Alleles In Space (AIS): Computer Software for the Joint Analysis of Interindividual Spatial and Genetic Information

M. P. MILLER

From the Department of Biology, 5305 Old Main Hill, Utah State University, Logan, UT 84322-5305.

Address correspondence to Mark P. Miller at the address above, or e-mail: Mark.Miller@usu.edu.

Genetic analyses of natural populations have historically relied on statistical procedures based on the concept that distinct "populations" of a species exist across a landscape. Invariably, commonly used analyses reduce to approaches that treat collections of individuals ("populations") as independent/causative variables and allele frequencies as dependent/response variables. Examples of these procedures include Wright's F_{ST} and its variants (Excoffier et al. 1992; Nei 1973; Slatkin 1995; Weir and Cockerham 1984), contingency table procedures (Raymond and Rousset 1995; Roff and Bentzen 1989), and measures of genetic distances among populations (e.g., Nei 1972, 1978; Reynolds et al. 1983). These analyses qualitatively or explicitly test null hypotheses of homogeneity of allele frequencies between or among populations.

Although almost universally applied, the analyses mentioned above are not necessarily appropriate in many situations. For example, highly mobile organisms such as large mammals or birds can occupy continuous habitats over large spatial scales. Plants may also occupy large continuous habitats, as can species inhabiting marine or aquatic systems. In these cases, objectively designating groups of individuals at population levels for use in genetic analyses may prove difficult, if not impossible. Clearly, an important consideration in these situations is the spatial extent of the "populations" that are chosen for analyses. If groups of organisms are defined over larger than appropriate spatial scales, resulting measures of genetic differentiation may actually provide ambiguous or misleading results (Miller et al. 2002).

To address many of these issues, I have developed a new software package entitled "Alleles In Space" (AIS). This program, rather than implementing methodology that relies on arbitrary groupings of individuals, instead has the ability to perform joint analyses of interindividual spatial and genetic information that can be applied at virtually any spatial scale. These approaches specifically lend themselves to analyses of genetic data when one or a few individuals are sampled from large numbers of collection sites. Moreover, the program is designed to handle a wide variety of genetic data types, including codominant marker systems, dominant marker systems, and DNA sequences. Thus AIS will likely be useful for elucidation of patterns in diverse study types ranging from local analyses of genetic structure, phylogeographical studies, and studies encompassing aspects of the emerging field of landscape genetics (Manel et al. 2003).

Program Description

Alleles In Space has a simple graphical interface that runs under any 32-bit Windows operating system (95/98/ME/NT/XP). A Pentium III processor with at least 64MB RAM is recommended. An approximately 4MB self-extracting installation file containing the executable program file, sample datasets, and documentation (in portable document format [PDF] format and a Windows help file) can be downloaded free of charge from http://www.marksgeneticsoftware.net. Two separate input files are used to perform analyses. One data file contains sets of spatial coordinates for each observation in the dataset, while the second contains genetic data for each individual analyzed. Once input files have been selected, users may specify any number of different options for the analyses they wish to perform. Following the analyses, new windows are displayed that contain text-based representations of analysis results and graphical depictions of the analyses (when appropriate). All text and graphics created by the program can be copied to the Windows clipboard and inserted in other electronic documents.

Analyses Implemented in AIS

Alleles In Space performs a number of different analyses that can be used to detect or characterize patterns of spatial genetic structure. For example, it can perform simple Mantel tests (Mantel 1967) to evaluate correlations between genetic and geographical distances of sampled individuals. Likewise, AIS can perform a generalized form of spatial autocorrelation analysis (Cliff and Ord 1973; Sokal and Oden 1978a,b) that permits detection of genetic structure and allows for inferences to be made about spatial scales over which the genetic structure occurs (Barbujani 2000; Clark and Richardson 2002; Manel et al. 2003).

Alleles In Space also implements a novel procedure based on the statistical concept of aggregation. Aggregation indices are commonly used in ecological studies to characterize spatial distributions of individuals across landscapes (Clark and Evans 1954; Hopkins and Skellam 1954; Pielou 1977) and have been widely used as measures of forest stand structure (Pommerening 2002), specifically with respect to describing the presence of either random, clumped, or uniform spatial distributions of individuals. AIS uses a modification of the aggregation index of Clark and Evans (1954) to perform an allelic aggregation index analysis (AAIA) that provides a basis for testing the null hypothesis that each allele at a locus is distributed at random across a landscape (i.e., no aggregation

or genetic structure) relative to the aggregation of the actual organisms sampled for analysis purposes.

Consider a set of N_j copies of allele j observed at a locus from a sample of n individuals collected at different locations across a landscape. It is possible to calculate R_j, an allele-specific aggregation index for allele j, as

$$R_j = \bar{d}_j^O / \bar{d}_j^E, \qquad (1)$$

where $\bar{d}_j^O$ represents the average nearest-neighbor distance for observations of allele j and $\bar{d}_j^E$ represents an expected average nearest-neighbor distance for N_j randomly distributed copies of allele j. $\bar{d}_j^E$ is calculated as

$$\bar{d}_j^E = 1/(2\sqrt{N_j/A}), \qquad (2)$$

where A represents the area over which all n sampled individuals were collected. Furthermore, the quantity R_j^{AVE} can be calculated over all alleles and loci as the arithmetic mean of all individual R_j values to serve as a global test statistic for the entire dataset. The significance of each individual R_j value and R_j^{AVE} can be evaluated through the use of a randomization procedure where individuals and genotypes are randomly redistributed among individual collection locations. R_j and R_j^{AVE} will be smaller than random expectations when alleles show a clumped (aggregated) spatial distribution, and in contrast, will be greater than random expectations when alleles show a tendency toward a uniform spatial distribution (Pielou 1977). Note that $\bar{d}_j^E$ is not affected by the randomization procedure. As a result, accurate area estimates (A) are not required to construct hypothesis tests. However, accurate estimates of A may facilitate comparisons of aggregation index values among datasets. AIS provides a number of different approaches that can be used to quantify A for a given dataset.

The analyses described above (Mantel tests, spatial autocorrelation analyses, and AAIA) provide a basis for determining if, on average, nonrandom patterns of genetic diversity exist over a landscape. However, over large spatial scales, considerable variation may exist in patterns of genetic structure due to vicariance or barriers to gene flow (Manel et al. 2003). Thus AIS includes two different procedures that may hold utility for researchers conducting phylogeographical analyses or other landscape-scale explorations of patterns of genetic diversity and structure. First, the program contains routines that implement Monmonier's algorithm (Monmonier 1973). This geographical regionalization procedure is increasingly being used to detect the locations of putative barriers to gene flow by iteratively identifying sets of contiguous, large genetic distances along connectivity networks (Doupanloup et al. 2002; Manel et al. 2003; Manni et al. 2004). In AIS, a Delaunay triangulation (Brouns et al. 2003; Watson 1992) is used to generate the connectivity network among collection sites. After analyses, a graphical representation of putative "barriers" inferred by the algorithm is superimposed over the connectivity network to assist with rapid identification of important geographical features reflected by the genetic dataset. A text-based representation of the search procedure is provided that contains quantitative information about detected barriers from each analysis.

Alleles In Space also implements a novel technique that can be used to obtain graphical representations of genetic distance patterns across landscapes. The three-dimensional surface plots generated by this procedure are referred to as "genetic landscape shapes." See Miller et al. (in press) for an example of this analysis applied to an empirical data set. Unlike Monmonier's algorithm, this procedure allows for qualitative characterization of all areas of a sampled landscape as opposed to solely identifying sets of sampling areas separated by contiguous, large genetic distances. The procedure is initiated by constructing a connectivity network of sampling areas and assigning calculated interindividual genetic distances (Z_i) to landscape coordinates at midpoints (X_i, Y_i) of the n connectivity network edges. Next, a simple interpolation procedure (inverse distance-weighted interpolation) (Watson 1992; Watson and Philips 1985) is used to infer genetic distances at locations on a uniformly spaced grid overlaid on the entire sampled landscape. For each grid coordinate (x, y), an inferred genetic distance, Z, is obtained from each of the $i = 1$ to n genetic distances (Z_i) assigned to the connectivity network as

$$Z = \frac{\sum_{i=1}^{n} w_i \times Z_i}{\sum_{i=1}^{n} w_i}, \qquad (3)$$

where w_i is a weighting function assigned to each Z_i that is inversely proportional to the geographical distance between a grid coordinate (x, y) and the actual geographical coordinates (X_i, Y_i) assigned to each of the n values of Z_i. The weighting function w_i is calculated as

$$w_i = \begin{cases} [(X_i - x)^2 + (Y_i - y)^2]^{-\frac{a}{2}} & \text{when } X_i \neq x, \ Y_i \neq y, \\ 1 & \text{when } X_i = x, \ Y_i = y \end{cases}, \qquad (4)$$

and a is a distance weighting value. Larger values of a cause interpolated values to be more influenced by close points, and smaller values of a ($\sim$0) allow all points to equally influence interpolated values. Some general guidelines for choosing appropriate interpolation parameters are provided in the program's documentation. After interpolation, AIS produces three-dimensional surface plots where X and Y coordinates correspond to geographical locations on the rectangular grid and surface plot heights (Z) reflect genetic distances. The resulting surface plots can be easily rotated in the program, and many additional aspects of the graph can easily be modified by users. Furthermore, the program generates a graphical representation of the connectivity network used in the interpolation procedure and provides a detailed text-based description of different steps performed during the analysis.

Acknowledgments

References

Barbujani G, 2000. Geographic patterns: how to identify them and why. Hum Biol 72:133–153.

Brouns G, De Wulf A, and Constales D, 2003. Delaunay triangulation algorithms useful for multibeam echosounding. J Surv Eng 129:79–84.

Clark PJ and Evans FC, 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453.

Clark SA and Richardson BJ, 2002. Spatial analysis of genetic variation as a rapid assessment tool in the conservation management of narrow-range endemics. Invert Syst 16:583–587.

Cliff AD and Ord JK, 1973. Spatial autocorrelation. London: Pion Limited.

Doupanloup I, Schneider S, and Excoffier L, 2002. A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11:2571–2581.

Excoffier L, Smouse PE, and Quattro JM, 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Hopkins B and Skellam JG, 1954. A new method for determining the type of distribution of plant individuals. Ann Bot 18:213–227.

Manel S, Schwartz ML, Luikart G, and Taberlet P, 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18:189–197.

Manni F, Guerard E, and Heyer E, 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum Biol 76:173–190.

Mantel N, 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.

Miller MP, Bellinger MR, Forsman ED, and Haig SM, in press. Effects of historical climate change, habitat connectivity, and vicariance on genetic structure and diversity across the range of the red tree vole (*Phenacomys longicaudus*) in the Pacific Northwestern United States. Mol Ecol.

Miller MP, Blinn DW, and Keim P, 2002. Correlations between observed dispersal capabilities and patterns of genetic differentiation in four aquatic insect species from the Arizona White Mountains, USA. Freshwater Biol 47:1660–1673.

Monmonier MS, 1973. Maximum-difference barriers: an alternative numerical regionalization method. Geogr Anal 5:245–261.

Nei M, 1972. Genetic distance between populations. Am Nat 106:283–292.

Nei M, 1973. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323.

Nei M, 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590.

Pielou EC, 1977. Mathematical ecology. New York: John Wiley & Sons.

Pommerening A, 2002. Approaches to quantifying forest structures. Forestry 75:305–324.

Raymond ML and Rousset F, 1995. An exact test for population differentiation. Evolution 49:1280–1283.

Reynolds J, Weir BS, and Cockerham CC, 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779.

Roff DA and Bentzen P, 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol Biol Evol 6:539–45.

Slatkin M, 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Sokal RR and Oden NL, 1978a. Spatial autocorrelation analysis in biology. 1. Methodology Biol J Linn Soc 10:199–228.

Sokal RR and Oden NL, 1978b. Spatial autocorrelation analysis in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. Biol J Linn Soc 10:229–249.

Watson DF, 1992. Contouring: a guide to the analysis and display of spatial data. New York: Pergamon Press.

Watson DF and Philips GM, 1985. A refinement of inverse distance weighted interpolation. Geo-Processing 2:315–327.

Weir BS and Cockerham CC, 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.