

PROGRAM NOTE

Nested clade analysis statistics

DAVID POSADA,* KEITH A. CRANDALL† and ALAN R. TEMPLETON‡

*Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, 36310 Vigo, Spain, †Department of Microbiology and Molecular Biology, Brigham Young University, 84602 Provo, Utah, USA, ‡Department of Biology, Washington University, St. Louis, Missouri 63130, USA

Abstract

Nested clade analysis (NCA) is a flexible and powerful method to study the phylogeography of species and populations, implemented in the software GEODIS. Despite the popularity of this method, an explicit description of the exact equations used to compute the NCA statistics has never been published. Given the importance of the methodology and increased interest in exactly how it works, here we describe the exact equations implemented in the program GEODIS for the calculation of these statistics.

Keywords: GEODIS, NCA, nested clade analysis, phylogeography, population genetics

Received 8 February 2006; revision accepted 17 February 2006

The nested clade analysis (NCA) (Templeton *et al.* 1995; Templeton 1998, 2004) is a widely used method in phylogeographic studies, currently implemented in the program GEODIS (Posada *et al.* 2000), available at <http://darwin.uvigo.es>. A search in the ISI Web of Science with the term 'nested clade analysis' retrieves 171 records in 51 different journals, of which 70 (40.9%) correspond to *Molecular Ecology*. On the other hand, the program GEODIS has been cited 265 times in 66 different journals, of which 97 (36.6%) correspond to *Molecular Ecology* articles, having almost 2000 registered users. Despite this popularity, and for editorial reasons, we have not previously published detailed descriptions of the NCA statistics. Given the importance of the methodology and increased interest in exactly how it works, we have been invited to describe the exact equations implemented in the program GEODIS for the calculation of the NCA statistics.

Before using the program GEODIS, the first step of the NCA is the estimation of a haplotype tree from restriction site data or from an alignment of nucleotide sequences. The estimation of this cladogram can be accomplished by the algorithm given in Templeton *et al.* (1992), implemented in the program rcs (Clement *et al.* 2000). Once this cladogram or set of cladograms has been estimated, they are converted to a nested design, in which haplotypes, 0-step clades, are grouped into 1-step clades, and these into 2-step clades and so on, until the next level of nesting includes the whole cladogram. The basic nesting rules are described in

Templeton *et al.* (1987) and refined in Templeton & Sing (1993) and Crandall (1996). This nesting design is robust to ambiguities in the cladogram including those generated by recombination (Templeton *et al.* 2000a, b; Antunes *et al.* 2002). However, the nesting rules given in Templeton & Sing (1993) that deal with cladogram ambiguity achieve invariance at the price of lower power. An alternative in dealing with cladogram ambiguity is to use only those nesting rules that apply to fully resolved cladograms, but apply them exhaustively to all possible nesting designs to determine which inferences are robust to cladogram ambiguity (Pfenninger & Posada 2002; Brisson *et al.* 2005).

The program GEODIS implements several statistical tests on the nested cladogram. A simplest test for the geographical association of haplotypes or clades of haplotypes is to treat sample locations as categorical variables. For a given clade at any nesting level, an exact permutational contingency test can be performed (Hudson *et al.* 1992), where a chi-squared value is calculated from a table in which rows are clades and columns are geographical locations. However, a more powerful analysis can be implemented by using information on geographical distances. 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. These statistics can be computed only for clades where both genetic and geographical variation exists. When the input includes the coordinates of each population, geographical distances are calculated using

Correspondence: David Posada, Fax: +34 986812556;

E-mail: dposada@uvigo.es

the standard formula for great circle distances. For each focal clade x , which without loss of generality is at the y -step level,

$$D_c = \sum_{i=1}^K a_i \cdot \omega \cdot \text{acos}[\sin \phi_i \cdot \sin \phi + \cos \phi \cdot \cos \phi_i \cdot \cos(\gamma - \gamma_i)] \quad (\text{eqn 1})$$

$$D_n = \sum_{i=1}^K a_i \cdot \omega \cdot \text{acos}[\sin \phi_i \cdot \sin \Phi + \cos \phi_i \cdot \cos \Phi \cdot \cos(\Gamma - \gamma_i)] \quad (\text{eqn 2})$$

where $a_i = \frac{x_i/n_i}{\sum_{i=1}^K x_i/n_i}$ (eqn 3)

$$\phi = \sum_{i=1}^K a_i \cdot \phi_i \quad (\text{eqn 4})$$

$$\gamma = \sum_{i=1}^K a_i \cdot \gamma_i \quad (\text{eqn 5})$$

$$\Phi = \sum_{i=1}^K \sum_{r=1}^R \frac{x_{ir}/n_i}{n_{y+1}} \cdot \phi_i \quad (\text{eqn 6})$$

and $\Gamma = \sum_{i=1}^K \sum_{r=1}^R \frac{x_{ir}/n_i}{n_{y+1}} \cdot \gamma_i$ (eqn 7)

where in turn $n_{y+1} = \sum_{i=1}^K \sum_{r=1}^R x_{ir}$ (eqn 8)

Here, K is the number of sampled locations, R is the number of clades nested in the $y + 1$ step clade, a_i is the relative abundance of clade x at location i , ω is the average radius of the Earth in kilometres, ϕ_i is the latitude of location i , ϕ is the average latitude of clade x weighted by the relative abundance a_i at each location, Φ is the average latitude of the $y + 1$ step clade within which the focal clade x is nested, γ_i is the longitude of clade x at location i , γ is the average longitude of clade x weighted by the relative abundance a_i at each location, Γ is the average longitude of the $y + 1$ step clade within which the focal clade x is nested, x_i is the number of copies of clade x in location i , x_{ir} is the number of copies in location i of clade r , n_i is the sample size at location i , and n_{y+1} is the total number of copies of the $y + 1$ step clade. Importantly, because sampling effort and sample size are seldom homogeneous in practice, in GEODIS we weight the contribution of each clade by its *relative abundance* (Fig. 1). Note that Fig. 1 in Templeton *et al.* (1995) was used only to illustrate the biological meaning of the distances and it does not make use of the relative abundances.

The calculations given previously assume that there are not restrictions for the movement between locations. However, many species may be constrained in their dispersal routes, and distances between locations are not adequately measured simply through geographical coordinates (e.g. Fetzner & Crandall 2003). In these cases, a matrix of pairwise distances among the different locations better describes their geographical distribution. Here, the analogue (but not identical) statistics are the average pairwise distances between members of the same focal clade, x , and the average pairwise distances between members of the focal clade with all members of the $y + 1$ step nesting clade (including the focal clade):

$$D_c = \frac{\sum_{i=1}^K \sum_{j \neq i}^K x_i x_j D_{ij}}{\sum_{i=1}^K \frac{x_i(x_i - 1)}{2} + \sum_{i=1}^K \sum_{j \neq i}^K x_i x_j} \quad (\text{eqn 9})$$

$$D_n = \frac{\sum_{i=1}^K \sum_{j \neq i}^K x_i X_j D_{ij}}{\sum_{i=1}^K \left[\frac{x_i(x_i - 1)}{2} + x_i(X_i - x_i) \right] + \sum_{i=1}^K \sum_{j \neq i}^K x_i X_j} \quad (\text{eqn 10})$$

where K is the total number of locations, D_{ij} is the user input distance between locations i and j , x_i is the number of copies of clade x at location i and X_i is the number of copies at location i of the $y + 1$ step clade within which the focal clade is nested.

In an intraspecific cladogram, interior clades tend to be older and more frequent than tip clades (Watterson 1976; Donnelly & Tavaré 1986). This information can be used for inferring population processes. So, within each nested category, an interior-tip statistic is estimated for both types of distances ($I-T_c$ and $I-T_n$), as the average interior distance minus the average tip distance. In this calculation, each clade distance is weighted by the number of copies in the focal clade relative to the total number of interior/tip copies in the nesting clade, respectively.

$$I-T_c = \sum_{r=1}^R \left[(1 - \beta_r) \cdot D_{c_r} \frac{\sum_{i=1}^K x_{ir}}{\sum_{s=1}^S \left((1 - \beta_s) \cdot \sum_{i=1}^K x_{is} \right)} \right] - \sum_{r=1}^R \left[\beta_r \cdot D_{c_r} \frac{\sum_{i=1}^K x_{ir}}{\sum_{s=1}^S \left(\beta_s \cdot \sum_{i=1}^K x_{is} \right)} \right] \quad (\text{eqn 11})$$

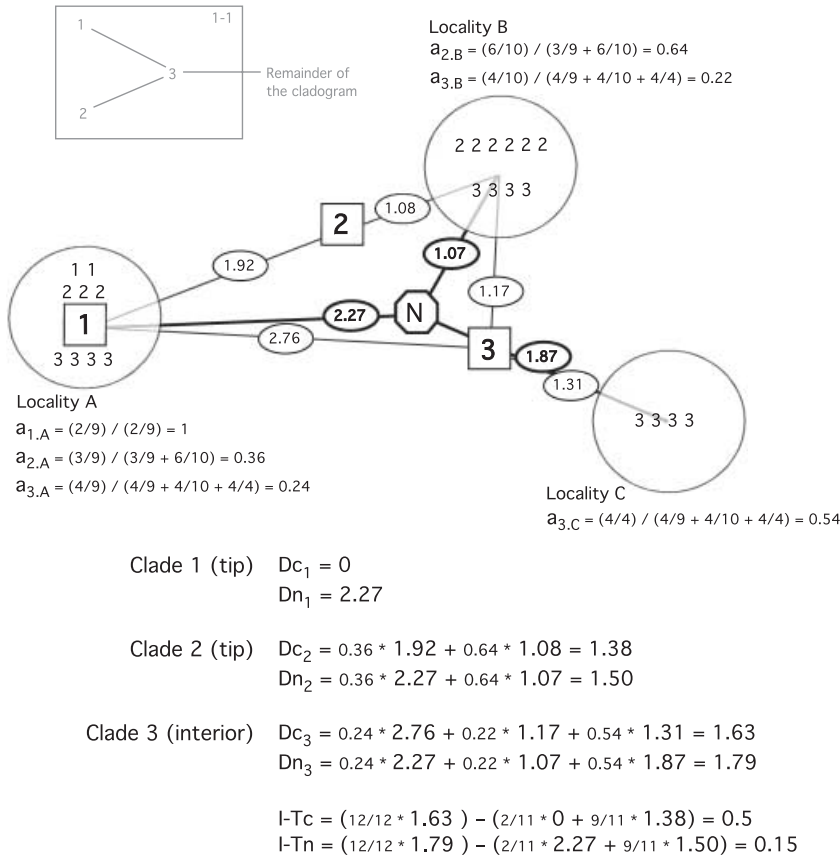


Fig. 1 Calculation of the NCA statistics. This example corresponds to the same data in Fig. 1 in Templeton *et al.* (1995), but in this case clade contributions are weighted by their relative abundance at each locality (a), as implemented in GEODIS. Three sampling areas (localities) on an island are indicated by the letters A, B and C. Within each locality, the number of times haplotypes of type 1, 2 and 3 sampled is indicated by the number of times their respective haplotype numbers appear within the circle centred at the locality. These three haplotypes are all within a common nested clade, as indicated at the top of the figure. Haplotypes 1 and 2 are tips and haplotype 3 is interior. The geographical centre of a particular haplotype is indicated by a box containing the number of the haplotype. The geographical centre of the entire nested clade is indicated by the hexagon enclosing the letter N. The great circle distance between these geographical centres and the localities are indicated by the numbers enclosed on an oval on the line connecting the geographical locations under consideration. Relative abundances for each haplotype at each location are indicated as $a_{haplotype,location}$. NCA statistics for each clade are indicated at the bottom.

$$I-T_n = \sum_{r=1}^R \left[(1-\beta_r) \cdot D_{n_r} \frac{\sum_{i=1}^K x_{is}}{\sum_{s=1}^S \left((1-\beta_s) \cdot \sum_{i=1}^K x_{is} \right)} \right] - \sum_{r=1}^R \left[\beta_r \cdot D_{n_r} \frac{\sum_{i=1}^K x_{ir}}{\sum_{s=1}^S \left(\beta_s \cdot \sum_{i=1}^K x_{is} \right)} \right] \quad (\text{eqn 12})$$

where $R = S$ are the number of focal clades r or s within the nested category, β_r is 0 if clade r or s is interior and 1 if clade r or s is terminal, K is the number of locations, x_{ir} is the number of copies at location i of clade r and x_{is} is the number of copies at location i of clade s . In cases where the cladogram is reliably rooted through an outgroup such that temporal polarity is known, the tips refer to the more recent clades within a higher order nesting category and the interiors refer to the older or oldest clade within a higher order nesting category.

Castelloe & Templeton (1994) provided a heuristic formula for calculating the probability of a haplotype being the root or outgroup for the cladogram. When root probabilities or outgroup weights for the cladogram are specified, the correlation of both distance measures with outgroup weights is also estimated within each nested category. These statistics, ρ_c and ρ_n , are defined as:

$$\rho_c = \frac{\sum_{r=1}^R (D_{c_r} \cdot w_r) - R \cdot \bar{D}_c \cdot \bar{w}}{\sqrt{\left[\sum_{r=1}^R (D_{c_r}^2) - R \cdot \bar{D}_c^2 \right] \cdot \left[\sum_{i=1}^C (w_i^2) - R \cdot \bar{w}^2 \right]}} \quad (\text{eqn 13})$$

$$\rho_n = \frac{\sum_{r=1}^R (D_{n_r} \cdot w_i) - R \cdot \bar{D}_n \cdot \bar{w}}{\sqrt{\left[\sum_{r=1}^R (D_{n_r}^2) - R \cdot \bar{D}_n^2 \right] \cdot \left[\sum_{i=1}^R (w_i^2) - R \cdot \bar{w}^2 \right]}} \quad (\text{eqn 14})$$

where R is the number of focal clades r within the nested category, and w_r is the outgroup weight for clade r and

$$\bar{D}_c = \frac{\sum_{r=1}^R D_{c_r}}{R} \quad (\text{eqn 15})$$

$$\bar{D}_n = \frac{\sum_{r=1}^R D_{n_r}}{R} \quad (\text{eqn 16})$$

$$\bar{w} = \frac{\sum_{r=1}^R w_r}{R} \quad (\text{eqn 17})$$

Finally, the statistical significance of all the described statistics is estimated through a Monte Carlo procedure. Null distributions are constructed by randomizing the data table for each clade and nesting level using the algorithm of Roff & Bentzen (1989), that preserves haplotype frequencies and sample sizes, and estimating again the test statistics for each randomized data set. An updated inference key that incorporates the effect of inadequate sampling is available at the GEODIS website to make the biological inference of the NCA statistics and their significance more consistent.

Acknowledgements

DP was supported by the 'Ramón y Cajal' programme of the Spanish government.

References

- Antunes A, Templeton AR, Guyomard R, Alexandrino P (2002) The role of nuclear genes in intraspecific evolutionary inference: genealogy of the transferrin gene in the brown trout. *Molecular Biology and Evolution*, **19**, 1272–1287.
- Brisson JA, De Toni DC, Duncan I, Templeton AR (2005) Abdominal pigmentation variation in *Drosophila polymorpha*: geographic variation in the trait, and underlying phylogeography. *Evolution*, **59**, 1046–1059.
- Castelloe J, Templeton AR (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution*, **3**, 102–113.
- Clement M, Posada D, Crandall KA (2000) rcs: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1659.
- Crandall KA (1996) Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. *Molecular Biology and Evolution*, **13**, 115–131.
- Donnelly P, Tavaré S (1986) The ages of alleles and a coalescent. *Advances in Applied Probability*, **18**, 1–19.
- Fetzner JW Jr, Crandall KA (2003) Linear habitats and the nested clade analysis: an empirical evaluation of geographic versus river distances using an Ozark crayfish (Decapoda: Cambaridae). *Evolution*, **57**, 2101–2118.
- Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.
- Pfenninger M, Posada D (2002) Phylogeographic history of the land snail *Candidula unifasciata* (Helicellinae, Stylommatophora): fragmentation, corridor migration, and secondary contact. *Evolution*, **56**, 1776–1788.
- Posada D, Crandall KA, Templeton AR (2000) GEODIS: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Molecular Ecology*, **9**, 487–488.
- 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 (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.
- Templeton AR (2004) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.
- Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.
- Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics*, **117**, 343–351.
- 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.
- Templeton AR, Clark AG, Weiss KM *et al.* (2000a) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *American Journal of Human Genetics*, **66**, 69–83.
- Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000b) Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics*, **156**, 1259–1275.
- Watterson GA (1976) Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theoretical Population Biology*, **10**, 239–253.