Journal of Theoretical Biology 398 (2016) 145–149

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A note on the path interval distance

Jane Ivy Coons [a], Joseph Rusinko [b,*]

[a] North Carolina State University, United States
[b] Department of Mathematics and Computer Science, Hobart and William Smith Colleges, 309 Lansing Hall Geneva, New York 14456, United States

HIGHLIGHTS

- Path interval distance analyzed as a method for comparing phylogenetic trees.
- Path interval distance captures global similarities between tree topologies.
- Random trees are not likely to be maximally distant under the path interval distance.

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form
11 March 2016
Accepted 17 March 2016
Available online 1 April 2016

Keywords:
Tree metrics
Phylogenetics
Cophylogenetics

ABSTRACT

The path interval distance accounts for global congruence between locally incongruent trees. We show that the path interval distance provides a lower bound for the nearest neighbor interchange distance. In contrast to the Robinson–Foulds distance, random pairs of trees are unlikely to be maximally distant from one another under the path interval distance. These features indicate that the path interval distance should play a role in phylogenomics where the comparison of trees on a fixed set of taxa is becoming increasingly important.

1. Introduction

Biologists display evolutionary relationships between operational taxonomical units, or taxa, as leaves of a tree. In many instances a collection of trees may be constructed from a single set of taxa. In other cases, while the taxa on two trees may be biologically distinct, one may wish to identify taxa on each tree to make a comparison. In this note we assume trees are constructed from a single set of taxa.

Tree metrics have been developed to compare the difference between trees generated using the same set of taxa and to assist when combining a set of trees to form a single tree which reflects the shared aspects of the elements of the set.

When testing reconstruction algorithms one compares a model tree to sample trees reconstructed from simulated data using the algorithm (Guindon et al., 2010; Lin et al., 2011; Salzburger et al., 2011). Alternatively, a step in many reconstruction algorithms involves finding a consensus tree among conflicting trees on the same set of taxa (Bansal et al., 2011; Bryant, 2003; Libeskind-Hadas et al., 2014). In the study of cophylogenetics one may identify the taxa on a host tree with taxa on a parasite tree if the parasite is known to live on a unique host. In this case the leaf labels refer to an identified pair of taxa (one host and one parasite) (Conow et al., 2010; Huggins et al., 2012; Ovadia et al., 2011). Trees with the same leaf set are also prevalent in phylogenomics when building a species tree from a set of gene trees (Heled and Drummond, 2010; Hughes et al., 2007; Liu and Yu, 2011; Mirarab et al., 2014).

Various factors may cause differences among trees on the same leaf set such as; gene species tree discordance under the coalescent model (Degnan and Rosenberg, 2009), a parasite's failure to speciate in reaction to a host speciation (Page, 2003), imperfect reconstruction algorithms, or incongruities among phylogenetic trees. However, given the shared evolutionary history, one expects to find a topological relationship between the trees.

Many tree metrics have been developed to analyze the differences between trees. Perhaps most famous among these methods are the Robinson–Foulds distance and the Nearest Neighbor Interchange (NNI) distance. Recently a new metric, *k*-interval cospeciation, was proposed in the context of cophylogeny (Huggins et al., 2012). As *k*-interval cospeciation has broad applicability in comparing trees with identified taxa sets, we refer to this metric instead as the *path interval distance* as this better reflects the generality of

* Corresponding author. Tel. +1 315 781 3607.
E-mail addresses: janyivycoons@gmail.com (J.I. Coons), rusinko@hws.edu (J. Rusinko).

http://dx.doi.org/10.1016/j.jtbi.2016.03.027
0022-5193/

the metric and its relationship to the path-difference metric (Steel and Penny, 1993) and the edge-difference distance (Williams and Clifford, 1971).

The path interval distance illuminates global similarities between trees that classical tree metrics do not uncover. We do not propose that the path interval distance supplants the Robinson–Foulds or NNI distances as their importance and statistical properties have been well documented (Bryant and Steel, 2009; Pattengale et al., 2007). However, alternative tree metrics may yield additional insight into the relationship between incongruous trees.

We compare the path interval distance to traditional tree metrics and explore its combinatorial and statistical properties. In particular we show that unlike the Robinson–Foulds metric, the probability that two randomly selected trees are maximally distant under the path interval distance goes to zero as the number of taxa increases. Thus the path interval distance may be useful as a tool for analyzing reconstruction algorithms on simulated data or developing consensus methods in phylogenomics.

2. Relationship between the path interval distance and other discrete tree metrics

We restrict our attention to unrooted binary trees where each non-leaf vertex has degree three. At each leaf of the tree is a *taxon* (plural: taxa), such as a gene or species. The number of unrooted binary trees on n taxa is denoted $ub(n)$ and given by,

$$ub(n) = (2n-5)!! = (2n-5) \times (2n-3) \times \cdots \times 5 \times 3 \times 1.$$

We define a *cherry* to be a pair of taxa with exactly two edges between them, and a *caterpillar tree* to be a phylogenetic tree with exactly two cherries. The *shortest path* between two taxa on a tree T is the smallest connected subgraph of T that contains both taxa. The *length* of the shortest path is the number of edges in the path. Since path interval distance is a discrete tree metric, we do not consider branch lengths.

Definition 1. Let T_1 and T_2 be phylogenetic trees on the taxon set X with $|X| = n$. Let $p_i(A,B)$ be the length of the shortest path between taxa A and B on T_i. The *path interval distance* between T_1 and T_2, denoted $d(T_1,T_2)$, is the maximum of $|p_1(A,B) - p_2(A,B)|$ over all pairs of taxa $\{A,B\} \in X^2$.

It follows that the path interval distance is a *tree metric* (Huggins et al., 2012) (where this is referred to as k-interval-cospeciation). See Xi et al. (2015) for a concise explanation of the relationship between the path interval distance and other tree metrics. In this note we contrast the path interval distance with several of the more popular discrete tree metrics.

Definition 2. A tree metric, d_a, *determines* d_b if there exists a bijection, $f : \mathbb{R} \to \mathbb{R}$ such that $f(d_a(T_1,T_2)) = d_b(T_1,T_2)$ for all n taxa trees T_1 and T_2.

We show that several commonly used discrete tree metrics do not determine the path interval distance. In fact, the path interval distance reflects global similarities between trees where other tree metrics do not.

One common tree metric is the nearest neighbor interchange distance. A *nearest neighbor interchange* (NNI) is a tree rearrangement operation in which we switch two subtrees of a tree that are joined by a single edge (Moore et al., 1973; Robinson, 1971). The nearest neighbor interchange distance between T_1 and T_2 is the smallest number of NNIs that takes to transform T_1 into T_2. Finding the NNI distance between two given trees is an NP-complete problem (DasGupta et al., 1997), while the path interval distance can be computed in $O(n^2)$ (Xi et al., 2015).

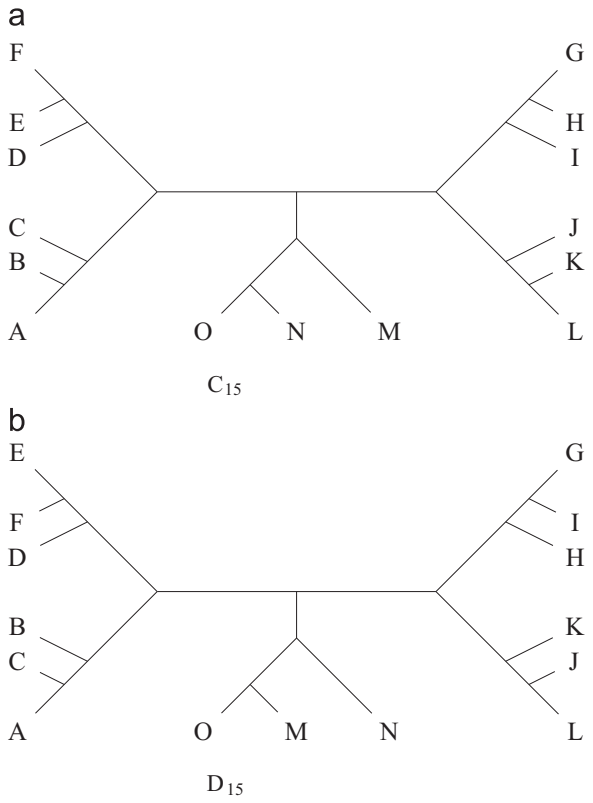

Fig. 1. Counterexample trees introduced in Huggins et al. (2012).

Huggins et al. (2012) proved that two trees are path interval distance one apart if and only if they are NNI distance one apart. However, they also introduced a counterexample to the possibility of the path interval distance providing an upper bound for the NNI distance for any $k > 1$.

Theorem 1. *The NNI distance does not determine the path interval distance.*

Proof. Let C_{3x} and D_{3x} be trees built by adding rooted triples on the same leaf set, but with differing cherries to the pendant edges of two $3x$-taxa caterpillar trees C and D (see Fig. 1(a) and (b)).

Since no NNI operation is able to alter two of these rooted triples at the same time, the NNI distance between C_{3x} and D_{3x} is x. However, the two trees have a path interval distance of two. Moreover by matching a rooted triple from C_{3x} with D_{3x} we maintain path interval distance of two, but decrease the NNI distance by one. Therefore, the NNI distance does not determine the path interval distance.□

Remark 1. Similarly, one can also show that path interval distance does not determine the Robinson–Foulds distance (defined in Robinson and Foulds, 1981), the path difference distance(defined in Steel and Penny, 1993), or the maximum agreement subtree distance (defined in Finden and Gordon, 1985).

The path interval distance between two trees describes global similarities between the trees that other common discrete tree metrics do not. One way of describing this global similarity is through the length of the longest path between two taxa on the trees. Notice that the path interval distance between two trees provides an upper-bound on the difference between the length of the longest path of the two trees. Thus, trees differing by a small path interval distance must share a similar global topology.

While trees C_{3x} and D_{3x} show that the path interval distance and many of the traditional tree metrics can differ greatly, the path