BIOINFORMATICS APPLICATIONS NOTE
Vol. 21 no. 8 2005, pages 1709–1710
doi:10.1093/bioinformatics/bti175

Phylogenetics

SYMMETREE: whole-tree analysis of differential diversification rates

Kai M. A. Chan[1] and Brian R. Moore[2,*]

[1]Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA and
[2]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA

Received on October 4, 2004; revised on November 17, 2004; accepted on November 22, 2004
Advance Access publication November 30, 2004

ABSTRACT

Summary: SYMMETREE implements several tests of differential diversification rates that exploit information on the topological distribution of species diversity throughout entire trees to address two general questions: (1) Has a given tree experienced significant variation in diversification rates among its branches? and (2) If so, along which branches have significant shifts in diversification rate occurred? These explicitly model-based methods are robust to uncertainty in estimates of branch length/duration and can accommodate incompletely resolved trees and other forms of phylogenetic uncertainty.

Availability: http://www.phylodiversity.net/bmoore/software.html

Contact: brian.moore@yale.edu

Phylogenies can potentially provide two sources of information relevant to the study of diversification rates: the topological distribution of species diversity across branches, and the temporal distribution of branching events through time (Sanderson and Donoghue, 1996). Two corresponding classes of methods have been developed to exploit these different sources of information: *topological* methods compare the observed difference in species diversity between two (or more) groups descended from a common node to a distribution of diversity differences generated under a stochastic model of diversification, whereas *temporal* methods compare the distribution of branching times (based on branch length/duration estimates) to that generated under a null model of random diversification.

Because they directly incorporate information on the timing of diversification, temporal methods are generally held to be more powerful than their topological counterparts. This perceived power advantage has, in turn, motivated the elaboration of temporal methods to address a relatively wide range of evolutionary questions related to diversification rates. Nevertheless, many inference problems will demand a topological approach. Some data types and/or inference procedures yield trees that are unsuitable for analysis by temporal methods. For example, phylogenies estimated from morphological data, estimated under parsimony, or derived from a super-tree approach typically lack reliable branch-length estimates. Furthermore, even when reliable branch-length estimates are available, it will occasionally be necessary to exclude these data from studies of diversification rates. For example, many evolutionary studies entail hypothesized associations between diversification rates and some other variable that is directly or indirectly dependent on branch lengths/durations. Attempts to understand the correlation of such variables to variation in rates of diversification will be confounded if both are conditioned on the same set of branch-length estimates.

Accordingly, the nature of the data at hand and/or the hypotheses of interest will often preclude the inference of diversification rates based on temporal data. These considerations motivated the development of topology-based methods that realize substantially increased power by virtue of incorporating information on the topological distribution of species diversity from the whole tree (Chan and Moore, 2002; Moore *et al*., 2004). These whole-tree methods, implemented in SYMMETREE, can be used to address two aspects of differential diversification rates: to detect significant among-lineage diversification rate variation and to locate significant diversification rate shifts.

These methods compare the observed topological distribution of species diversity to expectations under the equal-rates Markov (ERM) random-branching model—a continuous-time, discrete-state, pure-birth Markov process in which the probability of a branching event is constant for each tip in a growing tree at any instant in time (Yule, 1924). Under the ERM model, the cumulative probability of realizing a diversity partition as or more extreme than an observed partition of N species between two sister groups with l and r species is

$$P = \frac{2l}{(N-1)} \quad (1)$$

(unless $l = N/2$, in which case $P = 1$), where l is the number of species in the less diverse sister group (Slowinski and Guyer, 1989). Because their calculation incorporates minimal information on the topological distribution of species diversity, the sensitivity of individual ERM nodal probabilities is quite low. The premise of the whole-tree methods is to generalize individual ERM nodal probabilities to obtain methods that harness their collective statistical power.

SYMMETREE implements several whole-tree tests of diversification rate variation that variously incorporate information on the relative diversity of all internal nodes of a given tree: M_Π and M_Σ are based on the cumulative ERM probability derived from the product or sum of all individual nodal probabilities, respectively (Chan and Moore, 2002). M_Π^* and M_Σ^* are variants that differentially weight individual ERM nodal probabilities by their size (Moore *et al*., 2004). Other whole-tree tests implemented in the program include: M_R (Page, 1993); I_C (Heard, 1992) and B_1 (Shao and Sokal, 1990). The

*To whom correspondence should be addressed.

cumulative probability for each statistic is derived by Monte Carlo simulation: the value for a given statistic is calculated for the study tree and compared to its respective null distribution, which is approximated by repeatedly simulating trees of equal size under the ERM model and calculating values for the statistic for each simulated tree.

SYMMETREE also implements two whole-tree methods for locating significant shifts in diversification rate. The objective is to assess the probability of a shift along the lone internal branch of a three-taxon tree comprised of a local outgroup clade and the two basal-most subclades of the ingroup clade. The probability of a diversification rate shift along the internal branch is returned by a shift statistic, which is calculated as a function of two likelihood ratios. One likelihood ratio is calculated at the root of the three-taxon tree (involving the diversity partition between the outgroup and ingroup clades), the other at the root of the ingroup clade (involving the diversity partition between the left and right ingroup clades). Each likelihood ratio compares the likelihood of realizing the observed diversity partition under a homogeneous (one-rate parameter) model, in which both groups have the same branching rate, to that under a heterogeneous (two-rate parameter) model, in which the two groups have different branching rates.

In order to address the so-called 'trickle-down' problem—the tendency for a diversification rate shift along a given branch to create illusory shifts at more inclusive nodes in the tree (Moore *et al.*, 2004)—the shift statistics condition the likelihood of a diversification rate shift along the internal branch of the three-taxon tree by the likelihood of a shift within the ingroup clade. The two likelihood ratio-based shift statistics, Δ_1 and Δ_2, correspond to two different ways of conditioning the inclusive likelihood ratio by the nested likelihood ratio. The three-taxon evaluations are iterated over all internal branches to effectively survey the whole tree for diversification rate shifts.

Methods for detecting differential diversification rates typically require strictly dichotomous trees; the prevalence of polytomies has therefore proven to be a serious impediment to their application. Accordingly, an important feature of SYMMETREE is its ability to accommodate incompletely resolved trees. The program implements a number of random resolution algorithms and batch processing features that can estimate confidence intervals associated with polytomies and other forms of phylogenetic uncertainty. SYMMETREE is a freely available command-line program written in C++ and compiled for Macintosh OSX, Windows and Unix operating systems. Input files are read from the standard NEXUS format (Maddison *et al.*, 1997), and output is written to a tab-delimited text file. A manual and example data sets are included with the program distribution.

REFERENCES

Chan,K.M.A. and Moore,B.R. (2002) Whole-tree methods for detecting differential diversification rates. *Syst. Biol.*, **51**, 855–865.

Heard,S.B. (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, **46**, 1818–1826.

Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Moore,B.R., Chan,K.M.A. and Donoghue,M.J. (2004) Detecting diversification rate variation in supertrees. In Bininda-Emonds O.R.P. ed. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic, Dordrecht, the Netherlands, pp. 487–533.

Page,R.D.M. (1993) On describing the shape of rooted and unrooted trees. *Cladistics*, **9**, 93–99.

Sanderson,M.J. and Donoghue,M.J. (1996) Reconstructing shifts in diversification rates on phylogenetic trees. *Trends Ecol. Evol.*, **11**, 15–20.

Shao,K.-T. and Sokal,R.R. (1990) Tree balance. *Syst. Zool.*, **39**, 266–276.

Slowinski,J.B. and Guyer,C. (1989) Testing the stochasticity of patterns of organismal diversity: an improved null model. *Am. Nat.*, **134**, 907–921.

Yule,G.U. (1924) A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. *Phil. Trans. R. Soc. L. B*, **213**, 21–87.