Against Consensus

Martin Barrett; Michael J. Donoghue; Elliott Sober

Against Consensus

MARTIN BARRETT,[1] MICHAEL J. DONOGHUE,[2] AND ELLIOTT SOBER[1,3]

[1]Philosophy Department, University of Wisconsin,
Madison, Wisconsin 53706, USA
[2]Department of Ecology and Evolutionary Biology, University of Arizona,
Tucson, Arizona 85721, USA

In the project of phylogenetic inference, the idea of finding consensus trees has had an intuitive appeal. A single data set may yield several (often many) trees that are equally, or almost equally, parsimonious. The method of strict consensus instructs one to find those hypotheses about which all the best trees agree.

In addition, separate analyses of morphological and molecular data sets seem to provide a context in which consensus methods might appropriately be used. When the most parsimonious tree relative to morphological data differs from the most parsimonious tree constructed for a molecular data set, it has seemed reasonable to find the points of consensus. This will be a less than fully resolved tree that captures the hypotheses about which the two separate trees agree. Indeed, consensus techniques were originally designed to handle the problem of *different data sets*, rather than the problem of multiple trees for *a single data set* (Adams, 1972; Carpenter, 1988).

In both problems, finding the consensus tree has seemed to be a sensible, conservative strategy. By restricting one's final hypothesis to the points about which several competing hypotheses agree, one appears to run a reduced risk of being mis-

taken. The method of consensus has been viewed as a method of safety.

But why should the consensus tree be constructed from trees based on different data sets? Instead, why not pool the observations and find the most parsimonious tree for all of the data? Kluge (1983) suggested that when there are many molecular characters but few morphological ones, the result of pooling may be to "swamp" the morphological characters. Of course, characters can be pooled and a weighting scheme imposed (Miyamoto, 1985), but the worry has been that the weighting scheme cannot be objectively defended (e.g., Hillis, 1987). Consensus methods seem to possess the virtue of allowing biologists to avoid apparently unresolvable weighting problems.

Nonetheless, several commentators have suggested that the procedure is not without its difficulties. First, consensus trees are often highly unresolved, whereas for many purposes one would like a tree with as much resolution as possible. Second, if one has many molecular characters and few morphological ones, for example, the method of consensus appears to imply an equal weighting of data *sets* and hence an unequal weighting of the constituent characters (Cracraft and Mindell, 1989). If the two data sets count equally, then each molecular character would receive a lower weight than any morphological one. Here

---

the idea is that the consensus procedure implies weightings of its own, and these may be difficult to defend. A third problem is that the consensus tree will sometimes require more character-state changes than any of the separate trees require; when this is true, the consensus tree may be a misleading guide to patterns of character evolution (Miyamoto, 1985). One of the main points here, we believe, is that it may be inappropriate to interpret unresolved branch points as if they represented multiple simultaneous speciation events (Maddison, 1989). Finally, Kluge (1989) argued against consensus on the grounds that the choice among alternative consensus methods (see below) is effectively arbitrary.

In light of these pluses and minuses, Hillis (1987) recommended that classifications be based on consensus trees, but that the most parsimonious tree for the pooled data be used as the best estimate of the true phylogeny and as a guide to studying character evolution. This mixed strategy is intended to give due recognition to the fact that the tree obtained from the pooled data has "greater information content" and greater "global parsimony." Miyamoto (1985), while also assuming the safety (or stability) afforded by consensus methods, rejected their use in both phylogeny reconstruction and classification.

We disagree with the previously mentioned commentators in the following way. They assumed that consensus trees are conservative but noted some of the limitations that the strategy can engender. We argue that even this mixed assessment of the method is too generous. As others have observed, the consensus tree can be *different from* the tree for the combined data by virtue of being less resolved. What we show is that the consensus tree can positively *contradict* the most parsimonious tree obtained from the pooled data. The method of consensus is *not* a way to play it safe but involves committing to hypotheses that may not be sanctioned by (all) the data.

After describing consensus methods with more care, we present an example in which the consensus tree differs from the most parsimonious tree obtained from the

pooled data. Then we show that this problem is not limited to parsimony in particular or phylogenetic inference in general. We believe that in *all* inference problems, the best hypothesis is the one constructed in light of *all* of the data. Philosophers have called this *the principle of total evidence* (Carnap, 1950; Hempel, 1965; Good, 1983; also see Kluge, 1989).

CONSENSUS TECHNIQUES

There is a variety of consensus techniques, including strict (Sokal and Rohlf, 1981; Page, 1989), Adams (Adams, 1972), Nelson (Nelson, 1979; Page, 1989), and majority rule (Margush and McMorris, 1981). Of these, the most commonly used is strict consensus, mainly on the grounds that it is the most conservative. In contrast, Adams trees can contain components that are not present in any of the fundamental trees. Although this method is therefore often avoided, Funk (1985) and Hillis (1987) pointed out that it may be valuable in pinpointing taxa that are responsible for incongruence. Bremer (1990) observed that the Nelson consensus method also can contain components that contradict some of the fundamental trees, and (pers. comm.) that it assumes that there is a low probability that components will be replicated by chance alone. Majority rule has sometimes been used, both to find the best estimate of phylogeny and as a basis for studying character evolution (e.g., Jansen et al., 1991), but we know of no general justification for this approach.

In the case of comparing molecular and morphological results, the goal is often to obtain the most resolved tree that is compatible with both. This is very often at odds with strict consensus, which generates a tree that is least resolved. Hillis (1987) presented an example in which he combined trees for the genus *Rana*, choosing the more resolved solution wherever the fundamental trees were compatible but differed in resolution. A similar procedure was formalized by Bremer (1990) as combinable component consensus ("semistrict" consensus in PAUP, version 3.0; Swofford, 1990).
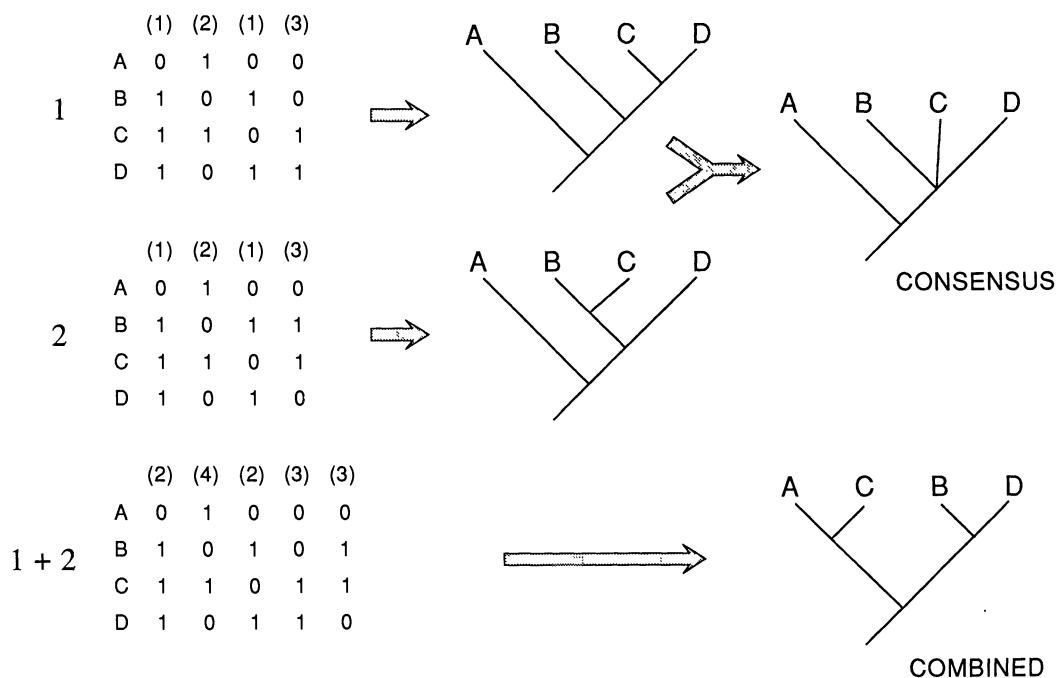
FIGURE 1. The consensus of the trees obtained from data sets 1 and 2 is incompatible with the tree obtained from the combined data set. See text for further explanation.

For our purposes, the main point is that these methods can give different results, but none will produce a result that is less resolved than, or incompatible with, the dictates of strict consensus. Consequently, of all the different methods, strict consensus is most likely to yield a result that is consistent with the tree produced from a combined data set. Thus, if we can show that a tree based on the combined data is incompatible with a strict consensus tree, then, in effect, we will have shown that *all* consensus methods have the same problem. As it happens, in the simple cladistic example presented below, all consensus techniques point to the same tree. It follows that all consensus methods can endorse trees that contradict the tree obtained from the pooled data.

A CLADISTIC EXAMPLE

It is not terribly difficult to concoct circumstances in which parsimony analysis of a pooled data set yields a tree that is positively at odds with the consensus based on the separately analyzed data sets. An example is shown in Figure 1, wherein Wagner analyses were carried out using the branch-and-bound option in PAUP (version 3.0; Swofford, 1990). In the two fundamental data sets, shown in the upper left, the same four taxa (A–D) are scored for seven characters. The number in parentheses above each column indicates the number of characters with that particular distribution of states. State 0 is assumed to be ancestral (plesiomorphic) and state 1 derived (apomorphic), and in the PAUP analyses trees were rooted by including a fifth taxon with the 0 state for each character.

The two fundamental data sets are identical except for the last pattern in each (i.e., the last three characters). In data set 1 the derived states are shared by C and D, whereas in data set 2 they are shared by C and B. Not surprisingly, data set 1 yields a tree in which C and D are united, whereas the tree derived from data set 2 unites C and B. In both cases the most parsimonious tree entails 10 steps, and uniting A

and C requires an additional step. The consensus of the two fundamental trees (using any of the methods discussed above) is shown in the upper right of Figure 1. Owing to the difference in the position of C, the B-C-D clade is collapsed to a trichotomy.

When the two data sets are combined and the resulting matrix is analyzed under Wagner parsimony, the tree in the lower right of Figure 1 is obtained. This requires 22 steps, whereas alternatives that unite either C and D or B and C require 1 more step. Note that the combined tree differs from the consensus tree in several ways. First, as in the example presented by Miyamoto (1985), the combined result is more parsimonious than the consensus. As Maddison (1989) pointed out, this comparison is not easy to make because there are different ways to optimize character-state changes on trees with polytomies. Nevertheless, any resolution of the B-C-D trichotomy requires at least 23 steps overall. And, if the trichotomy is interpreted as multiple speciation, as Miyamoto (1985) implicitly assumed in his example, a minimum of 26 steps is required. Second, and much more important for our purposes, the two topologies are positively at odds with one another regarding the position of taxon C. In fact, the one and only nontrivial component in the consensus tree (B-C-D) does not appear in the combined result.

Is this problem likely to arise in real cases? We have shown that it can occur, but we have not identified the general circumstances under which it is likely. However, the fact that such behavior can characterize small and rather simple data sets warns that it may also appear in complex real data. Furthermore, one can imagine real circumstances that would result in data sets like those in the example. For instance, in one set of data there may be convergence between two of the taxa (say, C and D evolved a set of traits related to a particular mode of pollination), whereas in a second set of data one of these two taxa has converged on a third (say, C and B evolved a set of nucleotides that enhance the function of a particular enzyme, or that

the rate of substitution was much increased in these lines, resulting in homoplasy). In each of the smaller data sets the characters that support the true relationship of C with A (let us assume) are outweighed by convergent characters. However, when the data are combined there is more support for the relation between C and A, and this arrangement wins out because C has converged on two different taxa and the homoplasy is therefore dispersed.

TWO STATISTICAL EXAMPLES

The example in the previous section raises the question of whether this flaw in consensus methods is peculiar to inference employing the principle of parsimony. Indeed it is not; nor is it restricted to phylogenetic inference in general. Consensus methods can be defined for extremely general statistical inference problems employing several data sets. In this section we show with two simple examples that even in this realm consensus methods can conflict with the principle of total evidence.

The first example is an urn model, which, despite its artificiality, exposes the pervasiveness of the problem. Imagine that the urn has been filled with six balls of different colors according to one of four distributions: (1) four white and two red balls, (2) two red and four green, (3) three white and three green, (4) three green and three blue. We consider the following two hypotheses:

W: There is at least one white ball in the urn.
R: There is at least one red ball in the urn.

Notice that each conjunction of W and R and their negations ($-W$ and $-R$) uniquely determines a single distribution. In other words, W&R is true if and only if distribution 1 obtains, and $-W$&R, W&$-R$, and $-W$&$-R$ correspond to distributions 2, 3, and 4, respectively. We suppose that our inference problem is to discover whether W is true and whether R is true.

We gather data by sampling with replacement and apply the method of max-

TABLE 1. Likelihood and consensus: an urn model.

| Joint hypotheses | Balls in urn by color | | | | Pr(D1/hyp) | Pr(D2/hyp) |
	w	r	g	b		
W&R	4	2			⅔	0
−W&R		2	4		0	⅔
W&−R	3		3		½	½
−W&−R			3	3	0	½

imum likelihood. Suppose our two data sets each consist of a single draw from the urn:

D1: A white ball is drawn.
D2: A green ball is drawn.

Based on D1, the likeliest joint hypothesis is W&R. That is, the conditional probability $Pr(D1/W\&R) = Pr(D1/distribution\ 1) =$ ⅔ exceeds the conditional probability for drawing white given any of the other three joint hypotheses. Based on D2, the likeliest joint hypothesis is −W&R, because $Pr(D2/−W\&R) = $ ⅔. All this is summarized in Table 1.

The data sets disagree on what the true joint hypothesis is. A consensus method would search for the points of agreement. In this example, both data sets agree that R is true while disagreeing about whether W is true. So consensus recommends that we make the "safe" conclusion that R is true. However, if we pool the data (D1 and D2), the likeliest joint hypothesis becomes W&−R. Indeed, this is the *only* joint hypothesis compatible with the data! Using all the evidence thus contradicts the "safe" conclusion.

The second example begins with polling data for preferences in a presidential election. Suppose that in four geographic regions of the country (Northwest, Northeast, Southwest, Southeast), preferences for a presidential candidate B and a vice-presidential candidate Q are ascertained. The distributions are shown in Table 2 (we assume that those who do not prefer B prefer the opponent, and the same for Q).

Suppose a bus is stopped somewhere and the riders are quizzed. The problem is to infer where the people on the bus came from, assuming that they all came from the same region. The two data sets consist of a sample of preferences from the people on the bus. The results are

D1: 56% prefer B to the opponent.
D2: 55% prefer Q to the opponent.

From which of the four regions should we infer that the people come? We assume that the data are binomially distributed and, for simplicity, that there are no correlations between presidential and vice-presidential preferences. We do not actually have to calculate the numerical values of the likelihoods of the joint hypotheses NW, NE, SW, and SE; instead, we can simply exploit the shape of the binomial distribution to select the hypothesis of maximum likelihood. Based on D1, we judge that NW is correct. Based on D2, we judge that SW is correct. The consensus conclusion is that the people come from the West. But based on the pooled data sets D1 and D2, SE is the likeliest hypothesis. Once again, using all the evidence leads to a hypothesis that contradicts the consensus.

CONCLUDING REMARKS

It is important to realize that consensus methods have been used to solve quite different problems. A correct assessment of those methods will be impossible, unless the different uses of consensus are disentangled. We already have mentioned two:

TABLE 2. Likelihood and consensus: election preferences.

	W	E
N	B: 56%	B: 45%
	Q: 37%	Q: 8%
S	B: 80%	B: 58%
	Q: 55%	Q: 51%

(1) given *a single data set*, find the points of consensus among various trees that are equally parsimonious, or nearly so; and (2) given *two or more data sets* that could be combined, find the points of consensus between the most parsimonious tree(s) constructed for each. Our objections to consensus pertain to the second problem, not to the first. If the consensus tree disagrees with the best tree inferred from the pooled data, the consensus tree cannot be considered to be the best inference. In this respect, we agree with Kluge (1989) and not with Carpenter (1988), who suggested that consensus is appropriate for different data sets but not for trees from a single data set.

It is worth emphasizing that the problem we have described here cannot arise when consensus methods are used to find points of agreement among multiple trees from the same data set (use 1, above). Indeed, far from wishing to criticize the use of consensus methods in this context, we believe that there is much to be said for them. If the most parsimonious tree, relative to all the data, requires 50 changes in character state and the next most parsimonious tree requires 51, one may wish to take seriously only those clades endorsed by both trees. Conservatism has a role to play in such cases, one that may be well served by the method of strict consensus.

Uses 1 and 2 of consensus must, in turn, be distinguished from several others: (3) Given that two data sets might disagree with each other (e.g., organellar and nuclear data), how much do they disagree? Here the problem is not to infer a single phylogeny, but to assess how much agreement there is between two types of data or between different gene trees. (4) How do the results differ when different methods of analysis (e.g., phenetic versus cladistic) are applied to the same data set? (5) How do the results obtained for different data sets compare when those data sets cannot in principle be combined (e.g., DNA–DNA hybridization and discrete characters interpreted by parsimony)?

Our point does not concern uses 3–5 of consensus methods. Rather, in discussing uses of type 2, we have shown that the consensus tree can contradict the tree inferred from the pooled data; when this is so, we believe that the consensus tree cannot be regarded as the best (or "safest") inference to make from the available data. In this, as in other inference problems, it is appropriate to abide by a principle of total evidence. Our suggestion, then, is that a "consistency check" be performed on consensus trees. This simply means checking whether the consensus tree is consistent with the best tree based on the pooled data, which in most cases is easily accomplished.

The same general conclusion has been endorsed by several other authors, notably Miyamoto (1985), Cracraft and Mindell (1989), Kluge (1989), and Donoghue and Sanderson (1992). Donoghue et al. (1989) made a similar argument with regard to simultaneous analysis of all relevant taxa, including fossils. However, we believe that the argument presented here is far more compelling than previous criticisms. For example, the fact that a consensus tree lacks resolution does not, in itself, impugn that method as a strategy for playing it safe. The same point applies to the use of consensus as a guide to character evolution; its implausibility in this context does not show that it fails to deliver a safe estimate of the tree topology.

We briefly discussed the question of how often a consensus tree will disagree with the tree inferred from the pooled data. Although we are not able to provide a precise answer to this question, our sense is that the circumstance may not be rare. We wish to emphasize, though, that it is foolhardy to assume without proof or without checking that what occurs in our simple cladistic example cannot arise in real data sets.

We suspect that some of the attraction of consensus methods has arisen in the following way. Morphological and molecular characters are often thought to be "independent" sources of evidence, and the idea has taken hold that their independence is best acknowledged by keeping them separate. We believe, however, that the independence of two pieces of data is not a reason for keeping them apart. When char-

acters are combined, it is *desirable* that they be independent of each other; in this way, their collective testimony will be stronger than what each says on its own. If independent morphological (or molecular) characters are to be combined with each other, it is hard to see why the "independence" of morphological and molecular characters is a reason for keeping them apart.

Many systematists suspect that when morphological and molecular characters are both available, agreement between these data sets is more significant than agreement within them. Perhaps morphological characters are often correlated with one another and molecular characters may be correlated with each other as well. When this is true, it is important that the inferred tree accommodate both character sets, not simply one of them or the other. We are not denying this judgment; rather, our point concerns how the judgment should be implemented. If *A* and *B* are highly correlated characters, whereas *C* evolved independently of *A* and *B*, then a tree based on all the data should be constructed with characters weighted to reflect such judgments of correlation and independence.

Systematists of various persuasions have been uncomfortable about combining morphological and molecular data for a second reason. If the data sets are kept separate, it is no great strain on one's credulity to accord equal weight to the characters within each data set. But this arrangement is more difficult to defend if the sets are combined. Equal weighting across all characters runs the risk of allowing the signal present in one data set to blot out the signal present in the other. This weighting scheme seems implausible, but it is hard to defend any particular scheme that assigns unequal weights. Better, then, to keep the data sets separate and to construct consensus trees.

This rationale is much less attractive than it first appears. We have shown that consensus trees do not always play it safe. What is more, consensus methods effectively induce a weighting between characters of different types because they must assume

some definite relationship between the separate data sets. If weighting is a problem, so too is any consensus method. In response to the fear that combining data sets will allow molecular characters to swamp morphological characters, it is worth observing that the sheer number of characters of different types is not as important in determining tree topology as the distribution of character support and homoplasy (Donoghue and Sanderson, 1992). The addition of even a small number of characters might change a topology based on a much larger data set.

Molecular evolutionists sometimes keep molecular and morphological characters "separate" by ignoring the latter altogether. Ignoring morphological data would be justified if such characters provided no evidence whatever about phylogenetic relationships or if such data were uninterpretable. We believe that neither of these conditions is satisfied. The principle of total evidence requires molecular biologists to take morphology into account or to provide an argument that shows why such data are devoid of evidential meaning.

Combining heterogeneous data is not a problem that suddenly arose when morphological and molecular characters confronted each other. It has been with us all along. Molecular characters can differ from each other in terms of their lability (e.g., different sites within codons or in transcribed versus nontranscribed regions), and the same is true of morphological characters. If combining the two different sorts of characters seems to involve an apples-and-oranges comparison, the same may be said of the heterogeneous data sets that systematists have always had to consider.

If molecular and morphological characters each provide evidence when taken separately, then the two types of evidence should be pooled and some weighting scheme defended. It is worth emphasizing that assigning equal weights involves biological assumptions just as much as assigning unequal weights (Sober, 1988). Systematists must defend whatever scheme they adopt by presenting a substantive argument (see, for example, a defense of equal

weighting in Donoghue [1989]). If consensus methods seem to allow us to avoid addressing such difficult issues, this is one good reason to be against consensus.

REFERENCES

ADAMS, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21:390–397.

BREMER, K. 1990. Combinable component consensus. Cladistics 6:369–372.

CARNAP, R. 1950. Logical foundations of probability. Univ. Chicago Press, Chicago.

CARPENTER, J. M. 1988. Choosing among multiple equally parsimonious cladograms. Cladistics 4:291–296.

CRACRAFT, J., AND D. P. MINDELL. 1989. The early history of modern birds: A comparison of molecular and morphological evidence. Pages 389–403 in The hierarchy of life (B. Fernholm, K. Bremer, and H. Jörnvall, eds.). Elsevier, Amsterdam.

DONOGHUE, M. J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. Evolution 43:1137–1156.

DONOGHUE, M. J., J. A. DOYLE, J. GAUTHIER, A. G. KLUGE, AND T. ROWE. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

DONOGHUE, M. J., AND M. J. SANDERSON. 1992. The suitability of molecular and morphological evidence in reconstructing plant phylogeny. Pages 340–368 in Molecular systematics in plants (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

FUNK, V. A. 1985. Phylogenetic patterns and hybridization. Ann. Mo. Bot. Gard. 72:681–715.

GOOD, I. J. 1983. Good thinking: The foundations of probability and its applications. Univ. Minnesota Press, Minneapolis.

HEMPEL, C. G. 1965. Aspects of scientific explanation. Free Press, New York.

HILLIS, D. M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18:23–42.

JANSEN, R. K., H. J. MICHAELS, AND J. D. PALMER. 1991. Phylogeny and character evolution in the Asteraceae based on chloroplast DNA restriction site mapping. Syst. Bot. 16:98–115.

KLUGE, A. G. 1983. Cladistics and the classification of the great apes. Pages 151–177 in New interpretations of ape and human ancestry (R. L. Ciochon and R. S. Corruccini, eds.). Plenum Press, New York.

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

MADDISON, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

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

MIYAMOTO, M. M. 1985. Consensus cladograms and general classifications. Cladistics 1:186–189.

NELSON, G. 1979. Cladistic analysis and synthesis: Principles and definitions, with historical notes on Adanson's Familles des Plantes (1763–1764). Syst. Zool. 28:1–21.

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

SOBER, E. 1988. Reconstructing the past: Parsimony, evolution and inference. MIT Press, Cambridge, Massachusetts.

SOKAL, R. R., AND F. J. ROHLF. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. Syst. Zool. 30:309–325.

SWOFFORD, D. L. 1990. PAUP, phylogenetic analysis using parsimony, version 3.0. Distributed by the author, Illinois Natural History Survey, Champaign.