A MIGRATION MATRIX MODEL FOR THE STUDY OF RANDOM GENETIC DRIFT

WALTER F. BODMER AND LUIGI L. CAVALLI-SFORZA

Department of Genetics, Stanford Medical Center, Palo Alto, California, and International Laboratory of Genetics and Biophysics, Pavia Section, c/o Institute of Genetics, The University of Pavia, Italy

Received December 18, 1967

THE buffering effect of cross migration on random genetic drift in partially isolated populations was first studied in a very simple model which SEWALL WRIGHT called the "island model" (1943, 1951). In this model every population receives a fraction m per generation of its genes from a common gene pool with a constant gene frequency. The name "island model" refers to the fact that the exchange among islands is effectively independent of the distance between them. In other words, every island exchanges genes equally with every other island, thus simulating immigration from a common and constant gene pool. The inadequacy of this model for the representation of variation due to geographic, or other sources of, isolation soon became apparent. SEWALL WRIGHT later suggested another model known as "isolation by distance" (1943, 1946, 1951). The population is then uniformly distributed and individual mobility is defined by a continuous distribution. A normal distribution of mobility is often used for simplicity though other distributions have also been considered. Two forms of the model were analyzed, one for geographic distribution of the population along a line, and the other on a surface. Results were given in terms of the mobility (as measured by the appropriate parameter or parameters of the mobility distribution) and of the population density per unit of length or surface area. These two quantities can be conveniently condensed into a single parameter, the "neighborhood size" which is sometimes sufficient to describe the situation. Similar continuous models have also been studied by MALÉCOT (1945, 1966) who has shown that, under the above assumptions, the correlation of gene frequencies in two groups separated by a distance, r, decreases exponentially with r.

An interesting and important class of models for the study of isolation by distance has been put forward by KIMURA and WEISS (1964) and by MALÉCOT (1950, 1962). These authors assume that the population is distributed discontinuously, in "colonies" which are of equal size at the points of an infinite regular lattice, and that there is a given rate of exchange between colonies i steps apart (the "stepping stone model"). In the simplest linear version of this model each colony has two neighbors with which it exchanges a fraction m of its gametes, $m/2$ to each. In the simplest two-dimensional case, the colony has four neighbors with each of which it exchanges $m/4$ of its genes. There are six neighbors in the three-dimensional version of the model.

The mathematical models analyzed by MALÉCOT and KIMURA differ slightly and give rise to slightly different predictions. These theories predict the expected equilibrium variance among an infinite number of *unrelated* colonies, and the correlations of gene frequencies between colonies as a function of their distance apart. The unit distance is defined as that between two neighboring colonies. Results are generally available for equilibrium conditions only and depend on three parameters, in addition to the number of dimensions: (i) the size of each individual colony; (ii) the migratory exchange between neighboring colonies, and (iii) a parameter (m_∞ for KIMURA, k for MALÉCOT) which measures collectively all constant stablizing factors such as recurrent mutation, "linearized" selection for the heterozygote, migration from a hypothetical constant external reservoir, etc.

The two types of models of isolation by distance, the continuous and discontinuous ones, suffer from several limitations. Neither can cope with the fact that real populations are almost always very irregular in their geographic distribution. Population size, density and mobility are not constant with respect to space and time. It is also usually difficult, in practice, to represent observed mobility in any given population by the functions which these theories employ.

In the present paper we develop a theory aimed at predicting the amount of variation to be expected between gene frequencies of a *finite* number of colonies of different sizes and the correlations between them, with a general migration pattern given in the form of a "migration matrix." The predictions are, of course, limited by the approximations made to obtain manageable results and by the fact that only one source of random variation is considered, namely that due to Mendelian segregation in a finite population. The major advantage to using migration matrices is that it makes it possible to use observed migration data in the model, without trying to force the migration pattern into a somewhat inflexible and usually inappropriate model. An approach to predicting the diffusion of consanguineous individuals and their probability of marriage, using migration matrices, has already been described by CAVALLI-SFORZA, KIMURA and BARRAI (1960).

THE MIGRATION MATRIX

A migration matriv is a formal representation of the displacement over one generation among a set of k colonies, villages or, more generally, groups of individuals of one species. The main requirement for the definition of a group to be useful is that it should have sufficient stability to be recognizable from one generation to the next. The matrix has k rows and columns, which correspond to the k colonies forming the population under investigation, arranged in an arbitrary order. Any one row and column corresponds to just one colony. The elements of an observed matrix are the numbers of individuals born in the i^{th} colony from parents born in the j^{th} colony (see the numerical example in Table 1).

Migration matrices obtained separately from fathers and mothers generally are different. For the present purpose one can sum or average data from fathers and mothers, as was done for Table 1C (see Table 1A and B). It can be proved that,

TABLE 1

A numerical example of an observed migration matrix and its forward and backward stochastic versions

A. Father-offspring
Birthplace of father

Birthplace of progeny	1	2	3	4	Total
1	170	3	12	3	188
2	2	106	4	2	114
3	8	0	158	5	171
4	0	1	3	181	185
Total	180	110	177	191	658

B. Mother-offspring
Birthplace of mother

Birthplace of progeny	1	2	3	4	Total
1	174	3	10	3	190
2	2	108	0	2	112
3	3	1	168	3	175
4	1	2	5	175	183
Total	180	114	183	183	660

C. Parent-offspring
Birthplace of parents

Birthplace of progeny	1	2	3	4	Total
1	344	6	22	6	378
2	4	214	4	4	226
3	11	1	326	8	346
4	1	3	8	356	368
Total	360	224	360	374	1318

D. Backward matrix
Birthplace of parents

Birthplace of progeny	1	2	3	4	Total
1	.9100	.0159	.0582	.0150	1.0000
2	.0177	.9469	.0177	.0177	1.0000
3	.0318	.0029	.9422	.0231	1.0000
4	.0027	.0082	.0217	.9674	1.0000

E. Forward matrix
Birthplace of parents

Birthplace of progeny	1	2	3	4
1	.9556	.0268	.0611	.0160
2	.0111	.9553	.0111	.0107
3	.0305	.0045	.9056	.0214
4	.0028	.0134	.0222	.9519
Total	1.0000	1.0000	1.0000	1.0000

Migration matrices are for the Lecce Province. For simplification original values for the single villages were condensed into four areas. Matrices A, B, C, indicate numbers of individuals born in the *i*th area to parents born in the *j*th area. Matrix C is the sum of A and B. Matrices D and E are the stochastic matrices derived from C. Data from MODIANO *et al.* (1965).

if the linear stabilizing factors (see below) are the same in both sexes, the average matrix is a very good representation of the total migration process (F. SCUDO, personal communication). It is also possible to use matrices in which rows and columns represent birth places of husband and wife, if this is the only type of data available. In this case, however, a correction is necessary and the resulting matrix is only approximate.

The observed matrix $\overline{M}_{ij}$ consists of the numbers of individuals which have been examined, classified according to birthplace of parent and birthplace of offspring. From such a matrix it is possible to form (see Table 1D, E) two stochastic matrices. A stochastic matrix has rows or columns summing to one, so that the elements can represent transition probabilities from one generation to the next. Matrices D and E of Table 1 are obtained by making rows and columns of matrix C, respectively, sum to one. We shall call them the "forward" and the "backward" matrices depending on whether the elements which sum to 1 are the columns or the rows. This definition refers to whether the transition is in the same direction as the time or opposite to it. In fact the terms of the forward migration matrix M^*_{ij} represent the probabilities that the children of parents born in colony *j go to* colony *i*, while in the backward migration matrix, the terms

are the probabilities M_{ij} that the children born in the i^{th} colony *come from* parents born in the j^{th} colony.

The two matrices are different unless the matrix $\overline{M}_{ij}$ is symmetrical. If there are no significant differences between the symmetrical elements of $\overline{M}_{ij}$ it is convenient to average them.

The prediction of the numerical composition of a population after t generations of migration depends on the forward matrix. If the initial distribution individuals at time t = 0 in colonies 1, 2, . . . , k, is $\boldsymbol{n}_0 = (n_1, n_2, \ldots, n_k)$ and they produce children who join the i^{th} colony with the probabilities given by the forward matrix, the expected composition of the next generation is given by

$$\boldsymbol{n}_1 = \boldsymbol{n}_0 M^* \tag{2.1}$$

Provided the migration matrix is constant, the expected composition of the population after t generations is

$$\boldsymbol{n}_t = \boldsymbol{n}_0 (M^*)^t \tag{2.2}$$

by simple iteration of Equation (2.1).

When t is sufficiently large, n_t, in general, approaches an equilibrium value as the product $(M^*)^t$ converges to a fixed matrix. The equilibrium composition is then independent of the initial state of the population and so can be predicted by powering the forward migration matrix and then pre-multiplying it by an arbitrary vector.

In order to predict the variation in gene frequencies within and between colonies, the backward matrix has to be used. The forward matrix was discussed because of its general relevance to our model. We may wish to test, for example, whether we can assume that the migration matrix estimated from the most recent generation is valid for a longer period of time. If the present relative proportions of sizes of colonies differ significantly from those predicted at equilibrium by this recent migration matrix, then the matrix must have differed in the more distant past from its present form.

Let us suppose that a pair of alleles A, a is segregating in each colony. Then the deterministic change in gene frequencies as a consequence of migration is given in terms of the gene frequencies of the A gene in colony i at the t^{th} generation, $p_i^{(t)}$, by

$$p_i^{(t+1)} = \sum_{j=1}^{k} p_j^{(t)} M_{ij} \tag{2.3}$$

where M_{ij} is the ij^{th} element of the backward migration matrix. This follows from the fact that the expected *number* of A genes in colony i at time t is

$$\sum_{j=1}^{k} M^*_{ij} n_j^{(t)} p_j^{(t)}$$

where M^*_{ij} is the forward matrix and $n_j^{(t)}$ is the number of individuals in colony j at time t. Thus the proportion of A genes in colony i at time $t+1$ is as given in Equation (2.3) where the terms of the backward migration matrix are given by

$$M_{ij} = \frac{M^*_{ij} n_j^{(t)}}{\sum_{j=1}^{k} M^*_{ij} n_j^{(t)}}. \tag{2.4}$$

The stochastic model we will use assumes that random sampling of genes, leading to random genetic drift, takes place for every colony at every generation after deterministic migration. The expected proportion of A genes is then computed deterministically following Equation (2.3). Colony sizes may vary but will usually, though not necessarily, be considered constant in time for any given colony.

The majority of areas investigated in practice are not totally closed to immigration from the outside. In our model the numbers of individuals, or genes, per colony and thus the total number also, are finite. In the absence of other stabilizing forces all alleles would, therefore, eventually always be fixed or lost. However, in the presence of the cumulative forces of recurrent mutation in either direction, migration from the outside and stabilizing selection, genes will not generally become fixed and the variances and covariances tend to non-trivial, finite values.

The stabilizing linear pressures of mutation, migration and (linear) selection can formally be introduced, in our model, in terms of a single parameter which we call "migration from the outside". The i^{th} colony has a fraction α_i of its genes coming from an external gene pool which may have, if desired, a characteristic gene frequency x_i. Since the calculation of the expected gene frequencies after migration involves the use of the backward migration matrix, the observed counterpart of the α_i are the frequencies of individuals in the i^{th} colony whose parents are born outside the total area investigated. In man, who is perhaps the organism most readily amenable to the study of migration, the simplest way of selecting individuals for such a study is to take residents born in the area and ask them their parents' birthplaces. This generates an extra column of M values from which α_i values can be computed.

The method developed in this paper is intended for application to observed migration matrices. However, for illustrative purposes and also for a comparison with other models, in particular the island and stepping stone models, it is of some interest to consider some simple hypothetical migration matrices. Numerical and analytical results for a collection of such matrices are discussed in the remaining sections of this paper.

MIGRATION MATRIX DRIFT THEORY

3.1. *Formulation of the mathematical model:* Suppose we have a group of k colonies, all segregating for two alleles A, a at a given locus and assume time is discrete, being measured in generations. We assume that mating is at random within each colony and that differential selection and other stabilizing factors affecting the alleles A and a act linearily. Let $p_i^{(n)}$ be the gene frequency of A in the i^{th} villages at the n^{th} generation, and let N_i be the population size of the i^{th} colony, assumed constant, where $i = 1, \ldots, k$. Let $[M_{ij}]$ be the backward migration matrix, such that M_{ij} is the proportion of individuals in colony i in the present generation who came from colony j, and so $\sum_{j=1}^{k} M_{ij} = 1$ (see Table 1). We assume in addition that, in each generation, a proportion α_i of the individuals

in the i^{th} colony came in from an external population with a constant A gene frequency of x_i. The parameters α_i and x_i represent the total effect of all the stabilizing factors taken into account by the model. The expected A gene frequencies in the n^{th} generation in terms of those in the previous generation are then given by

$$p_i^{(n)} = \sum_{j=1}^{k} (1-\alpha_i)\, M_{ij}\, p_j^{(n-1)} + \alpha_i\, x_i, \qquad \text{for } i = 1, \ldots, k, \tag{3.1}$$

as discussed in the previous section. A similar model has been formulated by MALÉCOT (1951). This equation is also analogous to the equations developed by MALÉCOT (1962) and KIMURA and WEISS (1964) for their migration models. To take account of random sampling variations from generation to generation (random genetic drift) we assume that the realized gene frequency in the i^{th} colony in the n^{th} generation is the result of a binomial sample of size $2N_i$ (the number of genes in the population) with parameter $p_i^{(n)}$, as given by equation (3.1). The $p_j^{(n-1)}$ used to obtain the expected frequencies in the n^{th} generation are the realized gene frequencies in the n-1^{th} generation. Thus we have

$$\begin{aligned} &\text{(a)}\quad E(p_i^{(n)} | p_j^{(n-1)}, j = 1, \ldots, k) = \sum_{j=1}^{k} m_{ij}\, p_j^{(n-1)} + \alpha_i x_i = P_i^{(n)}, \text{ say} \\ &\text{(b)}\quad \mathrm{V}(p_i^{(n)} | p_j^{(n-1)}, j = 1, \ldots, k) = \frac{1}{2N_i} P_i^{(n)} (1-P_i^{(n)}) \end{aligned} \tag{3.2}$$

for $i = 1, \ldots, k$, where $m_{ij} = (1-\alpha_i)\, M_{ij}$. This model assumes that migration occurs deterministically before mating and that population sizes remain constant, so that the only component of random drift taken into account is that due to genetic segregation in colonies with finite population sizes. When $M_{ij} = 0$ for all $i \neq j$, so that the only migration into any colony is from the general population, the model reduces to WRIGHT's "island model". We are interested in the variances and covariances of the gene frequencies in the n^{th} generation in terms of the initial gene frequencies.

3.2. *Use of the angular transformation for the derivation of the approximate variances and covariances of the gene frequencies:* Following FISHER and FORD (1947) and BODMER (1960), consider the application of the angular transformation $p = \sin^2 \theta$ or $\theta = \text{arc sin } \sqrt{p}$. It is well known that if $\tilde{p}$ is a binomial sample based on n observations, with expectation p and $\tilde{p} = \sin^2 \tilde{\theta}$, then $E(\tilde{\theta}) = \theta + 0(1/n)$ and $\mathrm{V}(\tilde{\theta}) = C/n + 0(1/n^2)$, which is effectively independent of θ and so of p, so long as p is not near 0 or 1. Here $C = 1/4$ if angles are measured in radians and $(90/\pi)^2$ if angles are measured in degrees. Let, now, $p_i^{(n)} = \sin^2 \theta_i^{(n)}$ and $P_i^{(n)} = \sin^2 \psi_i^{(n)}$ then

$$\begin{aligned} &\text{(a)}\quad E(\theta_i^{(n)} | \theta_j^{(n-1)}, j = 1, \ldots, k) = \psi_i^{(n)} + 0(1/N_i), \\ &\text{(b)}\quad \mathrm{V}(\theta_i^{(n)} | \theta_j^{(n-1)}, j = 1, \ldots, k) = C/2N_i + 0(1/N_i^2). \end{aligned} \tag{3.3}$$

If further $x_i = \sin^2 \eta_i$, then from equation (3.2a)

$$\sin^2 \psi_i^{(n)} = \Sigma\, m_{ij} \sin^2 \theta_j^{(n-1)} + \alpha_i \sin^2 \eta_i$$

or, since

$$\sin^2 \theta = 1/2\, [1 + 2\, (\theta - \pi/4) + 0(\theta - \pi/4)^3] = \theta + 1/2 - \pi/4 + 0(\theta - \pi/4)^3$$

when $\theta - \pi/4$ is small,

$$\psi_i^{(n)} = \Sigma\, m_{ij}\theta_j^{(n-1)} + \alpha_i\eta_i + 0(\theta - \pi/4)^3 . \tag{3.4}$$

This equation (3.4), is a good approximation to equation (3.2a) so long as all the angular transforms are near $\pi/4$, or equivalently, so long as all gene frequencies are near 1/2. The key to the usefulness of the angular transformation is that the variances of the angular values are dependent only on N_i and not on the angles themselves. The dependence of the covariance of the gene frequencies on the frequencies, and so on the total previous history of the process, is the major block to a simple analytical solution of the process defined by equation (3.2). In assuming the validity of the angular transformation throughout the whole process we are assuming that the stochastic variation is such that gene frequencies are never sufficiently far from ½ for the error terms in equations (3.3) and (3.4) to accumulate significantly. Numerical calculations (see below) suggest that this is reasonable as long as gene frequencies are, in fact, not very close to 0 or 1. KIMURA and WEISS' (1964) treatment of the stepping stone model involves a similar approximation. Thus in deriving their equation (1.3) they assume "that the product terms between $\tilde{p}$'s and ξ_i have expectation zero", where $\tilde{p}$ are the mean gene frequencies and ξ_i a random variable describing random variation from one generation to the next.

Ignoring the terms of $(0(\theta - \pi/4)^3$, equation (3.4), can be rewritten in the form

$$(\underline{\psi}^{(n)} - \underline{\xi}(I - \mathbf{m})^{-1}) = \mathbf{m}(\underline{\theta}^{(n-1)} - \underline{\xi}(I - \mathbf{m})^{-1}) \tag{3.5}$$

where $\mathbf{m}$ is the matrix (m_{ij}) and $\underline{\psi}^{(n)}$, $\underline{\theta}^{(n-1)}$, $\underline{\xi}$ refer to the vectors $\psi_i^{(n)}$, $\theta_i^{(n)}$ and $\alpha_i\,\eta_i$ respectively. If we write $\underline{\gamma} = \underline{\xi}(I - \mathbf{m})^{-1}$ and $\underline{\phi}^{(n-1)} = \underline{\theta}^{(n-1)} - \underline{\gamma}$, then

$$E(\underline{\phi}^{(n)}|\underline{\theta}^{(n-1)}) = \mathbf{m}\,\underline{\phi}^{(n-1)} . \tag{3.6}$$

Thus

$$E(\underline{\phi}^{(n)}|\underline{\theta}^{(n-2)}) = E[\mathbf{m}\,\underline{\phi}^{(n-1)}|\underline{\theta}^{(n-2)}] = \mathbf{m}^2\,\underline{\phi}^{(n-2)}$$

and so, by further iteration,

$$E[\underline{\phi}^{(n)}|\underline{\theta}^{(0)}] = \mathbf{m}^n\,\underline{\phi}^{(0)} . \tag{3.7}$$

Now, following FISHER and FORD (1947) and BODMER (1960) we can write, using equation (3.7)

$$\begin{aligned} V(\underline{\theta}_i^{(n)}|\underline{\theta}^{(0)}) &= E[(\phi_i^{(n)} - (\underline{m^n\,\phi}^{(0)})_i)^2|\underline{\theta}^{(0)}] \\ &= E[(\phi_i^{(n)} - (\mathbf{m}\,\underline{\phi}^{(n-1)})_i + (\mathbf{m}\,\underline{\phi}^{(n-1)})_i - (\mathbf{m}^2\,\underline{\phi}^{(n-2)})_i \\ &\quad + \ldots + (\mathbf{m}^{n-1}\,\underline{\phi}^{(1)})_i - (\mathbf{m}^n\,\underline{\phi}^{(0)})_i)^2|\underline{\theta}^{(0)}] \\ &= \sum_{r=1}^{n} E(V[(\mathbf{m}^{n-r}\,\underline{\phi}^{(r)})_i|\underline{\theta}^{(r-1)}]\,|\underline{\theta}_0) = \sum_{r=1}^{n} V[(\mathbf{m}^{n-r}\,\underline{\phi}^{(r)})_i|\underline{\theta}^{(r-1)}] \end{aligned} \tag{3.8}$$

This follows from equation (3.6), which gives

$$E[(\mathbf{m}^{n-r}\,\underline{\phi}^{(r)})_i|\underline{\theta}^{(r-1)}] = (\mathbf{m}^{n-r+1}\,\underline{\phi}^{(r-1)})_i$$

and from the fact that all difference pairs $(\mathbf{m}^{n-r}\,\underline{\phi}^{(r)})_i - (\mathbf{m}^{n-r+1}\,\underline{\phi}^{(r-1)})_i$ and $(\mathbf{m}^{n-s}\,\underline{\phi}^{(s)})_i - (\mathbf{m}^{n-s+1}\,\underline{\phi}^{(s-1)})_i$, $s \neq r$, are independent, because samples taken in successive generations are independent given the observed proportion in the preceding generation. For $r \neq n$

$$\begin{aligned} &V[\sum_{j=1}^{k} m_{ij}{}^{(n-r)}\,\phi_j^{(r)}|\underline{\theta}^{(r-1)}] \\ &\quad = \sum_{j=1}^{k} (m_{ij}{}^{(n-r)})^2\,V(\phi_j^{(r)}|\underline{\theta}^{(r-1)}) \end{aligned} \tag{3.9}$$

$$+\sum_{\substack{l=1\\l\neq m}}^{k}\sum_{m=1}^{k} m_{il}{}^{(n-r)}\, m_{im}{}^{(n-r)}\,\mathrm{Cov}(\phi_l{}^{(r)}, \phi_m{}^{(r)}|\underline{\theta}^{(r-1)}),$$

where $m_{ij}{}^{(r)}$ is the ij^{th} term of the matrix $\mathbf{m}^r$.

All the $\phi_j{}^{(r)}$, $j=1, \ldots, k$ are independent binomial samples given $\underline{\theta}^{(r-1)}$, the angular gene frequency transforms in the previous generation, so that all the covariance terms are zero. From the definition of $\underline{\phi}^{(r)}$ $(=\underline{\theta}^{(r)}-\underline{\gamma})$ we have

$$\mathrm{V}(\phi_i{}^{(r)}|\underline{\theta}^{(r-1)}) = V(\theta_i{}^{(r)}|\underline{\theta}^{(r-1)}) = C/2N_i + 0(1/N_i{}^2),$$

from equation (3.3b). Equation (3.9) gives, therefore, for $r \neq n$,

$$(3.10)\quad \mathrm{V}\left(\sum_{j=1}^{k} m_{ij}{}^{(n-r)}\phi_j{}^{(r)}|\underline{\theta}^{(r-1)}\right) = C/2 \sum_{j=1}^{k}\left\{\frac{[m_{ij}{}^{(n-r)}]^2}{N_j} + 0(1/N_i{}^2)\right\}.$$

When $r=n$, $\sum_{j=1}^{k} m_{ij}{}^{(n-r)}\,\phi_j{}^{(r)}$ reduces to $\phi_i{}^{(n)}$, so that from equations (3.8) and (3.10), for all $i=1, \ldots, k$

$$V(\theta_i{}^{(n)}|\underline{\theta}^{(0)}) = V(\phi_i{}^{(n)}|\underline{\theta}^{(0)}) = \frac{C}{2}\left[\frac{1}{N_i} + \sum_{r=1}^{n-1}\sum_{j=1}^{k} [m_{ij}{}^{(n-r)}]^2/N_j\right] + \text{terms of } 0\left(\frac{1}{N_j{}^2}\right).$$

Rearranging the double summation and ignoring terms of $0(1/N_j{}^2)$

$$(3.11)\qquad \mathrm{V}(\theta_i{}^{(n)}|\underline{\theta}^{(0)}) = \frac{1}{8}\left[\frac{1}{N_i} + \sum_{j=1}^{k}\frac{1}{N_j}\sum_{r=1}^{n-1}(m_{ij}{}^{(r)})^2\right]$$

if angles are measured in radians. This result is, of course, as discussed above only strictly valid when $\theta_i{}^{(n)}$ are near $\pi/4$ for all i, n and N, so all gene frequencies are near to $\frac{1}{2}$. Monte Carlo experiments using a computer indicate, however, that the approximation works well over a wide range of gene frequencies (see next section). Since $\mathrm{V}(p_i{}^{(n)}) = p_i{}^{(n)}\,(1-p_i{}^{(n)})\,V(\theta_i{}^{(n)})/C + 0(1/N_i{}^2)$, the gene frequency variances can easily be obtained from the variances of the angular transforms.

Gene frequencies are not usually observed directly, but are based on fitting expected to observed phenotype frequencies for a sample of the population, in terms of some specific genetic model which gives the expected phenotype frequencies in terms of gene frequencies. Let $\bar{\underline{\phi}}^{(n)}$ be the set of estimated frequencies in the n^{th} generation, which we assume to be unbiased. Then, following the procedure used in deriving equation (3.8) we can write

$$(3.12)\quad \mathrm{V}(\bar{\phi}_i{}^{(n)}|\underline{\theta}^{(0)}) = E[(\bar{\phi}_i{}^{(n)} - \phi_i{}^{(n)} + \phi_i{}^{(n)} - (\mathbf{m}\,\underline{\phi}^{(n-1)})_i + \ldots + (\mathbf{m}^{n-1}\,\underline{\phi}^{(1)})_i - (\mathbf{m}^n\,\underline{\phi}^{(0)})_i)^2|\underline{\theta}^{(0)}]$$

$$= \mathrm{V}(\bar{\phi}_i{}^{(n)}|\underline{\phi}^{(n)}) + \sum_{r=1}^{n}\mathrm{V}((\mathbf{m}^{n-r}\,\underline{\phi}^{(r)})_i|\underline{\theta}^{(r-1)}),$$

since $\bar{\phi}_i{}^{(n)}$ will be a set of random variables depending only on $\underline{\phi}^{(n)}$ and independent of the realized gene frequencies in previous generations. The only effect of the estimation procedure is, therefore, to add to the result of equation (3.11) the sampling variance of the estimated gene frequencies.

The covariance between the angular transforms of the gene frequencies in colonies i and j in the n^{th} generation can be obtained by the same general procedure used to obtain the variances. These covariances are independent of the sampling variances of the gene frequencies and in addition provide the basis for relating the correlation between gene frequencies in different colonies to their distance apart, as was done by MALÉCOT (1950, 1962) and KIMURA and WEISS (1964).

Following equation (3.8), for all i, j $(i \neq j)$

$$\begin{aligned}(3.13)\quad \mathrm{Cov}(\phi_i^{(n)}, \phi_j^{(n)} | \underline{\theta}^{(0)}) &= E[(\phi_i^{(n)} - (\mathbf{m}\,\underline{\phi}^{(n-1)})_i + (\mathbf{m}\,\underline{\phi}^{(n-1)})_i - (\mathbf{m}^2\,\underline{\phi}^{(n-2)})_i + \ldots) \\ &\quad \times (\phi_j^{(n)} - (\mathbf{m}\,\underline{\phi}^{(n-1)})_j + (\mathbf{m}\,\underline{\phi}^{(n-1)})_j - (\mathbf{m}^2\,\underline{\phi}^{(n-2)})_j + \ldots) | \underline{\theta}^{(0)}] \\ &= \sum_{r=1}^{n} \mathrm{Cov}[(\mathbf{m}^{n-r}\,\phi^{(r)})_i, (\mathbf{m}^{n-r}\,\phi^{(r)})_j | \theta^{(r-1)}] \ .\end{aligned}$$

The $\mathrm{Cov}(\phi_{l_1}^{(r)}, \phi_{l_2}^{(r)} | \underline{\phi}^{(r-1)}) = 0$ for all $l_1 \neq l_2$ and for all r, since gene frequency samples in the different colonies are independent, so that equation (3.13) reduces, after some rearrangement, to

$$(3.14)\quad \mathrm{Cov}(\theta_i^{(n)}, \theta_j^{(n)} | \underline{\theta}^{(0)}) = 1/8 \sum_{l=1}^{k} 1/N_l \sum_{r=1}^{n-1} (m_{il}^{(r)}\, m_{jl}^{(r)}), \text{ for all i, j } (\mathrm{i} \neq \mathrm{j}).$$

Estimates of gene frequencies in the different colonies will, usually, be independent, so that the covariance given by equation (3.14) is not affected by the substitution of estimates of the gene frequencies for the realized gene frequencies.

Since any set of observations of gene frequencies corresponds only to a single realization of the stochastic model, the only observation that can generally be made is that of the variance in the gene frequencies *between* colonies at any given time, while equation (3.11) gives the expected variance for any given colony based on repeated realizations of the stochastic process implied by the model.

The expected weighted variance in the gene frequencies between colonies in the n^{th} generation is given by

$$(3.15)\qquad V_B^{(n)} = E\left\{1/N \left\{ \sum_{i=1}^{k} N_i (\theta_i^{(n)} - \bar{\theta}^{(n)})^2 \right\} \right\}$$

where $N = \sum_{i=1}^{k} N_i$, $\bar{\theta}^{(n)} = 1/N \sum_{i=1}^{k} N_i\, \theta_i^{(n)}$ and E refers to expectations given $\underline{\theta}^{(0)}$, the initial frequencies. We have, therefore,

$$\begin{aligned}(3.16)\quad V_B^{(n)} &= 1/N \sum_{i=1}^{k} N_i \,(V(\theta_i^{(n)} - \bar{\theta}^{(n)}) + (E(\theta_i^{(n)} - \bar{\theta}^{(n)}))^2) \\ &= 1/N \sum_{i=1}^{k} N_i\, V(\theta_i^{(n)}) + V(\bar{\theta}^{(n)}) - 2/N \sum_{i=1}^{k} N_i\, \mathrm{Cov}(\theta_i^{(n)}, \bar{\theta}^{(n)}) \\ &\quad + 1/N \sum_{i=1}^{k} N_i \{E(\theta_i^{(n)}) - E(\bar{\theta}^{(n)})\}^2\end{aligned}$$

Here V and Cov refer to variances and covariances given $\underline{\theta}^{(0)}$.

The first term of equation (3.16) is simply the weighted mean of the individual variances within colonies, as given by equation (3.11). The second and third terms can be readily evaluated using equations (3.11) and (3.14). They will, however, in general be much smaller than the first term, especially for large k and n. The last term of equation (3.16) is the weighted variance of the *expected* angular transforms in the n^{th} generation. From equations (3.5) and (3.7), and the fact that the dominant eigenvalue of $\mathbf{m}$ must be less than unity (if α_i are not all zero),

$$E(\theta_i^{(n)}|\underline{\theta}^{(0)}) \longrightarrow \gamma_i \qquad \text{as } n \to \infty$$

where, as before, γ_i is the i^{th} component of the vector $\underline{\xi}(I - \mathbf{m})^{-1}$ and $\underline{\xi}$ is the vector $\alpha_i \eta_i$ (see equation (3.5)). For large n, therefore, the variance of the *expected* angular transforms in the n^{th} generation will not be far from that of the limiting expected angles, namely $1/N\{\sum_{i=1}^{k} N_i \gamma_i^2 - 1/N\{\sum_{i=1}^{k} N_i \gamma_i\}^2\}$. It can easily be shown that when $x_i = x$ for all i, $p_i = 1/k$ for all i is an equilibrium solution of equation (3.1), provided $x = 1/k$, in which case the γ_i are all equal and this last variance term is zero. Thus, for large n, the between colony variance is approximately the sum of the weighted mean of the individual colony variances and the weighted variances of the limiting angular transforms of the colony frequencies, which latter is zero if the x_i are all equal. When n is not large, the last term of equation (3.16) may make a significant contribution to $V_B^{(n)}$ even when the x_i are all equal, if the $\theta_i^{(0)}$ are unequal.

The combined results of equations (3.11), (3.12), (3.14) and (3.16) provide the basis for relating observed variances and covariances in the gene frequencies of a series of colonies to their expectations based on the model implied by equations (3.1) and (3.2). This in turn provides the basis, subject to the limitations of the model, for assessing the effects of various migration and linear selection patterns, both observed and hypothetical, on observed and expected variation and covariation in gene frequencies. In the remainder of this paper we analyze, numerically and analytically, the consequences of the model for a variety of specific migration patterns.

3.3. *The "total population variance":* The variance of the weighted mean frequency $\bar{\theta}^{(n)}$, which we call the "total population variance," is given by

$$(3.17)\quad V(\bar{\theta}^{(n)}) = \frac{1}{N^2}\left\{\sum_{i=1}^{k} N_i^2 V(\theta_i^{(n)}) + \sum_{\substack{i=1 \\ i\neq j}}^{k}\sum_{j=1}^{k} N_i N_j \operatorname{Cov}(\theta_i^{(n)}, \theta_j^{(n)})\right\}$$

$$= \frac{1}{8N^2}\left\{\sum_{i=1}^{k}\left\{N_i + N_i^2 \sum_{j=1}^{k}\frac{1}{N_j}\sum_{r=1}^{n-1}(m_{ij}^{(r)})^2\right\} + \sum_{\substack{i=1 \\ i\neq j}}^{k}\sum_{j=1}^{k} N_i N_j \sum_{l=1}^{k}\frac{1}{N_l}\sum_{r=1}^{n-1} m_{il}^{(r)} m_{jl}^{(r)}\right\}$$

$$= \frac{1}{8N^2}\left\{N + \sum_{r=1}^{n-1}\sum_{l=1}^{k}\frac{1}{N_l}\left\{\sum_{i=1}^{k} N_i m_{il}^{(r)}\right\}^2\right\}$$

on substituting from equations (3.11) and (3.14). If $\alpha_i = \alpha$ for all i and the

population sizes are at the equilibrium values given by repeated iteration of the forward migration matrix (see Section 2) then

$$\sum_{i=1}^{k} N_i \, m_{il}^{(r)} = (1-\alpha)^r N_l \qquad \text{for all } l. \tag{3.18}$$

Thus the total population variance becomes

$$V(\bar{\theta}^{(n)}) = \frac{1}{8N^2}\left(N + \sum_{r=1}^{n-1} N(1-\alpha)^{2r}\right) = \frac{1}{8N}\,\frac{1-(1-\alpha)^{2n}}{1-(1-\alpha)^2} \tag{3.19}$$

which is *independent of the form of the migration matrix*, M_{ij}. If all the N_i are equal, equation (3.18) holds without further restrictions.

This result (3.19) states that the variation in the total gene frequencies of sets of colonies connected by some migration pattern is independent of this pattern, and depends only on N, the total population size and α the rate of immigration from the outside. Equation (3.19) is in fact essentially equivalent to the result obtained by WRIGHT (1943 and later) for his island model provided $N\alpha$ is not small (see Section 3.5). As an example, suppose that each colony has a social structure with a defined pattern of migration. between social levels within the colony. Then (3.19) implies that, provided the migration between social levels within a colony is independent of the migration between colonies, the social structure has no effect on the variation in gene frequencies between the colonies.

3.4. *Convergence of gene frequency variances for large n:* It can be proved that the sequence of variances given by equation (3.11) converges as $n \to \infty$, though, in general, no simple explicit form for the limiting variances can be obtained. From the definition of m_{ij}, $\sum_{j=1}^{k} m_{ij} = 1-\alpha_i < 1$, so that the largest eigenvalue of the matrix $\mathbf{m}$, λ say, must be such that $\lambda < 1$. Let the matrix $\underline{\mu}$ be the normalized product of the left and right hand eigenvectors corresponding to λ so that, for large n, $\mathbf{m}^n \sim \lambda^n \underline{\mu}$. Then for large n_1 and n_2 $(n_1 > n_2)$, from equation (3.11) we have

$$\begin{aligned} V(\theta_i^{(n_1)}|\underline{\theta}^{(0)}) - V(\theta_i^{(n_2)}|\underline{\theta}^{(0)}) &= \frac{1}{8}\sum_{j=1}^{k}\frac{1}{N_j}\sum_{r=n_2}^{n_1-1}[m_{ij}^{(r)}]^2 \\ &= \frac{1}{8}\sum_{j=1}^{k}\frac{1}{N_j}\sum_{r=n_2}^{n_1-1}\lambda^{2r}\mu^2{}_{ij} + \text{terms } 0\left(\frac{\lambda_2}{\lambda}\right)^{n_2} \end{aligned} \tag{3.20}$$

where λ_2 is the second largest eigenvalue of the matrix $\mathbf{m}$. Thus, to this order of magnitude,

$$V(\theta_i^{(n_1)}|\underline{\theta}^{(0)}) - V(\theta_i^{(n_2)}|\underline{\theta}^{(0)}) = \frac{1}{8}\,\frac{\lambda^{2n_2}(1-\lambda^{2(n_1-n_2)})}{1-\lambda^2}\sum_{j=1}^{k}\frac{\mu^2{}_{ij}}{N_j} \tag{3.21}$$

which, since $\lambda < 1$, converges as $n_1 \to \infty$, and so the variance given by equation (3.11) also converges. This result is, of course, limited to the conditions under which the angular transformation remains valid throughout the whole process. When $\alpha_i = \alpha$ for all i, $\boldsymbol{m} = (1-\alpha)\,\mathbf{M}$, so that $\lambda = 1-\alpha$, and the ultimate rate of convergence depends directly on the proportion of immigrants from the general population, or the "stabilizing linear pressure".

It is clear that for convergence of the variances at least one of the α_i must be

non-zero. Without the balancing effect of the "stabilizing pressure" the variances would diverge until all loci were fixed, as they do in a single closed finite population. It is readily seen that the limiting behavior of the covariances will be closely analogous to that of the variances.

3.5. *Application to Wright's "island" model and to migration around a circle.* The migration matrix for Wright's island model takes the form $M_{ii} = 1$, all i and $M_{ij} = 0$ all $i \neq j$. Thus $m_{ii}^{(r)} = (1-\alpha)^r$ all i and $m_{ij}^{(r)} = 0$, $i \neq j$ and so, from equation (3.11)

$$(3.22)\quad V(\theta_i^{(n)}|\underline{\theta}^{(0)}) = \frac{1}{8}\left[\frac{1}{N_i} + \frac{1}{N_i}\sum_{r=1}^{n-1}(1-\alpha_i)^{2r}\right] = \frac{1}{8N_i}\,\frac{1-(1-\alpha_i)^{2n}}{1-(1-\alpha_i)^2}$$

which is essentially the same as (3.19). As $n \to \infty$

$$\lim_{n\to\infty} V(\theta_i^{(n)}|\underline{\theta}^{(0)}) \to 1/8N_i[1-(1-\alpha_i)^2] \sim 1/16N_i\alpha_i$$

if α_i is sufficiently small. This gives

$$V(p_i)/p_i(1-p_i) = 1/4N_i\alpha_i$$

which differs only slightly from WRIGHT's solution, $1/(4N_i\alpha_i+1)$, so long as $4N_i\alpha_i$ is appreciably larger than 1. It is clear from equation (3.14) that all the covariances are zero.

The circular migration matrix takes the general form shown in Table 2, namely

$$M_{ii} = 1-m,\ i = 1 \ldots = k,\ M_{ii+1} = m/2 = M_{ii-1},\ i = 2 \ldots k-1$$
$$M_{1k} = M_{12} = M_{k1} = M_{kk-1} = m/2 \quad \text{and} \quad M_{ij} = 0 \quad \text{otherwise}.$$

As k becomes large, this model tends to the linear "stepping stone" model analyzed by KIMURA and WEISS, and by G. MALÉCOT.

The derivation of the general form for the variances and covariances is given in the Appendix. From equations (A.6), (A.8), (A.11) and (A. 12), we have

$$(3.23)\quad V(\theta_i^{(n)}|\underline{\theta}^{(0)}) = \frac{1}{8N}\left[1 + \frac{1}{k}\sum_{l=0}^{k-1}\sum_{r=1}^{n-1}(1-\alpha)^{2r}\,(1-m+m\cos(2\pi l/k))^{2r}\right]$$
$$\to \frac{1}{8Nk}\sum_{l=0}^{k-1} 1/(1-(1-\alpha)^2\,(1-m+m\cos(2\pi l/k))^2) \qquad \text{as } n\to\infty,$$

and for all i, j such that $i-j = t$,

$$(3.24)\quad \mathrm{Cov}(\theta_i^{(n)}, \theta^{(n)}|\theta^{(0)}) = \frac{1}{8Nk}\sum_{l=0}^{k-1}\cos(2\pi lt/k)\sum_{r=1}^{n-1}(1-\alpha)^{2r}\,(1-m+m\cos(2\pi l/k))^{2r}$$

TABLE 2

General form of the circular migration matrix (k × k)

$1-m$	$m/2$	0	.....		0	$m/2$
$m/2$	$1-m$	$m/2$	0	.....	0	0
0	$m/2$	$1-m$	$m/2$	0	0	0
..	.....					
0	.....		0	$m/2$	$1-m$	$m/2$
$m/2$	0	.....		0	$m/2$	$1-m$

$$\rightarrow \frac{1}{8Nk} \sum_{l=0}^{k-1} \cos(2\pi lt/k)/(1-(1-\alpha)^2\,(1-m+m\cos(2\pi l/k))^2)$$

as $n \rightarrow \infty$

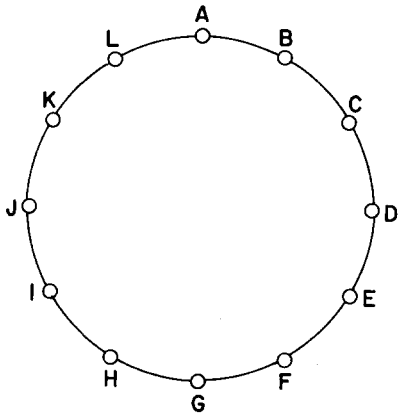
It can be shown (see Appendix) that when $k \rightarrow \infty$, the solutions given by equations (3.23) and (3.24) tend to the solutions given by KIMURA and WEISS (1964) except that, in our notation, m is replaced by $m(1-\alpha)$. This difference relates to a slight difference in the formulation of the migration models. The correspondence, in this case, of our solutions with those of KIMURA and WEISS (1964) confirms that the stochastic approximations made by them are essentially equivalent to the assumptions inherent in the use of the angular transformation. Numerical analysis of these models, discussed in the next section, shows that the infinite limit differs little from the finite models even for quite small values of k. Any differences are, of course, due to the edge effects associated with migration on a circle as opposed to migration along a line. The qualitative results of the circular migration model are then essentially the same as those of the linear stepping stone model as presented by KIMURA and WEISS (1964). Their main emphasis was placed on the approximately exponential decay of the co-variance, and therefore of the correlation, as a function of distance. Some numerical results showing the kinetics of approach to the equilibrium state and the dependence of the variance, and the exponent of decay of the correlation with distance, on m and α will be discussed in the next section.

NUMERICAL ANALYSIS

In this section we discuss some numerical investigations using theoretical matrices which are of interest particularly for showing the relationship of our model with other models and methods.

If the population is distributed at regular intervals on a line, it can be represented by a very simple migration matrix. This has elements $1-m$ on the principal diagonal, $m/2$ for positions with index i and $i \pm 1$, and zero in all other positions. In order to avoid edge effects, it is convenient to close the line to form a circle (see Figure 1) made up of k colonies. In this case, elements M_{1k} and M_{k1} are also equal to $m/2$ (see Table 2). All elements of the matrix are multiplied by $(1-\alpha)$ to take account of "migration from the outside" (see Equation (3.1)). Other symmetrical models can be built in order to simulate non-linear population distributions. Complete symmetry is only achieved by considering migration between colonies situated at the vertices of the regular solids, of which there are just five. We have used as an example for numerical analysis the icosahedron (see Figure 2) in which each of the 12 vertices, corresponding to one of the 12 colonies, is connected with five other vertices. This perhaps describes a situation intermediate between that of a two-dimensional or square lattice where each node (colony) is connected to the four neighboring nodes (as in the two-dimensional stepping stone model), and that of a three-dimensional or cube lattice in which each node is connected with six other nodes.

A finite two-dimensional lattice without edge effects was simulated by analogy



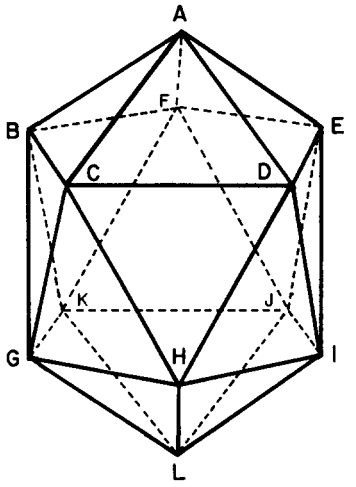


FIGURE 1.—An example of a circular pattern of colonies used as an approximation to the infinite linear stepping stone model.

FIGURE 2.—The icosahedral colony pattern (see text).

with the circular representation of an infinite line, by connecting the left and right sides of a finite square lattice, in addition to the upper and lower sides. This generates a *torus*. As a visual aid to the reconstruction of the connections in such a square lattice, it is convenient to repeat the first line of the lattice of colonies at the bottom and the first column at the right. Thus, for a 6×6 lattice, there are 36 colonies arranged as follows:

1	2	3	4	5	6	1
7	8	9	10	11	12	7
13	14	15	16	17	18	13
19	20	21	22	23	24	19
25	26	27	28	29	30	25
31	32	33	34	35	36	31
1	2	3	4	5	6	1

The migration matrix has $1-m$ on the diagonal and $m/4$ in the positions on each row corresponding to the immediate neighbors in the above arrangement. Thus for the first row $m/4$ occurs in columns 2, 6, 7 and 31, for the second row in columns 1, 3, 8 and 32 and for the 14^{th} in 8, 20, 13, and 15 and so on.

The choice of m and α values deserves some comment. There is very limited information on migration in real populations. Human populations probably lend themselves best to this type of investigation. Naturally, there may be some uncertainty as to what should be chosen as the unit corresponding to the colony. In humans the clustering is often sufficiently accentuated so that the choice of colony unit is quite clear cut. Some of the data in existence in the literature are worth discussing as an indication of the possible ranges for m and α in human populations. MODIANO *et al.* (1965) found for different villages in the province of Lecce, Italy, m values ranging between .23 and .08 for the father-offspring migration

matrix, and between .31 and .10 for the mother-offspring matrix. In the villages of the Parma valley, CAVALLI-SFORZA (1958) found "endogamy" values ranging from 34 to 77%. "Endogamy" was defined as the percentage of husband-wife pairs born in the same village. Such a value is likely to be somewhat higher than the product of the M_{ii} values from the father-offspring and mother-offspring matrices. This difference is due to the fact that further migration may take place after marriage and before birth of the progeny. Influences such as correlations in migratory potential between mates may further alter this relationship. Observations suggest, however, that m may be taken as approximately equal to $1 - \sqrt{\text{endogamy}/100}$, which lies between .4 and .13 for the Parma data. These data, thus, indicate rough limits for the range of m.

The range for α is more difficult to define. In a theoretical model which takes account only of mutation, α may be of the order of 10^{-5} or less. In a theoretical model which also includes some form of linear stabilizing selection, α may take on almost any value. A possible range of values might be between .0001 (nothing less would conceivably be measurable) and around .05. In any real population, however, it is almost always impossible, or at least very difficult, to exclude migration from outside the area being investigated and this may give rise to relatively high α values, of order m, for some of the colonies. Colonies at the periphery of a species distribution may tend to have lower α values (see MAYR (1965)). Inability to study an area completely may, at least in part, be compensated for by including in α the proportion of immigrants who come from parts of the area which have not been included in the analysis. Finally, the model may also be used to include local differences in selection, should data be available on this, by simulating with different x_i and α_i values. The above considerations justify the fairly wide range of m and α values which has been chosen for the numerical analysis.

The kinetics of change of the gene frequency variances in the colonies as a function of time, starting with an initially homogeneous population, was calculated for various α and m values and both the linear and icosahedral models, using Equation (3.11), with the results shown in Figure 3. The variance increases rapidly with time and is usually at some 90% of its asymptotic value after a period of time which is of the order of $1/\alpha$. The comparison of the linear and icosahedral models shows the effect of dimensionality. This is not pronounced unless m is large and even then it is not very striking.

The effect of m and α on the asymptotic value of the variance for the infinite linear model, calculated from Equation (A.15), is shown in Figure 4. The limiting variance decreases rapidly as both α and m increase. It should be noted that the values obtained from the circular model show very little effect of the number of colonies. Thus, values from small circular models are very close to those for an infinite line. The deviations from the approximation given by KIMURA and WEISS (1964) for α/m small, are equivalent to the deviations from linearity in the right lower part of the graph. The linear portions of the lines all have a slope of approximately $-1/2$, as expected if the variance is proportional to $1/\sqrt{\alpha}$ for any given value of m.

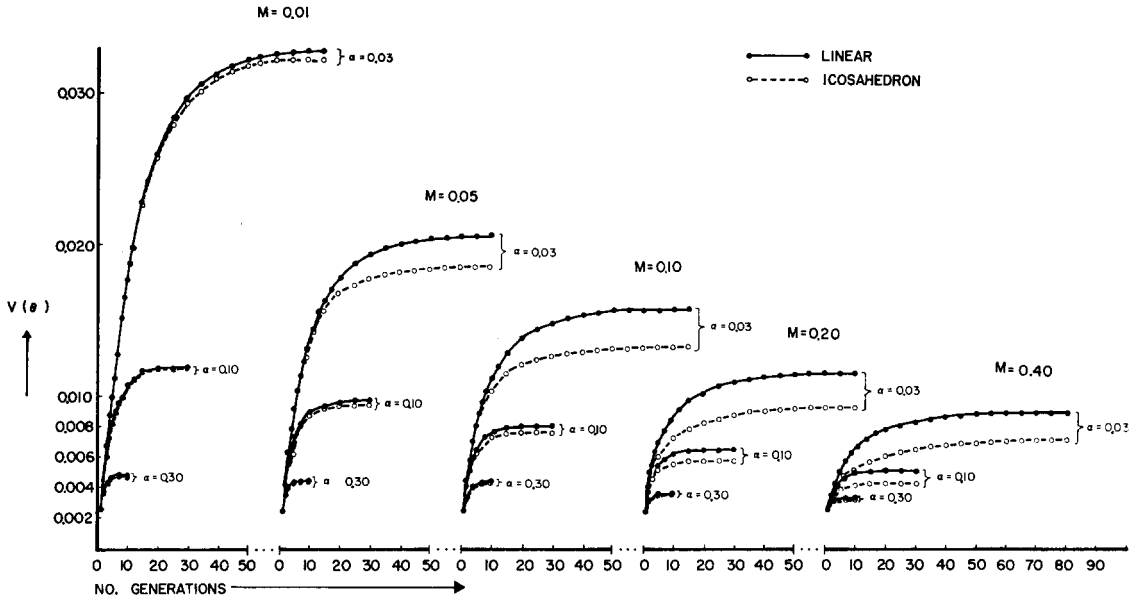


FIGURE 3.—The kinetics of the change in the variance between the gene frequencies as a function of time for 12 colonies arranged in a circular and an icosahedral pattern, for various values of m and α. Population size N is kept at a 100 in all cases.

A favorite way of studying the effect of distance is to follow the change in the correlation between the gene frequencies of different colonies as a function of their distance apart. The decrease of correlation with distance is expected to be a simple exponential (KIMURA and WEISS (1964, 1965), MALÉCOT (1962) after equilibrium has been reached, provided α/m is small.

The kinetics of the relationship between the correlation coefficient and the distance between colonies is shown in Figure 5 for the linear model with 24 colonies and $\alpha = m = 0.1$ (using equations 3.11 and 3.14). The relationship only becomes exponential at equilibrium and the intersection of the ultimate straight line with the ordinate is never at zero.

The asymptotic slopes for the infinite linear model with various m values are plotted as a function of α (using Equation (A.17)) in Figure 6. The slopes rapidly get steeper, corresponding to a rapid decline in the correlation with distance, as α and m increase. The deviations from linearity for given m are quite small even for relatively high α values. The slopes of these lines correspond to a dependance on $\sqrt{\alpha}$ for given m, as predicted by KIMURA and WEISS (1964) for small α/m. There is, however, a significant non-linearity for the correlation over short distances, which is the range over which correlations are higher and easier to measure. As the relationship is not strictly exponential and the cut-off is not at zero, the value of the first correlation coefficient (at distance one) is plotted separately in Figure 7 (calculated from Equation (A.16)). There is a very large non-linear effect of α on this correlation.

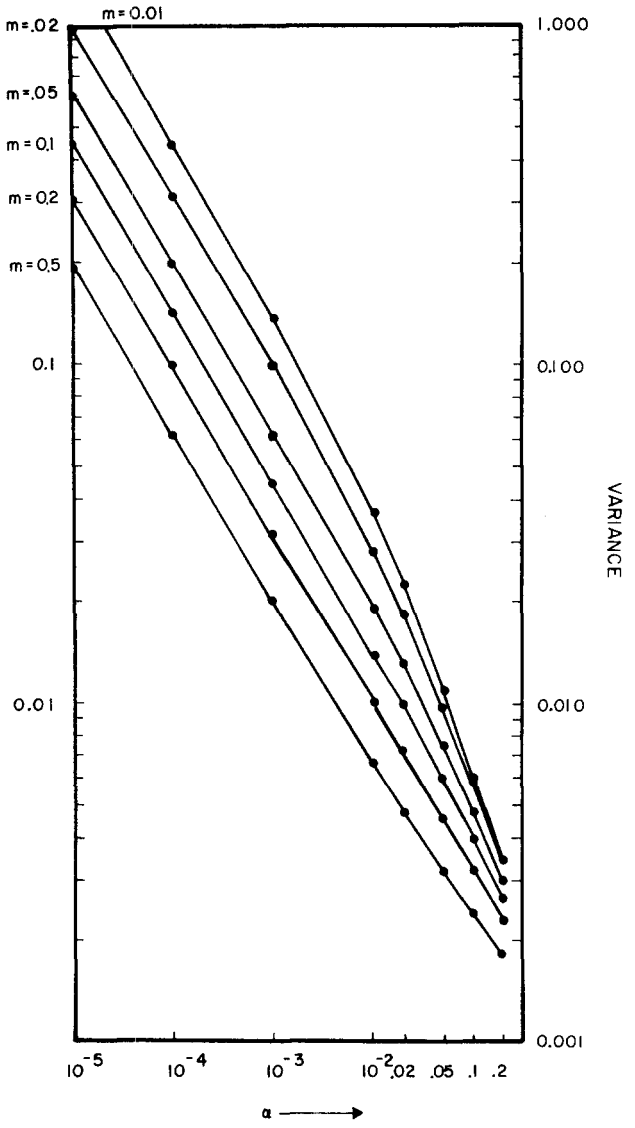


FIGURE 4.—Asymptotic values of the variance between gene frequencies for the infinite linear model, as a function of m and α.

The effect of dimensionality, shown in Figure 3, was further studied by calculating the dependence of the correlation coefficient on distance using the square lattice (Figure 8). The relationship is far from exponential. It should, however, be noted that a square lattice is a rather artificial model. The distance between two colonies is not usually uniquely determined except in limiting cases. There

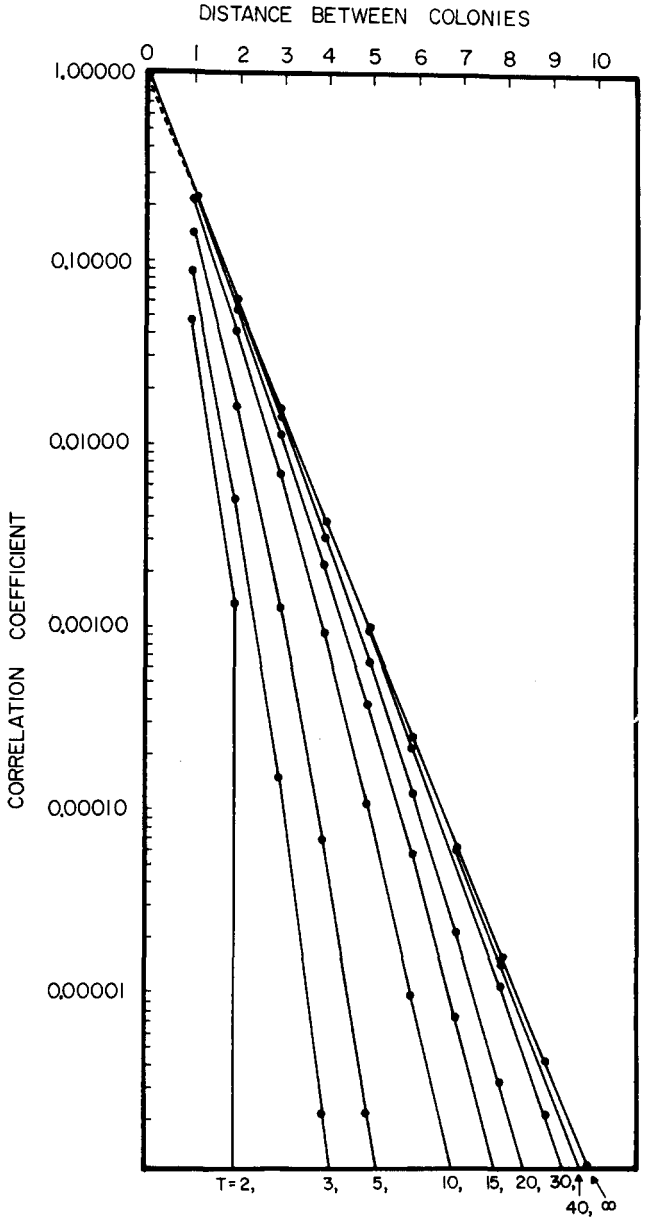


FIGURE 5.—The correlation coefficient between gene frequencies of colonies situated at the distance given in the abscissa, as a function of time : $m = 0.1$, $\alpha = 0.1$ and $k = 24$.

are colonies at a distance greater than one which are connected by a straight path while others are connected by a zigzag path. The covariances in these two types of cases are not identical even though the distance is apparently the same. The graph in Figure 8 was obtained by taking a weighted average of the probabilities for such cases.

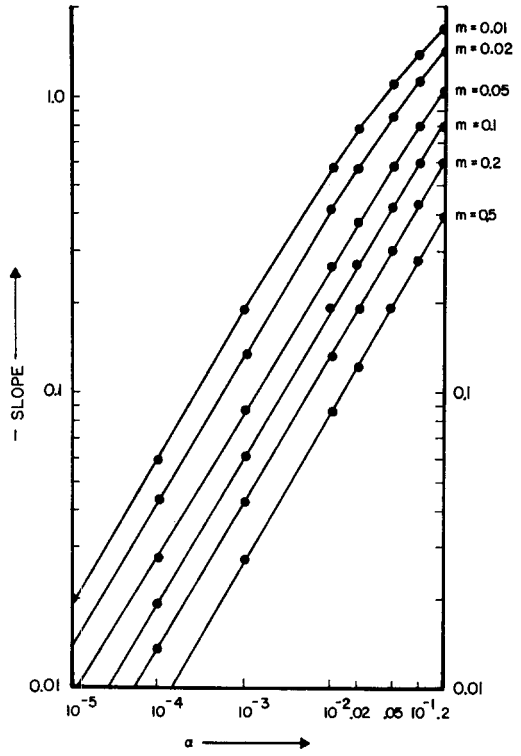


FIGURE 6.—The asymptotic slope of the exponential relationship between the correlation coefficient of gene frequencies in different colonies and the distance between them for the infinite linear model for various m and α values.

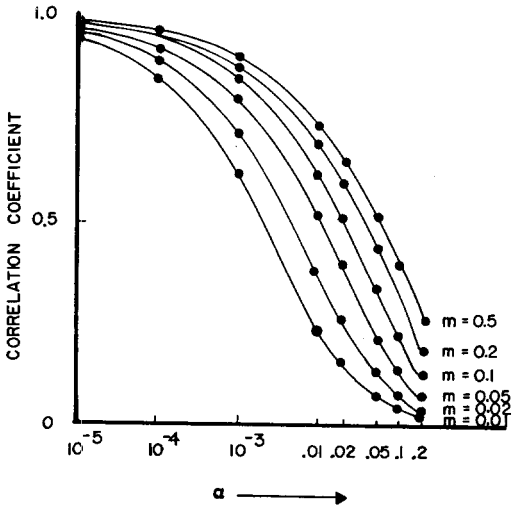


FIGURE 7.—The correlation coefficient between colonies at distance 1, for the infinite linear model, as a function of m and α.

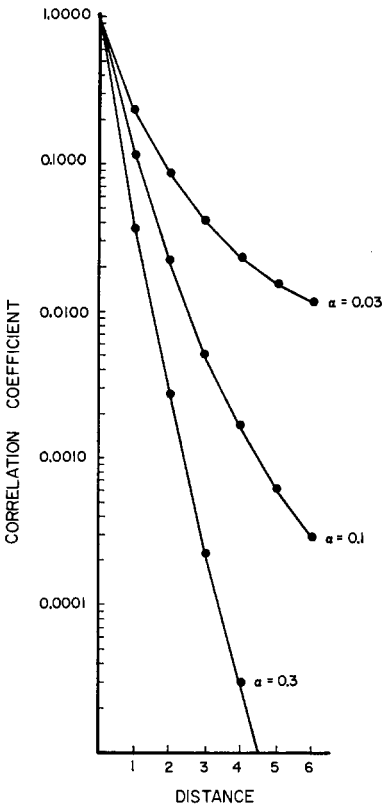


FIGURE 8.—The equilibrium relationship between the correlation coefficient and distance for a 6×6 square lattice with $m = 0.1$.

Interesting results can be obtained by varying α values between colonies. Figure 9 presents an extreme case in which only one of the ten colonies arranged in a circle receives immigration from the outside. This case is compared with that in which the same total amount of immigration is distributed equally between all the ten colonies ("isotropic immigration").

Colony A receiving all the immigrants is expected to have a lower *variance* than all the other colonies. The more distant colonies will have progressively higher variances (see Figure 9b). Thus a "cline" in the *variance* of gene frequencies is produced. There will not, in this case, be a cline in the expected gene frequencies. However, in reality, a gene frequency cline may well also be present, thus altering the correlation pattern and increasing, on the average, all the correlations. The relationship between the correlation coefficients and distance will then be different for every colony, and the correlations will in general be higher than those of the isotropic model because of the existence of a cline. This situation thus represents a potential pitfall for analysis based entirely on correlation coefficients.

Further numerical work was done with a view to testing the validity of the angular transformation. Random sampling experiments, using computer gener-

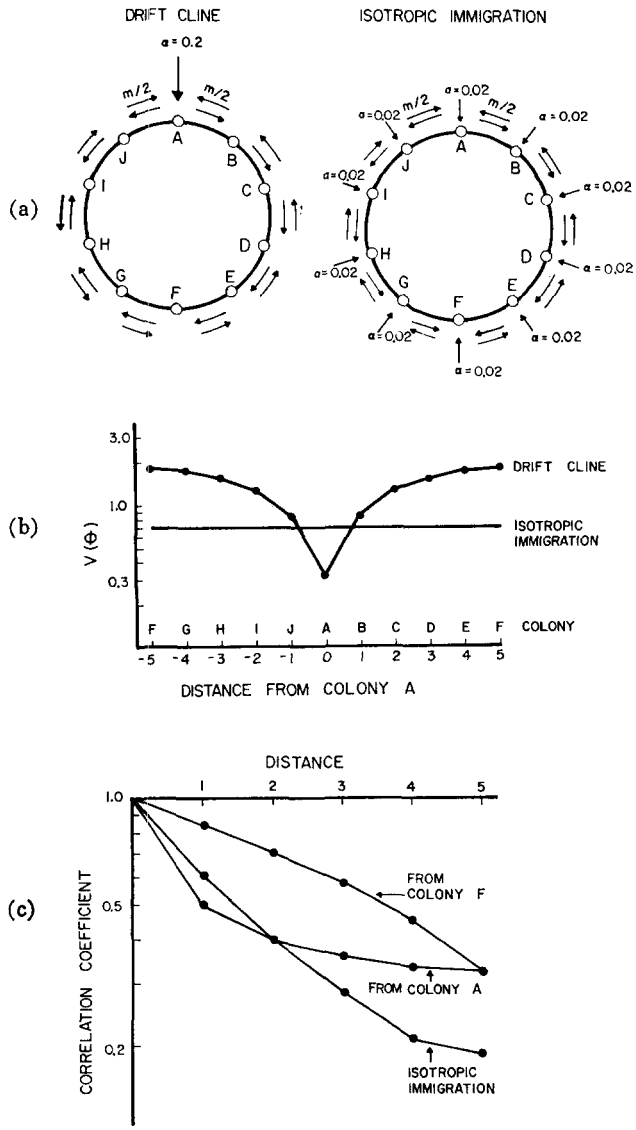


FIGURE 9.—A model of non-isotropic immigration from the outside, in comparison with the corresponding isotropic model. This gives rise to a cline of drift as shown in Figure 9b. The effects on the relationship between correlation and distance are shown in Figure 9c.

ated pseudorandom numbers were carried out following the requirements of the model (Equations (3.1) and (3.2)). At every generation, the expected gene frequency of each colony was computed as the weighted average of the gene frequencies of each colony in the preceding generation, using as weights the elements of the migration matrix corresponding to the colony to which migration takes place. A random sample of genes for that colony was then obtained. Twenty independent sampling experiments were done for each case corresponding to a

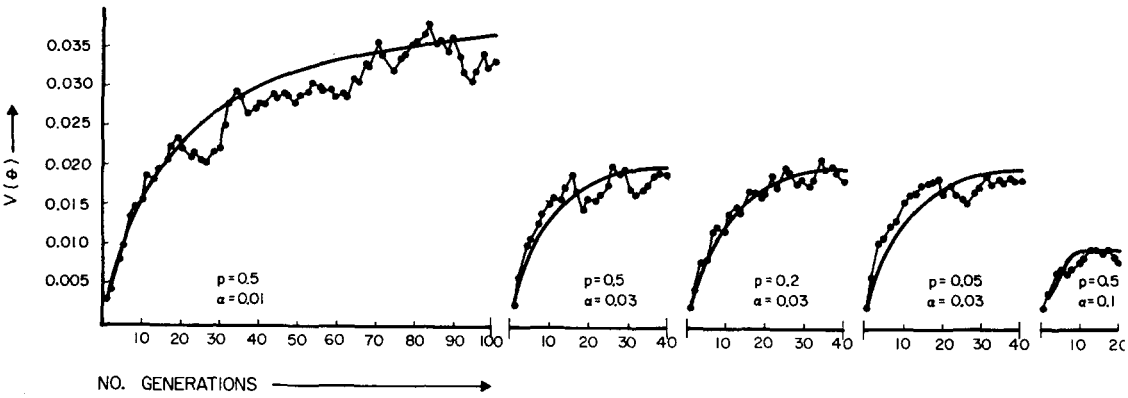


FIGURE 10.—Monte Carlo experiments to test the validity of the angular transformation (see text for further explanation).

single point on the curves shown in Figure 10. These points represent the mean variances of the angular values corresponding to the gene frequencies of each colony. Because of the nature of the experiment, there is a strong autocorrelation between successive values, which should be kept in mind when interpreting the slight divergence between the expected (solid lines) and observed (broken lines) mean variances. All experiments given in Figure 10 used a linear model with 12 colonies and with $m = .05$. The frequencies for the genes coming in from the outside (x_i) were put equal to the initial gene frequency. Thus, the mean gene frequency remains constant in each of these experiments. It can be seen that the approximation is quite satisfactory even for initial gene frequencies differing widely from 0.5 (for which the angular transformation is strictly valid) such as 0.2 and 0.05.

Some other possible factors of importance were also investigated, mainly the effects of variable colony size and of additional factors of diversification within the colonies. Variable colony size has little effect on the outcome. It is interesting, however, to consider the effect of additional isolation within the colonies such as could be due, for instance in human populations, to social stratification within a colony. As already discussed in the previous section, if the isolation thus generated is completely orthogonal to that due to distance, the variation between geographically isolated colonies is unaffected by the existence of social stratification. This was also observed in a Monte Carlo experiment with an artificial population (CAVALLI-SFORZA (1967)). An important factor requiring further investigation is the stratification of populations by age.

DISCUSSION

The main aim of our analysis has been to extend the earlier work of WRIGHT, MALÉCOT and KIMURA and WEISS to a model that can more readily make use of observed data on patterns of migration, which are generally quite irregular and involve a finite number of colonies. In so doing we have also attempted to test,

numerically, the robustness of KIMURA and WEISS's (1964, 1965) conclusions for the linear stepping stone with respect to more complex migration patterns and the validity range of the simple approximations for their equilibrium results. Their overall conclusions on the rate of decline of the variance with increasing m and α and on the approximately exponential rate of decline of the correlation in gene frequencies with distance seem to apply over a fairly wide range of α and m values. When α is comparable in magnitude to m, which may often be the case in real situations, correlations over short distances and gene frequency variances may show significant departures from the approximations based on assuming α/m is small. The rate of approach to the equilibrium conditions depends principally on $1/\alpha$. The difference between the infinite linear stepping stone and the finite circular migration models is, in all respects, quite small even for relatively few colonies. Non-linear symmetric migration patterns do not seem to have marked effects on the rate at which variances approach equilibrium or on their equilibrium values, though the correlations with distance may be far from exponential. Nonisotropic migration from the outside can clearly be a major disturbing factor to the simple conclusions derivable from the stepping stone model.

The angular transformation does not seem to be a major factor limiting the validity of our model so long as gene frequencies are not very near 0 or 1. It is possible to obtain exact variances and co-variances of the gene frequencies in terms of the elements of the square of the migration matrix (KARLIN, personal communication). The length of time needed for computation of these results would, however, increase very rapidly with increasing numbers of colonies. Undoubtedly explicit formulae for the variances and co-variances can be obtained with more complicated regular migration matrices than the circular matrix. In particular, any matrix whose rows are cyclic permutations of each other should be amenable to analysis.

There are, of course, a number of major factors which we have not taken into account, mainly because of their complexity, and which may seriously limit the extent to which the model we have proposed can account for observed variations in gene frequency. Two especially important limitations are the lack of a random element to migration and the ignoring of the population's age structure. The latter can, to some extent, be taken into account by the use of effective population sizes calculated according to KIMURA and CROW (1963). Further work is, of course, needed to assess the significance of these factors in the matching of theory and observation. Our model can, however, at least provide a theoretical framework for assessing the effects of a variety of migration patterns on random genetic drift.

The authors are indebted to G. ZEI for her help in preparation of computer programs, to PROFESSOR S. KARLIN for his advice on the solution for the circular migration matrix, and to PROFESSOR J. MCGREGOR for suggesting the "torus" migration pattern. This work was supported by grants from the U.S. Atomic Energy Commission, and by contract 012–61–12–B1A1 from EURATOM-CNEN (to LLCS) and by Research Grant GM 10452 from the Public Health Service, and Public Health Service Research Career Program Award (GM 35002–01) to WALTER BODMER. The authors would also like to thank MARC FELDMAN for his careful review of the

manuscript and his help in interpreting the approximations implied by the angular transformation.

SUMMARY

Migration between a finite set of discrete colonies can be specified in terms of a matrix. Assuming segregation for two alleles at a single locus, random mating within colonies and finite colony size, a stochastic model to describe gene frequency variation can be constructed with the use of the migration matrix. An important parameter of the model is the rate of immigration from an external gene pool representing a combination of linear stabilizing pressures which counteract the trend toward fixation due to random genetic drift. By use of the angular transformation, explicit expressions for gene frequency variances and co-variances after n generations of change can be obtained in terms of the initial gene frequencies, the migration pattern and the colony sizes.

The angular transformation is only strictly valid for gene frequencies not too far from $\frac{1}{2}$, though numerical calculations show that, in practice, it works well so long as the frequencies are not too near 0 or 1. The use of the angular transformation is critical to an analytical solution of the model. Further limitations to the model are that it does not take account of any random element in migration and ignores the population's age structure. All of these assumptions are, however, inherent in previous treatments of this problem. Subject to these limitations the gene frequency variances and co-variances at any given time can be obtained for an arbitrary migration pattern among a finite set of discrete colonies (see equations (3.11) and (3.14)). These results allow a systematic approach to the determination of an equilibrium state and the conditions for its existence. which are that at least one colony should be subject to migration from an outside source i.e. have a non-zero α. It is shown that internal subdivision of colonies which is independent of the migration pattern between the colonies, does not affect the results of the model. A complete explicit analytical solution can be obtained for migration between neighboring colonies on a circle. It can be shown that as the number of colonies in the circle tends to infinity, the analytical results become essentially equivalent to the equilibrium results given by KIMURA and WEISS (1964, 1965) for the stepping stone model. Numerical calculations indicate that circular migration models correspond closely to the infinite linear model even for a small number of colonies. These calculations further indicate that the approximate formulae based upon assuming α/m is small provided an adequate description of the linear stepping stone model for a fairly wide range of values of α and m. A major aim of the model is to evaluate the effects of more complex migration patterns on the variances and co-variances in gene frequency. Numerical results for icosahedral as opposed to circular migration models suggest that increasing dimensionality of the migration pattern does not markedly affect the results. The limiting variances generally decrease rapidly as both α and m increase. A special model in which only one of a series of ten colonies arranged in a circle receives immigration from the outside, gave rise to a cline in the variance of gene frequencies. This type of non-symmetrical heterogeneity cannot readily

be interpreted in terms of analyses based entirely on correlation coefficients between gene frequencies as a function of distance.

LITERATURE CITED

BODMER, W. F., 1960 Discrete stochastic processes in population genetics. Roy. Statist. Soc. B **22**: 218–236.

CAVALLI-SFORZA, L. L., 1958 Some data on the genetic structure of human populations. Proc. 10th Intern. Congr. Genet. **1**: 389–407. —— 1967 Human Populations, In *Heritage from Mendel*, edited by R. A. BRINK, University of Wisconsin Press.

CAVALLI-SFORZA, L. L., M. KIMURA, and I. BARRAI, 1966 The probability of consanguineous marriages. Genetics **54**: 37–60.

FISHER, R. A., and E. B. FORD, 1947 The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. Heredity **1**: 143–174.

KARLIN, S., 1966 *A First Course in Stochastic Processes.* Academic Press, New York.

KIMURA, M., and J. F. CROW, 1963 The measurement of effective population number. Evolution **17**: 278–288.

KIMURA, M., and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49**: 461–576.

MALÉCOT, G., 1945 La diffusion des gènes dans une population Mendélienne. Compt. Rend. Acad. Sci. **221**: 340–342. —— 1950 Quelques schemas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon, Sciences, Section A **13**: 37–60. —— 1951 Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique de populations. Ann. Univ. Lyon, Science, Section A **14**: 79–117. —— 1962 Migration et parenté génétique moyenne. Entretiens de Monaco en sciences humaines, 205–212. —— 1966 *Probabilité et Hérédité.* Press Universitaire de France.

MAYR, E., 1965 *Animal Species and Evolution.* Harvard University Press, Cambridge, Mass.

MODIANO, G., A. S. BENERECETTI-SANTACHIARA, F. GONANO, G. ZEI, A. CAPALDO, and L. L. CAVALLI-SFORZA, 1965 An analysis of ABO, MN, Rh, Hp, Tf and G-6-PD types in a sample from the human population of the Lecce province. Ann. Human. Genet. **29**: 19–31.

WEISS, G. H., and M. KIMURA, 1965 A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probab. **2**: 129–149.

WRIGHT, S., 1943 Isolation by distance. Genetics **28**: 114–138. —— 1946 Isolation by distance under diverse systems of mating. Genetics **31**: 39–59. —— 1951 The genetical structure of populations. Ann. Eugenics **15**: 323–354.

APPENDIX

Explicit Formulae for the Angular Variances and Covariances for the General "Circular" Migration Matrix

In this appendix we evaluate the variances and covariances of the angular transforms of the gene frequencies for the general circular model, using equations (3.11) and (3.14). We assume that $N_i = N$ and $\alpha_i = \alpha$ for all i and that the migration matrix $\mathbf{M}$ for k colonies takes the general form:

$$\begin{aligned} M_{ii} &= 1-m, && i = 1, \ldots, k \\ M_{ii+1} &= m/2 = M_{ii-1}, && i = 2, \ldots, k-1 \\ M_{1k} &= M_{12} = M_{k1} = M_{kk-1} = m/2 \\ M_{ij} &= 0 && \text{otherwise.} \end{aligned}$$

In order to apply equations (3.11) and (3.14) we need to find an explicit form for the n^{th} power of the matrix $\mathbf{M}$. The evaluation of the n^{th} power of the matrix

(A.1) $$\Delta = \mathbf{M} - (1-m)\,\mathbf{I}$$

which has $\Delta_{ii} = 0$ all i, but is otherwise the same as **M**, is a classic problem of matrix algebra (see e.g. KARLIN (1966, pp. 119–121). Using simple orthogonal trigonometric functions, it can be shown that

(A.2) $$\Delta^{(n)}_{uv} = \frac{1}{k} \sum_{l=0}^{k-1} (m \cos 2\pi l/k)^n \exp(2\pi l(u-v)i/k)$$

where, for $l = 0, \ldots, k-1$, $m \cos 2\pi l/k$, are the eigenvalues and $1/k \exp(2\pi l(u-v)\ i/k)$ are the terms of the normalized products of the left and right eigenvectors of the matrix Δ. From equation (A.1), it follows that the matrix **M** has eigenvalues $(1-m+m \cos 2\pi l/k)$ with the same eigenvectors as Δ. Thus the terms of the n^{th} power of the matrix

$$\mathbf{m} = (1-\alpha)\,\mathbf{M}$$

are given by

(A.3) $$m^{(n)}_{uv} = \frac{1}{k} \sum_{l=0}^{k-1} (1-\alpha)^n\,(1-m+m\cos(2\pi l/k))^n \exp(2\pi l(u-v)\ i/k)$$

In order to obtain the variance from equation (3.11) we need to evaluate the double sum

(A.4) $$\sum_{v=1}^{k} \sum_{r=1}^{n-1} (m^{(r)}_{uv})^2$$
$$= \frac{1}{k^2} \sum_{r=1}^{n-1} (1-\alpha)^{2r} \left[\sum_{l=0}^{k-1} (1-m+m\cos(2\pi l/k))^{2r} \sum_{v=1}^{k} \exp(4\pi l(u-v)\ i/k) \right.$$
$$+ \sum_{\substack{l=0 \\ l\neq l'}}^{k-1} \sum_{l'=0}^{k-1} (1-m+m\cos(2\pi l/k))^r\,(1-m+m\cos(2\pi l'/k))^r$$
$$\left. \times \sum_{v=1}^{k} \exp(2\pi(l+l')\ (u-v)\ i/k) \right].$$

Now $$\sum_{v=1}^{k} \exp(4\pi l(u-v)\ i/k)$$
$$= \exp(4\pi lui/k) \sum_{v=1}^{k} \exp(-4\pi lvi/k)$$
$$= \exp(4\pi lui/k) \exp(-4\pi li/k)\ (1 - \exp(-4\pi li))/(1 - \exp(-4\pi li/k))$$
$$= 0 \qquad \text{provided } l \neq 0 \text{ or } k/2$$
$$= k \qquad \text{if } l = 0 \text{ or } k/2.$$

Similarly, for $l \neq l'$,

$$\sum_{v=1}^{k} \exp(2\pi(l+l')\ (u-v)\ i/k) = 0 \qquad \text{if } l + l' \neq k$$
$$= k \qquad \text{if } l + l' = k.$$

Thus, since

(A.5) $$\cos(2\pi(k-l)/k) = \cos(2\pi - 2\pi l/k) = \cos(2\pi l/k)$$

we have, for all i,

(A.6) $$\mathrm{V}(\theta^{(n)}_i \,|\, \underline{\theta}^{(0)}) = \frac{1}{8N} \left[1 + \frac{1}{k} \sum_{l=0}^{k-1} \sum_{r=1}^{n-1} (1-\alpha)^{2r}\,(1-m+m\cos(2\pi l/k))^{2r} \right].$$

Now

(A.7) $$\lim_{n\to\infty} \sum_{r=1}^{n-1} (1-\alpha)^{2r}\,(1-m+m\cos(2\pi l/k))^{2r}$$
$$= \frac{(1-\alpha)^2\,(1-m+m\cos(2\pi l/k))^2}{1-(1-\alpha)^2\,(1-m+m\cos(2\pi l/k))^2} = \frac{1}{1-(1-\alpha)^2\,(1-m+m\cos(2\pi l/k))^2} - 1$$

so that

(A.8) $$\lim_{n\to\infty} \mathrm{V}(\theta^{(n)}_i \,|\, \underline{\theta}^{(0)}) = \mathrm{V} = \frac{1}{8Nk} \sum_{l=0}^{k-1} \frac{1}{1-(1-\alpha)^2\,(1-m(1-\cos(2\pi l/k)))^2}$$

The calculation of the covariance from equation (3.14) involves the double sum

(A.9) $$\sum_{w=1}^{k} \sum_{r=1}^{n-1} (m^{(r)}_{uw}\ m^{(r)}_{vw})$$

$$= \frac{1}{k^2} \sum_{r=1}^{n-1} (1-\alpha)^{2r} \left[\sum_{l=0}^{k-1} (1-m+m\cos(2\pi l/k))^{2r} \sum_{w=1}^{k} \exp(2\pi li(u+v-2w)/k) \right.$$
$$+ \sum_{\substack{l=0 \\ l\neq l'}}^{k-1} \sum_{l'=0}^{k-1} (1-m+m\cos(2\pi l/k))^{r} (1-m+m\cos(2\pi l'/k)^{r}$$
$$\left. \times \sum_{w=1}^{k} \exp(2\pi i(lu+l'v-(l+l')w)/k) \right].$$

Following similar arguments to those used in deriving equation (A.6), the first expression on the right of equation (A.9) is non-zero only if $l = 0$ or $k/2$ and the second expression only when $l' + l = k$. Now when $l + l' = k$ we have

(A.10) $$\sum_{w=1}^{k} \exp(2\pi i(lu + l'v - kw)/k)$$
$$= \exp(2\pi i(lu + (k-l)v)) \sum_{w=1}^{k} \exp(-2\pi iw)$$
$$= k \exp(2\pi i(l(u-v) + kv)/k) = k \exp(2\pi ilt/k)$$

where $u-v=t$. Thus from equations (A.5) and (A.10) the coefficient of $(1-m+m\cos(2\pi l/k))^{2r}$ in the second expression of equation (A.9) is, for $l \neq 0$ or $k/2$

$$k\,[\exp(2\pi ilt/k) + \exp(2\pi i(k-l)t/k)]$$
$$= k\,[\exp(2\pi ilt/k) + \exp(-2\pi ilt/k)]$$
$$= 2k\cos 2\pi lt/k\,.$$

Since equation (A.10) is also valid for $l = 0$ and $k/2$, the expression for the covariance becomes, from equation (3.14) and (A.9)

(A.11) $$\mathrm{Cov}(\theta_u^{(n)}, \theta_v^{(n)} \mid \underline{\theta}^{(0)}) = \frac{1}{8Nk} \sum_{l=0}^{k-1} \cos(2\pi lt/k) \sum_{r=1}^{n-1} (1-\alpha)^{2r} (1-m+m\cos(2\pi l/k))^{2r}$$
for all u, v such that $u-v = t < k$.

Thus, from equation (A.7) and since $\sum_{l=0}^{k-1} \cos(2\pi lt/k) = 0,\ t \neq k$

(A.12) $$\lim_{n\to\infty} \mathrm{Cov}(\theta_u^{(n)}, \theta_v^{(n)} \mid \underline{\theta}^{(0)})$$
$$= \mathrm{Cov}(t) = \frac{1}{8Nk} \sum_{l=0}^{k-1} \cos(2\pi lt/k) \Big/ (1-(1-\alpha)^2 (1-m+m\cos(2\pi l/k))^2\,.$$

Allowing $k\to\infty$, we obtain the linear "stepping stone" model analyzed by KIMURA and WEISS (1964, 1965) and by MALÉCOT (1962). In this case the sums in equations (A.8) and (A.12) become integrals in the variable $\theta = 2\pi l/k$ over the range 0 to 2π, where $d\theta \sim 2\pi/k$. The variance and covariance then take the form

(A.13) $$\overline{V} = \frac{1}{16N\pi} \int_0^{2\pi} d\theta/1 - (1-\alpha)^2 (1-m(1-\cos\theta))^2$$

and

(A.14) $$\overline{\mathrm{Cov}}\,(t) = \frac{1}{16N\pi} \int_0^{2\pi} \cos(t\theta)\, d\theta/1 - (1-\alpha)^2 (1-m(1-\cos\theta))^2$$

The result for the covariance is identical to that given by KIMURA and WEISS (1964, equation (2.1)) except that in our notation m is replaced by $m(1-\alpha)$. The result for the variance differs, in addition by an extra 1 in the denominator (KIMURA and WEISS (1964, equation (1.10))). The replacement of m by $m(1-\alpha)$ seems to arise from a slight difference in the formulation of the model by KIMURA and WEISS (1964). Where we have a factor $(1-\alpha)\,(1-m)$ they (equation (1.1)) have a factor $1-m-\alpha$. The evaluation of the integrals in equations (A.13) and (A.14) follows directly from WEISS and KIMURA (1965) to give

$$\overline{V} = \frac{1}{8NC_0} \qquad \text{where } \frac{1}{C_0} = \frac{1}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

(A.15)

$$R_1 = \sqrt{(2-\alpha-2m(1-\alpha))\ (2-\alpha)} \text{ and } R_2 = \sqrt{\alpha(\alpha+2m(1-\alpha)}$$

and

$$\text{(A.16)} \qquad \overline{\text{Cov}}(t) = \frac{1}{8N}\left[\frac{1}{2R_2}\left[1+\frac{\alpha}{m(1-\alpha)}-\sqrt{\frac{2\alpha}{m(1-\alpha)}+\frac{\alpha^2}{m^2(1-\alpha)^2}}\right]^t + \frac{1}{2R_1}\left[\sqrt{\frac{(2-\alpha-m(1-\alpha))^2}{m^2(1-\alpha)^2}}-1-\frac{2-\alpha-m(1-\alpha)}{m(1-\alpha)}\right]^t\right]$$

As pointed out by them, the second term in this equation (A.16) soon becomes negligible as t increases, at which time the covariance decreases exponentially with respect to the distance t, at a rate given by

$$\text{(A.17)} \qquad \log\left[1+\frac{\alpha}{m(1-\alpha)}-\sqrt{2\frac{\alpha}{m(1-\alpha)}+\frac{\alpha^2}{m^2(1-\alpha)^2}}\right] .$$

The differences between the results for finite k, and $k\to\infty$, reflect the edge effects due to the "circularization" of the migration matrix. The numerical results discussed in the body of the paper show that these effects are quite small even for fairly small values of k.