Syst. Zool., 39(1):1-13, 1990

A STATISTICAL FRAMEWORK TO TEST THE CONSENSUS OF TWO NESTED CLASSIFICATIONS

FRANÇOIS-JOSEPH LAPOINTE AND PIERRE LEGENDRE

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

Abstract.—We propose a method to compare rooted classifications when the fusion levels between OTUs are to be taken into account. This problem can be formulated as a statistical randomization test, that includes a double permutation procedure involving the generation of random dendrograms from ultrametric matrices. We test the null hypothesis stating that the two dendrograms under comparison are not more similar than dendrograms randomly generated in terms of three different aspects: topology, leaf positions, fusion level positions. The similarity beween nested trees is computed using a normalized form of the intermediate consensus index of Faith and Belbin (1986). A special case is discussed where limited permutations are required to test a conditional null hypothesis. This test is applied to kangaroo classifications to measure the congruence between dendrograms derived from different character sets. [Classification; consensus method; kangaroo; limited permutations; Macropodidae; nested tree; permutation test; statistical test.]

Résumé.—Nous proposons une méthode pour comparer des classifications enracinées lorsqu'on désire tenir compte des niveaux de fusion entre les objets. Ce problème peut s'exprimer en termes d'un test statistique impliquant une procédure à double permutation permettant la génération de dendrogrammes aléatoires à partir de matrices ultramétriques. Nous testons l'hypothèse nulle stipulant que les deux dendrogrammes comparés ne sont pas plus semblables entre eux que des dendrogrammes générés au hasard selon trois aspects: topologie, position des feuilles, position des niveaux de fusion. La similarité entre les arbres hiérarchiques emboîtés est calculée en utilisant une forme normalisée de l'indice de consensus intermédiaire de Faith et Belbin (1986). Nous discutons d'un cas particulier où il est nécessaire d'avoir recours à des permutations limitées pour tester une hypothèse nulle conditionnelle. Nous appliquons ce test à la classification des kangourous afin de mesurer la congruence entre des dendrogrammes basés sur différents jeux de caractères.

The task of early numerical taxonomy was to propose and evaluate methods of reconstructing evolution as well as techniques to classify present-day taxa. The next problem was to compare the methods that had been developed and the solutions they generated. This gave birth to several comparison schemes involving binary trees, nested trees, unrooted trees, classifications, networks and even unlabeled trees (Williams and Clifford, 1971; Adams, 1972; Hubert and Baker, 1977; Waterman and Smith, 1978; Margush and McMorris, 1981; Robinson and Foulds, 1981; Neumann, 1983; Stinebrickner, 1984; Day, 1985; Estabrook et al., 1985; Penny and Hendy, 1985). These methods were designed to find answers to an equally vast range of specific problems such as testing evolutionary hypotheses (Penny et al., 1982), evaluating taxonomic characters (Thorpe and Dickinson, 1988), comparing cladistic and phenetic methods (Sokal, 1983), measuring congruence for different datasets (Colless, 1980; Schuh and Polhemus, 1980), and so on. Notwithstanding this diversity of purposes in tree comparisons, most of the recent studies have revolved around two major questions:

1. Do different methods of tree reconstruction produce similar trees?
2. Do different groups of variables about the same OTUs lead to similar trees with a given method?

Consensus techniques were developed to answer these questions. Consensus indices measure the resemblance between trees, while consensus tree methods are used to produce a solution that reflects common aspects of two or more trees.

Despite all the work that has been done on consensus methods during the last decade, few statistical methods to compare taxonomic classifications have been proposed yet. Considering the recent developments in this area, the major concern of taxonomists should now be to test statistically their specific hypotheses about the consensus between trees. Some methods have already been developed to statistically compare special kinds of trees in taxonomy (Templeton, 1985, 1986; Shao and Sokal, 1986; Prager and Wilson, 1988) or in biogeography (Rosen, 1978; Simberloff, 1987; Page, 1988). The present paper introduces a new statistical framework designed to compare trees when fusion levels are to be taken into account. Some of the former methods required the generation of random trees; the procedure we propose proceeds instead to a special type of randomization test.

The position of the procedure proposed here is the following with respect to a nested series of possible comparisons of two datasets about the same OTUs. First, two data tables (same OTUs, different variables) can be compared by canonical correlation analysis or by redundancy analysis. The null hypothesis when testing canonical correlations for significance is that the two sets of variables are unrelated. In redundancy analysis, the null hypothesis can be stated in the regression framework: the variables in the first set do not explain a significant fraction of the variance of the variables in the second set (or conversely). Secondly, similarity or distance matrices can be computed for the two data tables, and compared using a Mantel test (1967). The null hypothesis in that case is that similarities (or distances) in the first matrix are no more (linearly) related to the similarities (or distances) in the second matrix than if the data vectors had been attributed at random to the OTUs. The third step consists of deriving dendrograms from the said similarity or distance matrices. This is the problem of consensus, discussed in the present paper. When testing for significance, the null hypothesis is that the two dendrograms do not exhibit a higher consensus than two randomly constructed dendrograms; "randomly constructed" will be discussed in some detail below. A fourth step would be to choose a partition of interest in each of the two dendrograms and to compare these partitions for information in common. This can be accomplished by contingency table analysis; the null hypothesis is then that there is no more information in common between the two partitions of interest than if the OTUs had been attributed at random to the groups. Alternatively, Nemec and Brinkhurst (1988) have proposed a test of significance comparing in turn all the cutting levels of two dendrograms, using Fowlkes and Mallows (1983) statistic computed on a matching matrix.

THE NEED FOR A SPECIFIC TEST TO COMPARE CLASSIFICATIONS

Let us come back to the formulation of the null hypothesis for comparing classifications. The objective is to test the significance of the resemblance—measured possibly by consensus indices—between two real classifications of the same OTUs. This objective may be made operational by testing whether two classifications based on real data are more closely related than two random dendrograms. The null hypothesis can then be stated as follows: the two classifications are no more similar than expected from pairs of dendrograms randomly selected from the populations to which the two real dendrograms pertain.

The approach used by Shao and Rohlf (1983) and by Shao and Sokal (1986) to test the significance of consensus indices between trees pertains to this family. They performed elaborate simulations to evaluate the sampling distribution of ten consensus tree methods and eight consensus indices. Their approach is designed, however, to compare labeled binary trees or *n*-trees. It does not take into consideration the hierarchical levels of fusion that form a prime component of dendrograms.

To overcome this shortcoming of most consensus techniques, Rohlf (1982) has proposed using matrix correlations (called "cophenetic correlations" in Sneath and Sokal, 1973) as consensus indices, since the direct comparison of ultrametric "cophenetic" matrices is equivalent to the comparison of the associated dendrograms.

Correlation statistics, either parametric or nonparametric, cannot be tested in the usual way, however, because individual values within distance matrices are not independent of one another. The Mantel (1967) randomization test represents a solution to the problem of comparing resemblance matrices. It is also called the Quadratic Assignment Procedure in psychometrics (Hubert and Schultz, 1976). Notice that the standardized form of the Mantel statistic is equivalent to a Pearson's correlation coefficient computed over the distance values in the two resemblance matrices. When this statistic is computed over cophenetic matrices, it is the same as the matrix correlation index advocated by Rohlf (1982). The difference with the classical test of the Pearson product-moment correlation coefficient lies in the fact that the Mantel test proceeds by randomization to assess the significance of the correlation statistic. Contrary to most consensus indices, both the fusion levels and the distribution of the matrix elements can be taken into account in the Mantel test. Accordingly, Hubert and Baker (1977) have proposed to extend the use of the Mantel method, originally designed to compare resemblance matrices, to measure the association between two cophenetic matrices. While this technique is directly applicable to dendrograms, it does not fully correspond to the null hypothesis that we would like to test. A permutation test of this kind simply rearranges the rows and columns of the matrices in all possible ways. This corresponds to a randomization of the OTU positions on the nested tree, while keeping its topology and fusion levels constant. Therefore, it does not test whether two trees are topologically similar, but instead it compares only the position of the OTUs on two dendrograms without any reference to the underlying topology. This does not correspond to the generation of completely random classifications.

Although the consensus tree distribution method and the cophenetic matrix technique do not solve, alone, the problem of comparing classifications, we would like to show that combining them does produce an acceptable test of the consensus between nested classifications.

DEFINITION OF A NESTED CLASSIFICATION

There are several types of classification (Sneath and Sokal, 1973). One of them can be represented by a dendrogram, which is a rooted tree of non-overlapping groups with fusion levels organized in a hierarchic way. This type of classification has three formal properties:

1. The tree topology and the position of the root must be specified. In evolutionary biology, this information conceptually represents the bifurcation sequence of the phyletic lineages leading to present-day taxa.
2. A scale must be given, to which the levels of hierarchic fusion are related; the position of those levels must be given. The scale provides information about the degree of resemblance between two sister-taxa. If we consider the scale to be time-related (constant evolutionary clock hypothesis), the fusion levels may then be thought of as representing the relative dates of divergence of the taxa.
3. The position of the OTUs on the terminal edges ("leaves") of this tree must be given; the tree must be labeled. Those labels are generally the species, or any other supraspecific taxa that we want to classify or the phylogeny of which we want to retrace.

All three aspects are essential. When one is missing, a classification can no longer be represented by a hierarchical dendrogram but degenerates into a scale-less binary or *n*-tree (no scale), an unlabeled nested tree (no specified leaves), a network (no root), or an ordination (no tree topology). Notice that a dendrogram is not necessarily a binary tree; it is called an *n*-tree (McMorris et al., 1983) when some fusion levels are equal.

Figure 1 illustrates four classifications that differ from one another in at least one aspect of the definition of a dendrogram.

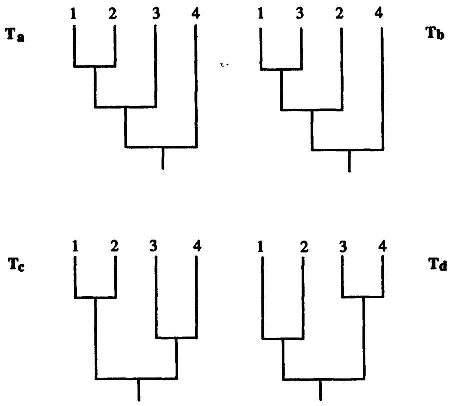

FIG. 1. Four nested classifications that differ in at least one aspect of the definition of a dendrogram. See text for details.

T_a and T_c have different topologies, T_a and T_b have different leaf positions, while T_c and T_d, though having the same topology and leaf positions, still differ because of their fusion levels. The two most dissimilar classifications are T_b and T_d because they are different in all three aspects.

Considering the three characteristics above, a randomization test comparing dendrograms requires that we generate classifications that are random with respect to all three criteria. A random classification, acceptable as a realization of the null hypothesis, will then have a random topology, random fusion levels and random leaf positions. Hubert and Baker (1977) randomized only the positions of the leaves, so that they did not meet this definition of a random dendrogram. Shao and Sokal (1986) generated random trees but did not randomize the fusion levels, which is an important aspect of the definition. We will propose now a randomization procedure for all three aspects of a nested classification, following a "double-permutation" procedure.

CONDITIONS TO GENERATE RANDOM DENDROGRAMS

When generating random trees, one has to conform to a null distribution. Various authors (Rosen, 1978; Savage, 1983; Page, 1988) proposed different distributions from which the trees can be sampled randomly. All agree with Simberloff et al. (1981) who defined three hypotheses of tree distribution: a) every topology is equiprobable, b) every tree is equally probable, and c) when growing a random tree, the location of the next branching node is equiprobably distributed among all growing tips (strictly Markovian dichotomous branching process with branching points equiprobable). The trees generated by the double permutation procedure (described beneath) are equally probable, as described in the second hypothesis of Simberloff et al. which is also called the proportional-to-distinguishable-types hypothesis. Therefore, the different topologies that result from the randomization process appear in proportion to the number of distinguishable trees with that same topology. Shao and Sokal (1986) also worked under the random tree hypothesis.

Notice that we don't want to simply generate random trees, but random dendrograms. This implies generating random fusion levels as well as random topologies. Those levels must be sampled from a distribution corresponding to the resemblance coefficients on which the dendrograms to be compared are based. Hajdu (1981) and Gower and Legendre (1986) have shown that the values produced by the various resemblance coefficients follow different distributions. Consequently, an easy way to insure that the random dendrograms will obey the correct fusion level sampling distributions is to simply randomize the actual fusion levels during the generation of random dendrograms. This constraint prevents fusion levels from taking incorrect values, thus removing that possibility of bias from the results of the test.

DOUBLE-PERMUTATION PROCEDURE FOR RANDOM DENDROGRAMS

As mentioned above, any classification can be translated into a "cophenetic" matrix; the classification and the cophenetic matrix both contain exactly the same information. Furthermore, if no reversal is present in the classification, the cophenetic matrix is also ultrametric; strictly hierarchical dendrograms do not contain reversals. Randomizing this matrix for all three aspects of our definition should generate a random classification which is acceptable as a realization of the null hypothesis. The following procedure, which allows every possible dendrogram to occur equiprobably, is carried out independently for each of the two matrices being compared.

1) Permutation of the tree structure (topology and fusion levels).—To fulfill the randomness requirement when permuting the tree structure, we will take advantage of the inequality defining the ultrametric axiom. The algorithm proceeds in three steps:

- *Read the fusion levels:* First, we construct a vector **V** containing the $(n - 1)$ similarity levels $V_{i:j}$ of hierarchic fusion between elements (leaves, OTUs, . . .) of a tree **T**. In Figure 2, for instance, the 4 fusion levels for 5 objects in tree T_a are 0.17, 0.44, 0.72 and 0.84.
- *Permute the fusion levels:* Permuting **V** at random enables us to generate any one member of the set of all possible random fusion level orders, and thus all topologies corresponding to this given set of hierarchic levels. These unlabeled trees appear in proportion to the number of different combinations of OTUs that they can distinguish. Suppose that a permutation of the vector **V** produces the values in the following order: 0.84, 0.72, 0.44 and 0.17. Let us write these values, in that order, in the off-diagonal of an empty cophenetic half-matrix. This defines the fusion levels between objects 1 and 2, 2 and 3, 3 and 4, and 4 and 5 respectively, as in Figure 2, tree T_c.
- *Fill the random matrix:* The well-known ultrametric property tells us that

$$V_{i:k} \geq \min(V_{i:j}, V_{j:k}) \text{ where } i, j \text{ and } k \in \textbf{OTU}$$

where $\textbf{OTU} = \{1, 2, \ldots, (n - 1), n\}$ is the set of n OTUs. From the similarity levels in vector **V**, the entire ultrametric matrix can be filled by using the ultrametric formula $V_{i:k} = \min(V_{i:j}, V_{j:k})$. In Figure 2, tree T_c, for instance, $V_{1:3} = \min(V_{1:2}, V_{2:3}) = \min(0.84, 0.72) = 0.72$. This first procedure fills the randomness requirements simultaneously for the topology and for the fusion levels by a simple permutation of fusion values. The leaf positions are not specified yet.

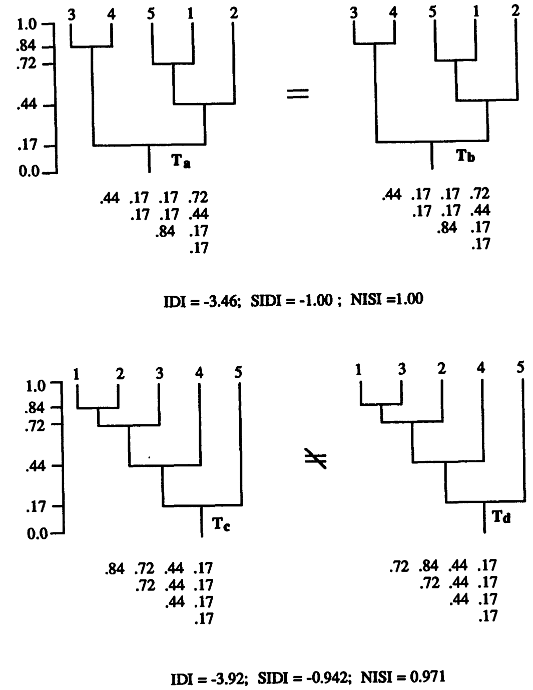

FIG. 2. The need for a standardization of the Intermediate Dissimilarity Index (*IDI*). Classifications T_a and T_b are identical but are considered less similar by *IDI* (higher value) than classifications T_c and T_d, which are different. The ultrametric *similarity* matrix is presented under each tree.

2) Permutation of the OTUs.—To obtain a random labeled dendrogram instead of a random unlabeled topology, the set of object numbers is permuted at random, and these numbers are attributed in that new order to the n terminal edges of the random dendrogram. Figure 2, T_d, shows a tree with a topology identical to T_c, but with different labels generated by the permutation of the OTUs. This random positioning of the leaves on the topology corresponds to the permutations of the rows and columns of the matrix in a Mantel test.

3) Final step.—The randomized ultrametric matrix is re-written in the original object order, so that the two matrices will remain comparable. For instance, Figure 2, could not easily be compared because the order of the objects is not the same. One of the two cophenetic matrices has to be re-written in a different order to make them comparable. Alternatively, in programming, one can use a vector of indirection indices to provide the order of the object numbers.

A statistical test for comparing classifications is obtained by adding two procedures to this algorithm:

- Choose and compute a consensus statistic between the pair of real classifications on the one hand, and between pairs of random tree ultrametric matrices on the other. The statistic described below pertains to the family of similarity functions.
- Compare the value of this consensus statistic, obtained for the pair of real classifications, to the distribution of values for some large number of pairs of random ultrametric matrices. If the actual value of the statistic is one likely to have occurred under the null hypothesis, then H_0 is accepted; if it is so large as to be considered an unlikely result under H_0, then H_0 is rejected.

This statistical scheme represents a general randomization test for the resemblance between dendrograms. Details of the double-permutation procedure are available upon request from the authors. A PASCAL program is also available to compute the double-permutation test.

As mentioned above, this procedure is not intended to produce all possible random classifications for *n* objects. It can generate only the dendrograms corresponding to the fusion levels of the real input matrices. This constraint, justified above, reduces the number of possible random trees when some fusion levels are equal. If we compared, for example, a hierarchical dendrogram to a "bush" (a tree with all fusion levels equal), we are not interested to know (H_1) whether these two dendrograms are more similar to one another than most pairs of random classifications, but instead (H_1') whether the fully hierarchical solution is among the closest approximations of a bush that one could get from a given vector of fusion levels. The null hypothesis tested with this algorithm asserts that the similarity value between the two real classifications is as small as the similarity between most pairs of random matrices containing the same levels of hierarchic fusion. The test must be one-tailed, because only the highest values of consensus, on a similarity scale, are reasons to reject the null hypothesis. The probability of the null hypothesis being true is obtained by counting the number of pairs of random ultrametric matrices with a similarity statistic as large as or larger than the similarity value for the pair of real classifications, divided by the total number of permutations; following Dwass (1957) and Hope (1968), the consensus value for the pair of real classifications is counted as one of the members of the reference distribution and is thus added to both the numerator and the denominator before computing the probability.

THE COMPARISON CRITERION, *NISI*

A choice has to be made between different measures of resemblance (consensus) for ultrametric matrices. Faith (1984) has defined three kinds of association measures for pairs of resemblance matrices, based on their pattern of sensitivity. The first type (equation 1 below) captures the similarity shared by two classifications, while the second (equation 2) measures the difference between the two matrices. These salient properties, found in all indices of type 1 or type 2, can be summarized as follows, with A_{ij} and B_{ij} representing *similarities* between OTUs *i* and *j* in the first and the second matrix respectively:

$$Co = \Sigma\Sigma \min(A_{ij}, B_{ij}) \qquad (1)$$

$$Cu = \Sigma\Sigma |A_{ij} - B_{ij}| \qquad (2)$$

where, following Day's (1983) nomenclature, *Co* is the organized complexity and *Cu* is the unorganized complexity. It is then obvious that these two types of measures evaluate different aspects of the resemblance between classifications. *Co* and *Cu* do not add up to one. Various authors have

TABLE 1. Terminologies used by different authors to refer to measures of association between classifications.

Similarity measure	Dissimilarity measure	References
Organized complexity (*Co*)	Unorganized complexity (*Cu*)	(Day, 1983)
Minimum value sensitivity	Separation sensitivity	(Faith, 1984)
Consensus similarity	Metric dissimilarity	(Faith and Belbin, 1986)

called these two types of measures by different names (Table 1).

The third pattern of association described by Faith (1984) combines the "separation sensitivity" (*Cu*) with the "minimum value sensitivity" (*Co*) properties (Table 1) into an intermediate type. These intermediate-type measures of association have the advantage of combining different association indices into simple linear algebraic expressions (Day and Faith, 1986). The Intermediate Dissimilarity Index (*IDI*) proposed by Faith and Belbin (1986), is an example of the third pattern of association:

$$IDI = \Sigma\Sigma[|A_{ij} - B_{ij}| - \min(A_{ij}, B_{ij})] \quad (3)$$

This equation could also be expressed as follows:

$$IDI^* = \Sigma\Sigma(Cu_{ij} - Co_{ij}) \quad (4)$$

Now, one could argue about which equation better reflects the association between two classifications. We have decided to use the intermediate dissimilarity index of Faith and Belbin (1986) in our test of the consensus between ultrametric matrices for dendrograms. This index (equation 3) seems to be appropriate for our study (see application, below) because, as argued by Faith and Belbin, their compromise-consensus measure is highly informative. In its actual form, the *IDI* coefficient is designed to compare similarity matrices scaled between 0 and 1. When using this index with distances matrices, one has to transform the fusion levels into similarities using a standardization of the level values between 0 and 1. The reader who prefers to do so is free to change this equation for one that he feels is more appropriate to his study, since the validity of our permutation test does not depend on the choice of a specific measure of consensus. Of course, changing the consensus measure may change the outcome of the test.

Since this index is a difference between two terms, the result may be positive or negative, depending on which is the largest term in the equation. The lowest and highest values that can be obtained from this measure are $-Co$ and *Cu*, when the other term equals zero. The more negative the *IDI* value is, the more similar are the two classifications.

We standardize the index between fixed minimum and maximum values, to insure that two different dendrograms will not appear to have a smaller dissimilarity than two identical classifications. Figure 2 shows an example where this could occur. Without standardization of the intermediate dissimilarity index, two identical trees could have been declared statistically different after the permutation test. To make sure that the lowest possible dissimilarity is that of two identical dendrograms during the generation of the reference statistical distribution of consensus index values, we divide the intermediate dissimilarity index by the largest of the two terms in equation 4, *Co* or *Cu*, to obtain a standardized intermediate dissimilarity index (*SIDI*) ranging from -1 to $+1$:

$$SIDI = \frac{\Sigma\Sigma[|A_{ij} - B_{ij}| - \min(A_{ij}, B_{ij})]}{MAX} \quad (5)$$

where $MAX =$

$$\max[\Sigma\Sigma|A_{ij} - B_{ij}|, \Sigma\Sigma \min(A_{ij}, B_{ij})] \quad (6)$$

The minimum *SIDI* value (-1) is found for two identical classifications, while the maximum value ($+1$) corresponds to the consensus between a hierarchical dendrogram and a bush whose single fusion level is zero (Fig. 3). The maximum distance

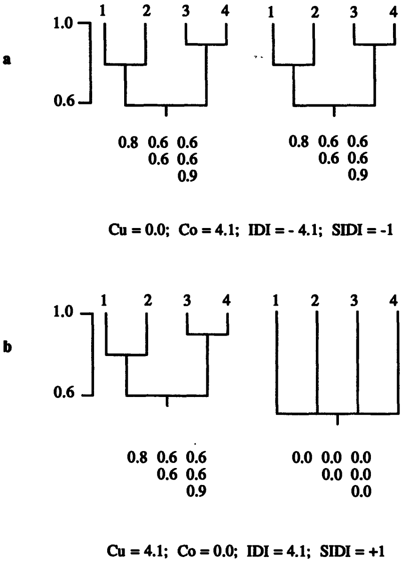

FIG. 3. Maximum and minimum values that could be taken by the Standardized Intermediate Dissimilarity Index (*SIDI*). (a) Maximum consensus. (b) Minimum consensus. The ultrametric *similarity* matrix is presented under each tree.

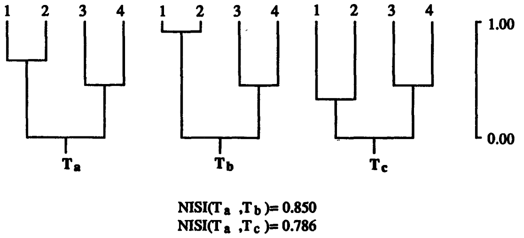

FIG. 4. Sackin's concept applied to the comparison of classifications.

(minimum consensus) will rarely be found in real taxonomic studies.

To normalize the index values between 0 and 1, the equation of the intermediate dissimilarity is modified into a Normalized Intermediate Dissimilarity Index as follows:

$$NIDI = \frac{1 + [(\Sigma\Sigma(|A_{ij} - B_{ij}| - \min(A_{ij}, B_{ij})))/MAX]}{2} \qquad (7)$$

Expressing this equation 7 in terms of similarity, we obtain:

$$NISI = 1 - NIDI \qquad (8)$$

where *NISI* is the Normalized Intermediate Similarity Index. Transformed in this way, the Faith and Belbin index loses its specificity to distinguish among identical trees those that share more "organized complexity" than others, but it retains all its properties when comparing non-identical dendrograms. Using this transformed consensus index *NISI*, two identical classifications always have a maximum similarity of 1 independent of the range of their fusion levels. The higher the consensus between two non-identical classifications, the larger the *NISI* value is. This is in agreement with the "good phenogram" concept of Sackin (1972) who stated that a classification with higher fusion levels is better than a classification with the same topology but lower hierarchic levels of fusion. The application of this assertion to the comparison of classifications means that in otherwise identical pairs of nested trees, T_a and T_b will be considered more similar than T_a and T_c if the (T_a, T_b) pair has higher organized complexity *Co* (higher value of $\Sigma\Sigma$ min[A_{ij}, B_{ij}]) (Fig. 4).

SPECIAL COMPARISONS REQUIRING LIMITED PERMUTATIONS

The general procedure requiring total permutations is not always adequate. It may prove to be irrelevant for special classification comparisons. Consider, for instance, the comparison of two dendrograms when both contain two identical, well-identified subsets of the same objects. An example is given in Figure 5. As will be shown below, a test of the consensus between these classifications, by complete randomization of both matrices, always rejects the null hypothesis of no consensus between these classifications, due to the trivial but predominant influence of the two-subset structure.

This type of situation is often found in taxonomic studies, when authors agree on the position of families but classify genera differently within these families. A test allowing genera of distinct families to be mixed by permutation will generate a vast majority of random trees less similar than

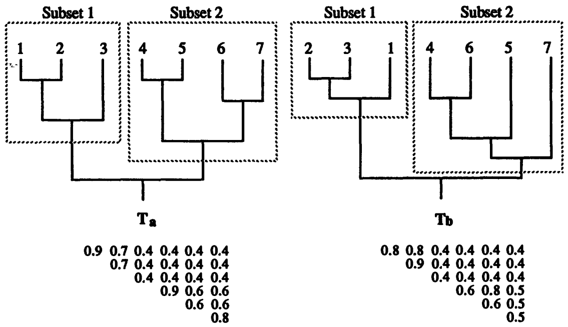

FIG. 5. A special case requiring limited permutations. The ultrametric *similarity* matrix is presented under each tree.

the two real dendrograms. Another situation would be the comparison of two area dendrograms concerning the same geographic areas, in biogeographic analysis, when a major vicariant event, such as continental drift, unquestionably creates two major groups of areas that form the main partition in each of the two dendrograms. To test correctly the consensus in these special conditions, we need either to test the two families or major areas separately, and then combine the probabilities, or more easily to limit the permutations to within the subsets only, instead of permuting all OTUs and fusion levels. These are the two ways of dealing with a conditional null hypothesis stating that within each subset the two trees do not share information. Let us illustrate this point.

Frank and Svensson (1981) have shown that for n objects, the number of possible dendrograms with distinct fusion levels (i.e., binary dendrograms) is equal to

$$n!(n-1)!/2^{n-1} \qquad (9)$$

if all the fusion levels are different. For the 7-OTU case of Figure 5, the number of random classifications generated by the permutation of a (7 × 7) matrix is 56,700. The total number of comparisons involving all permutations of both matrices is then $(56{,}700)^2$, or about 3.2×10^9. Under a conditional null hypothesis, the number of possible random trees is 54 for each classification when allowing only for limited permutations, which gives a total number of possible permutations of $(54)^2 = 2{,}916$. This means that virtually all the permutations produced by complete randomization are not relevant and should not be used to test the conditional null hypothesis of interest. Since most of these permutations would produce *NISI* values that are smaller than the index value for the real dendrograms, the test would always turn out to be significant. Empirical studies show that the consensus between pairs of classifications involving major common subsets of OTUs is always significant when the test is done by complete permutations.

Indeed, the classifications of Figure 5 prove to be statistically similar when compared by total permutations (*NISI:* 0.93204, $p(H_0) = 0.00599$), while limited permutations provide opposite results (*NISI:* 0.93204, $p(H_0) = 0.61067$). We could have obtained the same probability of the null hypothesis by comparing each subset independently before combining the probabilities: the product of the probabilities of the consensus associated with each subset ($p(H_0)_1 = 1.00$, $p(H_0)_2 = 0.60159$), obtained by sampling 10,000 times the set of the possible permutations, is equal to the probability of the limited permutation test ($p(H_0)_1 \times p(H_0)_2 = 0.60159 \approx 0.61067 = p(H_0)_{\text{limited perm.}}$). The slight discrepancy between the probability values is due to sampling error.

APPLICATION TO KANGAROO CLASSIFICATION

We offer the following example of an application of the test of consensus be-

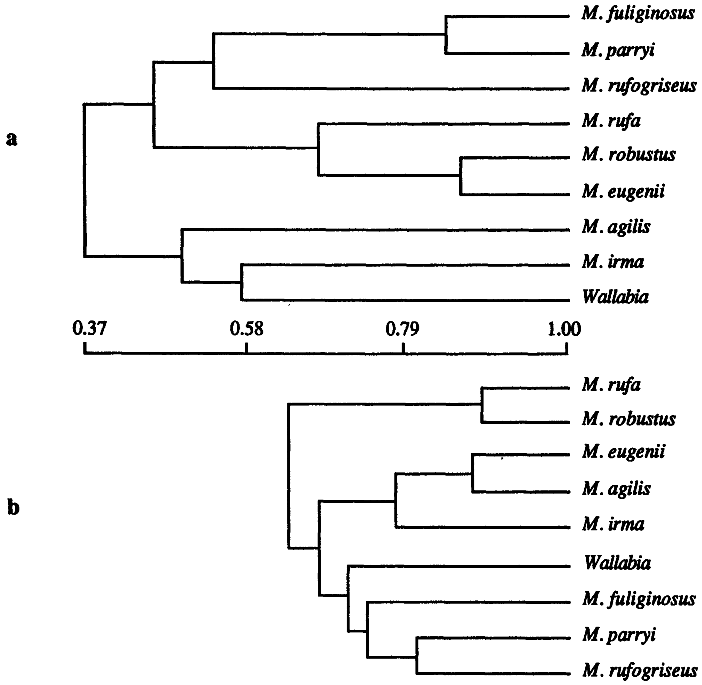

FIG. 6. Classifications of the Macropodidae derived from different datasets, nine species. (a) Electrophoretic data from Richardson et al. (1973). (b) Serological data from Kirsch (1977).

tween dendrograms, involving an unconditional null hypothesis. The taxonomic problem under study, that triggered our interest in classification comparison methods, is whether different published kangaroo datasets are congruent in the classifications they lead to.

Kirsch (1977) has studied serological relationships within several Marsupial families. His work is considered to be providing a sound model for Macropodid (kangaroo) taxonomy. Will different character sets also lead to Kirsch's classification? In particular, Avise (1974) and Baverstock et al. (1979) argued that electrophoresis does produce good taxonomic characters. Richardson et al. (1973) attempted to substantiate this assertion for the kangaroos, but their work was not statistically conclusive. We want to test here the statement that electrophoretic data and serological characters yield congruent dendrograms.

In a first-test, we chose nine species for comparison, all from the single sub-family *Macropodinae*; eight of them are from genus *Macropus* and the other one is a *Wallabia*; only nine species are used in this first example in order to avoid using limited permutations. We computed the electrophoresis similarity matrix from Richardson et al.'s published dataset, using Jaccard's similarity coefficient. Dendrograms were constructed using a variety of clustering methods; the one with the highest cophenetic correlation was obtained from the UPGMA method (Fig. 6a). The serological dendrogram (Fig. 6b), on the other hand, is given by Kirsch (1977). We compared these two classifications to measure the consensus between the two taxonomic solutions. In order to evaluate the distribution of the consensus statistic *NISI*, we computed 5,000 permutations out of the $(57{,}153{,}600)^2$, or about 3.27×10^{15}, that are possible when comparing two classifications of nine objects. The normalised intermediate similarity index for this pair of matrices proved not to be significant at level $\alpha = 0.05$ ($NISI = 0.82927$, $p(H_0) = 0.53589$: one-tailed test).

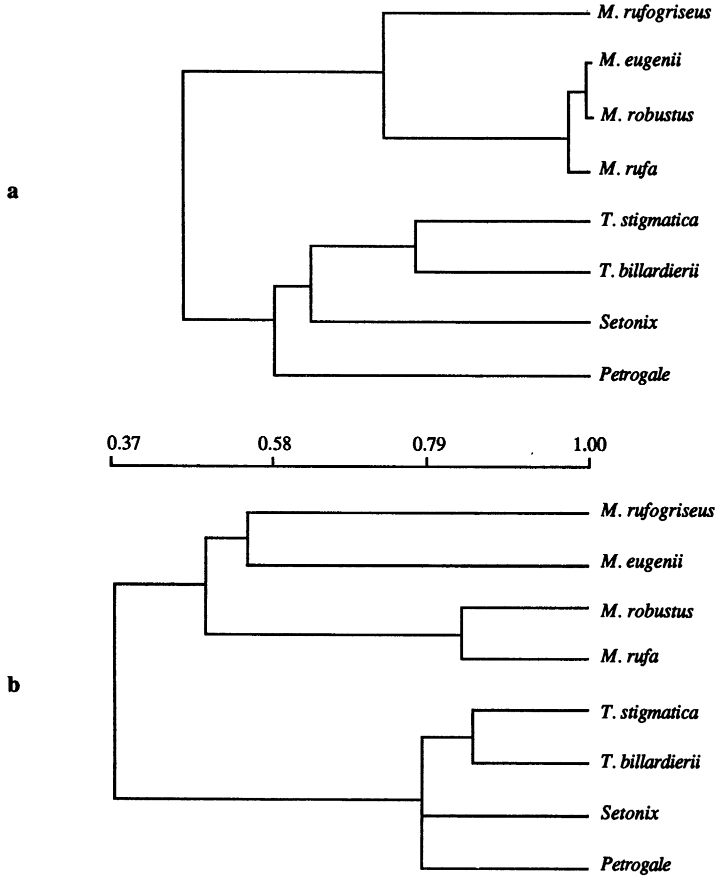

FIG. 7. Classifications of the Macropodidae derived from different datasets, eight species pertaining to two distinct groups. (a) Electrophoretic data from Richardson et al. (1973). (b) Serological data from Kirsch (1977).

We conclude from this statistical result that the electrophoretic and serological data did not produce statistically similar dendrograms. Those who believe that Kirsch's classification is correct are led to the conclusion that, contrary to the statement of Richardson et al. (1973), electrophoresis did not produce "good taxonomic characters" in this case.

To illustrate the use of limited permutations, the same datasets were compared again, for species pertaining to two distinct subsets (Fig. 7). For the purpose of the demonstration, we chose eight species divided equally in two subgroups. We would like to point out that we could have used more than two groups, and that the groups need not have the same cardinality. The first group contains four *Macropus* species, while the second one contains representatives of genera *Thylogale* (two species), *Setonix* and *Petrogale* (one species each). The classifications based on electrophoretic (Fig. 7a) and on serological (Fig. 7b) data both trace the dichotomy between the two subsets. We computed 5,000 permutations out of the $(396{,}900)^2$ possibilities for eight objects to evaluate the distribution of the statistic *NISI*. The general randomization test allowing every permutation to occur rejects the null hypothesis (stating that there is no consensus between these dendrograms) at level $\alpha = 0.01$ (*NISI* $= 0.83282$, $p(H_0) = 0.00979$). Limited permutations lead to the opposite conclusion (*NISI* $= 0.83282$, $p(H_0) = 0.60878$). Although both classifications find the same major partition, the position of the species in the two higher subsets is not statistically similar.

These applications illustrate the fact that total permutations do test mainly for the presence of similar partitions at higher levels of the hierarchy. The fine composition of these higher subsets may be analysed through the use of a conditional null hypothesis involving limited permutations.

CONCLUSION

A recent trend in evolutionary studies is to try to test evolutionary hypotheses, trading simple descriptions of taxonomic structures for predictions. Recent developments in consensus theory have led to some methods for the statistical comparison of evolutionary trees. These tests are designed to evaluate clearly specified hypotheses:

1. The comparison of binary trees on the sole basis of their bifurcation sequence, ignoring the fusion levels of the hierarchy (Shao and Sokal, 1986).
2. The comparison of partitions involving terminal edges of a tree, without reference to the underlying topology (Hubert and Baker, 1977).

In this paper, we proposed a test which is a combination of these two comparison schemes. It should be applied to the comparison of classifications when all three aspects of the two rooted trees are seen as essential components of the description of these classifications: topology, leaf positions, fusion levels. It is not meant to replace other methods. Instead, it is proposed as a refinement of earlier tests. A further point to remember is that it must be used only to compare classifications that are based on different datasets, and not to compare different methods of classification, because from the same data, different methods will produce solutions that are not independent, which makes the null hypothesis very unlikely to be supported indeed.

Further research is needed in this new area of the statistical comparison of trees. The next problem to be solved is the comparison of phylogenetic trees, that is, binary trees with patristic distances. To do this, one must first be able to produce random patristic matrices, in order to evaluate the distribution of the consensus statistic. Since no such test is available at the moment, one could use in the meantime our double permutation procedure and compare dendrograms derived from the patristic distance matrices.

ACKNOWLEDGMENTS

This work has been supported by NSERC grant No. A7738 to P. Legendre and by a NSERC scholarship to F.-J.L. Our thanks are due to Alain Vaudor, computer analyst at Université de Montréal, for assistance during the programming work, and to Dr. D. P. Faith for his helpful review of the manuscript.

REFERENCES

ADAMS, E. N., III. 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool., 21: 390–397.

AVISE, J. C. 1974. Systematic value of electrophoretic data. Syst. Zool., 23:465–481.

BAVERSTOCK, P. R., S. R. COLE, B. J. RICHARDSON, AND C. H. S. WATTS. 1979. Electrophoresis and cladistics. Syst. Zool., 28:214–219.

COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. Syst. Zool., 29:288–299.

DAY, W. H. E. 1983. The role of complexity in comparing classifications. Math. Biosci., 66:97–114.

DAY, W. H. E. 1985. Optimal algorithms for comparing trees with labeled leaves. J. Clasif., 2:7–28.

DAY, W. H. E., AND D. P. FAITH. 1986. A model in partial orders for comparing objects by dualistic measures. Math. Biosci., 78:179–192.

DWASS, M. 1957. Modified randomization tests for nonparametric hypotheses. Ann. Math. Stat., 28: 181–187.

ESTABROOK, G. F., F. R. MCMORRIS, AND C. MEACHAM. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool., 34:193–200.

FAITH, D. P. 1984. Patterns of sensitivity of association measures in numerical taxonomy. Math. Biosci., 69:199–207.

FAITH, D. P., AND L. BELBIN. 1986. Comparison of classifications using measures intermediate between metric dissimilarity and consensus similarity. J. Classif., 3:257–280.

FOWLKES, E. B., AND C. L. MALLOWS. 1983. A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc., 78:553–569.

FRANK, O., AND K. SVENSSON. 1981. On probability distributions of single-linkage dendrograms. J. Statis. Comput. Simul., 12:121–131.

GOWER, J. C., AND P. LEGENDRE. 1986. Metric and

Euclidean properties of dissimilarity coefficients. J. Classif., 3:5-48.

HAJDU, L. J. 1981. Graphical comparison of resemblance measures in phytosociology. Vegetatio, 48:47-59.

HOPE, A. C. A. 1968. A simplified Monte Carlo significance test procedure. J. Roy. Stat. Soc. Ser. B, 30: 582-598.

HUBERT, L. J., AND F. B. BAKER. 1977. The comparison and fitting of given classification schemes. J. Math. Psychol., 16:233-253.3.

HUBERT, L. J., AND J. SCHULTZ. 1976. Quadratic assignment as a general data analysis strategy. Br. J. Math. Statis. Psychol., 29:190-241.

KIRSCH, J. A. W. 1977. The comparative serology of Marsupialia, and a classification of Marsupials. Aust. J. Zool. Suppl. Ser., 52:1-152.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research, 27:209-220.

MARGUSH, T., AND F. R. MCMORRIS. 1981. Consensus *n*-trees. Bull. Math. Biol., 43:239-244.

MCMORRIS, F. R., D. B. MERONK, AND D. A. NEUMANN. 1983. A view of some consensus methods for trees. Pages 122-126 *in* Numerical taxonomy (J. Felsenstein, ed.). NATO Advanced Studies Institute, Ser. G (Ecological Sciences), No. 1. Springer-Verlag, Berlin.

NEMEC, A. F. L., AND R. O. BRINKHURST. 1988. The Fowlkes-Mallows statistic and the comparison of two independently determined dendrograms. Can. J. Fish. Aquat. Sci., 45:971-975

NEUMANN, D. A. 1983. Faithful consensus method for *n*-trees. Math. Biosci., 63:271-287.

PAGE, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool., 37:254-270.

PENNY, D., L. R. FOULDS, AND M. D. HENDY. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature, 297:197-200.

PENNY, D., AND M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool., 34:75-82.

PRAGER, E. M., AND A. C. WILSON. 1988. Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. J. Mol. Evol., 27: 326-335.

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. Ecol., 1:203-209.

ROBINSON, D. F., AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci., 53:131-147.

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

ROSEN, D. E. 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool., 27:159-188.

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. T., AND J. T. POLHEMUS. 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). Syst. Zool., 29:1-26.

SHAO, K. AND F. J. ROHLF. 1983. Sampling distributions of consensus indices when all bifurcating trees are equally likely. Pages 132-137 *in* Numerical taxonomy (J. Felsenstein, ed.). NATO Advanced Studies Institute, Ser. G (Ecological Sciences), No. 1. Springer-Verlag, Berlin.

SHAO, K., AND R. R. SOKAL. 1986. Significance tests of consensus indices. Syst. Zool., 35:582-590.

SIMBERLOFF, D. 1987. Calculating probabilities that cladograms match: A method of biogeographical inference. Syst. Zool., 36:175-195.

SIMBERLOFF, D., K. L. HECK, E. D. MCCOY, AND E. F. CONNOR. 1981. There have been no statistical tests of cladistics biogeographical hypotheses. Pages 40-63 *in* Vicariance biogeography: A critique (G. Nelson and D. E. Rosen, eds.). Columbia University Press, New York.

SNEATH, P. H., AND R. R. SOKAL. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

SOKAL, R. R. 1983. Taxonomic congruence in the Caminalcules. Pages 76-81 *in* Numerical taxonomy (J. Felsenstein, ed.). NATO Advanced Studies Institute, Ser. G (Ecological Sciences), No. 1. Springer-Verlag, Berlin.

STINEBRICKNER, R. 1984. S-consensus trees and indices. Bull. Math. Biol., 46:923-935.

TEMPLETON, A. R. 1985. The phylogeny of the hominoid primates: A statistical analysis of the DNA-DNA hybridization data. Mol. Biol. Evol., 2:420-433.

TEMPLETON, A. R. 1986. Further comments on the statistical analysis of DNA-DNA hybridization data. Mol. Biol. Evol., 3:290-295.

THORPE, P. A., AND W. J. DICKINSON. 1988. The use of regulatory patterns in constructing phylogenies. Syst. Zool., 37:97-105.

WATERMAN, M. S., AND T. F. SMITH. 1978. On the similarity of dendrograms. J. Theor. Biol., 73:789-800.

WILLIAMS, W. T., AND H. T. CLIFFORD. 1971. On the comparison of two classifications on the same set of elements. Taxon, 20:519-522.

Received 14 March 1989; accepted 25 September 1989