Theoretical Population Biology 100 (2015) 88–97

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Covariation of gene frequencies in a stepping-stone lattice of populations

Joseph Felsenstein*

Department of Genome Sciences, University of Washington, Box 355065, Seattle, WA 98195-5065, United States
Department of Biology, University of Washington, United States

ARTICLE INFO

Article history:
Received 14 November 2011
Available online 23 December 2014

This paper is dedicated to the memory of the late Takeo Maruyama (1936–1987)

Keywords:
Stepping-stone model
Genetic drift
Neutral mutation
Migration
Geographic differentiation

ABSTRACT

For a one- or two-dimensional lattice of finite length consisting of populations, each of which has the same population size, the classical stepping-stone model has been used to approximate the patterns of variation at neutral loci in geographic regions. In the pioneering papers by Maruyama (1970a, 1970b, 1971) the changes of gene frequency at a locus subject to neutral mutation between two alleles, migration, and random genetic drift were modeled by a vector autoregression model. Maruyama was able to use the spectrum of the migration matrix, but to do this he had to introduce approximations in which there was either extra mutation in the terminal populations, or extra migration from the subterminal population into the terminal population. In this paper a similar vector autoregression model is used, but it proves possible to obtain the eigenvalues and eigenvectors of the migration matrix without those approximations. Approximate formulas for the variances and covariances of gene frequencies in different populations are obtained, and checked by numerical iteration of the exact covariances of the vector autoregression model.

Stepping stone models of migration on rectangular lattices of populations have become of increasing interest as samples of many SNP loci have been collected in contiguous geographic areas, particularly in human populations. Stepping-stone models were pioneered independently by Malécot (1951) and by Kimura (1953); Kimura and Weiss (1964). They considered lattices of infinite numbers of populations connected by migration in one and two dimensions, and derived expressions for the genetic variability expected in the populations in a balance between mutation and genetic drift of neutral alleles.

Lattices of finite size are of greater practical interest as models of real populations. Malécot (1948, 1950) pioneered them, using a model of a torus or circle of a finite number of populations. Maruyama (1970a,b, 1971) was the first to consider stepping stone models for ordinary one- and two-dimensional linear or rectangular lattices of finite numbers of populations. He gave expressions for the variances and covariances of gene frequencies for arbitrary pairs of populations when there was a two-allele neutral mutation model in which the population had reached its equilibrium distribution. To do so he used an approximate model which had some lack of realism in the treatment of migration into the terminal populations of the lattice. He considered different variations in the way these terminal populations were modeled.

In the present paper I will treat a more exact model with a more realistic pattern of migration into the terminal populations. The results are similar, but not identical, to Maruyama's results.

Patterson et al. (2006) have pointed out the importance of using the eigenvectors and eigenvalues of the variation in gene frequencies in principal components analysis of data from different populations, an approach which goes back to Menozzi et al. (1978) and is reviewed by Cavalli-Sforza and Feldman (2003). An important issue is the interpretation of any significant patterns of geographic differentiation that are found. They do not necessarily indicate historical events such as waves of migration.

Novembré and Stephens (2008) have pointed out that the eigenvectors of a lattice model of migration are startlingly similar to principal components found for gene frequency patterns in geographic studies of genetic variation. When one of these principal components is seen, this makes it less obvious that it must arise from a historical invasion event, as the stepping stone models do not include historical invasions. They pointed out that migration matrices such as the ones Maruyama used fall into the class of Toeplitz matrices (Gray, 2006) for which the eigenvectors and eigenvalues can readily be computed. The more exact model which we use here has migration matrices that are not precisely Toeplitz matrices. It turns out that their eigenvalues have similarities with those of Maruyama's matrices, and their eigenvectors are readily found.

* Correspondence to: Department of Genome Sciences, University of Washington, Box 355065, Seattle, WA 98195-5065, United States.
E-mail address: joe@gs.washington.edu.

http://dx.doi.org/10.1016/j.tpb.2014.12.004
0040-5809/© 2015 Elsevier Inc. All rights reserved.

The immediate purpose of obtaining expressions for the covariances will be to approximate the joint distribution of gene frequencies in the populations by a multivariate normal distribution, for which the expectations and covariances are all that we need to determine the distribution. The expectations are easy to obtain; the covariances do still require one approximation but, when checked against an exact numerical solution of the equations, seem closer to the correct values than Maruyama's approximations are. The joint distribution of gene frequencies will not actually be normal, but for cases with small departures of gene frequencies from their expectations it will come close to being multivariate normal. Elsewhere I hope to discuss the development of an approximate maximum likelihood inference of the parameters of the model using this approximation. The formulas developed here may also be useful to others working with finite stepping-stone models who wish to have a closer approximation of the covariances of gene frequencies than has hitherto been available.

1. The model

The finite stepping-stone model is a linear lattice of n_1 populations (if in one dimension) or a rectangular lattices of $n_1 \times n_2$ populations (if in two dimensions). Analogous models can be erected in higher numbers of dimensions. Each population has the same size, N individuals. The model has discrete, nonoverlapping generations. In each generation an infinite number of offspring are produced in each population. From these a random N are chosen to survive to be the adults of the next generation. The population sizes thus remain constant.

Each individual among the offspring is diploid, and has two parents. Migration enters the picture in the locations of the parents. For most of the populations in a one-dimensional lattice, there is a probability $1 - m$ that a particular individual drawn to be a parent comes from the same population. In a one-dimensional lattice there is a probability $\frac{1}{2}m$ that this parent comes from the population to the left, and a probability $\frac{1}{2}m$ that it comes from the population to the right. The exception is when the offspring population is the first or the last in the lattice. Then there is a probability $\frac{1}{2}m$ that it comes from the adjacent population, and the rest of the time, $1 - \frac{1}{2}m$ of the time, it comes from the same population.

If the lattice is a two-dimensional one, the same process is imagined to occur in both dimensions, completely independently. Thus for a population that is not on a boundary of the lattice, a parent for the (i, j) population may have come from that population, or from any of the 8 populations surrounding that one, as shown in Fig. 1. It can have come from the four populations adjacent on diagonals with probability $\frac{1}{4}m^2$ each, and from the four populations adjacent in one direction with probability $\frac{1}{2}m(1-m)$ each. With probability $(1-m)^2$ it comes from population (i, j). If the population is on one of the boundaries of the two-dimensional lattice, but not at a corner, its parents come from the two adjacent side populations with probability $\frac{1}{2}m\left(1 - \frac{1}{2}m\right)$ each, from the adjacent interior population with probability $\frac{1}{2}m(1 - m)$, from the two nearest diagonal populations with probability $\frac{1}{4}m^2$ each, and hence from the same population the remaining $\left(1 - \frac{1}{2}m\right)(1-m)$ of the time. When the population is in a corner of the lattice, a parent comes from the two nearby populations with probability $\frac{1}{2}m(1 - \frac{1}{2}m)$ each, from the one population one step away on the diagonal with probability $\frac{1}{4}m^2$, and from that population itself with the remaining probability $\left(1 - \frac{1}{2}m\right)^2$.

These seemingly complicated patterns are really just the consequence of having $\frac{1}{2}m$ come from each neighboring population in one dimension, with the two-dimensional pattern being independent movement in the two dimensions.

Fig. 1 shows the migration rates in the present model in one- and two-dimensional lattices. This is slightly different from the migration pattern usually used in two-dimensional stepping-stone models. In most previous work, in two dimensions the migration can only come from one of the two populations immediately adjacent in one dimension, or immediately adjacent in the other dimension. Migrants could only come from populations $(i-1, j)$, $(i+1, j)$, $(1, j-1)$, or $(i, j+1)$. The present scheme, in which migration occurs or not in either dimension, independently, leads to greater mathematical tractability. It is worth noting that such a scheme is also implicit in Maruyama's papers.

The model follows the gene frequency of one allele at a locus, in the presence of migration, mutation, and genetic drift. The model of mutation has two alleles with mutation back and forth between them. If the mutation rate from A to a is μ, and the mutation rate from a to A is ν, the equilibrium gene frequency is $\bar{p} = \nu/(\mu + \nu)$. Mutation in such a one-locus model acts as if it were a form of migration. If we set an additional rate of migration into each population of $m_\infty = \mu + \nu$, and have these immigrant copies of the gene be drawn from a pool in which the gene frequency of A is $\bar{p}$, this will be indistinguishable from a model that has migration plus mutation between two alleles.

In the present model, parents are chosen according to the migration model, with the two parents of an individual independently drawn. Each parent contributes an allele to the offspring. Mutation occurs (or does not) for each copy of the gene. Each population thus has a pool of newborn offspring, whose genetic composition is characterized by the gene frequency of the A allele (which I will call p). Among these offspring, the model makes the presence or absence of the A allele at each copy from that pool independent of the other copies, so that we do not need to concern ourselves with the diploid genotype frequencies in the pool.

Genetic drift occurs by sampling N diploid individuals from the offspring pool, without replacement. As the presence of the A allele in each copy in the diploid individuals is independent, this has the same effect as drawing $2N$ times from a pool which has gene frequency p.

We can take the gene frequencies in the local populations and arrange them in a column vector, whose length is the number of populations. For the two-dimensional case this involves taking the gene frequencies in the rows of the array of populations, forming each into a column vector, and stacking them on top of each other, so that the first two entries in the vector are the gene frequencies for populations (1, 1) and (1, 2). Let the populations now be numbered in the order in which they appear in this vector.

For both one- and two-dimensional cases we will see below that we can use the migration pattern to create a migration matrix $\mathbf{M}$ whose elements m_{ij} are the probability that a copy of the gene found in the newborn offspring pool for population i came from a parent which was in population j, we can then write for population i the gene frequency in the next generation:

$$p'_i = (1 - m_\infty) \sum_{j=1}^{n} m_{ij}\, p_j + m_\infty\, \bar{p} + \varepsilon_i \tag{1}$$

where ε_i is the change due to genetic drift. The random variable p'_i is a binomial proportion after $2N$ trials with a probability of success equal to the sum of the first two terms on the right-hand side of this equation. The expectation of ε_i is thus zero.

This set of equations can be put into matrix form as

$$\mathbf{p}^{(t+1)} = (1 - m_\infty)\mathbf{M}\mathbf{p}^{(t)} + m_\infty \bar{p}\mathbf{1} + \boldsymbol{\varepsilon}^{(t)}, \tag{2}$$

where the vector $\mathbf{1}$ is a column vector of 1's and $\mathbf{p}^{(t)}$ is the vector of gene frequencies in the populations in the adult stage of generation t. This equation is true for any pattern of recurrent migration, not just for stepping-stone lattices.