Letter to the Editor

A Note on the Neighbor-Joining Algorithm of Saitou and Nei[1]

James A. Studier and Karl J. Keppler†*
*Department of Microbiology, University of Illinois, Urbana; and †Department of Mathematics, Shepherd College, Shepherdstown, West Virginia

Saitou and Nei (1987) present an algorithm, which they call the neighbor-joining (NJ) method, for estimating an additive tree from a distance matrix D. If D is treelike (i.e., if the distances in D correspond exactly to those in an actual tree), then the NJ method correctly reconstructs the tree from D. If D is not treelike (i.e., contains some noise), then there can be ambiguities in the estimated tree. Saitou and Nei simulate such data and verify that the accuracy of the NJ method is roughly equivalent to that of the Sattath and Tversky (1977) method.

The minimum running time of the algorithm as formulated by Saitou and Nei is unclear. We present an alternative formulation that runs in time $O(N^3)$, where N is the number of operational taxonomic units (OTUs). We consider the $O(N^3)$ running time to be useful in studies that involve a large number of OTUs, possibly in connection with reconstruction experiments using simulated or resampled (bootstrap, etc.) data.

The proof given by Saitou and Nei that the correct tree is recovered if D is treelike is incorrect. We describe the error and supply a correct proof below.

The modified algorithm is as follows:

A. For each pair i, j of OTUs, compute

$$S_{ij} = (N-2)D_{ij} - R_i - R_j , \tag{1}$$

where D is N by N and

$$R_i = \sum_k D_{ik} . \tag{2}$$

B. Pick a pair i, j for which S_{ij} is the smallest. Create a new node u and infer distances:

$$D_{iu} = \frac{1}{2}(D_{ik} + D_{jk} - D_{ij}) \text{ for } k \neq i, j . \tag{3}$$

[Formula (3) is not the one given by Saitou and Nei, but it is the correct one if D is treelike and i and j are neighbors.]

C. The branch lengths from the new node are

$$D_{iu} = \frac{1}{2(N-2)}[(N-2)D_{ij} + R_i - R_j] \tag{4a}$$

and

$$D_{ju} = \frac{1}{2(N-2)}[(N-2)D_{ij} - R_i + R_j] . \tag{4b}$$

1. Key words: phylogenetic tree, neighbor-joining method.

Address for correspondence and reprints: James A. Studier, Department of Microbiology, University of Illinois, 407 South Goodwin Avenue, Urbana, Illinois 61801.

Mol. Biol. Evol. 5(6):729–731. 1988.

0737-4038/88/0506-0009$02.00

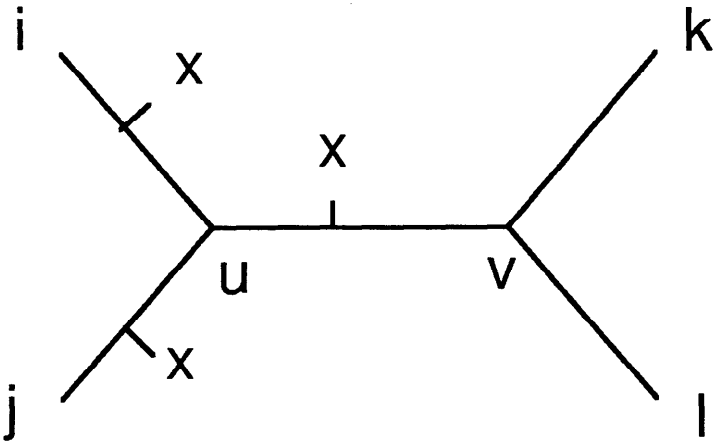

FIG. 1.—For the distinct OTUs i, j, k, and l, the subtree they determine includes two internal nodes u and v. Each arc shown is a sum of arcs from the overall tree. If k and l are neighbors, the x's represent locations at which other OTUs can intercept this subtree.

After deleting i and j from D and adding u, the process is repeated until the tree is complete.

From the definition of the S_{ij} one gets

$$S_{ik} - S_{ij} = \sum_{m \neq i,j,k} (D_{ik} + D_{jm} - D_{ij} - D_{km}) \tag{5a}$$

and

$$S_{kl} - S_{ij} = \sum_{m \neq i,j,k,l} [(D_{im} + D_{jm} - D_{ij}) - (D_{km} + D_{lm} - D_{kl})] . \tag{5b}$$

In formula (5a) i, j, and k represent three distinct OTUs, and in (5b) i, j, k, and l represent four distinct OTUs.

Assume that D is generated by a tree in which all branches are positive. Saitou and Nei (1987) prove the following lemma:

Lemma: If i and j are neighbors, then S_{ij} is the strictly least element in its row and column.

Proof: Each summand in formula (5a) is positive.

Theorem: If i and j are chosen so that S_{ij} is a minimum, then i and j are neighbors.

Saitou and Nei prove this theorem by induction, but they misapply the induction hypothesis. It is hard to see how to fill the resulting gap, so a different proof follows.

Proof of the theorem: The theorem is easily verified for any tree in which $N \leq 4$, so let $N \geq 5$. Suppose that i and j are not neighbors. By the lemma and the minimality of S_{ij}, neither i nor j has a neighbor. Let k and l be any pair of neighbors, so that i, j, k, and l are distinct and are represented by the tree in figure 1. Consider the sum in formula (5b), which is nonnegative. If m is a fifth OTU, then it joins the tree in figure 1 at a point x along one of the indicated arcs. Say that m is of type 1 if it joins the path from i to j at any node different from u and that m is of type 2 if it joins the path from i to j at node u.

If m is of type 1, then the corresponding summand in formula (5b) is $-2D_{ux} - 2D_{uv}$. If m is of type 2, then the corresponding summand in formula (5b) is $-4D_{ux} + 2D_{uv}$. For the sum in formula (5b) to be nonnegative, there must be at least as

many terms corresponding to OTUs m of type 2 as there are terms corresponding to OTUs m of type 1. It follows that there are more OTUs that join the path from i to j at u than there are OTUs that join that path at all other nodes combined.

Because neither i nor j has a neighbor, there must be a pair r, s of neighbors that meet the path from i to j at some node w that is different from u. By the above argument applied to w, there are more OTUs that join the path from i to j at w than there are OTUs that join that path at all other nodes combined. The conclusions about u and w contradict each other, and the theorem follows.

LITERATURE CITED

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SATTATH, S., and A. TVERSKY. 1977. Additive similarity trees. Psychometrika **42**:319–345.

WALTER M. FITCH, reviewing editor

Received February 22, 1988