Signal, Noise, and Reliability in Molecular Phylogenetic Analyses

D. M. Hillis and J. P. Huelsenbeck

DNA sequences and other molecular data compared among organisms may contain phylogenetic signal, or they may be randomized with respect to phylogenetic history. Some method is needed to distinguish phylogenetic signal from random noise to avoid analysis of data that have been randomized with respect to the historical relationships of the taxa being compared. We analyzed 8,000 random data matrices consisting of 10–500 binary or four-state characters and 5–25 taxa to study several options for detecting signal in systematic data bases. Analysis of random data often yields a single most-parsimonious tree, especially if the number of characters examined is large and the number of taxa examined is small (both often true in molecular studies). The most-parsimonious tree inferred from random data may also be considerably shorter than the second-best alternative. The distribution of tree lengths of all tree topologies (or a random sample thereof) provides a sensitive measure of phylogenetic signal: data matrices with phylogenetic signal produce tree-length distributions that are strongly skewed to the left, whereas those composed of random noise are closer to symmetrical. In simulations of phylogeny with varying rates of mutation (up to levels that produce random variation among taxa), the skewness of tree-length distributions is closely related to the success of parsimony in finding the true phylogeny. Tables of critical values of a skewness test statistic, g_1, are provided for binary and four-state characters for 10–500 characters and 5–25 taxa. These tables can be used in a rapid and efficient test for significant structure in data matrices for phylogenetic analysis.

An advantage of molecular phylogenetic studies is that comparable (orthologous) genes can be examined across virtually all living taxa to estimate evolutionary history. Unfortunately, the ease with which such comparisons can be made sometimes leads to analysis of molecular data sets that consist of little phylogenetic signal and considerable random noise (Fitch 1979, 1984; Hillis 1991). Random noise is generated if rates of change between nodes of a phylogenetic tree are high enough to effectively randomize the character states with respect to phylogenetic history (i.e., a probability of change of ≥75% between nodes for DNA sequences). It is often tacitly assumed that if DNA or protein sequences can be aligned, then the sequences are appropriate for phylogenetic analysis. However, some highly conserved positions in a given sequence may be sufficient to establish an unambiguous alignment, and yet the variable positions in the same sequence may be saturated by change. If all the variation among taxa is essentially random with respect to phylogenetic history (because of rapid evolution at the variable sites), then there is no basis to expect the most-parsimonious tree (or the best tree by any other optimization criterion; see Swofford and Olsen 1990) to be a good estimate of phylogeny. The question, then, is how can phylogenetic signal be distinguished from random noise in molecular (or other) data sets?

There are two commonly used, but never precisely described or defended, methods to evaluate data quality in phylogenetic studies using parsimony. The first is to evaluate the number of most-parsimonious trees. For instance, finding a single (or only a few) most-parsimonious tree(s) out of a large universe of possible trees is commonly taken to mean that the data contain considerable discriminatory power. A second criterion is sometimes used if a single most-parsimonious tree is found: if the most-parsimonious tree is several character-step changes away from the next best solution(s), this is usually taken as evidence of strong resolution.

In this article we show that even with

From the Department of Zoology, University of Texas at Austin, Austin, TX 78712. This paper was delivered at a symposium titled "Molecular Genetic Approaches to Phylogeny Reconstruction" sponsored by the American Genetic Association in Tucson, Arizona, April 20–21, 1991. We thank Jim Bull and Edna Huelsenbeck for assistance. This work was supported by a grant from the National Science Foundation (BSR 8657640). Address reprint requests to Dr. Hillis at the address above.

Journal of Heredity 1992;83:189–195; 0022-1503/92/$4.00

random data (i.e., all noise and no signal), one expects to pass these tests at relatively high frequency if the number of taxa is on the order common in most molecular systematic studies. We also show that an easily and rapidly calculated measure, based on the shape of the distribution of the lengths of all trees or a random sample thereof, provides a powerful and effective means of discriminating signal from noise in systematic data sets. Moreover, simulations of phylogeny demonstrate a close relationship between this measure and the effectiveness of parsimony to correctly estimate phylogeny.

Methods

We produced random data matrices to represent sequences of nucleotides (four character states) as well as binary characters. In both cases, each character state had an equal probability of appearing in a given cell of the character matrix. The dimensions of the character matrices generated in this study were varied for the number of taxa, the number of characters, and the number of possible character states. For both binary data and nucleotide sequences, we produced 100 data matrices for each of 5, 6, 7, 8, 9, 10, 15, and 25 taxa and 10, 50, 100, 250, and 500 characters.

We analyzed the 8,000 data matrices using PAUP (Phylogenetic Analysis Using Parsimony, version 3.0q; Swofford 1990). For data matrices of nine or fewer taxa, the exhaustive-search option (which analyzes all possible tree topologies) was used. For 10 or more taxa, the random-trees option was used to draw 10,000 trees at random (with replacement) from all possible tree topologies. Approximately 2×10^8 trees were examined among all the data sets. Several statistics were gathered from each analysis. For analyses in which an exhaustive search was performed, the number of most-parsimonious tree(s), the number of steps to the next-most parsimonious tree, and the g_1 value for the tree-length distribution were recorded. The most-parsimonious tree(s) were those tree(s) that minimized the number of character transformations (all characters were unordered). The number of character steps to the next-most parsimonious tree had a minimum value of 0 (i.e., if there was more than one most-parsimonious tree). The g_1 statistic is a measure of the skewness of a distribution (Sokal and Rohlf 1981) and is defined as the third central moment divided by the cube of the standard deviation:

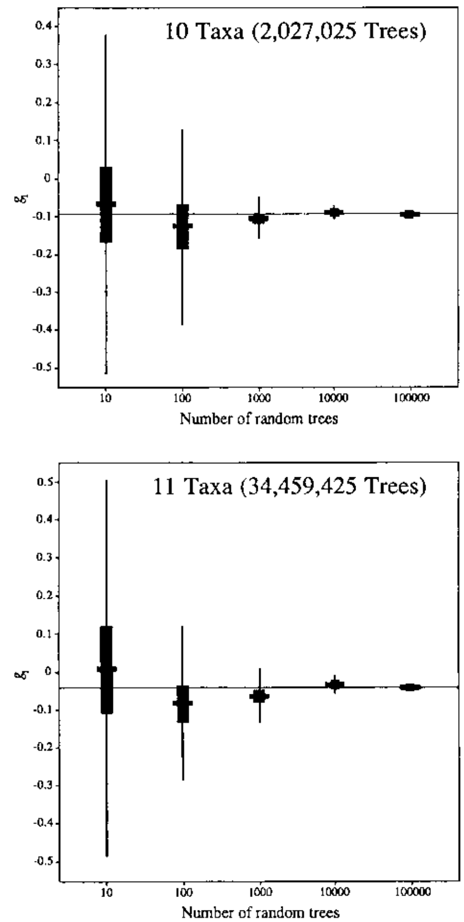

Figure 1. Estimation of skewness of tree-length distributions from random samples of trees. The g_1 statistic was calculated for 20 random samples for each category and compared to the true g_1 calculated from the entire distribution. The horizontal line represents the true g_1 as calculated from an exhaustive search of all possible unrooted trees for 10 and 11 taxa. The thin vertical line represents one standard deviation about the mean, and the black vertical box represents the standard error of the mean. The cross-bar indicates the mean.

$$\frac{\sum_{i=1}^{n} (T_i - \bar{T})^3}{n\, s^3}$$

where n is the number of trees of length T and s is the standard deviation of tree lengths. For a perfectly symmetrical tree-length distribution $g_1 = 0$, whereas a left-skewed distribution has a $g_1 < 0$ and a right-skewed distribution has a $g_1 > 0$.

For character matrices of 10 or more taxa, it was not practical to perform exhaustive enumerations of all possible trees (e.g., for 10 taxa, there are more than 2×10^6 unrooted tree topologies, and for 25 taxa, more than 2×10^{28} topologies). For this reason, we gathered only the g_1 value of the tree-length distribution, as estimated from a random subset consisting of 10,000 random trees out of the total number of possible trees. To test the validity of estimating the skewness of a large distribution from a subsample of 10,000 random trees, we performed exhaustive searches on two random data sets of 100 characters and 10 and 11 taxa, respectively. Each distribution was then sampled 20 times using PAUP's random trees option for 10, 10^2, 10^3, 10^4, and 10^5 random trees. Figure 1 shows the real g_1 of the distribution and the average, standard deviation, and standard error of the mean for the random-sampling experiments. As more trees are randomly sampled from the tree-length distribution, the estimate of g_1 improves. Furthermore, the number of taxa in the matrix being sampled does not seem to have a major effect on the accuracy of

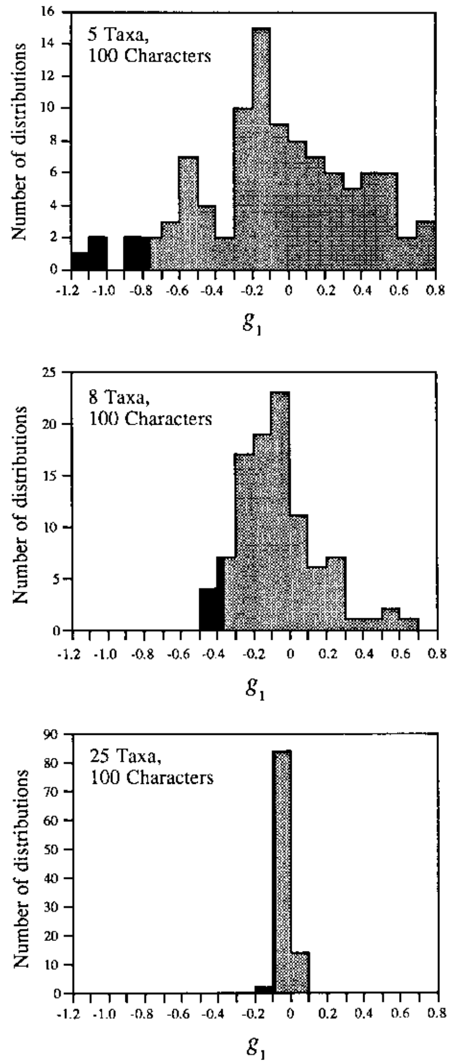

Figure 2. Estimation of 95% confidence limits from random data matrices. For each of three sets of 100 random data matrices (100 characters, and five, eight, and 25 taxa, respectively), the actual lower 5% of the distribution is shaded. The upper limit of this lower 5% of the distribution can be used as a critical value in a test to determine if a given data matrix is more structured than would be expected from random data.

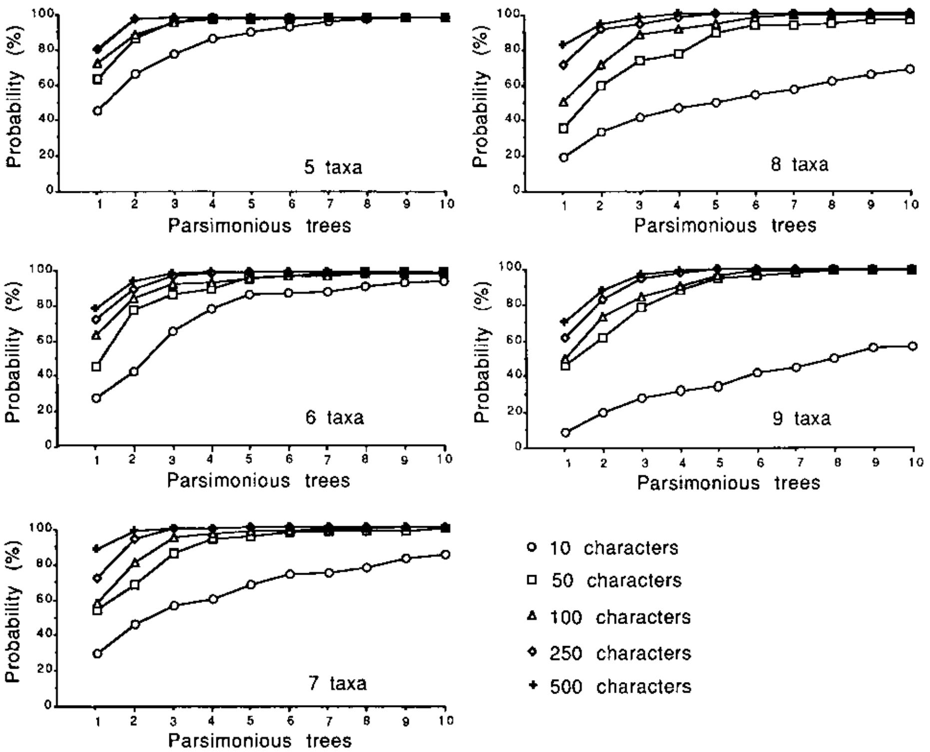

Figure 3. The cumulative probability of finding a given number of most-parsimonious trees for 10–500 random binary characters and five–nine taxa (based on 100 random data matrices for each combination of taxa and characters).

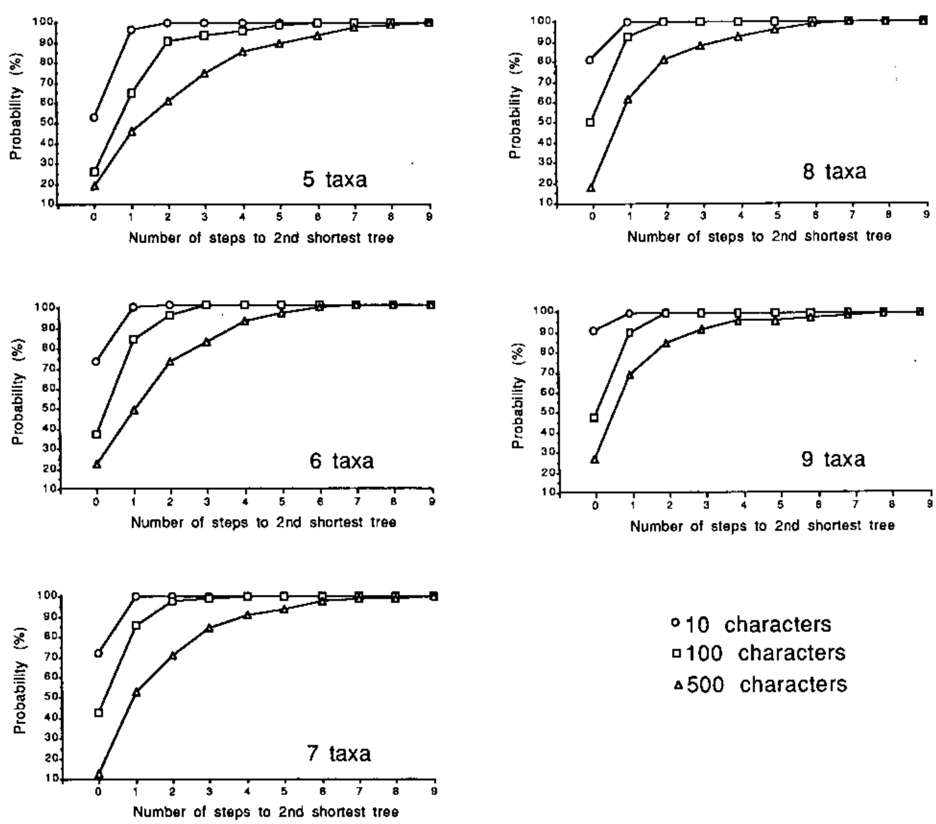

Figure 4. The cumulative probability of finding a second solution within a given number of steps from the most-parsimonious tree with 10–500 random binary characters and five–nine taxa.

g_1 estimates. A random sample of 10^4 trees, as used in this study, appears to provide an accurate estimate of g_1 (to within 0.05 of the true value in all replicates).

We determined critical values by the actual 95% and 99% limits of the distributions of g_1 produced from random data sets (see Figure 2). This direct calculation of confidence limits avoids making any assumptions about the underlying distributions, other than the assumption that the 100 replicates of each combination of characters and taxa are representative of those underlying distributions.

Number of Most-Parsimonious Trees

Finding a single solution to a problem with a given data set is often viewed as an indication that the data set has considerable resolving power. In phylogenetic analysis, one often finds many most-parsimonious trees for a given data set, particularly if the number of included taxa is great. Therefore, when only a single solution is found, it is intuitive that the data set must contain strong signal. However, this intuition is wrong. In Figure 3, the relationship between the probability of finding various numbers of phylogenetic trees is shown for five–nine taxa and 10–500 characters. If a single most-parsimonious tree is found with just 10 characters and nine taxa (for which there are 135,135 possible unrooted trees), one could still not reject the null hypothesis that the data were random with respect to phylogenetic history at $P < .05$, because this result occurs with random data about 10% of the time. With data sets in the size range typical for molecular studies, the most likely outcome is to find a single solution, even if the data are random. With nine taxa and 50 characters, a single solution is expected with random data nearly 50% of the time, and with 500 characters, over 70% of the time (Figure 3). The probability of finding a single solution with random data will drop with increasing number of taxa, but Figure 3 indicates that the number of optimal solutions found is a poor indicator of data quality in phylogenetic analysis.

Number of Steps to the Second-Best Solution

If finding a single solution is not unexpected even with random data, then what about the number of steps to the next-best solution? Several authors have used this measure to indicate the relative strength of their results in phylogenetic analysis (e.g., Hillis and Dixon 1989; Miyamoto and Boyle 1989). The probability of finding a given distance between the shortest and second-shortest trees is shown in Figure 4 for five–nine taxa and 10-500 random characters. With just 10 random characters, it is uncommon to find more than one step between the two best solutions, but with data sets more typical of molecular studies, it is not. With 500 characters, one could not reject the null hypothesis of random data (at $P < .05$) even if the second-best tree of nine taxa was six steps longer than the most-parsimonious tree (Figure 4). For five taxa and 500 random characters, one expects to see a difference of at least three steps between the best two solutions over 25% of the time (Figure 4). Thus, the number of steps between near-optimal solutions does not provide a very powerful test for discriminating between signal and random noise.

Skewness of Tree-Length Distributions
Several authors have suggested that the shape of a tree-length distribution provides a good indication of the presence of phylogenetic signal in a data set (Fitch 1984; Hillis 1985, 1991; Hillis and Dixon 1989; Huelsenbeck 1991; Le Quesne 1989). Distributions of tree lengths with a strong left skew, like that shown in Figure 5a, indicate that relatively few solutions exist near the optimal solution compared to elsewhere in the distribution. This, in turn, is an indication of correlation among characters beyond that expected at random. Characters may be correlated for reasons other than a common history; for instance, the assumption that the characters are independent may have been violated. Nonetheless, if the data do not show structure above that expected at random (e.g., the distribution shown in Figure 5b), there is little reason to expect that phylogenetic analysis will reveal information about historical relationships.

Hillis (1991) calculated critical values of g_1 for tree-length distributions of six, seven, and eight taxa represented by 100-bp random DNA sequences. He showed that small amounts of phylogenetic signal ($\leq$10%) were usually sufficient to skew distributions based on otherwise random data beyond the 95% confidence limits of the random distributions. Huelsenbeck (1991) simulated 300 eight-taxon phylogenies with varying rates of mutation, from 0.1% to 75% change between nodes, to test the relationship between skewness of the tree-length distribution and the performance of parsimony in finding the true phylogeny (Figure 6). For DNA sequences, a 75% mutation rate between nodes of a

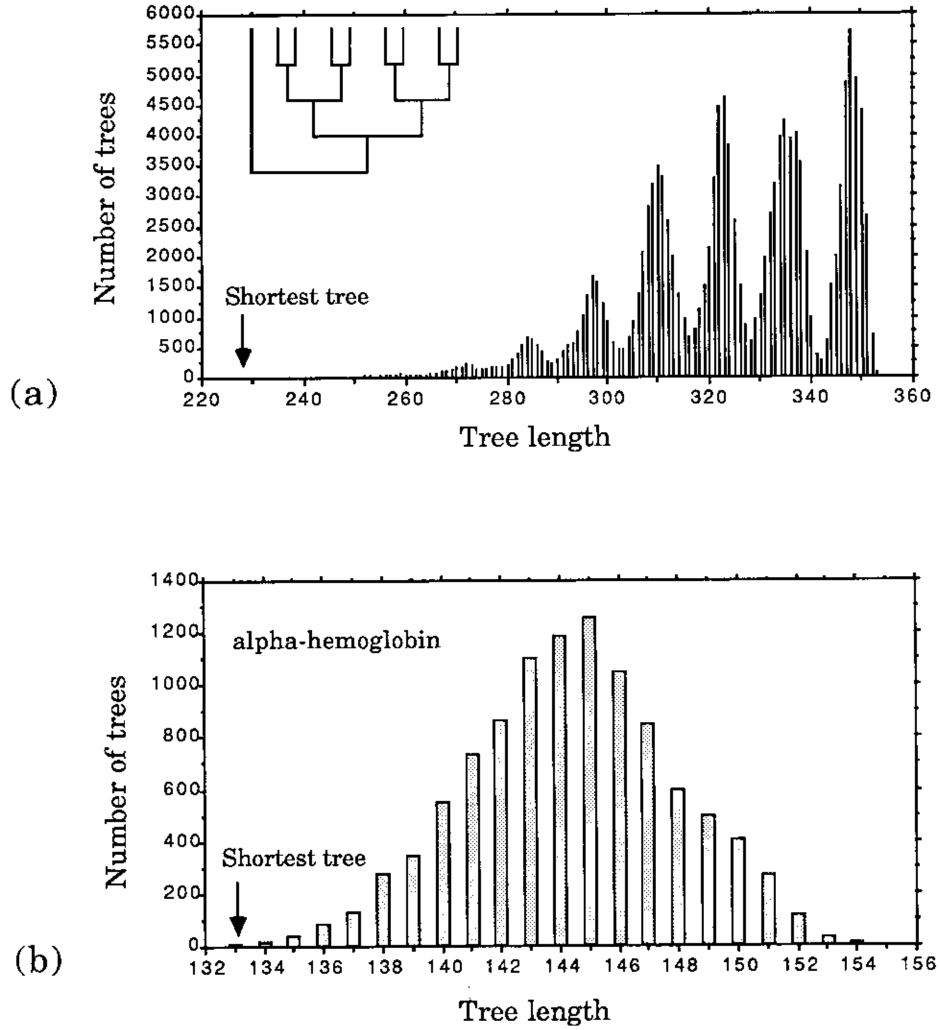

Figure 5. Examples of tree-length distributions: (a) Tree-length distribution from a significantly structured data set (from a laboratory-produced phylogeny of T7 bacteriophage, shown in inset; data from Hillis et al., 1992); (b) Tree-length distribution from a data set that is no more structured than random data (data from Fitch 1984).

Table 1. Critical values of g_1 for binary character data; data sets that produce g_1 values less than those shown (i.e., more negative) are significantly more structured than are the random data

No. of charac-ters	No. of taxa 5		6		7		8	
	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$
10	−1.12	−1.30	−0.75	−1.02	−0.63	−0.84	−0.37	−0.67
50	−0.88	−1.08	−0.67	−0.88	−0.39	−0.63	−0.37	−0.49
100	−0.77	−1.08	−0.59	−0.68	−0.37	−0.46	−0.37	−0.43
250	−0.94	−1.20	−0.74	−1.12	−0.37	−0.49	−0.33	−0.44
500	−0.60	−0.84	−0.53	−0.63	−0.35	−0.46	−0.31	−0.47

Table 2. Critical values of g_1 for four-state character data; data sets that produce g_1 values less than those shown (i.e., more negative) are significantly more structured than are the random data

No. of charac-ters	No. of taxa 5		6		7		8	
	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$
10	−0.95	−1.28	−0.70	−0.79	−0.59	−0.73	−0.51	−0.64
50	−0.78	−0.88	−0.58	−0.70	−0.45	−0.64	−0.37	−0.42
100	−0.66	−0.93	−0.56	−0.69	−0.40	−0.48	−0.31	−0.42
250	−0.81	−0.97	−0.43	−0.59	−0.39	−0.45	−0.26	−0.37
500	−0.73	−0.95	−0.43	−0.54	−0.27	−0.49	−0.29	−0.33

tree randomizes sequences with respect to history, because each nucleotide has an equal probability of occurring at a given position in the data matrix. As can be seen from Figure 6, the true phylogeny is likely to be represented by the most-parsimonious tree (or a nearly most-parsimonious tree) if the tree-length distribution is significantly more skewed than expected from random data. However, if the tree-length distribution is not significantly skewed (because the characters are nearly randomized with respect to history), there is little basis for expecting the most-parsimonious tree to represent the true phylogeny. In fact, under these conditions, the true phylogeny can even be very close to the least-parsimonious tree (Figure 6). Of course, one usually does not know the true phylogeny a priori, but g_1 can be calculated with ease (it is now calculated automatically for the exhaustive search and random-trees options in PAUP) and is routinely obtained in the course of phylogenetic analysis, so no additional analyses (beyond those usually conducted) are necessary to test for the presence of structured data.

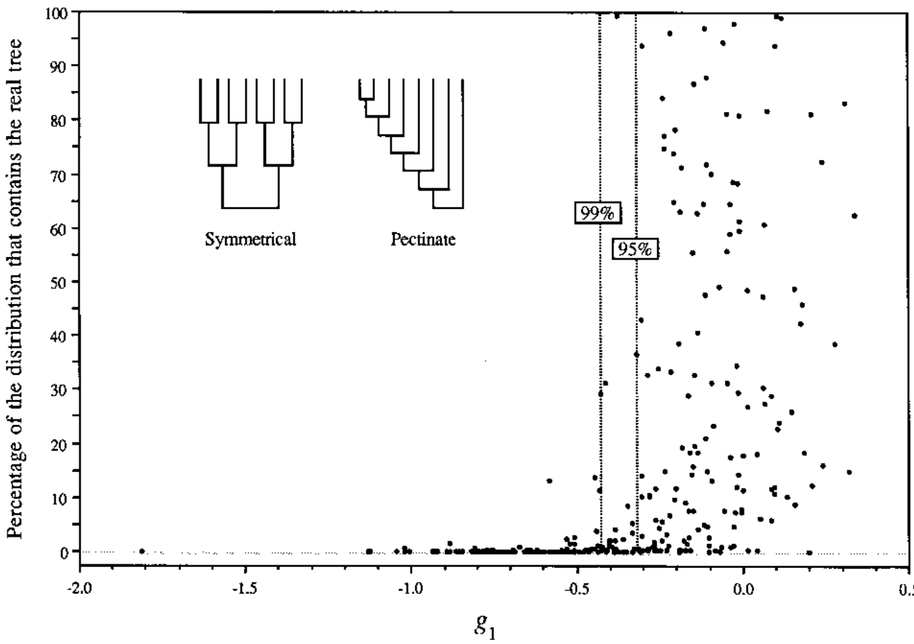

Figure 6. The relationship between g_1 and the performance of parsimony in finding the true phylogeny, based on 300 simulations of the two phylogenies figured, with rates of change between nodes of 0.1%–75% (see Huelsenbeck 1991). When distributions are significantly more skewed than expected from random data (i.e., scores to the left of the 95% or 99% confidence limits for random matrices), the performance of parsimony in finding the true phylogeny is excellent. When tree-length distributions are no more skewed than those obtained from random data (to the right of the confidence limits), the true phylogeny is often nowhere near the most-parsimonious solution.

Table 1. Extended

No. of taxa

9		10		15		25	
$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$
−0.48	−0.56	−0.44	−0.59	−0.28	−0.37	−0.18	−0.24
−0.31	−0.44	−0.35	−0.39	−0.16	−0.20	−0.10	−0.11
−0.33	−0.43	−0.26	−0.31	−0.15	−0.19	−0.09	−0.10
−0.29	−0.44	−0.22	−0.35	−0.15	−0.20	−0.08	−0.09
−0.29	−0.47	−0.20	−0.27	−0.10	−0.15	−0.08	−0.08

Table 2. Extended

No. of taxa

9		10		15		25	
$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$	$P = .05$	$P = .01$
−0.44	−0.51	−0.34	−0.42	−0.23	−0.25	−0.16	−0.18
−0.25	−0.30	−0.28	−0.36	−0.16	−0.19	−0.12	−0.13
−0.25	−0.33	−0.30	−0.33	−0.15	−0.20	−0.10	−0.12
−0.22	−0.30	−0.20	−0.27	−0.14	−0.16	−0.08	−0.09
−0.23	−0.30	−0.16	−0.27	−0.12	−0.15	−0.07	−0.09

Critical Values of g_1

In order to test data sets for nonrandom structure, one must take into account the number of taxa and the number of characters in the data matrix. We also explored the effects of number of states of the characters, from binary to four-state characters. This range encompasses that typical for most molecular studies. For DNA sequences, the four-state characters represent sequences with all four bases at equal frequencies, whereas the binary data represent the extremes of base-compositional or transition/transversion biases. Figure 7 shows that the number of character states (within the range examined) has little effect on the critical values of g_1, so it rarely is necessary to take base composition or transition/transversion bias into account when testing for structured data.

Table 1 (binary characters) and Table 2 (four-state characters) present the 95% and 99% critical values of g_1 for 5–25 taxa and 10–500 characters, each based on 100 replicates of random data. The critical values change very little beyond 15 taxa (Figure 8), so the values for 25 taxa can be used in a slightly conservative test for greater numbers of taxa with very little loss of power to discern structure. Note that the number of characters in the test is the number of variable positions (i.e., excluding invariant sites), rather than the total length of the sequence examined.

Tests with Real Sequences

Hillis (1991) examined several real data sets based on molecular sequences and noted that all of them, except one that consisted of α-hemoglobin sequences of mammals, showed more structure than expected at random. However, he also noted that the structure is not necessarily distributed throughout the branches of the estimated tree; a particular data set may contain historical information about one clade and none on another. To demonstrate this effect, consider a data set of mitochondrial cytochrome b sequences

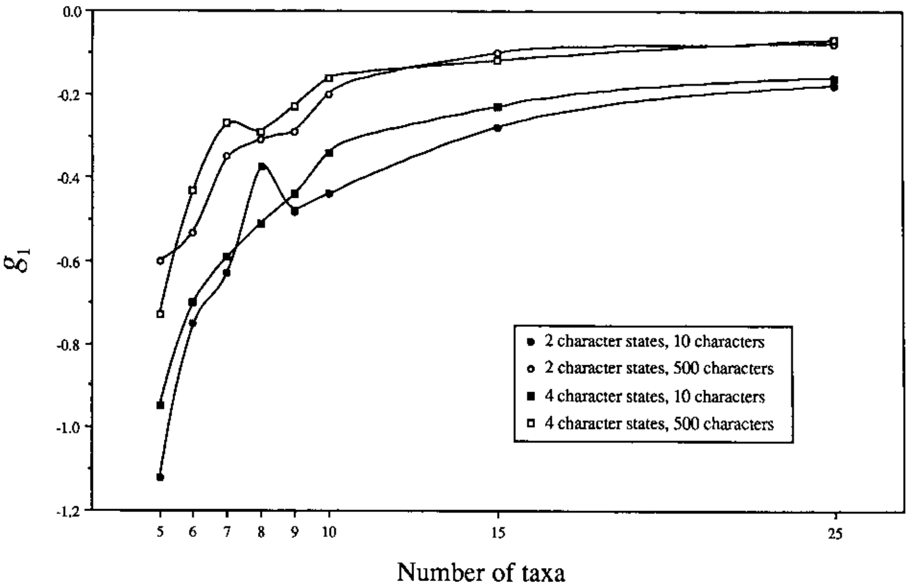

Figure 7. The effect of number of character states on critical values (95% confidence limits) of g_1. As the number of character states increases, the critical values tend to move closer to zero. Therefore, critical values for four-state characters can be used in a slightly conservative test for characters with more than four states.

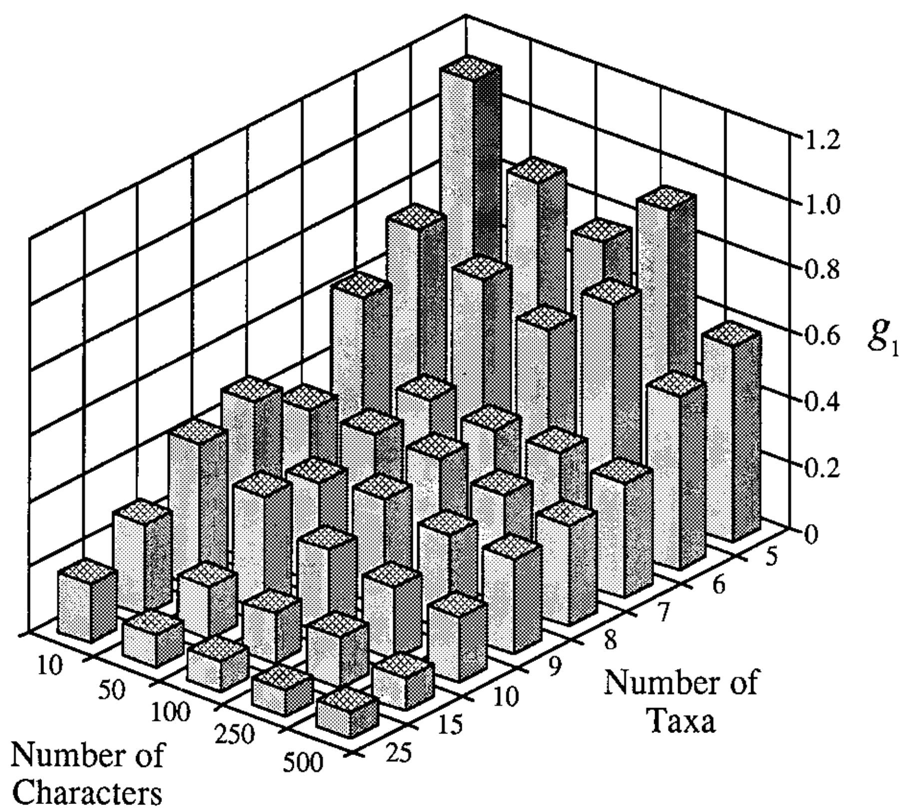

Figure 8. The relationship between critical values (95% confidence limits) of g_1 and size of the data set, from 5–25 taxa and 10–500 random binary characters.

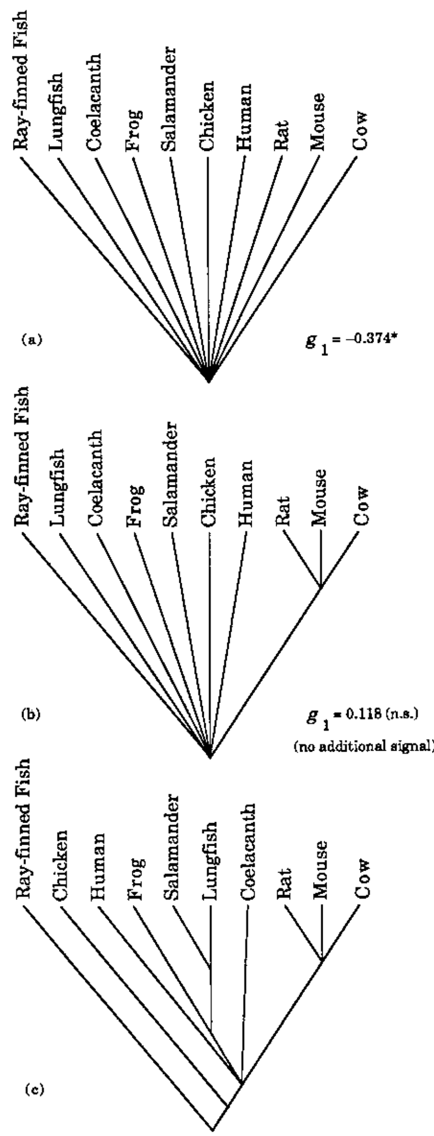

Figure 9. An example of testing a set of sequences for significant structure: (**a**) The tree-length distribution estimated from 100,000 random trees of all the taxa suggests signal is present ($P < .01$; number of variable positions = 126); (**b**) The most strongly supported branch unites rat, mouse, and cow. If the distribution of all the remaining trees is examined, it is no longer different from a distribution based on random data (in fact, it is slightly skewed to the right); (**c**) If the most-parsimonious trees are found for the sequences in spite of the apparent lack of signal, the resulting tree makes little sense compared to current ideas of vertebrate phylogeny (the consensus of the three most parsimonious trees is shown; the ray-finned fish was used as the outgroup). Cytochrome *b* sequences in this example are from GenBank and Meyer and Wilson (1990).

from 10 species of vertebrates (Figure 9). The tree-length distribution for all 10 species is significantly skewed ($g_1 = -0.374$, number of variable positions = 126, $P < .01$; Figure 9a). The best-resolved internal branch of the most-parsimonious tree from this data set (by number of supporting characters or bootstrapping) is the one that unites the two rodents and the cow together. One can now ask what the tree-length distribution looks like if these species are held together as a clade (Figure 9b); in other words, if we accept this grouping, is there still significant structure in the data set? The tree-length distribution of the constrained analysis is no longer skewed to the left ($g_1 = 0.118$). Therefore, virtually all the signal in this set of DNA sequences appears to be accounted for by this one clade, and there is little justification for resolving the tree any further. If we do so anyway (Figure 9c), none of the remaining branches is strongly supported, and the consensus of the three most-parsimonious trees bears little resemblance to traditional ideas about vertebrate phylogeny. However, the tree-length distribution indicates there is little basis for expecting any reasonable resolution at this point because the data appear to be no more structured than random sequences with respect to the remaining branches. This analysis indicates that cytochrome *b* genes are evolving much too rapidly to provide information about phylogeny this far back into time, in contrast to the conclusions reached by Meyer and Wilson (1990).

Conclusions

An often unstated assumption in phylogenetic analyses is that the characters being analyzed are evolving at a rate that is appropriate to accurately estimate phylogeny. However, this assumption is often not tested. DNA sequences, in particular, can be aligned with ease among distantly related organisms if some sites are invariant, even if the variable positions are randomized with respect to history (as in the cytochrome *b* example above). Phylogenetic analysis of such sequences is unlikely to reveal correct information about evolutionary history; instead, such analyses contribute to the inaccurate perception of considerable conflict between molecular and morphological studies (Hillis 1987). Tests are needed to distinguish sequences (or regions of sequences) that are significantly more structured than random, to avoid the problem of "reading beyond the signal" in phylogenetic analyses. Examining skewness of tree-length distributions is one approach to this problem that appears to be both fast and effective.

References

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

Fitch WM, 1984. Cladistic and other methods: problems, pitfalls, and potentials. In: Cladistics: perspectives on the reconstruction of evolutionary history (Duncan T and Stuessy TF, eds). New York: Columbia University Press; 221–252.

Hillis DM, 1985. Evolutionary genetics of the Andean lizard genus *Pholidobolus* (Sauria: Gymnophthalmidae): phylogeny, biogeography, and a comparison of tree construction techniques. Syst Zool 34:109–126.

Hillis DM, 1987. Molecular versus morphological approaches to systematics. Annu Rev Ecol Syst 18:23–42.

Hillis DM, 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. In: Phylogenetic analysis of DNA sequences (Miyamoto MM and Cracraft J, eds). New York: Oxford University Press; 278–294.

Hillis DM, Bull JJ, White ME, Badgett M, and Molineux IJ, 1992. Experimental phylogenetics: generation of a known phylogeny. Science 255:589–592.

Hillis DM and Dixon MT, 1989. Vertebrate phylogeny: evidence from 28S ribosomal DNA sequences. In: The hierarchy of life: molecules and morphology in phylogenetic analysis (Fernholm B, Bremer K, and Jörnvall H, eds). Amsterdam: Elsevier; 355–367.

Huelsenbeck JP, 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst Zool 40:257–270.

Le Quesne WJ, 1989. Frequency distributions of lengths of possible networks from a data matrix. Cladistics 5: 395–407.

Meyer A and Wilson AC, 1990. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J Mol Evol 31:359–364.

Miyamoto MM and Boyle SM, 1989. The potential importance of mitochondrial DNA sequence data to eutherian mammal phylogeny. In: The hierarchy of life: molecules and morphology in phylogenetic analysis (Fernholm B, Bremer K, and Jörnvall H, eds). Amsterdam: Elsevier; 437–450.

Sokal RR and Rohlf FJ, 1981. Biometry, 2nd ed. San Francisco: W. H. Freeman.

Swofford DL, 1990. PAUP: Phylogenetic Analysis Using Parsimony. Champaign: Illinois Natural History Survey.

Swofford DL and Olsen GJ, 1990. Phylogeny reconstruction. In: Molecular systematics (Hillis DM and Moritz C, eds). Sunderland, Massachusetts: Sinauer; 411–501.