Efficiency of Strict Consensus Trees

MARK WILKINSON[1] AND JOSEPH L. THORLEY[1,2]

[1]*Department of Zoology, The Natural History Museum, London SW7 5BD, UK; E-mail: marw@nhm.ac.uk*
[2]*School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK*

Consensus trees are used in phylogenetics as summaries or representations of sets of source trees. Here we ask 'How good are consensus trees?' in the sense of how well do individual consensus trees represent the set of source trees for which they stand? There are many different consensus methods and various contexts in which they may be used (Swofford, 1991; Wilkinson, 1994; Leclerc, 1998). Consequently, answers to our question must be specific as to both method and context. For example, majority-rule consensus trees (Margush and McMorris, 1981; Wilkinson, 1996) can provide useful graphical summaries of bootstrap or jackknife analyses but can be problematic when used to represent a set of equally optimal trees from the analysis of a single data set (Wilkinson and Benton, 1996). Here we focus on strict consensus methods sensu Wilkinson (1994), that is, methods that require unanimous agreement among the source trees, and on the contexts in which they are commonly used. Contexts in which strict consensus methods are used include the representation of the set of optimal trees for a single data set, the comparison of simulated trees and trees inferred from simulated data, and the quantification of the similarity of trees derived from different data sets in studies of taxonomic congruence. Here we describe a simple measure of consensus efficiency that allows us to say how well a particular strict consensus tree is doing its job of faithfully representing the source trees.

Consensus methods differ in the type of information they represent and the level of agreement required among the source trees for information of that type to be included in the consensus tree (Page, 1992). This is reflected in the consensus terminology of Wilkinson (1994), as we use here, in which the names of consensus methods combine descriptors of the type of information (e.g., component, Adams) and the level of agreement (e.g., strict, majority-rule).

Strict consensus trees provide information by permitting (or, conversely, prohibiting) a subset of the possible trees (Page, 1992; Wilkinson, 1994; Thorley et al., 1998). Consensus efficiency is a relation between the trees permitted by the consensus tree and the source trees. An ideal or maximally efficient strict consensus tree would permit only the source trees that it represents. Consensus trees might deviate from the ideal in two ways. First, they might permit trees that are not source trees, and second they might fail to permit some of the source trees. Both behaviors would reduce the correspondence between the consensus and the source trees the consensus is intended to represent and thereby would reduce the efficiency of the consensus tree. In practice, strict consensus trees must permit all the source trees. Thus the efficiency of consensus trees is maximal when it permits only the source trees and is reduced as it permits additional trees. A maximally inefficient consensus representation is a consensus tree that prohibits no trees (i.e., a bush) when the set of source trees does not include all possible binary trees. A measure of consensus efficiency (CE) that has these properties and that ranges between values of zero (minimal efficiency) and one (maximal efficiency) is given by:

$$CE = (\log T - \log C)/(\log T - \log S)$$

where T is the number of possible binary trees for the set of leaves (terminal taxa) under consideration, C is the number of binary trees permitted by the consensus tree, and S is the number of binary source trees. Where the source trees include polytomies, these are interpreted as standing for all the binary

trees that are possible resolutions of the poly-tomies. Formulae for T and C are given by Rohlf (1982). By convention, zero divided by itself is taken as unity (1). Thus if the consensus is a bush, its efficiency is either zero, when the source trees do not include all possible binary trees, or one, when the source trees include all possible binary trees. In the latter case, although the consensus tree is phylogenetically uninformative, it is a perfectly accurate representation of the source trees. CE is related to the measure of cladistic information content developed by Thorley et al. (1998), essentially being equivalent to the ratio of the cladistic information content of the consensus tree and the cladistic information content of the set of source trees.

CE ranges between zero and one only if consensus trees are interpreted correctly. Values of CE that exceed one indicate that a consensus tree has been incorrectly interpreted as permitting fewer trees than it actually permits. Such misinterpretations could happen if, for example, a strict Adams or a majority-rule component consensus tree is interpreted as if it is a strict component consensus tree. Such misinterpretations do sometimes occur. For example, Nielsen et al. (1996) interpreted an Adams consensus of equally parsimonious trees of metazoan phyla (their Fig. 1) as if it were a strict component consensus tree.

If consensus trees are interpreted correctly, then CE gives an indication of how satisfied we can be with any particular consensus tree. The CE is determined both by properties of the source trees and by properties of the consensus method. A low CE may indicate that the consensus method has undesirable properties or that the source trees are very dissimilar. In the latter case few relationships are true of all the source trees and only limited agreement can be represented in any consensus tree. Lack of resolution in strict component consensus trees is a common problem (Wilkinson, 1995)—sometimes due more to undesirable properties of this consensus method than to incongruence among the source trees (Wilkinson et al., 1996). A strict component consensus tree that has a high CE can be accepted as a good representation of the source trees. If its CE is low, other strict consensus methods, such as largest common pruned trees (Gordon, 1980), also known as greatest or maximum agreement subtrees, or reduced consensus meth-
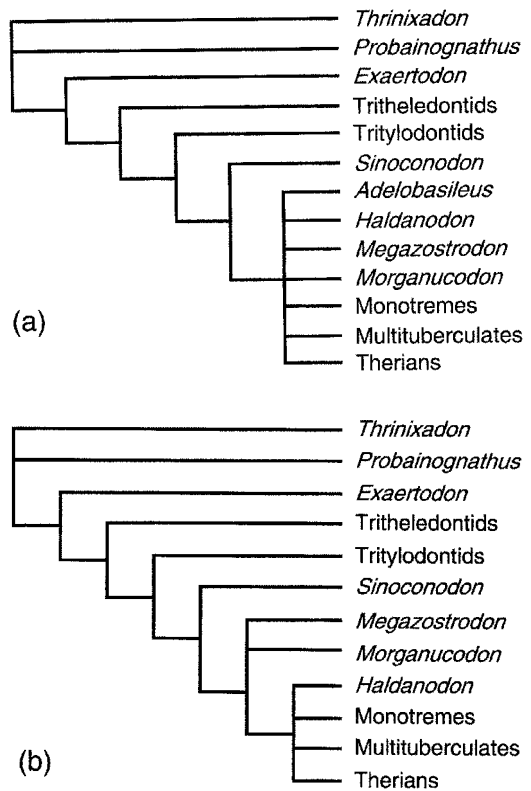

FIGURE 1. A strict RCC profile for 93 most-parsimonious binary trees supported by the data of Lucas and Luo (1993), comprising a strict component consensus tree (a) and a single additional RCC tree (b).

ods (Wilkinson, 1994) should be investigated. Only if all methods have a low efficiency can this result be confidently attributed to properties of the source trees, namely, their great dissimilarity or incongruence, or of the data on which they are based rather than to the limitations of the strict component consensus method.

Inferences from the lack of resolution of strict component consensus trees are widespread. A common use is to infer extensive incongruence in character data (e.g., Suter, 1994; Reeder, 1995) and such an inference is implicit in the use of strict component consensus trees in total support randomization tests of phylogenetic data (Källersjö et al., 1992). Where the strict component consensus is used for source trees from different sets of data, lack of resolution may be interpreted as indicating strong disagreement between the trees or between the separate data sets (e.g., Jenner and Schram, 1999). In a related context, resolution of a strict

component consensus of a simulated tree and of a tree inferred from simulated data is sometimes used to measure the accuracy of the inferred tree and to draw conclusions about the reliabilities of different methods of analysis (e.g., Wiens, 1998). Each of these uses is reasonable where CE is high, but where CE is low, additional work is needed to ensure that features of strict component consensus trees that are consequences of undesirable properties of that method are not wrongly attributed to properties of the source trees, the data, or the methods of analysis.

The underlying rationale of using CE as an indicator of possible problems with a consensus method is implicit in Wilkinson's (1999) reevaluation of Lucas and Luo's (1993) phylogenetic study of early mammals. He argued that a large difference between the number of most-parsimonious source trees and the number of trees permitted by their strict component consensus tree suggested that alternative consensus methods might produce better summaries of the most-parsimonious trees. CE provides a scaled measure of this difference. The data yield 93 most-parsimonious binary trees (the source trees), but their strict component consensus (Fig. 1A) permits 31,185 trees. For this consensus tree, CE = 0.735. Wilkinson (1999) applied the strict reduced cladistic consensus (RCC) method to the 93 source trees in an attempt to discover a better consensus tree. The RCC profile includes the strict component consensus (Fig. 1A) and one additional RCC tree (Fig. 1B) that is more resolved but includes one less leaf (*Adelobasileus* is excluded). The latter tree permits 3,105 binary trees (that include *Adelobasileus*). This RCC tree might be considered either better or worse than the strict component consensus tree, depending on whether resolution or inclusiveness is considered the more desirable. However, judged solely in terms of the efficiency of the consensus representation (CE = 0.840), it is a better representation of the source trees than is the corresponding strict component consensus tree. In this case, it is also possible to determine the efficiency of the entire RCC profile. Together, the two trees permit 1,485 trees, corresponding to a combined CE of 0.874.

To conclude, CE measures how well or how badly a given strict consensus tree represents a set of source trees. It informs us how satisfied we can be with a particular consensus tree and can affect what conclusions we can draw from a particular consensus tree. Inferring from poorly resolved consensus trees extensive conflict among the data may be reasonable when CE is high. When CE is low, however, the consensus tree is a poor representation of the set of source trees. Consequently, any lack of resolution can reflect the properties of the consensus method as much as the properties of the source trees or of the data. The methodological null hypothesis that a low CE is due to undesirable properties of the consensus method should be refuted before a poorly resolved consensus tree is interpreted as indicating highly incongruent trees or data. RadCon (Thorley and Page, 2000) computes CE measures for a variety of strict consensus methods when the source trees are binary.

REFERENCES

GORDON, A. D. 1980. On the assessment and comparison of classifications. Pages 149–160 *in* Analyse de Donnees et Informatique (R. Tomasone, ed.). INRIA, LeChesnay, France.

JENNER, R. A., AND F. R. SCHRAM. 1999. The grand game of metazoan phylogeny: Rules and strategies. Biol. Rev. 74:121–142.

KÄLLERJÖ, M., J. S. FARRIS, A. G. KLUGE, AND C. BULT. 1992. Skewness and permutation. Cladistics 8:275–287.

LECLERC, B. 1998. Consensus of classifications: The case of trees. Pages 81–90 *in* Advances in data science and classification (A. Rizzi, M. Vichi, and H.-H. Bock, eds.). Springer, Berlin.

LUCAS, S. G., AND Z. LUO. 1993. *Adelobasileus* from the Upper Triassic of West Texas: The oldest mammal. J. Vert. Paleonol. 13:309–334.

MARGUSH, T., AND F. R. MCMORRIS. 1981. Consensus n-trees. Bull. Math. Biol. 43:239–244.

NIELSEN, C., N. SCHARFF, AND D. EIBYE-JACOBSEN. 1996. Cladistic analysis of the animal kingdom. Biol. J. Linn. Soc. 57:385–410.

PAGE, R. D. M. 1992. Comments on the information content of classifications. Cladistics 8:87–95.

REEDER, T. W. 1995. Phylogenetic relationships among phrynosomatid lizards as inferred from mitochondrial ribosomal DNA sequences: Substitutional bias and the information content of transitions relative to transversions. Mol. Phylogenet. Evol. 4:203–222.

ROHLF, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. 59:131–144.

SUTER, S. J. 1994. Cladistic analysis of cassiduloid echinoids: Trying to see the phylogeny for the trees. Biol. J. Linn. Soc. 53:31–72.

SWOFFORD, D. L. 1991. When are phylogency estimates from molecular and morphological data incongruent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, Oxford.

THORLEY, J. L., AND R. D. M. PAGE. 2000. RadCon: Phylogenetic tree comparison and consensus. Bioinformatics 16:486–487.

THORLEY, J. L., M. WILKINSON, AND M. A. CHARLESTON. 1998. The information content of consensus trees. Pages 91–98 *in* Advances in data science and classification (A. Rizzi, M. Vichi, and H.-H. Bock, eds.). Springer, Berlin.

WIENS, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.

WILKINSON, M. 1994. Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43:343–368.

WILKINSON, M. 1995. Coping with missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501–514.

WILKINSON, M. 1996. Majority-rule reduced consensus methods and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.

WILKINSON, M. 1999. Choosing and interpreting a tree for the oldest mammal. J. Vert. Paleonotol. 19: 187–190.

WILKINSON, M., AND M. J. BENTON. 1996. Sphenodontid phylogeny and the problems of multiple trees. Philos. Trans. R. Soc. London B 351:1–16.

WILKINSON, M., S. J. SUTER, AND V. L. SHIRES. 1996. The reduced cladistic consensus method and cassiduloid echinoid phylogeny. Hist. Biol. 12:63–73.

Misleading Results from the Use of Ambiguity Coding to Score Polymorphisms in Higher-Level Taxa

NANCY B. SIMMONS

Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA; E-mail: simmons@amnh.org

The operational taxonomic units (OTUs) or terminals used in phylogenetic analyses range from individuals (e.g., haplotypes) to higher-level taxa (e.g., genera, families, orders), depending on goals and the availability of samples. Molecular systematic studies typically sample at the species level or below, but this is not true of all morphological studies, some of which include genera, families, or orders as OTUs. The latter approach is usually employed when goals include resolution of higher-level relationships within a group that is too species-rich to sample fully at lower taxonomic levels (e.g., Mammalia), or when available samples (e.g., fossils) do not permit scoring every species for every character. Scoring characters at the level of genera or families, for example, may permit complete scoring of OTUs in which different character systems have been studied in different species. This approach reduces missing data in the resulting matrix and may facilitate integration of difficult-to-obtain character data (e.g., details of fetal membrane development) that are available for only a limited number of species.

One problem that emerges when superspecific taxa are used as OTUs is taxonomic polymorphism, which occurs when characters exhibit two or more states within a higher-level taxon while remaining fixed within species (Nixon and Davis, 1991). Much has been written in the last decade about strengths and weaknesses of different methods of coding taxonomic polymorphisms for phylogenetic analysis (e.g., Nixon and Davis, 1991; Maddison and Maddison, 1992; Donoghue, 1994; Nixon and Carpenter, 1996; Rice et al., 1997; Bininda-Emonds et al., 1998; Wiens, 1998; Kornet and Turner, 1999; Prendini, in press). The goal of most approaches is to hypothesize the ancestral or plesiomorphic condition for each superspecific OTU, or to at least narrow the field of possibilities in ambiguous cases. On the basis of a series of computer simulations, Wiens (1998) concluded that coding a superspecific taxon based on the most common condition