Some Notes on the Nearest Neighbour Interchange Distance

Ming Li*
University of Waterloo

John Tromp**
University of Waterloo and CWI

Louxin Zhang***
University of Waterloo

Abstract. 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 $4n \log n$ [2] to $n \log n$. Second, we present a counterexample disproving several theorems in [13]. Roughly speaking, finding an equal partition between two trees doesn't 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.

1 Introduction

In a *free 3-tree*, n leaf nodes, labeled 1 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

* Supported in part by the NSERC Operating Grant OGP0046506, ITRC, a CGAT grant and DIMACS. Address: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: mli@math.uwaterloo.ca

** Supported by an NSERC International Fellowship. Address: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: tromp@daisy.uwaterloo.ca

*** Supported by a CGAT grant. Address: Deptartment of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: lzhang@daisy.uwaterloo.ca

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 [11] and [9]. The tree on the left of Figure 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 $\{A \cup D, C \cup B\}$.

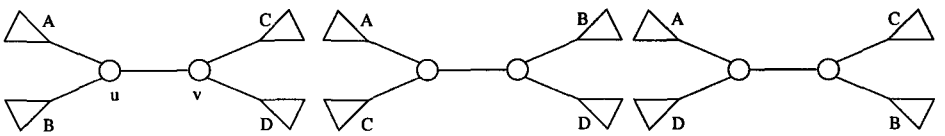

Fig. 1. The two possible nni operations on an internal edge (u, v).

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 iff 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. $|V| = 1 \cdot 3 \cdot 5 \cdots (2n-5)$
2. G is regular of degree $2(n-3)$
3. G is connected
4. $\frac{n-2}{4} \log(\frac{2\sqrt{2}}{3e}(n-2)) \leq \Delta(G) \leq 2n \log n + (4 - \log 3)n - 8$
5. $\Delta(G) \leq n \log n + O(n)$

The first three facts were established in [11], which also provided an asymptotically weaker upper bound of $\frac{1}{2}(n-2)(n-3)$ on $\Delta(G)$ (as did [13] independently). The last two facts 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.