Computer Note

IBD (Isolation by Distance): A Program for Analyses of Isolation by Distance

A. J. Bohonak

The genetic similarity among individuals or populations can be ascertained using a number of statistical techniques (reviewed by Bohonak 1999; Neigel 1997; Roderick 1996; Slatkin 1985). When populations can be defined a priori, one option is to analyze genetic "isolation by distance" (sensu Wright 1943) by plotting the genetic similarity (or distance) among population pairs as a function of the geographic distance between those pairs. Slatkin (1993) suggested the genetic distance $\hat{M} = (1/F_{ST} - 1)/4$ as an appropriate similarity measure, although other approaches are possible (e.g., Epperson and Li 1996).

Qualitative and statistical analyses of isolation by distance can reveal much about population genetic structure. The primary use for plots of (genetic) isolation by (geographic) distance is to assess whether more distant population pairs are more different genetically. However, these plots can also be used to test the validity of simpler models of population structure (e.g., island or hierarchical island models). Isolation by distance analyses may help separate the effects of population history from ongoing gene flow, and test the explanatory power of alternative dispersal pathways (Slatkin 1994). For example, one might assess whether the distance along a river or a topographic isocline is more biologically relevant than distance "as the crow flies." The influence of geographic features or specific life-history traits on population differentiation can also be tested. Peterson and Denno (1998) contrasted isolation by distance slopes and intercepts in species with different dispersal abilities.

IBD version 1.1 is a program written in C and compiled for Macintosh and Windows that can be used for analyses of isolation by distance. This program provides a number of unique features: isolation by distance slopes and intercepts are calculated using reduced major axis (RMA) regression, confidence intervals are generated based on several different assumptions regarding data structure, and statistical significance is determined using Mantel tests. The program is freely available at http://www.bio.sdsu.edu/pub/andy/IBD.html.

Rationale

Studies of isolation by distance typically seek to ascertain (1) whether there is a statistically significant relationship between genetic distance (or similarity) and geographic distance, and (2) the strength of this relationship. Significance is usually assessed by asking whether the pairwise genetic distance matrix is correlated with the pairwise geographic distance matrix using a Mantel test (see Manly 1994). For the genetic distance matrix **A** and the geographic distance matrix **B,** the test statistic is calculated as $Z = \Sigma_{i,j} A_{ij} B_{ij}$. IBD also reports an alternative statistic, *r,* which provides a standardized Z that ranges from -1 to 1 (Manly 1994). Significance is assessed by comparing Z_{actual} to a distribution of Z scores obtained by randomizing rows/columns of the **B** matrix and holding **A** constant. The IBD application provides one-tailed P values for this distribution. Because matrix rows (populations) are treated as a single unit, the Mantel test is a more appropriate way to assess significance than alternatives which assume that each population pair is independent.

A logical way to quantify the strength of the isolation by distance relationship is to calculate the slope and intercept of genetic similarity or distance against geographic distance. Based on simulations, Hellberg (1994) suggested that RMA regression is more appropriate for this purpose than standard ordinary least squares (OLS) regression. (In general, RMA is less biased when the independent variable is measured with error). McArdle (1988) suggests that RMA be used when the error rate in x exceeds one-third of the error rate in y. IBD calculates the RMA slope and intercept using the formulas provided by Sokal and Rohlf (1981).

Input File Format

IBD reads generic ASCII (text) files. The input file must be a text file named "IBD.#" (where # is replaced by a number) and must be in the same folder as the application. Two file formats are recognized; each can be generated by saving a text-only file from a spreadsheet application. For the pairwise distance format, genetic distances are entered on single lines as <population A> <population B> <genetic distance>, where "population A/B" is replaced by a number. Pairwise geographic distances are then entered line by line in the same manner.

For diploid, codominant markers (e.g., allozymes, microsatellites), genotypes may also be entered in a raw data format. Each line of the input file lists the genotypes at all loci for a single individual, beginning with the population number. Diploid genotypes at each locus are designated with two numbers separated by a comma (e.g., 1,1 2,3 4,6). Missing genotypes at a locus are coded 0,0. The end of the genetic data is indicated by a population number of 0, followed by genotypes of 0,0. The geographic distances between all population pairs are then entered as described for the pairwise distance format.

Program maxima for IBD consist of 100 populations, 30 loci, 20 alleles per locus, 1000 individuals per population, and 1×10^5 randomizations/bootstraps (see below). The IBD application folder contains

example files and a manual with more detailed information on input file formats.

Output

For the raw data format only, IBD provides (1) allele counts and heterozygosity for each population and locus, (2) locus-specific and overall F_{ST} for each population pair, estimated using the methods of Weir (1990), and (3) Slatkin's (1993) similarity measure $\hat{M} = (1/F_{ST} - 1)/4$ for each population pair.

If geographic distances are available from either input file format, IBD will perform a Mantel test as described above, using the number of matrix randomizations requested. The slope and intercept from RMA regressions are calculated following Sokal and Rohlf (1981). When raw genotypic data are entered, the dependent variable is $\hat{M}$, otherwise the genetic distances provided in the input file are used. Error estimation for RMA regression is considered using five methods:

1. Standard linear model formulas (Sokal and Rohlf 1981).
2. Jackknife over population pairs (i.e., each point on the graph): one-delete jackknife estimates of the slope, intercept, and associated standard errors are calculated following Weir (1990). The 95% and 99% confidence intervals are provided for each.
3. One-delete jackknife over populations.
4. Bootstrapping over population pairs: confidence intervals are calculated by creating new "pseudoreplicate" datasets, each with the same number of population pairs, by random sampling with replacement. The middle 95% and 99% of the bootstrap pseudoreplicates constitute the confidence intervals.
5. Bootstrapping over independent population pairs: random datasets are created by sampling completely independent population pairs. For p populations, each dataset will contain $p/2$ population pairs if p is even, or $(p - 1)/2$ pairs if p is odd. For example, if $p = 6$ populations, then one pseudoreplicate might consist of {(1,4), (3,6), (2,5)}.

Finally, the genetic and geographic distances are log-transformed (following Slatkin 1993) and the RMA analyses are repeated.

Statistical Considerations

As noted by numerous authors, pairwise contrasts in the isolation by distance relationship are not independent; a single population will be involved in multiple contrasts. The Mantel test is expected to provide an appropriate test of significance for isolation by distance because it appropriately considers the unit of replication to be a population (and not a pairwise contrast). Similarly, to generate confidence limits for the RMA slope or intercept, bootstrapping over independent population pairs would seem to be the most conservative approach (see above). For a small number of populations, jackknifing over populations provides the next best alternative. IBD is the only currently available software package that provides confidence limits for isolation by distance slopes and intercepts using the population as the unit of replication.

From the Department of Biology, San Diego State University, San Diego, CA 92182-4614. I thank Neil Davies, Francis X. Villablanca, and especially George Roderick for constructive commentary and support. IBD is written in C, compiled using CodeWarrior for Macintosh, and output is via CodeWarrior's SIOUX module. IBD can be downloaded from http://www.bio.sdsu.edu/pub/andy/IBD.html. Source code will be made available upon request. Address correspondence to A. J. Bohonak at the address above or e-mail: bohonak@sciences.sdsu.edu.

References

Bohonak AJ, 1999. Dispersal, gene flow and population structure. Q Rev Biol 74:21–45.

Epperson BK and Li T, 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proc Natl Acad Sci USA 93: 10528–10532.

Hellberg ME, 1994. Relationships between inferred levels of gene flow and geographic distance in a philopatric coral, *Balanophyllia elegans.* Evolution 48:1829–1854.

Manly BFJ, 1994. Multivariate statistical methods: a primer, 2nd ed. New York: Chapman & Hall.

McArdle BH, 1988. The structural relationship: regression in biology. Can J Zool 66:2329–2339.

Neigel JE, 1997. A comparison of alternative strategies for estimating gene flow from genetic markers. Annu Rev Ecol Syst 28:105–128.

Peterson MA and Denno RF, 1998. The influence of dispersal and diet breadth on patterns of genetic isolation by distance in phytophagous insects. Am Nat 152:428–446.

Roderick GK, 1996. Geographic structure of insect populations: gene flow, phylogeography, and their uses. Annu Rev Entomol 41:325–352.

Slatkin M, 1985. Gene flow in natural populations. Annu Rev Ecol Syst 16:393–430.

Slatkin M, 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264–279.

Slatkin M, 1994. Gene flow and population structure. In: Ecological genetics (Real LA, ed). Princeton, NJ: Princeton University Press; 3–17.

Sokal RR and Rohlf FJ, 1981. Biometry, 2nd ed. New York: WH Freeman.

Weir BS, 1990. Genetic data analysis: methods for discrete population analysis. Sunderland, MA: Sinauer.

Wright S, 1943. Isolation by distance. Genetics 28:114–138.

Received June 21, 2001
Accepted December 31, 2001

Corresponding Editor: Stephen J. O'Brien