Combining Distance Measures on Tumor Evolutionary Trees

1st Cecilia Ehrlichman
Department of Computer Science
Carleton College
Northfield, MN
ehrlichmanc@carleton.edu

2nd Layla Oesper
Department of Computer Science
Carleton College
Northfield, MN
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/BIBM55620.2022.9995343

Abstract—Tumors develop through cells acquiring heritable mutations as part of an evolutionary process and these histories can be represented using tree structures. This work expands and combines already existing distance measures between tumor trees to incorporate the advantages of multiple approaches.

I. Introduction

When studying tumors, it is useful to have algorithms that derive their evolutionary history from sequencing data. These histories can be represented using tree structures, where each node in the tree represents the acquisition of a mutation. Therefore, a cell that has acquired a certain mutation at a node also inherits all ancestral mutations. Distance measures between trees can be beneficial when benchmarking new inference methods on simulated data by evaluating which trees are more similar to the true tree. Many distance measures for tumor evolutionary trees already exist. Each of these have been shown to be useful in certain situations. This research explores the possible ways of combining distance measures, specifically the CASet and DISC distance measures [1], to reap the benefits of both.

II. Methods

Both the CASet and DISC distance measures consider mutations inherited by pairs of nodes across a pair of trees. CASet compares the sets of commonly inherited mutations for each pair of nodes across both trees, and DISC compares the sets of distinctly inherited mutations. Our combined distance measures works by partitioning all mutations into four sets for each pair of mutations in both trees. Three of these sets are derived from CASet and DISC, but we additionally include non-inherited mutations so that every mutation is included in the partition. We then compare these partitions across both trees using a variety of metrics including the method from [2], the Rand Index, and taking the average of the Jaccard Distance between each of the four subsets. An intrinsic flaw in partitioning these mutations is that we lose information embedded in the identity of the subsets. Therefore, we propose a modified approach on partition distance that encodes this information. Additionally, we investigated a version of the Jaccard distance approach that weights sets by their size.

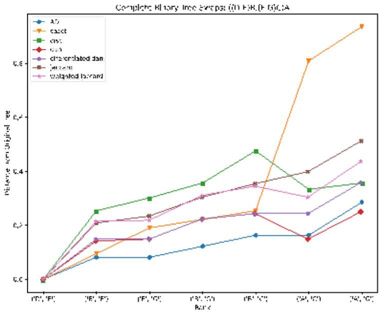

Fig. 1. Each pairwise swap compared to a complete binary tree. The swaps that vary most across distance measures are swaps with the root and another node.

III. Results

Our preliminary results indicate that pure partition distance is ineffective due to the lost information about the subsets. Still, we noticed that our new approach appears to combine elements of CASet and DISC and follows similar trends to previous distance measures. In one experiment we compared the distance between a base complete binary tree and each tree that was altered by a single pairwise swap (see Fig 1).

IV. Conclusions

Combining distance measures can be a useful technique when it is uncertain which factors are more useful for an instance, and therefore which distance measure to use. While basic partition distance is flawed, modified versions have the potential to be useful. Further directions for this could include taking a weighted average of the Jaccard distance that weights commonly inherited mutations over the other subsets.

References

[1] 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.

[2] D. Gusfield, ”Partition-Distance: A Problem and Class of Perfect Graphs Arising in Clustering,” Information Processing Letters 82(3), 159-164, May. 2002, doi:10.1016/S0020-0190(01)00263-0

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