Molecular Ecology (2000) **9**, 487–488

PROGRAM NOTE

GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes

D. POSADA,* K. A. CRANDALL* and A. R. TEMPLETON†

*Department of Zoology, Brigham Young University, Provo, UT 84602, USA,
†Department of Biology, Washington University, St.Louis, MO 63130, USA

Keywords: cladistic nested analysis, historical processes, phylogeography, population genetics, software

Received 21 June 1999; revision received 12 October 1999; accepted 23 October 1999

Correspondence: David Posada. Fax: +1 801 378 7423; E-mail: dp47@email.byu.edu

The central focus of population genetics is the study of the distribution of the genetic variation within and among populations. This endeavour has often been accomplished by the use of genealogies upon which geographical information is incorporated in the search of association among genetic variation and geographical distribution (see Avise 1998). However, a particular population genetic structure can be the result of distinct processes acting in different points through time and space and may reflect historical rather than ongoing population level processes (Gerber & Templeton 1996). Templeton (1993) and Templeton *et al*. (1995) describe a methodology (cladistic nested analysis) in which population structure can be separated from population history when it is assessed through rigorous and objective statistical tests upon an estimated nested cladogram (see Templeton *et al*. 1992).

GeoDis is a computer program that implements the cladistic nested analysis. The simplest test for geographical association is to treat sample locations as categorical variables. An exact permutational contingency test is performed for any clade at each nesting level. A chi-square statistic is calculated from the contingency tables in which rows are genetic clades and columns are geographical locations (see also software Chiperm, available at http://bioag.byu.edu/zoology/crandall_lab/programs.htm). A more elaborate analysis can also be carried out by using information on geographical distances. Using the geographical coordinates of each population two main statistics are calculated, the clade distance (D_c), which measures the geographical spread of a clade, and the nested clade distance (D_n), which measures how a clade is geographically distributed relative to other clades in the same higher-level nesting category. In the case of riparian or coastal species, or in the case of species with constrained dispersal routes, a matrix of pairwise distances among the different locations better describes their geographical distribution. The analogue statistics (D_{cl} and D_{nl}) are calculated as the average pairwise distances between members of the same focal clade and the average pairwise distances between members of the focal clade with all members of the nesting clade (including the focal clade). An interior-tip statistic (I-T) is also estimated within each nested category as the average interior distance minus the average tip distance. For the calculation of these averages, each clade distance is weighted by the number of copies in that focal clade relative to the total number of copies in the nesting clade. This tip vs. interior contrast corresponds to a young vs. old contrast and, to a lesser extent, rare vs. common (Crandall & Templeton 1993). If the haplotype tree is rooted, say by an out-group, the user can also specify which haplotype is the oldest by designating it as the 'interior', and regarding the younger haplotypes all as 'tips'. When root probabilities or out-group weights for the cladogram are specified (Castelloe & Templeton 1994), the correlation of both distance measures with out-group weights within each nested category is also estimated.

The significance of these statistics is estimated through a Monte Carlo procedure. Null distributions are constructed by randomizing the contingency data table for each clade and nesting level and estimating again the test statistics for each randomized data set. Matrix randomization is accomplished by using the algorithm of Roff & Bentzen (1989), which preserves the marginals of the table (clade frequencies and sample sizes), while permuting the individual cells. A minimum number of 1000 random permutations are recommended to make statistical inference at the 5% level of significance (Edgington 1986). The output of GeoDis consists of the calculated statistics and their associated permutational *P*-values. Templeton (1998) provides a key for the interpretation of these results in biological terms.

GeoDis has been written both in C and Java and includes new features, as weighted I-T statistics, and the possibility of using user-defined distances. A previous version of the program written in VAX/VMS Basic exists (AR Templeton). The C program prompts the user for all the options needed to run the program. The Java program provides an interface where the user selects the input and output files, the number of permutations, the possibility of using out-group weights, decimal degrees, and/or user-defined distances. The input file consists of the population information plus the description of the nested cladogram. Details are given in the program documentation. The GeoDis package, containing executables for Macintosh, PC, and Unix machines, documentation, and source code in Java and C is available for free from http://bioag.byu.edu/zoology/crandall_lab/programs.htm.

References

Avise JC (1998) The history and purview of phylogeography: a personal reflection. *Molecular Ecology*, **7**, 371–379.

Castelloe J, Templeton AR (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution*, **3**, 102–113.

Crandall KA, Templeton AR (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, **134**, 959–969.

Edgington ES (1986) *Randomization Tests*, 2nd edn. Marcel Dekker, New York.

Gerber AS, Templeton AR (1996) Population sizes and within-deme movement of *Trimerotropis saxatilis* (Acrididae), a grasshopper with a fragmented distribution. *Oecologia*, **105**, 343–350.

Roff DA, Bentzen P (1989) The statistical analysis of mitochondrial DNA polymorphisms: Chi-square and the problem of small samples. *Molecular Biology and Evolution*, **6**, 539–545.

Templeton AR (1993) The 'Eve' hypothesis: a genetic critique and reanalysis. *American Anthropologist*, **95**, 51–72.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.