Project Presentation

Minhyuk Park

4/20/21

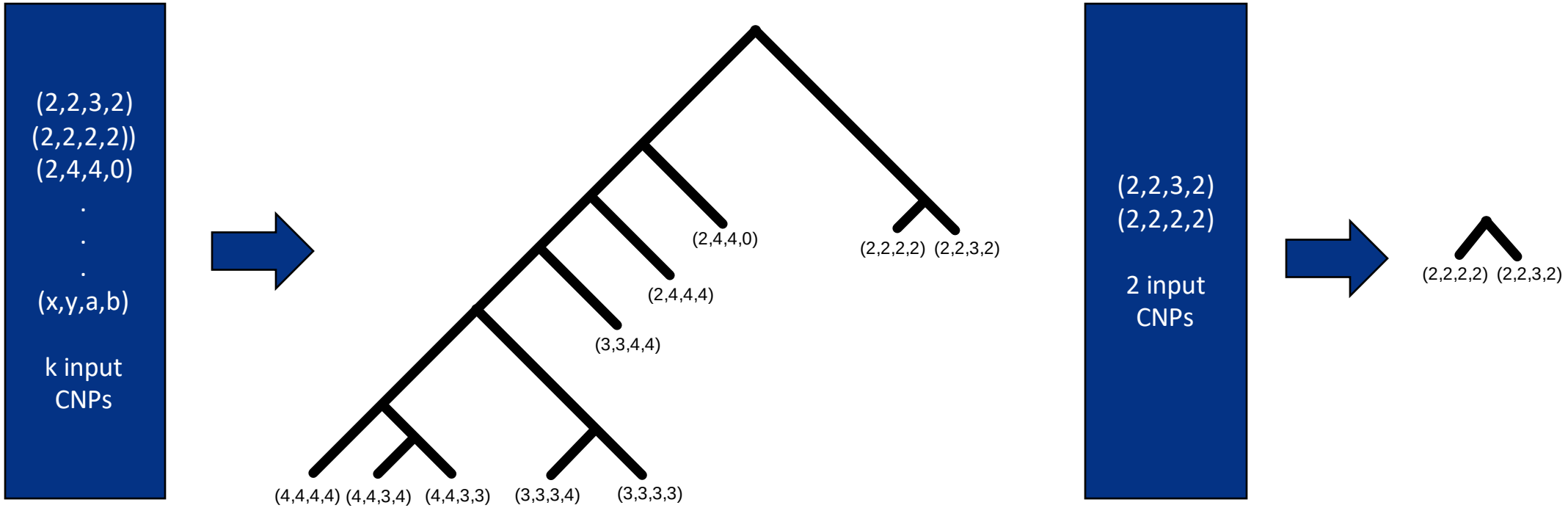
Citation

- El-Kebir, Mohammed, Benjamin J. Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira. Copy-Number Evolution Problems: Complexity and Algorithms. WABI 2016: 137-149

Background

- Copy Number Aberration (CNA):
  - Gains and losses of segments of the genome
  - Cancer evolutionary process
- CNT
  - Copy Number Tree
  - Finding a fully resolved tree whose leaves are copy number profiles
  - ILP solution exists
- CN3
  - Copy Number Triplet
  - Special case of copy number tree problem where there's 2 input copy number profiles
  - DP solution exists

CNT and CN3

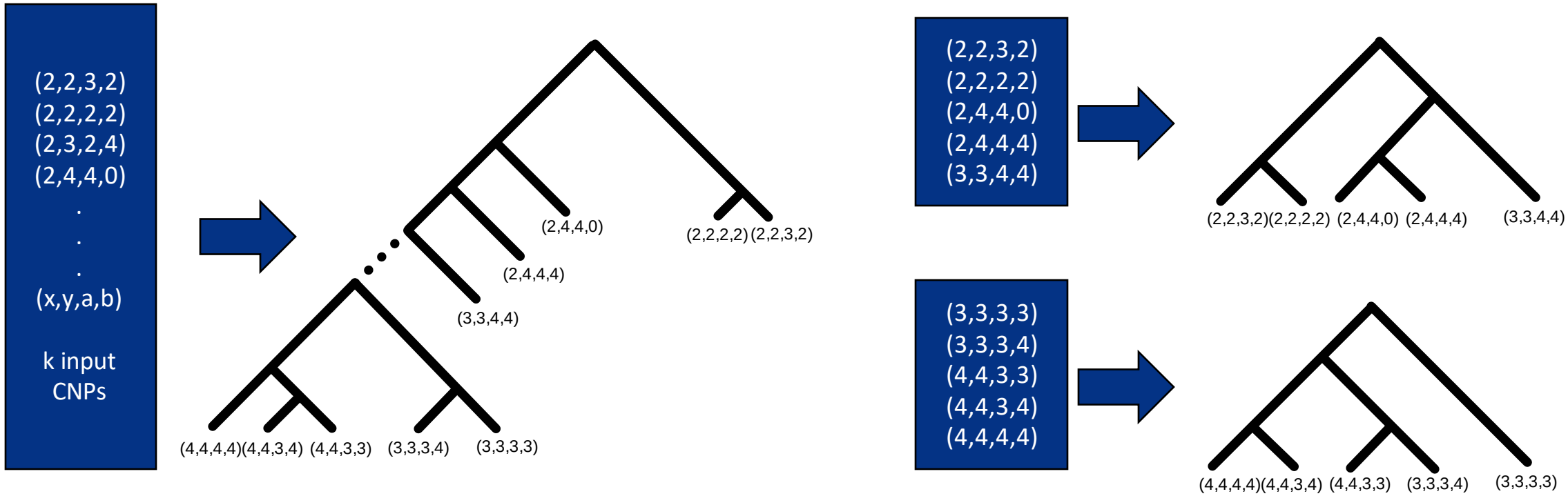


- We also get the internal node labeling
- Left is CNT; Right is CN3

CNT-ILP and CN3

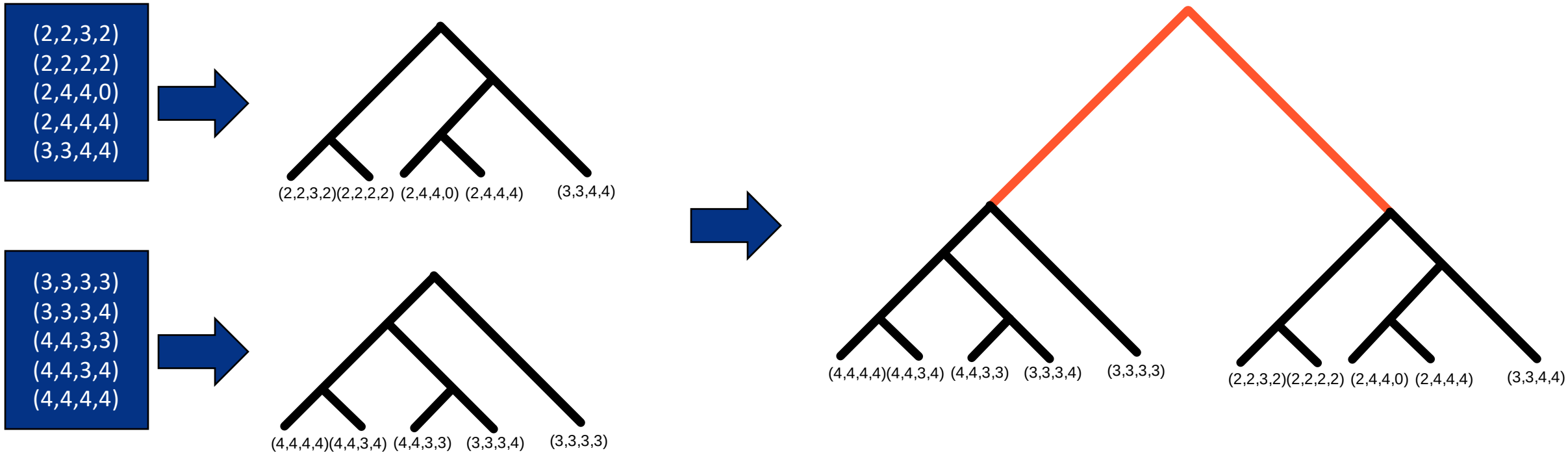
- CNT-ILP is a method that solves the CNT problem using integer linear programming
  - Computationally expensive
  - Can’t be used on very large instances
- CN3 is a dynamic programming algorithm that solves the CN3 problem using dynamic programming
  - Computationally cheap
  - Can only be used on instances where k = 2 (k is the number of leaves)

Divide-and-Conquer: Decomposition and Subtree Computation



- We first get an initial tree that helps us split the dataset into two
- Use CNT-ILP on each subset to get a better tree on those subsets compared to the initial tree

Divide-and-Conquer: Merging Subtrees



- Use CN3 on the roots of each subtree to obtain the root of the final tree
- Topology is fixed as two subtrees being siblings at the root node

Problem Statement

- Input:
  - Set of copy number profiles
- Output:
  - Tree topology, internal node copy number profiles
- Optimizing for:
  - Speed, no accuracy guarantee

Detailed Pipeline: Decomposition

- Our decomposition step relies on having an initial tree to split on in the first place
  - But the goal of the method is to find a tree on the full dataset
  - We need a very very quick method that can find a tree on the full dataset
- CN3 gives the minimum event distance between two copy number profiles
  - This allows us to create a distance matrix of minimum event distance for every pair of input copy number profiles
  - We can use this distance matrix to build a tree using a distance method (e.g. PAUP*'s NJ, FastME's BME, FastME's BioNJ)
- In our pipeline, we used the BioNJ tree from FastME
- We split the initial tree (BioNJ tree) at the centroid edge to create two subtrees

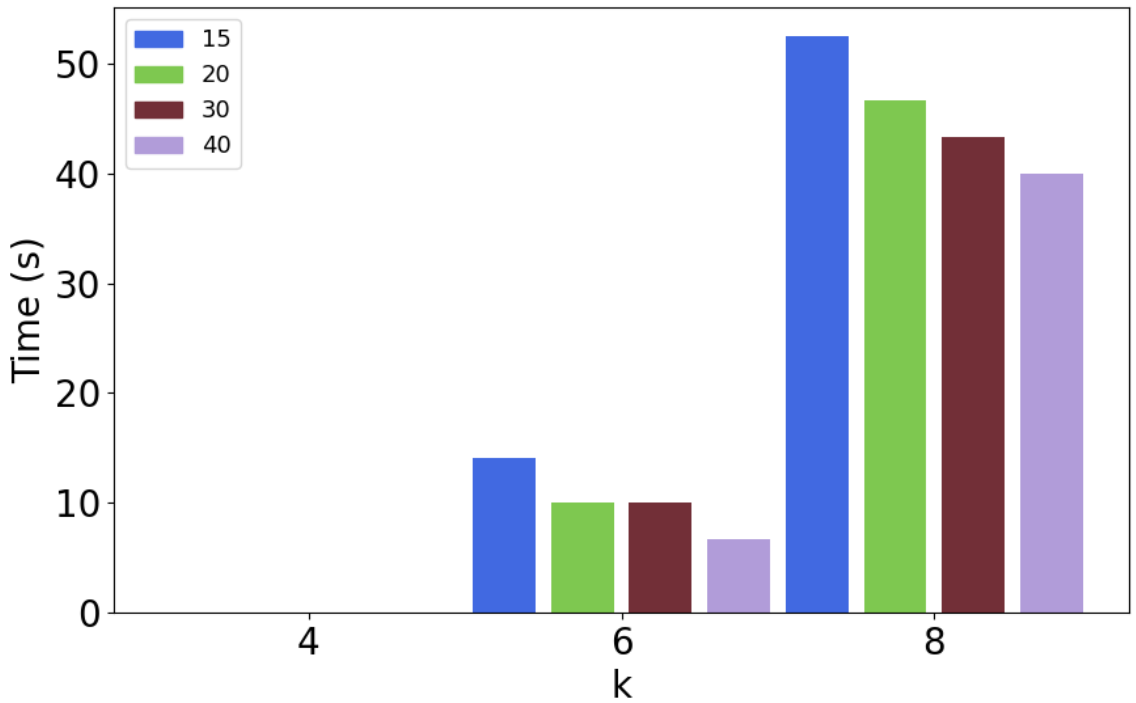
Detailed Pipeline: Constraint Computation

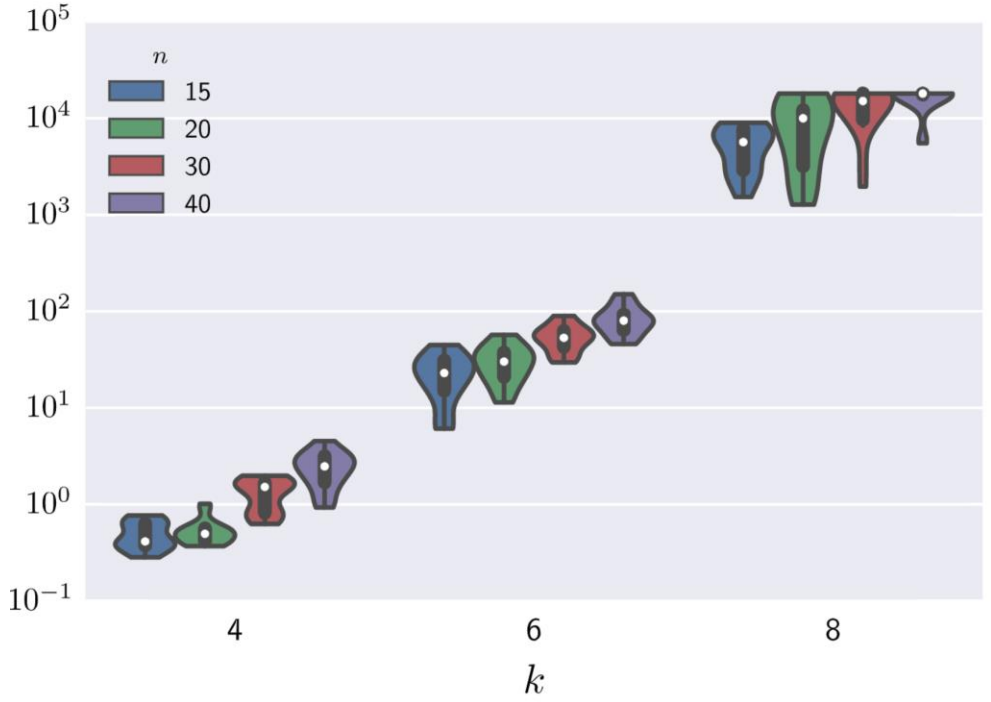
- This step is simply running CNT-ILP on the subset of taxa that is in each subtree

Detailed Pipeline: Merger

- This step connects the two CNT-ILP outputs by joining them at their roots and running CN3 DP algorithm to find the labeling at the root

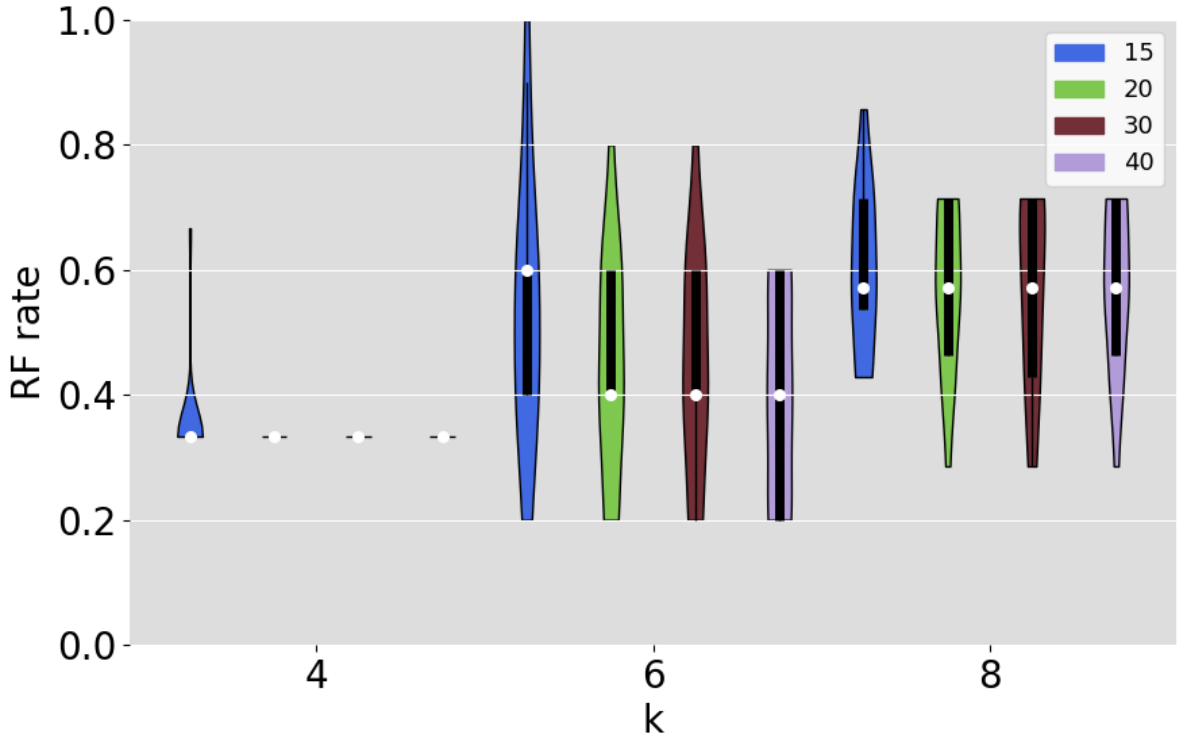
Results: Runtime

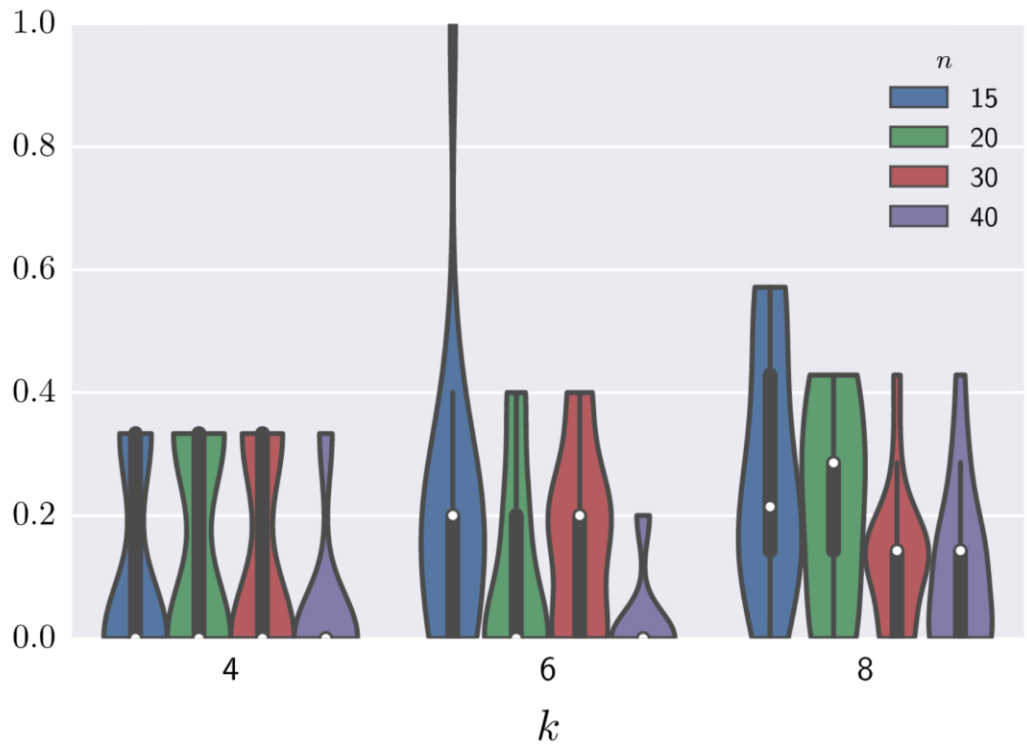




- Left: runtime output of divide-and-conquer pipeline; right: runtime plot from the original paper that introduced CNT-ILP
- K = number of input copy number profiles; [15,20,30,40] = length of the copy number profiles
- Runtime benefits are clear. This was expected

Results: RF rate





- Left: RF rate output of divide-and-conquer pipeline; right: runtime plot from the original paper that introduced CNT-ILP
- K = number of input copy number profiles; [15,20,30,40] = length of the copy number profiles
- Topological accuracy suffers, 20 to 40 percentage point difference across model conditions

Discussion

- Runtime, which is the objective we wished to improve on, was satisfyingly reduced.
  - Divide-and-conquer pipeline finishes in less than a minute on the hardest model condition
  - CNT-ILP takes more than 2 and a half hours
- However, the accuracy trade-off was also drastic
- Ideally, we would like to see the accuracy to stay competitive with the original approach while being faster

Future Direction

- Support more than two subsets or two subsets of different sizes
  - Allow for greater scalability and potentially accuracy improvements
- Better constraint method
  - A method more accurate than CNT-ILP but one that could not be used due to extreme computational requirements may be of use here
- Improving the merger
  - Merging while letting the constraint trees blend may be beneficial for the final topological accuracy
- Different ways of using the CN3 distance matrix
  - Minimum spanning tree
  - Simply use the FastME BioNJ tree

Grainger College of Engineering

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN