

A Linear-Time Approximation for Rotation Distance

Sean Cleary^{1,3} *Katherine St. John*^{2,3}

¹Department of Mathematics

City College of New York, City University of New York, New York, NY 10031

²Department of Mathematics & Computer Science

Lehman College & the Graduate Center, City University of New York, Bronx,
NY 10468

³The Graduate Center

City University of New York, 365 Fifth Avenue, New York, NY 10016

Abstract

Rotation distance between rooted binary trees measures the number of simple operations it takes to transform one tree into another. There are no known polynomial-time algorithms for computing rotation distance. In this short note, we give an efficient, linear-time approximation algorithm, which estimates the rotation distance, within a provable factor of 2, between ordered rooted binary trees.

| | | | |
|--------------------------------|--------------------------|-------------------------------|--------------------------|
| Submitted: July 2009 | Reviewed: August 2009 | Revised: August 2010 | Accepted: August 2010 |
| | Final: August 2010 | Published: November 2010 | |
| Article type: Concise paper | | Communicated by: T. Warnow | |

Research supported by Grants NSF #0811002 to Cleary and NSF #0513660 and #0920920 to St. John.

E-mail addresses: cleary@sci.ccnycuny.edu (Sean Cleary) stjohn@lehman.cuny.edu (Katherine St. John)

1 Introduction

Binary search trees are a fundamental data structure for storing and retrieving information [4]. Roughly, a binary search tree is a rooted binary tree where the nodes are ordered “left to right.” The potential efficiency of storing and retrieving information in binary search trees depends on their height and balance. Rotations provide a simple mechanism for “balancing” binary search trees while preserving their underlying order (see Figure 1). There has been a great deal of work on estimating, bounding, and computing rotation distances. By rotating to right caterpillar trees, Culik and Wood [5] gave an immediate upper bound of $2n - 2$ for the distance between two trees with n interior nodes. In elegant work using methods of hyperbolic volume, Sleator, Tarjan, and Thurston [12] showed not only that $2n - 6$ is an upper bound for $n \geq 11$, but furthermore that for all very large n , that bound is realized. In remarkable recent work, Dehornoy [7] gave concrete examples illustrating that the lower bound is at least $2n - O(\sqrt{n})$ for all n . There are no known polynomial-time algorithms for computing rotation distance, though there are polynomial-time estimation algorithms of Pallo [10], Pallo and Baril [1], and Rogers [11]. Baril and Pallo [1] use computational experimental evidence to show that a large fraction of their estimates are within a factor of 2 of the rotation distance. The problem has been recently shown to be fixed-parameter tractable in the parameter, k , the distance [3]. Li and Zhang [9] give a polynomial time approximation algorithm with approximation ratio depending on the shape of the tree, but strictly less than 2.¹ While Li and Zhang give an outline of the algorithm and do not analyze the time complexity, it can be easily shown that the running time is bounded by $O(n^3)$ where n is the number of leaves in the original trees and adjacency matrices are used to store the triangulated polygons used in the algorithm.

In this short note, we give a linear time approximation algorithm with an approximation ratio of 2, improving the running time at the very modest expense of the approximation ratio. This is accomplished by showing the distance between the trees is bounded below by $n - e - 1$ and above by $2(n - e - 1)$ where n is the number of internal nodes and e is the number of edges in common in the reduced trees. The number of common edges is equivalent to Robinson-Foulds distance, widely used in phylogenetic settings, which Day [6] calculates in linear time.

2 Background

We consider ordered, rooted binary trees with n interior nodes and where each interior node has two children. Such trees are commonly called *extended binary trees* [8]. In the following, *tree* refers to such a tree with an ordering on the leaves, *node* refers to an interior node, and *leaf* refers to a non-interior node.

¹Their exact ratio is stated for the equivalent diagonal flip distance for triangulated polygons. It is bounded by the maximum number of diagonals, d , allowed at any vertex of the polygons, and is $2 - \frac{2}{4(d-1)(d+6)+1}$.

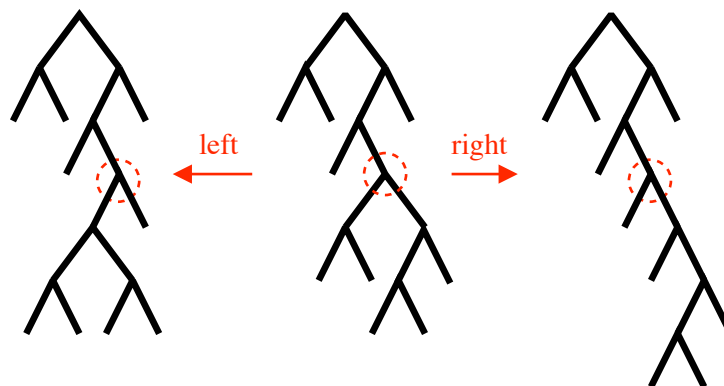


Figure 1: A (right) rotation at a node consists of rotating the right child of the left child of the node to the right child of the node. A left rotation is defined similarly by moving the left child of the right child of the node to the left child of the node. The circled node in the middle tree has been rotated right to yield the tree on the right, and similarly rotated left to yield the tree on the left.

Our trees will have $n + 1$ leaves numbered in left-to-right order from 1 to $n + 1$. The size of a tree will be the number of internal nodes it contains. Each internal edge in a tree separates the leaves into two connected sets upon removal, and a pair of edges e_1 in S and e_2 in T form a *common edge pair* if their removal in their respective trees, S and T , gives the same partitions on the leaves. In that case, we say that S and T have a *common edge*.

Right rotation at a node of a rooted binary tree is defined as a simple change to T as in Figure 1, taking the middle tree to the right-hand one. Left rotation at a node is the natural inverse operation. The *rotation distance* $d_R(S, T)$ between two rooted binary trees S and T with the same number of leaves is the minimum number of rotations needed to transform S to T .

The specific instance of the rotation distance problem we address is:

ROTATION DISTANCE:

INPUT: Two rooted ordered trees, S and T on n internal nodes,

QUESTION: Calculate the rotation distance between them, $d_R(S, T)$.

Finding a sequence of rotations which accomplish the transformation gives only an upper bound. The general difficulty of computing rotation distance comes from the lower bound.

3 Approximation Algorithm

The number of edges that differ between the two trees provides a useful bound on their rotation distance. We first give sharp bounds from which the approximation result follows easily.

Theorem 1 *Let S and T be two distinct ordered rooted trees with the same number of leaves. Let n be the number of internal nodes and e the number of common edges for S and T . Then,*

$$n - e - 1 \leq d_R(S, T) \leq 2(n - e - 1)$$

Proof: The lower bound follows from two simple observations. First, if we use a single rotation to transform T_1 to T_2 , all but one of the internal edges in each tree is common with the other tree. Second, every internal edge of S that is not common with an internal edge of T needs a rotation (possibly more than one) to transform it to an edge in common in T . The number of internal edges occurring only in S is $n - e - 1$ and, thus, is also a simple lower bound.

For the upper bound, we use two facts from past work on rotation distance. We first let $(S_1, T_1), (S_2, T_2), \dots, (S_{e+1}, T_{e+1})$ be the resulting tree pairs from removing the e edges S and T have in common, where we insert placeholder leaves to preserve the extended binary tree property. Let n_i be the size of tree S_i for $i = 1, 2, \dots, e + 1$. The first is the observation of Sleator *et al.* [12] used before: the rotation distance of the original tree pair (S, T) with a common edge is the sum of the rotation distances of the two tree pairs “above” and “below” the common edge. Extending this to e edges in common between S and T , we have

$$d_R(S, T) = \sum_{i=1}^{e+1} d(S_i, T_i) \leq \sum_{i=1}^{e+1} 2n_i - 2 = 2n - 2(e + 1) = 2(n - e - 1)$$

The inequality follows from the initial bound of $2n - 2$ on rotation distance between trees with n internal nodes of Culik and Wood [5].

Thus, $n - e - 1 \leq d_R(S, T) \leq 2(n - e - 1)$. \square

We note that using the sharper bound of $2n - 6$ for $n > 12$ from Sleator, Tarjan and Thurston [12] together with the table of distances for $n \leq 12$ can improve this further.

These reduction rules and counting the number of common edges can be carried out in linear-time [2, 6]. The efficiency comes from representing internal nodes in constant (versus linear) space: namely, in an ordered tree, the left-most and right-most children of any node uniquely identifies the node. This yields the corollary:

Corollary 1 *Let S and T be ordered rooted trees with n internal nodes. A 2-approximation of their rotation distance can be calculated in linear time.*

Proof: Let S and T be two distinct ordered rooted n -leaf trees. Let n be the number of internal nodes and e the number of edges in common for S and T . Then, by Theorem 1, $n - e - 1 \leq d_R(S, T) \leq 2(n - e - 1)$. Since this is within a linear factor 2 from both bounds, we have the desired approximation. \square

We note that this algorithm is easily adapted to give a sequence of rotations which realize the upper bound of the approximation, again in linear time. The specific problem is as follows:

ROTATION DISTANCE REALIZATION:

INPUT: Two rooted ordered trees, S and T on n internal nodes,OUTPUT: A sequence of rotations which converts S to T .

The approximation algorithm uses the Culik-Wood bound on potentially several pieces. On each piece, the $2n-2$ bound comes from rotating each internal node which is not on the right side of the tree to obtain a right caterpillar, and then rotating the caterpillar to obtain the desired tree. This can be accomplished simply in linear time.

The algorithm is as follows:

Step 1: Find the e common edges in S and T using the linear-time algorithms of Day [6] as described in Bonnet *et al.* [2]. Each common edge disconnects S and T , resulting in two families of trees $\{S_i\}$ and $\{T_i\}$, with the sizes $|S_i| = |T_i|$, and with $d(S, T) = \sum d(S_i, T_i)$.

Step 2: For each subtree pair (S_i, T_i) , use the method of Culik and Wood [5] to rotate from S_i to T_i through the all-right tree of the appropriate size. That is, for each non-right internal edge in S_i , rotate to reduce the number of non-right internal edges by one until the all-right tree is obtained. Perform the corresponding steps in reverse to obtain T_i from the all-right tree.

References

- [1] J.-L. Baril and J.-M. Pallo. Efficient lower and upper bounds of the diagonal-flip distance between triangulations. *Information Processing Letters*, 100(4):131–136, 2006.
- [2] M. L. Bonet, K. St. John, R. Mahindru, and N. Amenta. Approximating subtree distances between phylogenies. *Journal of Computational Biology*, 13(8):1419–1434 (electronic), 2006.
- [3] S. Cleary and K. S. John. Rotation distance is fixed parameter tractable. *Information Processing Letters*, 109:918–922, 2009.
- [4] T. Corman, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. McGraw-Hill, 1990.
- [5] K. Culik II and D. Wood. A note on some tree similarity measures. *Information Processing Letters*, 15(1):39–42, 1982.
- [6] W. H. E. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2:7–28, 1985.
- [7] P. Dehornoy. On the rotation distance between binary trees. Preprint, arXiv:math.CO/0901.2557.
- [8] D. E. Knuth. *The Art of Computer Programming. Volume 3*. Addison-Wesley, Reading, Mass, 1973. Sorting and searching.
- [9] M. Li and L. Zhang. Better approximation of diagonal-flip transformation and rotation transformation. In *COCOON '98: Proceedings of the 4th Annual International Conference on Computing and Combinatorics*, pages 85–94, London, UK, 1998. Springer-Verlag.
- [10] J. Pallo. An efficient upper bound of the rotation distance of binary trees. *Information Processing Letters*, 73(3-4):87–92, 2000.
- [11] R. Rogers. On finding shortest paths in the rotation graph of binary trees. In *Proceedings of the Southeastern International Conference on Combinatorics, Graph Theory, and Computing*, volume 137, pages 77–95, 1999.
- [12] D. D. Sleator, R. E. Tarjan, and W. P. Thurston. Rotation distance, triangulations, and hyperbolic geometry. *Journal of the American Mathematical Society*, 1(3):647–681, 1988.