

Weighted compromise trees: a method to summarize competing phylogenetic hypotheses

Michael J. Sharkey^{a,*}, Stephanie Stoelb^a, Daniel R. Miranda-Esquivel^b and Barabara J. Sharanowski^c

^aUniversity of Kentucky, Department of Entomology, S225 Agricultural Science Center North, Lexington, KY 40546-0091, USA; ^bLaboratorio de Sistemática y Biogeografía, Escuela de Biología, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia; ^cUniversity of Manitoba, Department of Entomology, Winnipeg, Manitoba, Canada, R3T 2N

Accepted 20 August 2012

Abstract

A new consensus method for summarizing competing phylogenetic hypotheses, weighted compromise, is described. The method corrects for a bias inherent in majority-rule consensus/compromise trees when the source trees exhibit non-independence due to ambiguity in terminal clades. Suggestions are given for its employment in parsimony analyses and tree resampling strategies such as bootstrapping and jackknifing. An R function is described that can be used with the programming language R to produce the consensus.

© The Willi Hennig Society 2012.

Nixon and Carpenter (1996) wrote that the only true consensus method is a strict consensus, whereas alternatives such as majority-rule consensus should rightfully be referred to as compromise trees; we follow their terminology here. These other methods, which include Adams (Adams, 1972), majority-rule (Margush and McMorris, 1981), Nelson (Nelson, 1979; Page, 1989; Nixon and Carpenter, 1996) and Bremer or semi-strict (Bremer, 1990; Swofford, 1990) compromise methods, are generally employed when a strict consensus is too poorly resolved for the purposes of the study and a compromise must be made to effect the desired resolution.

Consensus and compromise trees may also be useful for highlighting agreement or ambiguity among source cladograms, or summarizing phylogenetic information from multiple lines of evidence (Miyamoto, 1985; Carpenter, 1988; Anderberg and Tehler, 1990; Wilkinson, 1994). Strict consensus trees retain only those nodes that are present in all source trees, and thus represent unequivocal statements about phylogenetic relation-

ships (Nixon and Carpenter, 1996). However, almost all resolution of evolutionary relationships can be lost when a “rogue” or “wildcard” taxon is variably placed throughout the source trees (Page, 1989; Bremer, 1990; Nixon and Wheeler, 1991; Wilkinson, 1994; Sumrall et al., 2001; Wilkinson and Thorley, 2001). Adams compromise is particularly useful for identifying rogue taxa, but has been criticized (Page, 1989; Swofford, 1991) for recovering nodes that are not present in any of the source trees (but see Wilkinson, 1994). Majority-rule compromise (MRC) trees retain clades found in a user-defined frequency of source trees (typically 50%) and they are often more resolved than strict consensus trees. They are widely employed in statistically based phylogenetic methods such as Bayesian inference (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) as well as for summarizing trees from resampling techniques, such as bootstrap and jackknife (Efron, 1982; Felsenstein, 1985). However, the frequency of nodes present in the majority of source trees is not always an adequate criterion for inclusion in a compromise tree (Barrett et al., 1991; Sharkey and Leathers, 2001). Comprehensive reviews on consensus methods,

*Corresponding author:
E-mail address: msharkey@uky.edu

particularly for summarizing information from multiple data sets, can be found in Swofford (1991), Nixon and Carpenter (1996), Steel et al. (2000) and Bryant (2003).

A short definition is warranted to assist readers of this article. A component is any group of taxa defined by a node in a tree, regardless of the interrelationships among the taxa included in the component. Thus in Fig. 1 the clades (E(FG)) and (F(EG)) both define the component (EFG).

MRC has been criticized for including components that are only present in the majority of trees due to non-independence of the source trees (Sharkey and Leathers, 2001; Sumrall et al., 2001). To demonstrate the lack of independence in MRC, a contrived dataset is presented in Table 1, which results in three most-parsimonious trees (Fig. 1a, trees 1–3) from a traditional heuristic search in TNT (Goloboff et al., 2008) (50 replications RAS, 2 trees saved per replication, TBR, rseed = 0). The strict consensus and MRC trees are presented in Fig. 1b and c, respectively. An examination of the three source trees (Fig. 1a) reveals changes in the placement of terminal taxa C, D, E, F and G leading to little resolution in the strict consensus (Fig. 1b). In the MRC tree the relationships (BC) and (D(E(FG))) are recovered because they are present in two of the three trees (trees 1 and 2) and thus meet the majority criterion. Tree 1 and tree 2 differ only in the placement of taxon G. The basal split, (A((BC) (DEFG))), is identical in these two trees, but the interrelationships within the more apical component (EFG) are ambiguous. Source trees 1 and 2 are not independent, i.e. their shared basal split results in two solutions for the more derived clade (EFG). The alternative basal split, ((A)(DBCEFG)), found in source

Table 1
Contrived data matrix for the trees presented in Fig. 1

Taxon	111111111122222
	123456789012345678901234
A	0000000000000000000000
B	111001110111110000100000
C	111000010111110000011010
D	111110000001111000000101
E	111111010111111110111111
F	1111111010111011010101110
G	111111000011111101111110

tree 3 gives a unique solution to clade (EFG) and thus this basal topology is recovered only once as a most-parsimonious solution.

The lack of independence between trees 1 and 2 affects the MRC, leading to a higher frequency of the basal split shared by trees 1 and 2 in the source trees. Thus the relationship (A((BC)(D(E(FG)))) is recovered in the MRC (Fig. 1c). This can be argued in evolutionary terms. In this example there are two equally supported basal splits. The first is that (BC) is sister to (DEFG), and in this case the latter clade can be resolved in two maximally parsimonious ways. The second basal split, the alternative hypothesis (Tree 3), is that D is the sister of (BCEFG), and in this case the latter clade has a unique maximally parsimonious solution. We see no justification for equivocal terminal clades, as in trees 1 and 2, adding support to a basal split. Viewed from an historical perspective, there are two equally supported resolutions of the outgroup, and the resolution of the ingroup (BCDEFG) is dependent on each of these.

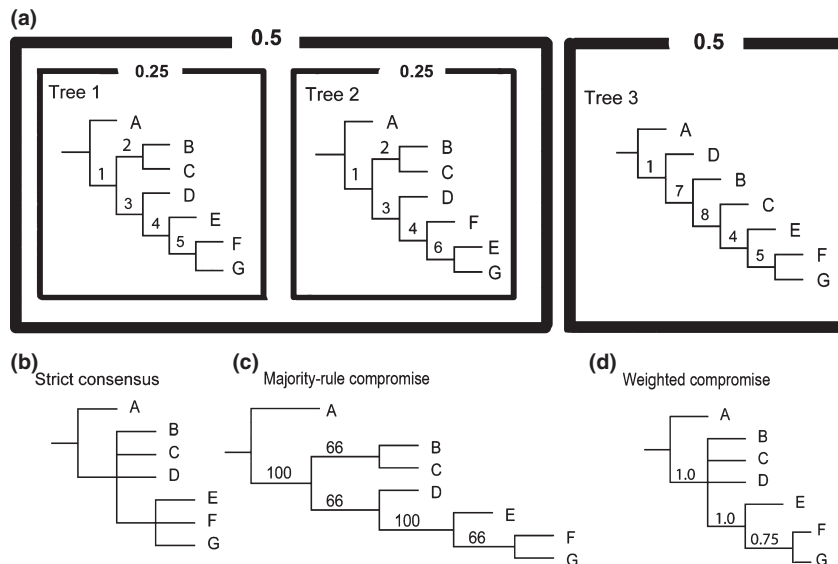


Fig. 1. (a) Three minimum-length trees from the data set of Table 1. Numbers in the boxes surrounding the trees are the weights of the trees. (b) Strict consensus of the three trees in (a). (c) Majority-rule compromise of the three trees in (a). (d) Weighted compromise of the three trees in (a).

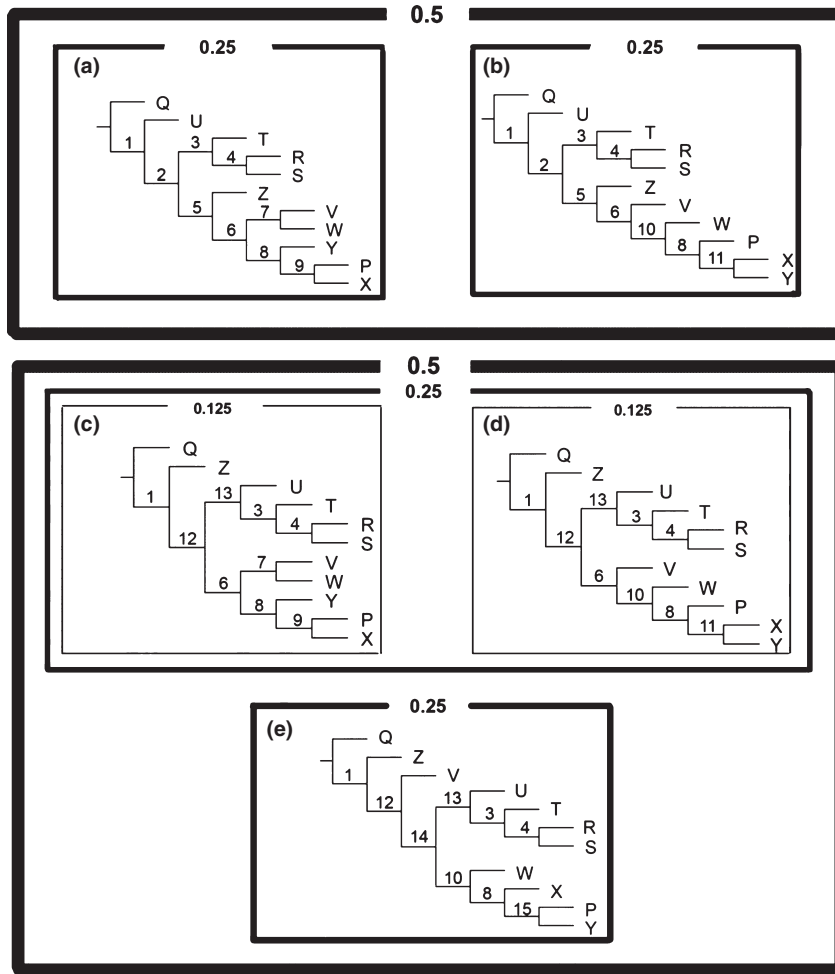


Fig. 2. The five minimum-length trees from the data set of Table 2. Numbers in the boxes surrounding the trees are the weights of the trees.

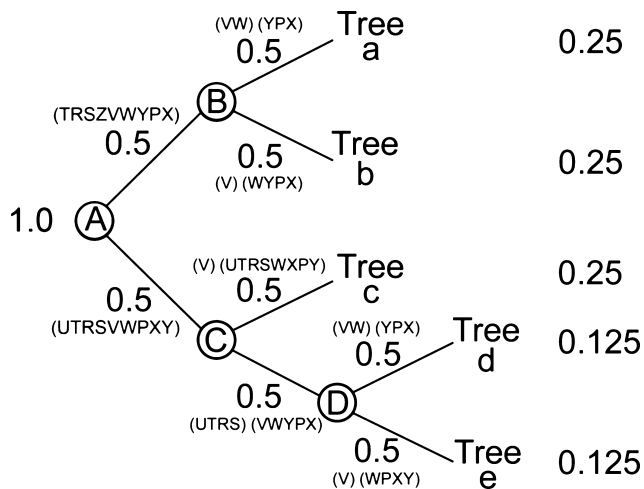


Fig. 3. Probability graph showing derivation of the weights of the five trees in Fig. 2. Nodes are identified by capital letters.

on resolution is set by the majority-rule tree and the lower limit by the strict consensus. The most basal components with conflicts in the source trees will never be resolved in a weighted compromise tree as there will always be two or more ties at this point. Given only two source trees, the strict consensus, the MRC and the weighted compromise will have the same topology.

Another approach to weighted compromise is to include all clades in the weighted compromise tree with the highest weights even if they are not over 0.5. For example, the highest weight for resolution within a clade might only be 0.33, but if this is higher than all other alternatives it may be included in the weighted compromise tree.

Like strict consensus trees and compromise methods, weighted compromise trees may not represent an optimal tree from which one can directly draw all phylogenetic conclusions, but may do so in many cases.

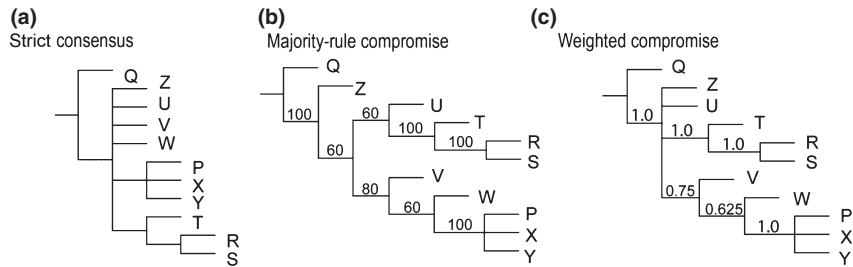


Fig. 4. (a) Strict consensus of the five source trees in Fig. 2. (b) Majority-rule compromise of the five source trees in Fig. 2. (c) Weighted compromise of the five source trees in Fig. 2.

Table 3
Weights for the five trees in Fig. 2 and weights for all components in the five trees

Component	Tree a weight	Tree b weight	Tree c weight	Tree d weight	Tree e weight	Component weight
1 (UTRSZVWYPX)	0.25	0.25	0.125	0.125	0.25	1.0
2 (TRSZVWYPX)	0.25	0.25	N/A	N/A	N/A	0.5
3 (TRS)	0.25	0.25	0.125	0.125	0.25	1.0
4 (RS)	0.25	0.25	0.125	0.125	0.25	1.0
5 (ZVWYPX)	0.25	0.25	N/A	N/A	N/A	0.5
6 (VWYPX)	0.25	0.25	0.125	0.125	N/A	0.75
7 (VW)	0.25	N/A	0.125	N/A	N/A	0.375
8 (YPX)	0.25	0.25	0.125	0.125	0.25	1.0
9 (PX)	0.25	N/A	0.125	N/A	N/A	0.375
10 (WPXY)	N/A	0.25	N/A	0.125	0.25	0.625
11 (XY)	N/A	0.25	N/A	0.125	N/A	0.375
12 (UTRSVWYPX)	N/A	N/A	0.125	0.125	0.25	0.5
13 (UTRS)	N/A	N/A	0.125	0.125	0.25	0.5
14 (UTRSWXPY)	N/A	N/A	N/A	N/A	0.25	0.25
15 (PY)	N/A	N/A	N/A	N/A	0.25	0.25
Total						9.00

N/A, not applicable.

Program

To calculate the weighted compromise, D.R.M.-E. developed the program *wconsensus*, using the programming language R (R Development Core Team, 2011). It can be implemented using the R package APE (Paradis et al., 2004). The function calculates the weighted compromise and presents either the MRC or the weighted compromise with the nodes collapsed according to a cut value (weighted compromise values at the nodes). The default cut value is 0.5, but this can be modified by the user. The input data (the initial trees in Newick format only) are entered as a multiphylo object, and a collapsed weighted compromise tree is output according to the calculated weights. This same output can be calculated by using the R-code (*wconsensus-test.R*) in the console command line or R command line. The *wconsensus* function and R code are available at <http://code.google.com/p/wconsensus/>, under a GPL 3.x licence, and instructions for their use are included.

Acknowledgements

We thank Mark Wilkinson, Graham Jones, Tom Shearin and members of the Hymenoptera Institute for comments on the manuscript; financial support was received via NSF grants EF-0337220 and DEB-0542864. D.R.M.-E. is indebted to División de Investigación y Extensión, Facultad de salud, Universidad Industrial de Santander (project 5658) and División de Investigación y Extensión, Facultad de Ciencias, Universidad Industrial de Santander (project 5132).

References

- Adams, E.N. III, 1972. Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.* 21, 390–397.
- Anderberg, A.A., Tehler, A., 1990. Consensus trees, a necessity in taxonomic practice. *Cladistics* 6, 399–402.
- Barrett, M., Donoghue, M.J., Sober, E., 1991. Against consensus. *Syst. Zool.* 40, 486–493.

- Bremer, K., 1990. Combinable component consensus. *Cladistics* 6, 369–372.
- Bryant, D., 2003. A classification of consensus methods for phylogenetics. In: Janowitz, M., Lapointe, F.J., McMorris, F.R., Mirkin, B., Roberts, F.S. (Eds.), *Bioconsensus*. DIMACS-AMS, pp. 163–184.
- Cadotte, M.W., Cardinale, B.J., Oakley, T.H., 2008. Evolutionary history and the effect of biodiversity on plant productivity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 17012–17017.
- Carpenter, J.M., 1988. Choosing among multiple equally parsimonious cladograms. *Cladistics* 4, 291–296.
- Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, SIAM, Philadelphia.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, 24, 774–786.
- Grande, C., Templado, J., Zardoya, R., 2008. Evolution of gastropod mitochondrial genome arrangements. *BMC Evol. Biol.* 8, 61.
- Highton, R., Hedges, S.B., Hass, C.A., Dowling, H.G., 2002. Snake relationships revealed by slowly-evolving proteins: further analysis and a reply. *Herpetologica* 58, 270–275.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–755.
- Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51, 673–688.
- Margush, T., McMorris, F.R., 1981. Consensus n-trees. *Bull. Math. Biol.* 43, 239–244.
- Martin, F.N., Tooley, P.W., 2003. Phylogenetic relationships among *Phytophthora* species inferred from sequence analysis of mitochondrially encoded cytochrome oxidase I and II genes. *Mycologia* 95, 269–284.
- Miyamoto, M.M., 1985. Consensus cladograms and general classifications. *Cladistics* 1, 186–189.
- Nelson, G.J., 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Families des Plantes (1763-1764). *Syst. Zool.* 28, 1–21.
- Nixon, K.C., Carpenter, J.M., 1996. On consensus, collapsibility, and clade concordance. *Cladistics* 12, 305–321.
- Nixon, K.C., Wheeler, Q.D., 1991. Extinction and the origin of species. In: Wheeler, Q.D., Novacek, M. (Eds.), *Extinction and Phylogeny*. Columbia University Press, New York, pp. 119–143.
- Page, R.D.M., 1989. Comments on component-compatibility in historical biogeography. *Cladistics* 5, 167–182.
- Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290.
- R Development Core Team. 2011. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, <http://www.R-project.org>.
- Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Sharkey, M.J., Leathers, J.W., 2001. Majority does not rule: the trouble with majority-rule consensus trees. *Cladistics* 17, 282–284.
- Steel, M., Dress, A.W.M., Bocker, S., 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Syst. Biol.* 49, 363–368.
- Sumrall, C.D., Brochu, C.A., Merck, J.W. Jr, 2001. Global lability, regional resolution, and majority-rule consensus bias. *Paleobiology* 27, 254–261.
- Swofford, D.L. 1990. PAUP: phylogenetic analysis using parsimony, version 3.0q. Illinois Natural History Survey, Champaign, IL.
- Swofford, D.L., 1991. When are phylogeny estimates from molecular and morphological data incongruent? In: Miyamoto, M.M., Cracraft, J. (Eds.), *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York, pp. 295–333.
- Wilkinson, M., 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic trees and profiles. *Syst. Biol.* 43, 343–368.
- Wilkinson, M., Thorley, J.L., 2001. Efficiency of strict consensus trees. *Syst. Biol.* 50, 610–612.