# Deriving consensus tumor trees using integer linear programming

1st Matthew Smith-Erb
Department of Computer Science
Carleton College
Northfield, Minnesota
mcse@maine.rr.com

1st Ziyun Guang
Department of Computer Science
Carleton College
Northfield, Minnesota
guang.cathy@gmail.com

2nd Layla Oesper
Department of Computer Science
Carleton College
Northfield, Minnesota
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/BIBM55620.2022.9995388

Abstract—The acquisition of somatic mutations by a tumor can be modeled by a type of evolutionary tree. Although many methods have been developed to infer a tumor's evolutionary history, they can produce conflicting results for a single patient. A consensus tree that reconciles these possible trees is important for understanding the tumor's evolutionary process. We use integer linear programming to find a consensus tree among multiple plausible tumor evolutionary histories.

Index Terms—Cancer, evolutionary history, consensus, integer linear programming

## I. Introduction

A tumor is the result of an evolutionary process, typically depicted as a rooted tree where the vertices represent tumor cell populations, and the edges indicate ancestral relationships. A better understanding of such histories may provide important insights for effective treatment plans for patients. Although there has been improved inference of tumor evolutionary histories, interpretation can be challenging due to these methods returning different trees. Here, we introduce an integer linear programming (ILP) method for reconciling a collection of plausible trees. We allow the assignment of confidence weights to each input tree. Our ILP also considers all topologies and clusterings of mutations to find the consensus tree that best represents the input trees.

## II. Methods

We pose the Weighted m-Tumor Tree Consensus Problem (W-m-TTCP) which incorporates confidence weight to a previous narrower consensus problem solved by GraPhyC [2]. Given: (i) a set of plausible tumor trees, (ii) confidence weights describing their relative importance, and (iii) a specific distance measure. Output: a consensus tree that minimizes the total weighted distance to each input tree. We describe an ILP approach to solve the W-m-TTCP with Ancestor Descendant (AD) distance metric from [2]. AD distance represents the number of ancestor-descendant relationships in one tree but not the other. In the ILP formulation, we introduce constraints to produce a feasible tumor tree. We create an objective function that penalizes the output tree when it has different ancestral pairs from the input trees, which minimizes the weighted AD distance between the output tree and all inputs.

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

## III. Results

On simulated data, we compare our ILP model to ConTreeDP [1] and GraPhyC [2]. In each trial, we generated a set of input trees by applying a process of random edits to a simulated true tree. We ran our ILP, ConTreeDP, and GraPhyC on each trial's input trees, and measured the distance between the outputted consensus tree of each model and the original true tree. The distances were calculated with the AD metric, and CASet and DISC which [1] used to benchmark ConTreeDP.

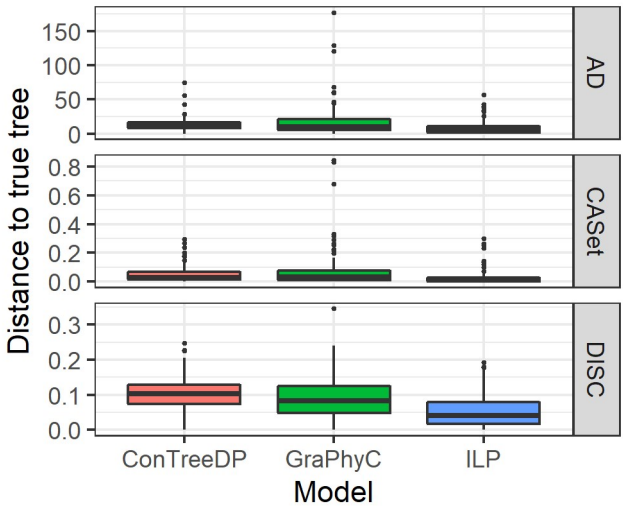


Fig. 1. Results showing our ILP's ability to uncover the true tree for the simulated data containing 30 mutations and 5 input trees.

In the simulations with 5 input trees, across 10, 20, and 30 mutations, our ILP outperforms ConTreeDP and GraPhyC for the three distance metrics. For instance, on 5 trees and 30 mutations, our ILP's average AD distance to the true tree was over 1.5⇥ closer than ConTreeDP's average, and over 2⇥ closer than GraPhyC's average. Additionally, our ILP outperformed the two consensus models for the CASet and

DISC metrics when benchmarked on simulated data containing 30 mutations and 5 trees.

We also analyzed the output of our ILP on data obtained from sequencing a patient with triple-negative breast cancer. Three possible trees for the patient's tumor were found by [3], and were used as the inputs to the consensus models. Our ILP and ConTreeDP generated identical consensus trees, differing from GraPhyC's tree by the relative placement of a single mutated gene, MAP3K4. However, the positioning of this mutation in our ILP's consensus tree may be more plausible considering the mutation's location in the input trees.

## IV. Conclusion

In this work, we introduce the W-m-TTCP, which we solve with an ILP to produce a consensus tumor tree. We demonstrate that our ILP was better at uncovering the true tree than other models on simulated data, and it produces more plausible results on real data. Future work includes more testing on real data sets and modifying our model to accept input trees with differing mutation sets.

## References


[1] X. Fu and R. Schwartz, "ConTreeDP: A consensus method of tumor trees based on maximum directed partition support problem," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 125-130.

[2] K. Govek, et al., "GraPhyC: Using Consensus to Infer Tumor Evolution," in IEEE/ACM Transactions on Computational Biology and Bioinformat-ics, vol. 19, no. 1, pp. 465-478, 1 Jan.-Feb. 2022.

[3] N. Karpov et al., "A multi-labeled tree edit distance for comparing", Proc. 18th Int. Workshop Algorithms Bioinf., pp. 22:1-22:19, 2018.