FORUM

Cladistics (1993) 9:93–99

ON DESCRIBING THE SHAPE OF ROOTED AND UNROOTED TREES

Roderic D. M. Page[1]

[1] *Biodiversity Programme, Department of Botany, Natural History Museum, Cromwell Road, London SW7 5BD, U.K.*

Received for publication 17 July 1992; accepted 27 November 1992

The shape of rooted trees is attracting increased attention in the literature, particularly in developing "null" models for testing hypotheses of taxon diversification (e.g. Slowinski and Guyer, 1989; Heard, 1992) and biogeography (e.g. Simberloff et al., 1981; Page, 1991). In view of the wide interest in the shape of trees it would be desirable to have a consistent notation for tree shape. Such a notation has already been developed by Harding (1972) and has been used by some workers (e.g. Slowinski and Guyer, 1989; Page, 1991). This notation also provides a natural definition of tree balance (Shao and Sokal, 1990) and can be extended, albeit less satisfactorily, to unrooted trees (Furnas, 1984). My purpose in this brief note is to draw attention to the papers and results of Harding and Furnas, and to illustrate some attractive properties of Harding's notation.

Harding's Notation

Harding's (1972) notation is best explained using an example. Figure 1 shows the 11 possible binary (i.e. fully resolved) rooted tree shapes of seven taxa. These 11 shapes are ordered 1 to 11 by their place in the *left-light rooted* (LLR) ordering defined by Furnas (1984:212; see also Harding, 1972:59–60). Given two trees, T and T', T' precedes T in the LLR ordering if:

1. $|T'| < |T|$, or
2. $|T'| = |T|$, and $T'_L < T_L$, or
3. $|T'| = |T|$, and $T'_L = T_L$, and $T'_R < T_R$,

where $|T|$ is the number of ultimate descendants in tree T, and T_L and T_R are the left and right subtrees, respectively, of T. Each tree in Fig. 1 has also been drawn in LLR form (Furnas, 1984), such that $T_L \leq T_R$, and both subtrees themselves are in LLR form. Using the three rules above we can see, for example, that tree 6 precedes tree 7 because the left subtree of tree 6 has fewer descendants than the left subtree of tree 7, and that tree 10 precedes tree 11, because although the left and right subtrees of both tree have the same number of ultimate descendants the right subtree of tree 10—(,(,(,))—precedes the right subtree of tree 11—((,),(,))—in the LLR order of four taxon trees (rule 3 above).

Given that we can define a unique ordering of tree shapes, Harding's (1972) notation represents the topology of a tree for n taxa as $\mathbf{n}_j$, where j is the jth shape in an ordered sequence of shapes. For the trees in Fig. 1, tree 1 as shape $\mathbf{7}_1$, tree 2 has shape $\mathbf{7}_2$, and so on.

0748-3007/93/010093+07$08.00/0

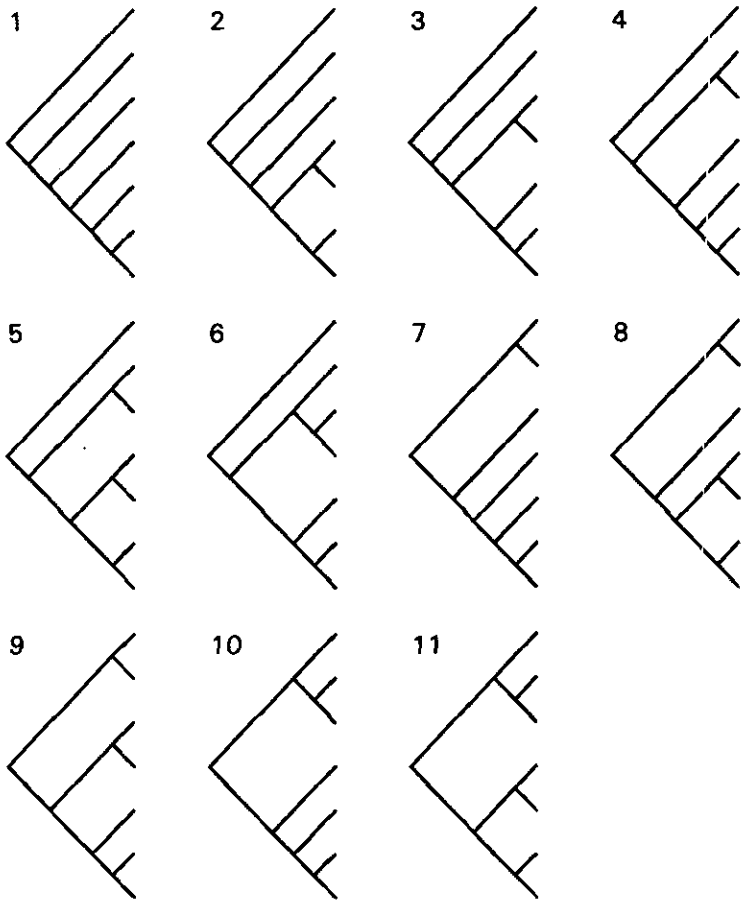

Fig. 1. The 11 possible tree shapes for binary-rooted trees for seven taxa draw in left-light rooted (LLR) form. In Harding's (1972) notation tree 1 has shape 7_1, tree 2 has shape 7_2, and so on. Note that the tree shapes are ordered from least balanced (7_1) to most balanced (7_{11}).

Utility

Computation

Any notation for tree shapes is in some sense arbitrary. Why might we want to adopt Harding's notation? One reason is ease of computation. The shape of a given rooted tree can be readily calculated using the formula given by Furnas (1984:213). It is also easy to "go backwards" and compute the tree that corresponds to a given position in the LLR order (Furnas, 1984:213–215). This makes it easy to automate the kind of studies done by Savage (1983) and Guyer and Slowinski (1991). To find the distribution of tree shapes one simply computes for each tree its place in the LLR order. As an example, Fig. 2 shows the frequency of each of the 11 tree shapes for seven taxa in two sets of 1000 random trees. One set of trees was generated using Harding's (1972) Markovian model, the other set was generated by assuming that all 10 395 labeled rooted trees or seven taxa are equally likely to occur [this is Slowkinski and Guyer's (1991:340) "proportional-to-distinguishable-arrangements" model].

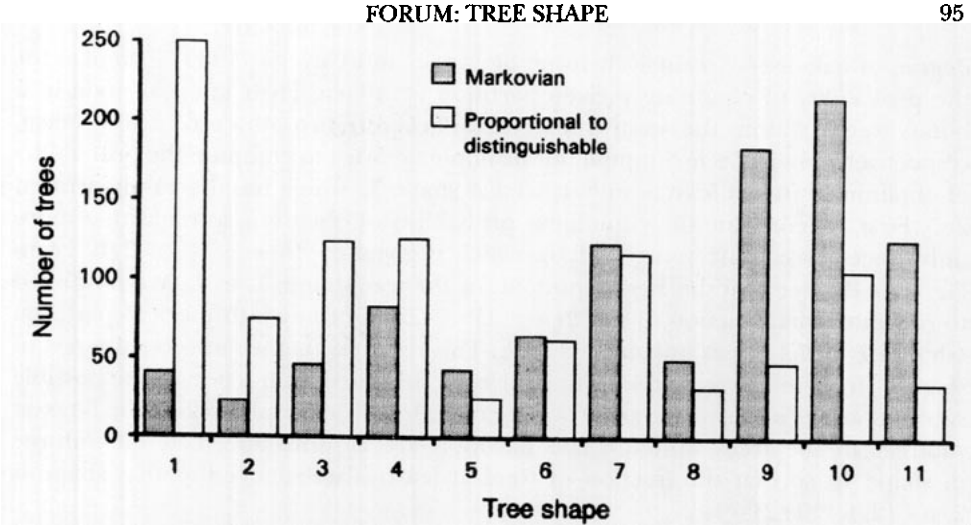

Fig. 2. The distribution of the tree shapes shown in Fig. 1 in two sets of 1000 random trees generated using the Markovian and the equiprobable models

Balance

The LLR ordering of trees also gives a ready definition of tree balance (Shao and Sokal, 1990). The trees in Fig. 1 can be regarded as ordered from least ($\mathbf{7}_1$) to most balanced ($\mathbf{7}_{11}$). We can see at a glance from Fig. 2 that balanced trees are more probable under the Markovian model than under the equiprobable model.

Kirkpatrick and Slatkin (in press) have proposed using the Markovian distribution of various measures of tree balance to test whether a given phylogeny is significantly more balanced or unbalanced than could be expected due to chance. A significant result can be due to variation in speciation and extinction rates between lineages (amongst other possibilities). This test requires knowing the distribution of tree balance. We can compute the exact distribution of tree balance (as defined by place in the LLR order) by computing for each shape its relative frequency in the set of labeled trees (Harding, 1972:66). Since for large numbers of taxa the number of tree shapes becomes prohibitively large we could also estimate the distribution by using samples of random trees, as I've done in Fig. 2. However, since we are only interested in the tails of the distribution we could perform an exact test of whether a tree is significantly (un)balanced by summing the number of ways of labeling each tree shape that occurs above (respectively, below) the observed tree's position in the LLR order until either the observed value or the desired significance level is reached. For example, suppose we observe a cladogram with shape $\mathbf{7}_2$. From Harding's (1972) table 2 under the Markovian model the probability of the shape $\mathbf{7}_1$ is 2/45, and the probability of shape $\mathbf{7}_2$ is 1/45, so the chance of getting a tree with the same or greater unbalance as shape $\mathbf{7}_2$ is 2/45 + 1/45 = 3/45 = 1/15 = 0·067. In contrast, there are 2520 cladograms with shape $\mathbf{7}_1$, and 630 with shape $\mathbf{7}_2$, so that under the assumption that all 10 395 labeled trees are equiprobable the probability of a tree with the same or greater unbalance as shape $\mathbf{7}_2$ is (2520 + 630)/10 395 = 0.303.

Felsenstein (pers. comm.) has pointed out that under Harding's (1972) model we can also compute the probability of obtaining a tree with the same or greater

degree of balance by recursively using Slowinski and Guyer's (1989) formulae for the probability of obtaining a given partition of n taxa. Each internal node in a binary tree partitions the set of n descendant taxa into two subsets of L and R taxa, respectively, where $L \leq R$. Suppose, for example, we wish to compute the probability of obtaining a tree at least as unbalanced as shape $\mathbf{7}_8$, which has the basal partition 2:5. First, we compute the cumulative probability of obtaining a tree with a more unbalanced basal partition (i.e. 1:6), which is given by $2L(n-1)^{-1} = 2/6$. From Fig. 1 we can see that this has accounted for the tree shapes $\mathbf{7}_1$ to $\mathbf{7}_6$. We now need to compute what fraction of the $2(n-1)^{-1} = 2/6$ of trees with the 2:5 partition (shapes $\mathbf{7}_7$ to $\mathbf{7}_9$) are as unbalanced as $\mathbf{7}_8$. This requires that we traverse the rest of shape $\mathbf{7}_8$ from left to right computing for each subtree what fraction of the possible subtrees are as unbalanced as those observed. In this example 1/2 of the 5-taxon subtrees in the trees with the 2:5 partition are as unbalanced as the subtree in shape $\mathbf{7}_8$, so that the fraction of trees at least as unbalanced as that shape is $2/6 + (2/6 \times 1/2) = 1/2$.

Unrooted Trees

Harding's (1972) notation applies only to rooted trees. However, it may be desirable to have a similar notation for unrooted trees, especially given the recent interest in distributions of unrooted tree lengths (e.g. Le Quesne, 1989; Hillis, 1991).

The number of unrooted tree shapes is less than the number of rooted tree shapes for the same number of taxa:

Taxa	3	4	5	6	7	8	9	10
Rooted shapes	1	2	3	6	11	23	46	98
Unrooted shapes	1	1	1	2	2	4	6	11

Furnas (1984) described an unrooted version of LLR ordering called *left-light*

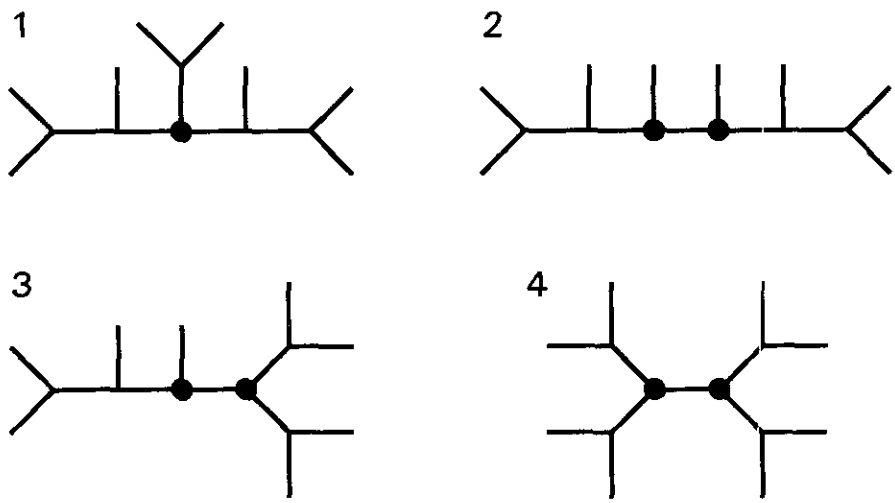

Fig. 3. The four possible unrooted tree topologies for eight taxa. The centroid nodes of each tree are marked (•) (see text).

centered (LLC) ordering. A center-rooted tree is an unrooted tree that has been rooted at the tree's center. Furnas (1984:217) terms a node of an unrooted tree a *centroid* if none of the branches that node lead to more than half the terminal nodes. A tree at least one and at most two centroids. As an example, Fig. 3 shows the four possible unrooted tree topologies for eight taxa and their centroids. If a tree has one centroid (e.g. tree 1 in Fig. 3) then that centroid is made the root of the tree and the tree T is termed "triple trunked", *trptr*(T), with three principal subtrees T_L, T_M and T_R (see tree 1 in Fig. 4). If the tree has two centroids (e.g. tree 2 in Fig. 3) then the tree is rooted on the branch between the two centroids yielding a "double trunked" tree, *dbltr*(T), with just two principal subtrees T_L and T_R (see tree 2 in Fig. 4).

If each principal subtree of the center-rooted tree is put in LLR form and the

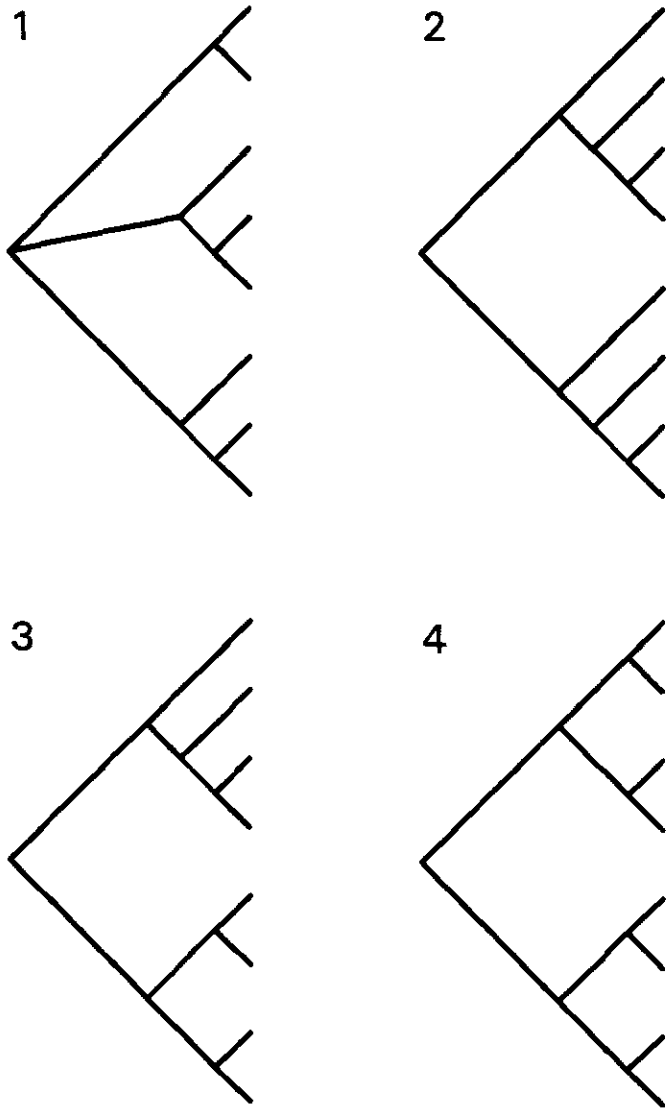

Fig. 4. The four unrooted trees in Fig. 3 drawn as rooted trees in left-light centered (LLC) form.

subtrees are ordered left to right by their place in the LLR ordering, then T' precedes T in the LLC ordering if (Furnas, 1984: 218):

1. *trptr*(T') and *trptr*(T) and either

   $T'_L < T_L$, or

   $T'_L = T_L$ and $T'_M < T_M$, or

   $T'_L = T_L$ and $T'_M = T_M$ and $T'_R < T_R$,

or

2. *trptr*(T') and *dbltr*(T) or,

3. *dbltr*(T') and *dbltr*(T) and $T' < T$.

As an example, Fig. 4 shows the four trees in Fig. 3 in LLC order.

While Furnas' LLC ordering allows each unrooted tree shape to be assigned a unique identification number, the ordering seems less natural than the LLR ordering for rooted trees. For example, tree 1 in Fig. 3 precedes tree 2, although intuitively tree 2 is less balanced than tree 1. Hence, the LLC order might best be used just as a notation for unrooted tree shapes, rather than as a measure of tree balance as well.

Software

Furnas' (1984) motivation for developing the LLR and LLC orderings was the generation of random rooted and unrooted trees with a uniform distribution of shapes (see also Simberloff et al., 1981). Routines to generate random trees using this distribution, and to compute the LLR and LLC order of a tree, are available in the program COMPONENT for Windows which can be obtained from the author.

Acknowledgments

I thank Joe Felsenstein, Pablo Goloboff and Chris Humphries for their comments on the manuscript. This research was supported by an Interdisciplinary Research Fellowship from the Natural History Museum, London.

REFERENCES

FURNAS, G. W. 1984. The generation of random, binary unordered trees. J. Classif. 1: 187–233.

GUYER, C. AND J. B. SLOWINSKI. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. Evolution 45: 340–350.

HARDING, E. F. 1972. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3: 44–77.

HEARD, S. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated trees. Evolution. 46: 1818–1826.

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, New York, pp. 278–294.

KIRKPATRICK, M. AND M. SLATKIN. In press. Searching for patterns in the shape of a phylogenetic tree. Evolution.

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

PAGE, R. D. M. 1991. Random dendrograms and null hypotheses in cladistic biogeography. Syst. Zool. 40: 54–62.

SAVAGE, H. M. 1983. The shape of evolution. Biol. J. Linn. Soc. 20: 225–244.

SHAO, K.-T. AND R. R. SOKAL. 1990. Tree balance. Syst. Zool. 39: 266–276.

SIMBERLOFF, D., K. L. HECK, E. D. MCCOY AND E. F. CONNOR. 1981. There have been no statistical tests of cladistic biogeographical hypotheses. *In:* G. Nelson and D. E. Rosen (eds). Vicariance Biogeography: A Critique. Columbia Univ. Press, New York, pp. 40–63.

SLOWINSKI, J. B. AND C. GUYER. 1989. Testing the stochasticity of patterns of organismal diversity: An improved null model. Am. Nat. 134: 907–921.