Copyright © 1996 by the Genetics Society of America

Inferring Patterns of Migration From Gene Frequencies Under Equilibrium Conditions

Jarle Tufto,* Steinar Engen† and Kjetil Hindar*

** Norwegian Institute for Nature Research, 7005 Trondheim, Norway and † Department of Mathematics and Statistics, Norwegian University of Science and Technology, 7034 Trondheim, Norway*

Manuscript received May 27, 1996
Accepted for publication August 23, 1996

ABSTRACT

A new maximum likelihood method to simultaneously estimate the parameters of any migration pattern from gene frequencies in stochastic equilibrium is developed, based on a model of multivariate genetic drift in a subdivided population. Motivated by simulations of this process in the simplified case of two subpopulations, problems related to the nuisance parameter q, the equilibrium gene frequency, are eliminated by conditioning on the observed mean gene frequency. The covariance matrix of this conditional distribution is calculated by constructing an abstract process that mimics the behavior of the original process in the subspace of interest. The approximation holds as long as there is limited differentiation between subpopulations. The bias and variance of estimates of long-range and short-range migration in a finite stepping stone model are evaluated by fitting the model to simulated data with known values of the parameters. Possible ecological extensions of the model are discussed.

THE pattern of migration (or gene flow) between a set of geographically separated populations reflects a large number of ecological and genetic processes. First, migration is often restricted to relatively short distances (*e.g.*, LEVIN and KERSTER 1974). Second, if carrying capacities vary between populations, optimality models predict that individuals should adopt a conditional dispersal strategy responding to the differences in local carrying capacities (HOLT 1985; JOHNSON and GAINES 1990). Finally, effective rates of migration may be modified by the breeding system (ANDERSSON 1994) and by geographic variation in selection (ENDLER 1986). Actual estimates of migration are important because they suggest how important migration is for the species, for example in limiting the development of local adaptations (*e.g.*, SLATKIN 1973; NAGYLAKI 1975).

In subdivided populations, when rates of migration are high, migration interacts with genetic drift occurring in each subpopulation, and the gene frequencies will then, after a number of generations, reach a stationary equilibrium distribution. Under the island model this distribution is the beta (WRIGHT 1931). For more complicated migration patterns such as stepping stone models (KIMURA and WEISS 1964) and models of spatially continuous populations (MALÉCOT 1975), only the variances around the equilibrium gene frequencies and the correlations between populations at different distances have been found analytically. If there is a limited amount of migration, and if effective population sizes are large, the number of generations necessary to reach equilibrium distribution may be very large, and the observed genetic structure may then reflect the initial historic genetic composition of the populations.

Evolutionary forces such as mutation and selection can also be important in determining geographic genetic variation. For many loci, and for many problems in population genetics, however, it is reasonable to assume that these forces are small compared to migration and drift, and that they therefore can be neglected (CROW 1985; AVISE 1994).

Given the large amount of already existing data on geographic genetic variation, it should be of great interest to develop models that make inferences about general migration patterns possible. Previous approaches to this problem have mostly been based on the expected amount of genetic differentiation under the island model as measured by the parameter F_{st} (WRIGHT 1951), estimated for all pairs of subpopulations or for all subpopulations taken together. Using the theory of the island model, some overall measure of gene flow is then calculated. While this approach has verified that the genetic correlations decreases with distance as predicted by *e.g.*, stepping stone models (SLATKIN 1993), it is clear that the assumptions of the island model are not valid in general.

The dependencies between the gene frequencies between the subpopulations are an important feature of the data that must be incorporated in a general model for the problem. Except under the island model, it is these dependencies that contain most of the information about the unknown migration pattern. Here, using some ideas in FELSENSTEIN (1982), we develop a model that can be used to estimate the parameters of any migration pattern by maximum likelihood. The model

Corresponding author: Jarle Tufto, Norwegian Institute for Nature Research, Tungasletta 2, 7005 Trondheim, Norway.
E-mail: jarle.tufto@nina.nina.no

is based on the underlying multivariate genetic drift process that generates the data. An attempt is made to eliminate the unknown equilibrium gene frequencies from the model by considering the distribution generated by the drift process, conditioned on the sufficient statistics for these nuisance parameters. Using simulations, we show that it is reasonable to approximate this conditional distribution by the multivariate normal. A method for calculating the covariance matrix of the distribution similar to COURGEAU (1974) is suggested, based on insights gained from further simulations. Finally, to evaluate the properties such as bias and efficiency of parameter estimates obtained using the model, we estimate the parameters of an example migration pattern, a finite stepping stone model, from simulated data.

THE GENERAL MODEL

Consider $n + 1$ populations indexed $i = 1, 2, \ldots, n + 1$. Let N_i be the variance effective size (EWENS 1979, eq. 3.96) of population i. We will only consider the simplest case of a single diallelic locus with two alleles A_1 and A_2. Let the elements of the column vector $\mathbf{p}_t$ represent the allele frequencies of A_1 in the populations in generation t, and let m_{ij} be the probability that an individual born in population i received a gene from a parent in population j. Each row of the $(n + 1) \times (n + 1)$ migration matrix $\mathbf{M}^* = [m_{ij}]$ therefore sums to one. The gene frequency in the $(n + 1)$th population remains constant and equal to q, that is, this population is of infinite effective size, and can thus be thought of as a large outside world population. The $(n + 1)$th column of $\mathbf{M}^*$ represents the immigration rates from this outside world into each subpopulation. These immigration rates can in general be different. We will let these immigration rates also include mutations, since the effects of mutations are indistinguishable from the effects of immigration from the outside world.

Apart from these assumptions, $\mathbf{M}^*$ can take any form depending on what assumptions we make about the underlying migration pattern. The migration matrix is generally a function of the parameters of some migration pattern model.

If we include genetic drift, the gene frequencies in generation $t + 1$ may be expressed as

$$\mathbf{p}_{t+1} = \mathbf{M}^*\mathbf{p}_t + \mathbf{e}, \tag{1}$$

where the elements of $\mathbf{e}$ represent the stochastic changes in the process. The elements of $\mathbf{e}$ are binomial variables rescaled to have zero expectations and variances equal to $p_{t,i}(1 - p_{t,i})/2N_i$, except e_{n+1} that always equals zero. Also note that we assume that the changes in the gene frequencies due to migration and drift are small so that the sequential order of the events in the life cycle can be ignored.

Substituting the gene frequencies with their deviation $x_i = p_i - q$ from the equilibrium gene frequency, we see that (1) can be rewritten as

$$\mathbf{x}_{t+1} = \mathbf{M}\mathbf{x}_t + \mathbf{e}, \tag{2}$$

where $\mathbf{x}_t$ is an n-dimensional column vector and $\mathbf{M}$ is a $n \times n$ matrix that equals $\mathbf{M}^*$ except that the $(n + 1)$th row and column have been dropped. The sum of each row of $\mathbf{M}$ is consequently equal to or less than one.

As noted by BODMER and CAVALLI-SFORZA (1968) and FELSENSTEIN (1982), the variances of the elements of $\mathbf{e}$ depend on the gene frequencies. To make the variances constant, one might use the arcsine square root transformation, but this also changes the expectations in the process. In fact, no transformation exists that will make both the variances constant and the expectations linear in p, and a more careful analysis will show that such a transformation is not necessary, at least to derive the covariance matrix of the stationary distribution of this multivariate process.

This derivation can be done as follows. We first want to find the recursion relation between the variances and covariances from one generation to the next. Multiplying each side of (2) with their own transposed yields

$$\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T = \mathbf{M}\mathbf{x}_t\mathbf{x}_t^T\mathbf{M}^T + \mathbf{M}\mathbf{x}_t\mathbf{e}^T + (\mathbf{M}\mathbf{x}_t)^T\mathbf{e} + \mathbf{e}\mathbf{e}^T. \tag{3}$$

The elements of the matrices in the second and third term on the right hand side, formed by taking products of row and column vectors, involve products of the stochastic variables e_i and $x_{j,t}$. Even though these stochastic variables are dependent for $i = j$, we always have $E(e_i x_{j,t}) = 0$ since $E(e_i \mid x_{j,t}) = 0$. If we take expectations of (3), these terms therefore vanish and we get

$$\mathbf{C}_{t+1} = \mathbf{M}\mathbf{C}_t\mathbf{M}^T + E(\mathbf{e}\mathbf{e}^T), \tag{4}$$

where $\mathbf{C}_t$ is the covariance matrix of the distribution at time t.

It remains to evaluate the expectation of the last term in (4). Because the elements of $\mathbf{e}$ are independent and with zero expectations, only the elements of the matrix $\mathbf{e}\mathbf{e}^T$ along the diagonal have nonzero expectations. If we first write each element of the column vector $\mathbf{e}$ as

$$e_i = e_{0,i}\sqrt{\frac{p_{t,i}(1 - p_{t,i})}{2N_i}}, \tag{5}$$

where $e_{0,i}$ is stochastic with variance equal to one and expectation equal to zero, then we see that

$$E(e_i e_i) = E\left(\frac{e_{0,i}^2}{2N_i} p_{i,t}(1 - p_{i,t})\right) = \frac{1}{2N_i}(Ep_{i,t} - E(p_{i,t}^2))$$

$$= \frac{1}{2N_i}(Ep_{i,t} - (Ep_{i,t})^2 - E(p_{i,t}^2) + (Ep_{i,t})^2)$$

$$= \frac{1}{2N_i}(q(1 - q) - c_{ii,t}), \tag{6}$$

since $E(p_{t,i}) = q$ and since $E(p_{i,t}^2) - (Ep_{i,t})^2 = \mathrm{Var}(p_{i,t}) = c_{ii,t}$. The average genetic drift in the process is thus reduced by an amount proportional to the variance c_{ii} around the equilibrium gene frequency q, as also shown by COURGEAU (1974) p. 365.

Substituting (6) into (4) and noting that the covariance matrix $\mathbf{C}_{t+1}$ must equal $\mathbf{C}_t$ as t tends to infinity and a stationary distribution is attained, we now know that the covariance matrix $\mathbf{C}$ of this stationary distribution must satisfy the equation

$$\mathbf{C} = \mathbf{M}\mathbf{C}\mathbf{M}^T + \mathbf{E}, \tag{7}$$

where the matrix $\mathbf{E} = E(\mathbf{e}\mathbf{e}^T)$. Note that $\mathbf{E}$ depends on $\mathbf{C}$. Equation 7 can be rewritten to a system of linear equations in the $n(n+1)/2$ unknown covariances c_{ij} and solved for $\mathbf{C}$.

A more formal proof of the existence of this limit is given in COURGEAU (1974). It should be noted that the solution of (7) is exact to the extent that the order of the events in the life cycle can be ignored.

FITTING THE MODEL TO A GENETIC SAMPLE

Some introductory remarks: Typically, in studies on genetic differentiation, a large number of individuals have been sampled from a set of subpopulations $i = 1, 2, \ldots, n$, and the frequencies in each subpopulation $p_1, p_2, \ldots, p_n$ of different alleles have been determined by *e.g.*, protein electrophoresis or restriction fragment length polymorphism (RFLP) analysis. Our interest is to make inferences about the parameters of the underlying migration pattern from these gene frequencies. Formally, this can be done by assuming that the population system has reached its stationary distribution, then use this stationary distribution of the process as the probability distribution for the data, and finally estimate the parameters of the model by maximizing the probability of the observations.

An additional difficulty is however introduced by the parameter q, the frequency of the long-range migrants, which in general will be unknown. We have no interest in making inferences about q, that is, q is a nuisance parameter in the model. Sufficient statistics are important in models containing such nuisance parameters because they contain all information in the data about the unknown parameter. The sampling distribution of the model, conditioned on some sufficient statistics t for the nuisance parameter θ, is the relevant distribution to consider, because this distribution, by definition, is independent of the true value of the nuisance parameter. By conditioning on t, we restrict our attention to only a small part of the sample space, and ignore other possible outcomes that are irrelevant for the problem. In the context of maximum likelihood estimation, this principle is called conditional likelihood (MCCULLAGH and NELDER 1989, ch. 7).

In the present case, with n subpopulations in the system, the relevant distribution that we seek, on which calculations of the likelihood must be based, is the $(n-1)$-dimensional stationary distribution of the process, conditioned on some sufficient statistic for q. Because of the complexity of the generated distribution, finding a sufficient statistic for q and finding the corresponding conditional distribution, will necessarily have to be based on some approximations.

Approximations for small fluctuations: If the fluctuations around q are small, for example if there is a high rate of immigration from the outside world into each subpopulation, then the genetic drift will be nearly constant, and the process can then be approximated by a multivariate autoregressive process. This process is known to have the multivariate normal as its stationary distribution. Since $E(p_i) = q$ for all populations, we know from the properties of the multivariate normal, that the weighted mean gene frequency

$$\bar{p} = \sum_{i=1}^{n} w_i p_i \tag{8}$$

is sufficient for q, provided that the weights $w_1, \ldots, w_n$ are chosen to minimize the variance of $\bar{p}$ (APPENDIX B).

It can also be shown that the multivariate normal distribution conditioned on the linear combination $\bar{p} = \sum_{i=1}^{n} w_i p_i$ is also multivariate normal (KENDALL *et al.* 1983, exercise 15.1). Also, if we as suggested by FELSENSTEIN (1982), work with the deviations of each gene frequency from the weighted sample mean, that is, an $(n-1)$-dimensional vector $\mathbf{y}$ where $y_i = p_i - \bar{p}$, then the distribution of $\mathbf{y}$ conditional on $\bar{p}$ is independent, not only of q, but also of $\bar{p}$ (APPENDIX C).

The vector $\mathbf{y}$ may be expressed as a linear matrix transformation of $\mathbf{p}$,

$$\mathbf{y} = \mathbf{K}\mathbf{p}, \tag{9}$$

where the $(n-1) \times (n)$ matrix

$$\mathbf{K} = \begin{bmatrix} 1-w_1 & -w_2 & \cdots & -w_{n-1} & -w_n \\ -w_1 & 1-w_2 & \cdots & -w_{n-1} & -w_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -w_1 & -w_2 & \cdots & 1-w_{n-1} & -w_n \end{bmatrix}. \tag{10}$$

The unconditional covariance matrix of $\mathbf{y}$ is then

$$\mathbf{C}_y = \mathbf{K}\mathbf{C}\mathbf{K}^T. \tag{11}$$

Again, since $\mathbf{y} \mid \bar{p}$ is independent of $\bar{p}$ (APPENDIX C) it follows that the conditional covariance matrix $\mathbf{C}_{y|\bar{p}}$ that we seek equals $\mathbf{C}_y$ given by (11).

We can therefore consider (11), when $\mathbf{C}$ is calculated from (7), to be a naive approximation of $\mathbf{C}_{y|\bar{p}}$. By making some further distributional assumptions, the likelihood can be calculated. Our main concern at this stage, however, is how well (11) approximates $\mathbf{C}_{y|\bar{p}}$ when the fluctuations around the equilibrium gene frequency q become large. This will be investigated in the next subsection. It still seems reasonable, however, to rely on the assumption that the distribution of $\mathbf{p}$, when conditioned on (8), is approximately independent of q, also more generally.

Simulations of the two population case: To get an impression of the behavior of the process, we will first simulate its stationary distribution. We will look at the