The Generalized Robinson-Foulds Metric*

Sebastian Böcker[1,**], Stefan Canzar[2,**], and Gunnar W. Klau[3,**]

[1] Chair for Bioinformatics, Friedrich Schiller University Jena, Germany
sebastian.boecker@uni-jena.de
[2] Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
canzar@jhu.edu
[3] Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
gunnar.klau@cwi.nl

Abstract. The Robinson-Foulds (RF) metric is arguably the most widely used measure of phylogenetic tree similarity, despite its well-known shortcomings: For example, moving a single taxon in a tree can result in a tree that has maximum distance to the original one; but the two trees are identical if we remove the single taxon. To this end, we propose a natural extension of the RF metric that does not simply count *identical* clades but instead, also takes *similar* clades into consideration. In contrast to previous approaches, our model requires the matching between clades to respect the structure of the two trees, a property that the classical RF metric exhibits, too. We show that computing this generalized RF metric is, unfortunately, NP-hard. We then present a simple Integer Linear Program for its computation, and evaluate it by an all-against-all comparison of 100 trees from a benchmark data set. We find that matchings that respect the tree structure differ significantly from those that do not, underlining the importance of this natural condition.

1 Introduction

In 1981, Robinson and Foulds introduced an intriguingly simple yet intuitively well-motivated metric, which is nowadays known as *Robinson-Foulds (RF) metric* [18]. Given two phylogenetic trees, this metric counts the number of splits or clades induced by one of the trees but not the other. The RF metric is highly conservative, as only perfectly conserved splits or clades do not count towards the distance. The degree of conservation between any pair of clades that is not perfectly conserved, does not change the RF distance. See Fig. 1 for an example of two trees that are structurally similar but have maximum RF distance.

Other measures for comparing phylogenetic trees do capture that the trees in Fig. 1 are structurally similar: The Maximum Agreement Subtree (MAST)

* This work is supported in part by the National Institutes of Health under grant R01 HG006677.

** Equal contribution.

A. Darling and J. Stoye (Eds.): WABI 2013, LNBI 8126, pp. 156–169, 2013.

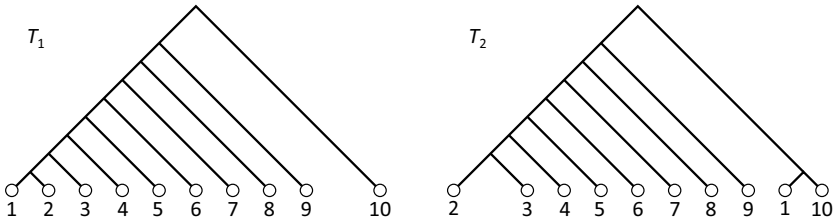

Fig. 1. Two rooted phylogenetic trees. Despite their high similarity, the RF distance of these two trees is 16, the maximum distance of two rooted trees with ten leaves.

score [11, 13] of the two trees is 9, where 10 is the highest possible score of two trees with 10 leaves. Secondly, the triplet distance counts the number of induced triplet trees on three taxa that are not shared by the two trees [2, 6]. Both measures are less frequently applied than the RF metric, and one may argue that this is due to certain "issues" of these measures: For example, if the trees contain (soft) polytomies or arbitrarily resolved polytomies, then we may have to exclude large parts of the trees from the MAST due to a single polytomy. Lastly, there are distance measures based on the number of branch-swapping operations to transform one tree into another; many of these measures are computationally hard to compute [1]. Such tree modifications are routinely used in local search optimization procedures, but rarely to compute distances in practice.

From an applied view, the comparison of two phylogenetic trees with identical taxa set has been frequently addressed in the literature [12, 16, 17]. This is of interest for comparing phylogenetic trees computed using different methods, output trees of an (MC)MCMC method, or host-parasite comparisons. Mutzner *et al.* [16] introduced the "best corresponding node" concept which, unfortunately, is not symmetric: Node a in the first tree may correspond to node b in the second, whereas b corresponds to a different node c in the first tree, and so on. Nye *et al.* [17] suggested to compute a matching between the inner nodes of the two trees, thereby enforcing symmetry. Later, Bogdanowicz [3] and, independently, Lin *et al.* [15] proposed to use these matchings to introduce a "generalized" version of the RF distance, see also [4]. Using matchings for comparing trees as part of MAST computations, was pioneered by Kao *et al.* [13].

Here, we present a straightforward generalization of the RF distance that allows us to relax its highly conservative behavior. At the same time, we can make this distance "arbitrarily similar" to the original RF distance. Unfortunately, computing this new distance is NP-hard, as we will show in Section 3. Our work generalizes and formalizes that of Nye *et al.* [17]: Their clade matching does not respect the structure of the two trees, see Fig. 1 and below. As a consequence, the matching distances from [3, 15] are no proper generalization of the RF distance: These distances treat the two input trees as collections of (unrelated) clades but *ignore the tree topologies*. In contrast, the RF distance does respect tree topologies, and so does our generalization.

In the following, we will concentrate on rooted phylogenetic trees.