Available online at www.sciencedirect.com

ScienceDirect

Information Processing Letters ••• (••••) •••–•••

Information Processing Letters

www.elsevier.com/locate/ipl

Approximating geodesic tree distance

Nina Amenta [a], Matthew Godwin [a,1], Nicolay Postarnakevich [a], Katherine St. John [b,c,*]

[a] Computer Science Department, University of California, Davis, 2063 Engineering II, One Sheilds Ave, Davis, CA 95616, USA
[b] Department of Mathematics & Computer Science, Lehman College–City University of New York, Bronx, NY 12581, USA
[c] Computer Science Department, CUNY Graduate Center, New York, NY 10016, USA

Received 10 January 2006

Communicated by F. Dehne

Abstract

Billera, Holmes, and Vogtmann introduced an intriguing new phylogenetic tree metric for weighted trees with useful properties related to statistical analysis. However, the best known algorithm for calculating this distance is exponential in the number of leaves of the trees compared. We point out that lower and upper bounds for this distance, which can be calculated in linear time, can differ by at most a multiplicative factor of $\sqrt{2}$.

Keywords: Analysis of algorithms; Phylogeny; Tree metric

1. Introduction

The evolutionary history of a set of species is fundamental to understanding the structural and functional similarity between species. This history is often represented by rooted trees, with the leaves labeled by extant (living) species. Evolutionary trees are regularly used to organize genetic information, for alignment, annotation, structure and function prediction [9]. Yet these evolutionary relationships are themselves estimates. Optimal or near-optimal trees are found via programs implementing sophisticated heuristic search strategies, such as MrBayes [10], for maximum likelihood analysis, and such as those in PAUP [13] and Ratchet [11], for parsimony. These methods generally output not a single "best" tree but a large family of equally, or almost equally, optimal trees. With current computational power, it is possible to calculate hundreds to thousands of possible evolutionary trees, or phylogenies, from input genomic sequence data from a set of species.

With these huge sets of trees comes the question of how to efficiently compare the tree in a biological relevant way; in particular, how do we find the distance between two trees, that is, what should be our *tree distance metric*? Usually a large set of trees is summarized using a consensus tree, a kind of representative average, which can be defined in various ways, depending on the metric. Also, the clustering and visualization analyses for tree distributions which we have been exploring [2,3,12], is

[*] Corresponding author at: Department of Mathematics & Computer Science, Lehman College–City University of New York, Bronx, NY 12581, USA.

E-mail addresses: amenta@cs.ucdavis.edu (N. Amenta), npost@ucdavis.edu (N. Postarnakevich), stjohn@lehman.cuny.edu (K.S. John).

[1] The second author did this work as a student at U.C. Davis, until his accidental death in March of 2005.

0020-0190/$ – see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.ipl.2007.02.008

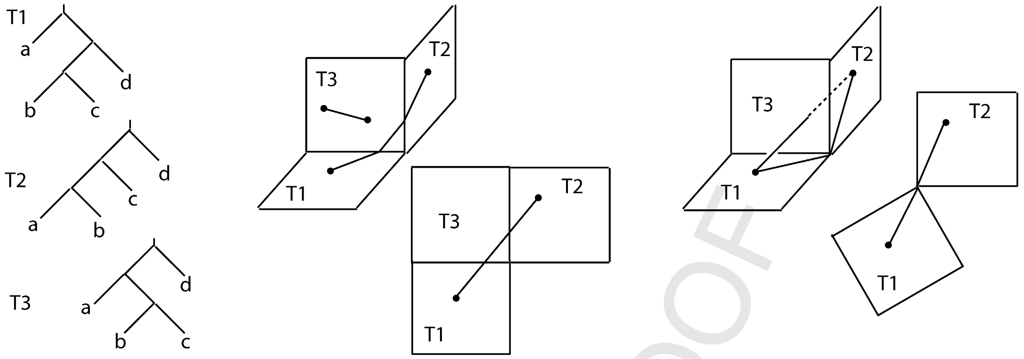

Fig. 1. On the left, three tree topologies; the lengths of the two non-terminal edges in each tree form the coordinates of a planar part of tree-space. In the middle, the shortest path between two trees with topology T_3 is a line segment, while between two trees with topologies T_1 and T_2 the shortest path includes some topology changes. The relevant parts of tree-space can be unfolded to straighten the path. On the right, the lower bound path is not constrained to lie in the tree-space. The upper bound path is constrained to go through the parts corresponding to T_1 and T_2 and the parts that they share. Again, the relevant parts of tree-space can be unfolded to straighten the path.

based on a distance metric on trees. We discuss the most popular tree metrics in Section 2.

The mathematicians Billera, Holmes and Vogtmann have recently proposed a new tree distance metric [4], which we shall call *geodesic distance*. Geodesic distance resembles the usual Euclidean distance metric used in geometry and statistics in an important way that the other existing metrics do not: there is a unique shortest path between two trees. The hope is that geodesic distance can be used to adapt statistical techniques from Euclidean geometry to compute better consensus trees or more generally to understand relationships between possible trees. In particular, the centroid of a set of trees seems like a logical choice for a consensus tree, and cetroids are well-defined with geodesic distance, but not for metrics which allow many shortest paths.

The drawback of geodesic distance is that it is not obviously easy to compute. However, in this note we observe that it is easy to approximate: we give simple upper and lower bounds which differ by a multiplicative factor of at most $\sqrt{2}$.

2. Tree metrics

This section includes a brief overview of distance metrics for phylogenetic trees. For a more detailed overview, see Hillis et al. [9] or Bryant [6]. Commonly used tree distances are Nearest Neighbor Interchange (NNI), Tree-Bisection-Reconnection (TBR), Subtree-Pruning-Regrafting (SPR), and Robinson-Foulds (RF). The first three metrics define the distance between unweighted trees and are defined as the minimal number of moves required to transform one tree into the other. The three metrics differ in how the moves are defined; for instance, in NNI, rotations are used. NNI [7] and TBR [1], as well as SPR for rooted trees [5] have been shown to be NP-hard, and SPR for unrooted trees is conjectured to be so also [1].

Since the other common metrics are intractable to compute, RF is often the measure of choice in practice. RF distance and geodesic distance are closely related, and both can be defined for either rooted or unrooted trees. From here on, we will consider rooted trees with n leaves.

We can define RF distance in terms of an edge-based representation for the trees. Each edge of a tree separates a subtree from the root, and this subtree contains a subset of the leaves; thus, we associate each edge with a subset of the n leaves. Consider an arbitrary ordering on the possible subsets of the n leaves. For each tree, we represent it as a point in which the ith coordinate is one if and only if the ith subset occurs in a subtree below an edge in the tree, and zero otherwise. For example, in Fig. 1 the leaf set is $\{a, b, c, d\}$, so, we can order the possible subsets as: $(\emptyset, a, b, c, d, ab, ac, ad, bc, bd, cd, abc, abd, acd, bcd)$ (omitting set bracket notation to improve readability). The tree T_1 in Fig. 1 contains edges corresponding to the subsets $a, b, c, d, bc, bcd, abcd$. If the edges are unweighted, we represent the tree with the point (0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1).

The Robinson–Foulds distance is just the L_1 norm on these points ($\sum_i |x_i - y_i|$, where x_i, y_i is the ith coordinate of x and y, respectively). Equivalently, the RF distance is the number of subsets which appear in only one of the trees, and not the other. For instance, the RF distance between trees T_1 and T_3 in Fig. 1 is 2. The RF distance is often normalized by dividing by n.

RF distance is not computed using this representation of trees as points, which would require time $O(2^n)$.

Please cite this article in press as: N. Amenta et al., Approximating geodesic tree distance, Information Processing Letters (2007), doi:10.1016/j.ipl.2007.02.008

Instead, Day [8] gives a linear-time algorithm for calculating RF distance.

3. Weighted edges, treespace and geodesic distance

The discussion up until this point has considered trees with unweighted edges. If there are weights associated with the edges, we can replace the non-zero elements in the representation above with the edge weights. This set as an embedding of the set of tress, or *tree-space*, into Euclidean space of dimension 2^n. Since a tree with n leaves has at most $2n-2$ edges, all but at most $2n-2$ coordinates of any point in the embedded tree-space are zero, although not all points with $2n-2$ non-zero coordinates correspond to trees.

Every fully-resolved tree topology with $2n-2$ edges is represented by a $(2n-2)$-dimensional affine subspace, parameterized by the weights of the $2n-2$ edges. Consider shrinking some subset of the edges in a fully-resolved tree to length zero, so that the tree develops polytomies (nodes of degree greater than three). This produces a tree with fewer edges, corresponding to a lower dimensional affine subspace. This lower-dimensional subspace on the boundary of several of the $(2n-2)$-dimensional subspaces corresponding to the different possible expansions of the polytomy. These shared subspaces connect the $(2n-2)$-dimensional spaces, forming an overal connected space.

Defining the distance between two trees with weighted edges provides a metric for this space. One way is to extend the notion of RF distance, defining the *weighted RF distance* to be the L_1 norm applied to the point representation of the trees. For most pairs of trees, there are an infinite number of "Manhattan" shortest paths, in which only one point coordinate changes at a time, which realize the L_1 distance. We could also consider the standard Euclidean distance between points in 2^n-dimensional space. In this case there is only one shortest path between two trees, but unless the two trees share the same topology, the path is not contained in the embedded tree-space. Since the Euclidean shortest path is the shortest path in the ambient space, its length is a lower bound on the length of any path connecting the two trees which is contained in tree-space.

Billera et al. [4] define the *geodesic distance* between two trees to be the length of the shortest path between the two trees which lies entirely within tree-space. When two fully-resolved trees T_1, T_2 have the same topology (i.e., they only differ in the weights assigned to their edges), the two corresponding points lie in an affine subspace of dimension $2n-2$ and the unique shortest path between them is simply a straight line; see middle picture in Fig. 1.

When the trees T_1 and T_2 have different topologies (e.g., right picture in Fig. 1), the shortest path between the corresponding points cannot lie in a single affine subspace, since it must include topology changes. If there are many possible ways to order the topology changes, finding the shortest path may be difficult. Billera et al. showed, using powerful techniques from mathematical topology, that the space with this metric is CAT(0), which implies that the shortest path is unique, and a geodesic. We call this *geodesic distance*. Because it gives unique geodesics, geodesic distance seems interesting as a potential tool for the statistical analysis of problems related to phylogenetic trees.

4. Bounds on geodesic distance

We compare two intuitive bounds on the geodesic distance. As mentioned above, the Euclidean distance between T_1 and T_2 is a lower bound on the geodesic distance. So we have

$$D_{lo}(T_1, T_2) = \sqrt{\sum_e \delta(e)},$$

where $\delta(e) = (w_1(e) - w_2(e))^2$ and $w_i(e)$ is the weight of edge e in tree T_i, and if e is not an edge of T_i, $w_i(e) = 0$. The unique shortest path corresponding to this distance metric does lie in tree-space, except when T_1, T_2 have the same topology.

The upper bound is given by the length of a particular path in tree-space. This path goes directly from T_1 to a *strict consensus tree* S of T_1 and T_2, and then to T_2. The strict consensus tree S is defined to be the tree containing only those edges that occur in both input trees. Strict consensus trees are usually used to summarize a set of phylogenetic trees. For that purpose, the weights on the edges can be found by averaging. Our goal, however, is to choose the edge weights so as to make the total path length as short as possible.

Let S_1 be the tree formed by shrinking all the edges of $T_1 - T_2$ to length zero, so that the lengths of the edges shared by T_1 and T_2 do not change at all. Define S_2 analogously, shrinking the edges of $T_2 - T_1$. Since S, T_1, and S_1 all lie in the same Euclidean subspace of tree-space (with S, S_1 on the boundary of the subspace), and similarly S, T_2, S_2. The segment T_1, S_1 is perpendicular to S_1, S, and similarly T_2, S_2 and S_2, S. So we can write the total length of the path from T_1 to S to T_2 as the sum of the two Euclidean lengths

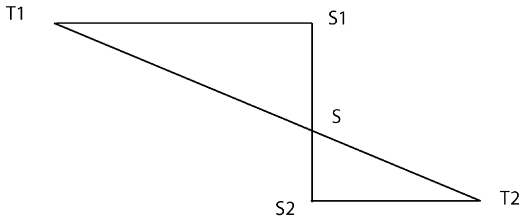

Fig. 2. The two-dimensional sub-planes containing T_1, L and T_2, L can be folded or unfolded along L without changing the intrinsic length of the shortest path from T_1 to T_2. So the length can be determined in the sub-plane formed by unfolding them to lie in the same plane.

$$
\begin{aligned}
D_{hi}(T_1,T_2) &= d(T_1,S)+d(S,T_2) \\
&= \sqrt{d^2(T_1,S_1)+d^2(S_1,S)} \\
&\quad + \sqrt{d^2(S,S_2)+d^2(S_2,T_2)}.
\end{aligned}
$$

Claim 1. *The choice of S minimizing D_{hi} must lie on the line connecting S_1 and S_2.*

Proof. Assume for the purpose of contradiction that there is some choice of S which minimizes D_{hi} which is not on the line through S_1, S_2. Now let S' be the projection of S to the line through S_1, S_2. The points S', S, S_1, S_2 all lie in a two-dimensional plane. Clearly $d(S_1, S') \leqslant d(S_1, S)$ and $d(S_2, S') \leqslant d(S_2, S)$, so using S' as the intermediate tree would give a shorter path, contradicting the assumption. □

We use the claim to show our main result:

Theorem 2. *The lower and upper bound on the geodesic difference between any two trees differs by a most a multiplicative factor of $\sqrt{2}$.*

Proof. From the claim, S lies on the line L containing S_1, S_2, and the entire path from T_1 to T_2 lies in the two two-dimensional subspaces containing, respectively, L, T_1 and L, T_2. We can visualize the path using two triangles in the same plane, as in Fig. 2. This corresponds to unfolding the two-spaces along L so that we can draw the path as a straight line; the intrinsic length of the path remains the same however we unfold the space. Notice we can re-write the distance as

$$
\begin{aligned}
D_{hi}(T_1,T_2) &= \big[\big(d(T_1,S_1)+d(S_2,T_2)\big)^2 \\
&\quad + \big(d(S_1,S)+d(S,S_2)\big)^2\big]^{1/2}
\end{aligned}
$$

which simplifies to:

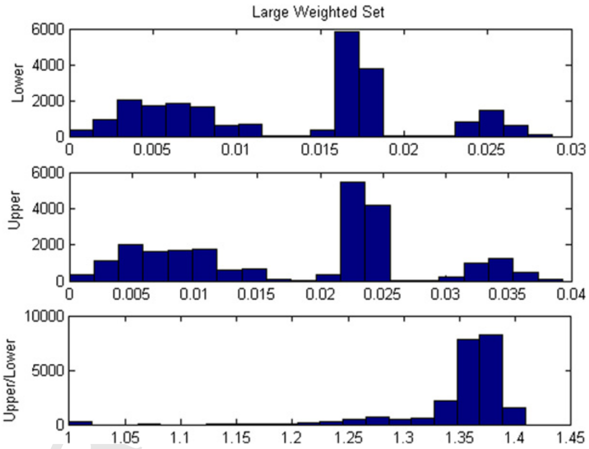

Fig. 3. Histogram of the pairwise distances of the animal dataset. Top graph shows the distances calculated with the lower bound approximation. The middle graph shows the upper bound approximation. The bottom graph shows a histogram of the ratio of the upper bound to lower bound for each pair of trees. Note that the maximum for the ratio is $\sqrt{2}$ which is achieved for a large fraction of the pairs of trees.

$$
\begin{aligned}
D_{hi}(T_1,T_2) = \Bigg[&\Bigg(\sqrt{\sum_{e\in(T_1-T_2)}\delta(e)}+\sqrt{\sum_{e\in(T_2-T_1)}\delta(e)}\Bigg)^2 \\
&+\sum_{e\in(T_1\cap T_2)}\delta(e)\Bigg]^{1/2}.
\end{aligned}
$$

Similarly, we can re-write

$$
\begin{aligned}
D_{lo} = \Bigg[&\sum_{e\in(T_1-T_2)}\delta(e)+\sum_{e\in(T_2-T_1)}\delta(e) \\
&+\sum_{e\in(T_1\cap T_2)}\delta(e)\Bigg]^{1/2}.
\end{aligned}
$$

Now we consider the ratio D_{hi}/D_{lo}. We would like to show that this ratio is never larger than $\sqrt{2}$. Let

$$
a=\Bigg(\sum_{e\in(T_1-T_2)}\delta(e)\Bigg)^{1/2}, \qquad b=\Bigg(\sum_{e\in(T_2-T_1)}\delta(e)\Bigg)^{1/2},
$$

and

$$
c=\sum_{e\in(T_1\cap T_2)}\delta(e).
$$

Then, the ratio can be written as: $D_{hi}/D_{lo} = ((a+b)^2 + c)/(a^2+b^2+c)$. Clearly this is maximized when $c = 0$, reducing the problem to the maximum of $(a+b)^2/(a^2+b^2)$. This is maximized when $a = b$. That is, the ratio between the upper and lower bounds is maximized when the common edges have the same weights in T_1 and T_2 (i.e., $c = 0$) and the edges in $T_1 - T_2$ contribute the same weight a as the edges in $T_2 - T_1$. In this case $D_{hi} = 2a$ and $D_{lo} = \sqrt{2}a$. □

Please cite this article in press as: N. Amenta et al., Approximating geodesic tree distance, Information Processing Letters (2007), doi:10.1016/j.ipl.2007.02.008

5. Running time

Note that both the lower and upper bounds can be calculated in time linear with respect to the number of leaves in the tree. For the lower bound, the simple Euclidean distance calculation is $O(n)$. The upper bound's running time calculates a constant number of Euclidean distances, and also calculates the strict consensus tree as assigns different weights to its edges to produce the trees S_1, S_2 and S. The strict consensus tree can be obtained in linear time again by using the techniques of Day [8]. This gives a linear running time for the upper bound algorithm.

6. Application

We have implemented our distance bounds and examined the inter-tree distances for a large sets of phylogenetic trees. The trees were derived from an RNA dataset for 48 animals. In particular, 18S (large ribosomal subunit) RNA was extracted for 48 animals, across the tree of life, from the European rRNA database (http://rrna.uia.ac.be/). A heuristic parsimony search, using PAUP* [13], was run and the 215 best scoring trees were saved.

7. Conclusion and future work

While the geodesic distance is hard to compute exactly, we give linear time algorithms that compute lower and upper bounds that differ by a constant factor. These approximations show promise for distinguishing more characteristics of the dataset. Future work includes finding tighter bounds on the geodesic distance, while keeping the efficient running time, to yield better distance methods.

Acknowledgements

This project was supported by NSF-ITR 0121651/0121682. The first author was also supported by an Alfred P. Sloan Foundation Research Fellowship. The first and fourth authors would like to thank the American Institute of Mathematics for the travel support and their organization of the conference "Geometric models of biological phenomena" (June 2003) that motivated this work. We would like to thank Derrick Zwickl of the Hillis Laboratory at the University of Texas, Austin for generating the dataset used in this paper.

References

[1] B.L. Allen, M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. of Combin. 5 (1) (2001) 1–15.

[2] N. Amenta, F. Clarke, K.St. John, A linear-time majority tree, in: Third International Workshop WABI 2003 (Workshop on Algorithms in Biology), Budapest, Hungary, in: Lecture Notes in Bioinformatics (subseries of Lecture Notes in Computer Science), vol. 2812, Springer, 2003, pp. 216–227.

[3] N. Amenta, J. Klingner, Case study: Visualizing sets of evolutionary trees, in: 8th IEEE Symposium on Information Visualization (InfoVis 2002), 2002, pp. 71–74. Software available at comet.lehman.cuny.edu/treeviz.

[4] L.J. Billera, S.P. Holmes, K. Vogtmann, Geometry of the space of phylogenetic trees, Adv. Appl. Math. 27 (2001) 733–767.

[5] M. Bordewich, C. Semple, On the computational complexity of the rooted subtree prune and regraft distance, Ann. of Combin. 8 (2005) 409–423.

[6] D. Bryant, Hunting for trees, building trees and comparing trees: theory and method in phylogenetic analysis, PhD thesis, Dept. of Mathematics, University of Canterbury, 1997.

[7] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, L. Zhang, On computing the nearest neighbor interchange distance, in: D.Z. Du, P.M. Pardalos, J. Wang (Eds.), Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications, in: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 55, American Mathematical Society, 2000, pp. 125–143.

[8] W.H.E. Day, Optimal algorithms for comparing trees with labeled leaves, J. Classification 2 (1) (1985) 7–28.

[9] D.M. Hillis, B.K. Mable, C. Moritz, Molecular Systematics, Sinauer Assoc., Sunderland, MA, 1996.

[10] J.P. Huelsenbeck, F. Ronquist, MrBayes: Bayesian inference of phylogeny, 2001.

[11] K.C. Nixon, The parsimony ratchet, a new method for rapid parsimony analysis, Cladistics 15 (1999) 407.

[12] C. Stockham, L.-S. Wang, T. Warnow, Statistically based post-processing of phylogenetic analysis by clustering, in: Proceedings of 10th Internat. Conf. on Intelligent Systems for Molecular Biology (ISMB'02), Edmonton, Canada, 2002, pp. 285–293.

[13] D.L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates, Sunderland, MA, 2002.

Please cite this article in press as: N. Amenta et al., Approximating geodesic tree distance, Information Processing Letters (2007), doi:10.1016/j.ipl.2007.02.008