PHYLOGENETIC ANALYSIS OF DNA SEQUENCES

Edited by

Michael M. Miyamoto
University of Florida
Gainesville, Florida

Joel Cracraft
University of Illinois
Chicago, Illinois

New York *Oxford*
OXFORD UNIVERSITY PRESS
1991

13

Discriminating Between Phylogenetic Signal and Random Noise in DNA Sequences

DAVID M. HILLIS

> Most parsimonious trees are really neat. I am all for them, but have you any idea about the distribution of all the trees in the universe that you might have sampled? Fitch (1984)

Nucleic acid sequences have become a major source of phylogenetic information. In order to use these data appropriately, it is critical to distinguish sequences that are saturated by change from those that are phylogenetically informative. In this chapter, I will call variation that is potentially informative about phylogenetic history "signal," and use "noise" to describe variation that is not informative. Distinguishing between signal and noise is not something that can be accomplished by use of alignment criteria, because two sequences may be identical at a large number of functionally constrained sites, and yet essentially randomized at sites where variation is tolerated (e.g., the third positions of codons). This chapter explores the use of the distributions of tree lengths as a guide to the detection of phylogenetic signal in comparative data sets.

In parsimony analysis, changes at nucleotide positions among aligned sequences are mapped onto a tree, and the number of evolutionary changes required to accommodate that tree with the data is calculated as the tree length. For any given data set, this procedure may be repeated for many thousands of trees; the tree that yields the shortest length is designated the optimal tree under the parsimony criterion. The optimal tree is thus the one that requires the fewest number of evolutionary changes. In this chapter, I argue that the shape of the distribution of tree lengths contains information useful in deciding whether or not the data set contains phylogenetic signal.

BACKGROUND

Fitch (1979) and Goodman et al. (1979a) were among the first authors to explore tree-length distributions in a phylogenetic context. These authors compared the relative phylogenetic information in two protein sequence data sets that had been used to infer relationships among orders of mammals. As can be seen in Figure 13-1, the distribution of tree lengths from the α-hemoglobin sequences is nearly perfectly symmetrical, whereas that from the α-crystallin sequences is strongly skewed with a long left tail. Therefore, there are many fewer solutions near the optimal (most-parsimonious) solution for the α-crystallin data set than for the α-hemoglobin data set. The shape of these tree-length distributions suggests that the α-

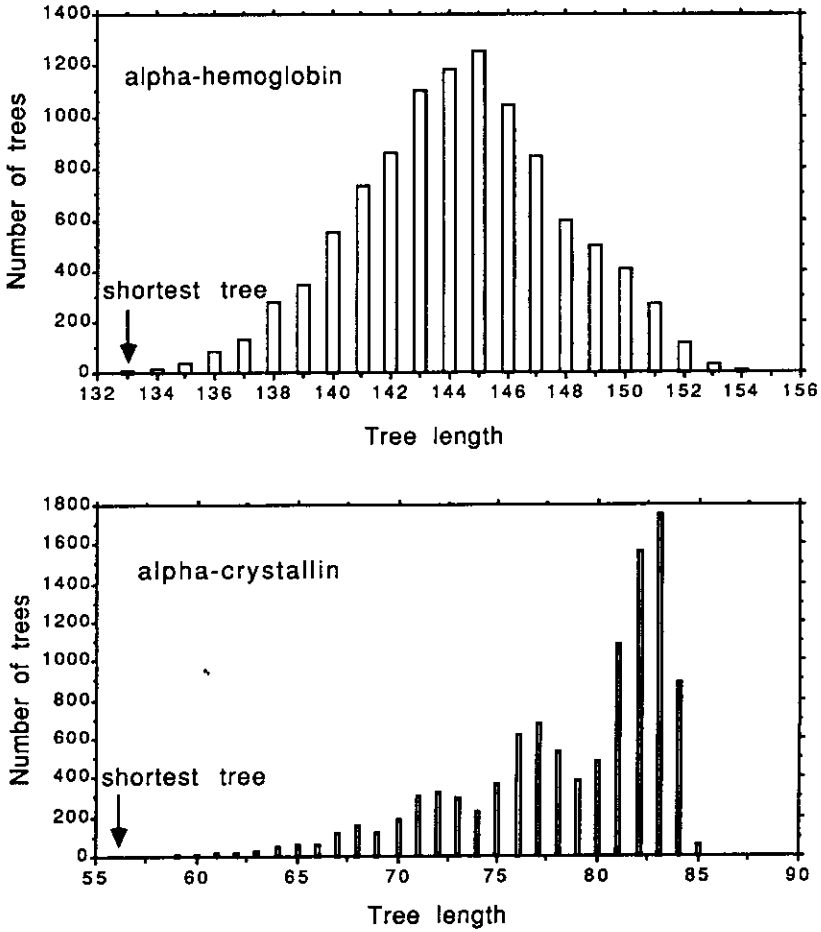


Figure 13-1 Tree-length distributions compared by Fitch (1979, 1984) and Goodman et al. (1979a).

crystallin data set is more informative (because it allows better discrimination among near-optimal solutions) than the α-hemoglobin data set.

Fitch (1984) suggested that symmetric distributions such as that for α-hemoglobin are "most likely for bushy trees where all lines emanate from a single point," whereas asymmetric distributions like that for α-crystallin are "most likely for the 'stringy' trees, those which are maximally asymmetric in their splitting and have long intervals between splits." Several authors have presented strongly skewed tree-length distributions and have suggested that such distributions are evidence for strong phylogenetic signal (e.g., Fitch, 1984; Goodman et al., 1979a; Hillis, 1985; Werman, 1986; Hillis and de Sá, 1988; Hillis and Dixon, 1989). I will argue that skewness of tree-length distributions can be a useful indicator of phylogenetic signal and is largely (but not completely) independent of tree topology.

Having looked at tree-length distributions for many data sets, I developed the view illustrated in Figure 13-2. If all tree topologies for a given set of sequences are equally optimal (i.e., of equal length), then there is neither signal nor noise in the data set. Such a data set would result in the completely unresolved bush shown in Figure 13-2a. If the real phylogeny is in the form of a bush, however, then an equal length for all possible trees is an unlikely outcome; random variation would result in some trees being shorter than others by chance alone. The distribution of trees would include some topologies that were shorter than average, and a similar number that were longer than average, so the distribution of these lengths would produce a nearly symmetrical distribution, such as those shown in Figure 13-2b and 13-2b′. The optimal topology (or topologies) in this case might be a completely symmetrical tree (Fig. 13-2b), a completely asymmetrical tree (Fig. 13-2b′), or something in between. However, if the true phylogeny was not a single polytomy and the sequences were constrained by history (i.e., contain phylogenetic information), then the tree-length distributions shown in Figures 13-2b and 13-2b′ are highly unlikely. Sequences constrained by history would produce distributions of tree lengths similar to those in Figure 13-2c and 13-2c′: highly asymmetrical distributions with few trees near the optimal solution. These asymmetric distributions are a consequence of congruence among characters as a result of a common phylogenetic history. Thus, many characters support the topology that reflects this history, and conflict with a large number of alternative trees.

If the view described above is correct, then it suggests a means of escaping a common problem in phylogenetic analysis. Any comparative data set can be subjected to phylogenetic analysis, even if the data contain no historical information (i.e., they are too noisy for meaningful phylogenetic analysis). Nonetheless, random variation will usually result in one or a relatively few optimal topologies (at least compared to the much larger number of possibilities). What is needed is a means of distinguishing such data sets from those that are informative about phylogenetic history. This chapter explores the causes of skewed tree-length distributions, examines the proposition

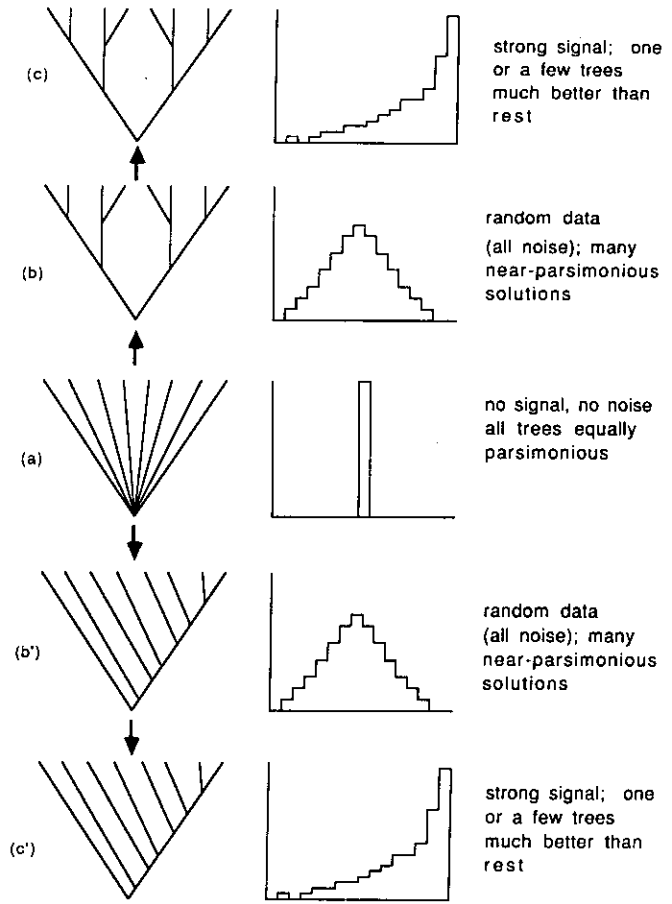


Figure 13-2 Hypothesis of the behavior of tree-length distributions under several model phylogenies with differing amounts of signal and noise. See text for explanation.

that random sequence data produce symmetrical tree-length distributions, explores possible tests for skewness, evaluates the amount of signal necessary to produce significant skewness, compares real data sets to random simulations, suggests a possible "stopping algorithm" to prevent the over-resolution of phylogenies, and points to areas in need of additional research.

THE CAUSES OF SKEWED TREE-LENGTH DISTRIBUTIONS

To understand why data sets with strong phylogenetic signal produce skewed tree-length distributions, consider a data set on eight taxa (Figs. 13-3–13-

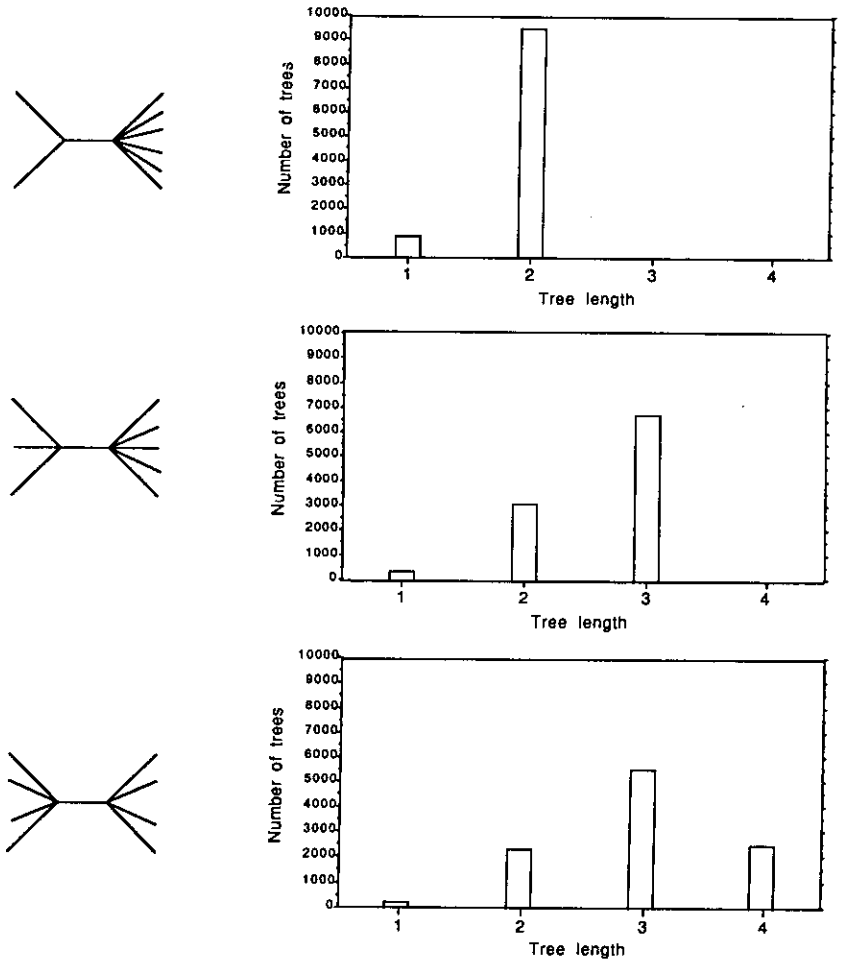


Figure 13-3 Classes of binary characters for eight taxa and their effects on tree-length distributions.

5).[1] For eight taxa, there are only three types of binary characters that affect tree length (Fig. 13-3): those for which two taxa have one state and the others have the alternative state ("two/six characters"); those for which the alternative states are distributed between three and five taxa, respectively ("three/five characters"); and those for which the alternative states are evenly divided between two subsets of four taxa each ("four/four characters"). Characters with states limited to a single taxon and characters that are uniform among all species are compatible with all possible tree topologies and thus do not affect the shape of the tree-length distribution.

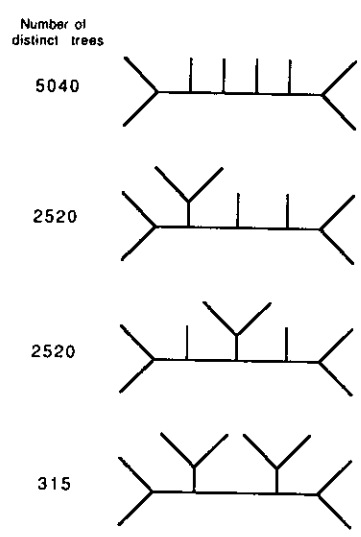
[1]Because tree lengths are unaffected by the rooting point of the tree, all examples in this discussion will concern unrooted trees. Also, for simplicity, examples in this section will involve ditypic (binary) characters, although the concepts can be extended easily to multistate characters.

Figure 13-3 shows the distributions of tree lengths for data sets limited to each of these types of characters. All three types of characters produce a tree-length distribution with a left-hand skew. Two/six characters produce distributions with the strongest skew but the smallest range; this is because only a very few trees connect the pair of species with the unique state, whereas all other trees require the character-state to evolve twice. Three/five characters also produce a skewed distribution, but more of the possible trees are compatible or partially compatible with this type of character. Although few trees unite the three species with the same state, many more unite at least two of the three species on the tree, and most trees split all three species. This produces a skewed distribution with a range of two steps (compared to a range of one step for a two/six character; see Fig. 13-3). The four/four characters produce the distribution with the greatest span but the least skewness, because there are fewer trees that unite none of the similar taxa than unite only two of the four similar taxa (Fig. 13-3). Therefore, any informative character can add to the skewness of a tree-length distribution, although different types of characters do not contribute to skewness equally.

The effects of tree topology on tree-length distributions can be assessed by considering the kinds of characters that would be produced by the different topologies. For eight taxa, there are only four unrooted, unlabeled topologies (Fig. 13-4). Each internal branch in the topology is supported by either a two/six, a three/five, or a four/four character; the numbers of each of these types of internal branches are shown in Figure 13-4. If one assigns two characters to each internal branch, the distributions for each topology would be as shown in Figure 13-5. Note that all of these distributions are strongly skewed, although not all to the same degree. A common measure of skewness is the g_1 statistic—the third central moment divided by the cube of the standard deviation (Sokal and Rohlf, 1981). For n trees of length T, g_1 is calculated as

$$g_1 = \frac{\sum_{i=1}^{n} (T_i - \bar{T})^3}{n\, s^3}$$

where s is the standard deviation of the tree lengths. This statistic is negative for distributions with left-skew, 0 for symmetric distributions, and positive for distributions with right-skew. The skewness of the distributions in Figure 13-5 varies from $g_1 = -1.096$ to $g_1 = -0.6637$. Therefore, although tree topology has a quantitative influence on skewness, the qualitative result (strong left-skew) is the same for all topologies. It is interesting to note that symmetry of the tree topology appears to have little influence on skewness.



Number of distinct trees	Type of internal branch 4/4	3/5	2/6	Relative span of tree lengths
5040	1	2	2	9
2520	1	1	3	8
2520	0	2	3	7
315	1	0	4	7
Total: 10,395				

Figure 13-4 Distinct unrooted tree topologies for eight taxa, numbers of internal branches supported by each class of binary character, and the relative spans of the tree-length distributions.

EXPERIMENTS WITH RANDOM SEQUENCES

To address the behavior of tree-length distributions in the absence of phylogenetic signal, random data matrices were constructed in the following manner (Fig. 13-6). One hundred matrices each with six, seven, or eight sequences, each 100 nucleotides long, were created using the random data generator in the computer program MacClade (version 2.97; Maddison and Maddison, 1991). This version of the program uses the random number generator in the Macintosh tool box to generate random data (D. Maddison, personal communication). The four nucleotides were given equal probabilities of occurring, so the frequency of each nucleotide was approximately 0.25 in each sequence. All possible tree topologies were analyzed for all 300 random matrices using Swofford's (1990) Phylogenetic Analysis Using Parsimony computer software package (PAUP version 3.0).

To address the effects of the length of sequences on skewness, two additional sets of 100 matrices each of random data were generated for eight taxa. These data sets consisted of sequences 30 positions long and 200 positions long, respectively.

Two examples of tree-length distributions from the eight-taxa data sets are shown in Figure 13-7; figured are the distributions with the strongest left-hand and right-hand skews, respectively. The distribution of g_1 scores for the tree-length distributions from all the random data sets are shown

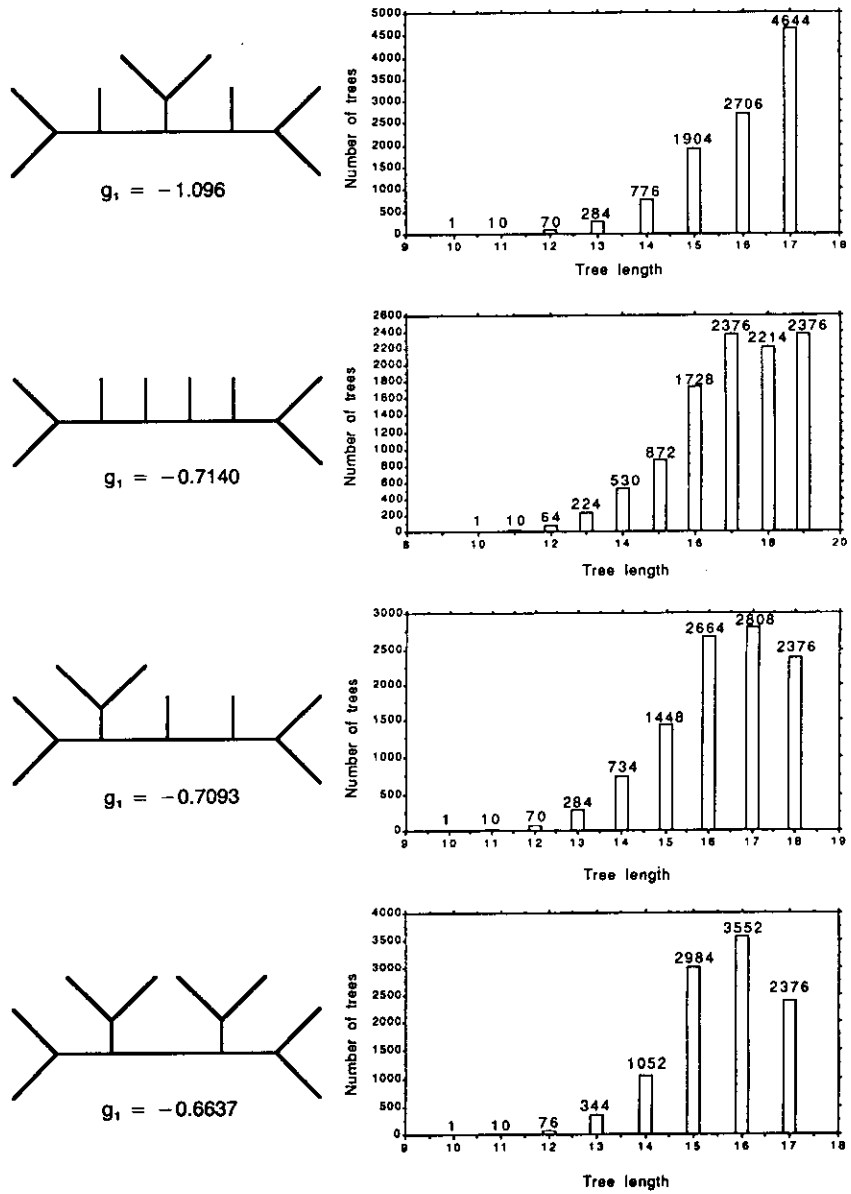


Figure 13-5 Effect of tree topology on tree-length distributions. Each internal branch is assigned a length of two character changes.

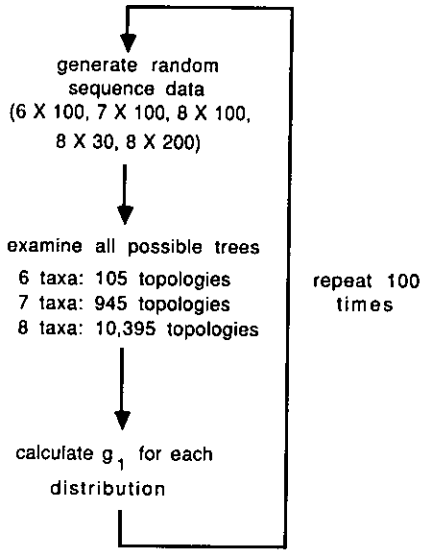


Figure 13-6 Protocol for the generation of random sequence-data matrices and analysis of skewness.

in Figure 13-8. Two trends are clear from these scores. First, the mean and modal values of g_1 across data sets are slightly negative, as predicted for binary data by Le Quesne (1989). Even with random data, approximately twice as many tree-length distributions are skewed slightly to the left as to the right. Second, the average value of the g_1 scores approaches 0 with increasing number of taxa: the tree-length distributions become closer to symmetrical.

Although the only published tests for skewness of tree-length distributions have used the normal distribution as a null model (e.g., Werman, 1986), the random data simulations indicate that this is an inappropriate test (Fig. 13-9). Only a small percentage of tree-length distributions based on random sequences fall into the 95% confidence limits of a normal distribution. The percentage of distributions that fall within the 95% confidence limits for skewness of a normal distribution drops with increasing number of taxa, despite the fact that the distributions become more symmetrical (Fig. 13-9). This is because the degree of expected departures from symmetry for a normal distribution decreases with increasing sample size, and the sample size (number of tree topologies) increases very rapidly with an increase in the number of taxa (Felsenstein, 1978). Thus, less than 20% of the tree-length distributions produced from random sequences fall within the expectations of a normal distribution with just eight taxa (Fig. 13-9).

The random data simulations provide a simple means for testing tree-length distributions for significantly greater left-handed skewness than would be expected from random data, without making any *a priori* assumptions about the shape of the distribution of expected g_1 values. The lower 5% and 1% of the g_1 distributions, shown in Figure 13-8, can be used to obtain

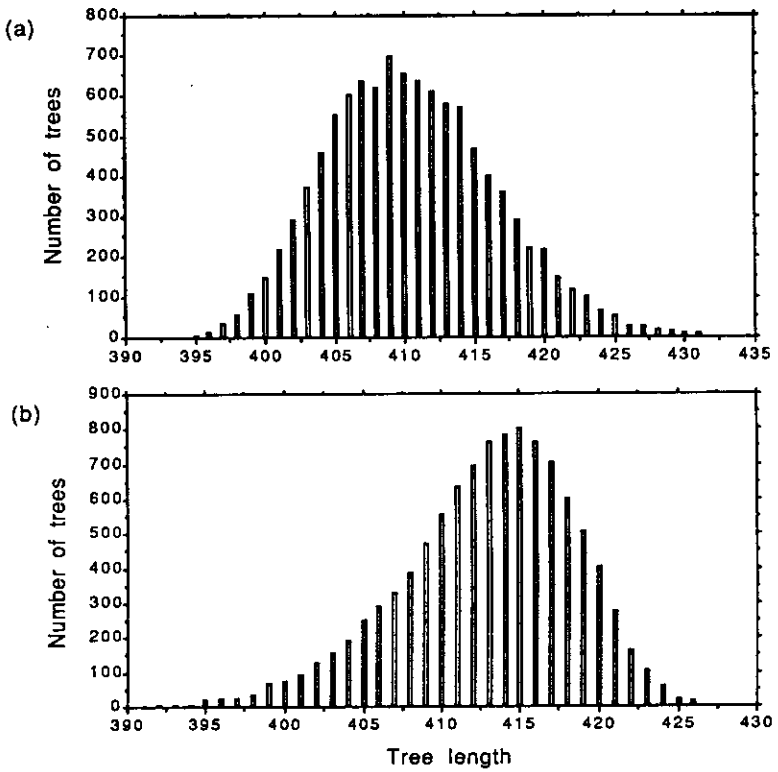


Figure 13-7 Examples of tree-length distributions from analysis of random sequence-data matrices for eight taxa. Shown are the most right-skewed (a; g_1 = 0.3055) and the most left-skewed (b; g_1 = −0.5278) distributions produced from 100 random matrices.

critical values of g_1 for a given number of taxa. A test employing these values (Table 13-1) requires a minimum of 100 variable sequence positions. For fewer than 100 positions, the critical values are slightly lower (more negative); there is very little change in the critical value for data sets with more than 100 sequence positions (Fig. 13-10).

How much phylogenetic signal is needed to produce significant skewness? To examine this question, I added characters consistent with the branches of a single tree (signal) to the eight-taxa random matrices (noise). For most data sets, 10% signal (e.g., 11 characters out of 110) was sufficient to produce tree-length distributions that were significantly skewed, as indicated by the test discussed above (see Fig. 13-11). The most right-skewed distribution examined (g_1 = 0.3055) required 20% signal to become significantly left-skewed (g_1 = −0.4037). Therefore, the test appears to be fairly sensitive; even small percentages of phylogenetic signal can be detected in otherwise random data sets.

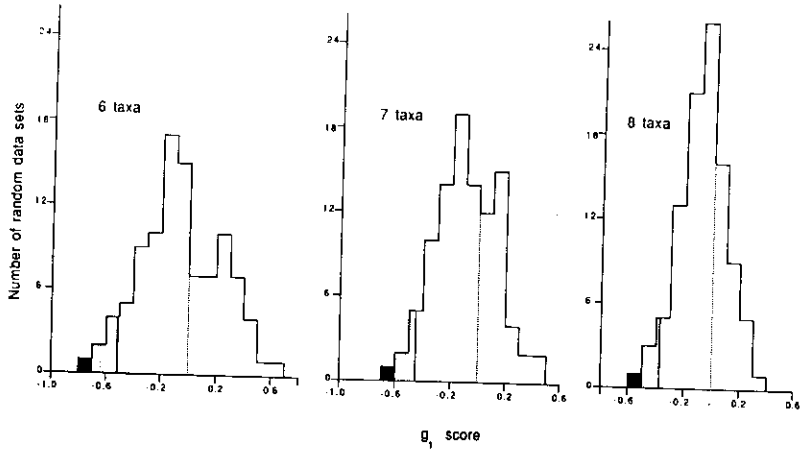


Figure 13-8 The distributions of skewness statistics (g_1) for the 300 random sequence matrices (six, seven, and eight taxa). The dashed vertical lines indicate the g_1 score for symmetrical distributions. The lower 5% (shaded) and 1% (black) of each distribution is indicated.

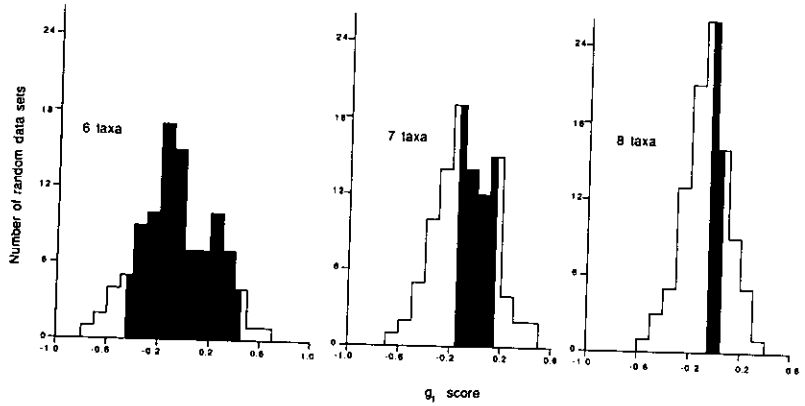


Figure 13-9 The proportions (shaded area) of skewness scores for the random sequence-data matrices that are not significantly different from scores for normal distributions (six, seven, and eight taxa).

Table 13-1 Critical Values for g_1 Statistics of Tree-Length Distributions for Six, Seven, and Eight Taxa*

	Number of Taxa		
p	6	7	8
0.05	−0.51	−0.45	−0.34
0.01	−0.67	−0.60	−0.47

*Scores lower than those shown are outside of the 95% or 99% confidence limits for distributions derived from random data.

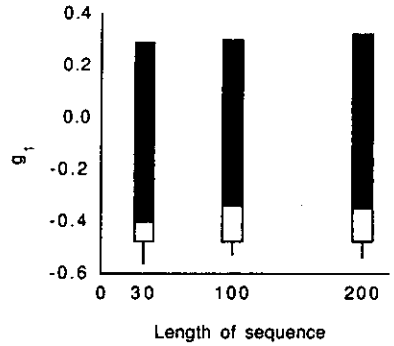


Figure 13-10 The effect of length of sequence on g_1 statistics. The three samples each contain 100 matrices of eight taxa; the lengths of the random sequences are 30, 100, and 200 nucleotides. The lowest 1% of the distribution is indicated by a single line, the lowest 5% of the distribution is indicated by an open bar plus the line, and the remaining portion of the distribution is indicated by a solid bar.

ANALYSIS OF REAL SEQUENCES

I have examined skewness of tree-length distributions from several real sequence data sets, all with at least eight taxa. Except for the α-hemoglobin data discussed by Fitch (1984) and above, all of these were significantly more skewed than the distributions from random data matrices ($p < 0.05$; Table 13-1 and Fig. 13-12). Only one of these distributions (that for mitochondrial cytochrome b genes) was even within the observed bounds of the g_1 values for the random matrices. The mitochondrial cytochrome b data set consists of sequences on eight species of vertebrates, including an actinopterygian fish, two amphibians, a bird, and four mammals. If one of the two most closely related taxa (sheep and cow) are removed from

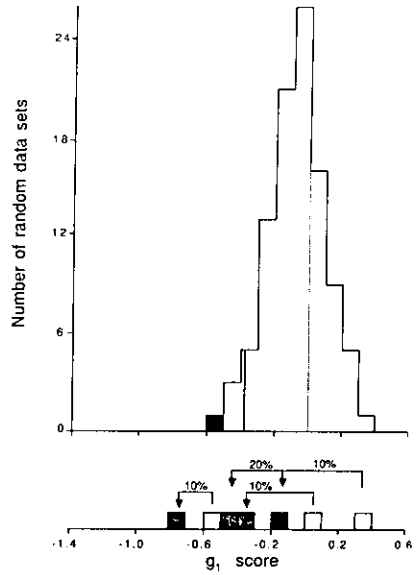


Figure 13-11 The effect on skewness of adding signal to otherwise random sequence matrices (see text). The g_1 scores represented by open boxes in the lower portion of the figure correspond to distributions produced from random data (the most left-skewed, symmetrical, and right-skewed distributions are used as examples). The shaded boxes indicate the g_1 scores for the distributions after the addition of 10% or 20% signal, as indicated.

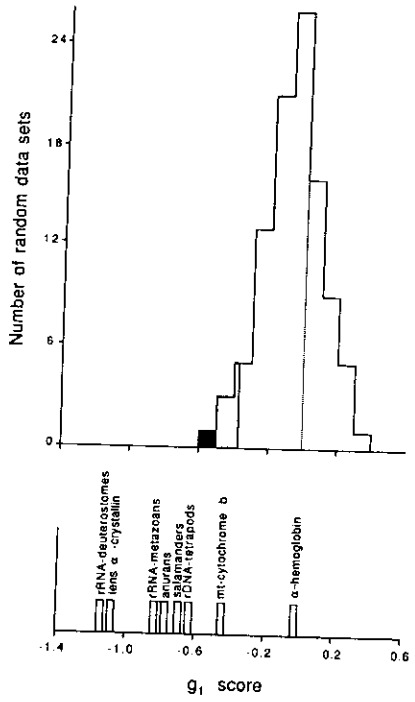


Figure 13-12 Skewness (g_1) scores for tree-length distributions produced from real data matrices (lower) compared to the distributions of g_1 scores produced from analysis of random data matrices (upper) for eight taxa. The rRNA data on deuterostomes (taxa shown in Fig. 13-13) and metazoans (representatives of eight phyla) are from Field et al. (1988); the lens α-crystallin data on eight orders of mammals are from De Jong et al. (1977); the α-hemoglobin data on eight orders of mammals are from Goodman et al. (1979b); the anuran and salamander data are from Hillis (1991); the ribosomal DNA data on tetrapods are from Hillis et al. (1991); and the mitochondrial cytochrome *b* sequences are eight species of vertebrates from Kocher and White (1989) and GenBank (IntelliGenetics, Mountain View, CA).

this data matrix, the tree-length distribution becomes much less skewed, indicating that it is the inclusion of these relatively closely related species that accounts for most of the phylogenetic signal in these sequences.

The observations from the mitochondrial cytochrome *b* data set suggest an algorithm that could be used to prevent over-resolution of phylogenies (reading beyond the signal) in sequence data sets. The execution of this algorithm is shown in Figure 13-13 for the ribosomal RNA (rRNA) sequences presented by Field et al. (1988) for eight species of deuterostomes.

1. Determine the skewness of the tree-length distribution for all possible trees (or a random sample thereof—MacClade and PAUP each have options for this purpose).
2. If the distribution is significantly skewed, then find the best-supported branch that unites two or more of the taxa (e.g., by bootstrapping or counting synapomorphies) and continue to step 3. If the distribution is no more skewed than would be expected from random data, then stop.
3. Repeat step 1 for all remaining tree topologies, given the branch determined in step 2.
4. Repeat step 2, saving any branches previously determined (continue repeating steps 1 and 2 until the distribution of all remaining trees is no more skewed than expected from random data).

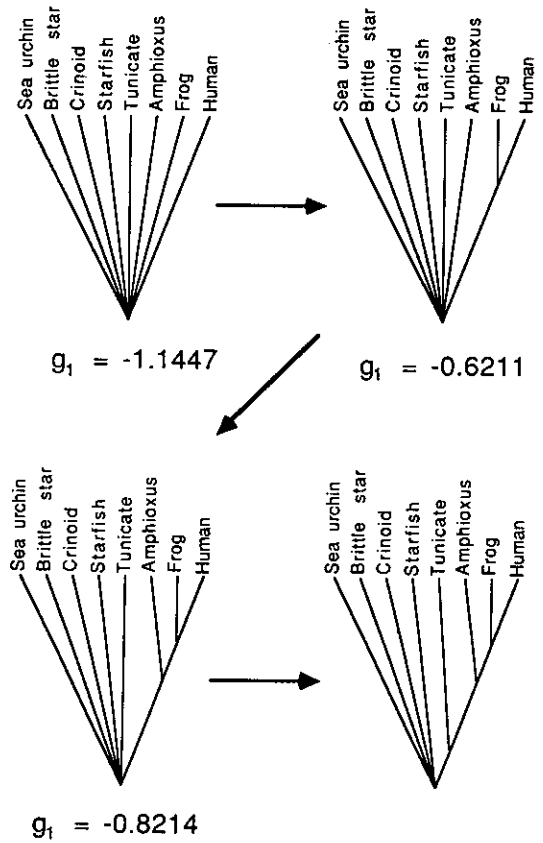


Figure 13-13 Resolution of relationships among eight species of deuterostomes (Field et al., 1988) using the "stopping algorithm" described in the text.

One objection to this algorithm is that tree resolution cannot proceed beyond a tree of five unresolved lineages (Fig. 13-13), because the number of possible tree topologies (15) is too small to examine skewness (see Table 13-2). However, I have found that for most sequence data sets, the tree-length distribution is no longer skewed well before this point; therefore, this limitation may not be too severe. Another objection is that the overall probability of making a type-I error increases as the number of resolved branches increases, so the initial level of α must be set quite low in order to keep the overall α below 0.05. This reduces the power of the test. There also may be objections that a single branch is determined at each step, whereas the phylogenetic information may be distributed across more than one branch. Nonetheless, I present this algorithm as an example of the possible uses of tree-length distributions, although I do not advocate its use until tree-length distributions have been studied in greater detail.

Table 13-2 Number of Distinct Unrooted Trees for Three or More Taxa, and the Recommended Methods of Analyzing Tree-Length Distributions

Number of Taxa	Number of Unrooted Trees	Method of Analysis
3	1	Too few trees
4	3	Too few trees
5	15	Too few trees
6	105	Exact enumeration
7	945	Exact enumeration
8	10,395	Exact enumeration
9	135,135	Exact enumeration or random sampling
10	2,027,025	Exact enumeration or random sampling
11	34,459,425	Exact enumeration or random sampling
≥12	≥654,729,075	Random sampling

IMPLEMENTATION AND DIRECTIONS FOR FUTURE RESEARCH

This study suggests that tree-length distributions can provide an accurate and sensitive indication of the presence of phylogenetic signal in comparative sequence data sets. Skewness statistics provide a simple means of evaluating these distributions, which are often determined as a by-product of phylogenetic analysis. In response to these findings, David L. Swofford and David R. Madison have recently modified PAUP and MacClade, respectively, so that these programs now provide g_1 statistics for tree-length distributions.

Although I have used the conventional measure of skewness (g_1) as a guide to the detection of phylogenetic signal, other measures (e.g., the other odd central moments) may prove to be even better indicators of signal. However, considerable additional research needs to be conducted before any measures of skewness can be used meaningfully with regularity. Simulations with random data (or exact solutions) need to be conducted for greater numbers of taxa (Table 13-2), the effects of base composition and length of sequence need to be studied in greater detail, and additional studies need to be conducted on the effects of different levels of signal and noise (or the effects of two sets of conflicting signal). Despite the need for these additional studies, this study indicates the general usefulness of tree-length distributions in phylogenetic analysis.

ACKNOWLEDGMENTS

I thank J. Coddington, B. I. Crother, M. T. Dixon, J. S. Farris, J. Felsenstein, K. Halanych, A. I. Hillis, W. M. Fitch, and M. White for helpful discussion, comments on the manuscript, and other assistance. D. R. Maddison and D. L. Swofford provided test versions of their respective programs, without which this study would not have been possible. J. J. Bull has provided constant input and advice on this work, and I am especially grateful for his efforts. This study was supported by National Science Foundation grants BSR-8657640 and BSR-8796293.

REFERENCES

De Jong, W. W., J. T. Gleaves, and D. Boulter. (1977) Evolutionary changes of α-crystallin and the phylogeny of mammalian orders. *J. Mol. Evol.* **10**:123–135.

Felsenstein, J. (1978) The number of evolutionary trees. *Syst. Zool.* **27**:27–33.

Field, K. G., G. J. Olsen, D. J. Lane, S. J. Giovannoni, M. T. Ghiselin, E. C. Raff, N. R. Pace, and R. A. Raff. (1988) Molecular phylogeny of the animal kingdom. *Science* **239**:748–753.

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

Fitch W. M. (1984) Cladistic and other methods: problems, pitfalls, and potentials. Pp. 221–252 in *Cladistics: Perspectives on the Reconstruction of Evolutionary History* (T. Duncan and T. F. Stuessy, eds.). Columbia University Press, New York.

Goodman, M., J. Czelusniak, and G. W. Moore. (1979a) Further remarks on the parameter of gene duplication and expression events in parsimony reconstructions. *Syst. Zool.* **28**:379–385.

Goodman, M., J. Czelusniak, G. W. Moore, A. Romero-Herrera, and G. Matsuda. (1979b) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* **28**:132–163.

Hillis, D. M. (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, D. M. (1991) The phylogeny of amphibians: current knowledge and the role of cytogenetics. Pp. 7–31 in *Amphibian Cytogenetics and Evolution* (D. M. Green and S. K. Sessions, eds.). Academic Press, New York.

Hillis, D. M., and R. de Sá. (1988) Phylogeny and taxonomy of the *Rana palmipes* group (Salientia: Ranidae). *Herpetol. Monographs* **2**:1–26.

Hillis, D. M., and M. T. Dixon. (1989) Vertebrate phylogeny: evidence from 28S ribosomal DNA sequences. Pp. 355–367 in *The Hierarchy of Life. Molecules and Morphology in Phylogenetic Analysis* (B. Fernholm, K. Bremer, and H. Jörnvall, eds.). Elsevier Science Publishers B. V., Amsterdam.

Hillis, D. M., M. T. Dixon, and L. K. Ammerman. (1991) The relationships of coelacanths: evidence from sequences of vertebrate 28S ribosomal RNA genes. In press in *Coelacanth Biology and Evolution* (J. A. Musick, ed.). Special volume of *Environ. Biol. Fishes* **32**.

Kocher, T. D., and T. J. White. (1989) Evolutionary analysis via PCR. Pp. 137–147 in *PCR Technology. Principles and Applications for DNA Amplification* (H. A. Erlich, ed.). Stockton Press, New York.

Le Quesne, W. J. (1989) Frequency distributions of lengths of possible networks from a data matrix. *Cladistics* **5**:395–407.

Maddison, W. P., and D. R. Maddison. (1991) *MacClade: Interactive Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland.

Sokal, R. R., and F. J. Rohlf. (1981) *Biometry*. 2nd ed. W. H. Freeman, San Francisco.

Swofford, D. L. (1990) *PAUP: Phylogenetic Analysis Using Parsimony*, Version 3.0. Illinois Natural History Survey, Champaign.

Werman, S. D. (1986) Phylogenetic relationships of the true viper, *Eristocophis mcmahoni*, based on parsimony analysis of allozyme characters. *Copeia* **1986**:1014–1020.

14

When are Phylogeny Estimates From Molecular and Morphological Data Incongruent?

DAVID L. SWOFFORD

Many sources of information may be available for use in inferring phylogenetic relationships among a group of taxa. Because different character sets share a common evolutionary history, we expect reliable methods of phylogenetic analysis to recover the correct evolutionary tree regardless of the kind of data employed. To the extent that character sets "tell the truth" about their past, phylogenies inferred from different character sets should be congruent with the true tree and therefore with each other. This line of reasoning suggests that congruence among data sets might provide the strongest achievable evidence that a proposed phylogeny is accurate (e.g., Penny and Hendy, 1986). In practice, however, the ideal of perfect congruence is frequently not achieved. Phylogenies estimated from additional character sets typically disagree in minor details and sometimes contain major discrepancies.

When phylogeny estimates from two character sets disagree in nontrivial ways, we are confronted with a dilemma: both of the estimates cannot be correct, so how do we reconcile the differences? One explanation is that one of the data sets is simply unreliable and that no method of phylogenetic reconstruction could be expected to recover the correct tree given the poor quality of the data. At one time or another, most of us have witnessed or participated in vigorous debates about the relative strengths of alternative sources of information or about the power versus futility of particular techniques. The increase in prominence of molecular approaches to systematics has increased the frequency of such exchanges, especially when molecular evidence controverts traditional notions of phylogenetic rela-