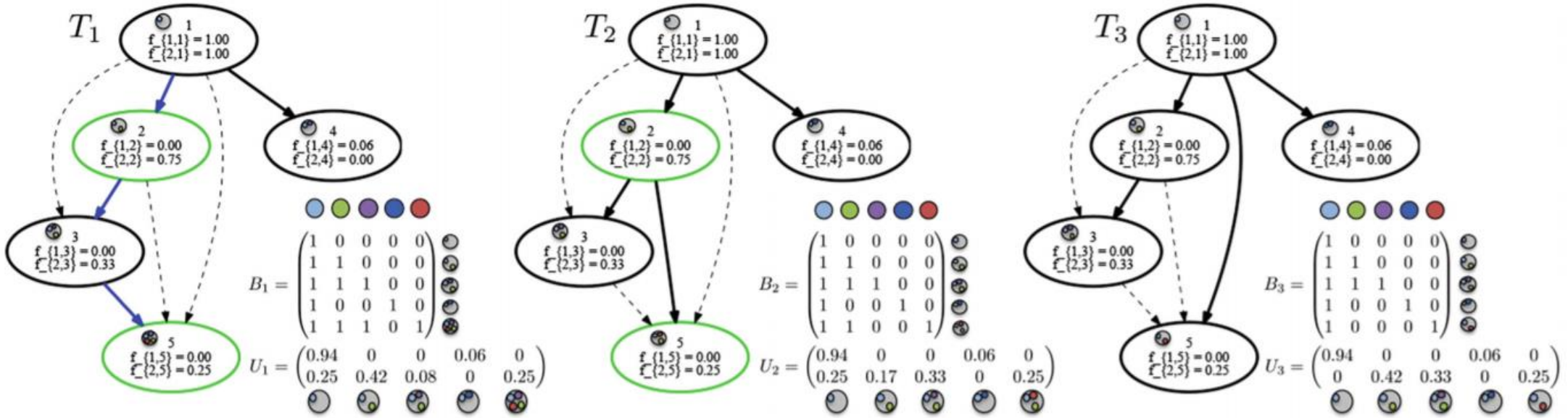
Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem.

Perfect Phylogeny Mixture (PPM) Problem

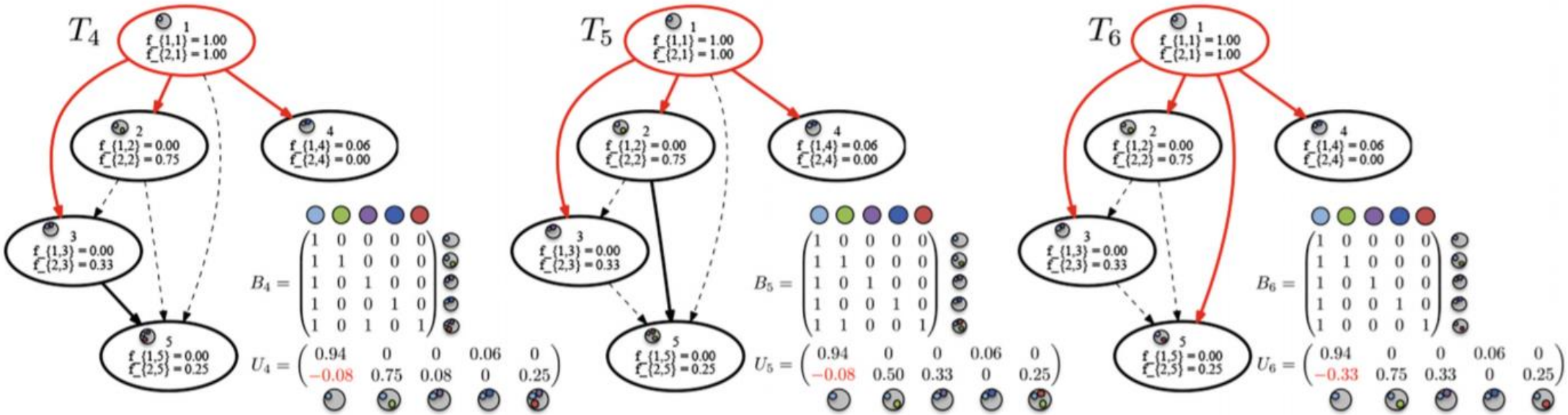
- Input: an $m \times n$ frequency matrix F
  - m: number of tumor samples
  - n: number of mutations (mutation clusters)
  - F: frequencies of variant reads
- Output: a binary $n \times n$ matrix B
  - $F = UB$
  - Each row of B represent a clone
  - Each B correspond to a phylogenetic tree
  - Infinite Site Assumption
- Deciding existence of a solution is NP-complete

Ancestry Graph

- A graphic view of PPM problem



$$F = \begin{pmatrix} 1 & 0 & 0 & 0.06 & 0 \\ 1 & 0.75 & 0.33 & 0 & 0.25 \end{pmatrix}$$



Sample or Counting the solutions of PPM

- #P: a complexity class: Counting version of NP problems
  - Example:
    - SAT: deciding existence of an assignment of all variables such that given CNF is true
    - #SAT: count the number of assignments of all variables such that given CNF is true
- #P-complete: the set of hardest problem in #P
- Proof of #P-completeness: Parsimonious reduction
  - Reduction need to preserve the number of solutions

Mono 1-in-3Sat

- Input: a Boolean formula in conjunctive normal form (CNF) where each clause has exactly three positive or three negative literals.
- Output: existence of a satisfying assignment with exactly one true literal in each clause
- Counting version proven to be #P-complete problem by Creignou, N., & Hermann, M. (1993)

#PPM is #P-complete

- Parsimonious reduction from #Mono 1-in-3Sat to #SubsetSum
- Parsimonious reduction from #SubsetSum to #PPM
- A special version of Subset sum is used here:
  - The given set of numbers are required to be unique and positive
- This statement holds even for m=2

Reduction from #Mono 1-in-3Sat to #SubsetSum

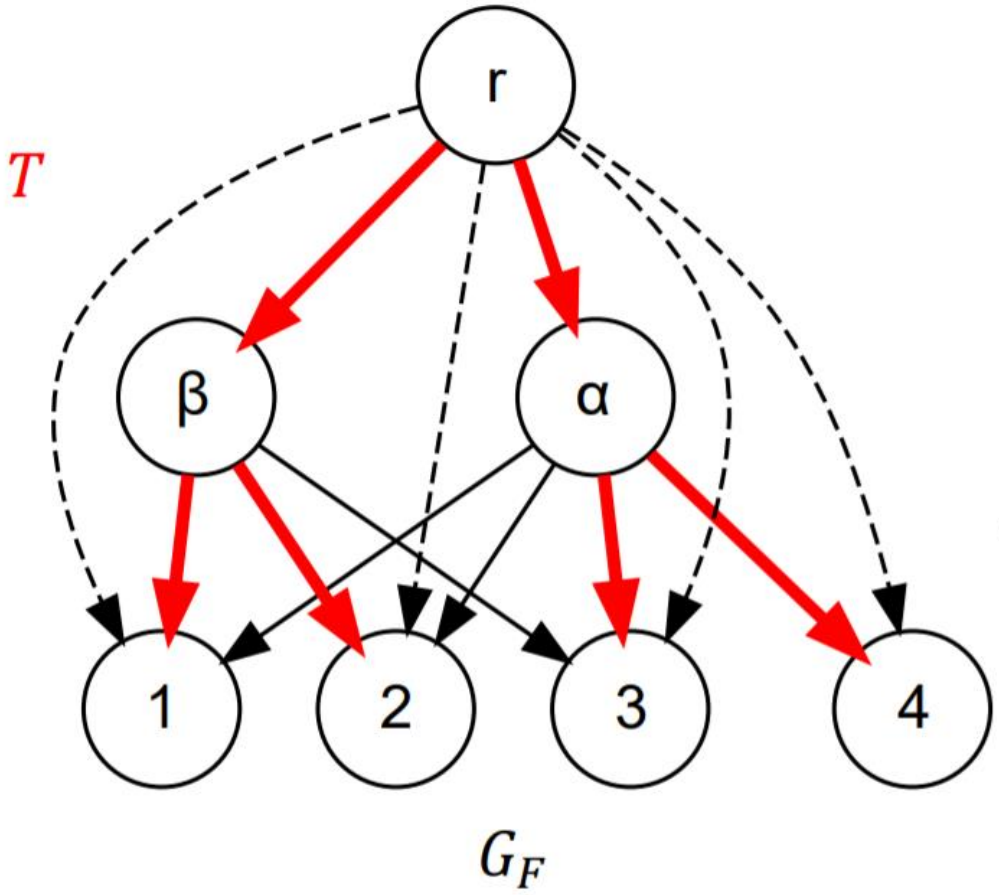
$$\phi = (x_1 \vee x_2 \vee x_3) \wedge (\neg x_2 \vee \neg x_3 \vee \neg x_4),$$

$$\theta = \{x_1 = false\ , x_2 = true, x_3 = false,\ x_4 = true\},$$

$$N = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, M = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} Q_2 \\ Q_3 \\ Q_6 \\ Q_7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Reduction from #SubsetSum to #PPM



$$S = \{1, 2, 3, 4\},\ t = 7,\ D = \{3, 4\}$$

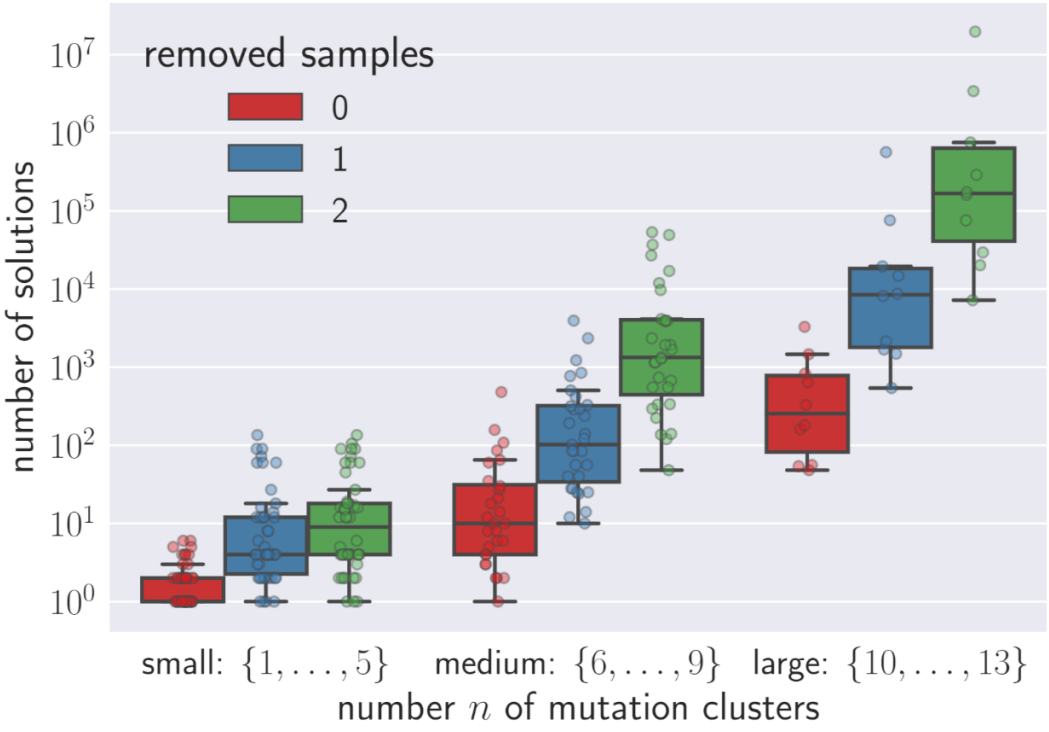
$$F = \frac{1}{10}\begin{pmatrix} 10 & 7 & 3 & 1 & 2 & 3 & 4 \\ 10 & 3 & 7 & 4\epsilon & 3\epsilon & 2\epsilon & \epsilon \end{pmatrix}$$

Approximation

- There is no fully polynomial randomized approximation scheme (FPRAS) for #PPM unless RP=NP
  - FPRAS: a ε-approximate algorithm that runs in polynomial in size of input (m, n) and ε^{-1}
- There is no fully-polynomial almost uniform sampler (FPAUS) for the solutions to PPM unless RP=NP
  - FPAUS: a sampler on a set S such that it returns each element in S with probalility in range $[\frac{1-\delta}{|S|}, \frac{1+\delta}{|S|}]$ that runs in polynomial in size of input (m, n) and δ^{-1}

TracerX: Lung Cancer

- Mutation clusters
- 90 % confidence interval



Following Work

- Consensus tree
- Develop a almost uniform sampler with SAT solver

Consensus Tree Problem

- Given a set of trees $\{T_1, T_2, \ldots, T_n\}$ on the same vertex set and a distance function $d(T, T')$, find a consensus tree T^* on the same vertex set, such that $\sum_{i=1}^{n} d(T^*, T_i)$ is minimized.
  - Here we will use parent-child distance
- A more general case:
  - Find multiple consensus trees instead of 1

Approximation to Consensus Tree Problem

- k-means
  - Starting with a k-partition of the trees
  - Find the consensus tree for each partition
  - Assign trees to one of the k consensus tree

Hardness of Consensus Tree Problem

- $k = 1$: solvable in polynomial time
- $k \geq 2$: expected to be NP-Hard
- Develop a exact solver with ILP

References

- Creignou, N., & Hermann, M. (1993). *On #P completeness of some counting problems* (Doctoral dissertation, INRIA).
- Govek, K., Sikes, C., & Oesper, L. (2018, August). A Consensus Approach to Infer Tumor Evolutionary Histories. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*(pp. 63-72). ACM.
- Jerrum, M. (2003). *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media.
- Pradhan, D., & El-Kebir, M. (2018, October). On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem. In *RECOMB International conference on Comparative Genomics* (pp. 277-293). Springer, Cham.