Computing the Quartet Distance Between Evolutionary Trees

David Bryant * John Tsang† Paul Kearney‡ Ming Li§

1 Introduction

The comparison of evolutionary trees is a fundamental problem in evolutionary biology. Different evolutionary hypotheses (or conflicting phylogenies) arise when different phylogenetic reconstruction methods are applied to the same data set, or when a single method is applied to different data sets (e.g. different genes). Several similarity metrics between evolutionary trees are currently in use [2]. In this paper, we study the quartet metric, which is based on common subtrees induced by four leaves. This metric has several attractive properties, though its use has been limited by the time required to compute the distance [7]. In this paper, we address this problem by describing an $O(n^2)$ algorithm that computes the quartet distance between two evolutionary trees.

Two general approaches are currently in use to resolve conflicting phylogenies. One method is to select a consensus tree (or trees) that best represents the information provided by each tree. The maximum agreement subtree (MAST) is an instance of this approach. A substantial amount of effort has been devoted to efficient algorithms for finding the MAST of two or many evolutionary trees, see [8] for a summary of results. A more quantitative approach is to define a similarity metric between trees to assess the stability of a solution by measuring the degree of similarity among the trees. The distributions of various tree similarity metrics are well-studied [7] and are very useful in testing statistical hypotheses.

An evolutionary tree represents the direction of evolution by the location of its root, the rate of evolutionary by its edge lengths and the history of speciation events by its branching pattern or topology. Biologists are often interested in the distance between two evolutionary trees independent of the direction and rate of evolution, which gives an indication of how similar two trees are in terms of the relationships among leaves. Various metrics have been proposed to measure the similarity based on the undirected tree topology. The symmetric difference metric (SM) [5], the nearest-neighbour interchange (NNI) metric [9], the subtree transfer distance (ST) [1], and the Robinson and Foulds metric (RF) [6] are examples of such measures. We study the quartet metric [4] in this paper.

For the duration of this paper let evolutionary trees be synonymous with degree-3 trees with leaves uniquely labeled by elements from a label set S where $|S| = n$. An unrooted (undirected) evolutionary tree induces a topology on any four labels from S, which we called a *quartet topology* (see Figure 1). Given two trees, the *quartet distance* between them is the number of quartet topology differences. It is well-known that the complete set of quartet topologies is unique for a given tree and the tree can be uniquely recovered from its set of quartet topologies in polynomial time [3]. More importantly, the quartet metric does not suffer from drawbacks of other distance metrics. For instance, metrics that are based on transformation operations, such as NNI, ST and RF, do not distinguish between rearrangements that affect the relationships between many leaves and rearrangements that affect only a few. In addition, metrics that are based on the number of split differences (e.g. SM) are unstable with respect to the placement of a few leaves. That is, they can make two highly similar trees very distant. But the quartet metric is more stable especially when n is large. Furthermore, the quartet metric has a far greater range than SM, and hence greater sensitivity [7].

The quartet distance between two trees can be easily obtained by comparing the quartets one by one. This takes $O(n^4)$ time as there are $\binom{n}{4}$ quartets. To our knowledge, the best existing result is an algorithm that runs in $O(n^3)$ time [7]. Our contribution is a simple algorithm that runs in $O(n^2)$ time. The algorithm can also return implicitly the set of quartet topologies shared by two trees.

For simplicity, let the input to the algorithm be two fully-resolved unrooted evolutionary trees T_1 and T_2 labeled by S. The algorithm can be easily extended to handle partially-resolved trees. Below is a brief overview of the algorithm. Further details

*Supported in part by a Bioinformatics Postdoc. Fellowship from the CIAR, Evolutionary Biology Program and by NSERC and CGAT grants to D. Sankoff. Address: CRM Université de Montréal. E-mail: bryant@CRM.UMontreal.CA

†Supported in part by NSERC. Address: Dept. of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. Email: jsctsang@math.uwaterloo.ca

‡Supported in part by a CITO grant and NSERC Research Grant 160321. Address: Dept. of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. E-mail: pkearney@math.uwaterloo.ca

§Supported in part by NSERC Research Grant OGP0046506, CITO, a CGAT grant, and the Steacie Fellowship. Address: Dept. of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. E-mail: mli@math.uwaterloo.ca

with proofs will be included in the full paper.

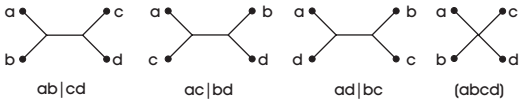

Figure 1: The possible quartet topologies labeled by $\{a, b, c, d\}$

2 The Algorithm

The algorithm was motivated by the following observation. An internal edge e of the tree partitions the leaf labels into two disjoint sets $A, B \subseteq S$ such that $S = A \cup B$. For any two labels a_i,a_j from A and b_i,b_j from B, we have the quartet topology $a_i a_j | b_i b_j$ and we say the quartet topology is *induced* by e. This association of quartet topologies to internal edges gives us a simple framework to count common quartets. We only need to consider the $O(n^2)$ internal edge pairings between T_1 and T_2. However, a quartet topology can be induced by more than one edge. To avoid double counting, we perform pre-processing on the input trees. In the pre-processing stage, each internal edge claims as many induced quartet topologies as possible as long as the quartets it claimed have not been claimed by any neighbouring edges. The quartet topologies claimed by each edge can be encoded by a constant number of sets. Hence, we can determine the common quartet topologies claimed by two edges by computing the sizes of certain set intersections. The set intersection operation can be done in constant time if we pre-compute all possible set intersections. This can be done in $O(n^2)$ time as follows. Given an evolutionary tree T, a *rooted subtree* of T given the directed edge (u, v) is the subtree of $T - \{(u, v)\}$ rooted at vertex v (see Figure 2). There are $O(n)$ such rooted subtrees for each input tree. The set intersection problem reduces to computing the common leaves for each of the $O(n^2)$ rooted subtree pairings (one from each input tree). We can process each pairing in constant time since we can first compute the pairings that involve their children. It follows that the sizes of all set intersections (also the intersections themselves) can be found in $O(n^2)$ time. Summing up the number of common quartet topologies between each pair of internal edges, one from each tree, gives the total number of agreed quartet topologies. This runs in $O(n^2)$ time since there are $O(n^2)$ internal edge pairings.

THEOREM 2.1. *Given two unrooted evolutionary trees T_1 and T_2, the number of quartet topologies shared by T_1 and T_2 can be determined in $O(n^2)$ time.*

REMARK 2.1. *The set of quartet topologies shared by T_1 and T_2 can also be determined in $O(n^2)$ time by the same algorithm.*

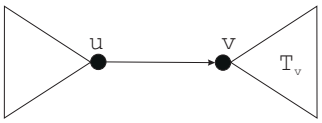

Figure 2: An example of a rooted subtree of T. T_v is obtained by removing the edge (u, v) and root it at vertex v

3 Acknowledgments

We thank I. Munro, N. Nishimura and H. Zhang for helpful discussions.

References

[1] B.L. Allen and S. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Research Report 170, Dept. of Math., University of Canterbury.*

[2] D. Bryant. *Building trees, hunting for trees, and comparing trees.* PhD thesis, University of Canterbury, 1997.

[3] P. Buneman. The recovery of trees from measures of dissimilarity. In Hodson, Kendall, and Tautu, editors, *Mathematics in the Archaeological and Historical Sciences.* Edinburgh University Press, 1971.

[4] G. Estabrook, F. McMorris, and C. Meacham. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst. Zool.*, 34(2):193–200, 1985.

[5] D. Robinson and L. Foulds. Comparison of weighted labelled trees. In *Lecture notes in mathematics*, pages 119–126. Springer-Verlag,Germany, 1979.

[6] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

[7] M. Steel and D. Penny. Distributions of tree comparison metrics–some new results. *Syst. Biol.*, 42(2):126–141, 1993.

[8] W.K. Sung. *Fast labeled tree comparison via better matching algorithms.* PhD thesis, University of Hong Kong, 1998.

[9] M Waterman and T Smith. On the similarity of dendrograms. *J. Theoret. Biol*, 73:789–800, 1978.