J. theor. Biol. (1996) **182,** 463–467

On the Nearest Neighbour Interchange Distance Between Evolutionary Trees

MING LI, JOHN TROMP† AND LOUXIN ZHANG

Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3GI, Canada

(*Received on* 20 *November* 1995, *Accepted in revised form on* 29 *May* 1996)

We present some new results on a well-known distance measure between evolutionary trees. The trees we consider are free 3-trees having n leaves labeled $0, \ldots, n-1$ (representing species), and $n-2$ internal nodes of degree 3. The distance between two trees is the minimum number of nearest neighbour interchange (NNI) operations required to transform one into the other. First, we improve an upper bound on the nni-distance between two arbitrary n-node trees from $4n \log n$ (Culik & Wood, 1982, *Inf. Pro. Letts.* **15,** 39–42) to $n \log n$. Second, we present a counterexample disproving several theorems in (Waterman & Smith, 1978, *J. theor. Biol.* **73,** 789–800). Roughly speaking, finding an equal partition between two trees does not imply decomposability of the distance finding problem. Third, we present a polynomial-time approximation algorithm that, given two trees, finds a transformation between them of length $O(\log n)$ times their distance. We also present some results of computations we performed on small size trees.

Introduction

In a free 3-tree, n leaf nodes, labeled one to n, are connected by a tree with $n-2$ internal nodes, all of degree 3. It follows that the tree has $n-3$ edges between internal nodes, the so-called internal edges. We study free 3-trees as representations of evolutionary trees, the main tool for modeling the evolutionary history of species. Much research in evolutionary genetics focuses on reconstructing the "correct" evolutionary tree for a set of species. However, the variety of methods and criteria available often lead to different evolutionary trees on the same set of species. In comparing such trees for similarity, several natural metrics have been defined. The measure we consider is derived from a simple tree transforming operation, the nearest neighbour interchange (nni), introduced independently by Robinson, (1971)/Moore *et al.* (1973). The tree on the left of Fig 1. has an internal edge (u, v) and four associated subtrees partitioned as $\{A \cup B, C \cup D\}$. An nni operation swaps two of the subtrees to create either of the trees on the right, with associated partitions $\{A \cup C, B \cup D\}$ or $\{A \cup D, C \cup B\}$.

We define the distance between two trees to be the minimal number of nni's needed to transform one into the other. This definition makes sense because the nni transformation is invertible. We can consider the collection of all 3-trees on n leaves as the vertices on a graph $G = G_n = (V, E)$, where an edge connects two 3-trees if they are one nni apart.

We summarize several facts about this graph [$\Delta(G)$ denotes the diameter of G, i.e. the maximum distance between any two trees. Also, all logs in this paper are in base 2]:

(1) $|V| = 1 \cdot 3 \cdot 5 \ldots (2n-5)$;
(2) G is regular of degree $2(n-3)$;
(3) G is connected;
(4) $(n-2)/4) \log[(2\sqrt{2}/3e)(n-2)] \leqslant \Delta(G) \leqslant n \log n + O(n)$.

† Author to whom correspondence should be addressed at: CW1, Kruislaan 413, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

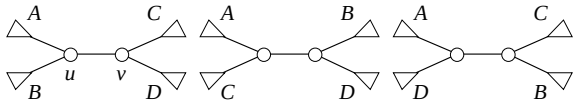

FIG. 1. The two possible nni operations on an internal edge (u, v).

The first three facts were established in Robinson (1971), which also provided an asymptotically weaker upper bound of $\frac{1}{2}(n-2)(n-3)$ on $\Delta(G)$ [as did Waterman & Smith (1978)]. The last fact will be established in the next section.

We wrote a C-program (available upon request) that uses space $|V|$ to find the distance of any tree to a given one. This was run for all possible non-isomorphic unlabeled trees to find the maximum distance, shown in Table 1 for trees up to size 11. Results for trees of size less than 11 were known previously. Computing the next value $\Delta(G_{12})$ requires 625 Mb, which is beyond our computing machinery.

Bounding $\Delta(G)$

The lower bound of $\Delta(G)$ in (4) follows from the following lemma.

LEMMA 1 The number of trees within distance m from any given tree is at most $3^{n-2}2^{4m}$.

PROOF: Analogous to Theorem 5.1 of the elegant work (Sleator *et al*. 1992), which states a $3^{2n-4}2^{3n}$ upper bound on the number of plane triangulations on n vertices within m "flips" of a given triangulation. Their result is proven by giving a short encoding of derivations on graphs. Their graphs and our trees (at least the internal nodes) share the property of being 3 regular, and their flips are nothing but a plane-oriented version of our nni. Thus, we can use the same encoding, with an extra bit per operation to specify which of the two nni's corresponding to an internal edge is used. That is why our bound uses 2^{4m} instead of 2^{3m}. The 3^{n-2} part corresponds to the initial 3-valued labels, one for each internal node of the tree. In Sleator *et al*. (1992), an initial label was needed for each of the $2n-4$ vertices in the dual graph of a triangulation.

An upper bound $\Delta(G) \leqslant 2n \log n + O(n)$ can be easily obtained by improving the analysis of Culik & Wood (1982). There, they use $2n-4$ steps (we conclude the stated $2n-6$ is a mistake) to transform any tree into one of minimal diameter $2 \log n/3$, then up to n swaps to permute its leaves, then another $2n-4$ steps to transform into arbitrary shape. But he swaps, each of which costs a number of steps equal twice the diameter, can be replaced by n moves, each only costing a number of steps equal to the diameter. By a completely different, and rather unexpected approach, we further improve the upper bound $2n \log n + O(n)$ to $n \log n + O(n)$ as in (4).

LEMMA 2 $\Delta(G) \leqslant n \log n + O(n)$.

PROOF: This lemma follows from the observation that we can simulate a merge-sort on degenerate trees. For convenience, we'll assume that the number of leaves, n is a 2-power. The i-th stage of the sorting process starts with a degenerate tree consisting of $n/2^i$ sorted blocks of 2^i leaves each, where the sorting order alternates between ascending and descending. Then, starting from the middle, adjacent blocks are merged together and pulled out, finally resulting in a degenerate tree consisting of $1/2 \cdot n/2^i$ sorted blocks of $2 \cdot 2^i$ leaves each, again alternating. The merging process is illustrated in Fig. 2.

The bound follows by noting that any tree can be converted into a degenerate one within n steps, and that our merge sort on trees uses no more than $n \log n$ nni's.

TABLE 1. Exact values for the diameter of G_n for small trees

n	3	4	5	6	7	8	9	10	11
$\Delta(G_n)$	0	1	3	5	7	10	12	15	18

Computing the nni-distance

The question of the computability of the nni distance measure, call it d, has generated much interest. As mentioned above, a brute force method can be employed which searches all (or a significant fraction of) trees in exponential time and space. In an attempt to improve efficiency, Waterman & Smith (1978) propose another distance measure, "closest partition" which they conjecture is actually equal to d. The closest partition distance $c(T, S)$ for trees sharing a partition is defined recursively as the sum of the two distances between the corresponding smaller parts resulting from splitting each tree into two. For trees T, S not sharing a partition it is defined as $k + c(R, S)$, where k is the minimum number of nni operations required to transform a tree T into a tree R that shares a partition with tree S. Note that the non-determinism in choosing R makes this measure somewhat ill-defined.

They base their conjecture on what Day (1983) aptly calls a decomposability property of nni, which can be stated as

Definition 1: $DP(d)$: $d(T_0 \cdot T_1, S_0 \cdot S_1) = d(T_0, S_0) + d(T_1, S_1)$, where $T \cdot S$ denotes the result of joining T with S at any pair of leaves (removing them and renumbering the rest in a consistent manner).

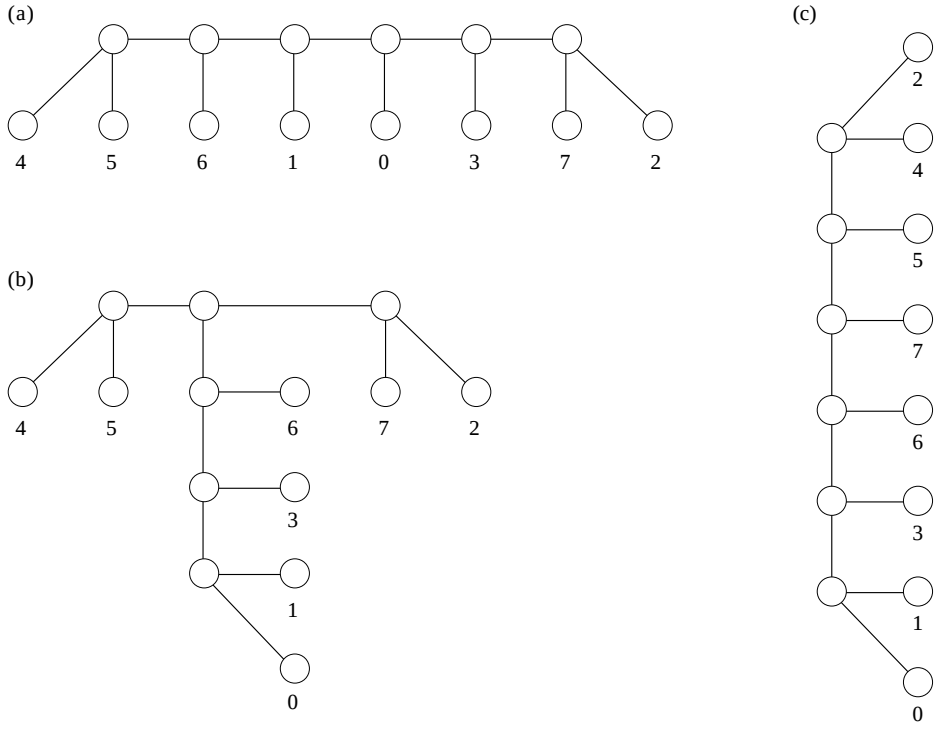

FIG. 2. Merging blocks of size two into blocks of size four.

Informally, $DP(d)$ says that if two trees can be split at some internal edge into identical subsets of leaves, then an optimal transformation of one into the other can be found in which no nni operation affects that internal edge. This claim appears in Waterman & Smith (1978) as Theorem 4. Its proof, however, appeals to their Theorem 3, which was shown invalid in Jarvis *et al*. (1983) with a six node counter-example. Consequently, Jarvis *et al*. (1983a) conclude that the status of Theorem 4 is unresolved, and observes that Theorem 5 of Waterman & Smith (1978) is a single step version of the Waterman & Smith's conjecture that c = d. This conjecture was shown to fail in Jarvis *et al*. (1983) and Boland *et al*. (1983) in a weak sense (for some choices that c allows), and shortly thereafter in Jarvis *et al*. (1983b) in a strong sense (for all choices in defining c). These papers also point out that computation of c appears to require exponential time as well, since there is no obvious bound on k in the definition of c.

In Day (1983), $DP(d)$ is presented as one of several conjectures, which are shown to be related and to hold for small trees of size up to 7.

King & Warnow (1994) showed a logarithmic gap between measures c and d. Their example is a pair of trees, each on $n = 2^k$ nodes equidistant from the central internal edge. In one tree, the leaves can be drawn in normal order, while in the other, the leaves can be draw in bit-reverse order (e.g. 0, 4, 2, 6, 1, 5, 3, 7). For this pair of trees one can show $d = \Theta(n)$, whereas $c = \Theta(n \log n)$ (in the weak sense at least).

In the opposite direction, Křivánek (1986) tried to show the NP-completeness of computing nni distance on unlabeled trees. The reduction is from the problem PARTITION, which, given a set of pairs of numbers, asks if we can select one number of each pair whose sum equals that of the unselected numbers. However, the NNI instance they produce is of size proportional to the numbers in the PARTITION instance, and hence exponential in their length. This obviously invalidates the whole reduction.

The important question of whether d is computable in polynomial time thus remains a major open problem.

Our next result resolves, for the first time, $DP(d)$, the decomposability of nni distance, serving as a counter example to all three theorems 3, 4, and 5 of Waterman & Smith (1978). This result suggests that there is no straightforward way to compute d efficiently.

First, we need to introduce some terminology. Recall that any edge in a tree partitions the set of n leaves into two subsets. An edge/partition in one tree is said to be shared with another tree if some edge in the other tree partitions the leaves into the same two

subsets. Remember that a sequence of nni-transformations is a path in graph G, and the distance between two trees is the length of the shortest such path.

LEMMA 3 There are trees T_0, T_1 sharing a partition which is not shared by any intermediate tree on a shortest path from T_0 to T_1.

PROOF: There are $n!/8$ trees whose internal nodes form a path since we can choose either end of the path, and either order of the two leaves at this end, and either order of the two leaves at the other end to produce one of $2 \times 2 \times 2 = 8$ permutations corresponding to a tree. By Lemma 1, it follows that there are two path-trees t_0, t_1 which are distance

$$d(t_0, t_1) \geqslant \frac{n}{4}\left(\log \frac{n}{3e}\right) = \Omega(n \log n)$$

apart, where $e = 2.718\cdots$, the base of the natural logarithm. Without loss of generality, we can assume t_0 to be the tree corresponding to the identity permutation.

We will construct trees T_0, T_1 on $2n-2$ leaves. Take one copy of t_0 and "cut off" leaf n. Take another copy of t_0 and increase each leaf number by $n-1$. Then cut off leaf $2n-1$ and replace it with the first copy. The result is shown on the left in Fig. 3. Tree T_1 is similarly constructed from t_1, though its central edge need not be at either end. Note that T_0 and T_1 share the central partition.

The obvious way to transform T_0 into T_1 is to transform both halves separately, at a cost of $2d(t_0, t_1)$. All intermediate trees on such a path from T_0 to T_1 share the same central partition. However, we can do better by "combining" the two halves of each tree. It takes $3n-8$ nni operations to transform T_0 into T_0', shown on the right. We repeatedly take the top leaf on the right, move it over to the left, and move down the central edge. For the similarly derived T_1', we have $3n\text{-}8 \leqslant d(T_1, T_1') \leqslant 4n-11$ (up to $n-3$ initial nni operations are needed to move the central edge to the top or bottom). But it is not hard to see

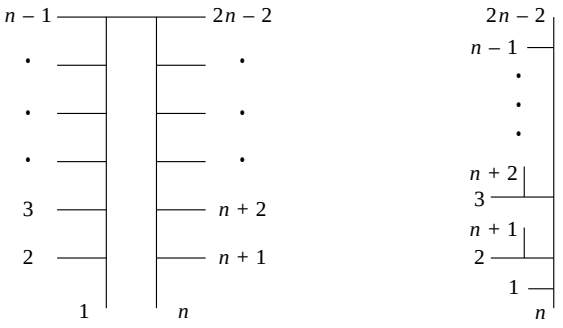

FIG. 3. T_0 and T_0'.

that $d(t_0, t_1) = d(T_0', T_1')$, the only difference being that each leaf i is replaced by a pair of leaves $\{i, i+n-1\}$.

Thus we find that $d(T_0, T_1) \leqslant d(t_0, t_1) + 7n - 19 < 2d(t_0, t_1)$, hence any shortest path must go through some trees that do not share the central partition. Consider a shortest path from T_0 to T_1 that maximizes the number of intermediate trees sharing the central partition. Then the trees on this path not sharing the central partition must form a subpath, and the trees next to either end of this subpath satisfy the requirements of the lemma.

This lemma shows that the shape of a shortest path between two trees can depend crucially on whether two subtrees are within a certain linear distance from each other, and gives the problem a sense of discontinuity. Possibly this phenomenon can be exploited to prove an NP-completeness result.

Logarithmic Approximation

Using the divide-and-conquer and balancing strategies, Brown & Day (1984) provided a heuristic algorithm for approximating the nni distance between two trees. Given two labeled trees S and T, the algorithm first divides S into at most three disjoint subtrees S_i $(i = 1, 2, 3)$ of roughly equal size and transforms T into a tree T', which has a decomposition T_i such that S_i and T_i have the same leaf set, and then transforms T_i into S_i for each i recursively. Brown & Day observed that the algorithm outputs a close approximation to the nni distance for small trees. But they did not give any approximation ratio analysis. In fact, it is not difficult to see that their algorithm suffers the same ill behavior as the decomposability property by Waterman & Smith (1978) as shown in the last section.

We now present an algorithm that approximates the nni distance within a logarithmic factor. Let us elaborate on what we mean by that. For algorithms computing the exact distance, it does not matter whether it actually produces a shortest path between them or only the length. Given the latter, one can repeatedly compute distances of neighbouring trees to trace out a shortest length path and thus get an algorithm of the former type. In case of approximation, however, such a self-reduction is not possible, and we must require an approximation algorithm to actually compute a path between two trees.

Given two trees, T_0, T_1, we first identify in T_0 those edges (partitions) that are not shared with T_1. These edges induce a subgraph of T_0 consisting of one or more components, each of which is a subtree of T_0. For example, the two trees in Fig. 4 give two

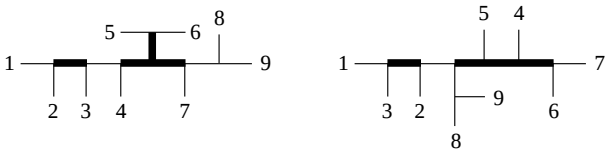

FIG. 4. Two trees with non-shared edges shown in thick.

components of non-shared edges, one single edge, and one triple edge component. Each non-shared-edge component links up the same set of neighbouring shared-edge-components in T_0 and T_1, but it does so in different ways.

Our algorithm transforms T_0 into T_1 by transforming each non-shared edge component separately. Consider a component consisting of k non-shared edges in T_0. This links up $k+3$ shared-edge-components, which we can consider as leaves for the purpose of linking them up differently. So we want to transform C_0 into C_1, where C_i is the $(k+3)$-tree corresponding to the component in T_i. By Lemma 2, the distance between C_0 and C_1 is at most $(k+3)\log(k+3)+O(k)$. On the other hand, it is clear that any transformation from T_0 into T_1 must use at least one nni operation on every non-shared edge.

The approximation factor of this algorithm is at most

$$\frac{\Sigma(k+3)\log(k+3)+O(k)}{\Sigma k} \leqslant \frac{n\log n+O(n)}{n-3}$$

since Σk is at most the number of internal edges, which is $n-3$.

LEMMA 4 Nni distance can be polynomial time approximated within a factor of $\log n + O(1)$.

Conclusion

The problem of efficiently computing the nni distance is surprisingly subtle given the history of a disproved conjecture, faulty NP-completeness proof, and our new result that invalidates some quite old and intuitively appealing theorems. It remains unclear whether the problem is NP-complete or not, either in the labeled or in the unlabeled case. Also, the question of whether nni distance can be approximated within a constant factor is still open. In practice, the nni distance can be found exactly if the "non-shared-edge" components are of size at most 11 (taking about 33 Mb), since the phenomenon of Lemma 3 can only occur on impractically large trees.

We thank Valerie King and the referees for their comments and pointers to the literature. M. Li was supported in part by the NSERC Operating Grant OGP0046506, ITRC, a CGAT grant and DIMACS; J. Tromp was supported by an NSERC International Fellowship; and L. Zhang was supported by a CGAT grant.

REFERENCES

BOLAND, R. P., BROWN, E. K. & DAY, W. H. E. (1983). Approximating minimum-length-sequence metrics: a cautionary note. *Math. Soc. Sci.* **4,** 261–270.

BROWN, E. K. & DAY, W. H. E. (1984) A computationally efficient approximation to the nearest neighbour interchange metric, *J. Classification* **1,** 93–124.

CULIK II, K. & WOOD, D. (1982). A note on some tree similarity measures, *Inf. Proc. Letts.* **15,** 39–42.

DAY, W. H. E. (1983). Properties of the Nearest Neighbour Interchange Metric for Trees of Small Size *J. theor. Biol.* **101,** 275–288.

GAREY, M. R. & JOHNSON, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

JARVIS, J. P., LUEDEMAN, J. K. & SHIER, D. R. (1983a). Counter-examples in measuring the distance between binary trees, *Math. Soc. Sci.* **4,** 271–274.

JARVIS, J. P., LUEDEMAN, J. K. & SHIER, D. R. (1983) Comments on computing the similarity of binary trees, *J. theor. Biol.* **100,** 427–433.

KŘIVÁNEK, M. (1986). Computing the Nearest Neighbour Interchange Metric for Unlabeled Binary Trees is NP-Complete, *J. Class.* **3,** 55–60.

KING, V. & WARNOW, T. Nov. 4 (1994). On Measuring the nni Distance Between Two Evolutionary Trees, *DIMACS mini workshop on combinatorial structures in molecular biology*, Rutgers University.

MOORE, G. W., GOODMAN, M. & BARNABAS, J. (1973). An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets, *J. theor. Biol.* **38,** 423–457.

PAPADIMITRIOU, C. H. & YANNAKAKIS, M. (1991). Optimization, approximation, and complexity classes, *J. Comp. Syst. Sci.* **43,** 425–440.

ROBINSON, D. F. (1971). Comparison of Labeled Trees with Valency Three, *J. Combin. Theor.* **11,** 105–119.

SLEATOR, D., TARJAN, R. & THURSTON, W. (1992). Short encodings of evolving structures, *SIAM J. Discrete Math.* **5,** 428–450.

WATERMAN, M. S. & SMITH, T. F. (1978). On the Similarity of Dendrograms, *J. theor. Biol.* **73,** 789–800.