THE AVERAGE CONSENSUS PROCEDURE: COMBINATION OF WEIGHTED TREES CONTAINING IDENTICAL OR OVERLAPPING SETS OF TAXA

FRANÇOIS-JOSEPH LAPOINTE1 AND GUY CUCUMEL2

1Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale centre-ville, Montréal, Québec H3C 3J7, Canada; E-mail: lapoinj@ere.umontreal.ca
2École des sciences de la gestion, Université du Québec à Montréal, C.P. 8888, Succursale centre-ville, Montréal, Québec H3C 3P8, Canada; E-mail: cucumel.guy@ugam.ca

Abstract.—The average consensus procedure, originally proposed to combine dendrograms (i.e., ultrametric trees), is extended here to apply to any type of tree with branch lengths (ultrametric or not) containing identical, inclusive, or overlapping sets of taxa. The method proceeds in two steps. First, the average of the path-length matrices corresponding to the trees to be combined is computed. Then, a least-squares algorithm is applied to this average matrix to obtain a consensus solution. The average consensus tree is a solution that minimizes the sum of squared distances between the consensus and the trees in the input profile. An application of the method to combine phylogenetic hypotheses for kangaroos is presented. [Consensus method; kangaroo; Macropodidae; phylogeny; supertree; weighted trees.]

Given a set of phylogenies, how could we combine them to obtain a consensus solution that is representative of the entire set? Different methods could lead to various consensus solutions that may be considered representative (for a discussion, see Barthélémy et al., 1986). Whatever the approach selected, it seems illusory to believe that there exists a single consensus method adapted to every problem. For that matter, it is important to know exactly what we want in a consensus method. For example, the median procedure (Barthélemy and McMorris, 1986) is an optimization-based technique that returns a consensus tree that is in some sense “closest” to the set of input trees. In our case, we want an optimization-based method that returns a consensus weighted tree (i.e., a tree with branch lengths) that is “closest” to a set of input weighted trees. We introduce here the average consensus procedure as a method to accomplish that goal. Considering branch lengths can sometimes lead to consensus solutions differing from those produced by methods that ignore them. Also, the average consensus can be applied to any type of weighted trees, whether ultrametric or not (i.e., with or without equal rates of evolution in the different lineages). We extend the method to special cases involving trees bearing nonidentical sets of leaves (see also Gordon, 1986; Brossier, 1990; Lapointe and Cucumel, 1991; Baum, 1992; Ragan, 1992; Steel, 1992; Purvis, 1995).

THE PROBLEM

Any phylogeny can be decomposed into three distinct parts: topology, label positions, and branch lengths (see Lapointe and Legendre, 1990). The topology is the shape of the tree (e.g., symmetrical or asymmetrical, comblike, starlike) depicting the branching pattern of the different lineages. The labels attached to the leaves (or tips) of the branches refer to taxon names, and their relative positions provide information about their phylogenetic relationships. Branch lengths represent the amount of evolutionary change (relative or absolute) of the various clades and can sometimes refer to dates of separation among lineages. Depending on the comparison criterion selected, trees can differ in terms of topological relationships, label positions, or branch lengths. Likewise, depending on the consensus procedure employed, some aspects of a tree are given more or less weight when building the consensus tree. Accordingly, the results will differ among methods that account for
different aspects of the trees combined. Two families of consensus methods for trees are thus distinguishable: those that do not consider branch lengths versus those that do. The strict (Sokal and Rohlf, 1981), semistrict (Bremer, 1990), majority rule (Margush and McMorris, 1981), and Adams (1972) consensus methods are among those that ignore branch lengths. Stinebrickner (1984), Lefkovitch (1985), and Brossier (1990) have proposed alternative methods that account for branch lengths; these methods are restricted to dendrograms (i.e., ultrametric trees), however. We were looking for a consensus technique capable of taking into account the branch lengths while combining any type of weighted trees, including ultrametric and more general nonultrametric additive trees.

**THE SOLUTION**

The average consensus procedure, originally proposed by Cucumel (1990) to combine dendrograms, is a method specifically designed to compute the consensus of trees with branch lengths; the solution is returned as a weighted tree of the same kind as the trees that were combined. The average procedure takes as input a set of several weighted trees and returns a weighted tree that is in some sense "closest" to the entire input set. Given that there exists a one-to-one correspondence between any weighted tree $T$ and its associated path-length matrix $M$ (Hartigan, 1967; Buneman, 1971), it is equivalent to deal with trees or with their matrix representations. Accordingly, the average procedure returns a path-length matrix, associated with a consensus weighted tree that is closest to the entire input set.

Let $S = \{1, \ldots, i, \ldots, j, \ldots, n\}$ be a set of taxa. Let $T_1$ and $T_2$ be two weighted trees defined on $S$; $M_1$ and $M_2$ are their associated path-length matrices. The distance $\Delta$ between the trees $T_1$ and $T_2$ is defined as follows (Hartigan, 1967):

$$\Delta(T_1, T_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} [d_1(i, j) - d_2(i, j)]^2$$  \hspace{1cm} (1)

for every $i$ and $j$, where $d_1(i, j)$ and $d_2(i, j)$ are the path-length distances corresponding to $M_1$ and $M_2$, respectively.

Let us now consider a profile $P = \{T_1, \ldots, T_n, \ldots, T_r\}$ of $t$ weighted trees defined on a common set of taxa $S$. The average consensus tree $T_c$ is defined as the solution that minimizes the following loss function:

$$\sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} [d_1(i, j) - d_k(i, j)]^2$$  \hspace{1cm} (2)

for every $i$, $j$, and $k$, where the $d_1(i, j)$ and $d_k(i, j)$ are the path-length distances in the corresponding matrices $M_1$ and $M_k$. With no loss of generality, one could consider this consensus procedure as one involving a profile of scaled trees $T_k$ represented by their matrices $M_k$ taking their values in the interval [0, 1].

Computation of an average consensus involves two simple steps. First, the average of the input path-length matrices is computed. By definition, this average matrix represents the centroid of the set of input path-length matrices. The second step consists of finding the closest path-length matrix from this average matrix using a least-squares algorithm (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967; De Soete, 1983). The weighted tree associated with this path-length matrix is the average consensus tree.

Let $\bar{M}$ represent the average distance matrix containing distances $\bar{d}(i, j)$, such that

$$\bar{d}(i, j) = \frac{1}{t} \sum_{k=1}^{t} d_k(i, j)$$  \hspace{1cm} (3)

for every $i$, $j$, and $k$, where the $d_k(i, j)$ are the path-length distances corresponding to the different $M_k$. Because the average matrix $\bar{M}$ is likely to be nonultrametric or additive, the average consensus $T_c$ (see Eq. 2) is then obtained as the solution that minimizes the following function:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} [d(i, j) - \bar{d}(i, j)]^2$$  \hspace{1cm} (4)

for every $i$ and $j$, where the $d(i, j)$ are the
path-length distances associated with matrix $M_c$. Thus, we have shown (see proof in the Appendix) that using any least-squares algorithm on the average matrix, $\bar{M}$, the returned weighted tree $\bar{T_c}$ is an average consensus.

### The Average Consensus Supertree

Given the general form of the average consensus procedure (Eq. 4), the same method can be extended to cases involving weighted trees bearing partly overlapping or inclusive sets of taxa. Let us consider a profile $P = \{T_1, \ldots, T_k, \ldots, T_t\}$ of $t$ weighted trees defined on $t$ sets of taxa $S_i$, such that $S = S_1 \cup \ldots \cup S_k \ldots \cup \ldots \cup S_t$. The average procedure takes as input the different weighted trees and returns a weighted supertree on $S$.

Contrary to the simpler case of weighted trees defined on the same set of taxa, the average consensus supertree is usually obtained in three steps. First, average path-length distances defined on $S$ need to be computed. To account for the fact that some taxa may not be represented in all trees, a weighted average matrix $M^*$ is required here:

$$
\bar{d}^*(i, j) = \frac{1}{W(i, j)} \sum_{k=1}^{t} d_k(i, j)
$$

for every $i, j$, and $k$, where $d_k(i, j)$ are the path-length distances in $M_k$, and $W(i, j)$ represent a set of weights counting the number of trees that include both $i$ and $j$.

For trees defined on inclusive sets of taxa, the average consensus supertree $T^*_c$ is directly derived from $M^*$ (see Eq. 4). However, with trees defined on overlapping sets of taxa, the resulting matrix $M^*$ will bear missing cells for pairs of taxa that never occurred together in any tree $T_k$ of $P$. In such cases, an additional step is required to estimate these values prior to phylogenetic reconstruction. De Soete (1984), Lapointe and Kirsch (1995), and Landry et al. (1996) have proposed different methods for estimating missing cells using the ultrametric or additive property of path-length distances. When the average matrix $\bar{M}^*$ is filled, the average consensus supertree $T^*_c$ can finally be derived using any least-squares algorithm.

### Application

To illustrate the computation of the average consensus, we used the method to combine a pair of weighted trees depicting the phylogenetic relationships among different genera of kangaroos. The first tree (Fig. 1a), reproduced from Kirsch et al. (1995), is based on DNA-hybridization data. The second tree (Fig. 1b) is the result of the immunological work of Bav-erstock et al. (1989). In both cases, only one exemplar of each genus was selected for comparisons, leaving 8 and 12 taxa in the corresponding trees. Four genera included in Baverstock et al.'s tree were not represented in Kirsch et al.'s study, and one genus in Kirsch et al.'s tree (Thylol- gale) was used only as an antigen by Bav-erstock et al. Seven taxa were included in both trees. We first computed the average consensus tree for this common set of taxa, comparing the pruned trees. Then, a consensus supertree defined on the entire set of 13 genera was obtained by combining the two original trees. The consensus trees were obtained in several steps.

The first requirement before computing an average consensus is to make sure that the trees are comparable, i.e., they need to be standardized. Because relative distances are more important than their absolute values, matrices must always be scaled before computation of the average matrix. One way of doing this is to transform the original path-length matrices so that distances become bounded between 0 and 1. This was done for the path-length matrices representing the complete and pruned trees (Fig. 1).

Then, average path-length distances can be computed for pairs of taxa occurring in all the trees. In the present case, seven genera were involved in the averaging operation computed over the reduced matrices. For the consensus supertree, the additional distances in the average matrix are the same as those in the original path-length matrices corresponding to the trees in
FIGURE 1. Proposed phylogenies of kangaroos, from Kirsch et al. (1995) (a) and Baverstock et al. (1989) (b). The trees were arbitrarily rooted between the potoroines (Aepyprymnus, Potorous, Bettongia) and macropodines.

which the relevant taxa were present. However, because one taxon (Thylogale) was not included in the larger phylogeny (Fig. 1b), distances involving that genus and the five others not included in the smaller tree (Fig. 1a) had to be estimated; we used the algorithm proposed by Landry et al. (1996). This method uses the four-point condition of path-length distances to fill missing cells in the average matrix. It was possible to estimate every missing distance by recursively applying the procedure to every quartet of taxa in the matrix. Once filled, the average matrix could be used in a phylogeny reconstruction program.

The final step consists of finding the weighted tree closest to the profile of trees. This step is necessary because the average of path-length distances computed from the original matrices are not necessarily path-length distances themselves, i.e., the average matrix is not always ultrametric or additive. Thus, the consensus solution is obtained by applying a least-squares algorithm to the average matrix. We used the Fitch and Margoliash (1967) method (FITCH program in PHYLIP; Felsenstein, 1993) to compute the consensus tree and the supertree of the kangaroo trees.

In the present application, the average consensus (Fig. 2a) obtained is very similar to the input trees (see Fig. 1). In fact, the consensus represents a mix of the clades represented in the reduced trees. Macropus and Wallabia form a clade as in the initial trees. However, Dendrolagus and Petrogale, as well as Setonix and Dorcopsulus, are not
sister taxa in the average consensus tree. The branch lengths of the consensus also differ from those observed in any of the initial trees.

The consensus supertree (Fig. 2b) is even more telling about the relationships among the 13 genera of kangaroos. As suggested by Baverstock et al. (1989), Dendrolagus, Petrogale, and Thylogale should form a clade, which is what happened when we added Thylogale to their tree. The positions of the other taxa in the consensus were not affected by the averaging operation (i.e., the tree is identical to Baverstock et al.’s tree; Fig. 1b); however, the relative branch lengths are different.

If path lengths had been ignored when constructing the consensus, the results might have been less resolved. For instance, relationships among Dorcopsis, Setonix, and the Wallabia-Macropus clade in both consensus trees would be unresolved on the basis of topological information alone. Accounting for path-length distances does make a difference. If one desires to capture every aspect of the original trees in a consensus representation, methods that deal with weighted trees should be seriously considered. The average procedure described here is such a method.

ACKNOWLEDGMENTS

We thank Marc Bourdeau, Gildas Brossier, Bill Day, Geert De Soete, Allan Gordon, John Kirsch, Bruno Leclerc, Pierre Legendre, Buck McMorris, Mike Steel, Ralph Stinebrickner, and two anonymous reviewers for commenting on this or earlier versions of the manuscript.
REFERENCES


Received 16 February 1996; accepted 5 December 1996
Associate Editor: David Baum

Appendix

Here is the detailed development of the function to minimize Equation 2, which shows that the average consensus T, is the least-squares tree derived from the average path-length matrix M for every i, j, and k, given

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (d_{(i,j)} - d_{(i,j)})^2 \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} [d_{(i,j)} - 2d_{(i,j)}d_{(i,j)} + d_{(i,j)}^2] \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ k(d_{(i,j)} - 2d_{(i,j)}) \sum_{i=1}^{n} d_{(i,j)} + \sum_{i=1}^{n} d_{(i,j)}^2 \right\} \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \left( d_{(i,j)} - \frac{1}{k} \sum_{i=1}^{n} d_{(i,j)} \right)^2 \right] \]

\[ - \frac{1}{k} \sum_{i=1}^{n} d_{(i,j)}^2 + \frac{1}{k} \sum_{i=1}^{n} d_{(i,j)}^2 \].

Given that
Now, setting (see Eq. 3)
\[ \bar{d}(i, j) = \frac{1}{t} \sum_{i=1}^{n} d_s(i, j), \]
then \( d_s \) is the solution that minimizes (see Eq. 4)
\[ \sum_{i=1}^{n} \sum_{j=1}^{n} [d(i, j) - \bar{d}(i, j)]^2. \]