



Short Communication

PRAP—computation of Bremer support for large data sets

Kai Müller*

Nees-Institute for the Biodiversity of Plants, University of Bonn, Meckenheimer Allee 170, 53115 Bonn, Germany

Received 6 June 2003; revised 5 November 2003

1. Introduction

The Bremer support (BS, Bremer, 1988, 1994; Källersjö et al., 1992), also known as “decay index” (Olmstead et al., 1993), “length difference” (Faith, 1991), or “support index” (SI, Davis, 1993; Kluge and Farris, 1969), is a measure of branch support frequently seen in the recent phylogenetic literature. The strengths and limits of this index have lately been examined (DeBry, 2001; Oxelman et al., 1999) and it was generally concluded that care must be taken when interpreting the values. In particular, the calculation of BS becomes problematic for intermediate to large data sets. Either, the resulting support values will be sometimes drastic overestimates of support (Bremer, 1994) or, if more thorough search strategies are invoked to assess support of all the individual branches, the process will become a very time consuming task (Oxelman et al., 1999). Various procedures have been suggested for computationally more demanding data sets (Davis, 1995; Morgan, 1997), and the reverse constraint method has emerged as most effective (Morgan, 1997). The power of all approaches, however, strongly depends on the ability of the applied heuristic search strategy to find shortest trees. With the common search strategies and the currently available processor speed, this ability shrinks rapidly as matrix sizes grow to more than 100–150 taxa (although strongly depending on the data set). While finding the most parsimonious (MP) tree for a large data set is time consuming itself, heuristic searches have to be repeated N times if N is the number of internal branches to be tested ($N \leq \text{number of terminals} - 2$). Thus, a compromise has to be found between elaborateness of the search per branch and the time invested. If the search parameters used are too lax, the BS values will probably

be an overestimation, since the shortest trees under the constraint will not be found, and, thus, the step difference to the actual shortest tree will be too high. For these reasons, BS has rarely been considered for sizeable data sets, and never for large scale analyses such as on angiosperm phylogenetics (e.g., Soltis et al., 2000).

To speed up the search for the shortest tree for large data sets, a variety of methods have been described recently (e.g., Goloboff, 1999; see also Moret et al., 2002). The parsimony ratchet (Nixon, 1999) was among the first to be developed, and its impressive efficiency compared with previous search schemes has been reported a number of times (e.g., Goloboff, 1999; Hilu et al., 2003; Soltis et al., 2000). These methods, however, have been implemented in software available for WINDOWS/DOS only, for example TNT (Goloboff et al., 1999), and to the author’s knowledge, these programs do not support use of the respective algorithms during the calculation of BS. The most commonly used phylogenetic software, PAUP* (Swofford, 1998), is available for virtually all computer platforms, and is not restricted to purely cladistic methods, instead featuring maximum likelihood methods, and numerous statistical tests. PAUP*’s data and tree file format (NEXUS; Maddison et al., 1997) is understood or shared by a variety of other programs (Huelsenbeck and Ronquist, 2001; Maddison and Maddison, 1992; Müller and Müller, 2003a). However, as of this writing, neither the calculation of BS nor the parsimony ratchet or other fast algorithms are available in PAUP*. While for the older Mac OS systems (up to OS9), software exists that simplifies application of the reverse constraint method for small to intermediate data sets (e.g., AutoDecay, Eriksson, 1998), no program appears to be at hand that attempts to implement fast algorithms into the search strategies during heuristic searches under constraints. In a study employing branch support analysis with help of AutoDecay on a 82-taxa data set (Oxelman et al., 1999), several search strategies as supported by PAUP* were

* Fax: +49-228-73-3120.

E-mail address: kaimueller@uni-bonn.de.

applied to each constrained node, including TBR branch swapping starting from the unconstrained topology and saving only few trees, and running a low number of random addition cycles with TBR on only one tree. It became apparent that more rigorous settings yield lower BS values in a significant proportion of branches; however, none of the today known much faster algorithms could be tested in this study. For these reasons, an application appeared to be useful that enables users to perform calculation of BS using fast search strategies on their preferred platform and in combination with the most widely used phylogenetic software.

PRAP (Parsimony Ratchet Analyses using PAUP*) is a new application with a user-friendly graphical interface written in JAVA, running on each system with a Java Virtual Machine (JVM) installed. The JVM can easily be downloaded for practically all platforms. PRAP allows the calculation of BS (Decay values) using the parsimony ratchet algorithm. Alternatively, BS can also be computed without using the ratchet. The support values are written to tree files that can have the widely used NEXUS format or, to facilitate publication of the tree, the new TGF format interpreted by the program TreeGraph. (Müller and Müller, 2003b; uses TGF files to generate more complex postscript trees that can include bootstrap percentages, posterior probabilities, branch lengths, clade annotations, and graphical elements and are easy to edit, print, or embed in a manuscript.) PRAP reads a NEXUS tree file supplied by the user. Then it creates constraint trees for each branch and writes commands for PAUP* into a command file. Instead of simple heuristic search statements, a series of commands is written for each branch, corresponding to the parsimony ratchet procedure. The user can supply parameters influencing efficiency and search time of the ratchet. When executing this command file, PAUP* calculates MP trees using the reverse constraint option. A log file is created and parsed by PRAP to determine how many more steps a tree requires that constrains a

given branch. This value, corresponding to the BS, is assigned to the branch. This is done for each branch in the tree. The whole tree and its BS values are finally written to a file: (i) as a NEXUS tree with the BS values written as taxon labels that can be viewed and printed with the program TreeView (Page, 1996); and/or (ii) as a TGF file (see above).

Beyond calculating BS, PRAP can also be used to run parsimony ratchet searches in order to find the shortest (unconstrained) trees for a given data set. In addition to simple ratchet searches, random addition cycles of the ratchet are possible, running several series of ratchet cycles from different starting trees. Even for simple non-ratchet searches the application can be used as graphical interface for PAUP* under Windows and UNIX-based systems for which no menu driven PAUP* versions exist.

Three data sets were used to assess the effect of using the ratchet during calculation of BS (details available on request): (i) Lamiales (angiosperms), 89 taxa, plastid *trnK* intron sequences, data set from Müller et al. (in press); (ii) pleurocarpous mosses, 86 taxa, plastid *trnL-F* and *rps4* sequences, data set from Buck et al. (2000); and (iii) angiosperms, 385 taxa, plastid *matK* sequences, data set from Hilu et al. (2003). For all datasets, the tree to be evaluated was a strict consensus tree of all shortest trees found during 10 random addition cycles of 200 ratchet iterations each (i + iii: as published; ii: recalculated here). The settings chosen in the non-ratchet searches (simple addition and saving of up to 100 trees) were designed to require computation times roughly comparable to those needed with the ratchet (50 iterations, each assigning double weight to 25% of the characters). For data set (i) the branch support was overestimated in 81% of the branches. For these 81%, an analysis applying the ratchet found BS at least 1 step lower, averaging 0.96 steps lower over all branches (including those that showed no difference). In data set (ii) 71% of branches obtained a lower BS value after using the

Table 1

Performance divergence between (i) the random addition search strategy (RAS) of Oxelman et al. (1999) for the calculation of Bremer support (BS) in large trees; and (ii) the parsimony ratchet approach (PR), observed for the first 15 min of calculating BS for 10 randomly selected nodes of tree 'B' and the data set in Hilu et al. (2003)^a

Node	Time for 25 RAS cycles	PR found same BS after	Difference of BS after 10 min	Difference of BS index after 15 min
15	0:09:52	0:01:23	5	5
46	0:09:30	0:01:44	2	5
133	0:10:32	0:05:01	1	3
148	0:10:48	0:00:47	5	5
153	0:09:50	0:02:12	4	4
175	0:09:58	0:01:01	5	7
241	0:09:58	0:00:48	5	5
255	0:11:07	0:00:21	5	5
295	0:10:14	0:01:03	3	3
305	0:09:25	0:00:43	4	4

^a Calculated with PAUP* 4.0b10 using a 850 MHz Pentium III and Windows XP.

ratchet; the difference was 1.34 steps on average. Since both data sets are rather moderate in size, the advantage of using the ratchet here is not particularly significant. More thorough search schemes without the ratchet, applying random addition cycles and saving more trees, would certainly decrease the reported differences in BS, but only at the cost of computation times far exceeding those needed with the parsimony ratchet.

For the considerably larger data set (iii) the performance of both approaches was studied in more detail (Table 1). Here, the random addition search strategy (RAS) of Oxelman et al. (1999; 25 cycles, no multrees) was compared to the parsimony ratchet during the first 15 min of search time for each of 10 randomly selected nodes. The lowest BS indices RAS found during its 25 cycles (~10 min) on average were encountered within the first 2–3 iterations (<1.5 min) of the ratchet. More importantly, the ratchet found BS indices 4–5 steps lower within the first 15 cycles (~10 min) and continued to find still lower ones afterwards (Table 1). The performance difference is expected to increase with the number of taxa, the number of replications and saved trees during RAS, and, up to a data-dependent saturation point, the number of ratchet iterations used (see also Nixon, 1999).

PRAP can be downloaded from <http://www.botanik.uni-bonn.de/system/downloads/PRAP>, together with documentation and sample data files, or can directly be requested from kaimueller@uni-bonn.de.

Acknowledgments

The program was designed to facilitate the reconstruction and statistical evaluation of large phylogenetic trees compiled in the project “Systematics of Amaranthaceae and evolution of pollen characters” funded by the Deutsche Forschungsgemeinschaft (DFG), Grant No. Bo 1815/1-1 to T. Borsch. The author would like to thank T. Borsch, W. Barthlott, K.W. Hilu, D. Quandt, C. Löhne, A. Worberg, and two anonymous reviewers, for helpful comments and testing of the program.

References

- Bremer, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstructions. *Evolution* 42, 795–803.
- Bremer, K., 1994. Branch support and tree stability. *Cladistics* 10, 295–304.
- Buck, W.R., Goffinet, B., Shaw, A.J., 2000. Testing morphological concepts of orders of pleurocarpous mosses (Bryophyta) using phylogenetic reconstructions based on trnL-trnF and rps4 sequences. *Mol. Phylogenet. Evol.* 16, 180–198.
- Davis, J.I., 1993. Character removal as a means for assessing stability of clades. *Cladistics* 9, 201–210.
- Davis, J.I., 1995. A phylogenetic structure for the monocotyledons, ascertained from chloroplast DNA restriction site variation, and a comparison of measures of clade support. *Syst. Bot.* 20, 503–527.
- DeBry, R.W., 2001. Improving interpretation of the decay index for DNA sequence data. *Syst. Biol.* 50, 742–752.
- Eriksson, T., 1998. AutoDecay, Program distributed by the author. Available from <http://www.bergianska.se/personal/TorstenE/>.
- Faith, D.P., 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* 40, 366–375.
- Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* 15, 415–428.
- Goloboff, P.A., Farris, J.S., Nixon, K.C., 1999. TNT: Tree Analysis Using New Technology, Available from www.cladistics.com.
- Hilu, K.W., Borsch, T., Müller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M., Powell, M., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A., Rohwer, J.G., Campbell, C.S., Chatrou, L., 2003. Angiosperm phylogeny based on matK sequence information. *Am. J. Bot.* 90, 1758–1776.
- Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Källersjö, M., Farris, J.S., Kluge, A.G., Bull, C., 1992. Skewness and permutation. *Cladistics* 8, 275–287.
- Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18, 1–32.
- Maddison, D.R., Swofford, D., Maddison, W.P., 1997. NEXUS: an extensible file format for systematic information. *Syst. Biol.* 46, 590–621.
- Maddison, W.P., Maddison, D.R., 1992. MacClade. Sinauer Associates, Sunderland.
- Moret, B.M.E., Bader, D.A., Warnow, T., 2002. High-performance algorithm engineering for computational phylogenetics. *J. Supercomp.* 22, 99–110.
- Morgan, D., 1997. Decay analysis of large sets of phylogenetic data. *Taxon* 46, 509–517.
- Müller, J., Müller, K., 2003a. QuickAlign: a new alignment editor. *Plant Mol. Biol. Rep.* 21, 5.
- Müller, J., Müller, K., 2003b. TREEGRAPH: generating complex postscript trees using an extensible tree description format, Program distributed by the authors, Botanical Institute, University of Bonn. Available from <http://www.botanik.uni-bonn.de/system/downloads/TreeGraph>.
- Müller, K., Borsch, T., Legendre, L., Porembski, S., and Barthlott, W., in press. Evolution of carnivory in Lentibulariaceae and the Lamiales. *Plant Biol.*
- Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15, 407–414.
- Olmstead, R.G., Bremer, B., Scott, K.M., Palmer, J.D., 1993. A parsimony analysis of the asteridae s.l. based on rbcL sequences. *Ann. Mo. Bot. Gard.* 80, 700–722.
- Oxelman, B., Backlund, M., Bremer, B., 1999. Relationships of the Buddlejaceae s.l. investigated using parsimony jackknife and branch support analysis of chloroplast ndhF and rbcL sequence data. *Syst. Bot.* 24, 164–182.
- Page, R.D.M., 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12, 357–358.
- Soltis, D.E., Soltis, P.S., Chase, M.W., Mort, M.E., Albach, D.C., Zanis, M., Savolainen, V., Hahn, W.H., Hoot, S.B., Fay, M.F., Axtell, M., Swensen, S.M., Prince, L.M., Kress, W.J., Nixon, K.C., Farris, J.S., 2000. Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Bot. J. Linn. Soc.* 133, 381–461.
- Swofford, D.L., 1998. PAUP*. Phylogenetic Analysis using Parsimony (* and other methods). Sinauer Associates, Sunderland.