Cladistics (1996) 12:83–92

FORUM

ON CONSENSUS, CONFIDENCE, AND "TOTAL EVIDENCE"

Roderic D. M. Page

Department of Zoology, University of Oxford South Parks Road, Oxford OX1 3PS, England, U.K.

Received for publication 21 May 1995; accepted 23 February 1996

Introduction

The issue of whether it is best to combine separate data sets ("total evidence"), or to combine trees ("taxonomic congruence") has received considerable attention in the recent literature (e.g. Kluge, 1989; Barrett et al. 1991; Kluge and Wolf, 1993; Miyamoto and Fitch, 1995). In this note I discuss some recent arguments that have been made regarding this issue, starting with Barrett et al.'s (1991) argument against consensus methods. I then touch on bootstrapping and the problem of confidence sets of trees (Sanderson, 1989), before considering the implication of the difference between gene trees and species trees for the concept of total evidence. The common theme underlying these topics is the role of consensus in systematics. Consensus trees are very useful tools, but they have their limits. However, the nature of these limits is still being explored.

For and Against Consensus

In their provocatively titled paper Barrett et al. (1991) presented an example where the consensus of the most parsimonious trees for two data sets contradicted the tree found for the two data sets combined. They constructed two data sets each comprising four weighted, polarised characters for four taxa. Table 1 lists the lengths of all 15 possible rooted binary trees for four taxa for each data set, and for the combined data (which is simply the sum of the length for the two data sets). Plotting this information (Fig. 1) reveals that Barrett et al.'s example is cleverly constructed. For both data sets 1 and 2 the minimal length tree (8 and 7, respectively) has a length of 10 steps. However, tree 8 has a length of 13 steps for data set 2, and tree 7 has a length of 13 steps for data set 1; hence, each tree has a length of 23 steps for the two data sets combined. If 23 steps was also the minimum possible length for the combined data then "total evidence" and consensus would arrive at the same result: two equally good trees whose shared information may be summarised by the consensus tree (A, (B, C, D,)).

Logically the shortest possible tree for the two data sets combined is 20 (=10+10)

Present address: Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K. Email: r.page@bio.gla.ac.uk.

0748-3007/96/010083+10/$18.00/0

steps. Hence, there is a zone (Fig. 1) of tree lengths ≥20 and <23, in which if one or more trees were found they would be more parsimonious for the combined data than the best trees for the separate data sets. If such a tree existed, and it had a length of >10 steps for both data sets 1 and 2 then that tree would contradict the minimal trees for the two separate data sets (this must be so as it is a different tree). Barrett et al.'s (1991) example contains just such a tree; tree 12 has a length of 11 steps for both data sets 1 and 2, giving it a combined length of 22 steps, one step less than trees 7 and 8, which are each optimal for one of the separate data sets.

For Barrett et al.'s (1991) example to work there must be at least one tree in this zone. This has the following, perhaps paradoxical implication. Some advocates of total evidence favour rejecting any sub optimal hypothesis from consideration (e.g. Kluge and Wolf, 1993: 190–191) as this can quickly lead to descent down a "slippery slope". Yet Barrett et al.'s argument for the superiority of combining data sets over using consensus trees depends on the existence of sub optimal trees. If there are no sub optimal trees in the zone shown in Figure 1 then combining data sets cannot lead to a different result from taxonomic congruence. Hence, advocacy of total evidence amounts to admitting the possibility that a sub optimal tree for a given data set may turn out to be better supported by subsequent data. If this is not the case, there are no grounds for favouring total evidence over consensus; total evidence and consensus would produce the same result.

Consensus and Distances

In discussing the role of consensus in systematics, Barrett et al. (1991: 491) agreed that consensus methods could be used to address the problem of "How do the results obtained for different data sets compare when those data sets cannot in principle be combined (e.g. DNA-DNA hybridisation and discrete characters interpreted by parsimony)?" Miyamoto and Fitch (1995: Table 1) stated that consensus was the only option, unless we "deny the informativeness of some categories of evidence." Figure 1 suggests a third approach; provided we can assign a number to each tree that describes the relationship between that tree and some data (e.g. least squares fit, sum of branch lengths) then in principle we could compare two (or more) data sets and discover whether sub optimal trees exist that are "better" overall than the best trees for the separate data sets. Of course, this raises the question of whether different measures are comparable, but in principle this problem is encountered whenever we compare different data sets. Goodman et al. (1990) found a good correlation between percent sequence divergence and $\Delta T_{50}H$ measures of DNA–DNA hybridisation distance, suggesting that DNA–DNA distances can be expressed in the same units as discrete sequence data. The apparent noncombinability of discrete and distance data often referred to in the literature is due as much to limitations in existing phylogenetic software as to any logical difficulties.

For small number of taxa, diagrams of the kind shown in Fig. 1 can be readily produced by obtaining the values of the two optimality criteria (e.g. number of steps, sum of least squares branch lengths) for each tree and plotting them. For larger numbers of trees where heuristic methods are employed, this approach requires software that can save sub optimal trees. This facility is available in current

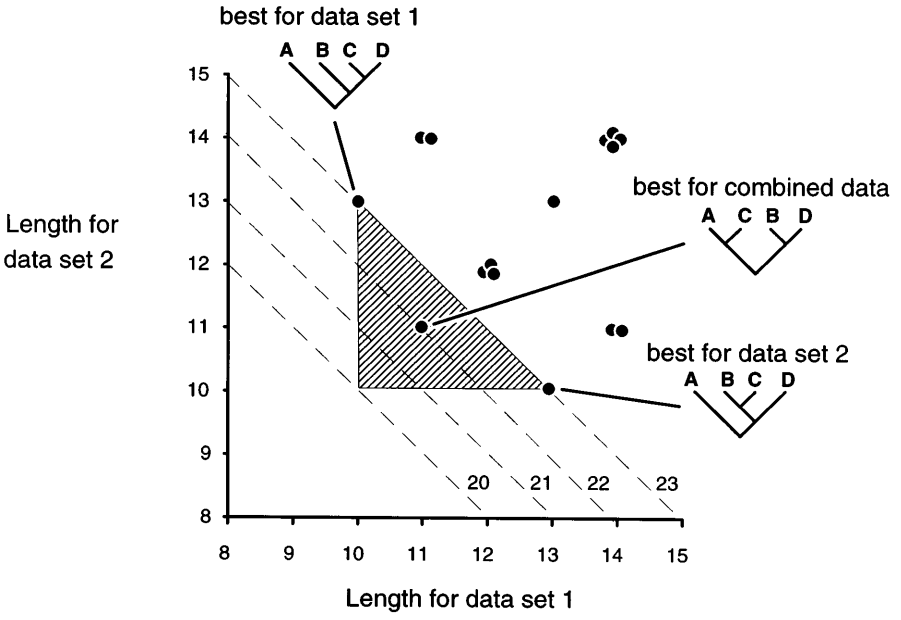

Fig. 1. Tree lengths for all 15 possible trees for Barrett et al.'s (1991) data sets, with the best trees for the individual and combined data sets indicated. The hatched zone encompasses the possible combination of lengths for trees that are suboptimal for either data set taken separately, but are more parsimonious for the combined data than either of the best trees for the data sets considered separately. Trees falling on the same dashed line (sloping down from left to right) will have the same length for the combined data sets. This length is indicated at the bottom of each line.

implementations of the parsimony criterion, and the next release of PAUP (Swofford, 1990) will extend this feature to other criteria.

Confidence

To illustrate his arguments on total evidence, Kluge (1989) presented an analysis of biochemical and morphological data for the snake genus *Epicrates*. Kluge showed that the most parsimonious trees for these two data sets differed, and advocated combining the two data sets. He suggested that the incongruence was primarily due to some morphological characters being jointly affected by paedomorphosis, whereas Swofford (1991) suggested that some of the biochemical characters were not independent.

Table 1
Lengths for all 15 possible trees for four taxa for Barrett et al.'s (1991) data sets 1 and 2, and for the two data sets combined. The shortest trees are indicated in **boldface**.

Tree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Data set 1	14	11	14	13	14	14	13	**10**	12	14	12	11	12	11	14
Data set 2	14	14	14	13	14	11	**10**	13	12	11	12	11	12	14	14
Combined	28	25	28	26	28	25	23	23	24	25	24	**22**	24	25	28

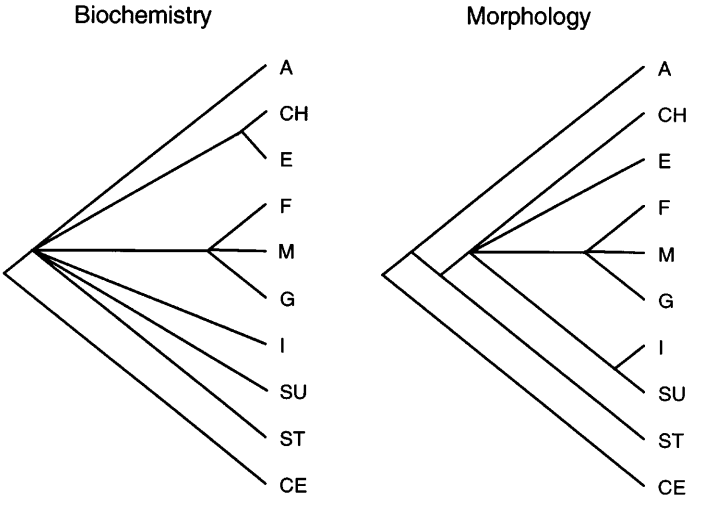

Fig. 2. $M_{0.95}$ consensus trees (i.e. containing only those clusters found in ≥95% of the input trees) for 1000 bootstrap trees computed for biochemical and morphological data sets for *Epicrates*. Note that the two trees are consistent with each other; they either resolve the same groupings or one tree contains groups compatible with the other (see text).

A third possibility is sampling error. If a data set contains homoplasy then different characters support different trees, hence which tree (or trees) a given data set supports will depend on which characters have been sampled. As a consequence estimates of phylogeny based on samples are accompanied by sampling error. Patterson et al.'s (1993) review of congruence between morphological and molecular data sets concluded that incongruence is common, and clearly this is a crucial problem for phylogenetics. However, it is important to distinguish between genuine cases of incongruence, which may tell us about evolutionary processes and our ability to reconstruct phylogeny, and spurious cases of incongruence due to inadequate samples.

Bootstrapping (Felsenstein, 1985) is one method that is often used to estimate sampling error in phylogenetics. The results of bootstrapping are frequently summarised by a consensus tree, such as a majority-rule tree. This tree belongs to the M_l family of consensus trees (McMorris and Neumann, 1983), where l is the proportion of trees in which a cluster must appear in order to be included in the consensus tree. For the majority-rule consensus tree $l>0.5$, for the strict consensus tree, $l=1.0$. Hence, if we were to adopt Felsenstein's (1985) rule-of-thumb and regard only those clusters with bootstrap proportions ("*P* values") of 0.95 or greater as well supported, we could use the $M_{0.95}$ consensus tree to summarise the bootstrap trees. Any tree that was a refinement of the $M_{0.95}$ tree (i.e. contained all the clusters in the $M_{0.95}$ tree together with one or more other compatible clusters) would be regarded as consistent with the data. As an example, Fig. 2 shows the $M_{0.95}$ tree for 1000 bootstrap trees computed for the morphological and biochemical data for *Epicrates*. The two trees are mutually consistent, that is, no clusters in either tree are incompatible with clusters in the other tree. This suggests the dis-

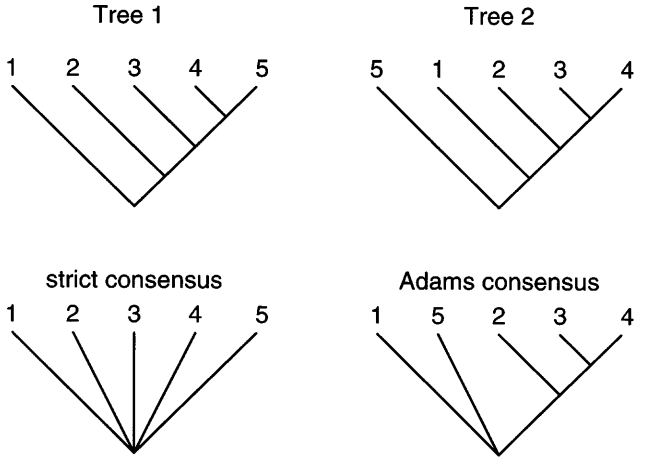

Fig. 3. Two trees that have no cluster in common and their strict and Adams consensus. Note that the Adams consensus tree captures more of the similarity between the two trees.

crepancy between the trees inferred from these two data sets could be attributed to sampling error.

One difficulty with the example presented above is that although the consensus tree summarises the structure common to the bootstrap trees, not all trees consistent with the consensus tree need be in the set of bootstrap trees (here called the bootstrap *profile*). Hence, even if two consensus trees are consistent, this does not demonstrate that the two sets of trees from which the consensus trees were constructed contain trees in common.

Another difficulty is that the majority-rule tree has the undesirable property (which it shares with other cluster-based consensus methods such as strict and Nelson consensus (Page, 1989)) that two trees may have no clusters in common yet differ in the placement of only a single terminal (Fig. 3). Hence, even if the bootstrap trees share considerable structure in common, all it takes is one sequence whose position is highly uncertain to collapse the *P* values. Other consensus methods, such as Adam's (1986) consensus are less sensitive to this phenomenon. The practical implication is that low bootstrap values by themselves need not indicate that the data set as a whole is poor. It could be that the data permit a precise estimate of relationships, but for only a subset of the sequences.

Confidence Sets of Trees

Sanderson (1989) suggested an alternative way of summarising the results of the bootstrapping that retains the idea that the confidence interval around an estimate of the phylogeny is a set of trees. We can visualise the bootstrap profile as a cloud of trees in "tree space" (Fig. 4). At the centre of this cloud is the "best estimate" of the phylogeny. If we measure the distance between this central tree and all the other trees using a tree comparison metric (Penny and Hendy, 1985; Steel and Penny, 1993) then the percentage (p) of bootstrap trees closest to the central tree estimates the p confidence interval around that tree.

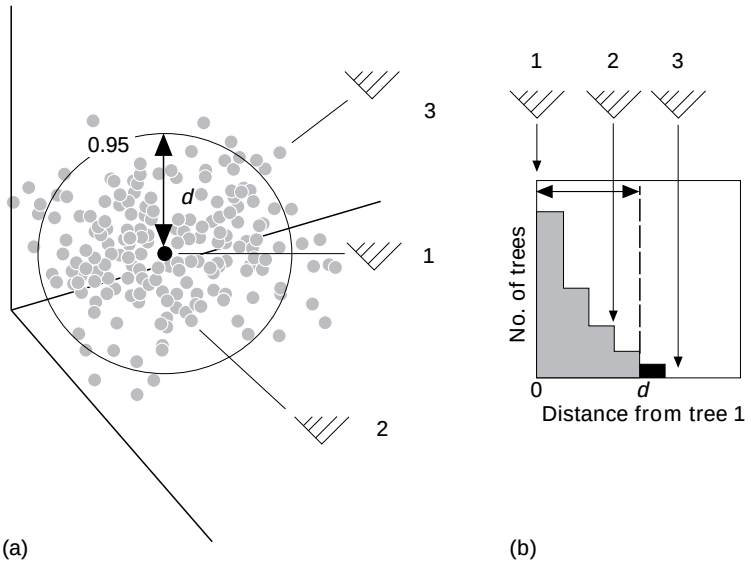

Fig. 4. (a) The set of bootstrap trees represented as a cloud of trees in "tree space". Tree 1 is at the centre of the cloud and represents the best estimate of the phylogeny. Of the bootstrap trees 95% are within distance *d* of tree 1 (b) and comprise the 95% confidence sets of trees. Tree 2 is a member of this set, but tree 3 is not.

Sanderson (1989) suggested using the most parsimonious tree as the central tree. However, if we use the partition metric, d_s, (Penny and Hendy, 1985) as our measure of tree similarity then the natural choice is the majority-rule consensus tree as it is a median tree (Barthélemy and McMorris, 1986). Given a set of trees (the "profile"), a median tree is a tree which has the smallest average dissimilarity to all the trees in the profile.

If the majority-rule consensus tree is not binary (i.e. not fully resolved), and all trees in the bootstrap profile are binary, then the consensus tree itself may not be a member of the bootstrap profile. This may seem undesirable, but a simple analogy with sample of integers shows that this impression is misleading. Consider the set of integers $s = \{6, 7, 7, 8, 9, 12\}$ sampled from a larger set. Set *s* has a mean of 8.167 and a median of 7.5, yet neither 8.167 nor 7.5 are elements of *s*. Note that were we to insist that the median tree was a member of the bootstrap profile, we would have to compute the median binary tree (assuming the bootstrap trees are themselves all fully resolved). Computing this tree is an NP-hard problem (McMorris and Steel, in press), and so is considerably less attractive computationally than the majority-rule tree.

To compute the confidence set of tree we first find the majority-rule consensus tree, and then compute the distance between that consensus tree and every tree in the profile. We then decide on the desired confidence interval (e.g. 95%) and keep the corresponding proportion of trees closest to the consensus tree. We need not be limited to just the partition metric, indeed other metrics may be more discriminating (Sanderson 1989; Page, 1993a).

This approach has received little attention (but see Rodrigo et al., 1993; Hillis, 1995), perhaps in part due to the lack of software for comparing trees. Such

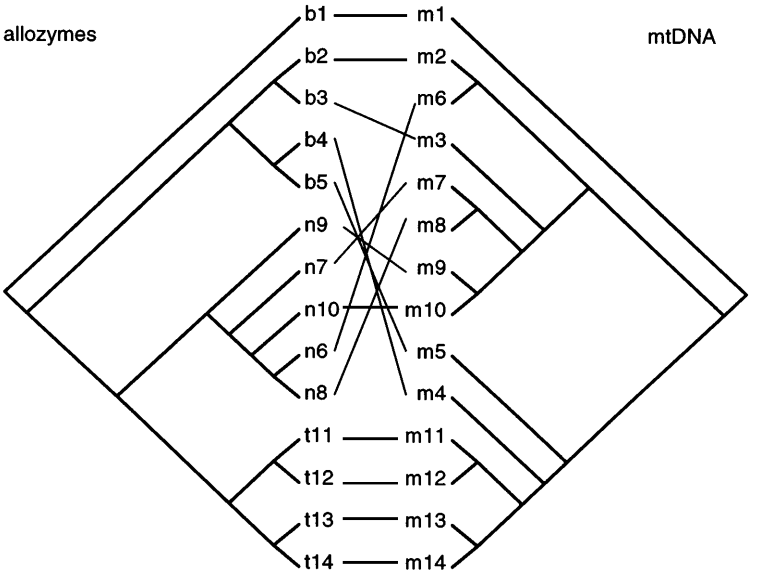

Fig. 5. Trees for 14 populations of *Thomomys bottae* and *T. townsendi* pocket gophers inferred from allozymes and from mitochondrial DNA sequences (after Patton and Smith, 1994).

software is now readily available: COMPONENT (Page, 1993a) can compute partition, nearest neighbour, quartet, triplet, and agreement subtree metrics (PAUP (Swofford, 1990) can also compute the partition metric). The manual for COMPONENT contains a worked example of this approach.

Given two data sets for which the phylogenies are incongruent, we could see whether this incongruence might plausibly be due to sampling error by seeing whether their bootstrap profiles overlapped. As a rule of thumb, if there is a tree (or trees) common to the two bootstrap profiles then sampling error is a reasonable explanation of the disagreement between the two data sets. However, if the bootstrap profiles do not overlap then sampling error is insufficient to explain the disagreement. Rodrigo et al. (1993) used this idea to compare bootstrap trees for morphological and molecular data sets for sponges; however, they compared the entire bootstrap profiles, rather than a subset of the trees as described here. Effectively, Rodrigo et al. (1993) used a confidence interval of 100%.

Bull et al. (1993) have argued that different data sets for the same taxa should be combined only when the data sets are homogeneous. Two data sets with overlapping bootstrap profiles are homogenous in the sense that there are trees both data sets support. In such cases we would be justified in combining the two data sets in the hope of obtaining a more precise phylogenetic estimate. Of course, this does not imply that both data sets need to be analysed under the same evolutionary model.

Gene Trees and Species Trees

Generally, we expect congruence between different data sets. This is particularly the case when the data sets are drawn from the same system, e.g. different genes

from the mitochondrial genome (e.g. Árnason and Johnsson, 1992) which is inherited as a single unit without recombination, hence all genes are tracking the same phylogeny. However, our expectation of congruence will be unjustified if the entities described in the data sets are tracking different histories. Patton and Smith (1994) provide a recent and sobering example of the possible complexity of relationships between gene and species trees (Fig. 5).

Doyle (1992) made use of the possible discrepancy between different gene trees to argue that each gene (or set of linked genes) constitutes a single character, hence, nucleotide sites are characters of genes, not of organisms. In one sense this suggestion is nothing new; after all, several authors have advocated this treatment for allozyme data where individual loci are characters and alleles are character states (e.g. Mickevich and Mitter, 1981). The treatment of such data as unordered multistate characters reflects our ignorance of the character state tree connecting the alleles. Nucleotide sequences provide that information and hence allow us to construct character state trees for alleles at a locus (Swofford and Berlocher, 1987: 313).

While some authors have used the gene tree/species tree distinction (among other considerations) to argue for consensus methods (e.g. de Queiroz, 1993; Miyamoto and Fitch, 1995), consensus trees are likely to be a poor choice for the study of historically associated lineages. Intraspecific polymorphism and duplicated genes can result in complicated gene trees that violate a basic requirement of current consensus methods; that the trees contain the same number of terminals. Furthermore, genes may be transferred horizontally. These kinds of problem are familiar to biogeographers and parasitologists (e.g Nelson and Platnick, 1981; Brooks and McLennan, 1991; Page, 1994), and parallels between these problems and molecular systematics are becoming increasingly appreciated (Page, 1988, 1993b; Baum, 1992; Doyle, 1992). Given these parallels, molecular systematics may come to resemble cladistic biogeography (Nelson and Platnick, 1981) more than orthodox systematics. Individual gene phylogenies would be combined to provide the best estimate of organismal phylogeny. Simply concatenating nucleotide sequences under the maxim "total evidence" may not be the best use of the evidence provided by those sequences. Surely all systematists want to use all the available evidence; the question is how best to interpret that evidence.

Acknowledgements

Much of this paper is based on a talk presented at the XIIIth meetings of the Willi Hennig Society in Copenhagen, 1994. I thank Ole Seberg for the invitation to attend, and for suggesting the topic of congruence. The helpful, candid comments of two anonymous referees were much appreciated.

REFERENCES

ADAMS, E. N. 1986. *N*-trees as nestings: Complexity, similarity, and consensus. J. Classif. 3: 299–317.

ÁRNASON, Ú. AND E. JOHNSSON. 1992. The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina*. J. Mol. Evol. 34: 493–505.

BARRETT, M., M. J. DONOGHUE, AND E. SOBER. 1991. Against consensus. Syst. Zool.40: 486–493.

BARTHÉLEMY, J. -P. AND F. R. MCMORRIS. 1986. The median procedure for *n*-trees. J. Classif. 3: 329–334.

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.

BROOKS, D. R. AND D. A. MCLENNAN. 1991. Phylogeny, Ecology, and Behavior. Chicago University Press, Chicago.

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD, AND P. J. WADDELL. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42: 384–397.

DE QUEIROZ, A. 1993. For consensus (sometimes). Syst. Biol. 42: 368–372.

DOYLE, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17: 144-163.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.

GOODMAN, M., D. A. TAGLE, D. H. A. FITCH, W. J. BAILEY, J. CZELUSNIAK, B. F. KOOP, P. BENSON, AND J. L. SLIGHTOM, 1990. Primate evolution at the DNA level and a classification of the hominoids. J. Mol. Evol. 30: 260–266.

HILLIS, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44: 3–16.

KLUGE, A. G. 1989. A concern for evidence and phylogenetic hypothesis of relationship among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38: 7–25.

KLUGE, A. G. AND A. J. WOLF. 1993. Cladistics: What's in a word? Cladistics 9: 183–199.

MCMORRIS, F. R. AND D. NEUMANN. 1983. Consensus functions defined on trees. Math. Social Sci. 4: 131–136.

MCMORRIS, F. R. AND M. A. STEEL. (In press) The complexity of the median procedure for binary trees. *In:* Proceedings of the 4th Conference of the International Federation of Classification Societies. Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Berlin.

MICKEVICH, M. F. AND C. MITTER. 1981. Treating polymorphic characters in systematics. *In*: Funk, V. A. and D. R. Brooks (eds). Advances in cladistics: Proceedings of the first meeting of the Willi Hennig Society. New York Botanical Garden, New York, pp. 45–60.

MIYAMOTO, M. M. AND W. M. FITCH. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44: 64–76.

NELSON, G. AND N. I. PLATNICK. 1981. Systematics and biogeography: Cladistics and vicariance. Columbia University Press, New York.

PAGE, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool. 37: 254–270.

PAGE, R. D. M. 1989. Comments on components-compatibility in historical biogeography. Cladistics 5: 167–182.

PAGE, R. D. M. 1993a. COMPONENT. Ver. 2.0. The Natural History Museum, London.

PAGE, R. D. M. 1993b. Genes, organisms, and areas: the problem of multiple lineages. Syst. Biol. 42: 77–84.

PAGE, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43: 58–77.

PATTERSON, C., D. M. WILLIAMS AND C. J. HUMPHRIES. 1993. Congruence between molecular and morphological phylogenies. Annu. Rev. Ecol. Syst. 24: 153–188.

PATTON, J. L. AND M. F. SMITH. 1994. Paraphyly, polyphyly, and the nature of species boundaries in pocket gophers (genus *Thomomys*). Syst. Biol. 43: 11-26.

PENNY, D. AND M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. 34: 75–82.

RODRIGO, A. G., M. KELLY-BORGES, P. R. BERQUIST, AND P. L. BERQUIST. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. N. Z. J. Bot. 31: 257–268.

SANDERSON, M. J. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5: 113–129.

STEEL, M. A. AND D. PENNY. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. 42: 126–141.

SWOFFORD, D. L. 1990. PAUP—Phylogenetic Analysis Using Parsimony. Ver. 3.0. Illinois Natural History Survey, Champaign, Illinois.

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? *In*: Miyamoto, M. M. and J. Cracraft (eds). Phylogenetic analysis of DNA sequences. Oxford University Press, New York. pp. 295–333.

SWOFFORD, D. L. AND S. H. BERLOCHER. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst. Zool. 36: 293–325.