

**A Statistical Framework to Test the Consensus Among Additive Trees
(Cladograms)**



Francois-Joseph Lapointe; Pierre Legendre

Systematic Biology, Vol. 41, No. 2 (Jun., 1992), 158-171.

Stable URL:

<http://links.jstor.org/sici?sici=1063-5157%28199206%2941%3A2%3C158%3AASFTTT%3E2.0.CO%3B2-7>

Systematic Biology is currently published by Society of Systematic Biologists.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ssbiol.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

A STATISTICAL FRAMEWORK TO TEST THE CONSENSUS AMONG ADDITIVE TREES (CLADOGRAMS)

FRANCOIS-JOSEPH LAPOINTE AND PIERRE LEGENDRE

Département de Sciences biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec H3C 3J7, Canada

Abstract.—A statistical framework to test the consensus of dendrograms (=phenograms of some authors) is extended to the comparison of cladograms (additive trees). Additive trees can be compared through their associated path-length matrices. The procedure calls for a decomposition of additive trees into ultrametric and star components that are independently permuted and summed together to obtain randomized path-length matrices. This triple permutation test evaluates the null hypothesis that the trees under comparison are no more similar than random additive trees with a random topology, random labels, and randomized branch lengths. Along with the global test, the integrated approach rationalizes the simultaneous use of tests involving each component separately. The method is applied to kangaroo phylogenies to measure the congruence among trees derived from different character sets. [Additive trees; cladograms; consensus; dendrograms; kangaroo; Macropodidae; permutation test; statistical test; star trees.]

In a recent paper (Lapointe and Legendre, 1990), we introduced a statistical approach to test the consensus of two independent dendrograms (=phenograms of some authors) based on the comparison of their cophenetic matrices. We tested the null hypothesis that the two dendrograms under comparison are no more similar than randomly selected ultrametric trees, using a double permutation procedure involving the fusion levels as well as the positions of the labels on the leaves. Some workers have advocated, however, that dendrograms be used only when a molecular clock hypothesis is assumed (Blanken et al., 1982). In dendrograms, we consider all present-day taxa to be the same distance from the root representing the ancestor. On the other hand, additive trees (which we call cladograms) are more appropriate in evolutionary studies where one assumes that the branches of the trees might have evolved at different rates (Tateno et al., 1982). The tests available so far to compare cladograms are designed to take only the topology into account (Shao and Sokal, 1986; Page, 1988), with no consideration for the lengths of the branches. In this paper, we propose an extension of our earlier test that will allow the comparison of path-length matrices associated with the corresponding additive trees, taking into

consideration the topology, the label positions, and the branch lengths.

The problem of comparing cladograms can be treated as a randomization test. The statistical procedure consists of evaluating whether the trees under comparison are more similar than would be expected by chance alone. Such a randomization framework therefore requires an algorithm capable of generating random trees or their corresponding path-length matrices in an equiprobable manner (Furnas, 1984).

Another possibility would be to use the double permutation test (initially designed for the comparison of dendrograms) to compare cladograms. To do so, a prior transformation of the path-length matrices is required to obtain ultrametric approximations of the additive trees under comparison. An easy way to perform this task is to apply a clustering algorithm on the path-length matrices and to use the resulting cophenetic matrices in the comparison test. Another way would be to correct the actual cladograms for unequal rates of evolution (Klotz et al., 1979). The result of such a transformation is a dendrogram with a topology identical to that of the initial additive tree but with branch lengths satisfying the ultrametric property. Both of these methods to transform cladograms into dendrograms have drawbacks, however,

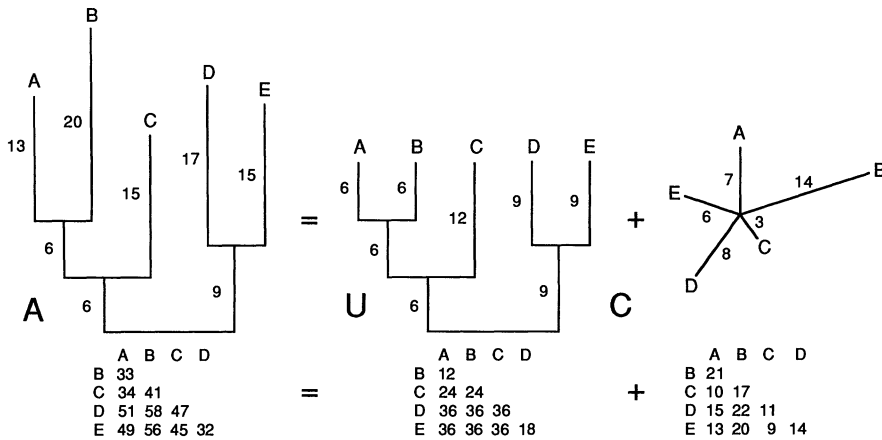


FIGURE 1. An additive tree and its associated path-length matrix A is decomposed into a dendrogram and its corresponding ultrametric matrix U plus a star tree along with its corresponding star distance matrix C.

because they do nothing more than compare degenerate additive trees. We prefer a procedure that directly generates random path-length matrices. Such matrices would be the basis for a simple test for cladogram comparison. This randomization is possible and straightforward when using the decomposition property for additive distances. An extension of the double permutation algorithm will then allow the generation and comparison of random path-length matrices representing random cladograms.

THE DECOMPOSITION PROPERTY FOR ADDITIVE DISTANCES

A given path-length matrix A (see definitions in Appendix) representing an additive tree (Fig. 1) can always be decomposed into ultrametric and star components U and C (Carroll, 1976; Sattah and Tversky, 1977; Carroll et al., 1984; Brossier, 1985; Furnas, 1989):

$$A_{ij} = U_{ij} + C_{ij} \tag{1}$$

for all pairs *i* and *j*, where

$$U_{ij} \leq \max(U_{ik}, U_{kj}) \tag{2}$$

for all triplets *i*, *j*, and *k*, and

$$C_{ij} = d_{ic} + d_{cj} \tag{3}$$

for all pairs *i* and *j*, where *c* is the root of

the tree. The decomposition is not unique, however, because it depends on the position of the root or decomposition center *c* (Klotz and Blanken, 1981). The root can be placed at any node of the tree or even between any two nodes; thus there are an infinite number of ways to decompose an additive tree, but with the following consequence (Carroll et al., 1984):

It may be the case, however, that for some path-length or additive trees there is no placement of the root that will result in a decomposition such that both the ultrametric and star component are nonnegative. In such cases there is a fundamental "tradeoff"—one may choose to make the ultrametric component positive but allow some negative star distances [see Equation 5 below], or make the star component positive but allow some negative values in the ultrametric component [Equation 6].

Yet, one can always find a positive constant to add to all branches of the additive tree, which will be sufficiently large to guarantee full positivity of both the ultrametric and star components. Therefore a matrix D* containing negative values is defined to be a relative distance matrix (Brossier, 1985) if and only if there exists a positive constant *k* such that

$$D^*_{ij} + k \geq 0 \tag{4}$$

for all pairs *i* and *j*. Similarly, one can de-

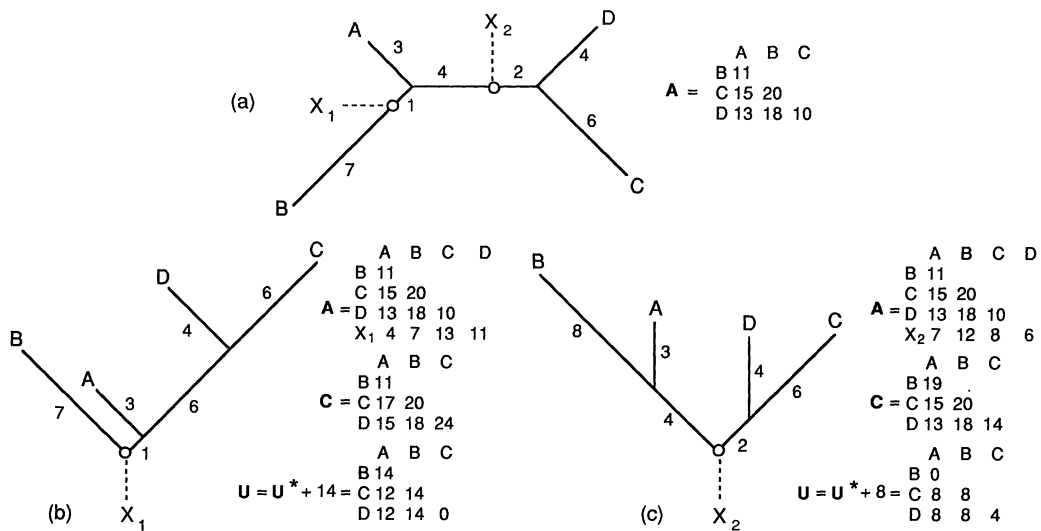


FIGURE 2. An unrooted additive tree (a) with its corresponding path-length matrix. The same additive tree is rooted at node X_1 (b) and at node X_2 (c) and represented with the matrices A , C , and U . The constants added to the matrices U^* in (b) and (c) are the smallest values that make all values of the resulting U matrices positive. Values C_{ij} are then computed as $A_{ij} - U^*_{ij} = A_{ij} - U_{ij} + k$.

fine the relative matrices A^* , U^* , and C^* up to an additive constant k .

The decomposition property holds in cases of relative distances if and only if either one of the following equations is verified:

$$A_{ij} = U_{ij} + C^*_{ij} \quad (5)$$

for all pairs i and j , or

$$A_{ij} = U^*_{ij} + C_{ij} \quad (6)$$

for all pairs i and j .

This property can be translated into an algorithm to decompose additive trees given a decomposition center selected among the nodes of the tree.

The Decomposition Algorithm

The decomposition of additive distances has been used in several algorithms to fit additive trees to distance matrices D (Carroll et al., 1984; Brossier, 1985; De Soete et al., 1985). In cases where D is actually satisfying the four-point metric (Equation 17, Appendix), the end purpose of the decomposition is to produce an additive tree representation of the given path-length matrix A . The "root and trim" procedure is performed in three steps (Furnas, 1989): (1)

selection of a decomposition center to root the tree, (2) pruning the branches of the additive tree at a fixed radius from the root to obtain a spherical tree satisfying the ultrametric inequality, and (3) rooting the pendant branches on a common center to form a star tree. After this decomposition, the topology is derived from the ultrametric fraction, following which the star part is added onto the branches. The result is an additive tree perfectly fitting the corresponding path-length matrix (Fig. 1).

Selecting the Decomposition Center

The decomposition property relies on a rooting operation performed on additive trees. Carroll et al. (1984) have shown that the position of the root will not affect the decomposition because the distances are correctly recovered in every case. Although there is no mathematical problem in the choice of any root, some useful criteria have been suggested to select the center of a path-length matrix.

The simplest case concerns rooted trees (i.e., cladograms) where the "real root" is actually known; the decomposition is performed from that root (Klotz et al., 1979).

Indeed, most studies in systematics are not dealing with undirected trees but rather with rooted phylogenetic trees. This root must be taken into account not only for decomposition purposes but also as a part of the tree topology. It is possible for a pair of rooted trees to have identical path-length matrices **A** but different directed topologies induced by the root position (Figs. 2b, 2c). This extra information does not explicitly appear in a standard patristic matrix (Fig. 2a). To allow the decomposition of directed additive trees, a supplementary object must be added to the path-length matrix to represent the root. The placement of this root induces a direction to the edges of an otherwise undirected tree but does not affect the other additive distances in the path-length (=patristic) matrix. Matrix **A** therefore contains an additional line representing the distances to the center from which the decomposition is performed (Figs. 2b, 2c).

All other situations concern unrooted trees; the center may therefore be selected among the actual nodes (terminal or internal) for decomposition purposes. In such cases, the root must be chosen so as to optimize the decomposition of a path-length matrix. One choice is based on selecting one of the objects as the "present-day ancestor" (Klotz and Blanken, 1981), or including an outgroup in the analysis that will allow rooting the tree at the junction between the outgroup vertex and the remainder of the tree. The alternative is to root the tree using some objective function. Brossier (1985) proposed such a mathematical criterion, which consists of rooting the tree at the node that minimizes the sum of distances to all other vertices. The center selected in this way represents the node closest to all the others. One may choose the opposite but more biologically relevant criterion that roots the tree at the node farthest from all others; that node is more likely to correspond to an "outgroup" decomposition center.

A TEST OF STATISTICAL SIGNIFICANCE

We suggest a statistical framework based on the decomposition described above to

test the consensus of cladograms displayed as path-length matrices. Any distance matrix (**D**, **A**, **U**, or **C**) can be treated the same way in a comparison test. The null hypothesis to be evaluated is that the matrices under comparison are no more similar to each other than randomly selected matrices would be. The test, therefore, evaluates whether the trees under comparison are more similar than expected by chance alone. The statistical evaluation proceeds as follows:

1. Compute a reference statistic (see below) between the actual input matrices.
2. Run the randomization procedure to generate a pair of random matrices representing the corresponding type of tree.
3. Compute the reference statistic between the random matrices.
4. Add one to a counter variable (originally set at zero) if the value of the statistic for the random trees is as high as or higher than the actual reference statistic between the trees under comparison.
5. Repeat steps 2–4 a large number of times (e.g., 999).
6. Add one to the counter variable to include the actual value of the statistic among those that could have been obtained under H_0 . Likewise, add one to the number of repetitions (Hope, 1968; Edgington, 1987).
7. Divide the counter variable by the number of repetitions (1,000 in this example) to obtain the probability that the null hypothesis is true.

This test must be one-tailed because we are only interested in knowing whether the true statistic is more extreme than for most of the trees generated at random (H_1). This is why we count only the values of the statistic that are larger than or equal to the reference value. The similarity of the initial pair of trees is declared statistically significant when most of the random trees are less similar than the pair under study.

NISI, THE COMPARISON CRITERION

We recommended in our double permutation test use of the *NISI* statistic (Nor-

malized Intermediate Similarity Index) to measure the association between two ultrametric matrices (Lapointe and Legendre, 1990). Here, the use of this composite index is extended to the comparison of additive distance matrices. Notice, however, that any other comparison statistic designed to take the metric values into consideration could be used in this test. For instance, the matrix correlation advocated by Rohlf (1982), which is also known as the standardized Mantel (1967) statistic, could have been computed directly on the distance matrices without any transformation. Another adequate consensus index would be the sum of squared differences.

The *NISI* coefficient (Equation 7.1) is derived from the Intermediate Dissimilarity Index of Faith and Belbin (1986) for which we have proposed a standardization form that produces values ranging from 0 to 1. In the following equations, **A** and **B** represent cophenetic matrices containing similarity values.

$$NISI = \frac{1 + [\sum(Cu - Co)/MAX]}{2}, \quad (7.1)$$

where

$$Cu = \text{minimum}(A_{ij}; B_{ij}) \quad (7.2)$$

for all *i* and *j*,

$$Co = |A_{ij} - B_{ij}| \quad (7.3)$$

for all *i* and *j*, and

$$MAX = \text{maximum}(Cu; Co). \quad (7.4)$$

This index is highly informative because it combines an unorganized complexity part (*Cu*) that reflects the common aspect of the matrices under comparison (Equation 7.2) with an organized complexity component (*Co*) measuring the differences of the two matrices (Equation 7.3). The *NISI* statistic was used in the case of dendrograms to compare cophenetic matrices containing similarity values. In the case of cladograms, it is conceptually easier to consider distances because the path-length matrices generally represent the number of mutations (character-state changes) along the branches connecting two objects. Notice however that the *NISI* coefficient is

designed to compare proximity matrices bounded between 0 and 1. The distance matrices under comparison must therefore be standardized a priori to provide meaningful *NISI* values.

Another modification of *NISI* is required to deal with distance matrices representing trees. The unorganized complexity part (Equation 7.2) of the index must be inverted to allow the comparison of distance matrices. The modified unorganized complexity measure therefore becomes

$$Cu^* = 1 - \text{maximum}(A_{ij}; B_{ij}) \quad (8)$$

for all *i* and *j*. The organized complexity component of the coefficient (Equation 7.3) remains unchanged. If we assemble all parts of *NISI*, we obtain a modified index *NISI** designed to compare distance matrices. Its simplest form is

$$NISI^* = \frac{1 + [\sum(Cu^* - Co)/MAX^*]}{2}, \quad (9.1)$$

where

$$MAX^* = \text{maximum}(Cu^*; Co). \quad (9.2)$$

The properties of *NISI** computed over the standardized matrices **A** and **B** are not modified in this form adapted for distances. In particular, *NISI** is still a similarity coefficient, so that identical trees will always have an *NISI** value of 1.

SEPARATE TESTS FOR THE ULTRAMETRIC AND STAR COMPONENTS

One interesting aspect of the decomposition method resides in its capacity to separate additive distances in two components that can be tested individually, if the problem calls for that. The evolutionary biologist may wish to compute the test over one component only or over both separately. This dual approach is quite appealing because it can provide more information than comparisons involving the path-length matrices alone and may thus allow a better evaluation of the aspects that the evolutionary trees under comparison have in common.

Comparing the Ultrametric Components

Whereas the star distances do not provide information about the hierarchical

subsets of objects, the ultrametric component illustrates the bifurcation sequence of the evolutionary tree; it portrays the topology of the relationships among the taxa, but it also represents something else. There is a relationship between the ultrametric component and Farris's (1977, 1979) special similarity,

$$T_{ij} = (C_{ij} - A_{ij})/2 \quad (10)$$

for all pairs i and j , which in our case is simply half the value of $-U^*_{ij}$. For a rooted tree, T_{ij} is the path length from the root to the most recent common ancestor of i and j (MRCA $_{ij}$). Testing the ultrametric fraction of two cladograms requires comparison of the tree topologies (branching patterns) and the relationships between MRCAs independently from the rate of evolution of the terminal branches. The null hypothesis in that situation is that the two dendrograms under comparison (each represented by its ultrametric U or special similarity T matrix) are no more similar to each other than are random dendrograms sampled equiprobably from a uniform distribution of trees. Such random dendrograms can be generated equiprobably (Lapointe and Legendre, 1991) using vectors of random distances (complete randomization), or they can result from a permutation of the actual vectors of branch lengths read from the input trees (constrained randomization). The comparison in the constrained situation is performed through the double permutation test proposed by Lapointe and Legendre (1990), which simply permutes the ultrametric fraction and the labels of the tree without any consideration for the star component; the reference statistic is computed over the ultrametric component matrices only. The probability for acceptance of the null hypothesis is calculated by counting the number of random dendrogram pairs with a value of the statistic as high as or higher than the actual trees and dividing by the total number of permutations.

Comparing the Star Components

The star distances, on the other hand, represent the different rates of evolution of the present-day taxa (including the rates

of evolution of both the terminal and the internal taxa) without information regarding the topological relations among them (Klotz et al., 1979). Testing the consensus between two star distance matrices evaluates whether the two cladograms have present-day taxa that have evolved at the same rate under different character sets, subject to random statistical error in the evaluation of these rates. The null hypothesis is that the two star components are no more similar than random star distance matrices would be if sampled uniformly throughout their distribution. Here again, one could call for a complete or a constrained randomization approach for the star trees. In the constrained situation, the test is reduced to its most simple expression because only one procedure is required to perform this "single permutation test." One can perform a permutation of vector S containing the distances from the root to every taxa; the permuted star distances are used thereafter to rebuild a random matrix C . This procedure is sufficient to randomize and generate any star-distance matrix corresponding to a given vector S . The test is computed in the usual way, using any given consensus index (e.g., *NISI*) and calculating the probability of H_0 by dividing the counter variable by the number of permutations. Notice that a Mantel test (1967) based on the star matrices C evaluates exactly the same null hypothesis. The Mantel procedure compares two distance matrices by permuting at random the rows and columns of these matrices. In our situation involving a star tree, such a randomization of the matrix is identical to relabeling the leaves of the tree at random, except the statistic that we use may differ.

THE COMBINATION PROPERTY AND THE COMPARISON OF PATRISTIC MATRICES

We have proposed so far a pair of tests to evaluate the consensus of components extracted from additive trees by decomposition. We now use the reciprocal of the decomposition property (Equation 1) to introduce a test to compare path-length ma-

trices globally. The combination property states that the sum of any ultrametric and star distance matrices represents an additive distance matrix (Brossier, 1985):

$$U_{ij} + C_{ij} = A_{ij} \quad (11)$$

for all pairs i and j . That means that the generation of random patristic matrices can be accomplished by generating random ultrametric and star matrices, followed by a recombination into a path-length matrix or by the construction of an equivalent additive tree from the permuted U and C components, as described above. It is therefore possible, now that we know how to generate random additive trees, to design a randomization test to compare cladograms. This global consensus test evaluates the same general null hypothesis as the component tests: the pair of cladograms under comparison are no more similar than randomly selected pairs of additive trees would be.

The Combination Algorithm

The complete randomization design proceeds in (1) generating a random dendrogram, (2) generating a random star tree, (3) summing the two matrices computed in steps 1 and 2 to obtain an additive distance matrix A , and (4) assigning the objects randomly onto the terminal nodes of the tree, which simply corresponds to labeling the rows and columns of the matrix at random. This procedure produces random path-length matrices, each representing a unique cladogram with a random topology, random length of the branches, and random position of the labels on the leaves. In statistical testing, the patristic distances generated should not take any possible value; instead, they should be comparable to the actual matrices under study: Constraints, therefore, must be added to the randomization algorithm. To insure that the random matrices remain comparable with the actual matrices, the random trees are constrained to have the same number of objects n as the real problem and the random matrices to have a fixed sum of distances. Actually, we are trying to distribute the character-state changes (mutation events)

at random onto the branches of the tree under the constraints that the total number of changes is constant over all random trees and is the same as that of the input trees and that branch lengths are nonnegative:

$$\sum \sum A_{ij} = a, \quad (12)$$

where a is a constant and $A_{ij} \geq 0$, for all pairs i and j in the matrix. The constrained randomization results from merging the decomposition and combination procedures. First, the additive matrices to be compared are read in, from which the ultrametric and star distances are extracted using the decomposition algorithm. The two distance components (U and C) are then permuted at random and recombined, followed by random relabeling, to generate random additive matrices comparable to the actual ones. The resulting assemblage is a triple permutation procedure, as compared with the double permutation process (Lapointe and Legendre, 1990), because three randomizations are necessary here. The complete algorithm, which is applied in turn to each tree, can be summarized as follows:

1. Input the path-length matrix A .
2. Choose the decomposition center c .
3. Read the n star distances from c to the n terminal nodes and store the values in vector S .
4. Compute the star matrix C from vector S .
5. Remove the star component C from the additive distance A to obtain U .
6. Read the $(n - 1)$ fusion level values from the resulting ultrametric fraction U of matrix A .
7. Execute the double permutation procedure on the vector of fusion values (see Lapointe and Legendre [1990, 1991] for details):
 - a. Permute at random the vector of fusion levels from U .
 - b. Rebuild the matrix corresponding to the permutation order using the ultrametric property.
 - c. Permute at random the rows and columns of the random matrix or, alternatively, relabel the leaves at random.

8. Permute the star distance vector \mathbf{S} at random; recompute the star matrix \mathbf{C} .
9. Add the star \mathbf{C} and ultrametric \mathbf{U} components to obtain the randomized matrix \mathbf{A}' .

The triple permutation algorithm does not necessarily guarantee that all branch lengths will be positive in sign. The constraints imposed can lead to negative path lengths that violate the conditions characterizing distance matrices. Even positive matrices may fail to satisfy the triangle inequality implied by the four-point metric (Furnas, 1989:45). Three options are thus open to deal with such matrices:

1. Negative branch lengths are allowed in the generation process.
2. Matrices bearing negative values, or not satisfying the triangle inequality for some other reason, are deleted from the generation process (R. D. M. Page, pers. comm.). This can be done by testing for positivity (or the triangle inequality) after each permutation or by doing restricted permutations and combinations only among branches of \mathbf{U} and \mathbf{C} that will produce satisfying path-length matrices.
3. A positive constant k is added to the initial distances so that all permutations and combinations lead to matrices that satisfy the triangle inequality; k is chosen to be larger than or equal to the absolute value of the largest negative value in \mathbf{U}^* or \mathbf{C}^* , thus making all branch lengths positive.

By definition, the four-point metric that characterizes additive trees is only fixed up to an additive constant, so that negativity need not be taken as a contradiction of the four-point metric broadly defined (Carroll et al., 1984). Such additive trees bearing negative branch lengths can be interpreted as relative path-length matrices with distances measured on a scale where k , which is the origin of the distance scale, will have a negative value instead of zero (Brossier, 1985). The triple permutation test remains invariant over any value of k when relative path-length matrices are compared be-

cause the input matrices must be standardized between 0 and 1 before computing the *NI* statistic, so that options 1 and 3 lead to exactly the same results. Restricting the generation process to positive trees only (option 2) may lead to a statistically different outcome, however, because that test is based on a different reference population. All three options lead to the same result if and only if $U_{ij} \leq C_{ij}$ for every i and j .

A compiled computer program for performing the triple permutation procedure (for Macintosh) and a detailed PASCAL source code listing are available upon request from the first author. The algorithm produces random patristic matrices from which the test is computed.

THE INTEGRATED APPROACH

Different hypotheses can be assessed using either the global test or each of the simple component tests. A complete statistical comparison of the phylogenetic trees is obtained by computing all of these. The integrated approach therefore consists of simultaneously computing the global test and both component tests. The rationale for such an integrated comparison scheme becomes obvious when all the statistically possible outcomes are examined. Three tests are applied; for each of them, two alternatives are possible (H_0 is rejected or H_0 is accepted). The number of possible combinations of results is $2^3 = 8$. Even though all eight situations are theoretically possible, it is impossible in practice for the global test to accept H_0 when both component tests have previously rejected their null hypotheses. All the other situations can occur in actual problems. The following examples illustrate these seven cases.

Consider a vector $\mathbf{V} = \{1, 3, 4, 7\}$, representing the fusion levels of a dendrogram, and a vector $\mathbf{S} = \{0, 1, 4, 6, 8\}$, representing the branch lengths of a star tree; also consider a set of matrices \mathbf{A} that result from the combinations of the randomized component vectors \mathbf{V} (used to construct \mathbf{U}) and \mathbf{S} (used to construct \mathbf{C}). Each random additive tree (Fig. 3) is obtained by summing the randomized component matrices. To compare these additive trees using

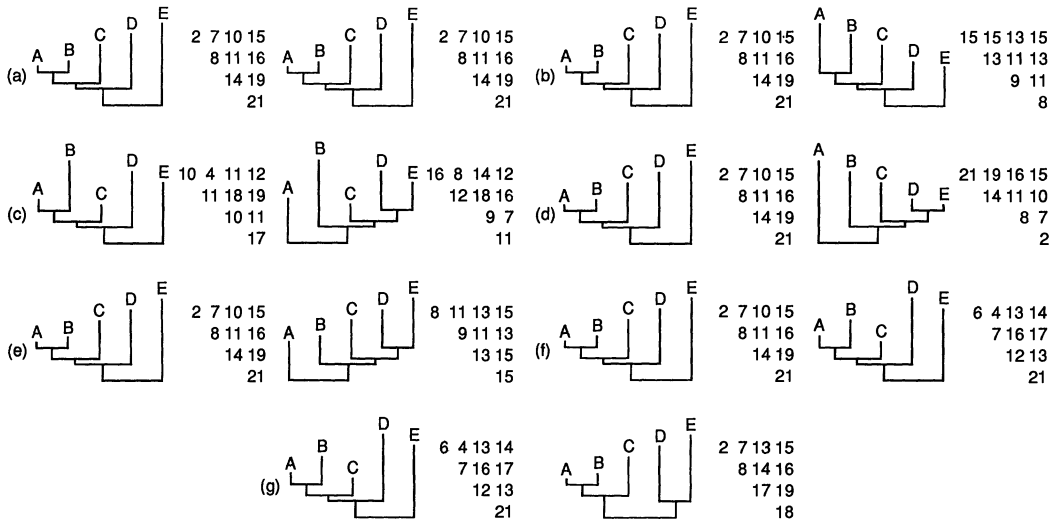


FIGURE 3. Seven pairs of additive trees compared using the integrated approach (see text for details and Table 1 for results).

the integrated approach, the path-length matrices are decomposed, then the ultrametric and star fractions are tested individually along with a global test of the additive trees. Table 1 presents the results of the seven integrated tests comparing the pairs of trees depicted in Figure 3, showing that all seven situations can actually occur.

As illustrated by this example, the "combined" results provide more information than does the global test alone. The results of the simultaneous tests become important when attempting a biological interpretation of the results. One can conclude for instance that two phylogenetic trees are

different although they contain species that have evolved at the same rate (Fig. 3c) or that a pair of trees should be considered similar even though they possess different branching patterns (Fig. 3e). In any case, we recommend using the integrated approach in real problems to bring more depth to the comparison of additive trees.

APPLICATION TO KANGAROO EVOLUTIONARY TREES

The following example is a typical application of the triple permutation test to compare additive trees. Following the example developed by Lapointe and Legen-

TABLE 1. Results (Pearson product-moment correlations, r , and associated probabilities, P) of the tests comparing the trees presented in Figure 3 (a-g). In the summary of the three tests, a significant result is represented by a 1.

Trees	Component tests				Global test, path-length matrix A		Significance summary		
	Ultrametric matrix U		Star matrix C		r	P	U	C	A
	r	P	r	P					
a	1.000*	0.000	1.000*	0.000	1.000*	0.000	1	1	1
b	1.000*	0.000	0.007	0.992	0.167	0.920	1	0	0
c	0.264	0.998	1.000*	0.000	0.804	0.080	0	1	0
d	0.264	0.998	0.007	0.992	0.051	0.992	0	0	0
e	0.264	0.998	1.000*	0.000	0.952*	0.010	0	1	1
f	1.000*	0.000	0.810	0.134	0.910*	0.022	1	0	1
g	0.821	0.084	0.810	0.134	0.854*	0.036	0	0	1

* Significant values of the Pearson correlation coefficient tested by randomization at the alpha significance level of 0.05; 499 permutations were performed for each test.

dre (1990), we want to verify whether kangaroo phylogenies based on different character sets are statistically congruent. Three different trees are compared in this example; two of them are additive trees, and the last one is a dendrogram.

Comparing a Pair of Cladograms

Baverstock et al. (1989) recently published a reanalysis of the kangaroo phylogeny based on albumin immunological relationships. Their paper provides an additive tree representing the evolutionary distances among 14 species of the family Macropodidae (Fig. 4a). We want to test the resemblance between this immunology-based tree and another one based on electrophoretic characters. This second tree (Fig. 4b) was derived from the results published by Richardson et al. (1973) concerning kangaroo electrophoresis patterns for seven proteins; to obtain that additive tree, we applied the Fitch and Margoliash (1967) method to a distance matrix computed over the authors' data set (1 - Jaccard similarity). The path-length matrices were computed for the 10 species that both trees have in common and were compared using the integrated approach of the triple permutation test. The decomposition center of each matrix was selected to represent the terminal taxon farthest from all other species: the "Baverstock tree" was thus rooted on *Aepyprimmus* and the "Richardson tree" on *Petrogale*. Five thousand permutations were calculated for the tests. The results presented in Table 2 show that the global test is significant ($P < 0.001$) even though both component tests accepted the null hypothesis ($P > 0.05$). Similar results were obtained using *NISI**, the cross-product-

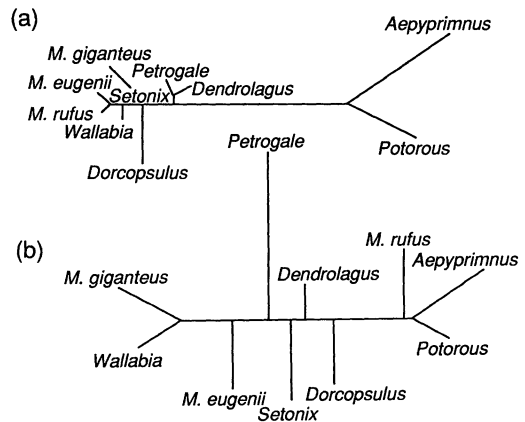


FIGURE 4. Additive tree representations of the evolutionary relationships among 10 species of the Macropodidae. (a) From Baverstock et al. (1989). (b) Derived from the Richardson et al. (1973) data set.

moment correlation, and the sum of squared differences.

This statistical conclusion might have been the result of the different decomposition centers of the matrices, which resulted in different component residuals. To assess whether an identical decomposition modifies the outcome of the tests, we ran another set of tests after selecting the same root (i.e., *Aepyprimmus*) for both trees. The new results (Table 2) show significance for all tests. This apparent contradiction between the two sets of tests illustrates the importance of the selection of an appropriate root in the triple permutation test (see Fig. 2). Different roots modify the component matrices and, consequently, also the results of the component tests. However, the global test will give the same results in all cases because the path-length matrix is invariant under different decomposition centers.

TABLE 2. Results of the tests comparing the trees presented in Figure 4 (a, b).

Roots	Component tests				Global test, path-length matrix A	
	Ultrametric matrix U		Star matrix C		<i>NISI*</i>	<i>P</i>
	<i>NISI*</i>	<i>P</i>	<i>NISI*</i>	<i>P</i>		
<i>Aepyprimmus</i> (a) and <i>Petrogale</i> (b) ^a	0.313	0.112	0.206	0.955	0.871	0.001
<i>Aepyprimmus</i> (both) ^b	0.301	0.002	0.364	0.039	0.871	0.001

^a Trees (a) and (b) were decomposed based on the *Aepyprimmus* and *Petrogale* root, respectively.

^b Both trees were decomposed using *Aepyprimmus* as the root.

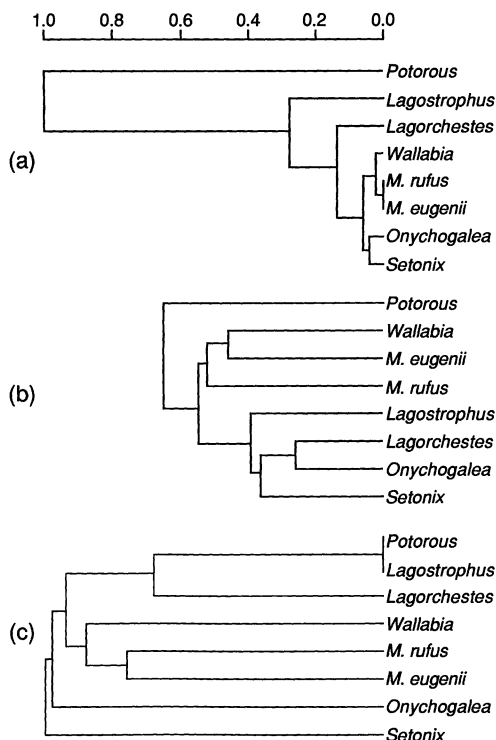


FIGURE 5. Dendrograms depicting evolutionary distances among seven species of Macropodidae. (a) The ultrametric component extracted from the Baverstock et al. (1989) additive tree, rooted on the farthest species (*Potorous*). (b) Modified from Kirsch (1977). (c) The ultrametric component extracted from the Baverstock et al. (1989) additive tree, rooted on the most central species (*Setonix*).

Comparing a Cladogram to a Dendrogram

In this second example, we compare the Baverstock et al. (1989) additive tree to the serology-based dendrogram proposed by Kirsch (1977). That specific problem calls for a special test designed to compare only the ultrametric fraction of the Baverstock et al. (1989) path-length matrix (Fig. 5a), after correction for unequal rates of evolution (Klotz et al., 1979), to the cophenetic matrix associated with the Kirsch tree (Fig. 5b). After pruning the trees to keep only the species in common, we compared the pair of matrices representing the evolutionary relationships among eight species of kangaroos. Because ultrametric matrices were to be compared, the double permutation test (Lapointe and Legendre, 1990) was used instead of the triple permutation

procedure. The reference statistic was computed over 5,000 pairs of randomized dendrograms to obtain a distribution of the comparison index. The comparison test was significant ($NISI^* = 0.590$; $P = 0.016$) when *Potorous* was selected (Fig. 5a) to decompose the additive tree of Baverstock et al. (1989); other rootings lead to the opposite results. When *Setonix* was chosen (the most central species) as the root (Fig. 5c) instead of *Potorous* (the farthest species), the null hypothesis was accepted ($NISI^* = 0.096$; $P = 0.188$), implying that the "Baverstock ultrametric component" is not more related to the "Kirsch ultrametric tree" than random pairs of dendrograms would be.

These contradictory results illustrate once again the influence of the decomposition center on the outcome of the test. In the case of additive tree comparisons, root selection is less important because the global tests are not affected by the decomposition. However, ultrametric components may exhibit very different topologies, depending on selection of a "central" or an "outgroup" root. One possible way to deal with this situation may be to select each species in turn to decompose the tree; if at least one of the n tests rejects the null hypothesis, then the test may be declared significant. The alternative is to compute the test using a reference distribution based on random comparisons of dendrograms against cladograms (Lapointe and Legendre, 1992).

ACKNOWLEDGMENTS

We acknowledge the helpful comments of R. D. M. Page and J. A. W. Kirsch on earlier drafts of this manuscript. This work was supported by a NSERC scholarship to F.-J. Lapointe and by NSERC grant No. OGP0007738 to P. Legendre. This is contribution No. 384 of the Groupe d'Écologie des Eaux Douces, Université de Montréal.

REFERENCES

- BARTHÉLEMY, J.-P., AND A. GUÉNOCHE. 1988. Les arbres et les représentations des proximités. Méthode Programmes, Masson, Paris.
- BARTHÉLEMY, J.-P., AND N. X. LUONG. 1986. Représentations arborées de mesures de dissimilarité. Stat. Anal. Données 11(1):20-41.
- BAVERSTOCK, P. R., B. J. RICHARDSON, J. BIRRELL, AND M. KRIEG. 1989. Albumin immunologic relationships of the Macropodidae (Marsupialia). Syst. Zool. 38:38-50.

- BLANKEN, R. L., L. C. KLOTZ, AND A. G. HINNEBUSH. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. *J. Mol. Evol.* 19:9-19.
- BROSSIER, G. 1985. Approximation des dissimilarités par des arbres additifs. *Math. Sci. Hum.* 91:5-21.
- BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387-395 in *Mathematics in archeological and historical sciences* (F. R. Hodson, D. G. Kendall, and P. Tautou, eds.). Edinburgh Univ. Press, Edinburgh.
- BUNEMAN, P. 1974. A note on the metric properties of trees. *J. Comb. Theory (B)* 17:48-50.
- CARROLL, J. D. 1976. Spatial, non-spatial and hybrid model for scaling. *Psychometrika* 41:439-463.
- CARROLL, J. D., L. A. CLARK, AND W. S. DESARBO. 1984. Tree representations of three-way proximity data. *J. Classif.* 1:25-74.
- DE SOETE, G., W. S. DESARBO, AND J. D. CARROLL. 1985. Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. *J. Classif.* 2:173-192.
- DOBSON, A. J. 1974. Unrooted trees for numerical taxonomy. *J. Appl. Prob.* 11:32-42.
- EDGINGTON, E. S. 1987. *Randomization tests*, 2nd edition. Marcel Dekker, New York.
- FAITH, D. P., AND L. BELBIN. 1986. Comparison of classifications using measures intermediate between matrix dissimilarity and consensus similarity. *J. Classif.* 3:257-280.
- FARRIS, J. S. 1973. On comparing the shape of taxonomic trees. *Syst. Zool.* 22:50-54.
- FARRIS, J. S. 1977. On the phenetic approach to vertebrate classification. Pages 823-850 in *Major patterns in vertebrate evolution* (M. K. Hecht, P. C. Goody, and B. M. Hecht, eds.). Plenum, New York.
- FARRIS, J. S. 1979. The information content of the phylogenetic system. *Syst. Zool.* 25:260-270.
- FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. *Science* 155:279-284.
- FURNAS, G. W. 1984. The generation of random, binary unordered trees. *J. Classif.* 1:187-233.
- FURNAS, G. W. 1989. Metric family portraits. *J. Classif.* 6:7-52.
- GOWER, J. C., AND P. LEGENDRE. 1986. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3:5-48.
- HARTIGAN, J. A. 1967. Representation of similarity matrices by trees. *J. Am. Stat. Assoc.* 62:1140-1158.
- HOPE, A. C. A. 1968. A simplified Monte Carlo significance test procedure. *J. R. Stat. Soc. Ser. B* 30:582-598.
- JARDINE, C. J., N. JARDINE, AND R. SIBSON. 1967. The structure and construction of taxonomic hierarchies. *Math. Biosci.* 1:173-179.
- JOHNSON, S. C. 1967. Hierarchical clustering schemes. *Psychometrika* 32:241-254.
- KIRSCH, J. A. W. 1977. The comparative serology of Marsupialia, and a classification of marsupials. *Aust. J. Zool. Suppl. Ser.* 52:1-152.
- KLOTZ, L. C., AND R. L. BLANKEN. 1981. A practical method for calculating evolutionary trees from sequence data. *J. Theor. Biol.* 91:261-272.
- KLOTZ, L. C., N. KOMAR, R. L. BLANKEN, AND R. M. MITCHELL. 1979. Calculation of evolutionary trees from sequence data. *Proc. Natl. Acad. Sci. USA* 76:4516-4520.
- LAPOINTE, F.-J., AND P. LEGENDRE. 1990. A statistical framework to test the consensus of two nested classifications. *Syst. Zool.* 39:1-13.
- LAPOINTE, F.-J., AND P. LEGENDRE. 1991. The generation of random ultrametric matrices representing dendrograms. *J. Classif.* 8:177-200.
- LAPOINTE, F.-J., AND P. LEGENDRE. 1992. Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Syst. Biol.* 41 (in press).
- LE CALVÉ, G. 1985. Distances à centre. *Stat. Anal. Données* 10(2):29-44.
- MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27:209-220.
- PAGE, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. *Syst. Zool.* 37:254-270.
- PATRINOS, A. N., AND S. L. HAKIMI. 1972. The distance matrix of a graph and its tree realization. *Q. Appl. Math.* 30:255-269.
- PHIPPS, J. P. 1971. Dendrogram topology. *Syst. Zool.* 20:306-308.
- PRUZANSKY, S., A. TVERSKY, AND J. D. CARROLL. 1982. Spatial versus tree representations of proximity data. *Psychometrika* 47:3-19.
- RICHARDSON, B. J., P. G. JOHNSTON, P. CLARK, AND G. B. SHARMAN. 1973. An evaluation of electrophoresis as a taxonomic method using comparative data from the Macropodidae (Marsupialia). *Biochem. Syst.* 1:203-209.
- ROHLF, F. J. 1982. Consensus indices for comparing classifications. *Math. Biosci.* 59:131-144.
- SATTAH, S., AND A. TVERSKY. 1977. Additive similarity trees. *Psychometrika* 42:319-345.
- SHAO, K., AND R. M. SOKAL. 1986. Significance tests of consensus indices. *Syst. Zool.* 35:582-590.
- SIBSON, R. 1972. Order invariant methods for data analysis. *J. R. Stat. Soc. Ser. B* 34:311-349.
- SMOLENSKII, Y. A. 1969. A method for linear recording of graphs. *USSR Comput. Math. Phys.* 2:396-397.
- SNEATH, P. H. A., AND R. R. SOKAL. 1973. *Numerical taxonomy*. W. H. Freeman, San Francisco.
- SOKAL, R. R., AND F. J. ROHLF. 1962. The comparison of dendrograms by objective methods. *Taxon* 9:33-40.
- TATENO, Y., M. NEI, AND F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* 18:387-404.
- WATERMAN, M. S., T. F. SMITH, M. N. SINGH, AND W. A. BEYER. 1977. Additive evolutionary trees. *J. Theor. Biol.* 64:199-213.
- ZARETSKII, K. 1965. Constructing a tree on the basis of a set of distances between the hanging vertices. *Usp. Mat. Nauk* 20:90-92 [in Russian].

Received 7 October 1991; accepted 16 January 1992

APPENDIX
CLADOGRAMS, DENDROGRAMS,
AND BUSHES

Consider a matrix **D** containing values d_{ij} for every pair of objects i and j . This matrix is defined as a dissimilarity matrix when the following properties are satisfied: symmetry,

$$d_{ij} = d_{ji} \quad (13)$$

for all i and j ; positiveness,

$$d_{ij} > 0 \quad (14)$$

for $i \neq j$; and definiteness,

$$d_{ij} = 0 \quad (15)$$

for $i = j$. Furthermore, we consider the dissimilarity matrix **D** to represent a distance matrix **M** when it meets the triangle inequality:

$$d_{ij} \leq d_{ik} + d_{kj} \quad (16)$$

for all triplets i, j , and k . Because the triangle inequality property defines metric dissimilarities, **D** can then also be said to be metric (see, for example, Sneath and Sokal, 1973). When the fourth condition is not satisfied, the distance is said to be a semimetric.

Any matrix **D** meeting those four properties always allows an Euclidean representation of $\sqrt{\mathbf{D}}$ in a space of dimension $n - 1$ (Gower and Legendre, 1986), represented here by an Euclidean matrix **E**, where n is the number of points (objects, OTUs, . . .) under study. A nonspatial alternative representation of matrix **D** (Pruzansky et al., 1982) is a tree, which is a connected graph without cycle. Not all distance matrices can be exactly represented by trees, however. Additional properties should be respected depending on which tree representation is sought. Three types of trees will be considered here: additive trees, dendrograms, and star trees. For a complete description of these trees (definitions, properties, and relationships), see Barthélemy and Guénoche (1988).

Additive Trees

A distance matrix **D** can be represented in the form of an additive tree (Fig. 1) if the four-point condition (Equation 17) is satisfied in addition to properties 13-16 mentioned above (Buneman, 1971):

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk} \quad (17.1)$$

for all 4-tuples i, j, k , and l , or

$$d_{ij} + d_{kl} \leq \max[(d_{ik} + d_{jl}), (d_{il} + d_{jk})] \quad (17.2)$$

for all 4-tuples i, j, k , and l . Statement 17.1 implies that any quadruplet of objects can be labeled so as to satisfy that condition for the three possible sums of two distances (there are six distances among four objects); 17.2 implies that the three sums of the opposite edges of the tetrahedron $\{i, j, k, l\}$ define an isosceles triangle (Patrinos and Hakimi, 1972; Buneman, 1974; Dobson, 1974). The four-point condition or quadruplet inequality also implies the triangle inequality (condition 16) and can thus be called the four-point metric.

If the branches of an additive tree are weighted (i.e., have values attached to the branches), the distance between two terminal nodes may be defined as the sum of the lengths of the branches connecting these two nodes. The resulting additive matrix **A** (Fig. 1) containing all such distances is the path-length matrix. When the branches of the tree are not weighted, the additive distance between two objects is simply the topological distance counting the number of edges (=branches) between two terminal nodes (Phipps, 1971; Farris, 1973). The distance matrix **B** in that simpler case is the branch-distance matrix.

Rooted trees have a node defined as the root, which is the ancestor to all other terminal nodes. Roots are selected among internal nodes but can also be associated with terminal nodes in specific cases. Cladograms used in cladistics, as well as other forms of phylogenetic trees, are usually represented in the form of rooted additive trees where the distance between two objects corresponds to the sum of the character-state changes along the path of branches connecting these two objects ("patristic" distances). Path-length matrices associated with cladograms are often called patristic matrices.

Dendrograms

A matrix **D** can be represented in the form of a dendrogram (Fig. 1) if the following ultrametric inequality is met (Hartigan, 1967; Johnson, 1967):

$$d_{ij} \leq \max(d_{ik}, d_{kj}) \quad (18)$$

for all triplets i, j , and k . This condition implies that for every triplet of points, the two largest distances are equal (i.e., every triangle $\{i, j, k\}$ is isosceles). The distances meeting this condition are said to represent spherical trees, because every terminal node of the tree is equidistant from the root of the dendrogram. We use here the term dendrogram in a graph theoretic sense (Barthélemy and Guénoche, 1988:24) and not in a biological sense, where the term phenogram has been used to define these kinds of trees (Sneath and Sokal, 1973:58). A distance matrix (Fig. 1) meeting condition 18 is called an ultrametric **U** or a cophenetic matrix (Sokal and Rohlf, 1962). Notice that matrices meeting the ultrametric inequality (Equation 18) also satisfy the four-point condition (Equation 17) as well as the triangle inequality (Equation 16). Therefore, from all the matrices defined so far, we can characterize an inclusion order that encompasses all types of representations (Le Calvé, 1985):

$$\mathbf{D} \supset \mathbf{M} \supset \mathbf{E} \supset \mathbf{A} \supset \mathbf{U}. \quad (19)$$

It follows from that order that dendrograms meeting the ultrametric inequality (**U**, Equation 18) are special cases of additive trees satisfying the four-point metric (**A**, Equation 17) and that patristic matrices **A** are less restrictive than ultrametric matrices **U** (Sibson, 1972).

Previous workers have shown that every dendrogram is associated with a unique ultrametric (=cophenetic) matrix (Jardine et al., 1967) and that additive trees also are in one-to-one correspondence with path-length matrices (Waterman et al., 1977) or branch-distance matrices (Zaretzkii, 1965; Smolenskii, 1969).

From the latter considerations, any ultrametric distance can be uniquely defined by $(n - 1)$ values only, whereas $(2n - 3)$ distances are required for additive trees. Generating dendrograms therefore represents a simpler problem (fewer parameters to adjust) than generating additive trees (Barthélemy and Luong, 1986).

Star Trees

We also consider the star tree (Fig. 1) or singular tree (Sattah and Tversky, 1977). This representation corresponds to a tree with only one interior node (Le Calvé, 1985). This unique node is the center c of the tree. In star trees, the path between any two objects i and j must pass through the center c . A matrix \mathbf{D} is a star distance matrix, symbolized by \mathbf{C} (Fig. 1), when the following equation holds:

$$d_{ij} = d_{ci} + d_{cj} \quad (20)$$

for all i and j . Likewise, any star tree can be represented by a star distance vector \mathbf{S} containing the distances between the center c and every terminal node. Star distances therefore are uniquely defined by n values associated with the branches of the corresponding tree. Any matrix \mathbf{C} that satisfies the star condition also satisfies the four-point condition. Thus, another inclusion order characterizing the position of star distances can be established (Le Calvé, 1985):

$$\mathbf{D} \supset \mathbf{M} \supset \mathbf{E} \supset \mathbf{A} \supset \mathbf{C}. \quad (21)$$

That order shows that star trees are no more than degenerate additive trees. Also, no inclusion relation exists between star distances and ultrametric distances.