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

***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

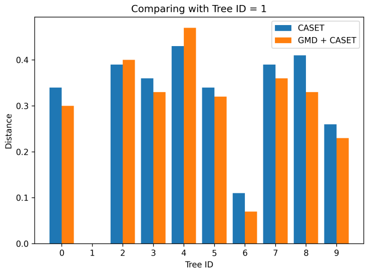
We devise a way to apply existing distance measures to *k*-Dollo phylogenies by transforming *k*-Dollo phylogenies into ones valid under the ISA. Formally, the problem we are trying to solve is as follows. **Generalized Matching Distance (GMD) Problem:** Input: Tumor phylogenies T and T'. A distance function *dist* designed for ISA-phylogenies. Output: Two trees $\overline{T}$ and $\overline{T'}$ that have been relabeled based on the matching M in the matching graph $G_{(T,T')}$ that minimizes the *dist* between T and T'. The *matching graph* of two phylogenies T and T' is a bipartite graph $G_{(T,T')} = (A \cup B, E)$ whose vertices A (B) correspond to gains and losses in T (T'), 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.

We propose a way to approximate the GMD that explores applying different weighting schemes to $G_{(T,T')}$ and finds a min-cost matching with the Hungarian Algorithm.

This project is supported by NSF award CAREER-IIS-2046011

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 figure below). 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).



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.

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.