Cladistics (1992) 8:275–287

SKEWNESS AND PERMUTATION

Mari Källersjö[1,4], James S. Farris[1,2], Arnold G. Kluge[3] and Carol Bult[4]

[1] *Naturhistoriska riksmuseet, Molekylärsystematiska laboratoriet, Box 50007, S-104 05 Stockholm, Sweden*

[2] *Department of Entomology, American Museum of Natural History, Central Park West at 79th St, New York, New York 10024, U.S.A.*

[3] *Department of Reptiles and Amphibians, Museum of Zoology, The University of Michigan, Ann Arbor, Michigan 48109, U.S.A*

[4] *Laboratory of Molecular Systematics, Smithsonian Institution, Washington, D.C. 20560, U.S.A.*

Received for publication 17 January 1992; accepted 2 June 1992

Abstract—The skewness criterion of phylogenetic structure in data is too sensitive to character state frequencies, is not sensitive enough to number of characters (degree of corroboration) and relies on counts of arbitrarily-resolved bifurcating trees. For these reasons it can give misleading results. Permutation tests lack those drawbacks and can be performed quickly by using approximate parsimony calculations, but the test based on minimal tree length can imply strong structure in ambiguous data. A more satisfactory test is obtained by using a support measure which takes multiple trees into account.

Introduction

Following Fitch's (1979) early suggestion, Le Quesne (1989), Huelsenbeck (1991) and Hillis (1991) have all recommended assessing the phylogenetic structure in systematic data according to the skewness of the distribution of tree lengths. We point out here that such evaluations can be misleading; arguments for that approach are not well-considered.

The permutation method of Archie (1989) and Faith and Cranston (1991) seems more promising for this purpose, but it requires excessive effort and may underrate the departure of data from randomness. We describe a much less laborious implementation and a stronger evaluation. Most importantly, the existing method may suggest significant structure for quite ambiguous data.

To resolve this problem we introduce a test based on a new measure—total support—which takes multiple most parsimonious trees into account. Our fast method for approximating support may prove useful in analyses of very large data matrices.

Skewness

The distribution of tree lengths (DTL) is obtained by finding the length[1] of each bifurcating tree for the data; multifurcating trees are not included. When there are too many terminals to evaluate every tree, trees are sampled at random.

[1] Throughout, the smallest number of steps needed for the characters as coded and weighted to evolve on the tree.

0748-3007/92/020275+13$08.00/0

The g_1 index (of Sokal and Rohlf, 1981) is used to quantify skewness. It is typically negative when the distribution is left-skewed, that is, when the median exceeds the mean.

The skewness index itself is routine statistics; the further interpretation is not. Degree of negative skewness of the DTL is supposed to indicate strength of phylogenetic signal (as Hill calls it). But it need not do so, as can be seen on comparing results from data matrices One and Two.

One	Two
1 1 1 1 1 0 0 0 0 0	1 1 0 0 0 0 0 0 0 0
1 1 1 1 1 0 0 0 0 0	0 0 1 1 0 0 0 0 0 0
1 1 1 1 1 0 0 0 0 0	0 0 0 0 1 1 0 0 0 0
1 1 1 1 1 0 0 0 0 0	0 0 0 0 0 0 1 1 0 0
1 1 1 1 1 0 0 0 0 0	0 0 0 0 0 0 0 0 1 1
0 0 0 0 0 1 1 1 1 1	1 0 0 0 0 0 0 0 0 1
0 0 0 0 0 1 1 1 1 1	0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1	0 0 0 1 1 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1	0 0 0 0 0 1 1 0 0 0
0 0 0 0 0 1 1 1 1 1	0 0 0 0 0 0 0 1 1 0

Rows are characters; columns are terminals.

All 10 characters of One match the same division of the terminals. While One does not determine a fully resolved tree, the signal that it does provide is strong and definite.

Two yields two distinct most parsimonious trees (using the exact-solution, i.e. command of Hennig86), each supported by five of the 10 characters. Those trees have no informative groups in common: their consensus (from the Hennig86 nelsen command) is entirely unresolved. Two could be said to mix two incongruent signals, or to lack a coherent signal. Either way, the net signal is certainly weak.

Exhaustive enumeration of bifurcating trees, the root (outgroup) being held fixed, provides the DTL of One.

Steps	10	20	30	40	50
Trees	11 025	154 350	668 250	936 000	257 400

This distribution was obtained using Farris' dtl program. The skewness is $g_1 = -0.288$.

Similar processing of Two gives:

Steps	15	16	17	18	19	20
Trees	30	2475	37 050	237 525	755 100	994 845

Here, $g_1 = -0.959$. The skewness criterion produces the thoroughly unreasonable conclusion that ambiguous Two has a stronger signal than One.

It is seen that DTL skewness can be more strongly influenced by the frequencies of states within characters than by congruence among characters. Characters dividing the terminals into large groups of similar size (as in One) make the DTL more symmetrical. Characters setting off small groups (as in Two) increase left-skewness. The implications of this effect have been neglected, as will be seen.

DTL skewness also has the property of insensitivity to the number of characters.

The matrix comprising five copies of each of the characters of Two has the same skewness as Two. One shows just one type of character distribution.[2] The DTL skewness is the same, whether that character distribution is represented one, 10 or 50 times.

The skewness-based assessment of weak signal in One may be reasonable when there is just one such character, but it is highly implausible when there are 50. A measure of strength of phylogenetic structure in data must surely reflect the degree to which conclusions are corroborated, but DTL skewness does not seem to do this.

Permutation

The approach of Archie (1989) and Faith and Cranston (1991) (hereinafter AFC) provides a significance test for phylogenetic structure. The underlying mathematics is discussed in more detail by Farris (1991), whose treatment we generally follow.

Permutation methods compare the observed data to randomizations of those data. A randomization is a matrix generated by permuting (rearranging) the entries within each row (character) of the original data matrix. A separate permutation is chosen at random for each character, so that congruence among characters in a randomization is just that produced by chance associations.

In the AFC procedure, congruence is assessed simply from the length of the most parsimonious tree(s) for a matrix, here for brevity termed the minimal length (ML). MLs are calculated for the observed data and for each of a sample comprising a number W of randomizations. The MLs for some number E of those randomizations exceed that for the observed data. If the lower tail probability (error rate) $\alpha' = 1 - E/(W+1)$ is small enough (no greater than 5%, say), the data differ significantly from randomizations.[3]

Evaluated against $W = 999$ randomizations (using the kara program, discussed later), Two yields $\alpha' = 470/1000$—comfortably far from significance. The level of incongruence in Two is near the median of that resulting from random association of characters. One gives the very highly significant $\alpha' = 1/1000$, a more satisfactory assessment than that suggested by skewness.

Unlike DTL skewness, α' is sensitive to the number of characters. Matrices having one, two and three of the characters of One give α' values 1000/1000, 7/1000 and 1/1000, respectively, so reflecting degree or corroboration. But as α' can be no less than $1/(W+1)$, that sensitivity is limited. It can be improved by the standardized score methods described below.

This procedure lacks the peculiar sensitivity to state frequencies shown by DTL skewness because permutation does not change those frequencies. Randomizations differ from the observed data only in the joint—not marginal—frequencies of states. This is like the usual chi-square test of independence, in which expectations of joint frequencies are calculated with the marginals held at their observed values. It is instructive to extend that comparison.

The chi-square could be performed by randomization, but that is unnecessary as distribution tables are available. Randomization is used with ML because of the

[2] There might be two if apomorphies were specified.

[3] Archie's (1989) wording incorrectly implied significant structure when α' is *large* enough.

difficulty of calculating the distribution directly. The two tests have the same formal null hypothesis. They employ different measures of departure from independence (randomness) because the chi-square is designed for a broader class of alternatives. The ML criterion is intended to identify hierarchic structure in particular.

It might be supposed[4] that holding the marginal frequencies fixed rests on the assumption that they are fixed in nature. But, in fact, in both the chi-square and ML tests the observed values of those frequencies are used in order to assess just the correlations among variables. The null hypothesis, that is, postulates only independence, not anything about the marginal distributions.

It is possible, of course, to include specific marginal expectations in a null hypothesis, but this is not a useful way to study congruence, which is a kind of correlation. A test of such a hypothesis would reject on data not matching the hypothesized marginals, regardless of correlations among variables.

Models

Hillis (1991) proposed a skewness-based significance test. Conclude significant structure when g_1 for the data DTL is below the fifth percentile (say) of DTL g_1 for matrices produced under his null model. The characters of such matrices are generated randomly and independently, with all states having the same expected frequency.

Data depart from that model when characters are highly congruent, but also simply when states have different frequencies. As skewness is influenced by both congruence and state frequency, Hillis' test confounds the two effects.

A character with equally abundant states might, of course, be poorly correlated with phylogenetic relationships. But it also might well distinguish a large monophyletic group. It is obvious that characters whose states depart from equal frequency occur in real data. But that leaves open the question of whether those characters are congruent, and it is nonsense to view them as providing a strong phylogenetic signal when they are poorly congruent.

Reckoning phylogenetic signal by departure of states from equal frequency is thus surely ill-founded. Further, if for some reason one wanted to test such departure, a conventional chi-square test of equality of marginal frequencies would suffice; no new method would be needed.

Huelsenbeck's (1991) advocacy of skewness was based on results from simulations. The most parsimonious tree for simulated data, he found, is likely to be accurate (match the simulated tree) when the DTL is strongly left-skewed, less so otherwise.

In those simulations, all branches of the tree had the same probability of character change. Accuracy of the most parsimonious tree is determined by that probability; it is best for intermediate values. If the change probability is too small, the simulated characters are likely to be invariable or autapomorphic. If it is too large, the character distributions become independent of the tree.

Skewness is likewise determined by the change probability. It is 0 when all characters are invariable or autapomorphic. It is strongest for intermediate change probabilities, when states are most likely to depart from equal frequencies.

---

[4] Both W. Maddison and D. Faith (!) did so at the 1991 meeting of this Society.

Skewness is weak when the change probability is large, for then the expected frequencies of states become equal—as in Hillis' null model.

Accuracy of the most parsimonious tree is thus correlated with DTL skewness in Huelsenbeck's study. But that correlation is not general, for it results from a restriction of his simulations: that all branches have the same change probability.

Data such as One might, for example, be found when the studied species comprise two anciently separated but recently diversified groups. Then character change would be much more likely in the tree's basal branches than in others, so that Huelsenbeck's assumption would not apply. Under those circumstances parsimony analysis might well correctly identify those groups, despite the weak skewness of the DTL.

Le Quesne (1989) compared real to randomly generated matrices, finding that DTLs of the latter showed weaker skewnesss and lower variance. His random matrices were produced by permutation (although he did not use that term) of the real data, so that state–frequency differences did not confound his contrasts.

Nonetheless, Le Quesne did not draw his conclusion carefully enough (p. 406):

> "When real data show both a large variance and large negative skewness the data are more likely to be informative than when only one of these values is large".

Large, it must be added, compared to random data with the same array of state frequencies. Without that qualification, the difficulty exemplified by One and Two can easily arise.

Speed

While free of that difficulty, the permutation method has problems of its own, the most obvious of which is effort. Since α' can be no smaller than $1/(W+1)$, MLs must be found for at least 99 randomizations to demonstrate a highly significant departure from randomness—and more for higher significance.

Archie (1989) and Faith and Cranston (1991) calculated most parsimonious trees as exactly as they could with PAUP: by branch and bound for small matrices, by global branch-swapping for larger ones. They had to prepare a separate input matrix for each randomization of data; α' values were obtained by gleaning results from PAUP outputs. For even a moderately large data matrix, all this might easily take a week.

Most of that work is unnecessary. The test requires only the length of a most parsimonious tree for each matrix, whereas much of the time expended by PAUP was spent on identifying multiple trees. For exact parsimony calculations it would be several times faster to use the Hennig86 ie command, which is designed to find just one most parsimonious tree.

But Hennig86's single-pass hennig command is much faster still, while the length that it yields seldom departs from the exact ML by more than a few percent. The test uses only the number of randomizations whose MLs exceed that for the data. If the same method is applied to both kinds of matrices, such small approximation differences are unlikely to have much effect on that number.

Finally, a suitable program can generate and process randomization internally, obviating the need to handle numerous input matrices and output listings. We have combined these features with a simple hennig algorithm in the kara program, part of whose output is illustrated in Fig. 1. This is a bar-chart of the distribution of

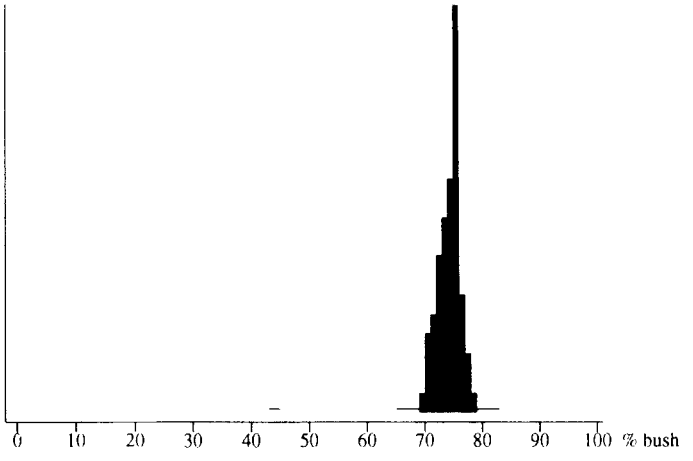

Fig. 1. ML distribution for the data of Crisci et al. (1990).

MLs obtained from 9999 randomizations of the restriction-site data of Crisci et al. (1990). MLs are shown as fractions of the length of the unresolved tree.[5]

The left-most bar of that chart represents the ML for the observed data. As this is well separated from the rest of the distribution, there is little chance that the approximate parsimony calculations have produced an incorrect conclusion. The approximation is equally safe for poorly-structured data such as Two: the data ML is well within the distribution. It might be more problematical in cases of marginal significance—but then those are problematical anyway.

This approach makes the permutation method much easier to use. For the data of Crisci et al., which have $n = 65$ characters and $t = 11$ terminals, 10 000 ML calculations require just 915 seconds on a 20 megahertz 80386DX (to which timings refer throughout). Young's (1981) data, with $n = 41$ and $t = 34$, take 577 seconds for 1000.[6] Dahlgren and Bremer's (1985), with $n = 61$ and $t = 50$, take 342 seconds for 100. The present program is a prototype, and we expect that its speed can be further improved.

Faith and Cranston (1991) suggested (but did not pursue) the alternative of comparing quickly approximated MLs for randomizations to the (or a more) exact ML for the observed data. They felt that this would provide a conservative test, that is, one less likely to yield specious conclusions of significant difference. But if anything their proposal would have the opposite effect.

In a case of near-significance the observed data would show smaller ML than most of the randomizations, supposing all MLs to be exact. As approximate MLs exceed corresponding exact values, Faith and Cranston's suggested procedure would increase the apparent difference between real and randomized matrices. This would worsen the risk of a false conclusion of significant congruence.

---

[5] For given data, no tree is longer than the bush, and permutation does not affect bush length (cf. Farris, 1991).

[6] In view of Riggins and Farris' (1983) comments on Young's coding, his data are treated as nonadditive throughout.

Departure

Ideally the permutation test would employ the exact distribution of MLs for the statistical population of possible randomizations. A sample of randomizations is used instead because it is not presently feasible to calculate that exact distribution in any but simple cases.

A test based on the exact distribution would have as its error rate lower tail probability α, the population probability that a randomization of the observed data yields ML no greater than that for the data. For some limited number W of randomizations, this population α may be much less than $1/(W+1)$. When this occurs, the α' value from the sampling technique will understate the departure of the data from randomness.

An improved evaluation can sometimes be obtained by approximating the population α. We will discuss methods based on the standardized score $Z = (A - L)/S$. Here, L is the ML for the observed data; A and S are, respectively, the mean and the standard deviation of MLs from the sample of W randomizations. Suppose that W is reasonably large (99 or more), so that this sample Z is likely to be close to its population value.

Archie's (1989) suggested use of Student's t-test is in this category: the t statistic amounts to Z. Unfortunately, the standard t tables would yield accurate tail probabilities (significance levels) only if lengths from randomizations were normally distributed. Archie correctly noted that requirement, but did not maintain that it is satisfied. That it is not generally satisfied can be seen from the exact distributions tabulated by Farris (1991) and from the sample distribution figured here.

Those distributions are left-skewed (this has little to do with DTL skewness), with a left tail thicker than that of a normal distribution. Tail probabilities from the standard t tables may then be considerably smaller than the accurate values; their use would increase the chance of an erroneous finding of significant congruence. It is safer to employ a conservative approximation to the tail probability, that is, one bounded below by the population α.

A very conservative approximation $\alpha'' = (1/Z)^2$ is given directly by Chebyshev's theorem (cf. Walpole, 1983), which uses no information about the form of the exact distribution. A much closer, though typically still quite conservative, value $\alpha^* = e^{-z}$ can be obtained by using the fact that the exact distribution falls off faster than exponentially in the left tail. The latter approximation is not reliably conservative if there are too few informative characters, but this is seldom a drawback in practice.

Neither of these formulae is useful when Z is small or negative. But when Z is big enough, using α'' or α^* can considerably reduce the effort of establishing high significance for a large matrix. An example is provided by Hamby's (1990) ribosomal RNA sequence data.

These data have 471 sites for 60 terminals, and it takes nearly 38 minutes to find $\alpha' = 1/100$. But in the bar-chart (Fig. 2) the ML for the observed data is widely separated from a narrowly concentrated distribution of MLs for randomizations. This makes for a large $Z = 65.7$, so that even α'' is only about 0.00023. To obtain such a significant α' would require $W > 4300$ and over 27 hours of computation. No feasible amount of computer time would suffice for α' to match $\alpha^* = 2.9 \times 10^{-29}$.

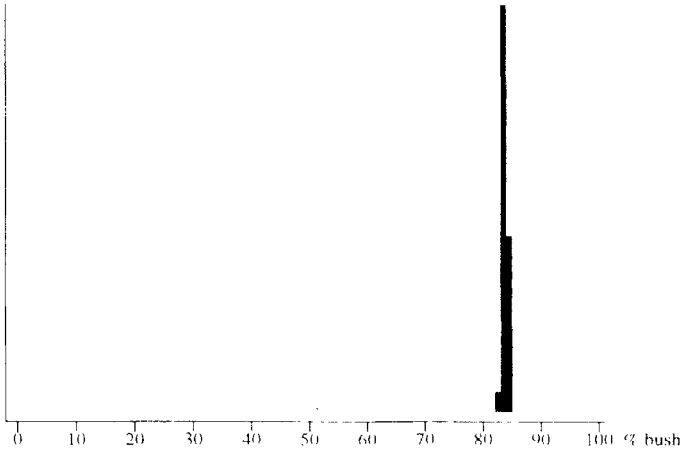

Fig. 2. ML distribution for the data of Hamby (1990).

Because α'' and α^* are conservative, the sample α' can be smaller than either; the smallest value should be used to determine the significance level for a data matrix. As an example take Young's (1981) morphological data (Fig. 3). In this case the ML for the real data is much nearer the distribution of MLs for randomizations, so that Z is only 8.06, $\alpha' = 0.015$, and $\alpha^* = 3.2 \times 10^{-4}$. The α' found from 9999 randomizations is more significant at 10^{-4}.

Z continues to increase as congruent characters are added to data. Matrices having five, 10, 20 and 50 of the characters of One yield Z values of 6.3, 10.5, 17.1 and 29.0, respectively. Using 9999 randomizations all those matrices have the same α' value, 10^{-4}. When the number of characters is large, the standardized score thus provides a more sensitive indication of strength of corroboration.

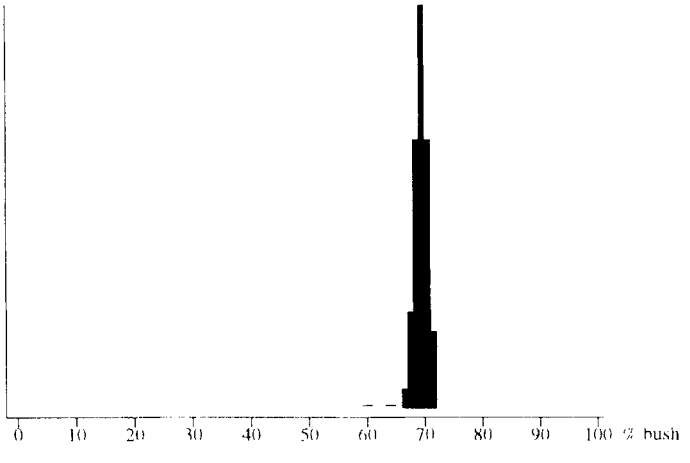

Fig. 3. ML distribution for the data of Young (1981).

Faith and Cranston (1991) advocated $(1 - \alpha')$ as an index of "cladistic covariation"[7] (hierarchic structure). This would be 0.99 for Hamby's data, 0.9999 for Young's. The latter is larger because the feasible number of randomizations is greater for the smaller matrix, but that hardly indicates better structure. Taking the precaution of fixing W at (say) 100, $(1 - \alpha')$ would be the same for the two matrices.

That evaluation does not reflect the greater departure from randomness in Hamby's data, evident on comparing Figs 2 and 3. It would probably be more useful to base an index on standardized scores. Here, however, we concentrate on tests rather than indices. Choice of a measure of structure should in any case take the results of the next section into account.

Support

A different kind of difficulty with the AFC method is illustrated by data matrix Three. As before, rows are characters.

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

Three yields 10 distinct most parsimonius trees, whose consensus is entirely unresolved. No well-defined hierarchic structure is present. Yet $\alpha' = 1/1000$, showing a very highly significant departure from randomness.

While Three indeed gives shorter trees than do most of its randomizations, this need not mean that the data show unambiguous hierarchic structure. What is needed is a measure that reflects such structure. We will derive one from methods used in earlier work.

Ambiguity is usually detected by finding multiple most parsimonious trees, the unambiguous part of the structure (that common to the several trees) being recovered as a consensus tree. Sometimes trees longer by some amount are included as well, although the choice of such a value is seldom explained.

Farris et al. (1982) approached that problem of choice by formalizing earlier ideas from distance analyses. Adding trees in order of decreasing goodness of fit, they calculated a series of consensus trees, noting the level of fit at which groups were lost from the consensus. They suggested using the consensus of those trees separated from others by a large gap in goodness of fit (cf. Farris, 1985).

Bremer (1988) employed similar series of consensus trees in parsimony analyses, adding trees in order of increasing length. Unlike Farris et al., he used "strict" consensus trees. That type of consensus is easy to calculate, and we use it here.

[7] Perhaps by analogy with Farris' (1969) hierarchic correlation.

It provides a minimal assessment of common structure—perhaps too much so—but other available consensus techniques have faults as well.

On this view, a group on a considered most parsimonious tree is supported by strong evidence when a large increase in length of included trees is required before that group is lost in the consensus. The strict consensus lacks any group absent from any included tree. We thus define the *Bremer support* of such a group as the difference in length between the considered tree and the shortest tree(s) lacking that group.[8]

Assessing support in this way offers some advantages. Strength of evidence for monophyly of a group is usually equated to number of synapomorphies. But that number may not be clearly defined if there are several parsimonious reconstructions of character states for the stem species of the tree. One might take evidence as the minimum number of synapomorphies among reconstructions. But even this may overstate the case, if some most parsimonious tree lacks the group.

The support measure takes both these problems into account. When characters are perfectly congruent and the reconstruction unique, the Bremer support of a group is the (possibly weighted) sum of character changes that set off that group, that is, the number of synapomorphies. Otherwise, this amount is reduced to the degree that alternative groupings or character interpretations are parsimonious. It is zero when the group is absent from some most parsimonious tree.

Faith (1991) has described a support-based[9] test for evidence of monophyly of specific groups. We will not pursue that subject here, but employ support to evaluate hierarchic structure in the data as a whole. For this purpose we use *total support*, the sum of group supports.

Total support is typically greater in well-structured data than in randomizations. Bearing this is mind, the new measure is easily incorporated into a permutation test. If a number X of the W randomizations yield total support no less than of the observed data, then the error rate on concluding significant structure is upper tail probability $\alpha_s' = (X+1)/(W+1)$.

Standardized scores for total support can be used in Chebyshev and exponential approximations to the population upper tail probability α_s. Unlike the MI test, we have not yet encountered a real data matrix for which α_s^* is much less than α_s', but this lack seems unlikely to be permanent.

For a data matrix of any great size, it is not practical to evaluate support exactly within a permutation method. Fortunately, an approximation may reasonably be used. Notice that α_s' depends just on the number of randomizations showing total support no less than than of the observed data. A small α_s' should then give a reliable indication of significant structure, provided only that the approximated total support is large only for a well-structured matrix.

To obtain a fast approximation in kara we use a simplified branch-swapping algorithm. It considers just trees that can be obtained from the hennig tree by replacing branches one at a time. With $W = 99$ the support test of Young's (1981) data requires 244 seconds, less than 2.5 seconds per matrix. In contrast, the bb

---

[8] Donoghue et al. (1992) called Bremer's length difference "decay". That seems an unfortunate choice: the most strongly supported groups would be most decayed.

[9] Not mentioning consensus trees or connected work, Faith attributed the measure to a suggestion by Felsenstein.

command of Hennig86 takes 184 seconds just to produce 100 most parsimonious trees for the observed data.

While no doubt capable of improvement, that method seems to give satisfactory results. For ambiguous Three kara gives the thoroughly non-significant $\alpha_s' = 1000/1000$. Whenever the consensus of most parsimonious trees is unresolved, total support is zero, so that the population α_s is necessarily unity. For One $\alpha_s' = 1/1000$, as is surely appropriate.

For a practical example, compare Young's (1981) morphological data with Hamby's (1990) rRNA sequence data. Both pertain to relationships among higher groups of angiosperms. The ML test assesses Young's matrix as very highly significantly structured. But Riggins and Farris (1983), who analyzed Young's data in detail, found them quite feebly-structured.

The support test of Young's data gives $\alpha_s' = 84/100$—worse than the majority of randomizations and far from significance. For Hamby's data $\alpha_s' = 1/100$, showing highly significant hierarchic structure. As with Three, the support test seems better able to recognize poorly structured data than is the test based only on minimal length.

With extremely large matrices it has until now often been impractical to find more than a single approximately most parsimonious tree. As this gives no indication of ambiguity, even an approximation to support would be a considerable benefit. Because of its speed, kara's support approximation is feasible for large matrices and so may prove valuable in such cases.

A full assessment of strength of evidence should take the reliability of characters into account, but evaluation of characters may itself be influenced by congruence. While the present program does not do so, support can be calculated with congruence-based weights. This would complicate the test procedure somewhat, as weights would vary among randomizations. It nonetheless seems feasible to use a weighted support measure, and we plan to investigate this possibility elsewhere.

Statistics

A last argument for skewness (cf. Hillis, 1991) fits here because it concerns multiple trees, but it also serves to summarize the thinking behind the skewness approach.

A strongly left-skewed DTL (this reasoning runs) will have a long, thin left tail, in which case relatively few trees will be most parsimonious or nearly so. If the DTL is less skewed, the left tail will be less attenuated, so that relatively many trees have minimal or near-minimal length. DTL skewness would thus seem to measure ambiguity of data in the same sense that support does.

But skewness does not measure just the ends of tails. Determined predominantly by the central mass of the distribution, skewness is negative when the median exceeds the mean, whether the left tail is greatly attenuated or not. That difficulty is particularly obvious when there are many terminals and DTL skewness is estimated from a random sample of possible bifurcating trees.

The skewness index itself can be estimated quite precisely from such a sample. But for as few as 20 terminals, there are over 2.2×10^{20} bifurcating trees, and even a sample of 1 000 000 trees would comprise only a tiny fraction of the possibilities.[10]

[10] In PAUP, recommended for this purpose by Hillis (1991) and Huelsenbeck (12991), the default is 1000 trees.

If the left tail is very attenuated, the sample is unlikely to capture much of it, let alone the nearly and most parsimonious trees in particular.

The very trees on which the reasoning depends have little chance of affecting the estimated skewness. To make such an argument properly, a more suitable index for assessing the DTL would need to be developed. But there is little to be gained by doing so, for the DTL itself can be misleading. According to the DTLs, highly-structured One has a much greater number of most parsimonious trees than does ambiguous Two.

The DTL is based only on bifurcating trees. One's single most parsimonious tree is poorly resolved, and so corresponds to many "distinct" bifurcating arrangements. Each of Two's most parsimonious trees is better resolved, and they are represented in the DTL by fewer bifurcating schemes. Counting arbitrary resolutions as distinct leads to exactly the wrong assessment of ambiguity in these matrices.

Interpreting DTL skewness as strength of phylogenetic structure, in short, consists of using a poorly-chosen statistic to summarize a poorly-chosen distribution.

Recent emphasis on statistical methods has a parallel in earlier stress on quantitative approaches. Valuable as it was, the earlier idea nonetheless fostered a plethora of now-vanished phenetic techniques. All the methods discussed here are statistical in some sense. That does not mean that they are all equally useful in phylogenetic systematics.

Acknowledgments

Financial support was provided for M. Källersjö by a postdoctoral grant from the Swedish Natural Science Research Council. P. Goloboff, M. Novacek and G. Naylor made constructive and valuable suggestions. K. Hamby and E. Zimmer provided data from a manuscript in preparation.

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.

CRISCI, J. V., E. A. ZIMMER, P. C. HOCH, G. B. JOHNSON, C. MUDD AND N. S. PAN. 1990. Phylogenetic implications of ribosomal DNA reconstruction site variation in the plant family Onagraceae. Ann. Missouri Bot. Gard. 71: 633–699.

DAHLGREN, R. AND K. BREMER. 1985. Major clades of the angiosperms. Cladistics 1: 349–359.

DONOGHUE, M. J., R. G. OLMSTEAD, J. F. SMITH AND J. D. PALMER, 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Missouri Bot. Gard. (in press).

FAITH, D. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40: 366–375.

FAITH, D. AND P. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? Cladistics 7: 1–28.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18: 374–385.

FARRIS, J. S. 1985. Distance data revisited. Cladistics 1: 67–85.

FARRIS, J. S. 1991. Excess homoplasy ratios. Cladistics 7: 81–91.

FARRIS, J. S., A. G. KLUGE AND M. F. MICKEVICH. 1982. Immunological distances and the phylogenetic relationships of the *Rana boylii* species group. Syst. Zool. 31: 479–491.

FITCH, W, M. 1979. Cautionary remarks on using gene expression events in parsimony procedures. Syst. Zool. 28: 375–379.

HAMBY, R. K. 1990. Ribosomal RNA and the early evolution of flowering plants. Ph.D. thesis. Louisiana State University, Baton Rouge.

HUELSENBECK, J. P. 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst. Zool. 3: 257–270.

HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. *In:* M. M. Miyamoto and J. Cracraft (eds). Phylogenetic Analysis of DNA Sequences. Oxford Univ. Press, Oxford.

LE QUESNE, W. J. 1989. Frequency distributions of lengths of possible networks from a data matrix. Cladistics 5: 395–407.

RIGGINS, R. R. AND J. S. FARRIS. 1983. Cladistics and the roots of angiosperms. Syst. Bot. 8: 96–101.

SOKAL, R. R. AND F. J. ROHLF. 1981. Biometry. Freeman, San Francisco.

YOUNG, D. A. 1981. Are the angiosperms primitively vesselless? Syst. Bot. 6: 313–330.

WALPOLE, R. E. 1983. Elementary Statistical Concepts. Macmillan, New York.