WILEY-BLACKWELL

Cladistics 26 (2010) 114–116

Cladistics

10.1111/j.1096-0031.2009.00280.x

Letter to the Editor

# Program note: Cladescan, a program for automated phylogenetic sensitivity analysis

Accepted 16 June 2009

**Abstract**

Examination of trees for the presence of particular nodes is a fundamental aspect of systematics, and is the basis of phylogenetic sensitivity analysis, but becomes unwieldy when performed manually for complex nodes or over large numbers of trees. The program Cladescan is presented here as a stand-alone application to facilitate the detection of nodes in such situations. Cladescan includes features useful for phylogenetic sensitivity analysis, such as automatic generation of ''Navajo rug'' sensitivity plots. In addition, researchers may find it useful for general comparisons among large data sets.



Sensitivity analysis measures how variation in output can be apportioned to variation in input (Saltelli, 2000). In phylogenetics, sensitivity analysis has been used to refer to how relationship hypotheses (phylogenetic trees) vary in response to different tree-search methods [e.g. maximum likelihood (ML), Bayesian likelihood, and maximum parsimony (MP)], optimality parameters (e.g. cost matrices in maximum parsimony or models of evolution in maximum likelihood), or character weighting (Wheeler, 1995; Giribet, 2003). As originally proposed by Wheeler (1995), parameter sensitivity analysis was combined with measures of character and taxon congruence among hypotheses as a method for choosing optimal cost parameters for parsimony analysis. However, sensitivity analysis may also be used independently of hypothesis selection to illustrate the stability of a hypothesis to changes in the underlying assumptions, a process implicit in the common practice of displaying both ML and MP bootstrap values on nodes of published trees.

More formally, this leads to the notion of ''nodal stability'', a measure of robustness to input assumptions which can be seen as complementary to measures of nodal support, such as bootstrap values and Bremer support (Giribet, 2003). Support and stability measures are frequently correlated, but when divergent may help to identify nodes of particular interest to the investigator. Although this approach has drawn criticism on epistemological grounds (Grant and Kluge, 2003, 2005), researchers may continue to find the technique useful (see D'Haese, 2002; Giribet and Edgecombe, 2005).

As currently implemented, however, the process of determining nodal stability requires unwieldy manual examination of multiple trees. Although this may be trivial for small numbers of taxa and trees, as data sets increase in size and the number of trees grows, manual examination becomes time consuming and prone to human error.

The program ''Cladescan'' was written to facilitate such comparisons. Cladescan takes as input one or more sets of trees and a configuration file identifying a node or nodes of interest, records whether each node of interest occurs in each tree, and outputs a summary of the results. This functionality dramatically increases the speed and accuracy of nodal stability estimates compared with manual examination, and may be used as a general tool for quickly comparing topologies among trees.

## Program description

Cladescan is written in Perl, and should run on any Unix-based computer with Perl 5.8.x or higher. The program, along with detailed configuration instructions

Corresponding author:
*E-mail address:* jsanders@oeb.harvard.edu

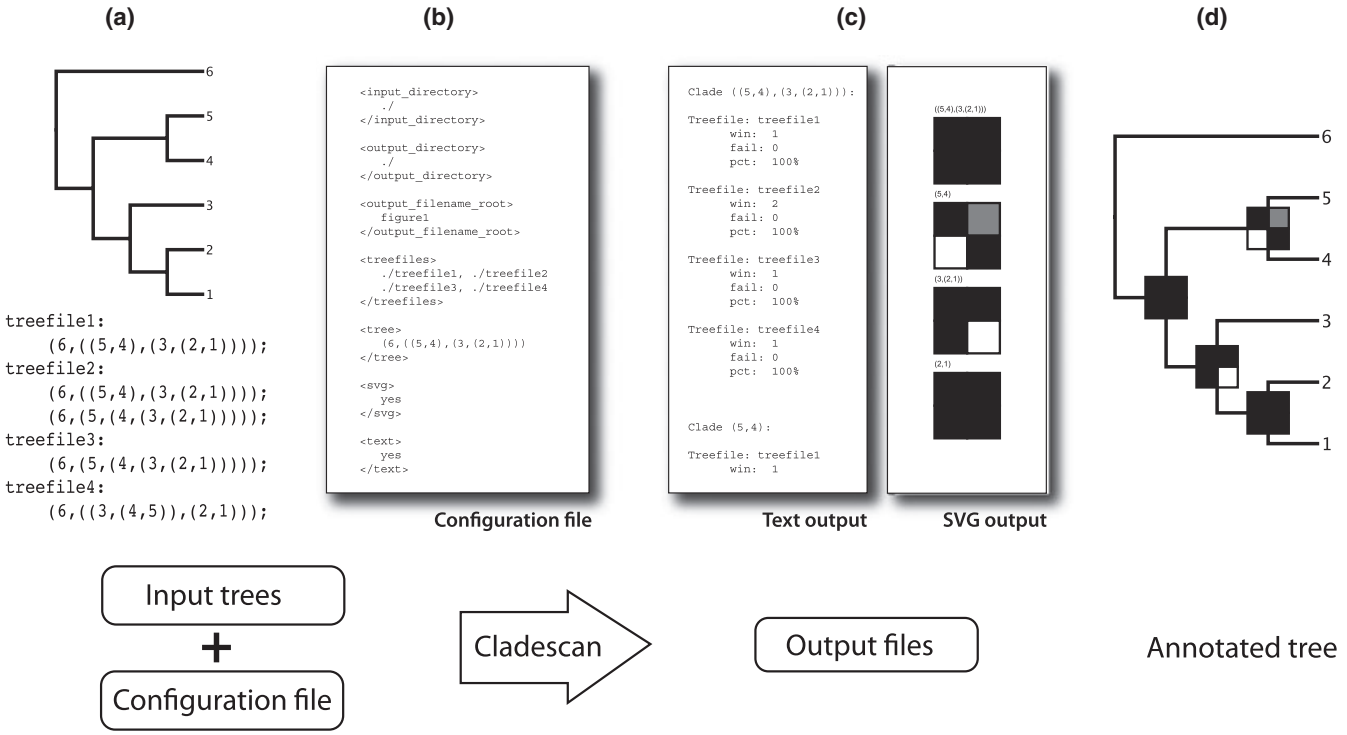


Fig. 1. Example implementation of Cladescan on a simplified data set. (a) Maximum-parsimony analysis of a six-taxon data set under four different character weightings has resulted in four most-parsimonious tree files, one of which contains two equally parsimonious trees. The investigator wishes to determine whether the nodes from the illustrated tree were recovered in each of the four parsimony condition sets. (b) A configuration file is written, directing the program to the location of the input tree files, specifying which nodes to search for, and indicating the desired output formats. (c) Cladescan searches the input tree files for the specified nodes, then outputs results both as a text file containing detailed information for each node and tree file, and as a graphical representation in SVG format. (d) SVG illustrations can then be manually placed on the tree to illustrate nodal stability across cost matrices.

and sample input files, is available for free download under the GNU General Public Licence at http://rc.fas.harvard.edu/cladescan.

Input to Cladescan consists of one or more tree files and one configuration file. Tree files may contain multiple rooted or unrooted trees, which must be in parenthetical format and delimited by semicolons (Fig. 1). Each tree file is taken to represent one condition in the analysis (e.g. the most parsimonious trees from one particular set of cost parameters); thus, a separate tree file is required for each condition. Output options are specified in the configuration file. The user may indicate a number of specific nodes to annotate, in which case the program will scan for any monophyletic grouping of those terminals; alternatively, one may supply a parenthetical tree, in which case the program will scan for each node of the given tree in turn.

As output, Cladescan produces a text file with the results for each target node and tree file, indicating the number of trees in each file containing each target node, the number not containing the node, and the percentage of trees in that tree file containing the node. Optionally, the program can graphically illustrate these results as "Navajo rug" sensitivity plots (sensu Giribet, 2003) in scalable vector graphics (SVG) format, suitable for import into vector-based graphics programs such as Adobe Illustrator (Fig. 1).

## Acknowledgements

I wish to thank Ronald Clouse and Gonzalo Giribet for discussion and helpful critiques of this manuscript. This work was supported by a National Science Foundation Graduate Research Fellowship.

## References

D'Haese, C.A., 2002. Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment. Proc. R. Soc. Lond. B Biol. Sci. 269, 1143–1151.

Giribet, G., 2003. Stability in phylogenetic formulations and its relationship to nodal support. Syst. Biol. 52, 554–564.

Giribet, G., Edgecombe, G.D., 2005. Conflict between datasets and phylogeny of centipedes: an analysis based on seven genes and morphology. Proc. R. Soc. Lond. B Biol. Sci. 273, 531–538.

Grant, T., Kluge, A.G., 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics 19, 379–418.

Grant, T., Kluge, A.G., 2005. Stability, sensitivity, science and heurism. Cladistics 21, 597–604.
Saltelli, A. 2000. What is sensitivity analysis? In: Saltelli, A., Chan, K., Scott, E.M. (Eds.), Sensitivity Analysis. John Wiley & Sons, Chichester, pp. 3–13.
Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.

Jon G. Sanders
*Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, 31 Oxford St, Cambridge, MA 02138, USA*