5643
	2·0	0·3983 × 10^{-3}	0·5926 × 10^{-4}	0·9007 × 10^{-4}	0·2261	0·6078
	10·0	0·3755 × 10^{-3}	0·1175 × 10^{-4}	0·2960 × 10^{-4}	0·0788	0·7198
10	0·0	0·4625 × 10^{-3}	0·1629 × 10^{-3}	0·2054 × 10^{-3}	0·4442	0·5604
	2·0	0·4133 × 10^{-3}	0·6241 × 10^{-4}	0·9431 × 10^{-4}	0·2282	0·6044
	10·0	0·3888 × 10^{-3}	0·1224 × 10^{-4}	0·3074 × 10^{-4}	0·0791	0·7174
10^3	0·0	0·4650 × 10^{-3}	0·1643 × 10^{-3}	0·2070 × 10^{-3}	0·4452	0·5600
	2·0	0·4152 × 10^{-3}	0·6282 × 10^{-4}	0·9485 × 10^{-4}	0·2284	0·6040
	10·0	0·3905 × 10^{-3}	0·1231 × 10^{-4}	0·3089 × 10^{-4}	0·0791	0·7171
			$4N_e v = 0{\cdot}2$			
2	0·0	0·2216 × 10^{-1}	0·4914 × 10^{-2}	0·7126 × 10^{-2}	0·3215	0·6285
	2·0	0·2127 × 10^{-1}	0·2400 × 10^{-2}	0·4080 × 10^{-2}	0·1919	0·6648
	10·0	0·2061 × 10^{-1}	0·5684 × 10^{-2}	0·1535 × 10^{-2}	0·0745	0·7590
10	0·0	0·3012 × 10^{-1}	0·8166 × 10^{-2}	0·1111 × 10^{-1}	0·3690	0·5996
	2·0	0·2827 × 10^{-1}	0·3637 × 10^{-2}	0·5860 × 10^{-2}	0·2073	0·6393
	10·0	0·2710 × 10^{-1}	0·7947 × 10^{-3}	0·2075 × 10^{-2}	0·0766	0·7418
10^3	0·0	0·3141 × 10^{-1}	0·8748 × 10^{-2}	0·1181 × 10^{-1}	0·3760	0·5956
	2·0	0·2937 × 10^{-1}	0·3844 × 10^{-2}	0·6150 × 10^{-2}	0·2094	0·6358
	10·0	0·2812 × 10^{-1}	0·8308 × 10^{-3}	0·2160 × 10^{-2}	0·0768	0·7394

Likewise, the expectation of homozygosity of chromosomes with two specified sites p and q,

$$F_{(p,q)} = \sum_{i_p=1}^{K_p} \sum_{i_q=1}^{K_q} E\{x_i^2\},$$

where
$$i = (i_p, i_q),$$

is
$$F_{(p,q)} = \frac{1}{|M|}[a'\beta\gamma + C'\{b'C' + (2b'+a')\gamma - 4a'\}]. \tag{19}$$

Appropriate measures of non-random association between linked sites other than σ_d^2 were considered by Hill (1975) and Ohta (1980), independently. They are

$$\sum_{i_p=1}^{K_p} \sum_{i_q=1}^{K_q} E\{X_{(i_p,i_q)}\}/\{(1-F_p)(1-F_q)\} \quad \text{and} \quad F_{(p,q)} - F_p F_q.$$

In particular, the latter, which Ohta called the identity excess, is closely correlated to the expectation of total squared linkage disequilibrium, $E\{D^2_{(p,q)}\} = \Sigma_i E\{Z_i\}$. We can show that the ratio $R = E\{D^2_{(p,q)}\}/(F_{(p,q)} - F_p F_q)$ is limited to a narrow range from $\frac{5}{9}$ to 1 when recombination is absent. Although it was not verified that such a simple relationship holds for any value of C, it is expected to be true and several numerical calculations support this finding, (Table 1).

(ii) *Completely linked multiple site model* ($r > 2$, $c = 0$)

(a) *Mean homozygosity and maximum identity excess*

Hill (1974*b*) studied the linkage disequilibrium in the case of three linked loci by using a moment generating matrix. As pointed out there, however, it is not easy to treat the problem through the diffusion equation method. This is mainly because it is necessary to know all the moments of order lower than those of the quantities to be obtained while Hill's method can obtain them rather efficiently. If we want to consider many more sites simultaneously and the higher order linkage disequilibrium between them (see, for example, Franklin & Lewontin, 1970; Slatkin, 1972, for the effect of epistatic selection on the linkage disequilibrium), the computations involved become prohibitive even under the neutral mutation hypothesis. Therefore, let us consider a special case of multiple site models.

When the sites are completely linked, the diffusion operator (6) reduces to

$$L_r = \sum_i \left\{ \sum_{p=1}^{r} \theta_p x_{i,p} - \sum_{p=1}^{r} K_p \theta_p x_i \right\} \frac{\partial}{\partial x_i} + \sum_i \sum_{i'} x_i(\delta_{ii'} - x_{i'}) \frac{\partial^2}{\partial x_i \, \partial x_{i'}}. \tag{20}$$

We first study (7) in the case of $f = \Sigma_i x_i^2$. Let us denote by $F_{(1,2,\ldots,r)}$ the expectation of $\Sigma_i x_i^2$ and by $\mathrm{F}_{(1,2,\ldots,\underset{\wedge}{p}\ldots,r)}$ that of $\Sigma_{i_1} \ldots \Sigma_{i_{p-1}} \Sigma_{i_{p+1}} \ldots \Sigma_{i_r} x_{i,p}^2$. Then, we have the recurrence equation

$$\frac{dF_{(1,2,\ldots,r)}}{dT} = 2\left\{1 + \sum_{p=1}^{r} \theta_p F_{(1,2,\ldots \underset{\wedge}{p} \ldots,r)} - \left(1 + \sum_{p=1}^{r} K_p \theta_p\right) F_{(1,2,\ldots,r)}\right\}, \tag{21}$$

and at equilibrium

$$F_{(1,2,\ldots,r)} = \frac{1 + \sum\limits_{p=1}^{r} \theta_p F_{(1,2,\ldots \underset{\wedge}{p} \ldots,r)}}{1 + \sum\limits_{p=1}^{r} K_p \theta_p}. \tag{22}$$

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

We can determine $F_{(1,2,\ldots,r)}$ from the equations of lower order than r successively. In the case of $K_p = K$ and $\theta_p = \theta$, (22) can be represented by a simple formula,

$$F_{(1,2,\ldots,r)} = a(r)+b(r)\,a(r-1)+b(r)\,b(r-1)\,a(r-2)+\ldots+b(r)\,b(r-1)\ldots b(1) \tag{23}$$

where $$a(r) = \frac{1}{1+Kr\theta} \quad \text{and} \quad b(r) = \frac{r\theta}{1+Kr\theta} \quad \text{for} \quad r \geqslant 1.$$

Using (22) or (23), we can estimate the maximum identity excess for r sites, Δ_r, as

$$\Delta_r = F_{(1,2,\ldots,r)} - \prod_{p=1}^{r} F_p. \tag{24}$$

Table 2. *Dependence of the values of* $F_r = F_{(1,2,\ldots,r)}$, H_r *and the identity excess on* r *obtained from* (23), (24) *and* (25), *assuming that* $K = 4$ *and* $v = 10^{-8}$ *at all sites*

$4N_e$	r	1	2	3	4	6	10	100	300
10^5	F_r	0·9990	0·9980	0·9970	0·9960	0·9940	0·9901	0·9091	0·7693
	H_r	0·9990	0·9980	0·9970	0·9960	0·9940	0·9901	0·9091	0·7692
	Δ_r	0	$9·9 \times 10^{-7}$	$3·0 \times 10^{-6}$	$5·9 \times 10^{-6}$	$1·5 \times 10^{-5}$	$4·4 \times 10^{-5}$	0·0042	0·0283
10^6	F_r	0·9901	0·9805	0·9710	0·9617	0·9436	0·9093	0·5004	0·2502
	H_r	0·9901	0·9804	0·9709	0·9615	0·9434	0·9091	0·5000	0·2500
	Δ_r	0	$9·5 \times 10^{-5}$	$2·8 \times 10^{-4}$	$5·5 \times 10^{-4}$	0·0013	0·0038	0·1295	0·1991
10^7	F_r	0·9118	0·8375	0·7741	0·7195	0·6302	0·5044	0·0912	0·0323
	H_r	0·9118	0·8333	0·7692	0·7143	0·6250	0·5000	0·0909	0·0323
	Δ_r	0	0·0061	0·0161	0·0284	0·0557	0·1074	0·0911	0·0323

Of particular interest is the relationship between formula (23) and that predicted by the conventional infinite allele model of Kimura & Crow (1964) which has been extensively studied and used from various statistical points of view. When a gene contains r selectively neutral nucleotide sites, the expected homozygosity is

$$H_r = \frac{1}{1+4N_e vr} \tag{25}$$

at equilibrium based on the infinite allele model where v is the mutation rate per site. This can be contrasted with (23) for $K = 4$ since each site consists of four kinds of nucleotide bases, (see Table 2 and Discussion).

(*b*) *Genetic distance and evolutionary distance*

Let us consider the genetic distance and evolutionary distance between related species. As shown in Table 2, the infinite allele model can predict the amount of genetic variability at the nucleotide level fairly well, and therefore we may expect that the genetic distance and evolutionary distance based on this model also provide good approximations to nucleotide differences between related species. However, the reality of the infinite allele model clearly depends on the number

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

of nucleotide sites in question since each site can take only four states. Furthermore, convenient detection of genetic variation within and between species is now made by using various species of restriction enzymes. Most enzymes recognize 4 to 6 nucleotides in a DNA segment and cleave them. Therefore, to analyse the data observed by such a technique, it is necessary that we construct a theory incorporating the real situation of the genome.

Following Nei (1972), suppose that a population splits into two isolated populations and thereafter no migration occurs between them, and assume that the effective sizes of the two populations are equal and constant in time. An appropriate diffusion operator denoted by T_r is then given by

$$T_r = \theta\sum_i\left\{\sum_{p=1}^{r} x_{i,p} - Krx_i\right\}\frac{\partial}{\partial x_i} + \theta\sum_i\left\{\sum_{p=1}^{r} y_{i,p} - Kry_i\right\}\frac{\partial}{\partial y_i}$$
$$+\sum_i\sum_{i'} x_i(\delta_{ii'}-x_{i'})\frac{\partial^2}{\partial x_i\,\partial x_{i'}} + \sum_i\sum_{i'} y_i(\delta_{ii'}-y_{i'})\frac{\partial^2}{\partial y_i\partial y_{i'}}. \quad (26)$$

In (26), we assumed that θ_p and K_p are constant and equal to θ and K, respectively. The frequency of A_i in one population is denoted by x_i and that of A_i in the other population by y_i. The subscript p in x_i and y_i has the same meaning as in (1). The probability of identity of homologous DNA segments with r sites sampled from two populations is $j_r = \Sigma_i x_i y_i$. From (7) and (26), we have the recurrence equation for $J_r = E\{j_r\}$

$$\frac{dJ_r}{dT} = 2\theta r(J_{r-1} - KJ_r) \quad (27)$$

for $r \geqslant 1$ under initial conditions for $J_p(0)$, $(p = 1, 2, \ldots, r)$ and a boundary condition of $J_0(t) = 1$. Noting that the equilibrium value of $J_p(\infty)$ is equal to $\frac{1}{K^p}$ and that $T = \frac{t}{4N_e}$, we have as the solution for (27)

$$J_r(t) = \frac{1}{K^r}\sum_{p=0}^{r}\left[{}_rC_p\left\{\sum_{q=0}^{p} {}_pC_q(-1)^q K^{p-q} J_{p-q}(0)\right\}e^{-2pKvt/(K-1)}\right], \quad (28)$$

where ${}_rC_p$ is the binomial coefficient. If the probability of identity within a population is constant in time, $J_r(0)$ is given by (23) in both populations. Thus, the normalized identity is obtained by dividing (28) by $J_r(0)$ or $F_{(1,2,\ldots,r)}$, and the genetic distance becomes

$$G_r = -\log\left(J_r(t)/J_r(0)\right). \quad (29)$$

In particular, as K becomes indefinitely large $J_r(t) = J_r(0)\,e^{-2rvt}$ so that $G_r = 2rvt$. This is equivalent to the formula originally demonstrated by Nei (1972). However, if we regard a 'site' as a nucleotide site, the above limit has no biological meaning. Thus the genetic distance in terms of nucleotide differences must be calculated by (29), substituting (23) for $J_r(0)$ and (28) for $J_r(t)$, respectively (Table 3).

On the other hand, if the initial population is monomorphic, i.e. $J_p(0) = 1$ ($p = 1, 2, \ldots, r$) (28) reduces to

$$J_r(t) = \left\{\frac{1}{K}+\left(1-\frac{1}{K}\right)e^{-2Kvt/(K-1)}\right\}^r \tag{30}$$

and the evolutionary distance K_{nuc} defined by $2vt$ can be solved as

$$K_{nuc} = -\frac{K-1}{K}\log\left(\frac{KJ_r^{1/r}(t)-1}{K-1}\right). \tag{31}$$

Formula (31) is equivalent to that of the evolutionary distance given by Jukes & Cantor (1969) if we set $r = 1$ and $K = 4$, and that devised for data obtained by restriction enzymes (Aoki, Tateno & Takahata, 1981).

Table 3. *Relationship between the degree of polymorphism, F_r, the evolutionary distance, $K_{nuc} = 2vt$ and the genetic distance, G_r*

(The degree of polymorphism is assumed constant in time since the divergence of two isolated populations.)

$4N_e v$	$K_{nuc} = 2vt$	0·0001	0·001	0·01	0·02	0·04	0·1	0·2
0 ($F_4 = 1$)	J_4	0·9996	0·9960	0·9609	0·9234	0·8531	0·6749	0·4620
	G_4	0·0004	0·0040	0·0399	0·0797	0·1589	0·3932	0·7722
10^{-3} ($F_4 = 0·9960$)	J_4	0·9956	0·9920	0·9570	0·9197	0·8497	0·6723	0·4603
	G_4	0·0004	0·0040	0·0399	0·0793	0·1586	0·3930	0·7719
10^{-2} ($F_4 = 0·9617$)	J_4	0·9613	0·9578	0·9241	0·8882	0·8208	0·6499	0·4455
	G_4	0·0004	0·0041	0·0399	0·0795	0·1584	0·3919	0·7695
0 ($F_6 = 1$)	J_6	0·9994	0·9940	0·9419	0·8873	0·8360	0·5544	0·3140
	G_6	0·0006	0·0060	0·0599	0·1196	0·1791	0·5899	1·158
10^{-3} ($F_6 = 0·9940$)	J_6	0·9934	0·9881	0·9363	0·8820	0·8311	0·5513	0·3123
	G_6	0·0006	0·0060	0·0598	0·1195	0·1790	0·5895	1·160
10^{-2} ($F_6 = 0·9436$)	J_6	0·9430	0·9379	0·8889	0·8375	0·7893	0·5242	0·2975
	G_6	0·0006	0·0061	0·0597	0·1193	0·1786	0·5878	1·154

3. DISCUSSION

Linkage disequilibrium

The two site (locus) multiple state (allele) model was studied to reveal the relationship between the previous work of Ohta & Kimura (1969*b*) and Hill (1975) by using the diffusion equation method. The present formulae are similar to those derived in Ohta & Kimura but nevertheless different even if we substitute $v_p/(K_p-1)$ ($p = 1$ and 2) for the backward mutation rates in their formulae. This is because the quantity given in (14) depends upon the number of possible states (or alleles). One exception is, however, the standard linkage disequilibrium (16) which can be obtained directly from their formula through the above substitution.

In contrast, we can easily obtain the formulae analogous to (10) of Hill (1975) if we take the limit of K_p to infinity in (15) after summing over all the possible states. The slight difference is due to the different methods used, but the agreement between the two sets of values is again satisfactory as pointed out in the case of the two locus two allele model (Ohta & Kimura, 1969*b*). Formulae are tabulated in Table 1 to show their K_p dependence, assuming that the values of $4N_e v$ and $4N_e c$ are given. We can see that unless $4N_e v$ is very small, the total squared linkage disequilibrium $\Sigma_i E\{Z_i\}$ increases as $K(K_1 = K_2)$ becomes large particularly when linkage is weak, whereas $\Sigma_i E\{X_i\}$ is rather insensitive to changes of K. As a result or directly from (16), σ_d^2 increases with K. A most marked difference between the values of σ_d^2 for $K = 2$ and for $K = 10^3$ is shown in the case of $4N_e v = 0{\cdot}2$ and $4N_e c = 0{\cdot}0$. The latter is therein about 1·17 times larger than the former. (Note here that the looser the linkage, the smaller the discrepancy between the above two opposite cases.) In other words, recombination can, in general, break down linkage disequilibrium more easily when K is greater than 2 than in the case of $K = 2$. This finding is consistent with our intuition though the difference is not large.

In Table 1, the ratio of the total squared linkage disequilibrium to the identity excess, R, is also presented. Most values are very close to $\frac{5}{9}$ or a little greater, irrespective of the value of K and $4N_e c$. In addition, preliminary Monte Carlo experiments indicate that the above relationship holds true even under a multiple site model. Therefore, we can conclude that the identity excess is closely correlated to the total squared linkage disequilibrium and this in turn suggests that we can use it as an appropriate measure for linkage disequilibrium.

Next, let us examine the results for the completely linked multiple site model in relation to the infinite site and infinite allele models (Kimura & Crow, 1964; Kimura, 1969; Watterson, 1975; Li, 1977; Ewens, 1979 and others). If we regard a 'site' as a nucleotide site in a cistron or a small part of genome, then the assumption of complete linkage between sites is realistic but if a 'site' is referred to as a gene, then the assumption may be unrealistic. Henceforth we will use the term to indicate a nucleotide site and therefore K_p can be regarded as 4 for any site. Watterson (1975) studied the distribution for the number of segregating sites at stationary state using a model which is quite similar to the present one. But they are different from each other in the sense that he assumed that cistrons each contain infinitely many sites and that at each site there are only two possible nucleotides. The difference is not important when $4N_e v$ is small, although the present model is the real one. The assumption of no recombination between sites is common to our models. In this context, of particular importance is his criticism of the assumption of independence of sites employed by many authors (Watterson, 1975). The infinite allele model is more appropriate, however, when we wish to consider a certain region of the genome since it takes into account complete linkage between sites. The fact can be checked partly by comparing $F_{(1,2,\ldots,r)}$, F_r in short, in (23) with H_r in (25). From Table 2, we can see that there is no significant difference between the two without regard to the value of r when $4N_e v$ is small.

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

Since the mutation rate per site is very small, presumably 10^{-8} or less, and it is unlikely that the effective population size N_e is much greater than 10^8 in most organisms, we can expect that the infinite allele model predicts quite satisfactorily the amount of genetic variation observed at the nucleotide level. This fact *does not* necessarily rule out the possibility that some other quantities may depend heavily on the number of sites in question.

Table 2 also contains various values of identity excess, which tends to be equal to F_r as r increases since the product of identity probability per site, F_1^r, decreases rapidly. Furthermore, the identity excess provides an upper limit to the expected total squared linkage disequilibrium in r linked sites under the neutral mutation hypothesis. Clearly, the assumption of independence of sites is not warranted in a situation where the identity excess differs significantly from zero.

Genetic distance

Usually, the genetic distance of Nei (1972) and others is applied to allele frequencies in related species observed by electrophoresis while the evolutionary distance is calculated by comparison of homologous nucleotide sequences. The former is based on polymorphism in both populations, whereas the latter tacitly assumes monomorphism or ignores polymorphism by regarding the two nucleotide sequences, each randomly sampled from individuals, as representative for each species (Jukes & Cantor, 1969; Kimura, 1980, 1981; Takahata & Kimura, 1981). Therefore, unless we assume the infinite allele model, as does Nei (1972), both distances do not coincide and the relationship seems rather obscure since one might suspect that they are different measures based on different assumptions. In addition, several authors have recently devised evolutionary distance for data obtained from application of restriction enzymes (Upholt, 1977; Nei & Li, 1979). In this case, new models incorporating the real situation for DNA sequence are required to study the distance because we are usually dealing with at most a hexanucleotide.

Although the above measures were studied independently according to the characteristics of methodology used, they can be represented by a single formulation. To do so, we make use of the definition of genetic distance given by Nei (1972) based on the present multiple site model. The identity probability between the two isolated populations is a key quantity to be formulated. The formula at any time is given in (28) in terms of the initial condition, the number of sites and the number of states per site. As mentioned earlier, (28) reduces to the formula obtained by Nei as an increase of the number of states while it is equivalent to those obtained by Jukes & Cantor (1969) and Aoki *et al.* if we assume that the initial population before splitting was monomorphic, and that $K = 4$. Thus, all measures concerning distances can be connected with each other through (28).

Assuming that $J_r(0)$ is equal to the expected identity probability at equilibrium, F_r, the theoretical relationship between the genetic distance (29) and the evolutionary distance $K_{nuc} = 2vt$ for $r = 4$ and 6 is given in Table 3. First, when K_{nuc}

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

is small, the genetic distance G_r is approximately equal to rK_{nuc} which is the expected value from the infinite allele model. In addition, G_r is surprisingly stable to the change of the degree of polymorphism. In other words, for a given value of K_{nuc} it can predict almost the same value even if F_r is greatly changed, and this is a desirable property for a measure of evolutionary distance. But G_r is, in general, smaller than rK_{nuc} because of back mutations at each site, the extent of which depends not only on K_{nuc} but also on F_r.

Where we want to estimate the distances from the values of F_r and J_r, Table 3 can be looked at slightly differently. We note that the estimated distances are strongly affected by the level of polymorphism within populations when two populations compared are closely related. For example, for a given value of J_4, say 0·961, we would estimate K_{nuc} (G_4) as 0·01 (0·04) if we assume that populations have been monomorphic. In contrast, for the same value of J_4 we would estimate K_{nuc} (G_4) as 0·0001 (0·0004) if we take some polymorphism into account, say $F_4 = 0{\cdot}961$. Thus, it is very important to know how to incorporate the degree of polymorphism into the theory correctly, otherwise we will grossly overestimate the distances. After there have been many nucleotide substitutions, however, the problem of polymorphism is, in accord with intuition, not critical.

The restriction enzyme technique is now applied mainly to mitochondrial DNAs in closely related populations, and Brown (1980) revealed that human mitochondrial DNA is polymorphic even within a race (see also Avise *et al.* 1979*a*, *b* and Shah & Langley, 1979). For a theoretical analysis of the extent of genetic variability maintained in extranuclear genome, see Takahata & Maruyama (1981). Although no observation of this kind has yet been made with respect to nuclear DNA, we can expect the same situation as in mitochondrial DNA, unless populations are quite small, and apply the present formulae to such data.

In conclusion, the present model is general for linked genes or a part of the genome and has the possibility that various problems can be formulated in a unified way. Although in the future a slight modification may be necessary with respect to the mutation scheme or crossing over, the simple multiple site multiple state model treated here is worth detailed investigation from various points of view.

I would like to thank an anonymous referee for his helpful comments and Dr R. C. Griffiths for sending me a manuscript in which he has analysed the neutral two locus multiple allele model. This research was supported by Grant 439016 from the Ministry of Education, Japan.

REFERENCES

Aoki, K., Tateno, Y. & Takahata, N. (1981). Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary G+C content. *Journal of Molecular Evolution*. (In the Press.)

Avise, J. C., Lansman, R. A. & Shade, R. O. (1979*a*). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *peromyscus*. *Genetics* **92**, 279–295.

Avise, J. C., Giblin-Davidson, C., Laerm, J., Patton, J. C. & Lansman, R. A. (1979*b*). Mitochondrial DNA clones and matriarchal phylogeny within the among geographic popu-

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

lations of the pocket gopher. *Geomys pinetis. Proceedings of the National Academy of Sciences, U.S.A.* **76**, 6694–6698.

Brown, W. M. (1980). Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 3605–3609.

Ewens, W. J. (1979). *Mathematical Population Genetics.* Berlin, Heidelberg, New York: Springer-Verlag.

Franklin, I. & Lewontin, R. C. (1970). Is the gene the unit of selection? *Genetics* **65**, 701–734.

Haldane, J. B. S. (1931). A mathematical theory of natural and artificial selection. viii. Metastable populations. *Proceedings of the Cambridge Philosophical Society* **27**, 137–142.

Hill, W. G. (1968). Population dynamics of linked genes in finite populations. *Proceedings of the XII International Congress of Genetics* **2**, 146–147.

Hill, W. G. (1974*a*). Disequilibrium among several linked neutral genes in finite population. I. Mean changes in disequilibrium. *Theoretical Population Biology* **5**, 366–392.

Hill, W. G. (1974*b*). Disequilibrium among several linked neutral genes in finite population. II. Variances and covariances of disequilibria. *Theoretical Population Biology* **6**, 184–198.

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics (Der Züchter)* **38**, 226–231.

Jukes, T. H. & Cantor, C. H. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, (ed. H. N. Munro), pp. 21–123. New York: Academic Press.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232.

Kimura, M. 1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Kimura, M. (1981). On estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 454–458.

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

Li, W.-H. (1977). Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* **85**, 331–337.

Maxam, A. & Gilbert, M. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 560–564.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.

Nei, M. & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 5269–5273.

Ohta, T. (1980). Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. *Genetical Research* **36**, 181–197.

Ohta, T. & Kimura, M. (1969*a*). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

Ohta, T. & Kimura, M. (1969*b*). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**, 229–238.

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.

Robertson, A. (1961). Inbreeding in artificial selection programmes. *Genetical Research* **2**, 189–194.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science, U.S.A.* **74**, 5463–5467.

Shah, D. H. & Langley, C. H. (1979). Inter- and intraspecific variation in restriction maps of *Drosophila* mitochondrial DNAs. *Nature* **281**, 696–699.

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press

SLATKIN, M. (1972). On treating the chromosome as the unit of selection. *Genetics* **72**, 157–168.

TAKAHATA, N. & KIMURA, M. (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*. (In the Press.)

TAKAHATA, N. & MARUYAMA, T. (1981). A mathematical model of extranuclear genes and the genetic variability maintained in a finite population. *Genetical Research* **37**, 291–302.

UPHOLT, W. B. (1977). Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucleic Acids Research* **4**, 1257–1265.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

https://doi.org/10.1017/S0016672300020747 Published online by Cambridge University Press