



## PERMUTATIONS

James S. Farris,<sup>1,2</sup> Mari Källersjö,<sup>1</sup> Arnold G. Kluge<sup>3</sup> and Carol Bult<sup>4</sup>

<sup>1</sup> *Naturhistoriska riksmuseet, Molekylärsystematiska laboratoriet, Box 50007, S-104 05, Stockholm, Sweden,* <sup>2</sup> *Department of Entomology, American Museum of Natural History, Central Park West at 79th St, New York, New York 10024, U.S.A.,* <sup>3</sup> *Department of Reptiles and Amphibians, Museum of Zoology, The University of Michigan, Ann Arbor, Michigan 48109, U.S.A. and* <sup>4</sup> *Institute for Genomic Research, 932 Clopper Rd, Gaithersburg, Maryland 20878, U.S.A.*

*Received for publication 28 October 1993; accepted 13 March 1994*

### Introduction

In an earlier paper (Källersjö et al., 1992) we called attention to a previously unreported characteristic of Archie's (1989) "randomization test for phylogenetic information in systematic data". That test may indicate significance in data too ambiguous to yield any resolved grouping. To assess unambiguous resolution we therefore introduced a new randomization test, the total support test.

Faith and Cranston (1991) had proposed Archie's test under the new name PTP, and so Faith and Ballard (1994) now protest that our comments on the Archie/Faith/Cranston (AFC) technique were unfair: that test properly indicates "overall" hierarchic structure, structure "in a PTP sense". In their view, furthermore, our test is both markedly inferior to and "nearly identical to the 'all-groups' form of the T-PTP test", which they promise to "describe in detail".

Which they do not, nor do they perform either total support or all-groups (AG) tests, and attending to these details leads to rather different conclusions. Nor do Faith and Ballard manage to describe any kind of hierarchic information that conforms to the AFC method, as will be seen here.

### Reduction

Faith and Ballard start their discussion of AFC by presenting ours backwards:

"Källersjö et al. claim that if their TS [total support] test suggests no significant structure (as for their data set "Three"; see our Table 1), but the PTP [AFC] test nevertheless implies strong structure, then the PTP test must be flawed".

In fact, we discovered the behavior of AFC<sup>1</sup>, then sought a new method. Although highly significant by AFC, Three yields 10 distinct most-parsimonious trees, the consensus of which is entirely unresolved. Our conclusion was based directly on the demonstrable inadequacy of AFC, not on comparing AFC to other procedures.

According to Faith and Ballard, however, we had missed the true meaning of AFC:

"Källersjö et al. interpret this as indicating a fault in PTP [AFC] testing, but this data set [Three] properly does have structure in a PTP sense. There is no one definitive topology among the 10 equally most-parsimonious trees, but these trees nevertheless definitely delimit a family of similar

<sup>1</sup> PTP (permutation tail probability) would describe any of the tests discussed here.

topological relationships among the taxa (Fig. 3 shows the first two MPTs [most parsimonious trees]). The fact that there might be hierarchical information here should not be surprising; data set "Three" has reduced the number of considered trees from 2 027 015 down to 10.<sup>2</sup>

As Three has 10 terminals, the number of possible bifurcating trees is 2 027 025 when the outgroup is fixed.<sup>2</sup>

Those 10 trees might be considered topologically similar in that all are pectinate. But they show different relationships among the taxa; the two that Faith and Ballard figure have no informative groups in common. Evidently, the "structure in a PTP sense" or "hierarchical information" present amounts to the reduction in the number of trees. This seems indeed to be Faith and Ballard's intention, for they explain earlier:

"Suppose we have character data for 4 taxa, and that only 2 of the 3 possible quartets [resolved undirected trees] are implied by these characters. There is no resolution in a strict consensus tree (when half the characters support each quartet), but this data set nevertheless has hierarchical information—it excludes the third quartet".

Broad though it is, this concept of information still provides no defense of the AFC test, since reduction in the number of trees is not generally related to AFC significance. Our 1992 matrix Two (p. 276) has 10 terminals but yields only two most-parsimonious trees. Yet it is nowhere near significance by AFC, the tail probability being approximately 0.47.

Neither of those two trees is fully resolved, and Faith and Ballard might insist on counting bifurcating resolutions, of which there are 30 for Two. In that case, consider our 1992 matrix One (p. 276), which also has 10 terminals. One is very highly significant by AFC, but has 11 025 bifurcating most-parsimonious trees.

Faith and Ballard's paper furnishes more examples. Their matrices X14 and X00 each have six terminals. X14 is far from significance by AFC, the tail probability being 0.7983. It has just one (fully resolved) most-parsimonious tree. X00 has significant AFC tail probability 0.0364, but two distinct most parsimonious trees<sup>3</sup>.

Cases such as Three will also be of interest, but Three itself is inconveniently large for some of the calculations used later. Some of its characteristics are repeated on a smaller scale by matrix Four.

```

1 1 1 1 0 0 0 0
0 1 1 1 1 0 0 0
0 0 1 1 1 1 0 0
0 0 0 1 1 1 1 0

```

Columns are terminals 0–7 in order; rows are characters. To incorporate duplication of characters (as in the original Three), use matrix Dfour. This has eight characters, two each of those tabulated here. Much like Three, either Four or Dfour yields eight distinct most-parsimonious trees, the consensus of which is unresolved.

This provides a last example concerning Faith and Ballard's tree-counting con-

<sup>2</sup> Throughout, a "bifurcating" ("fully resolved") tree is allowed a basal trifurcation where the outgroup joins.

<sup>3</sup> Faith and Ballard give those tail probabilities as 0.83 and 0.08, respectively; the latter misleadingly suggesting poor significance. The inaccuracy is attributable to sampling error. We use 9999 randomizations with exact calculations of minimum lengths and support values throughout. They apparently use only 99 randomizations, probably because their method of computation is relatively laborious. It involves reading each randomization into PAUP as a separate matrix.

cept of information. Dfour has strongly significant AFC tail probability 0.0194. But Four has just the same suite of most-parsimonious trees, while its AFC tail probability is quite non-significant at 0.6160. Again there is no connection between AFC significance and reduction of the number of trees.

### Tests

Faith and Ballard base a further defense of AFC on the premise that Three provides significant evidence against monophyly of group  $(a + c + e + g)$ .

"The T-PTP test for non-monophyly of this group based on data set "Three" is highly significant (length difference equals  $28 - 18 = 10$ ; T-PTP = 0.01 based on an *a priori* test, Faith, 1991). It would be incorrect therefore to claim that "Three" has no significant hierarchical information or structure. Similar results of course could be found in tests on a number of other taxon groups; the original PTP [AFC] test on these data has provided a correct indication of overall hierarchical structure".

They go on to stress that non-monophyly constitutes hierarchic information, but this actually adds nothing to their argument. Specifying that a certain group is not monophyletic (or that it is, for that matter) rules out some otherwise possible trees, so that this idea of information fits into the earlier one. The novel element is the T-PTP test. AFC is now to be seen as correctly indicating "overall" structure because the *a priori* T-PTP agrees with the AFC assessment of significance.

Such use of *a priori* T-PTP tests to demonstrate structure is more problematic than Faith and Ballard suppose, as will become apparent later. That aside, this attempt to justify AFC has the same weakness seen before, that AFC does not generally agree with the proposed criterion. Performing an *a priori* T-PTP on the marsupial RNA sequence data of Thomas et al. (1989), Faith (1991) found significant support for monophyly of (*Thalacinus* + *Dasyuridae*). But, as he pointed out, those data are not significant by AFC.

For a further example we use an *a priori* AG test, which (as is discussed below) is a T-PTP for a whole tree, rather than a single group. When applied to matrix Four and tree  $[0\ 7\ 6\ 5\ 4\ 3\ 2\ 1]^4$ , this test yields strongly significant tail probability 0.0161. By AFC, again, Four is thoroughly non-significant at 0.6160. The indication from AFC is once more incorrect.

As Faith and Ballard present the matter, our 1992 discussion rested on mistaking the purpose of AFC:

"While that random-covariation null model was originally used in a test [AFC] for overall data structure (regardless of topology), Källersjö et al. were interested in tests relating to support for tree topology".

By which they mean our total support test, whereas:

The PTP [AFC] test should be viewed as quite distinct from such topology-dependent tests—PTP provides a general indicator of the strength of hierarchical information.

To give their story the needed substance, Faith and Ballard try to come up with a concept of hierarchic information that corresponds to AFC. They do not succeed; AFC conforms to their criteria—tree counts and T-PTP—in the one case that they discuss, but not in other cases. Even when judged by Faith and Ballard's measures of structure, the AFC test seems to serve no purpose.

<sup>4</sup> The tree notation is as supported by Hennig86.

But then, if Faith and Cranston (1991) had already worked out a test for “overall” structure, why do Faith and Ballard now have such trouble describing that quantity? Why not just quote the earlier paper—as distinguished from citing it?

“The PTP [AFC] test in contrast [to total support] does not focus on particular topological hypotheses (Faith and Cranston, 1991). Indeed, it is possible to have strong cladistic structure indicated by PTP and not have one definitive topology implied by the data”.

That happens in Three, but not in any example of theirs. Faith and Ballard offer no quotation now because the necessary passage was never there. Not that Faith and Cranston (1991) described a concept like ours, one of unambiguous structure. But neither did they insist on ambiguity; they simply did not consider the distinction.

Faith and Cranston (1991) can hardly be criticized for failing to take a clear stand on an issue that did not arise until we found examples like Three. Faith and Ballard, however, seem subject to criticism on other grounds. Their last-quoted comment gives the impression that they knew all along what was in fact discovered only recently.

### Structure

There is much the same flavor to Faith and Ballard’s portrayal of our method:

“[Källersjö et al.’s] proposed test (“total support”, TS) is nearly identical to the “all-groups” form of the T-PTP test referred to above”.

Their claim may be regarded with some skepticism, not least because they evidently do not know how our test is performed. They suppose, for instance:

“This same example reveals another aspect of the TS test’s behaviour. Note that the MPT for set [matrix] X14 (Fig. 2a) has a high TS [value] of 3, but this same tree topology has a TS value of 0 for data set X00”.

The second “TS” is actually 2<sup>5</sup>, not 0, but that hardly matters. Their subsequent criticism of those values, even if correct in itself, would not bear on an “aspect of the TS test’s behaviour” because no such comparison is employed in our test.

In our method, data are considered significantly structured if the total support from the real matrix exceeds that from a great-enough proportion of randomized matrices. For any one matrix, real or randomized, the total support value used is that of the most-parsimonious tree(s) for *that same matrix*. The second “TS” that Faith and Ballard refer to here is evaluated on matrix X00, but the tree used is not most parsimonious for that matrix. That “TS” value would thus never arise in our procedure.

The comparisons used in our method are chosen according to its purpose, which is to assess strength of hierarchic structure in data. The reason for our choice of comparisons can be illustrated in the simplest case, using undirected trees for four terminals, a, b, c, d. Say that matrix X has 99 abs (binary characters uniting a with b) and 1 ac, while matrix Y has the opposite configuration. A third matrix Z has 51 abs and 49 acs. Each matrix is a possible randomization of the others.

For X, tree (ab)(cd) has total support 98, while for Y (ac)(bd) has that value, and for Z (ab)(cd) has total support 2. Although they give different trees, X and Y are

<sup>5</sup> We thank Dr K. Rognes for calling this to our attention.

both strongly structured, much more so than Z. Of course other comparisons are conceivable; a given tree could be evaluated on different matrices. Z could be said to support (ab)(cd) more strongly than does Y. But that does not reflect the strength of structure in either matrix, since Y and Z yield different trees.

The comparison in Faith and Ballard's remark is of that latter kind. While it does not occur in our procedure, that type of comparison is employed in some AG tests, as will be discussed presently. What grounds Faith and Ballard might have for attributing that approach to us is not clear; certainly they quote no such prescription from our paper. At most they offer that we "were interested in tests relating to support for tree topology" rather than strength of structure.

But that characterization itself is Faith and Ballard's, not ours, and, further, it closely resembles their way of describing their AG test. This mistaken objection to our method, then, seems to arise just from their notion that our test is "nearly identical" to theirs.

### Strength

Most of Faith and Ballard's comments, however, have little to do with confused identity, but simply ascribe imagined faults to our method. First, they think that our test may be too lenient in judging structure:

"Källersjö et al. claim that TS is a "stronger" test. . . . A strong test is one that is able to detect correctly a data set/tree that does not show good support. . . . Here, we have shown that Källersjö et al.'s claim of TS test strength in any case requires further evaluation. In our earlier example, TS gave a high support value for a tree (the MPT for X14) having high homoplasy".

Attributing that use of "stronger" to us is merely fanciful, but the criticism itself is no more substantial. They cite no other such examples, and for X14 they report only the total support value, not the result of a total support test. For X14 we find a total support tail probability of 0.2998, quite far from both significance and the impression that Faith and Ballard try to create.

Not satisfied with claiming that our test finds structure too readily, Faith and Ballard also suggest the opposite.

"But [strength] reflects only one important evaluation criterion. A good test method naturally must also be able to detect a data set that does have good structure, else one could trivially construct a "strong" test by making it always fail to reject the null hypothesis!".

Naturally, they do not mention that we found highly significant total support in matrix One and in Hamby's (1990) rRNA sequence data<sup>6</sup>. Nor do they perform any total support tests themselves.

The idea that "a good test method" should "detect a data set that does have good structure" could also refer to one of the arguments quoted earlier. Faith and Ballard take a significant *a priori* T-PTP to show "significant hierarchical information" in Three, which lacks significant total support. They feel that this illustrates a weakness of our approach, but another example will instead reveal a peculiarity of theirs.

For their example Faith and Ballard select group (a + c + e + g) so as to yield significance in an *a priori* T-PTP test of Three. It is not obvious that this conforms to

<sup>6</sup> Of 30 real data matrices used by A. Mooers (pers. comm.), 20 show significant total support at the 5% level.

Faith's (1991) distinction, discussed below, between *a priori* and *a posteriori* tests. But in any case they use that method now, and so to illustrate its pitfalls we shall do the same in selecting trees for a *a priori* AG T-PTP tests of Four.

The consensus of its most-parsimonious trees being unresolved, Four is like Three in lacking significant total support. According to our test, Four has no unambiguous structure. But as was seen above, according to the *a priori* AG T-PTP test, Four shows strongly significant support for tree [0 7\6\5\4\3\2\1]. Faith and Ballard would interpret this as demonstrating that Four does possess "significant hierarchical information".

An *a priori* AG test of Four with tree [0 1\7\6\5\4\3\2], however, also shows strongly significant support, with a tail probability of 0.0171. This tree has no informative groups in common with the first one. A similar conclusion, moreover, is obtained for any of the eight trees that are most parsimonious for Four, although those trees are mutually contradictory<sup>7</sup>.

The paradoxical conclusions offered by the *a priori* AG test would appear to provide no sensible basis for criticizing other methods. But even if it were granted that those trees are all strongly supported in some special sense, there is no evident way of choosing among them. Our test seems correct, then, in judging that matrix ambiguous.

### Posteriori

Faith (1991) distinguished *a priori* from *a posteriori* T-PTP tests; the two types differ in both method and application. Although Faith and Ballard promise that they will "describe in detail the "all-groups" version of the T-PTP tests", they do not place their AG test in either category. It will be illuminating to explore this subject before proceeding.

As Faith (1991: 371) put it:

"The basis of the distinction is that the *a priori* test is a result of a hypothesis of monophyly derived from extraneous sources, whereas the *a posteriori* hypothesis has arisen because of the apparent support for the particular group in the cladistic analysis of the data".

This refers to an alternative—rather than null—hypothesis. The procedure carried out is in either case a statistical test of the null hypothesis of randomness.

Faith (1991: 368) described the *a priori* T-PTP method:

"The evidence for monophyly can be evaluated by comparison of the observed difference in minimum length under the monophyly versus nonmonophyly constraints, with the corresponding difference values that can be found for randomized data sets".

Monophyly is considered significantly supported if the real length difference exceeds enough (95 out of 99, say) of the length differences from randomized matrices.

In the *a posteriori* method the length difference for the observed data is as before, but randomized matrices are handled differently (Faith, 1991: 371):

"For each randomized data set, the length difference is calculated as the largest value that could be achieved for any monophyletic group of the same size. That is, a difference value, contrasting the length achievable under monophyly versus nonmonophyly, is calculated for each (e.g.) pair of taxa".

<sup>7</sup> At the 1993 Nordic Phylogenetic Systematics Network meeting, L. Struwe suggested (and Faith denied) that similar contradictions can arise with T-PTP tests of single groups.

This being understood, significance is assessed as before.

The advantage of controlling for group size—maximizing over groups of the same size rather than all possible groups—is that it typically improves the power<sup>8</sup> of the test. Faith (1991) did not discuss this, probably because in the same paper he objected to bootstrapping on the extraordinary grounds that it has too much power.

Faith and Ballard's characterization of AG calculations has the virtue of brevity:

"The critical length difference is given by the length of the considered tree, subtracted from the length of the most parsimonious of all trees constrained not to have any of the original groups".

The original groups are those on the considered tree. This defines the AG length difference for a particular tree and matrix, but it does not actually describe any test.

An *a priori* AG test for tree *T* (Faith and Ballard do not say) would be obtained by finding the AG length difference for both real and randomized matrices, one same considered tree—*T*—being used for each matrix. This is the method that we employ here for *a priori* AG tests.

An *a posteriori* test would use one considered tree *T* for the real matrix, but would require maximizing the length difference over considered trees "of the same size" for randomized matrices. Faith and Ballard do not mention this at all, let alone disclose what collection of trees that would be. Any tree for the matrix would have the same number of terminals, but trees can differ in shape, and controlling for such factors might well affect the result of the test. Lacking any guidelines, however, for *a posteriori* AG tests we simply maximize the length difference over all possible trees.

Applying these tests is useful in clarifying parts of Faith and Ballard's discussion. They maintain, for example, that:

"[Matrix] X14 shows a dramatic departure from this congruence [seen in matrix X52] in results for the two tests".

The tests they mean are AG and total support, but they perform neither, and while doing so does show a departure, it is not they one that Faith and Ballard suggest. Neither matrix shows significant total support. Applying an *a priori* AG as in the previous section yields significance for any most parsimonious tree for either matrix, while according to the *a posteriori* AG test none of those trees is significant.

This makes it seem all the more curious that Faith and Ballard never say which type of AG test they intend. As for the difference between X14 and X52, it does not concern outcomes of tests, but rather individual AG and total support values, which subject we take up next.

### Measures

Faith and Ballard base their main objection to total support on results from selected randomizations of matrix X00.

Matrix	Minimum length	Total support	AG
X00	9	4	3,4
X52	10	2	2
X14	11	3	1

<sup>8</sup> The probability with which a test avoids the error of failing to reject a null hypothesis that is in fact false.

In contrast to the case of confused identity discussed before, these are correct total supports, calculated from most-parsimonious trees for each matrix. These AG values are also so calculated, and we do the same in our example below. Most-parsimonious trees for a single matrix can differ in AG value, as for X00 here, though not in total support.

According to Faith and Ballard:

"The TS value is inflated by independently counting the steps for different groups. Consequently, although the tree length for data set X14 is higher by one step relative to X52 (reflecting greater homoplasy and higher PTP [AFC] score), the TS value is actually greater (i.e. suggesting greater support for the topology)".

"Counting the steps for different groups" refers to the fact that total support is found by summing Bremer<sup>9</sup> (group) supports, one for each group of a most-parsimonious tree. (Any most-parsimonious tree for the same matrix yields the same total.) This is supposed to produce the undesirable behavior of total support; why that behavior is undesirable Faith and Ballard explain further by continuing:

"This result might be justified if the MPT topology for X14 was [sic] in some way better supported than the MPT topology for X52—but for each of X52's MPTs, it takes 2 steps to remove all the component groups, while for the more-homoplasious X14 it takes only 1 step. Of course, this difference in support is reflected in the corresponding AG values; AG reports the MPT for X14 as having less support".

The number of steps needed to "remove all the component groups", however, is readily recognized as the definition of their AG length difference. Their comment, then, amounts merely to insisting that "support", by definition, must refer to their measure rather than ours. It is ironic that Faith and Ballard rely on such reasoning, considering that (as quoted at the beginning) they wrongly accuse us of having committed a similar fallacy in our 1992 discussion of AFC.

Plays on definitions aside, Faith and Ballard's criticism is that total support, being—unlike AG—the sum of Bremer supports, may increase with homoplasy. This argument has a weakness by now familiar: it conforms to the one case that they present, but not to other cases. The opposite result is found on comparing Four with one of its randomizations, Xfour:

```

0 1 0 0 1 0 1 1
1 1 0 0 0 1 1 0
1 1 0 1 1 0 0 0
0 1 1 0 1 1 0 0

```

The values are:

Matrix	Minimum length	Total support	AG
Four	7	0	0
Xfour	9	0	1

27 distinct trees for Xfour have AG = 1. No tree for Four has AG greater than 0; we have checked them all. In this case AG increases with homoplasy while total support remains innocently at 0. AG thus shows the very behavior that Faith and Ballard consider objectionable—when it occurs in total support. And this time it is obviously not the summing of Bremer supports that produces that behavior.

<sup>9</sup> We called them Bremer supports because Bremer (1988) had employed an equivalent measure. Faith (1991) used that measure too, but did not mention Bremer's work.

That Faith and Ballard's argument would work against their own measure reflects a much more general defect of their position. They assume, without offering any justifying reasoning, that a measure of support should not increase with homoplasy. But that presumption, if taken seriously, would rule out any worthwhile test of support.

Among randomizations of one matrix, the situation that Faith and Ballard consider, amount of homoplasy differs from minimum tree length by a constant (cf. Farris, 1991). If some measure  $S$  of "support" never increased with rising homoplasy, it would have a perfect—though negative—monotonic (rank-order) relationship to minimum length among such randomizations. A test employing  $S$  would then amount to AFC, which is based directly on comparing minimum lengths from real and randomized matrices.

But examples such as Three show that AFC does not provide an adequate test of unambiguous structure, and not even Faith and Ballard dispute that finding as such. While they attempt to rehabilitate AFC by inventing new interpretations for it, they themselves plainly do not believe that it indicates support for unambiguous grouping.

"The PTP [AFC] test evaluates the level of parsimony achieved by a particular data set, but a significant tree length in this sense does not necessarily imply significant support for particular monophyletic groups".

Our test can judge Three non-significant while AFC does not, precisely because total support is not equivalent to minimum length. But this only illustrates what should be obvious. To test support—rather than minimum length—requires at the least a measure that is not perfectly related to minimum length. Yet any such measure would, according to Faith and Ballard's present argument, show objectionable behavior. As an approach to evaluating measures of support, that is to say, their argument is completely futile.

While Faith and Ballard do not explain their reasons for proposing that argument, they presumably wish to restrict the effects of homoplasy. If so, they overlook a more promising possibility, pointed out in our 1992 paper. A better assessment of evidence could be obtained with a weighted total support, the weights being based on congruence among characters as in the successive weighting method of Farris (1969)<sup>10</sup>. Such weighting naturally reduces the influence of more homoplasious characters. That method is still under development, and we plan to pursue it elsewhere.

### Purpose

There is a peculiarity in the logic of the AG test, seen on comparing the AG length difference to the length difference used in the T-PTP for monophyly.

In the test for monophyly, the length difference employed is that between the shortest tree having the group and the shortest tree lacking the group. The test is intended to evaluate the evidence favoring a particular group, and that length difference seems just the right one for the purpose. A group should be considered weakly supported if there is any arrangement that both denies the group and fits the data well.

<sup>10</sup> Gustafsson and Bremer (1994) employ such weighted Bremer supports, though not in a permutation test.

The AG test is supposed to evaluate support for a particular tree, but here the same reasoning does not seem to apply. Logically, the evidence in favor of a particular tree should be weak whenever any other tree conforms well to the data. But the AG length difference is not that between the initial tree and the shortest possible alternative. Instead, the alternative tree is required to differ from the initial one in every group.

This makes quite a difference to the interpretation. If support for a particular tree were truly being evaluated, then significant support, errors of inference aside, would suggest that particular tree is correct. But AG instead assesses support for any among a class of trees, all those trees that share one or more groups with the initial tree. Significant AG, then, suggests that some member of that class is correct, but does not actually say which one.

The correct interpretation is necessary to avoid contradictory conclusions. A suitable example is provided by matrix FB, which is obtained from Faith and Ballard's X52 by adding three copies of each character, the new number of characters then being 28.

FB has four distinct most-parsimonious trees, each with AG = 8. Each of those four trees is highly significant according to the *a posteriori* AG test, the tail probability being 0.0044. If those tests were taken as indicating significant support for those particular trees, then the results would have to be regarded as nonsense: four different trees cannot each be correct.

The contradiction vanishes on recognizing that the support actually applies to classes of trees, for in this case the classes overlap. The four most-parsimonious trees have an informative group (a + d + g + s) in common. And the Bremer support for that group is 8. All this corresponds to a sensible conclusion: what matrix FB does support is the information on grouping common to those four trees.

While it seems curious in isolation, Faith and Ballard's misinterpretation of their own test can be readily traced to another part of their position:

"One purpose of our paper is to point out that Källersjö et al. unfortunately ignored an existing family of [T-PTP] tests for topology support which, if considered in their work, might have precluded any need for development of a new test".

Our test, recall, is "nearly identical" to AG, and we "were interested in tests relating to support for tree topology".

To support those claims, Faith and Ballard had to come up with the necessary AG. Faith (1991: 372) had not actually constructed any such test, and his general discussion provided some latitude:

"The most general form of the T-PTP test is one in which any combination of compatible constraints on the tree topology is made, and the minimum length is compared to that found for trees not having the these topological characteristics".

That could mean lacking one or more of the original groups, and a length difference so calculated would indeed assess support for a particular tree. But that measure would not behave at all like total support. So the other possible reading of "not having these topological characteristics" had to be used instead, whence the present AG.

### Synthesis

Faith and Ballard's evidence for their view, that our test is "nearly identical" to theirs, comprises agreement on significance.

"The desire to have Three show non-significant structure in some sense did not require, as implied by Källersjö et al. (1992: 283), any new measure; the T-PTP (AG) on this data set is also [non-significant]."

They anticipate the outcome of that AG test, rather than actually performing it. But even if they are right, this is only one case, and it will be useful to consider another.

Faith and Ballard do not say whether they mean an *a priori* or an *a posteriori* AG test. But the claim of near identity is certainly not true of the former, as is clear from examples already presented. For an example of the *a posteriori* test use matrix Yfour. This is obtained from Xfour by adding seven copies of each character, so that the new number of characters is 32.

Yfour has 27 distinct most parsimonious trees with AG = 8 (all other trees have lower AG). Each of those trees is highly significant by an *a posteriori* AG test, the tail probability being 0.0040 in each case. The total support for the same matrix is 0, which is entirely non-significant, the tail probability being 1.0.

So the two tests can hardly be described as "nearly identical", but Faith and Ballard have a back-up position prepared: their test is better:

"Based on these initial comparisons, there appears to be no advantage in Källersjö et al.'s ignoring the existing family of T-PTP tests, and creating what amounts to a poor approximation to the all-groups version of T-PTP tests".

"These initial comparisons" are those based on Faith and Ballard's thoroughly self-defeating idea that support must not increase with homoplasy. But Faith and Ballard have recourse to other arguments. By analogy with their discussion of (a + c + e + g), the AG test shows that Yfour has "significant hierarchical information", which our test fails to find.

Which only means that their concept of "hierarchical information" is nebulous to the point of inscrutability. Whether AG support is for particular trees or for classes of trees does not matter this time. Those 27 trees have no informative groups in common, and the AG test provides no way of choosing among them. Just as was seen earlier with Four, our test is entirely correct in judging Yfour ambiguous.

Our test is meant to assess strength of unambiguous hierarchic structure in data, and that is what it does.

### Acknowledgments

This work was supported by NFR grants 10204-300 to J. S. Farris and 09858-303 to M. Källersjö. Drs V. Albert, K. Bremer, J. Carpenter, D. Lipscomb and W. Wheeler offered many helpful suggestions and much warm encouragement. We particularly thank Dr K. Rognes for his extensive and painstaking analyses and his detailed commentary.

### REFERENCES

- ARCHIE, J. W. 1989. A randomization test for phylogenetic information in systematic data. *Syst. Zool.* 38: 219-252.
- BREMER, K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 795-803.
- GUSTAFSSON, M. H. G. AND K. BREMER. 1994. Morphology and phylogenetic interrelationships of the Asteraceae, Calyceraceae, Campanulaceae, Goodeniaceae, and related families (Asterales). *Am. Jour. Bot.* In press.

- FAITH, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* 40: 366-375.
- FAITH, D. P. AND P. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? *Cladistics* 7: 1-28.
- FAITH, D. P. AND J. W. O. BALLARD. 1994. Length differences and topology-dependent tests: a response to Källersjö et al. *Cladistics* 10: 57-64.
- FARRIS, J. S. 1969. A successive approximations approach to character weighting. *Syst. Zool.* 18: 374-385.
- FARRIS, J. S. 1991. Excess homoplasy ratios. *Cladistics* 7: 81-91.
- HAMBY, R. K. 1990. Ribosomal RNA and the early evolution of flowering plants. Ph.D. thesis. Louisiana State University, Baton Rouge.
- KÄLLERSJÖ, M., J. S. FARRIS, A. G. KLUGE AND C. BULT. 1992. Skewness and permutation. *Cladistics* 8: 275-287.
- THOMAS, R. H., W. SCHAFFNER, A. C. WILSON AND S. PÄÄBO. 1989. DNA phylogeny of the extinct marsupial wolf. *Nature* 340: 465-467.