

# Points of View

*Syst. Biol.* 52(6):849–851, 2003  
Copyright © Society of Systematic Biologists  
ISSN: 1063-5157 print / 1076-836X online  
DOI: 10.1080/10635150390252251

## Branch Lengths and Support: Revisited

DAVID A. MORRISON

*Department of Parasitology, National Veterinary Institute and Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden;  
E-mail: david.morrison@vmm.slu.se*

Wilkinson et al. (2003) took issue with Farris et al. (2001) over the concept of whether the branch lengths of phylogenetic trees can be used to indicate support for the component clades. I do not wish to enter into this particular debate here, but instead I point out that some of the details discussed by Wilkinson et al. are apparently based on a misinterpretation of the information provided by Farris et al. In particular, two pieces of information presented by Farris et al. are apparently ambiguous and have led Wilkinson et al. to some incorrect argumentation. This does not necessarily change the substantive conclusions reached by Wilkinson et al., but it does change the details of their argument.

Figure 1 of Farris et al. (2001) shows part of an artificial DNA sequence data matrix (for taxa labeled A–H), along with two optimal phylogenetic trees derived from analysis of the complete data matrix (which contains 60 repetitions of the small matrix shown). The data analysis was performed using the default settings of the DNAML program of Felsenstein (1993), and thus the trees have maximum likelihood under the chosen evolutionary model. Farris et al. (2001:298) claimed that the consensus tree for the data is “unresolved,” whereas Wilkinson et al. (2003) pointed out that the strict component consensus of the two trees shown in figure 1 of Farris et al. does in fact have one informative component. Interestingly, both parties are making statements that are uncontradicted by the information provided by Farris et al. This arises because Wilkinson et al. based their discussion solely on the trees presented by Farris et al., and these trees do not accurately reflect the data matrix.

First, the trees as shown by Farris et al. (2001) may be superficially interpreted as being fully resolved binary trees. However, this interpretation is incorrect because not all of the internodes shown on the trees represent evolutionary branches. In the text, Farris et al. (2001:298) explicitly stated, “With the exception of the ABEF/CDGH split, however, all the interior branches of both trees have the same length, about 0.16.” Note that four of the internodes shown on the trees are referred to as “branches” with a specified branch length, whereas the fifth internode is referred to as a “split” with no specified branch length. Inspection of the data matrix

reveals that this fifth internode represents a split (or bipartition) that is neither supported nor contradicted by the sequence data. In other words, if the internode is interpreted as a branch, then it has zero branch length. It is, perhaps, worth noting that the DNAML program actually reports this internode as having a length of 0.00006, but the likelihood-ratio test makes it clear that this estimated length is not significantly different from zero.

Therefore, it would be more accurate to represent this part (of both trees) as a four-way polytomy; the chosen presentation is ambiguous because the trees as shown involve an apparently arbitrary resolution of a polytomy. While phylogenetic trees showing branches with zero length are not unknown in the literature (e.g., fig. 5 of Kluge and Farris, 1969, has two such branches), these unsupported resolutions of multifurcations have been severely criticized in the context of searching for and presenting phylogenetic trees (e.g., Nixon and Carpenter, 1996; Farris and Källersjö, 1998) — quite literally in this case, apparent branch lengths (on the diagram) do not indicate real support (in the data).

Second, from the context it is possible to interpret the two trees presented by Farris et al. (2001) as being the only two optimal trees that can be derived from the data matrix because Farris et al. (2001:298) stated that “this matrix has two different maximum-likelihood trees, as shown.” In fact, the matrix has four different trees with the same likelihood under this particular evolutionary model, as revealed by analysis of the data using the PAUP\* program (Swofford, 2002). Use of this program obviates the methodological difficulties referred to by Farris et al. (2001:298) because the data matrix is small enough to be analyzed using the exhaustive search option (there is no branch-and-bound procedure known for maximum likelihood, and PAUP\* defaults to performing an exhaustive search if this option is chosen). That all four of these trees are optimal was independently confirmed by inputting them to DNAML as user trees under the default settings (the strategy used by Farris et al.).

The four maximum-likelihood trees are shown here in Figure 1. There are only eight splits (or bipartitions)

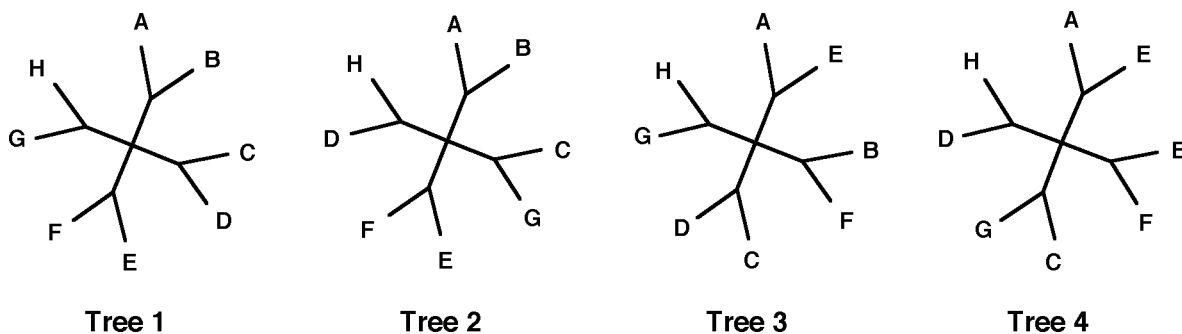


FIGURE 1. The four optimal trees found by maximum likelihood analysis of the data matrix discussed by Farris et al. (2001). Trees 1 and 4 are the two trees presented by Farris et al. The trees are shown here as unrooted. Farris et al. did not discuss the rooting of their trees, but the trees as shown in their figure could be interpreted as being rooted on the branch leading to taxon A.

that have support in the data matrix, and these form two nonoverlapping subsets of the taxa consisting of two alternative pairings for each taxon. These pairings can each be combined in two noncontradictory ways between the two subsets, thus forming the four trees, as shown in Table 1. As an aside, these four trees are the optimal trees regardless of what size of data matrix is used for the maximum likelihood analysis. Changing the matrix size changes the estimated likelihood value and the estimated branch lengths, but the same four tree topologies result from analysis using the small matrix shown in figure 1 of Farris et al. (2001) or the larger one that is referred to in their text. This is to be expected because the repetition of the data does not change the relative weights of the 16 distinct data patterns present in the matrix.

The consensus tree of these four optimal trees is a polytomy "bush." This topology is true regardless of what form of consensus tree is used (e.g., Wilkinson, 1994; Nixon and Carpenter, 1996), because the contradictory components revealed by the four optimal unrooted trees lead to the same consensus tree under all consensus methods. Thus, the consensus tree derived from the data matrix is unresolved, as claimed by Farris et al. (2001), but the consensus of the two trees shown in their figure 1 does have a resolved component, as claimed by Wilkinson et al. (2003). Note that an unresolved consensus is not true for nestings if the trees are treated as rooted; e.g., if

the trees are rooted on taxon A, then the Adams consensus tree is  $(A(EB(CDFGH)))$ .

As a final point, it is worth noting that the number of optimal trees found by PAUP\* for this data set may depend to some extent on the evolutionary model chosen. Different evolutionary models (e.g., differing in how variation in the base frequencies, nucleotide substitutions, and among-site rate variation is modeled) are known to produce different likelihood estimates for any specified data set (e.g., Posada and Crandall, 2001), so it is important to choose an appropriate model when performing maximum likelihood analysis. The model used by the DNAML program is known in PAUP\* as the F84 model (Kishino and Hasegawa, 1989), although the default values do not involve maximum likelihood estimation of the model parameters but instead use both a transition:transversion ratio of 2.0 and the empirical base frequencies as heuristic estimates. However, the same four trees are also found as the optimal trees using any evolutionary model that assumes either equal base frequencies or empirical base frequencies (which are equivalent in this case, given the artificially constructed nature of the data matrix). Changing the model changes the estimated likelihood value for the trees and changes the estimated branch lengths, but the same four tree topologies result from each analysis.

Choosing an evolutionary model in an arbitrary fashion is usually not recommended for maximum likelihood analysis, particularly "using the default models implemented in standard computer programs for phylogenetic estimation" (Posada and Crandall, 2001:580). Inspection of the Farris et al. (2001) data matrix shows that the following expectations should apply to any model used to analyze the data, as a direct result of the artificial patterns that have been built into the sequences: equal base frequencies (all bases are represented with exactly equal frequency), unequal transition:transversion ratio (all 16 of the data patterns involve transversions), and equal transversion rates (all four possible transversions are represented by four symmetrical data patterns each). This approach matches the K80 (or K2P) model (Kimura, 1980). Maximum likelihood estimation should be used

TABLE 1. Taxon splits (or bipartitions) that are supported by the data matrix discussed by Farris et al. (2001), along with the combinations in which they occur in the four optimal trees resulting from maximum likelihood analysis of the data (as shown in Fig. 1).

Split	Tree 1	Tree 2	Tree 3	Tree 4
{A,B}{C,D,E,F,G,H}	*	*		
{E,F}{A,B,C,D,G,H}	*	*		
{A,E}{B,C,D,F,G,H}			*	*
{B,F}{A,C,D,E,G,H}			*	*
{C,D}{A,B,E,F,G,H}	*		*	
{G,H}{A,B,C,D,E,F}	*		*	
{C,G}{A,B,D,E,F,H}		*		*
{D,H}{A,B,C,E,F,G}		*		*

for the transition:transversion ratio because this value is unlikely to be 2.0.

#### ACKNOWLEDGMENTS

I thank Mike Steel and Mark Wilkinson for helpful comments on the manuscript.

#### REFERENCES

- FARRIS, J. S., AND M. KÄLLERSJÖ. 1998. Asymmetry and explanations. *Cladistics* 14:159–166.
- FARRIS, J. S., M. KÄLLERSJÖ, AND J. E. DE LAET. 2001. Branch lengths do not indicate support—Even in maximum likelihood. *Cladistics* 17:298–299.
- FELSENSTEIN, J. 1993. PHYLIP: Phylogeny inference package, version 3.5c. Documentation and program. Department of Genetics, Univ. Washington, Seattle.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–120.
- KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29:170–179.
- KLUGE, A. G., AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18:1–32.
- NIXON, K. C., AND J. M. CARPENTER. 1996. On consensus, collapsibility, and clade concordance. *Cladistics* 12:305–321.
- POSADA, D., AND K. A. CRANDALL. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50:580–601.
- SWOFFORD, D. L. 2002. PAUP\*: Phylogenetic analysis using parsimony (\*and other methods), version 4.0b10. Sinauer, Sunderland, Massachusetts.
- 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., F.-J. LAPOINTE, AND D. J. GOWER. 2003. Branch lengths and support. *Syst. Biol.* 52:127–130.

*First submitted 23 April 2003; reviews returned 11 June 2003;  
final acceptance 28 July 2003  
Associate Editor: Mike Steel*

*Syst. Biol.* 52(6):851–852, 2003  
Copyright © Society of Systematic Biologists  
ISSN: 1063-5157 print / 1076-836X online  
DOI: 10.1080/10635150390252260

## Phylogenetic Methods and Aetosaur Interrelationships: A Rejoinder

SIMON R. HARRIS,<sup>1,2</sup> DAVID J. GOWER,<sup>2</sup> AND MARK WILKINSON<sup>2</sup>

<sup>1</sup>*Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, U.K.; E-mail: simon.harris@bristol.ac.uk*

<sup>2</sup>*Department of Zoology, The Natural History Museum, London SW7 5BD, U.K.*

In a previous paper (Harris et al., 2003), we discussed the treatment of intraorganismal homology in character construction. Our aims were to highlight alternative approaches and to investigate their theoretical foundations and analytical consequences, and we used the phylogeny of aetosaurian reptiles as an example. In a previous study, Heckert and Lucas (1999) employed several characters describing variation in aetosaur osteoderms. Harris et al. (2003) noted that parallel variations (e.g., presence of either radial or random patterning) in the osteoderms of different body regions were represented as separate but covarying characters, implying independent evolutionary changes in different regions. We suggested that osteoderms are intraorganismal homologues and that a single change affecting multiple osteoderm regions, represented by a single composite character, was a plausible alternative to the previous more reductive interpretation. We demonstrated that alternative reductive and composite approaches to constructing characters from variation in osteoderms impacts inferred relationships and their apparent levels of support.

In response, Heckert and Lucas (2003:253) dismissed our interpretation of osteoderms as intraorganismal homologues, arguing that we made “an unverifiable assumption about underlying genetic control” and that we ignored known intraorganismal variation in osteoderms. In fact, we explicitly discussed (2003:244) such

variation. Importantly, we argued that osteoderms are intraorganismally homologous, not that they are intraorganismally homogenous. Heckert and Lucas’s (2003) preferred, more reductive, coding also entails assumptions that they did not discuss. For example, their use of separate characters to represent parallel variation in cervical and dorsal osteoderms assumes homology within, but nonhomology between, these regions. The latter entails that any changes occurring in both regions are coincidental and homoplastic. As we pointed out, this assumption contravenes Hennig’s auxiliary principle (Hennig, 1966). Heckert and Lucas (2003) did not discuss the theoretical issues we raised and dismissed our work as teaching us “not much” about aetosaur phylogeny.

Heckert and Lucas (2003) argued that the alternative character constructions make no difference to inferred relationships, yielding trees that are “remarkably similar,” and claimed (2003:253) that the tree obtained with our more composite characters “is still the same as that published by us (compare Harris et al., 2003: fig. 2a with Heckert and Lucas, 1999: fig. 9).” This assertion is incorrect. The cited trees are not the same—they differ in the placement of three of eight ingroup taxa (Fig. 1). Furthermore, the cited tree (Harris et al., 2003: fig. 2a) was based on a reductive, not a composite coding. In fact, composite and reductive approaches yielded substantially