DISTANCE METHODS AND THE APPROXIMATION OF MOST-PARSIMONIOUS TREES

DANIEL P. FAITH

CSIRO Division of Water and Land Resources, G.P.O. Box 1666, Canberra, A.C.T., 2601, Australia

Abstract.—Distance data have posed a number of problems for phylogenetic analysis. Among these are the loss of information about individual character states, and the frequent departures from metric properties of distance matrices derived by molecular techniques. The common degree-of-fit methods for the analysis of such data imply possibly unrealistic assumptions about these distances. As an alternative, a minimum-length criterion is considered. This has the appeal of requiring more conservative assumptions about distance data and represents the equivalent criterion to that for the analysis of character data by numerical cladistic techniques. Based upon its similarity to the character-Wagner algorithm, the distance-Wagner algorithm is suggested as a possible heuristic method for the approximation of most-parsimonious trees from distance data. Both the distance-Wagner algorithm and a recent modification have weaknesses in this role. Computer simulations demonstrate that the new algorithm developed in this study compares favorably with not only the distance-Wagner algorithm but also the character-Wagner algorithm in the approximation of most-parsimonious trees. [Metric; molecular distances; Wagner algorithm; parsimony; cladistics; immunological distance.]

Among current methods for the inference of phylogenetic trees, those relying on the use of character data seem to contrast increasingly with those methods being developed that depend only upon the analysis of a matrix of pairwise distances among taxa. The contrast is not only in terms of the mathematics of the respective algorithms, but also apparently in terms of the goal of the analysis; what is being optimized may be quite different in the two cases. In the first case, for character data, an approximation to a phylogenetic tree is usually derived by the application of some form of a parsimony criterion (for a recent review of parsimony methods see Felsenstein, 1982). For example, the character-Wagner algorithm of Farris (1970) is used to approximate an overall most-parsimonious tree (Wagner tree) by a sequence of individually most-parsimonious additions of taxa to the tree or network. The best approximation is equivalent to the tree that has minimum length. This length is computed using a Manhattan-metric measure based upon the character data.

The second group of methods, using only distances among taxa as its input data, arises especially out of the need to analyze various types of molecular data, such as those resulting from immunological work. This need has led to a profusion of new methods for the inference of phylogenetic trees (see Felsenstein, 1982). One of the most popular, the distance-Wagner algorithm, was developed by Farris (1972) as an extension of the character-Wagner algorithm to infer a phylogenetic tree using only distance data and without the assumption of equal rates of evolution in different phyletic lines.

The most common criterion to be optimized in these distance methods is some measure of the degree of fit between the initial distances among the taxa and the path-length distances derived from the resulting tree (Swofford, 1981; Felsenstein, 1982). This criterion is different from the criterion of parsimony employed in the analysis of character data. Phylogenetic inference based upon a fit criterion will not necessarily yield the same result as inference by a minimum-length criterion (Farris, 1983). While the parsimony criterion corresponds to the minimization of the number of "extra steps," or the total homoplasy, degree-of-fit may only imply the minimization of the sum of all pairwise homoplasy values (sensu Farris, 1967).
Perhaps as a result of the occasional use of the term “homoplasy” in both contexts, the criteria of distance-fitting and parsimony are sometimes confused (e.g., Berlocher, 1984).

The notion of degree-of-fit as a criterion to be optimized in phylogenetic inference has raised a number of issues. An obvious one is the proper choice of a particular measure of fit; the choice varies among current distance methods. For example, when the method of Fitch and Margoliash (1967) is used, one particular measure of fit between input and derived distances is implied. Farris (1981) and Felsenstein (1984) discussed the problems of choosing a fit measure in the context of expected patterns of error in the input distances.

Another issue that naturally follows on the first is the determination of the best algorithm for optimization based upon a particular measure. Recent work (Farris, 1981; Swofford, 1981) has shown that the distance-Wagner algorithm can be used to initially derive a tree topology, and branch-length optimization can be performed afterwards to produce better fit than with alternative methods such as that of Prager and Wilson (1978).

In parallel with the development of better methods of optimization, there has been increasing criticism of the degree-of-fit approach as applied to molecular-distance data (Farris, 1981, 1983, pers. comm.). According to this view, the critical defect of such molecular-distance analyses is that the input distances are not Manhattan-metric distances and, therefore, cannot be interpreted as amounts of evolutionary change for any possible underlying character data. This implies that the interpretation of the branch lengths obtained by a fit of derived (path-length) distances to these input distances is also meaningless. Beyer et al. (1974) presented an example of this problem.

In a defense of such distance analyses, Felsenstein (1984) has argued that, if we can regard the input distances as having additivity (and therefore metric conditions) in expectation, then in practice the input distances need not appear to be perfectly metric. In effect, only their “true” values are metric. Felsenstein’s arguments are, to a degree, persuasive. Clearly, we could imagine taking perfectly metric, additive distances derived from a phylogenetic tree and, with some perturbation function, distorting these distance values so that departures from metric conditions appeared. The initial tree, however, might still be recoverable by an algorithm based upon some understanding of the nature of the perturbation function used to distort the distances.

The assumption that the input distances approximate the true patristic distances is critical to a justification of degree-of-fit methods. In Felsenstein’s (1984:20) words, “the model assumes that distances are expected to be additive along the true tree.” Let the input distances and derived patristic (path-length) distances between taxa \(i\) and \(j\) be represented by \(D(i,j)\) and \(P(i,j)\), respectively. Each of these can be regarded as an estimate of a true value, call these \(\Delta(i,j)\) and \(\Pi(i,j)\). The true phylogenetic distance between \(i\) and \(j\) is \(\Delta(i,j)\), based upon some universe of characters. It represents the expectation over a large number of characters, while \(D(i,j)\) is the actual phenetic difference for the observed set of characters. The true patristic distance \(\Pi(i,j)\)—the patristic distance defined over the same universe of characters—must be greater than or equal to the true phylogenetic distance because of convergences and reversals of characters (homoplasy). These character-state changes must contribute to patristic distance and will in effect reduce the phenetic distance relative to patristic distance. This observation is consistent with Farris’ (1967) definition of the pairwise homoplasy for two taxa as the difference between their phenetic and patristic distances. If the true value of pairwise homoplasy is \(\Theta(i,j)\), then

\[
\Delta(i,j) = \Pi(i,j) - \Theta(i,j). \tag{1}
\]

This equation is revealing in terms of the assumptions of fit methods. Felsenstein’s general argument is that degree-of-fit measures may be employed under a variety of particular assumptions about the
relationship between input distances, \( D(i,j) \), and the true patristic distances, \( \Pi(i,j) \). As an example, one may have

\[
D(i,j) \sim N[\Pi(i,j), V(i,j)],
\]

so that the observed phenetic distances are normally distributed with variance \( V(i,j) \) and mean \( \Pi(i,j) \).

Certain types of molecular distance data may well match this type of model. The distances produced by DNA-hybridization studies, for example, produce matrices that appear to be very close to ultrametric (e.g., Sibley and Ahlquist, 1984). To the extent that these matrices are ultrametric, the patristic distances can be modelled as proportional to the input distances and a degree-of-fit approach is justified. For other types of data, it may be possible to devise a correction factor for the amount of homoplasy or "multiple hits" (e.g., Holmquist, 1972), so that at least the transformed input distances can be considered to be estimates of \( \Pi(i,j) \) values. Again a fit approach is well justified.

More generally, the use of a model like that of equation (2) will be justified either when there is little or no pairwise homoplasy, or when the expected pattern of homoplasy over the tree is understood, so that some function of the input distances can be viewed as an approximation of the true patristic distances. Homoplasy would be built into a statistical model relating input distances to true patristic distances. This raises a limitation of the distance fitting approach, however. Suppose that, rather than the model implied by equation (2) one has a simpler model where:

\[
D(i,j) \sim N[\Delta(i,j), V(i,j)].
\]

\( D(i,j) \) is an approximation of the true phenetic distance. This is, of course, a quite plausible model for \( D(i,j) \). It has an important difference from a model of the form of equation (2). \( D(i,j) \) values here need not show any simple relationship to the \( \Pi(i,j) \) values, because \( \Delta(i,j) \) is a function of both \( \Pi(i,j) \) and \( \Theta(i,j) \). Equation (3) represents the more conservative stance; one may be able to expect only that the input distances approximate the true number of character differences between taxa.

Even this more conservative assumption, that molecular distances are reasonable approximations to phenetic differences between taxa, may be problematic when there are observed departures of the distances from the metric conditions. Perhaps the most notorious distance data in this respect have been immunological distances between taxa (for a review, see Romero-Herrera et al., 1978). For these types of distance data, some model of the distances is needed in order to explain their relationship to true phenetic or true patristic distances. I (1985) have developed a simple model of immunological distances that may account for the departures of these distances from metric properties and, more importantly, may explain their relationship to underlying phenetic distances. If the model is realistic, then immunological distances sometimes can be expected to be rather poor approximations of phenetic (and certainly patristic) distances between taxa. However, in exposing the bias in these distances, the model suggests at the same time that a transformation of the initial distances to phenetic (not patristic) distance approximations is possible.

The lack of justification of fit methods in the absence of a model that relates the input distances to true patristic distances (and not just true phenetic distances) can be understood by analogy to the analysis of character data by phylogenetic methods. Phenetic (usually Manhattan-metric) distances computed from these data are generally not regarded as reflecting degree of patristic dissimilarity. However, if UPGMA clustering is used for phylogenetic inference, the distances are viewed as reflecting "patristic association" as measured along an ultrametric scale. The effect of UPGMA is to achieve a "fit" between these input distances and the derived ultrametric distances (Farris, 1979). It is argued (e.g., Farris, 1983:29) that the presence of homoplasy makes this assumed relationship misleading. Now con-
sider the case where the expected values of the input distances are true phenetic distances. The same argument holds for the fit of these input distances to derived path-length distances; the presence of homoplasy also makes this assumed relationship misleading. The attempt to approximate phylogenetic trees by degree-of-fit methods leads to a bias analogous to that implied by the analysis of the distances in UPGMA clustering for phylogenetic inference.

Homoplasy will therefore be misleading for degree-of-fit methods, unless it can be taken into account in achieving a model relating input to true patristic distances. A method of inference is required for the more restrictive case frequently implied by molecular distances, where input distances may be supposed only to approximate phenetic distances in expectation and little is known about the distribution of homoplasy over the tree. An alternative to fit methods for the analysis of distance data is exactly the same strategy as employed for character data, where a minimum-length tree based upon the given data is considered the best means of inference of the true phylogenetic tree (e.g., Kluge and Farris, 1969; Eldredge and Cracraft, 1980). For character data, the character-Wagner algorithm is used to build a tree topology such that when the character data are assigned to the tree, the sum of the Manhattan distances along segments is approximately minimal. The distances measured in this way between taxa will be greater than or equal to the Manhattan distances computed from the character data, because of homoplasy. The corresponding distance problem, at least for molecular distances that are approximations to phenetic, Manhattan distances, can be posed analogously: Can a topology be found such that, when distances are assigned to the segments under the constraint that the $P(i,j)$ values are greater than or equal to the $D(i,j)$ values, the sum of the distances along the segments is approximately minimal?

While such a definition of optimality for phylogenetic inference from distance data is analogous to that for character data, the loss of information inherent in distance data poses algorithmic problems. A likely algorithm for the calculation of such a topology is the distance-Wagner algorithm of Farris (1972). While this method bears some resemblance to the character-Wagner algorithm, the relationship between the two is unclear. Note that the use of the distance-Wagner algorithm need not imply that the Wagner criterion (parsimony) is optimized. Both Farris (1972, 1981) and Felsenstein (1982) characterized the distance-Wagner algorithm as one not related to parsimony but rather as an algorithm attempting a fit between input and derived distances. Felsenstein (1982:395) stated,

Farris (1972) presented a method, the distance Wagner method, which has been widely used. Despite its name, it is not related to the Wagner parsimony method. It fits expected to observed distances, though without an explicit measure of goodness of fit.

In the following sections, I will argue that there is a useful relationship between the two methods in that the distance-Wagner algorithm can be viewed as attempting the same sort of stepwise parsimonious tree-building as the character-Wagner algorithm. The modified Farris method of Tateno et al. (1982) will then be compared with the distance-Wagner algorithm in terms of approximation of most parsimonious trees. I suggest an alteration of the distance-Wagner method that may lead to improved approximations. In the final section, the methods will be evaluated by simulations in terms of their recovery of a tree topology corresponding to a shortest-length tree rather than in terms of degree-of-fit.

**THE DISTANCE-WAGNER METHOD**

The distance-Wagner algorithm, like the character-Wagner algorithm, leads to the construction of a tree by the progressive addition of taxa (OTUs; Farris, 1970, 1972). A new OTU, C, can be attached between any two nodes, G and H, on the tree forming a new node, F, between C, G and H.
G and H can be OTUs or nodes previously created by the addition of an OTU to the tree (Farris [1970] called these latter types “HTUs” for hypothetical taxonomic units). G and H are chosen in order to minimize the “special similarity” (SS) between G and H, with C as a reference point. The equation is

\[
SS[C,(G,H)] = \frac{1}{2}[D(C,G) + D(C,H) - D(G,H)],
\]

(4)

where \( D \) is the phenetic (in this case Manhattan) distance between OTUs:

\[
D(A,B) = \sum |A_i - B_i|
\]

for \( n \) characters.

\( SS[C,(G,H)] \) can be viewed as a measure of the distance between object C and segment (GH), or equivalently between C and the new HTU, F. Although such a distance is computed between C and the segment GH, this value need not be assigned to the new segment formed by the addition of C to the tree. Instead, the distance can be considered simply as a criterion for addition of an OTU to the tree. While the distance-Wagner algorithm uses this distance in both ways, the algorithm developed here uses the distances solely as an addition criterion. In the usual character-Wagner algorithm, equation (4) is similarly used to provide an addition criterion, indicating the number of extra steps (character-state changes) that would necessarily have to exist if the OTU were added in a particular place. The choice of an addition to the tree to minimize this quantity makes the algorithm “greedy” (sensu Klee, 1980). One hopes to achieve an overall minimum-length tree by a succession of most-parsimonious additions of new OTUs to the tree.

The meaning of equation (4) is best understood initially in terms of an example using character data with a single binary character having states \( a \) and \( b \). In Figure 1, let G, H, etc., represent particular nodes (OTUs or HTUs) on the network, and let C be the OTU which is to be added. Equation (4) could then be used, for example, to evaluate the distance of C to the segment GH, based upon the given binary character. Table 1 gives the resulting values for \( SS[C,(G,H)] \) for different values of the character on G, C and H. If Figure 1 is considered as a segment of a phylogenetic tree, then it can be seen that \( SS[C,(G,H)] \) equals 1 when an extra character-state transition would be implied by the addition of OTU C to the branch GH. Over a number of characters, the sum of these values for a particular branch gives the total number of such extra steps. The choice of G and H that minimizes \( SS[C,(G,H)] \) then corresponds to the choice of G and H that minimizes the number of extra steps. In this way, the use of equation (4) implies stepwise parsimony (minimization of the number of implied character-state transitions) in building the tree. A limitation of equation (4) in this context should be noted. The equation should be applied only to taxonomically close neighbors, because in providing an estimate of the number of extra steps it implies an assumption of no homoplasy in segments CG, CH, and GH. For example, \( D(G,H) \) is taken as a good estimate of \( II(G,H) \).

In the character-Wagner algorithm, character states are assigned to HTUs as they are created, so that \( SS[C,(G,H)] \) may be computed in the same way when G and/or H are HTUs. Given only a matrix of pairwise distances, \( D \), the computation described above for C, G and H will not be straightforward. Because the character states for the HTUs are unknown, they cannot be used to compute \( D(C,G) \) and \( D(C,H) \) for equation (4); as a result, these values must be estimated by the algorithm. In the distance-Wagner algorithm, Farris (1972) estimated the distance of all unattached OTUs to an HTU at the time of formation of that HTU on the tree. For example, if OTU M had just been added to the tree in Figure 1, HTU G would be newly created and the distance of G to any unattached OTU, such as C, would then be estimated (distances within the tree from any node, say Q, to the new HTU are basically computed using equation 1; see Farris, 1972). As an estimate of \( D(C,G) \), Farris used, for all other nodes, N, currently on the tree:
\[ D(C,G) = \text{maximum over } N \text{ of } \left[ D(C,N) - D(N,G) \right]. \] (5)

Each node on the tree provides an estimate of \( D(C,G) \) and the maximum of all of these estimates is chosen. This forms the basis for the distance-Wagner method. I will use, for example, \( D[C,(G,H)] \) rather than \( SS[C,(G,H)] \) to refer to cases where phenetic (input) distances are unknown and estimates of these values were used in the application of equation (4).

The use of equation (5) for the estimation of a distance such as \( D(C,G) \) explains the relationship of this algorithm to degree-of-fit methods. Farris (1972) noted that any path-length distance, \( P(A,B) \), derived by the algorithm is effectively calculated as a sum of such greatest-lower-bound (glb) quantities. This implies that the pairwise homoplasy, \( H(A,B) = P(A,B) - D(A,B) \), is also a glb approximation. Such values therefore "represent the minimum amounts of homoplasy between OTUs A and B that are consistent with a particular data matrix of phenetic differences" (Farris, 1972:650). In this sense, the pairwise homoplasy is minimized, under the constraint that \( P(A,B) \) is greater than or equal to \( D(A,B) \). This is equivalent to the minimization of the sum over all pairs of OTUs, of \( P(A,B) - D(A,B) \), where \( P(A,B) \) is computed by glb approximations. Farris (1972) in fact used this quantity as a means of evaluation of the success of the algo-

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|}
\hline
OTU & G & H & C & SS[C,(G,H)] \\
\hline
\hline
\[ a \] & \[ a \] & \[ b \] & \[ b \] & \[ 1 \] \\
\hline
\[ a \] & \[ b \] & \[ b \] & \[ 0 \] \\
\hline
\[ b \] & \[ b \] & \[ b \] & \[ 0 \] \\
\hline
\end{tabular}
\caption{The value of \( SS[C,(G,H)] \) for different character-state combinations for C, H, and C.}
\end{table}

rithm. Because the sum of all differences between \( P(A,B) \) and \( D(A,B) \) is minimized, the algorithm can be regarded as attempting a fit between patristic and input distances.

The estimation of \( D(C,G) \), for the application of equation (4) as an addition criterion, also suggests a relationship of the distance-Wagner algorithm to the problem of the approximation of a minimum-length tree. Equation (4) is effectively used for this purpose in the character-Wagner algorithm. The success of such an approximation in the distance case, however, depends upon how well the unknown distances like \( D(C,G) \) can be approximated. Clearly, for the data of Table 1, \( D[C,(G,H)] \) will still reflect the number of extra steps as long as \( D(C,G) \) and \( D(C,H) \) are both approximately 1.

How well can the distance-Wagner algorithm approximate a most-parsimonious tree? It will be revealing to examine an example of the application of the algorithm in which the initial distances are derived from data on a single binary character with two states, \( a \) and \( b \). Knowledge of these states can then be used to evaluate the behavior of the algorithm. Figure 2 shows a partially constructed tree for OTUs M, N, L and H, with HTUs J and G. The tree construction is assumed to be based upon a set of characters, but the example will focus on a single character showing some homoplasy in order to understand its contribution to the computed distance from an unattached taxon, C, to a segment, GH. The character states of the OTUs are shown in circles on the figure.

Suppose that L and then H have just been added to the tree in Figure 2. Suppose also that C may be added to the tree
and that one wishes to determine the distance of C to segment GH. Using equation (5), $D(C,G)$ was computed (after H joined the tree) as 1. $D(H,G)$ was computed, using equation (4), as 0. $D(C,H)$ is found in the original distance matrix as 1. Therefore, applying equation (4),

$$D[C,(GH)] = \frac{1}{2}(1 + 1 - 0) = 1.$$ 

This value can be interpreted to mean that the addition of OTU C to segment GH implies an extra step for this character. This is reasonable from direct examination of the tree with the character states shown and would correspond to the result for the character-Wagner method which uses the actual character data. A change from a to b, or vice versa, now would have to occur twice on the tree. The distance-Wagner algorithm in this example has effectively counted the number of extra steps (in this case, for a single character) in computing the distance of C to the interval GH. Thus, this algorithm appears to treat the distances in such a way that a parsimonious estimate is made for some underlying, unobserved, character data.

The estimates provided by the distance algorithm are frequently rather poor. Suppose that, after OTUs L and H joined the tree, two other OTUs P and Q, both with state b, joined the tree between J and G, creating new HTUs, I and K (Fig. 3). How should this affect the potential addition of OTU C on segment HG? Examination of the tree suggests that the addition of C would not imply any extra steps ($D[C,(GH)] = 0$). However, if equation (4) is reapplied with distance-Wagner estimates, none of the values in the equation will have changed, because the distances of all unattached OTUs to G would be computed before P and Q joined the tree. The calculated distance from C to GH then is still 1, wrongly implying an extra step in terms of this character. This sort of error implies that the algorithm can fail to make the most-parsimonious addition at any given step and, in fact, can make an addition that is purely an artifact of the algorithm.

The problem demonstrated above seems to be a result of a failure of the distance algorithm to take into account other pertinent information on the tree at the time a new OTU is to be added. The distance of a new OTU to a segment seems to be wrongly biased by the order in which the OTUs were added to the tree. A problem may also exist in the choice of the maximum value in equation (5). Suppose that nodes P and Q are included in the computation of $D(C,G)$ using equation (5). The maximum value for $D(C,G)$ is still 1, as
given by the choice of N equal to H. This implies that \( D[C,(GH)] \) will still be 1.

It appears that the distance-Wagner algorithm is not sensitive to changes to the tree that should affect the estimation of distances such as \( D(C,G) \). The following sections serve to develop an alternative approach to the effective estimation of distances for the application of equation (4).

**THE TATENO, NEI, AND TAJIMA METHOD**

Tateno et al. (1982) outlined a modified Farris method that seems to avoid the problem of estimation of distances based on only part of the existing tree and uses a form of equation (5) without the maximum. Their algorithm will be revealing in terms of possible alterations to the distance-Wagner algorithm for the approximation of most-parsimonious trees. However, Tateno et al. considered the case where a transformation of the distances could be assumed to be approximations to true path-length distances, justifying a distance-fitting approach. Their algorithm was designed to provide better estimates of branch lengths for this purpose.

Tateno et al. misrepresented the distance-Wagner algorithm in their paper. I will first clarify this point and then will derive a mathematical expression for their algorithm before it is evaluated in terms of approximation of most-parsimonious trees. Figure 4 shows an example which is a redrawing of part of their Figure 1. OTU A has joined the tree and one wishes to compute \( D[B,(XY)] \). Tateno et al. argued that in an application of the distance-Wagner method, the usual expression for the distance of B to segment XY (from equation [4]).\[ D[B,(XY)] = \frac{1}{2}[D(B,X) + D(B,Y) - D(X,Y)] \] (6)

would not be used. Instead they claimed that this distance would have been computed, before A (and therefore Y) had joined the tree, as:

\[ D[B,(XY)] = \frac{1}{2}[D(B,3) + D(B,X) - D(3,X)] \] (7)

Tateno et al. correctly pointed out that this estimation does not take into account some changes in the tree (namely, the addition of A and Y), and they went on to recommend a different equation which is equivalent to equation (6). While the claim that the distance-Wagner algorithm is not properly sensitive to changes in the tree during its construction may be true, Tateno et al. used an incorrect form for equation (6) and in fact recommended the current form as an alteration. The confusion here may be due in part to the fact that, if the maximal estimate of \( D(B,Y) \) is \( D(B,3) - D(Y,3) \) when applying equation (5) (see Fig. 4), equation (6) actually is equivalent to equation (7).

Tateno et al. went on to suggest an alteration to the distance-Wagner algorithm that does alter the sensitivity of the algorithm to changes in the tree as it is being constructed. In the distance-Wagner algorithm, \( D(B,X) \) would be computed before A and Y had joined the tree (Fig. 4). The calculation of \( D[B,(XY)] \) by equation (6) then does not reflect the information gained by the addition of A and Y to the tree. The modified distance-Wagner method of Tateno et al. does make use of the entire tree topology at the time of computation of a distance like \( D[B,(XY)] \) and so avoids one weakness of the original distance-Wagner method. Their exact algorithm is only outlined in their paper, so I derive an algebraic expression for it here. Referring again to Figure 4, \( D[B,(XY)] \) will
be computed using equation (6). The estimation of the unknown quantities, \(D(B,X)\) and \(D(B,Y)\), are computed differently from the distance-Wagner algorithm. Tateno et al. computed \(D(B,Y)\) as the average of \(D(B,\alpha) - D(\delta,X)\) over all OTUs, \(\alpha\), on the Y side of the tree from Z (in this case, OTUs A and 3; see Fig. 4). Similarly, \(D(B,X)\) will be the average of \(D(B,\delta) - D(\delta,X)\) over all \(\delta\) on the X side of the tree from Z. In contrast to equation (5), then, an average of several estimates is used, rather than a maximum. The two sets of OTUs can be designated as \(\{\alpha\}\) and \(\{\delta\}\). The equations for \(D(B,Y)\) and \(D(B,X)\) can be written:

\[
D(B,Y) = \frac{1}{\tau} \sum_\alpha [D(B,\alpha) - D(\alpha,Y)] 
\]  

(8)

and

\[
D(B,X) = \frac{1}{\mu} \sum_\delta [D(B,\delta) - D(\delta,X)]
\]  

(9)

where \(\tau\) is the number of OTUs in \(\{\alpha\}\), and \(\mu\) is the number of OTUs in \(\{\delta\}\). Similarly, \(D(X,Y)\) would be computed as an average of all possible estimates:

\[
D(X,Y) = \frac{1}{\mu} \sum_\delta \left( \frac{1}{\tau} \sum_\alpha [D(\alpha,\delta) - D(\alpha,Y) - D(\delta,X)] \right)
\]  

(10)

where \(\{\alpha\}\) and \(\{\delta\}\) are defined as above. Substituting these values into equation (6) and simplifying:

\[
D[B,(XY)] = \frac{1}{\mu \tau} \left[ \frac{1}{\mu} \sum_\alpha [D(B,\alpha) - D(\alpha,Y)] \right]
\]

\[
+ \frac{1}{\tau} \sum_\delta [D(B,\alpha) - D(\alpha,Y)]
\]

\[
- \frac{1}{\mu} \sum_\delta \left( \frac{1}{\tau} \sum_\alpha [D(\alpha,\delta) - D(\alpha,Y) - D(\delta,X)] \right).
\]

(11)

The distance of B to Z—\(D[B,(XY)]\)—is seen to be the average of all possible special similarities between \(\alpha\) and \(\delta\) with B as a reference point, with the restriction that \(\delta\) and \(\alpha\) are any OTUs currently on the tree with Z on the path between them. Equation (11) is quite simple computationally if the OTUs can be easily sorted into \(\alpha\) and \(\delta\) sets. The computation of the branch length for the new segment BZ is given by the same value computed for \(D[B,(XY)]\). In this way, the estimation of branch lengths and the criterion for OTU addition both make use of the entire topology at the time of addition and, in fact, only need to use the distances between OTUs.

How well does this method approximate most-parsimonious trees? One indicator will be the degree to which the most-parsimonious addition is likely to be made at each step. The method applied to the example of Figure 3 gives a result of
\( D[C,(GH)] = \frac{1}{2} \), which represents at least a small improvement over the value of 1 for the Farris algorithm, recalling that the distance should be 0 if it is to reflect the fact that no extra steps would be implied. Interestingly, if P and Q are deleted from the tree, the Tateno et al. method yields \( D[C,(GH)] = \frac{1}{2} \), which is not as good as the value of 1 for the Farris’ method, since this is exactly the desired value if P and Q are not on the tree.

A general weakness of the Tateno et al. modification of the Farris method is revealed if Figure 3 is changed by adding several OTUs on to branches MJ and NJ, all having the state \( a \) (Fig. 5). Using the modified method, as the number of new branches with state \( a \) increases, the value of \( D[C,(GH)] \) approaches 1! Clearly, however, no matter how many OTUs to the left of P and Q have the \( a \) state, the \( D[C,(GH)] \) value should be 0, or at least close to this value as an estimate. The addition of OTU C to the tree at GH does not imply any extra steps. The method of Tateno et al., then, cannot be guaranteed to estimate distances that can be used as addition criteria for approximating most parsimonious trees.

A NEW ALGORITHM

The preceding mathematical representation of the Tateno et al. algorithm suggests a better method for the calculation of distances like \( D[C,(GH)] \). I derive a new method of approximation by repeated application of equation (4). Figure 5 can be used as an example; again OTU C may potentially join the tree at segment GH. Initially, the estimate of \( D[C,(GH)] \) is expressed as (equation [4]):

\[
D[C,(GH)] = D[C,F] = \frac{1}{2} [D(C,G) + D(C,H)] - D(G,H). \tag{12}
\]

\( D(C,G) \) in Figure 5 can be estimated using the same basic equation:

\[
D(C,G) = D[C,(KL)] = \frac{1}{2} [D(C,K) + D(C,L)] - D(K,L), \tag{13}
\]

and similarly,

\[
D(G,H) = D[H,(KL)] = \frac{1}{2} [D(H,K) + D(H,L)] - D(K,L). \tag{14}
\]

This process of substitution can be repeated for all the remaining unknown terms (all those distances between two HTUs or an HTU and an OTU) until the initial expression (equation [4]) is expressed only in terms of known distances among OTUs on the tree. This process of substitution leads to the following general equation:

\[
D[C,(GH)] = \frac{1}{2} \left\{ \sum_{\alpha} \sum_{\delta} \frac{1}{2^{N(\delta)}} \cdot \frac{1}{2^{N(\alpha)}} \cdot D(\alpha,C) + D(\delta,C) - D(\alpha,\delta) \right\}, \tag{15}
\]

where \( \alpha \) is any OTU on the G side of C and \( \delta \) is any OTU on the H side of the tree from C. For any given \( \alpha \), \( N(\alpha) \) is the number of existing nodes between F and \( \alpha \). For example, in Figure 5, \( N(\alpha) \) equals 2 when \( \alpha = Q \).

Because \( \alpha \), \( \delta \), and C are all OTUs the above equation can be re-expressed as

\[
D[C,(GH)] = \sum_{\alpha} \sum_{\delta} \frac{1}{2^{N(\alpha)+N(\delta)}} \cdot SS[C,(\alpha,\delta)]. \tag{16}
\]

The distance from C to GH is then seen to be a weighted average of the special similarity, with C as a reference point, of all pairs of OTUs currently on the tree, one member of the pair from each side of the tree (i.e., the two members would have F [Fig. 5] on the path between them). The general formula is equivalent to computing \( D[C,(GH)] \) by successive applications of the basic special similarity formula.

This method, applied to the example of Figure 3, yields a value for \( D[C,(GH)] \) of \( \frac{1}{2} \), which is closer to the desired value of 0 than are the estimates by Farris’ method (1) or the modified Farris’ method (\%). The same estimate would be obtained for the example of Figure 5, which also has the desired value of 0. In general, this method should better estimate most-parsimonious trees using distance data. The
above simple examples suggest that neither the method of Farris nor that of Tateeno et al. adequately copes with the contribution of homoplasious characters to the distance of any unattached OTU to a given segment.

Further contrasts of my algorithm with those of Farris (1972) and Tateeno et al. (1982) are instructive, particularly as they relate to the choice of a goal of approximating a most-parsimonious tree rather than a goal of distance fitting. In both the distance-Wagner and modified-Farris methods, the value of $D[C,(GH)]$ is used for two purposes: it acts as a criterion for addition of C to the segment GH; and it acts as a length of that segment if C does join at GH. For the new estimation method given by equation (16), only the role as an addition criterion is to be emphasized. The entire tree topology can be created without the assignment of any branch lengths along the way. I emphasize the recovery of tree topology over the approximation of branch lengths. Naturally, given a final tree topology and the initial distances, branch lengths may be estimated by a linear-programming approach. The overall tree length can be minimized under the constraint that the patristic distances are greater than or equal to the phenetic distances. Such a procedure has been applied to tree topologies by Beyer et al. (1974).

**MONTE CARLO COMPARISON OF ALGORITHMS**

Under the assumption that a "greedy" algorithm will lead to the better estimation of a minimum-length tree, the new algorithm might be expected to outperform the usual distance-Wagner algorithm by better estimating the number of extra steps implied by any potential addition of an OTU to a tree. Two simulation studies were performed to compare the algorithms. Binary-character data were simulated for a set of OTUs and, from the resulting data, a Manhattan-metric distance matrix was computed. The calculation of the topology in each case relied only upon input distance matrices and the topologies produced by the two algorithms were optimized using the character data. Optimization of the tree provided the actual length of the tree implied by the particular topology. For comparison, the character-Wagner algorithm was used to approximate a tree using the character data. Distance- and character-Wagner programs were provided by Don Colless, CSIRO Division of Entomology, Canberra.

1. Random data.—In this aspect of the study, 1,000 data matrices were produced, each being 10 OTUs by 10 characters. The actual character values were coded as 0 or 1, and in this first study were assigned randomly to the matrix. As expected, this produced data with little structure; the average length of the most-parsimonious tree as approximated by the character-Wagner algorithm was 17.8 over a sample of 128 trees.

Each algorithm used the same basic criterion in building the tree. In all runs, an OTU, C, was chosen to be added next to the tree at a segment, GH, so that $D[C,(GH)]$ was minimized. The algorithms were given a random starting pair to initiate the tree.

The results are presented in Table 2. The new algorithm was separately compared to the distance- and character-Wagner programs. The entries in the table give the number of times out of 1,000 trees that one algorithm or the other found a shorter tree; on all other occasions the algorithms found trees of the same length. The new algorithm bettered both the distance and character algorithms for these random starts, though its advantage over the character algorithm is negligible. It is a satisfying result, however, that the particular brand of "greediness" implicit in this new distance algorithm does just as well as the algorithm that uses the character data itself.

The basic character-Wagner algorithm presents the possibility of further refinements that can improve its performance. For example, it is possible to re-optimize the tree at each step of the procedure (i.e., character states can be re-assigned to the
HTUs after new OTUs are added to the tree). This strategy may produce better approximations (D. Swofford, pers. comm.). It is also possible to try so-called multiple-tree approaches where each tree is optimized and the shortest one chosen. Using the same 1,000 data sets, a multiple-tree approach was created by running the algorithms with all possible starting pairs. The shortest tree, as determined by optimization of the topology with the character data, was then chosen. Table 2 shows these results. Over all possible starting pairs, the advantage of the new algorithm over the character algorithm disappears. For both comparisons, the number of occasions in which one algorithm produced a shorter tree than the other is greatly reduced from the case where a single starting pair was used. The new algorithm and the distance-Wagner algorithm are now hardly differentiated.

II. Structured data.—The random data described above resulted in a high degree of homoplasy, as indicated by the average tree length for the 10 characters of 17.8. The algorithms were investigated also over data sets that were somewhat more structured. These data sets, again 10 OTUs by 10 characters, were created from initial data sets corresponding to trees with no homoplasy. Then a random “shotgun” effect was used to arbitrarily change some 0s to 1s, and vice versa. A different initial tree was used for each 100 runs. The average length of a sample of 100 trees was 15.3 as indicated by the character-Wagner algorithm. For a total of 1,000 trees, the results are shown in Table 2. Overall, there was slightly less disagreement among the algorithms than for the random data. For the random starting pair, the advantage of the new algorithm over both the distance-Wagner and the character-Wagner algorithm is greater than for the random data, running approximately 2:1 in favor of the new algorithm when different-length trees were found. Again, this advantage disappears for the search over all possible starting pairs. The character algorithm in particular generally found the shorter tree when the two algorithms differed in the lengths of their best approximations, though the two only found different-length shortest trees 2% of the time.

### Table 2. Comparison of new algorithm with distance-Wagner and character-Wagner results. Values give number of trees out of 1,000 that were found to be shorter by each algorithm for each pairwise comparison.

<table>
<thead>
<tr>
<th>Type of data</th>
<th>New algorithm</th>
<th>Distance-Wagner algorithm</th>
<th>New algorithm</th>
<th>Character-Wagner algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>145</td>
<td>84</td>
<td>113</td>
<td>105</td>
</tr>
<tr>
<td>Structured</td>
<td>109</td>
<td>46</td>
<td>99</td>
<td>52</td>
</tr>
<tr>
<td>All starting pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>23</td>
<td>19</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>Structured</td>
<td>7</td>
<td>12</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

### DISCUSSION

Several limitations were noted in the simulations described above. First, the evaluation of any algorithm is most appropriately carried out in terms of the true optimum value rather than with respect to the result of another algorithm. That one algorithm “betters” another 50% of the time means very little if both algorithms do a very poor job of estimating the optimal solution. In this study, the favorable comparison of the new algorithm to two widely used algorithms, that are considered to perform moderately well, is seen as a useful and promising initial test of the algorithm.

The success of the new algorithm relative to the character algorithm for these data sets is interesting in the context of the design of greedy algorithms. In the character-Wagner algorithm, each character indicates (generally) a “yea” or “nay” on the question of whether the addition of a new OTU at a particular point implies an extra step. The answer is based upon a most parsimonious interpretation of the character data with the addition of the new OTU. This answer can of course change with the addition of other OTUs to the
tree; this is the obvious danger of a stepwise parsimony approach. The new algorithm may be slightly more robust to these stepwise problems. Equation 13 does not provide an all-or-nothing statement about whether an extra step will be implied by an addition to the tree. Instead it provides in some sense a likelihood of an extra step based upon the entire tree structure present at the time the new OTU is to be added. This slightly more "laid-back" greediness may account for the greater success of the new algorithm, using only distances, compared to the character algorithm for randomly chosen starting pairs.

Any enthusiasm for the new algorithm must be tempered somewhat by the results for the search over all possible starting pairs. These results suggest that the starting point can make a big difference in the length of tree discovered, and that the character algorithm in particular benefits from trying many different initial starts.

The utility of the new algorithm and, for that matter, the basic distance-Wagner algorithm is limited without the addition of other search procedures, such as branch swapping, to these basic algorithms. The work presented here does suggest that the new algorithm may provide a useful starting point for other procedures. However, the use of a multiple-tree or a branch-swapping approach with a distance algorithm may be impractical because in practice it is likely to involve repeated application of time-consuming linear-programming computations. Therefore, for all cases except those with a small number of taxa, the new algorithm appears to represent some progress in the development of useful heuristic methods for the analysis of distance data. An alternative approach to the same problem that requires linear programming at each step can be found in Waterman et al. (1977).

Finally, to emphasize a point made earlier, the role of the algorithm presented here is the calculation of a tree topology that can yield a good approximation to a most-parsimonious tree. The algorithm does not estimate any intervening branch lengths in deriving the topology. The resulting tree length will only be estimated in practice, given a topology and the input distances, by a linear-programming approach. The best tree is viewed as the shortest-length tree, and not the tree that in some sense has the best fit to the input distances. This paper hopefully will assist in the further development of distance algorithms for the approximation of most parsimonious trees.

ACKNOWLEDGMENTS

I wish to thank Mike Austin, Charles Marshall, Don Colless, Joe Felsenstein, Dave Swoford, Sam Rogers and Margaret Adomeit for their comments on earlier versions of the paper.

REFERENCES


Received 2 January 1985; accepted 29 April 1985.