Discrete Applied Mathematics 217 (2017) 405–426

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Discriminative measures for comparison of phylogenetic trees

Omur Arslan *, Dan P. Guralnik, Daniel E. Koditschek

Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 3 May 2016
Accepted 1 October 2016
Available online 27 October 2016

Keywords:
Phylogenetic trees
Evolutionary trees
Nearest Neighbor Interchange
Comparison of classifications
Tree metric

ABSTRACT

In this paper we introduce and study three new measures for efficient discriminative comparison of phylogenetic trees. The *NNI navigation dissimilarity* d_{nav} counts the steps along a "combing" of the Nearest Neighbor Interchange (NNI) graph of binary hierarchies, providing an efficient approximation to the (NP-hard) NNI distance in terms of "edit length". At the same time, a closed form formula for d_{nav} presents it as a weighted count of pairwise incompatibilities between clusters, lending it the character of an edge dissimilarity measure as well. A relaxation of this formula to a simple count yields another measure on *all* trees — the *crossing dissimilarity* d_{CM}. Both dissimilarities are symmetric and positive definite (vanish only between identical trees) on binary hierarchies but they fail to satisfy the triangle inequality. Nevertheless, both are bounded below by the widely used Robinson–Foulds metric and bounded above by a closely related true metric, the *cluster-cardinality metric* d_{CC}. We show that each of the three proposed new dissimilarities is computable in time $O(n^2)$ in the number of leaves n, and conclude the paper with a brief numerical exploration of the distribution over tree space of these dissimilarities in comparison with the Robinson–Foulds metric and the more recently introduced matching-split distance.

1. Introduction

1.1. Motivation

A fundamental classification problem common to both computational biology and engineering is the efficient and informative comparison of hierarchical structures. In bioinformatics settings, these typically take the form of phylogenetic trees representing evolutionary relationships within a set S of taxa. In pattern recognition and data mining settings, hierarchical trees are often used to encode nested sequences of groupings of a set of observations. Dissimilarity between combinatorial trees has been measured in the past literature largely by recourse to one of two separate approaches: comparing edges and counting edit distances. Representing the former approach, a widely used tree metric is the Robinson–Foulds (RF) distance, d_{RF}, [30] whose count of the disparate edges between trees requires linear time, $O(n)$, in the number of leaves, n, to compute [18]. Empirically, d_{RF} offers only a very coarse measure of disparity, and among its many proposed refinements, the recent matching split distance d_{MS}, [8,24] offers a more discriminative metric albeit with considerably higher computational cost, $O(n^{2.5}\log n)$. Alternatively, various edit distances have been proposed [29,26,1,20] but the most natural variant, the Nearest Neighbor Interchange (NNI) distance d_{NNI}, entails an NP-complete computation for both labeled and unlabeled trees [17].

[*] Corresponding author.
E-mail addresses: omur@seas.upenn.edu (O. Arslan), guralnik@seas.upenn.edu (D.P. Guralnik), kod@seas.upenn.edu (D.E. Koditschek).

http://dx.doi.org/10.1016/j.dam.2016.10.003
0166-218X/© 2016 Elsevier B.V. All rights reserved.

1.2. Results

Our main contribution is the introduction of a dissimilarity measure on the space $\mathcal{BT}_S$ of labeled binary trees which bridges the above approaches by what is, effectively, a solution to the NNI navigation problem in $\mathcal{BT}_S$:

Problem 1 (*NNI Navigation Problem*)**.** Given a target $\tau \in \mathcal{BT}_S$, provide an efficient algorithm $\mathcal{A}_\tau$ which, for any $\sigma \in \mathcal{BT}_S$, computes a Nearest Neighbor Interchange to be performed on σ while guaranteeing that successive application of $\mathcal{A}_\tau$ terminates in τ.

This problem is motivated by applications in coordinated robot navigation [2–5], where a group of robots is required to reconfigure reactively in real time their (structural) adjacencies while navigating towards a desired goal configuration. Thus, our particular formulation of the problem is inspired by the notion of reactive planning [12], but may likely hold value for researchers interested in tree consensus and averaging as well.

Of course, since computation of d_{NNI} is NP-hard, one cannot hope for repeated applications of $\mathcal{A}_\tau$ to produce NNI geodesics without incurring prohibitive complexity in each iteration. However, as we will show, constructing an efficient navigation scheme is possible if we allow the algorithm to produce less restricted paths: for $|S| = n$, our navigation algorithms require $O(n)$ time for each iteration and produce paths of length $O(n^2)$ (as compared to the $O(n \log n)$ diameter of d_{NNI} — see (19)).

Additional insight into the geometry of the space $(\mathcal{BT}_S, d_{\mathrm{NNI}})$ is gained by recognizing a significant degree of freedom with which our navigation algorithm may select the required tree restructuring operation at each stage. As it turns out, for any given target τ, the repeated application of $\mathcal{A}_\tau$ to a tree σ until reaching τ will yield paths of equal lengths regardless of any choices made along the way. This length, by definition, is the navigation dissimilarity $d_{nav}(\sigma, \tau)$ (and is obtained, in the manner described, in $O(n^3)$ time, though more efficient implementations will guarantee $O(n^2)$). At the same time, a closed form formula we derive for d_{nav} allows us to avoid computing a navigation path when only the value of d_{nav} is needed, and computes it in $O(n^2)$ time. Surprisingly, despite the asymmetric character of its construction, d_{nav} is a symmetric (and positive definite) dissimilarity on $\mathcal{BT}_S$, though it fails to be a metric.

Although d_{nav} does not satisfy the triangle inequality, it is related to the well accepted Robinson–Foulds distance by the following tight bounds:

$$d_{RF} \le d_{nav} \le \frac{1}{2}d_{RF}^2 + \frac{1}{2}d_{RF}. \tag{1}$$

We find it useful to introduce a "relaxation" of d_{nav}, the *crossing dissimilarity* d_{CM}. This dissimilarity simply counts all the pairwise cluster incompatibilities between two trees, hence it is symmetric, positive-definite, and computable in $O(n^2)$ time. In fact, the two dissimilarities are commensurable, leading to similar bounds in terms of d_{RF}:

$$d_{RF} \le d_{nav} \le \frac{3}{2}d_{CM}, \qquad d_{RF} \le d_{CM} \le d_{RF}^2. \tag{2}$$

Finally, we introduce a true metric whose spatial resolution and computational complexity is comparable to those our new dissimilarities. Exploiting a well known relation between trees and ultrametrics [14], we also introduce *the cluster-cardinality distance* d_{CC} – constructed as the pullback of a matrix norm along an embedding of hierarchies into the space of matrices and computable in $O(n^2)$ time – which is a true metric bounding d_{CM} from above (and hence also d_{nav}, up to a constant factor). Thus, cumulatively we obtain:

$$\frac{2}{3}d_{RF} \le \frac{2}{3}d_{nav} \le d_{CM} \le d_{CC}. \tag{3}$$

We have surveyed some of the new features of our tree proximity measures that might hold interest for pattern classification and phylogeny analysis relative to the diverse alternatives that have appeared in the literature. Closest among these many alternatives [23,15,10], d_{nav} has some resemblance to an early NNI graph navigation algorithm, d_{ra} [10] which used a divide-and-conquer approach with a balancing strategy to achieve an $O(n \log n)$ computation of tree dissimilarity. Notwithstanding its lower computational cost, in contrast to d_{nav}, the recursive definition of d_{ra}, as with many NNI distance approximations [23,15,10], does not admit a closed form expression.

It is often of interest to compare more than pairs of hierarchies at a time, and the notion of a "consensus" tree has accordingly claimed a good deal of attention in the literature [11]. For instance, the majority rule tree [25] of a set of trees is a median tree respecting the RF distance and provides statistics on the central tendency of trees [6]. When d_{nav} and d_{CM} are extended to degenerate trees they fail to be positive definite, and thus their behavior over (typically degenerate) consensus trees departs still further from the properties of a true metric. However, it turns out that both notions of a consensus tree (strict [22], and loose/semi-strict [9]) behave as median trees with respect to both our dissimilarities. In fact, the loose consensus tree is the maximal (finest) median tree with respect to inclusion for both d_{nav} and d_{CM}.

The paper is organized as follows. Section 2 briefly summarizes the necessary background while introducing the notation used throughout the sequel. Section 3 introduces and studies the cluster-cardinality distance d_{CC} and the crossing dissimilarity d_{CM}. In Section 4 we present a solution of the NNI navigation problem and study properties of the resulting NNI navigation dissimilarity d_{nav} and its relations with other tree dissimilarity measures. Section 5 discusses the relation between