Reprinted from *The American Mathematical Monthly*, Vol. 87, No. 7, Aug. - Sept. 1980.

HOW ALIKE ARE TWO TREES?

TEMPLE F. SMITH AND MICHAEL S. WATERMAN

Let τ_n be the set of binary trees with the same n labeled terminal vertices. (Such trees are connected graphs which have no cycles and with each vertex of valence one or three.) Let e be an interior edge joining vertices u and v. Deletion of e, u, and v produces four subtrees T^1, T^2, T^3, and T^4, where both T^1 and T^2 are adjacent to u and both T^3 and T^4 are adjacent to v. A **nearest neighbor interchange about** e consists in making T^1 adjacent to v and making one of T^3 or T^4 adjacent to u. See Figure 1.

Define the distance between tree $T \in \tau_n$ and $S \in \tau_n$ to be the minimum number of nearest neighbor interchanges to change tree T into tree S. An example of two trees in τ_7 with a distance of two is shown in Figure 2. Questions of interest include (i) an efficient algorithm to compute the distance and (ii) characterizing the pairs of trees in τ_n which maximize the distance. These problems have received some attention [2], but only partial results are known. Even the maximum distance attainable for two elements of τ_n is unknown.

There are many applications of such trees; the one motivating the problems here is that of representing possible evolutionary relationships of contemporary organisms. Many schemes exist for reconstructing evolutionary relationships, and they often give distinct binary trees. It is of interest, then, to compare these trees. The distance in this paper implies that we weigh speciation equally whether it occurred early or relatively recently.

Dobson [1] has given a survey of techniques to compare trees. None of the measures she gives seems to us to be satisfactory from a biological point of view, and therefore we devised the new metric on τ_n [2]. While we are satisfied with our metric, it lacks an efficient computation algorithm. For $n=6$, the metric has been analyzed by Fitch and Siegel [personal communication].

For a given interior edge e, there is an associated *partition* $\pi_e = \{A, B\}$ of the terminal vertices

into two sets. It is routine to show that the collection of all such partitions represents the tree [2].

We did devise a technique which was feasible for computation, but we could not prove that it always calculates the required distance. It searches for the least number of nearest neighbor interchanges to achieve a partition in T identical with that in S. Then the algorithm considers each "side" of the partition independently. Some results [2] given below suggested this technique. An optimal path is any sequence of nearest neighbor interchanges changing tree T into tree S which achieves the distance between T and S.

(i) If π^T is a partition in T and $\pi^S = \pi^T$ is a partition in S, then π^T will not change in any optimal path.

(ii) If an optimal path has partition π_e associated with edge e which is equal to a partition in S, then e is crossed at most once by any end vertex.

(iii) If T and S have no equal partitions but some nearest neighbor interchange in T yields an equal partition, then that nearest neighbor interchange is on some optimal path.

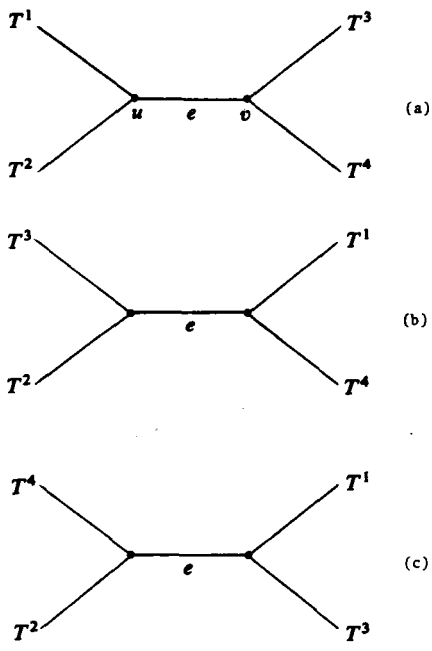

FIG. 1. Binary tree with subtrees T^1, T^2, T^3, and T^4 (a) and the two resulting trees from nearest neighbor interchanges about edge e.

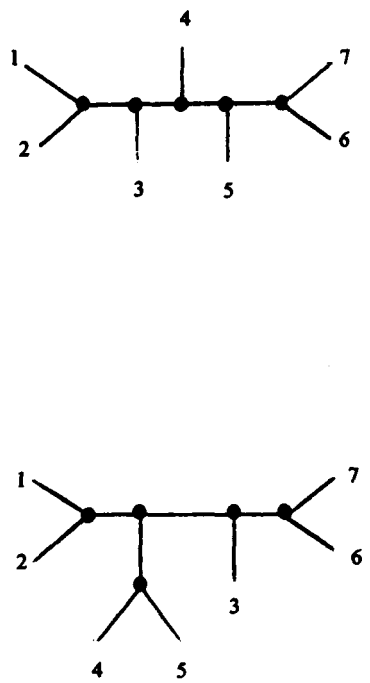

FIG. 2. Two binary trees a distance of two apart.

References

1. A. J. Dobson, Comparing the shapes of trees, in Combinatorial Mathematics 3, Lecture Notes in Mathematics, vol. 452, Springer-Verlag, New York, 1975, pp. 95–100.

2. M. S. Waterman and T. F. Smith, On the similarity of dendrograms, J. Theor. Biol., 73 (1978) 789–800.

PHYSICS DEPARTMENT, NORTHERN MICHIGAN UNIVERSITY, MARQUETTE, MI 49855.
DEPARTMENT OF MATHEMATICS, UNIVERSITY OF HAWAII, HONOLULU, HI 96822.