BIOINFORMATICS APPLICATIONS NOTE

Vol. 21 no. 3 2005, pages 390–392
doi:10.1093/bioinformatics/bti020

Clann: investigating phylogenetic information through supertree analyses

C. J. Creevey* and J. O. McInerney

Bioinformatics and Pharmacogenomics Laboratory, Department of Biology,
National University of Ireland, Maynooth, Co. Kildare, Ireland

Received on March 11, 2004; revised on September 3, 2004; accepted on September 6, 2004
Advance Access publication September 16, 2004

ABSTRACT

Summary: Clann has been developed in order to provide methods of investigating phylogenetic information through the application of supertrees.

Availability: Clann has been precompiled for Linux, Apple Macintosh and Windows operating systems and is available from http://bioinf.may.ie/software/clann. Source code is available on request from the authors.

Supplementary information: Clann has been written in the C programming language. Source code is available on request.

Contact: chris.creevey@may.ie

The aim of constructing phylogenetic supertrees is to combine the information contained in source trees with partially overlapping leaf-sets. Supertree methods can combine the information from trees with no taxa in common as long as additional trees that overlap both exist. Increasingly many methods for supertree construction exist (see Bininda-Emonds *et al*., 2002 for a review), and there is a need for a tool that permits the exploration of the congruence across the input data and the quality of the hypotheses that are derived from the data. In this manuscript we report one such software product. Some desirable properties of supertrees have been described elsewhere (Wilkinson *et al*., 2004), however, no method is guaranteed to have all these properties. As a result we need to explore the data and trees using a variety of methods, each with different properties. This amounts to a sensitivity analysis to examine which hypotheses of relationships are most frequently supported by the different methods and therefore more likely to be the correct relationships.

At present there are four supertree methods implemented in Clann: Matrix Representation using Parsimony (MRP); Most Similar Supertree (MSSA) (Creevey *et al*., 2004); Maximum Quartet Fit (QFIT) and Maximum Splits Fit (SFIT). With MRP, the Baum and Ragan coding scheme, which is additive and binary, is used to create a matrix from the set of source trees (Baum, 1992; Ragan, 1992). This matrix consists of rows representing each taxon and columns representing each internal branch from each source tree. Each internal branch of a source tree divides the taxa into two groups (those descended from the branch versus those ancestral to it). Scoring the taxa with either ‘1’ or ‘0’ according to the group in which they are found represent the hypotheses of relationships defined by each internal branch. If a taxon is not present in a source tree, it is scored with a ‘?’. Parsimony analyses are then used to reconstruct the supertree from this data. The parsimony step must be carried out by ‘PAUP*’ (Swofford, 2002) as Clann writes a nexus formatted file containing the MRP coding scheme and commands for PAUP* to carry out the analysis.

The MSSA scoring method compares each source tree separately to the supertree by comparing the path length distance matrix (Steel and Penny, 1993) derived from a source tree to another distance matrix derived from a pruned supertree. The differences between the matrices are scored and the sum of the scores from all the comparisons is calculated. The user can choose to impose several weighting schemes on this score to adjust for the influence of differential tree size. The weighted or un-weighted sum is the score assigned to the supertree. This sum is used as an optimality criterion to determine the supertree that best fits the set of source trees. This method is related to the average consensus method (Lapointe and Cucumel, 1997) with branch lengths set to unity and as such is also related to MRP (Lapointe *et al*., 2003).

With both the QFIT and SFIT method, each source tree is individually compared to a proposed supertree by determining all the quartets (relationships between any four taxa) (QFIT) or splits (components) (SFIT), respectively for both the source tree and appropriately pruned supertree. A score is then calculated which is defined by the number of quartets or splits that are shared between the supertree and the set of source trees. The sum of the scores calculated for all the source trees is used as an optimality criterion to determine the optimal supertree (the supertree that shares the most quartets or splits with the set of source trees).

For each of the optimality criteria, several different methods of searching tree-space and analysing the underlying

*To whom correspondence should be addressed.

phylogenetic information are implemented in Clann. These methods include complete exhaustive searches of tree-space, heuristic methods of searching tree-space (though not for MRP), methods of bootstrapping the trees to examine the underlying support for any hypothesis and methods for determining whether any phylogenetic signal present in the data is better than would be expected from random data.

Two heuristic algorithms for searching supertree-space are implemented in Clann. They are nearest neighbour interchange (NNI) and sub-tree pruning and re-grafting (SPR) as described and implemented in PAUP* (Swofford, 2002).

Bootstrapping is a statistical technique for empirically estimating the variability in an estimate. It assumes that the samples are independent and identically distributed (Efron, 1979). In a phylogenetic context, bootstrapping allows the estimation of support for a phylogeny. This can be extended to the supertree context as implemented in Clann, by considering the source trees as one possible set of trees that could have been used in the analysis. Choosing a slightly different set of source trees may result in a different optimal supertree. In order to estimate the likely nature of the universe of optimal supertrees, the source trees may be bootstrapped. For each bootstrap replicate, the source trees are sampled with replacement until a new dataset is created with the same number of source trees as the original dataset. This means that some source trees may be represented in the dataset more than once, while others may not be represented at all. For each repetition, the supertree that best represents this (bootstrapped) set of source trees, according to the chosen optimality criterion, is determined. Repeating this procedure a large number of times gives an indication as to how much support there is for the clades in a supertree (Purvis, 1995). If during any bootstrap replicate taxon is not represented (due to the initial low occurrence of the taxon), the software will alert the user to the unsuitability of the data to bootstrapping and refuse to continue.

A randomization method to test the null hypothesis that the phylogenetic signal in the source trees is no better than random is also implemented in Clann. This test has been implemented for all the supertree methods except MRP where a normal Permutation Tail Probability (PTP) (Archie, 1989; Faith and Cranston, 1991) test is available. We have called this method the YAPTP (Yet Another Permutation Tail Probability) test (Creevey *et al.*, 2004). For each repetition of the test, each source tree is replaced with a randomly chosen topology for the same leaf-set. This removes any congruent phylogenetic signal between source trees, while leaving the numbers and sizes of source trees, the frequency with which any particular taxon was found across the source trees and the frequency of cooccurrence of any group of taxa within source trees unaltered. A search of tree space can then be carried out and the score of the best supertree recorded. The user can repeat this test as many times as required and the distribution of the resulting scores can be compared to the score of the real data (or the distribution of scores from bootstrapping) to assess if the real data contains a signal that is better than random. Permutation tests of this kind are extremely forgiving in nature. Passing them may however be considered a minimal requirement for any dataset to be considered for further analysis.

While both bootstrapping and the YAPTP test provide means of assessing the results of the supertree analysis, it must be pointed out that such assessments must be regarded within the context of what the supertree analysis was trying to achieve and the methods used to achieve them. For instance, was the goal of the analysis to reconstruct a phylogeny, test for tree-likeness of the data, to assess the support for particular clades or to reconstruct a historical timeline? Then, how do the methods chosen to carry out these analyses affect the interpretation of the results? Clann provides a necessary tool to help achieve these and other goals in a supertree context.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Mark Wilkinson for reading the manuscript, the four anonymous referees for their helpful comments and for the many users who contributed suggestions and reported bugs in earlier versions of the software.

REFERENCES

Archie,J.W. (1989) A randomisation test for phylogenetic information in systematic data. *Syst. Zoo.*, **38**, 239–252.

Baum,B.R. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, **41**, 3–10.

Bininda-Emonds,O.R.P., Gittleman,J.L. and Steel,M. (2002) The (super)tree of life: procedures, problems and prospects. *Ann. Rev. Ecol. Syst.*, **33**, 265–289.

Creevey,C.J., Fitzpatrick,D.A., Philip,G.K., Kinsella,R.J., O'Connell,M.J., Pentony,M.M., Travers,S.A., Wilkinson,M. and McInerney,J.O. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc. R. Soc. Lond. B. Biol. Sci.*, (in press).

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

Faith,D.P. and Cranston,P.S. (1991) Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics*, **7**, 1–28.

Lapointe,F.-J. and Cucumel,G. (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst. Biol.*, **46** 306–312.

Lapointe,F.J., Wilkinson,M. and Bryant,D. (2003) Matrix representations with parsimony or with distances: two sides of the same coin? *Syst. Biol.*, **52**, 865–868.

Purvis,A. (1995) A composite estimate of primate phylgoeny. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **348**, 405–421.

Ragan,M.A. (1992) Matrix representation in reconstructing phylogenetic-relationships among the eukaryotes. *Biosystems*, **28**, 47–55.

Steel,M. and Penny,D. (1993) Distributions of tree comparison metrics—some new results. *Syst. Biol.*, **42**, 126–141.

Swofford,D.L. (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*And Other Methods). Version 4.* Sinauer Associates, Sunderland, MA.

Wilkinson,M., Thorley,J.L., Pisani,D., Lapointe,F.-J. and McInerney,J. (2004) Some desiderata for liberal supertrees. In Bininda-Emonds,O.R.P. (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic, Dordrecht.