An Approach to Relax the Infinite Sites Assumption in Tumor Phylogeny Distance Measures

Quoc Nguyen
Department of Computer Science
Carleton College
Northfield, Minnesota, USA
nguyenq2@carleton.edu

Layla Oesper
Department of Computer Science
Carleton College
Northfield, Minnesota, USA
loesper@carleton.edu

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9994895

Abstract—Tumor phylogenies representing the evolutionary history of a tumor can be inferred from sequencing data. We propose a matching-based framework which allows existing distance measures to be applied to transformations of phylogenies that do not adhere to the Infinite Sites Assumption.

Index Terms—tumor phylogeny, clonal tree, tumor tree, Infinite Sites Assumption

I. Introduction

Methods have emerged that can infer tumor evolutionary histories in the form of tumor phylogenies from sequencing data. An assumption often used in tumor phylogeny inference is the Infinite Sites Assumption (ISA) [1] which disallows homoplasy and back mutations within the genome. Distance measures have been designed for comparing tumor phylogenies under the ISA, but there is a movement toward more relaxed models of tumor evolution [2]. The k-Dollo model allows for mutational losses, but most existing distance measures are not fit to compare k-Dollo phylogenies. We analyze why existing distance measures don't work on k-Dollo phylogenies and introduce a framework to resolve this problem.

II. Methods

We devise a way to apply existing distance measures to k-Dollo phylogenies by transforming k-Dollo phylogenies into ones valid under the ISA. Formally, the problem we are trying to solve is as follows. Generalized Matching Distance (GMD) Problem: Input: Tumor phylogenies T and T^0. A distance function dist designed for ISA-phylogenies. Output: Two trees $\overline{T}$ and $\overline{T^0}$ that have been relabeled based on the matching M in the matching graph $G_{(T,T^0)}$ that minimizes the dist between T and T^C. The matching graph of two phylogenies T and T^0 is a bipartite graph $G_{(T,T^0)} = (A \cup B, E)$ whose vertices A (B) correspond to gains and losses in T (T^C), and whose edge set E is composed of edges (a, b) such that a and b correspond to either two gains or two losses of the same character, one in each tree. A visual depiction of GMD is provided in Fig 1.

We propose a way to approximate the GMD that explores applying different weighting schemes to $G_{(T,T^0)}$ and finds a min-cost matching with the Hungarian Algorithm. We consider the following weighting schemes (visualized in Fig 2).

- depth: an edge (a, b) of the matching graph $G_{(T,T^0)}$ has weight equal to the difference between the depths of a in T and b in T^0.
- parent: an edge (a, b) of the matching graph $G_{(T,T^0)}$ has weight of either 1, indicating that a and b don't share the same parent mutation, or 0 if they do.
- lineage: an edge (a, b) of the matching graph $G_{(T,T^0)}$ is weighted as the cardinality of the symmetric difference between the set of all ancestor and descendant mutations for a and b.

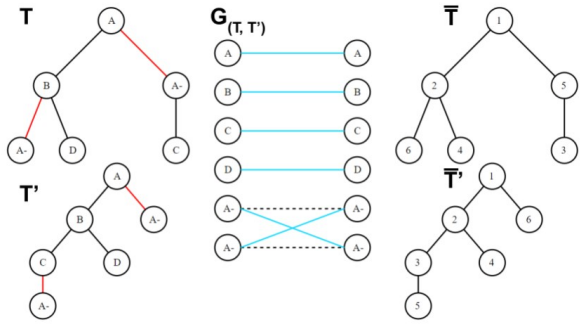

Fig. 1. We start with two 2-Dollo tumor phylogenies T and T^0 where the losses are labeled by red edges. Using the muatations of those phylogenies, we create the matching graph $G_{(T,T^0)}$. We then find an optimal matching M, represented by the blue edges, that minimizes some input distance function dist. From this matching, we can relabel T and T^0 to become $\overline{T}$ and $\overline{T^0}$, which both conform to the ISA.

III. Results

Early experiments show that the distance measure CASet [3] (not designed to handle ISA), when applied to k-dollo trees, often results in a higher distance than when combined with our GMD approach (see Fig 3). We tested four different distance measures combined with three different weighting schemes on the matching graph. We found that certain distance measures combined with specific weighting schemes often led to optimally solving the GMD (e.g., the parent weighting scheme combined with parent-child distance).

This project is supported by NSF award C AREER-IIS-2046011

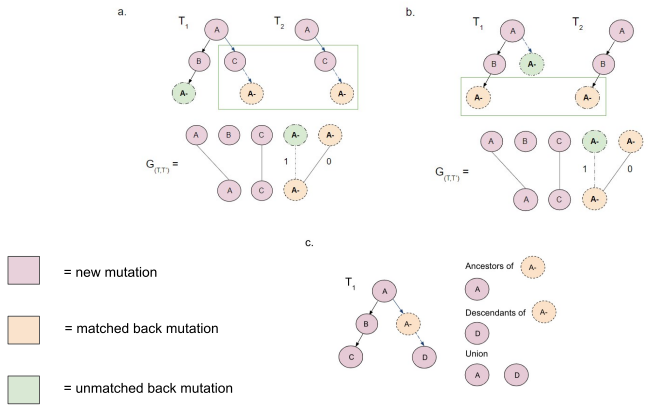

Fig. 2. (a) The parent weighting scheme and resulting matching graph $G_{(T,T')}$ for two trees. (b) The depth weighting scheme and the resulting matching graph $G_{(T,T')}$ for two trees. (c) The lineage weighting scheme. Provided is a description of the sets being compared in the lineage weighting scheme. Specifically, the ancestor set and the descendant set for a given node is unioned in order to form its lineage set.

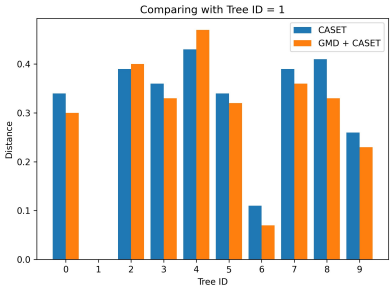

Fig. 3. This figure shows pairwise comparisons between 10 simulated 4-Dollo tumor phylogenies compared against a single 4-Dollo phylogeny identified with Tree ID 1. Comparisons were made with (1) an ISA specific distance function (CASet) by itself and (2) combining that distance function with the GMD framework. We see that (1) generally resulted in higher distances when comparing the phylogenies.

IV. Conclusion

There is a need for distance measures designed for phylogenies that don't adhere to the ISA. We have created a framework enabling existing distance measures designed for ISA-phylogenies to compare k-Dollo phylogenies (losses are allowed). We have shown that our heuristic approach works well, even optimally, in some cases. Future directions include experiments on larger trees and those with more losses.

V. Acknowledgement

We thank Mohammed El-Kebir for useful conversations in the intial stages of this project.

References

[1] M. Kimura, "The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations," Genetics, vol. 61, no. 4, pp. 893–903, Apr. 1969.

[2] Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. 2017. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Res 27, 11 (November 2017), 1885–1894. DOI:https://doi.org/10.1101/gr.220707.117

[3] Z. DiNardo, K. Tomlinson, A. Ritz, and L. Oesper, "Distance measures for tumor evolutionary trees," Bioinformatics, vol. 36, no. 7, pp. 2090–2097, Apr. 2020, doi: 10.1093/bioinformatics/btz869.