BIOINFORMATICS APPLICATIONS NOTE Vol. 00 no. 00 2006, pages 1–2
doi:10.1093/bioinformatics/btl614

Phylogenetics

The automation of Nested Clade Phylogeographic Analysis

Mahesh Panchal
School of Biological Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK

Received on September 21, 2006; revised on November 25, 2006; accepted on November 28, 2006
Advance Access publication . . .
Associate Editor: Keith A Crandall

ABSTRACT

Summary: ANeCA is a fully automated implementation of Nested Clade Phylogeographic Analysis. This was originally developed by Templeton and colleagues, and has been used to infer, from the pattern of gene sequence polymorphisms in a geographically structured population, the historical demographic processes that have shaped its evolution. Until now it has been necessary to perform large parts of the procedure manually. We provide a program that will take data in Nexus sequential format, and directly output a set of inferences. The software also includes TCS v1.18 and GeoDis v2.2 as part of automation.

Availability: The software is available free of charge from http://www.rubic.rdg.ac.uk/~mahesh/software.html. The program is written in Java and requires the Java 1.4 Runtime Environment (or later) to run. The source code is included in the package, and includes the source from TCS and GeoDis. ANeCA, TCS and GeoDis are released under the GNU General Public License.

Contact: m.panchal@rdg.ac.uk

Nested Clade Phylogeographic Analysis (NCPA) (Templeton *et al.*, 1995) was proposed as a method for disentangling the historical processes that determine a species evolution and geographical distribution. It attempts to distinguish between a wide variety of processes such as restricted gene flow, fragmentation and range expansion. These inferences are made with a published inference key available at http://darwin.uvigo.es/software/geodis.html.

In recent years there has been some debate as to effectiveness of NCPA (e.g. Knowles and Maddison, 2002; Templeton, 2004). We describe here software that automates the complex NCPA analysis, written as part of a larger project to evaluate the NCPA methodology (see M. Panchal and Beaumont, submitted for publication). The software is written in Java, and requires the Java 1.4 Runtime Environment or better to run. To fully automate NCPA, TCS v1.18 (Clement *et al.*, 2000) and GeoDis v2.2 (Posada *et al.*, 2000) are incorporated, to avoid re-programming commonly used software in the analysis (the latest versions of each application are available at http://darwin.uvigo.es/). TCS is used to generate a haplotype network using Statistical Parsimony (Templeton *et al.*, 1992), although this is not the only method of haplotype network estimation. Ways in which other methods of haplotype network estimation can be applied are included in the documentation. GeoDis calculates distance statistics and their significance (Posada *et al.*, 2006), which are used by the inference key to determine the events and processes that determined the species history. The default parameters of GeoDis have been changed to read longitude and latitude as decimal degrees, and the number of permutations have been increased from 1000 to 10 000, in order to reduce the variance associated with the estimated *P*-values. The role of ANeCA is to automate the nesting algorithm and the published inference key, and to link together the various components. An alternative partially automated inference key, AUTOINFER 1.0 (Zhang *et al.*, 2006), uses a similar approach but allows user-input as inference key questions are answered (ANeCA requires this information specified before application of the automated key or uses a default value).

The first stage utilizes TCS. The input to TCS is a Nexus sequential file that has been modified in the following way. The label for each individual DNA sequence is appended with a full stop and the identification number of the location it came from. This has no effect on the TCS analysis, but neatly includes the geographic information for the nesting algorithm stage. The identification number is specified in another file that contains geographic information. The geographic file contains the identification number, shorthand name, sample size, latitude, longitude and radius of each sampled location and unsampled habitable area within the study area. The longitude and latitude should be written in decimal degrees and the radius specified in km. The radius measures the spread of an area.

The next stage in ANeCA is the nesting algorithm. This requires the graph file from TCS and the geographic information file. The nesting algorithm reconstructs the haplotype network from the graph file and creates a population distribution profile for each haplotype. Following the rules described in M. Panchal and Beaumont (submitted for publication), the nesting algorithm proceeds to create a nested design, by hierarchically clustering haplotypes known as 'clades' (Fig. 1 illustrates the nesting algorithm on a simple haplotype network). From this nested design, four output files are created. The first is the input file for GeoDis where geographic information is given as latitude and longitude in decimal degrees format. The second file is also a GeoDis input file, but geographic information is given in the form of a matrix of (great circle) distances (in km). The third output file is a representation of the nested design in the GML syntax (the format used by TCS to write the graph file). The fourth file (the nest file) is optionally written, and contains a summary of the nested design. Currently there is no software to visually inspect the nested design. As a result it must be manually drawn onto a diagram of the haplotype network (an option in TCS is available to save a diagram of the haplotype network). The nest file provides the user with the identification numbers' of each haplotype/clade and the order each haplotype/clade was nested, allowing them to verify the nested design.

The third stage calculates the distance statistics with GeoDis (Posada *et al.*, 2006), and their significance through a permutation

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

procedure. The final stage is to run the automated inference key. This requires the GeoDis input, the GeoDis output, the GML nested design and the geographic information file. The automated inference key is a modified version of the inference key dated July 14, 2004. Modifications include introducing quantifiers into questions, calculating clade boundaries as convex hulls of the locations represented within a clade and rewording questions to remove subjectivity within them (see M. Panchal and Beaumont, submitted for publication). The automated inference key produces two output files. The first is a summary of the population distributions, and the distance statistics and their significance. The second output file contains the series of questions followed and the inference for each clade Movement of individuals/groups is not included and must be traced manually using the summary file.

ANeCA is simple to apply includes both graphical and command line interfaces for ease of use. The software makes applying NCPA much faster, repeatable and opens up new scope for investigation. Uses of the software include analysis of various species, exploration of demographic models via simulations and the improvement of NCPA by providing a benchmark with which to make comparisons. The software is freely available to download at http://www.rubic.rdg.ac.uk/~mahesh/software.html.

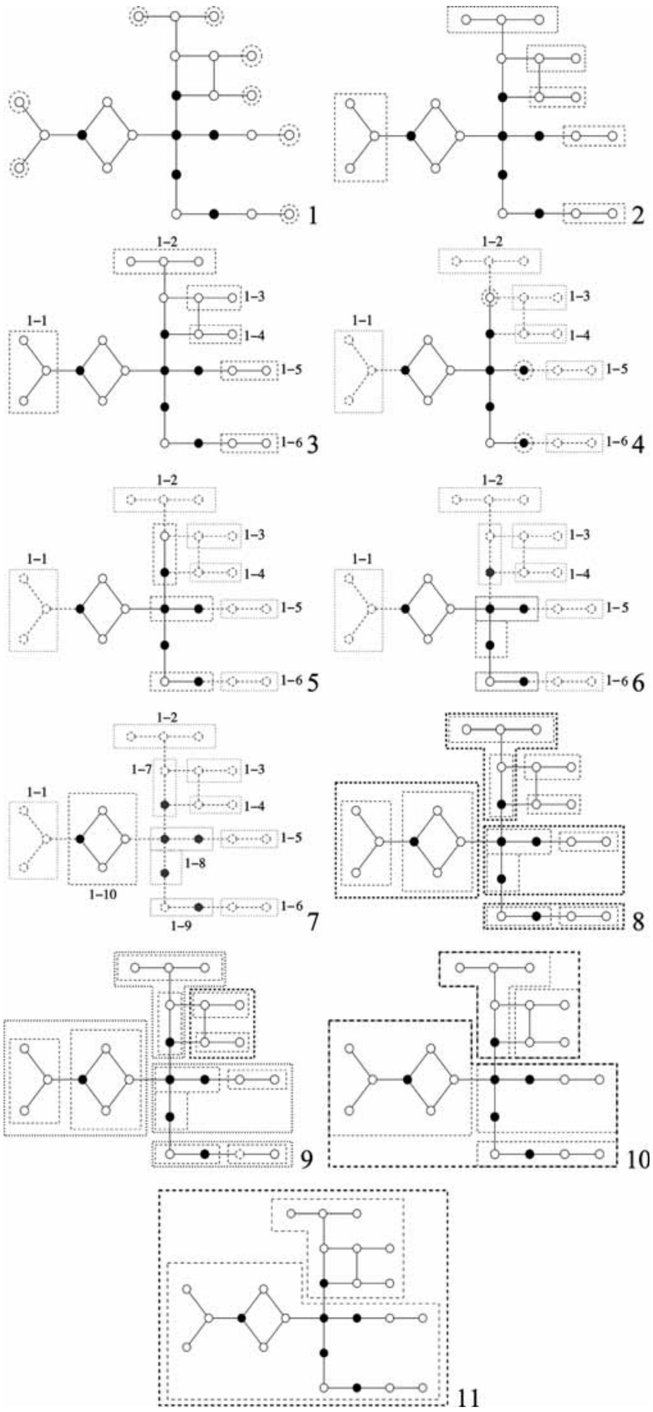

Fig. 1. An application of the nesting algorithm to a simple haplotype network. White circles indicate sampled haplotypes and black circles are missing intermediate haplotypes inferred by a program such as TCS. (1) Identify the tips. (2) Nest the tips to the connecting haplotype. (3) Label the clades. (4) Prune the clades and identify the next set of 'tips'. (5) Nest the 'tips' to the connecting clades. (6) Nest the symmetrically stranded clade to the connecting clade with the lowest sample size. (7) Nest remaining loop in a clade of its own. (8) Begin nesting again at the next level. (9) Prune and nest again. (10) Proceed to nesting at the next level. (11) Terminate the nesting at the Total Cladogram, which encompasses all the haplotypes.

ACKNOWLEDGEMENTS

I would like to thank M. A. Beaumont for his guidance and comments on this article. I would also like to thank M. Pulquério, O. S. Paulo, A. R. Templeton, K. A. Crandall, D. Posada, J. Gómez-Zurita, S. Creer, I. Carbone, P. Mardulyn, L. L. Knowles and the anonymous participants in the questionnaire regarding the use of NCPA, for their discussions on the methodology and application of NCPA. Further gratitude goes to M. Pulquério, O. Moya, I. Phillipsen, S. O'Loughlin, J. Gómez-Zurita and G. Segelbacher for testing and providing feedback regarding the software. I would also like to thank K. Crandall and two anonymous reviewers for providing comments on the manuscript and software. The Biotechnology and Biological Sciences Research Council is also gratefully acknowledged for funding.

Conflict of Interest: none declared.

REFERENCES

Clement,M., Posada,D. and Crandall,K.A. (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1659.

Knowles,L.L. and Maddison,W.P. (2002) Statistical phylogeography. *Mol. Ecol.*, **11**, 2623–2635.

Posada,D., Crandall,K.A. and Templeton,A.R. (2000) GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.*, **9**, 487–488.

Posada,D., Crandall,K.A. and Templeton,A.R. (2006) Nested clade analysis statistics. *Mol. Ecol. Notes*, **6**, 590–593.

Templeton,A.R., Crandall,K.A. and Sing,C.F. (1992) A Cladistic analysis of Phenotypic Associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. 3. cladogram estimation. *Genetics*, **132**, 619–633.

Templeton,A.R., Routman,E. and Phillips,C.A. (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,A.R. (2004) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Mol. Ecol.*, **13**, 789–809.

Zhang,A.-B., Tan,S. and Sota,T. (2006) Autoinfer 1.0: a computer program to infer biogeographical events automatically. *Mol. Ecol. Notes*, **6**, 597–599.