Syst. Zool., 39(3):266–276, 1990

TREE BALANCE

KWANG-TSAO SHAO[1] AND ROBERT R. SOKAL

Department of Ecology and Evolution, State University of New York,
Stony Brook, New York 11794-5245

Abstract.—Hierarchic classifications can differ with respect to tree balance—the degree to which branches divide the subtended taxa into subsets of equal size. Several indices, sensitive to different aspects of tree balance, are compared. Extensions of these indices to multifurcating trees and to trees with varying numbers of OTUs are proposed, and suggestions for employing these indices are furnished. Different tree-forming algorithms may result in trees with differing degrees of balance no matter which index is computed. Tree balance is an important consideration for phylogenetic systematics, because balance of the true phylogeny will affect the accuracy of its estimates. [Tree balance; tree symmetry; tree asymmetry; comparing classifications.]

Trees differ in the degree to which branches divide the subtended taxa into subsets of equal size. Several terms have appeared in the literature to describe this aspect of tree form: shape, form, symmetry (or asymmetry), balance (or imbalance), and skewness. Rohlf and Fisher (1968) used the general terms *form* and *shape* to cover the meaning of tree balance. Some authors have used *shape* to refer to tree topology (Farris, 1973; Dobson, 1975; Smith and Waterman, 1980; Fowlkes et al., 1983). However, these papers were not concerned with the *degree* of balance or shape, but with the consensus or distance between two trees. Savage (1983) estimated probabilities associated with various tree shapes or topologies of binary trees in three different ways. Mickevich and Platnick (1989) took the *symmetry* or the *pectinate arrangement* into consideration in their studies on the information content of classifications. We prefer the term *tree balance* as being truest to the property we describe below and least likely to lead to semantic confusion with other properties of trees.

Common sense notions of tree balance lead to the recognition of balance as indicating equal numbers of included terminal nodes for both branches of the various furcations (interior nodes) of a dendrogram. By contrast, imbalance is the opposite property—unequal numbers of included terminal nodes. By such a criterion, tree 1 in Figure 1 is the most balanced and tree 6 the most unbalanced tree among the first six trees, which have the same number of interior nodes. Trees showing the most unbalanced or asymmetrical structure are often referred to as pectinate, comblike, chained, or linear. A more precise definition of balance or imbalance will vary depending on the specific aspects of this property measured by a given index. Such indices and the nature of the tree balance or imbalance they describe will be discussed below.

Several indices have been proposed to describe the balance or imbalance of a dendrogram. Sackin (1972) used a *b* (for branching) vector to characterize a phenogram and measure its "useful properties." He claimed that "The balanced phenogram looks 'better' than the skewed one because (1) fewer taxonomic categories need to be postulated; and (2) the cluster sizes at any category level are more constant." Colless (1982) proposed an index to demonstrate that the Hennigian cladograms in Wiley's (1981) book are highly unbalanced. Both of these indices measure tree imbalance, because their values will be higher the more unbalanced the trees. Astolfi et al. (1981), using a linearity index, demonstrated that tree form will affect the accuracy of reconstructing an unrooted tree. F. Murtagh (in an unpublished manuscript dated

[1] Present address: Institute of Zoology, Academia Sinica, Nankang, Taipei, Taiwan, R.O.C.

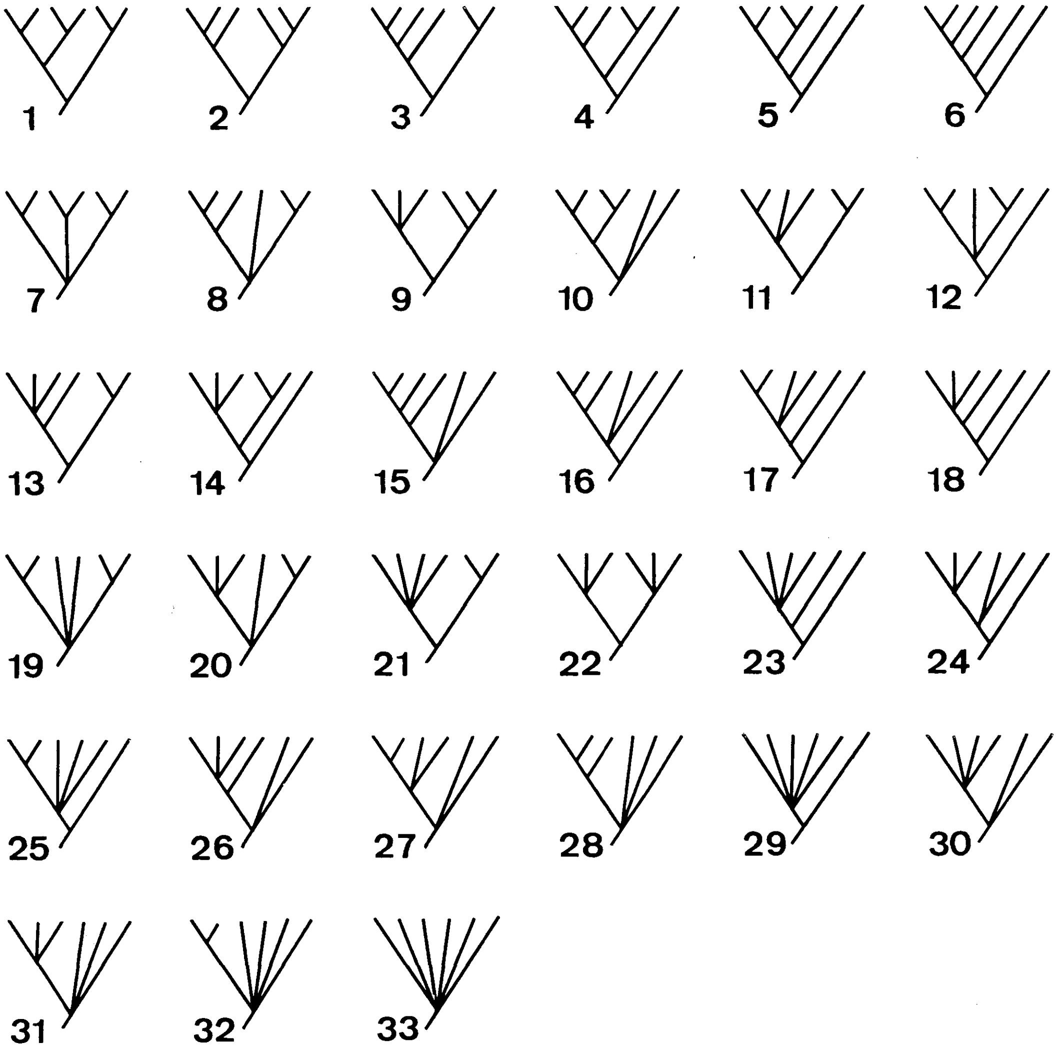

FIG. 1. All 33 possible unlabeled rooted trees for six OTUs. The numbers of interior nodes (including the root) for trees 1–6, 7–18, 19–28, 29–32, and 33 are 5, 4, 3, 2, and 1, respectively.

1982) developed a complicated "structure coefficient" to measure tree balance and prove that different phenetic clustering methods will result in phenograms with different shapes. The last two indices are not considered further in this paper because the former was originally defined for an unrooted tree only and the latter requires a fully bifurcating tree, so its usefulness is restricted. Fowlkes et al. (1983) developed a parameter *a* to specify tree shape and demonstrated by simulation that single linkage clustering will result in a more unbalanced tree than complete linkage clustering. This parameter is also not employed in this paper because (1) it is computationally complex and (2) it applies only to binary trees with particular tree topologies.

Tree balance plays an important role in various types of numerical taxonomic studies. The balance of the trees being analyzed can affect the results when different numerical taxonomic methods are compared (Rohlf et al., 1983; Shao, 1983; Sokal, 1983). Sokal (1983) included shape measurements among a series of tree characteristics to argue that the true tree of the Caminalcules, as well as estimates thereof, does not differ from estimated trees constructed for real organisms. Tree balance may also be interpreted phylogenetically. For example, it may reflect differences in extinction or speciation rate during phylogeny—a balanced tree implies more equal extinction and speciation rates than an unbalanced tree. Grant (1963) has proposed "directed speciation" as a mechanism resulting in a more unbalanced tree. By contrast, Raup and Gould's (1974) "phylogenetic drift," in which the direction of speciation is random, appears to yield a more balanced tree. Stanley (1979:135) questioned the likelihood of a comblike tree existing in the real world. To obtain such a tree, one branch at each furcation should have a high probability of speciation soon after forming, followed by a low speciation probability after forming the new lineage. Phylogeneticists who construct their classifications based on synapomorphies frequently obtain cladograms in the shape of a "Hennigian comb." Such trees are often produced by paleontologists as well. It is of great interest to systematists whether such tree shapes reflect patterns of speciation or possible biases of the estimation method. It is obvious that neither of these aspects of tree estimation can be investigated unless a suitable measure of tree balance is developed. Of equal concern to those wishing to estimate phylogenies is the finding by Rohlf et al. (1990) that the amount of imbalance of true phylogenetic trees affects the accuracy with which they can be estimated by different methods.

Other recent papers relating to tree balance are connected to consensus methods used to compare classifications. For instance, two consensus indices—Mickevich's (1980) index (CI_M) and levels sum (Schuh and Farris, 1981)—have a maximum possible value that is a function of shape of the consensus tree (Rohlf, 1982). Consensus indices related to term information (Nelson and Platnick, 1981) are biased by tree balance (Shao and Rohlf, 1983), termed tree shape in this reference. Shao (1983) showed that the three consensus indices—term information, total information (Nelson and Platnick, 1981), and levels sum—are all algebraically derivable from one of the imbalance indices discussed below. In view of their dependence on tree balance, the usefulness of some consensus indices for measuring consensus information can be questioned. Furthermore, understanding how to measure tree balance is an essential step toward understanding why different numerical taxonomic techniques will result in differing distributions of consensus indices.

In this paper we propose two balance indices and compare their properties with those of two imbalance indices introduced, respectively, by M. J. Sackin and D. H. Colless.

FOUR MEASURES OF TREE BALANCE

We shall designate balance indices by B and indices measuring imbalance by I. For an explanation and comparison of various balance and imbalance indices, consider Figure 1, which shows all 33 possible unlabeled rooted trees for six OTUs, excluding trees with vertices of degree 2 (i.e., with nodes connected to only two other nodes). Four different indices are examined here. Two indices, $I_S(1)$ and $I_C(1)$, measure the imbalance of the tree; the other two, $B_1(1)$ and $B_2(1)$, proposed here, are balance indices. The postscript (1) refers to their original formulations. Other formulations, designated (2), (3), and (*), will be introduced later.

Let t be the number of terminal nodes (OTUs) in a rooted tree and k the number of furcations (interior nodes) including the root. Quantity k can vary between 1 and $t - 1$. Its upper bound will depend on the resolution of the tree, with the maximum value, $t - 1$, reached only in a fully bifurcating tree. We designate N_i as the number of interior nodes between terminal node i and the root, which is included in the count. Thus, $N_i = 4$ for the leftmost terminal node in tree 3 of Figure 1. Then Sackin's (1972) index is defined as

$$I_S(1) = \sum_i N_i \qquad i = 1, \ldots, t. \qquad (1)$$

We designate as T_j the absolute difference in number of terminal nodes subtended by the two branches of bifurcation j. Thus, T_j for the root of tree 3 in Figure 1 has the value $|4 - 2| = 2$. The index described by Colless (1982) can be stated as

$$I_C(1) = \sum_j T_j \qquad j = 1, \ldots, k(3), \qquad (2)$$

where $k(3)$ is the number of interior nodes of degree 3 (bifurcations). In a fully bifurcating tree $k = k(3)$.

The third index is based on the $k - 1$ interior nodes excluding the root. For each furcation j, let us consider that furcation to be the root of the subset of terminal nodes (OTUs) subtended by it. We define $M_j = \max N_i$ (as defined above), where the maximum is computed over the subset of OTUs i subtended by j. For the subset including the first five OTUs in tree 17 of Figure 1, $M_j = 3$; for a similar subset in tree 6, $M_j = 4$. The index is defined as

$$B_1(1) = \sum_j (1/M_j)$$

$$j = 1, \ldots, k - 1 \ (j \neq \text{root}), \qquad (3)$$

where the summation is over all interior nodes other than the root of the entire tree. For tree 17 in Figure 1, we compute (from the top) $B_1(1) = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} = 1.833$.

The fourth index evaluates the probability of reaching terminal node i starting at the root, assuming equiprobable branching at each interior node. The probability is

$$P_i = \prod_j [1/(d_j - 1)] \qquad j = 1, \ldots, N_i, \qquad (4)$$

where d_j is the degree of interior node j. In the special case of a completely bifurcating tree, the formula simplifies to

$$P_i = (\tfrac{1}{2})^{N_i}, \qquad (4A)$$

because all furcations are of degree 3. Using Equation 4 we obtain P_i for the leftmost terminal node of tree 16 in Figure 1 as (from the root) $\frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{24}$. For the leftmost terminal node of tree 6 in Figure 1 (by Equation 4A because it is fully bifurcating), we obtain $P_i = (\frac{1}{2})^5 = 0.03125$. To develop an index of balance, we compute the Shannon-Wiener information function as

$$B_2(1) = -\sum_i P_i \log P_i \qquad i = 1, \ldots, t. \qquad (5)$$

This measures the equitability of the probabilities of arriving at the terminal nodes of the tree. For tree 1 in Figure 1 this would yield $-[4(\frac{1}{8} \log \frac{1}{8}) + 2(\frac{1}{4} \log \frac{1}{4})] = 0.7526$.

Indices I_S, I_C, and B_1 have been referred to previously as BSUM, COLLESS or SI_b, and SHAO, respectively (Sokal, 1983; Rohlf et al., 1990). $I_C(1)$ was originally defined for fully bifurcating (binary) trees, i.e., trees where all internal nodes are of degree 3. The index has been modified for nonbinary trees (those with some internal nodes of degree >3) by ignoring multifurcating nodes. The values of these four indices for the 33 trees of Figure 1 are given in Table 1. Readers can confirm their understanding of Equations 1 through 4 by working out several of the values. Note that $B_1(1)$ and $B_2(1)$ were designed to measure balance, whereas the other two indices measure imbalance.

The results in Table 1 show that the various indices measure tree balance or imbalance differently. For instance, $I_S(1)$ yields different values for trees 3–5 and 15–18, but $B_1(1)$ yields the same value for the trees in each of these subsets; $I_S(1)$, $I_C(1)$, and $B_2(1)$ yield equal balance for trees 1 and 2, but index $B_1(1)$ indicates tree 1 as more balanced. These inconsistencies are not inherent faults of the indices, but reflect the fact that they embody different definitions of balance. Because the evaluation of balance or imbalance involves abstract and subjective issues, it is unlikely that everyone will agree on the same definition. However, it is still of interest to measure tree balance by these indices because (1) different indices will maintain a common order when certain trees show an unambiguous ordering with respect to balance (in Table 2, e.g., the following trees are unambiguously ordered by common sense definitions of balance and by all balance or imbalance indices: 1 > 5 > 6, 7 > 11 > 15, and 19 > 26) and (2) different indices

TABLE 1. Balance and imbalance indices for the 33 trees in Figure 1. For explanation of indices see text.

Tree no.	No. of interior nodes	$I_S(1)$	$I_C(1)$	$B_1(1)$	$B_2(1)$	$I_S(*)$	$B_1(*)$
1	5	16	2	3.500	0.753	0.000	1.000
2		16	2	3.000	0.753	0.000	0.647
3		17	5	2.833	0.715	0.250	0.529
4		18	6	2.833	0.640	0.500	0.529
5		19	7	2.833	0.602	0.750	0.529
6		20	10	2.083	0.583	1.000	0.000
7	4	12	0	3.000	0.778	0.000	1.000
8		13	1	2.500	0.728	0.167	0.571
9		14	1	2.500	0.765	0.333	0.571
10		14	0	2.500	0.678	0.333	0.571
11		14	2	2.500	0.740	0.333	0.571
12		15	4	2.500	0.640	0.500	0.571
13		15	4	2.500	0.640	0.500	0.571
14		16	5	2.500	0.646	0.667	0.571
15		15	3	1.833	0.653	0.500	0.000
16		16	5	1.833	0.615	0.667	0.000
17		17	7	1.833	0.596	0.833	0.000
18		18	9	1.833	0.586	1.000	0.000
19	3	10	0	2.000	0.753	0.000	1.000
20		11	0	2.000	0.737	0.200	1.000
21		12	2	2.000	0.753	0.400	1.000
22		12	0	2.000	0.778	0.400	1.000
23		15	7	1.500	0.602	1.000	0.000
24		14	4	1.500	0.619	0.800	0.000
25		13	4	1.500	0.646	0.600	0.000
26		13	2	1.500	0.657	0.600	0.000
27		12	0	1.500	0.670	0.400	0.000
28		11	1	1.500	0.715	0.200	0.000
29	2	11	4	1.000	0.651	1.000	0.000
30		10	u[a]	1.000	0.678	0.667	0.000
31		9	u	1.000	0.721	0.333	0.000
32		8	0	1.000	0.759	0.000	0.000
33	1	6	u	0.000	0.778	0.000	0.000

[a] u means the index is undefined for that particular tree because it lacks any bifurcating subsets.

emphasize different aspects of balance and one can choose among them based on one's judgment of which aspects of balance are more important in a given application.

COMPARING THE BALANCE OF DIFFERENT TREES

When the investigator wishes to compare different trees for balance or imbalance, the four indices can be used as given above, so long as the number of OTUs t in the trees to be compared is the same. However, when trees with different numbers of terminal taxa are compared they are no longer comparable, because the maxima of the indices change monotonically with an increase in t. Let us first consider binary (fully bifurcating) trees. The obvious remedy is a normalization by dividing I(1) or B(1) by its maximum value for a given t. We designate such normalized indices I(2) or B(2). However, this does not completely solve the problem because the minimum values of the normalized indices still vary with OTU number t. For example, the minimum normalized index $I_S(2)$ for 12 OTUs equals 0.5714; for 16 OTUs, it equals 0.4740; the value declines monotonically as t increases. Therefore, a further correction is necessary. We compute

$$[I(1) - \min I(1)]/[\max I(1) - \min I(1)]$$

or

TABLE 2. Ordering of the four balance measures for the 33 trees in Figure 1. Within each group, trees are ordered from the most to the least balanced. Note that the ordering of I(*) or B(*) equals that of I(1) and B(1).

Number of interior nodes	Index	Ordering of trees by their balance
5	$I_S = I_C = B_2$	$1, 2 > 3 > 4 > 5 > 6$
	B_1	$1 > 2 > 3, 4, 5 > 6$
4	I_S	$7 > 8 > 9, 10, 11 > 12, 13, 15 > 14, 16 > 17, 18$
	I_C	$7, 10 > 8, 9 > 11 > 15 > 12, 13 > 14, 16 > 17 > 18$
	B_1	$7 > 8, 9, 10, 11, 12, 13, 14 > 15, 16, 17, 18$
	B_2	$7 > 9 > 11 > 8 > 13 > 10 > 15 > 14 > 12 > 16 > 17 > 18$
3	I_S	$19 > 20, 28 > 21, 22, 27 > 26 > 24 > 25 > 23$
	I_C	$19, 20, 22, 27 > 28 > 21, 26 > 24, 25 > 23$
	B_1	$19, 20, 21, 22 > 23, 24, 25, 26, 27, 28$
	B_2	$22 > 19, 21 > 20 > 28 > 27 > 26 > 25 > 24 > 23$
2	$I_S = B_2$	$32 > 31 > 30 > 29$
	I_C	$30, 31, 32 > 29$
	B_1	$29 = 30 = 31 = 32$
1	$I_S = I_C = B_1 = B_2$	33

$$[B(1) - \min B(1)]/[\max B(1) - \min B(1)]$$

and call those corrected indices I(3) or B(3), respectively. These indices will range from zero to one. In the above formulas, find min I(1) and max B(1) by computing these indices for the most balanced tree with a given number *t* of OTUs. Similarly, find max I(1) and min B(1) by employing the most unbalanced tree for the same number of OTUs.

Let us now consider multifurcating trees. With these, I(3) or B(3) occasionally will become negative, when the most balanced or unbalanced binary trees needed for computing min I(1) and min B(1) are more resolved (i.e., have more interior nodes) than the multifurcating tree being evaluated. Note that I(1) or B(1) values are strongly positively correlated with tree resolution. For instance, in Table 1, maximal and minimal index values decline with decreasing interior node number. Consequently, I(1) or B(1) for a largely unresolved tree will be smaller than min I(1) or min B(1) for a fully resolved tree. To solve both the problems of standardization and multifurcations, and to have the index values range between 0.0 and 1.0, the maximum and minimum index values in the above formula for I(3) or B(3) should be redefined so as to refer to a set of trees with the same number of interior nodes as the tree being measured, rather than to a completely binary tree. An algorithm for calculating the maximum or minimum index values for trees with the same number of interior nodes is furnished in the Appendix. The final formulations are denoted as I(*) and B(*).

COMPARISON OF BALANCE INDICES

As already stated, different indices measure different aspects of balance (see Tables 1, 2). Table 3a presents the correlation matrix among nine indices based on 100 randomly generated multifurcating trees for $t = 10$ OTUs. (An algorithm for randomly and equiprobably generating trees containing multifurcations as well as bifurcations was given by Oden and Shao [1984].) Table 3b is the correlation matrix among the six indices of Table 2 over the 33 trees of Figure 1. The correlations among I(1), I(2), and I(3), or B(1), B(2), and B(3), for each index are all 1.0 (the second formulation for each index is not shown to conserve space). This means that the three versions of each index preserve the property of each balance measure without distortion. Even for the final formulations I(*) and B(*), the correlation with I(1) and B(1) is still high (Table 3a), unless there are many unresolved trees. Such trees tend to lower the correlation between I(1) and I(*) or B(1) and B(*) because low tree resolution (number of interior nodes) tends to

TABLE 3. Correlation matrix among (a) nine indices over 100 randomly generated multifurcating trees; (b) six indices over the 33 trees of Figure 1.

(a)

	$I_S(1)$	$I_C(1)$	$B_1(1)$	$B_2(1)$	$I_S(3)$	$B_1(3)$	$B_2(3)$	$I_S(*)$
$I_C(1)$	0.965							
$B_1(1)$	−0.444	−0.519						
$B_2(1)$	−0.799	−0.794	0.712					
$I_S(3)$	1.000	0.965	−0.444	−0.799				
$B_1(3)$	−0.444	−0.519	1.000	0.712	−0.444			
$B_2(3)$	−0.799	−0.794	0.712	1.000	−0.799	0.712		
$I_S(*)$	0.863	0.891	−0.775	−0.930	0.863	−0.775	−0.930	
$B_1(*)$	−0.667	−0.702	0.934	0.785	−0.667	0.934	0.785	−0.845

(b)

	$I_S(1)$	$I_C(1)$	$B_1(1)$	$B_2(1)$	$I_S(*)$
$I_C(1)$	0.810				
$B_1(1)$	0.657	0.183			
$B_2(1)$	−0.629	−0.819	0.058		
$I_S(*)$	0.497	0.764	−0.198	−0.877	
$B_1(*)$	0.051	−0.313	0.690	0.556	−0.523

affect the (1) formulation but not the (*) formulation. In Table 3b, the low correlation between $I_S(*)$ and $B_1(*)$ with the other I(1) or B(1)s is due to this phenomenon, because a large proportion of the 33 trees of Figure 1 is unresolved. However, when trees with the same interior node number are compared, the ordering with respect to balance is inconsistent only among different indices and not within each index. For example, the $I_S(*)$ and $B_1(*)$ values for the 33 trees of Figure 1 show the same inconsistencies as in their original forms, $I_S(1)$ and $B_1(1)$ in Table 2.

On examining the relationships among different indices, we find I_S and I_C to be highly correlated; B_2 is correlated more with I_S and I_C than with B_1. These relationships are stronger for large numbers of OTUs. In general, the indices can be placed into two groups—(I_S, I_C, B_2) and (B_1)—each measuring a different aspect of balance. One major difference between the two groups is that the indices in the first group take the position or the size (i.e., the number of OTUs) of the subsets into account, but the second group (B_1) measures only the graph diameter of each taxonomic subset (the maximum distance from the root of the subset to a subtended OTU) without allowing for the size of each subset (see trees 3–5 of Fig. 1). This can also be seen in trees 29–32 of Figure 1, where the indices of the first group yield different values but B_1 yields identical values (Table 1).

In Table 1, I(2) or B(2) and I(3) or B(3) are omitted because all 33 trees have the same OTU number and therefore the normalizations would be simply proportional. Only I_S and B_1 were calculated in the I(*) and B(*) formulation because the other indices have the following restrictions or undesirable properties. I_C has the disadvantage of an irregular trend of minimum $I_C(1)$ values as the number of OTUs increases. This in turn will affect the final normalized index values, causing them to fluctuate. For example, the minimum $I_C(1)$ for OTU numbers $t = 7$, 8, and 9 are 2, 0, and 3, respectively. B_2 is deficient because its value becomes very small and the range between maximum $B_2(1)$ and minimum $B_2(1)$ shrinks as the OTU number increases; besides, the range of minimum $B_2(1)$ for different interior node numbers will also decrease and the value of minimum $B_2(1)$ will be fixed at 0.6021 (=2 log 2) no matter how many furcations the tree has, as long as the OTU number t is greater than 35. Based on the above considerations, our choice of indices for most studies would be in the following order of preference: B_1, I_S, I_C, and B_2.

THE CHOICE OF A BALANCE INDEX

It would not be difficult to develop other balance or imbalance indices to quantify tree balance. However, it is doubtful that the problem of disagreement among different indices could be avoided. On the basis of our experience we suggest the following.

1. Choose those indices, such as I_S and B_1, with fewer mathematical drawbacks and that are less restricted to certain types of trees.
2. Choose those indices that will assign equal values to ambiguous trees. For example, trees 3–5 or 15–18 (in Fig. 1) are ambiguous with respect to tree balance; under the circumstances, B_1 seems better than I_S because B_1 will give identical values for each of these groups of trees, but I_S will not.
3. Choose the formulation I(*) or B(*) when comparing trees that are not all binary. For computational simplicity, one can use I(3) or B(3) when the trees are all binary but differ in OTU number, or even I(1) or B(1) when the trees are all binary and have the same OTU number.
4. Choose more than one index if possible. The algorithms for computing these indices should be as uncorrelated as possible, i.e., the more different aspects for defining tree balance that are covered, the better. The correlation matrices among some indices in Table 3 could be used to choose uncorrelated indices. If the results from uncorrelated indices are consistent, it suggests that the trees examined differ unambiguously in balance. If there are inconsistencies, conclusions on balance should be made cautiously.

ACKNOWLEDGMENTS

This paper represents contribution 765 from the Program in Ecology and Evolution at the State University of New York at Stony Brook. This study was supported by NSF grants DEB 8306004 and BSR 8822491 to Robert R. Sokal. We are grateful to Drs. F. J. Rohlf, M. Bell, R. Grimson, W. H. E. Day, and C. A. Meacham for their valuable comments on earlier versions of the manuscript and to F. J. Rohlf for a critical reading of the present version. Much advice and computational assistance was given by Dr. Kent Fiala, Gene Hart, and Barbara Thomson. Computer time was furnished at the Biological Sciences Computation Center of the State University of New York at Stony Brook.

REFERENCES

ASTOLFI, P., K. K. KIDD, AND L. L. CAVALLI-SFORZA. 1981. A comparison of methods for reconstructing evolutionary trees. Syst. Zool., 30:156–169.

COLLESS, D. H. 1982. Phylogenetics: The theory and practice of phylogenetic systematics II [book review]. Syst. Zool., 31:100–104.

DOBSON, A. J. 1975. Comparing the shapes of trees. Pages 95–100 *in* Combinatorial mathematics 3. Lecture notes in mathematics. Volume 452. Springer-Verlag, New York.

FARRIS, J. S. 1973. On comparing the shape of taxonomic trees. Syst. Zool., 22:50–54.

FOWLKES, E. B., C. L. MALLOWS, AND J. E. MCRAE. 1983. Some methods for studying the shape of hierarchical trees. J. Am. Stat. Assoc., 78:553–569.

GRANT, V. 1963. The origin of adaptation. Columbia Univ. Press, New York. 606 pp.

MICKEVICH, M. F. 1980. Taxonomic congruence: Rohlf and Sokal's misunderstanding. Syst. Zool., 29:162–176.

MICKEVICH, M. F., AND N. I. PLATNICK. 1989. On the information content of classifications. Cladistics, 5: 33–47.

NELSON, G. J., AND N. PLATNICK. 1981. Systematics and biogeography: Cladistics and vicariance. Columbia Univ. Press, New York. 567 pp.

ODEN, N. L., AND K. SHAO. 1984. An algorithm to equiprobably generate all directed trees with *k* labeled terminal nodes and unlabeled internal nodes. Bull. Math. Biol., 46:379–387.

RAUP, D. M., AND S. J. GOULD. 1974. Stochastic simulation and evolution of morphology—Towards a nomothetic palaeontology. Syst. Zool., 23:305–322.

ROHLF, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci., 59:131–144.

ROHLF, F. J., W. S. CHANG, R. R. SOKAL, AND J. KIM. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution, 44 (in press).

ROHLF, F. J., AND D. R. FISHER. 1968. Tests for hierarchical structure in random data sets. Syst. Zool., 17:407–412.

ROHLF, F. J., J. KISHPAUGH, AND D. KIRK. 1983. NT-SYS, numerical taxonomy system of multivariate statistical programs. Technical report. State Univ. New York, Stony Brook.

SACKIN, M. J. 1972. "Good" and "bad" phenograms. Syst. Zool., 21:225–226.

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

SCHUH, R. J., AND J. S. FARRIS. 1981. Methods for investigating taxonomic congruence and their application to the Leptopodomorpha. Syst. Zool., 30: 331–351.

SHAO, K. 1983. Consensus methods in numerical taxonomy. Ph.D. Dissertation, State Univ. New York, Stony Brook. 290 pp.

SHAO, K., AND F. J. ROHLF. 1983. Sampling distribution of consensus indices when all bifurcating trees are equally likely. Pages 132–137 *in* Numerical taxonomy. Proceedings of a NATO Advanced Study Institute (J. Felsenstein, ed.). NATO Adv. Study Inst. Ser. G (Ecol. Sci.), No. 1. Springer-Verlag, Berlin.

SMITH, T. F., AND M. S. WATERMAN. 1980. How alike are two trees? Am. Math. Mon., 87:552–553.

SOKAL, R. R. 1983. A phylogenetic analysis of the Caminalcules I. The data base. Syst. Zool., 32:159–184.

STANLEY, S. 1979. Macroevolution, pattern and process. W. H. Freeman, San Francisco. 332 pp.

WILEY, E. O. 1981. Phylogenetics: The theory and practice of phylogenetic systematics. John Wiley and Sons, New York. 439 pp.

Received 8 March 1990; accepted 11 June 1990

APPENDIX

Algorithms for Calculating the Maximum and Minimum Index Values of $I_S(1)$ and $B_1(1)$ for a Tree with Given Numbers of OTUs and Interior Nodes

I. Hand Calculation

It is not difficult to calculate balance or imbalance indices by hand if there are few trees based on a small number of OTUs. To calculate them by hand you may have to explore several trees, because sometimes it may not be obvious which is the most balanced tree and this needs to be discovered by calculating the index values of the various contenders.

1. Draw the most unbalanced and the most balanced *binary* (fully bifurcating) trees with the same number of OTUs t as the tree to be evaluated.

2.1. If the tree to be evaluated is binary, go to step 4.

2.2. If the tree to be evaluated is multifurcating, go to step 3.1.

3.1. Obtain the most unbalanced multifurcating tree with the same number of interior nodes k as the tree to be evaluated as follows. Change the resolved subsets of the most unbalanced binary tree from step 1 into unresolved subsets, starting from the smallest subset on the top (i.e., upper level) of the tree down to the largest subsets on the bottom (i.e., lower level) of the tree, one by one, until the number of interior nodes is reduced to k (see Fig. 2a).

3.2. Obtain the most balanced multifurcating tree with the same number of interior nodes k as the tree to be evaluated as follows. Change the resolved subsets of the most balanced binary tree from step 1 into unresolved subsets by directly connecting them to the root, starting from the largest subset near the root of the tree to the smallest subsets at the top of the tree. If the sizes of subsets being chosen are equal, then choose any one of them (see Fig. 2b).

4. Calculate the max $I_S(1)$ or min $B_1(1)$ values from the most unbalanced tree and the min $I_S(1)$ or max $B_1(1)$ values from the most balanced tree obtained via steps 1 or 3.1 and 3.2 by using the formulas given in the text. The index values for $I_S(1)$ and $B_1(1)$ calculated from these two trees will be the maximum or minimum values among all possible trees with the same number of interior nodes. Thus, the most unbalanced or balanced trees obtained from the above steps will satisfy the definition of these two indices.

II. Computer Calculation

To evaluate any one tree, binary or multifurcating, with t OTUs, compute its number of interior nodes k and index values of $I_S(1)$ and $B_1(1)$. Then,

1. The max $I_S(1)$ or min $B_1(1)$ index values for the most unbalanced tree with the same t and k as the tree to be evaluated can be calculated directly from

$$\max I_S(1) = t + \sum_j (t - j) \qquad j = 1, \ldots, (k - 1);$$

$$\min B_1(1) = \sum_j (1/j) \qquad j = 1, \ldots, (k - 1).$$

For example, for tree c of Figure 3, max $I_S(1) = 11 + 10 + 9 + \ldots + 4 = 60$; min $B_1(1) = 1/1 + 1/2 + 1/3 + \ldots + 1/7 = 2.5929$. The index values calculated by these formulas are identical to the values calculated by the hand method I.3.1 for any given number of OTUs and number of interior nodes.

2. The min $I_S(1)$ or max $B_1(1)$ values for the most balanced tree with the same t and k as the tree to be evaluated can be computed by the following formulas, in which INT and mod stand for the integer and remainder portions of fractions, respectively:

$$\min I_S(1) = t + \sum_{i=1}^{l-1} \sum_{j=1}^{m_i} F(t, i); \qquad \text{(A1)}$$

$$\max B_1(1) = \sum_{i=1}^{l-1} \sum_{j=1}^{m_i} i^{-1}, \qquad \text{(A2)}$$

where $l = \text{INT}(\log t / \log 2 + 0.9999)$, $m_i = \text{INT}[(t + 2^{i-1} - 1)/2^i]$, and

$$F(t, i) = \begin{cases} 2^i & \text{if } t \text{ is even and can be factorized by } 2^{i-1}, \text{ i.e., } (t \bmod 2^{i-1}) = 0, \\ \begin{cases} 2^i & \text{for all but the last } F(t, i) \text{ for a given } i, \\ 2^{i-1}[(t \bmod 2^{i-1}) + 2^{i-1}] & \text{for the last } F(t, i). \end{cases} \end{cases}$$

For multifurcating trees with $k < (t - 1)$ internal nodes, sum only the first $k - 1$ terms of Appendix Equations A1 and A2.

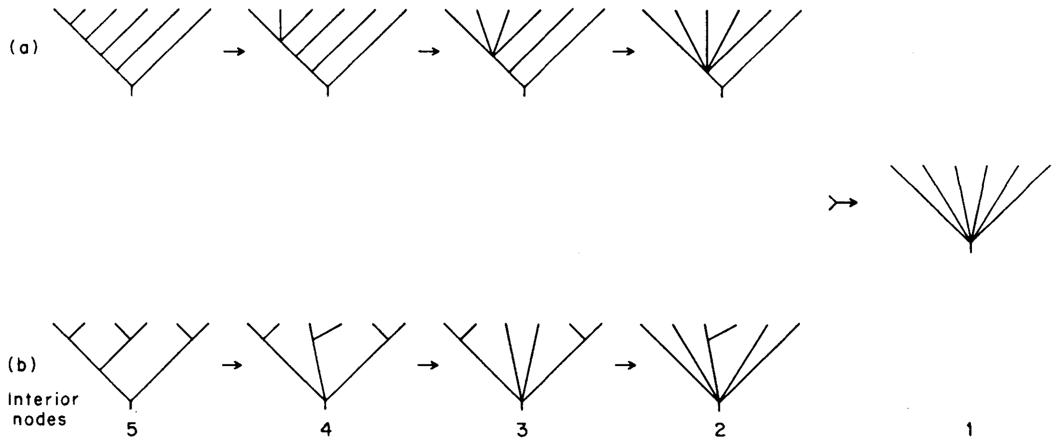

FIG. 2. Changing resolved subsets into unresolved ones to obtain the completely unresolved tree starting with the most unbalanced tree (a) and the most balanced tree (b) for a given number of interior nodes. Note that the tree with two interior nodes of (b) is isomorphic with tree 32 of Figure 1.

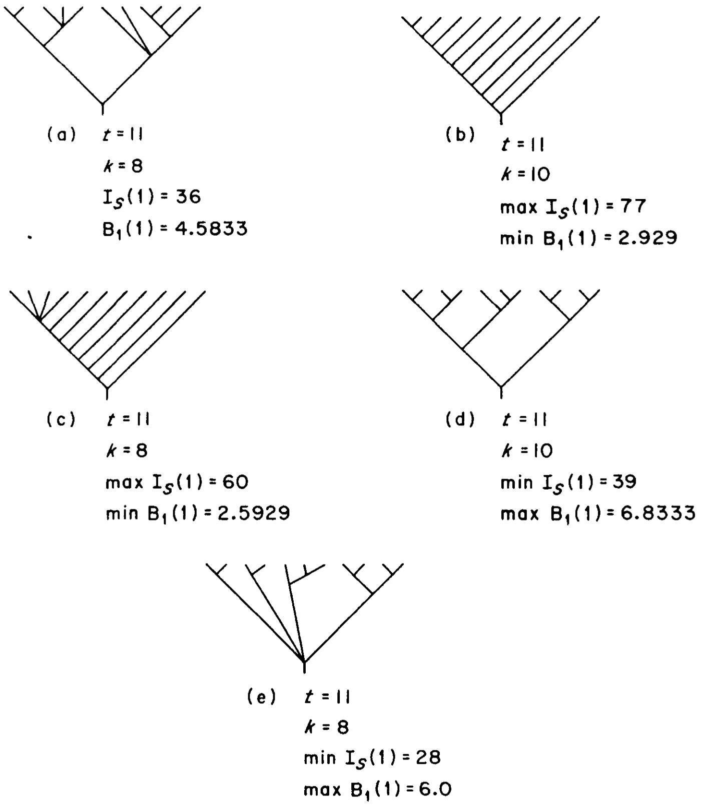

FIG. 3. An example to demonstrate how to draw the most unbalanced (b) and most balanced (d) binary trees, and their most unbalanced multifurcating tree (c) and most balanced multifurcating tree (e) with the same number of interior nodes as the tree to be evaluated (a). The numbers *t*, *k*, and various indices of I_S and B_1 for each tree are computed based on the algorithms in the text or Appendix. Resolved trees (b) and (d) are transformed into unresolved ones (c) and (e), respectively, following the procedure shown in Figure 2.

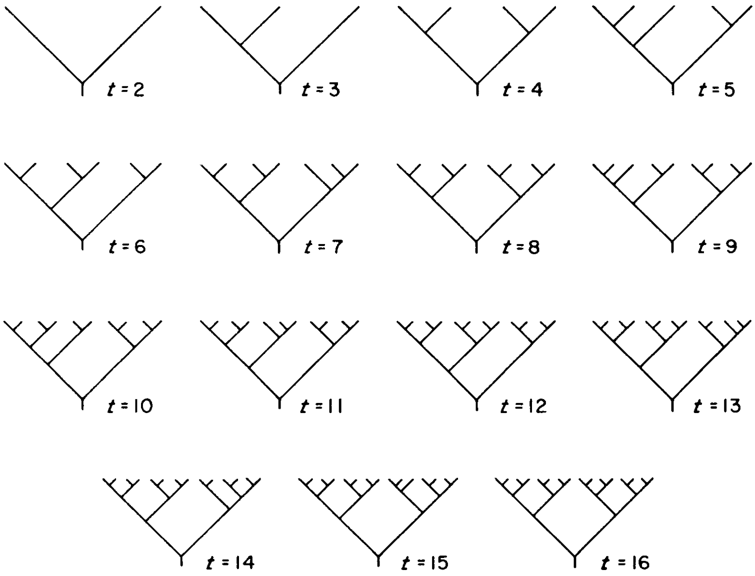

FIG. 4. Examples of the most balanced binary trees for OTU numbers t from 2 to 16.

In the above equations l is the total number of hierarchic levels (including the root) of the most balanced *binary* tree. It is equivalent to max N_i as this symbol is employed in Text Equation 1. Figure 4 is an example of the most balanced trees for t from 2 to 16. *Perfectly* balanced trees are restricted to OTU numbers 2, 4, 8, 16, 32, . . . , 2^l ($l = 1, \ldots, \infty$). Adding any one additional OTU to 2^l will increase the total number of levels by one, i.e., $l + 1$. So, for any tree with a given number of OTUs t, between $2^{l-1} + 1$ and 2^l, the total number of levels should equal l. From these relations we obtain the expression for l given above. Thus, for tree d in Figure 3, l = INT(log 11/log 2 + 0.9999) = INT(4.4593) = 4.

Quantity m_i given above is the number of subsets on each level i of the most balanced *binary* tree. For any tree, the size of subsets on the first level always equals 2. So, the total number of subsets m_1 on the first level is always equal to INT(t/2). The formula given above furnishes the number of subsets m_i for any level i.

For tree d in Figure 3 as an example, m_1 = INT[(11 + 2^0 − 1)/2^1] = INT(5.5) = 5; m_2 = INT(3.0) = 3; m_3 = INT(1.75) = 1; m_4 = INT(1.125) = 1. Then the sequence of all subsets j of the tree can be numbered from 1 to k (=t − 1), starting from the subsets on level 1 to the subsets on level l.

Function F(t, i) gives the size for any one subset j on its level i for the most balanced *binary* tree.

As an example of applying Appendix Equations A1 and A2, we show, for tree d in Figure 3, min $I_5(1)$ = 11 + 2 + 2 + 2 + 2 + 2 + 4 + 4 + 3 + 7 = 39; max $B_1(1) = \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} =$ 6.833. For tree e, min $I_5(1)$ = 11 + 2 + 2 + 2 + 2 + 2 + 3 + 4 = 28; max $B_1(1) = \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{2} =$ 6.0.