Letter to the Editor

A Note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's Algorithms for Inferring Phylogenies from Evolutionary Distances

Olivier Gascuel
Département d'Informatique Fondamentale, LIRMM

Introduction

Agglomerative algorithms iteratively pick a pair of taxa, create a new node that represents the cluster of these taxa, and compute a new distance matrix with reduced size where both taxa are replaced by this node. The cycle is repeated until the number of taxa becomes three (or two for rooted trees). This general scheme was first applied to additive unrooted trees by Sattath and Tversky (ADDTREE method; 1977) in the context of mathematical psychology. The neighbor-joining (NJ) method of Saitou and Nei (1987) widely popularized this approach in the phylogenetic study. This method is based on the minimum evolution principle and provides trees with near-minimal sum of branch-length estimates. An alternative formulation of the NJ method with reduced computational complexity was given by Studier and Keppler (SK method; 1988), while Rzhetsky and Nei (1992, 1993) clarified the theoretical foundation of the minimum evolution principle. Several simulations (Saitou and Nei 1987; Nei 1991) have shown a high relative efficiency of ADDTREE and of the NJ method in recovering the true topology. These studies have also shown that ADDTREE and the NJ method, whose principles seem very different, are in fact close and usually provide identical or similar trees. For example, they obtain the same tree with Case's (1978) data. The explanation for this proximity was given by Saitou and Nei (1987) for four taxa. In this note, we account for this proximity regardless of the number of taxa, and we show that the minimum evolution principle, as employed in the NJ method, is very close to the neighborliness used by Sattath and Tversky (1977) and by Fitch (1981) in a nonagglomerative way. In the following, we recall the principles of the ADDTREE, NJ, and SK methods and explain why they are so close. We will only be concerned with the construction of the tree shape, and not with branch-lengths estimation. For the latter aspect, we refer the reader to the original papers and to

Rzhetsky and Nei (1993), which detail an interesting solution, similar to ADDTREE's estimation procedure (which is only outlined in Sattath and Tversky 1977). Straightforward estimation procedures based on non-negative least-squares regression are given by Lawson and Hanson (1974, pp. 158–165) and by Barthélemy and Guénoche (1991, pp. 62–66).

The ADDTREE Method

In an additive tree, any four distinct objects, x, y, i, j, appear in one of the configurations of figure 1. The pattern of distances that correspond to these configurations are, respectively,

$$d_{xy} + d_{ij} < d_{xi} + d_{yj} = d_{xj} + d_{yi},$$

$$d_{xi} + d_{yj} < d_{xy} + d_{ij} = d_{xj} + d_{yi},$$

$$d_{xj} + d_{yi} < d_{xy} + d_{ij} = d_{xi} + d_{yj}.$$

The task is to select the most appropriate configuration on the basis of an observed evolutionary distance ∂. When inequality

$$\partial_{xy} + \partial_{ij} \leqslant \mathrm{Min}(\partial_{xi} + \partial_{yj}, \partial_{xj} + \partial_{yi})$$

holds, the first configuration above is the most appropriate, and the objects x and y (as well as i and j) are "neighbors." For each pair x, y, ADDTREE examines all objects i, j and counts the number of quadruples in which x and y are neighbors. It thus obtains the "neighborliness" of the pair x, y, denoted as N_{xy} in the following. For the sake of simplicity, let $x = 1$ and $y = 2$. Then we have

$$N_{1,2} = \sum_{r \geqslant i > j \geqslant 3} [H(\partial_{1i} + \partial_{2j} - \partial_{1,2} - \partial_{ij}) \\ H(\partial_{1j} + \partial_{2i} - \partial_{1,2} - \partial_{ij})], \tag{1}$$

where H denotes the Heaviside function (if $t \geqslant 0$, then $H(t) = 1$; else $H(t) = 0$) and r is the number of remaining taxa. The pair x, y with the highest neighborliness is selected, and its members are combined to form a new element z that replaces x and y in the subsequent analysis. The evolutionary distance between z and any other
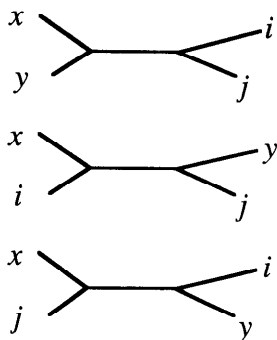
FIG. 1.—The three possible configurations of four objects in an additive tree.

element u is set to $(\partial_{ux}+\partial_{uy})/2$. This process is repeated until only three objects remain. At each stage of clustering ADDTREE examines every quadruple, so that the whole complexity is in $O(n^5)$, where n is the number of taxa.

The NJ Method

The NJ algorithm is similar to ADDTREE, but it uses the minimum evolution principle instead of the neighborliness to agglomerate pairs of nodes. This principle consists in choosing the tree with the smallest sum of branch lengths. Rzhetsky and Nei (1993) have shown that, under some mild assumptions, the minimum evolution principle is well founded when the lengths are estimated by the ordinary least-squares method. The NJ method proceeds in a heuristic manner and guarantees that a short tree is found, but not the shortest. At each stage of clustering, NJ considers that data are starlike, as shown in figure 2a. Then, it extracts the pair x, y which minimizes the length of the tree shown in figure 2b. Let the least-squares estimate of the length of the tree in figure 2b be denoted as S_{xy}. When $x = 1$, and $y = 2$, we have

$$S_{1,2} = \frac{1}{2(r-2)} \sum_{r \geq i \geq 3} (\partial_{1i}+\partial_{2i})$$
$$+ \frac{1}{2} \partial_{1,2} + \frac{1}{r-2} \sum_{r>j>i\geq 3} \partial_{ij}. \quad (2)$$

The pair x, y with the smallest value of S_{xy} is selected, and its members are combined in the same way as in ADDTREE. Computing the last term of formula (2) requires examining every pair i, j. Therefore, finding the best pair to be clustered has an $O(r^4)$ complexity, and the whole algorithm is in $O(n^5)$.

The SK Method

In a short note, Studier and Keppler (1988) proposed a new version of the NJ method. They suggest to replace the S criterion defined in equation (2) by the Q criterion defined as follows (having $x=1$ and $y=2$):

$$Q_{1,2} = (r-2)\partial_{1,2} - \sum_{r \geq i \geq 1} \partial_{1i} - \sum_{r \geq i \geq 1} \partial_{2i}. \quad (3)$$

But they do not provide any comparison between the S and Q criteria. In fact, S and Q are strongly related. Using equation

$$\sum_{r \geq j > i \geq 3} \partial_{ij} = \sum_{r \geq j > i \geq 1} \partial_{ij} - \sum_{r \geq i \geq 1} (\partial_{1i}+\partial_{2i}) + \partial_{1,2}, \quad (4)$$

we obtain

$$S_{1,2} = \frac{1}{2(r-2)} Q_{1,2} + \frac{1}{(r-2)} \sum_{r \geq j > i \geq 1} \partial_{ij}.$$

The last term in the latter equation is a constant. Therefore, minimizing S or Q is equivalent, and both criteria always select the same pair. The difference is that computing S_{xy} for any given pair x, y requires $O(r^2)$ time while Q_{xy} may be computed in $O(1)$. To achieve this goal, it is sufficient to compute all the terms $\sum_{r \geq i \geq 1} \partial_{zi}$ appearing in equation (3) before selecting the best pair, which may be done in $O(r^2)$. Finally, each clustering stage requires $O(r^2)$ time, and the whole algorithm is in $O(n^3)$.

The algorithms SK and NJ (and ADDTREE) differ in their way of combining the elements x, y of the selected pair. In SK, the distance ∂_{zu} between the new node z and any other element u is set to $(\partial_{ux}+\partial_{uy}-\partial_{xy})/2$. It follows that the reduced distance matrices computed by SK and by NJ are not the same. But this difference does not affect the further steps of the algorithms. In fact, it is easily seen from equation (3) that, when adding a constant k (here, $k=\partial_{xy}/2$) to every distance ∂_{zu} $(u \neq z)$, we obtain a new value Q' that satisfies $Q'_{ij} = Q_{ij} - 2k$ for any pair i, j. Since k is a constant, minimizing Q or Q' is equivalent. This shows that the same tree shape would be obtained by SK if it reducing matrices in the manner of NJ. Equivalently, the same tree shape would be obtained by NJ when employing SK's reduction, and it easily seen from equation (1) that the same holds for ADDTREE. What may be said is that SK's reduction simplifies branch-lengths estimation because
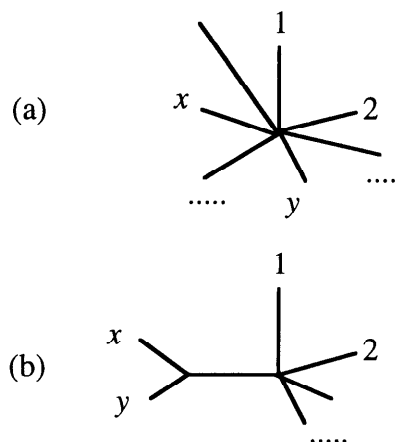

FIG. 2.—The starlike tree (a) with no hierarchical structure, and the tree (b) in which x and y are clustered.

$(\partial_{ux}+\partial_{uy}-\partial_{xy})/2$ is, under some assumptions, the least-squares estimate of the distance between z and u.

Elements given in this section prove that NJ and SK always obtain the same tree shape. Simple considerations show that both algorithms also provide identical branch lengths.

From the Neighborliness to the Minimum Evolution Principle

The ADDTREE method has a dichotomous view of neighborliness: the objects x and y either are or are not neighbors for a given pair i, j. It is possible to have a more gradual vision. The objects x and y may be seen as more or less neighboring, depending on whether the sum $(\partial_{xy}+\partial_{ij})$ is much smaller or not than both $(\partial_{xi}+\partial_{yj})$ and $(\partial_{xj}+\partial_{yi})$. A simple mathematical translation of this view consists in suppressing the Heaviside function in equation (1) and in considering that the "gradual neighborliness" of the pair x, y for the pair i, j is represented by the quantity $(\partial_{xi}+\partial_{yj}+\partial_{xj}+\partial_{yi}-2\partial_{xy}-2\partial_{ij})$. The greater is this quantity, the more neighboring are x and y (the same holds for i and j). A related approach proposed by Fitch (1981) is based on an "interior distance" between taxa and on the quantity $(\max(\partial_{xi}+\partial_{yj}, \partial_{xj}+\partial_{yi}) - (\partial_{xy}+\partial_{ij}))$. The overall gradual neighborliness, denoted as N', is defined as follows (having $x=1$ and $y=2$):

$$N'_{1,2} = \sum_{r\geqslant j>i\geqslant 3} (\partial_{1i}+\partial_{2j}+\partial_{1j}+\partial_{2i}-2\,\partial_{1,2}-2\,\partial_{ij}). \quad (5)$$

This expression may be simplified. We have

$$N'_{1,2} = (r-3) \sum_{r>i\geqslant 3} (\partial_{1i}+\partial_{2i})$$
$$- (r-2)(r-3)\partial_{1,2} - 2 \sum_{r\geqslant j>i\geqslant 3} \partial_{ij},$$

and using equation (4) we obtain

$$N'_{1,2} = (r-1) \sum_{r\geqslant i\geqslant 1} (\partial_{1i}+\partial_{2i})$$
$$- (r-1)(r-2)\partial_{1,2} - 2 \sum_{r\geqslant j>i\geqslant 1} \partial_{ij}.$$

Finally, according to equation (3), we get the following equation

$$N'_{1,2} = - (r-1)Q_{1,2} - 2 \sum_{r\geqslant j>i\geqslant 1} \partial_{ij}.$$

Since the last term is a constant, we have the desired result: maximizing the gradual neighborliness (N'), minimizing the sum of branch lengths (S), and minimizing Studier and Keppler's criterion (Q) are strictly equivalent. ADDTREE's neighborliness (N) is not identical to our gradual neighborliness (N'), but maximizing one or the other usually leads to the same tree

or to similar ones. This explains why all these methods are so close.

Conclusion

The contribution of the NJ method as formulated by Studier and Keppler (1988) remains important from a practical point of view. An $O(n^3)$ algorithm opens the way to large data matrices, containing, say, a few hundred taxa, and greatly facilitates reconstruction experiments using simulated or resampled (bootstrap, etc.) data.

This note establishes some relationships between the (gradual) neighborliness and the minimum evolution principle. Other properties are given by Rzhetsky and Nei (1992, 1993), while we have shown (Gascuel and Levy, in press) that, in the case of four taxa, the neighborliness and the minimum evolution principle are equivalent to the least-squares criterion with the positivity constraint on branch-length estimates. Our feeling is that much theoretical work remains to be done to better understand these criteria, to exhibit new, better ones, and finally to build more efficient methods.

LITERATURE CITED

BARTHÉLEMY, J. P., and A. GUÉNOCHE. 1991. Trees and proximity representations. Wiley, Chichester.
CASE, S. M. 1978. Biochemical systematics of members of the genus *Rana* native to western North America. Syst. Zool. **27**:299–311.
FITCH, W. M. 1981. A non-sequential method for constructing trees and hierarchical classifications. J. Mol. Evol. **18**:30–37.
GASCUEL, O., and D. LEVY. A reduction algorithm for approximating a (nonmetric) dissimilarity by a tree distance. J. Classif. (in press).
LAWSON, C. M., and R. J. HANSON. 1974. Solving least squares problems. Prentice Hall, Englewood Cliffs, N.J.
NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. Pp. 90–128 in M. M. MIYAMOTO and J. L. CRACRAFT, eds., Phylogenetic analysis of DNA sequences. Oxford University Press, Oxford.
RZHETSKY, A., and M. NEI. 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum evolution methods of phylogenetic inference. J. Mol. Evol. **35**:367–375.
———. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073–1095.
SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
SATTATH, S., and A. TVERSKY. 1977. Additive similarity trees. Psychometrika **42**:319–345.
STUDIER, J. A., and K. J. KEPPLER. 1988. A note on the neighbor-joining method of Saitou and Nei. Mol. Biol. Evol. **5**:729–731.