Extending the Variant Allele Frequency Factorization Problem (VAFFP) to Multi-Patient Data

By: Peyton Wiecking

Course: CS 5124 Algorithms in Bioinformatics – Spring 2025

Instructor: Dr. Palash Sashittal

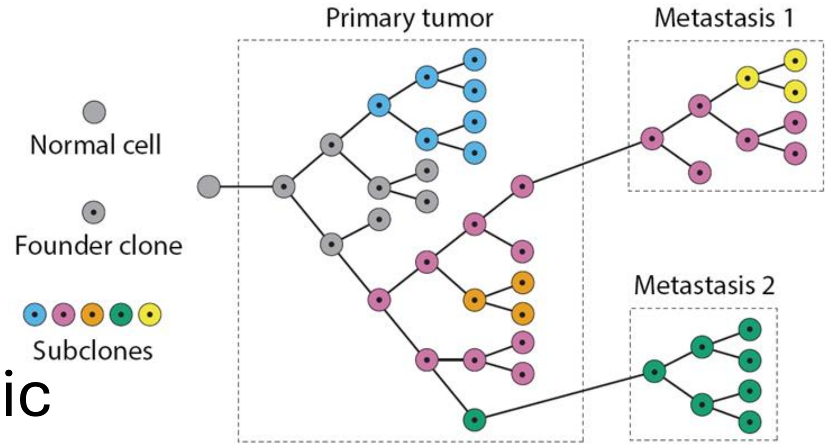
Date: 01MAY25

Overview

- Background
- Motivation: Why Model Conserved Evolution Across Patients?
- VAFFP: Single-Patient Formulation
- Problem Statement: Multi-Patient VAFFP Extension
- Comparative Analysis
    - TreeMHN
    - Sapling
    - MASTRO
- Results
- Conclusion
- Image References and Paper References

NOTE: All images used in this presentation are the property of their respective sources and are cited at the end of the presentation. No ownership is claimed.

Background: Tumor Evolution



- Cancer arises through the accumulation of somatic mutations in a single founder cell [1]*
- Tumor growth leads to the formation of genetically distinct subclones, resulting in intra-tumor heterogeneity [1, 2]
- Understanding the evolutionary relationships among subclones helps interpret tumor progression, resistance, and treatment response [1]
- These relationships are typically represented as rooted phylogenetic trees, where:
    - Nodes represent subclones or sets of mutations
    - Edges represent ancestral lineage.

*NOTE: References are included at the end of the Presentation

Background: Variant Allele Frequency Factorization Problem (VAFFP)

VAFFP models tumor evolution from bulk sequencing data for a single patient [3]

- Input: Mutation frequency matrix $F \in R^{m \times n}$
  - *m:* tumor samples
  - *n:* somatic mutations
- Goal: Factorize $F \approx U \times B$
  - $U \in R^{m \times k}$: usage matrix (proportion of clones per sample)
  - $B \in \{0,1\}^{k \times n}$: binary mutation matrix (which clone has which mutations)

Visual Example on Next Slide!

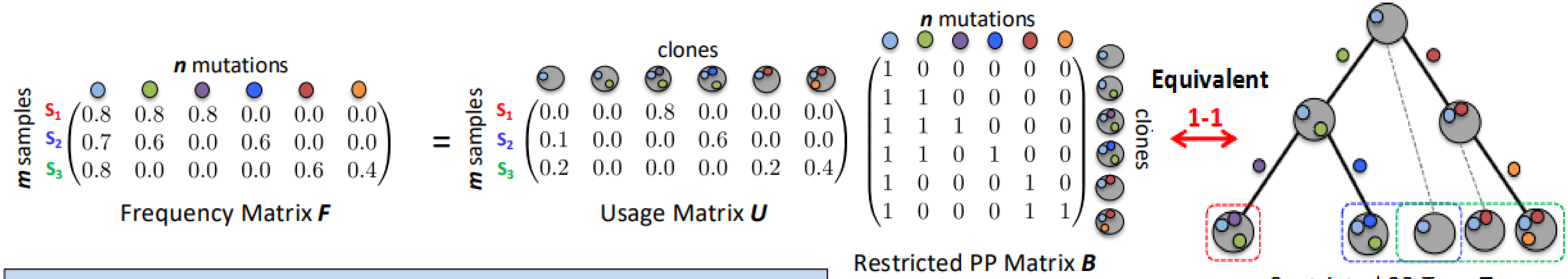
- Constraint: Perfect Phylogeny
  - **Formal:** A binary matrix B admits a perfect phylogeny if, for every pair of mutations, the subclones carrying them are either nested or disjoint (does not contain: (1,1),(1,0),(0,1)) [7]
  - **Simple:** Each mutation appears once and is inherited by all descendants
- Constraint: Sum Condition
  - **Formal:** $F_{pj} = \sum_k U_{pk} B_{kj}$
  - **Simple:** The observed frequency of a mutation equals the sum of contributions from all subclones that carry it [3].
- Constraint: Ancestry Condition
  - **Formal:** If mutation *i* is ancestral to j, $F_{pj} \geq F_{pk}$ for all *p*
  - **Simple:** A mutation seen earlier in evolution should never be less frequent than one of its descendants [3]

Assumptions:

- Infinite sites assumption: a character changes state once
- **Error-free data**

Sum Condition:

Given ***F*** and ***T***, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$



$$\begin{array}{c|cccccc} & & & & & & \\ S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array} = \begin{array}{c|cccccc} & & & & & & \\ S_1 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Frequency Matrix ***F***

Usage Matrix ***U***

Restricted PP Matrix ***B***

Restricted PP Tree ***T***

Theorem 1:

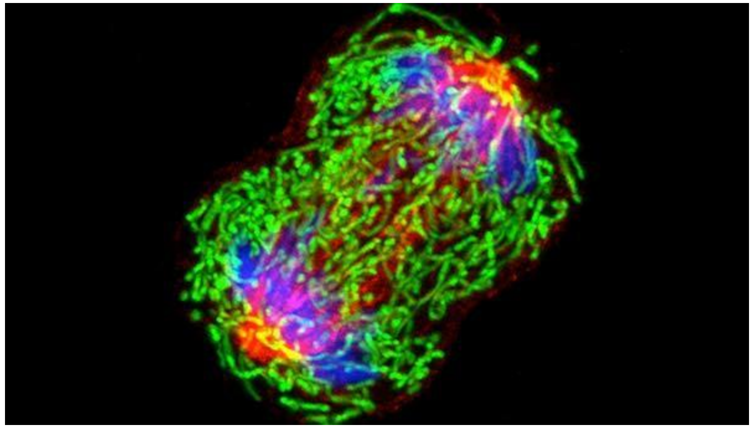
T is a solution to the VAFFP if and only if ***T*** is a spanning tree of ***G*** satisfying the sum condition

Theorem 2:

VAFFP is NP-complete

Motivation for Multi-Patient Modeling

- Certain mutational trajectories recur across patients with the same cancer type [1, 2]
- Existing models (e.g., VAFFP, Sapling, TreeMHN) are designed for single-patient analysis.
- Key biological questions remain unanswered:
  - Are tumor evolutionary paths conserved across individuals?
  - Can we detect shared early drivers or mutation sequences?
- Modeling across patients allows us to:
  - Identify common progression patterns
  - Uncover population-level therapeutic targets
- Requires a new formulation that:
  - Jointly models multiple patients
  - Preserves perfect phylogeny and other biological constraints
  - Operates on raw frequency data, not just inferred trees



Proposed Extension: Multi-Patient VAFFP

- Goal: Capture shared evolutionary structures across multiple patients
- Key Idea:
  - Infer a shared binary mutation matrix B across patients
  - Allow patient-specific usage matrices $U^{(i)}$
- Preserve biological constraints:
  - Perfect phylogeny
  - Sum condition
  - Ancestry condition
- Input: Patient-specific mutation frequency matrices $F^{(i)}$

Problem Statement (VAFFP on Multiple Patients)

Given a set of mutation frequency matrices $F^{(i)} \in R^{m_i \times n}$ *for multiple patients* $i = 1, 2, \ldots, N$ *where* m_i*is the number of samples for patient* i *and* n *is the number of mutations, find a shared binary matrix* $B \in \{0,1\}^{k \times n}$ *representing the presence or absence of mutations in inferred subclones across patients, and a set of patient-specific usage matrices* $U^{(i)} \in R^{m_i \times k}$ *representing the proportions of different tumor clones in each sample, such that* $F^{(i)} \approx B \times U^{(i)}$ *for all patients and that the matrices satisfy the perfect phylogeny, sum, and ancestry constraints.*

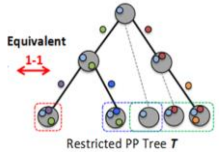
$$F^{(i)} \approx B \times U^{(i)}$$

n mutations

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}\right)$$

Frequency Matrix ***F***



clones

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{array}\right)$$

Usage Matrix ***U***

n mutations

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}\right)$$

Frequency Matrix ***F***

$$\approx$$

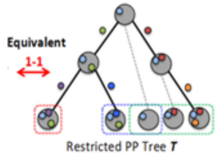
n mutations

$$\left(\begin{array}{cccccc|c} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & \bullet \\ 1 & 1 & 0 & 0 & 0 & 0 & \bullet \\ 1 & 1 & 1 & 0 & 0 & 0 & \bullet \\ 1 & 1 & 0 & 1 & 0 & 0 & \bullet \\ 1 & 0 & 0 & 0 & 1 & 0 & \bullet \\ 1 & 0 & 0 & 0 & 1 & 1 & \bullet \end{array}\right)$$

clones

Restricted PP Matrix ***B***

$$\times$$



clones

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{array}\right)$$

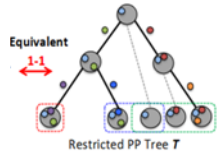
Usage Matrix ***U***

n mutations

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}\right)$$

Frequency Matrix ***F***



clones

m samples

$$\left(\begin{array}{c|cccccc} & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline S_1 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{array}\right)$$

Usage Matrix ***U***

NOTE: These matrices are for example purposes only. The number are not necessarily correct

Comparisons to Existing Methods

- How does the proposed multi-patient VAFFP formulation differ from existing methods?
- Compare against:
  - TreeMHN – focuses on mutual hazard dependencies between mutations
  - Sapling – summarizes single-patient uncertainty via backbone trees
  - MASTRO – mines shared mutation orders across trees
- For each method, the following will be addressed:
  - A brief conceptual overview
  - The problem statement
  - Visual representation
  - A side-by-side comparison to multi-patient VAFFP

TreeMHN – Introduction

- TreeMHN models cancer progression using pre-inferred tumor mutation trees [2]
- It constructs a Mutual Hazard Network (MHN):
  - A weighted graph representing how mutations influence the likelihood of acquiring others
- Captures:
  - Co-occurrence: some mutations tend to appear together
  - Mutual exclusivity: some mutations rarely co-occur
- Output: a network (rate matrix Θ) estimating inter-mutation influence
- Learns from population-level patterns using maximum likelihood estimation

TreeMHN – Problem Statement

Given a set of rooted tumor mutation trees $T = \{T_1, \ldots, T_N\}$, *where each tree* T_i *represents the evolutionary relationships among mutations in a patient's tumor, the goal of TreeMHN is to infer a Mutual Hazard Network characterized by a rate matrix* $\theta \in \mathbb{R}^{n \times n}$. *Each entry* θ_{ij} *quantifies the influence of mutation* i *on the hazard rate of acquiring mutation* j. *The objective is to estimate the mutual hazard rates by maximizing the regularized likelihood of the observed tumor trees under a probabilistic model of mutation accumulation* [2]

Simple Objective: Estimate Θ by maximizing the regularized likelihood of the observed trees under a probabilistic generative model [2]

TreeMHN – Visual Diagram [2]

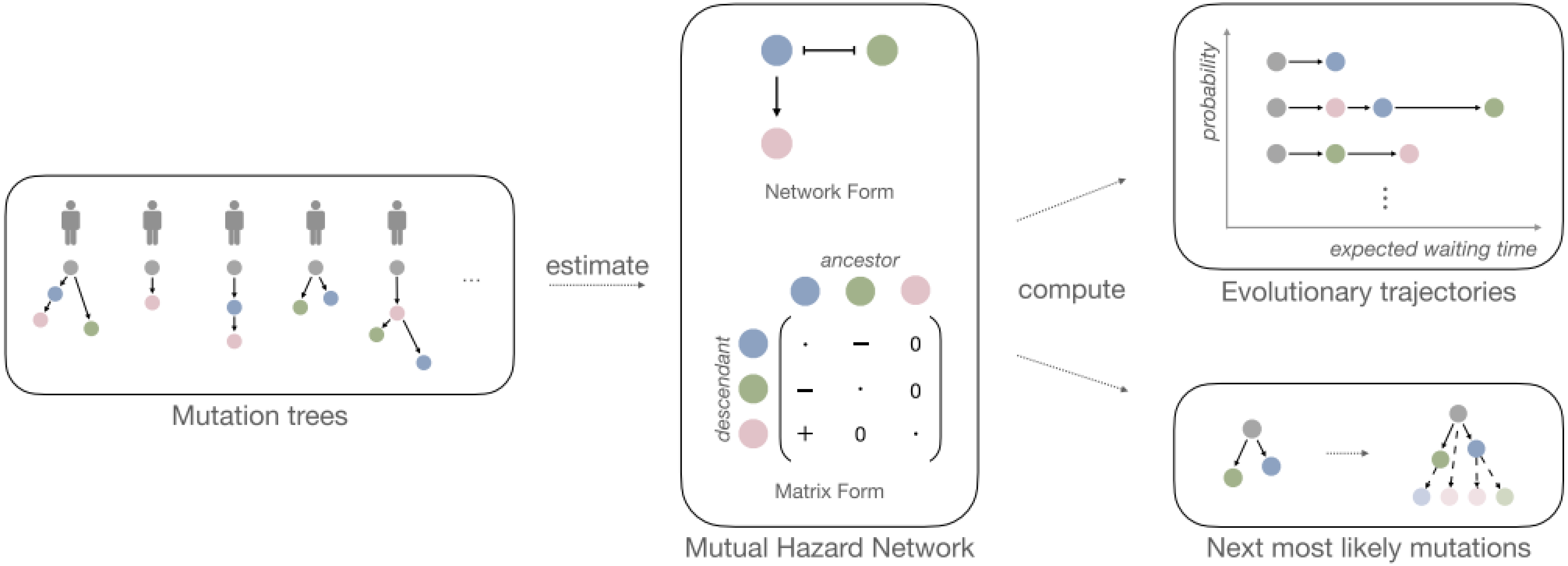


Fig. 1 | Inference with TreeMHN. The middle panel is a Mutual Hazard Network (MHN) for three distinct mutations (colored differently) represented as a network and equivalently as a matrix. The edges represent the co-occurring (→) or exclusive (⊣) stochastic dependencies among the mutations, corresponding to positive and negative off-diagonal entries in the matrix respectively. The diagonal entries are the baseline rates of fixation and have no influence on the network structure (therefore depicted as dots). Given a set of heterogeneous tumor mutation trees (left panel), we estimate the dependency parameters of an unknown underlying MHN. From the estimated MHN, we can compute the probability and the expected waiting time of any evolutionary trajectory (upper right panel). Additionally, we can compute the most probable next mutational events given an existing tree (lower right panel).

TreeMHN vs Multi-Patient VAFFP

Aspect	TreeMHN [2]	Multi-Patient VAFFP
Input	Pre-inferred mutation trees (1 per patient)	Raw mutation frequency matrices $F^{(i)} \in R^{m_i \times n}$ from multiple patients
Model	Learns a Mutual Hazard Network modeling inter-mutation influence	Jointly factorizes patient matrices using a shared *B* and per-patient $U^{(i)}$
Output	Rate matrix $\theta \in \mathbb{R}^{n \times n}$	Shared binary mutation matrix *B*; usage matrices $U^{(i)}$
Constraints	No perfect phylogeny; allows mutation loss and recurrence	Enforces perfect phylogeny, sum condition, and ancestry condition [3, 7]
Complexity	No NP-hardness or runtime bounds stated in the paper	NP-complete [3]
Data Layer	Operates on inferred tree structures	Operates directly on raw sequencing data
Scope	Designed for population-level trend discovery across inferred patient-specific trees; focuses on inter-mutation relationships	Designed for multi-patient modeling of conserved evolutionary trajectories

Sapling – Introduction

- Sapling models uncertainty in tumor evolution based on bulk sequencing read counts [5]
- Rather than outputting a single tree, it samples a large number of plausible evolutionary trees per patient
- It then constructs a backbone tree
    - A summary that shows recurring ancestral relationships among subclones
- Designed for single-patient analysis, handling noisy or ambiguous data
- Useful for understanding intra-patient variation, but does not model across patients

Sapling – Problem Statement

Given read count matrices $A \in \mathbb{N}^{m \times n}$ and $D \in \mathbb{N}^{m \times n}$ for variant and total reads, respectively, where m is the number of tumor samples and n is the number of mutations, the goal of Sapling is to infer a set of backbone trees that summarize the solution space $T(\rho)$ of plausible tumor evolutionary trees with likelihood above a threshold ρ. The problem consists of two main tasks: (1) Backbone Tree Inference from Reads, where the objective is to find a minimal set of representative trees summarizing likely mutation orders, and (2) Backbone Tree Expansion, where a backbone tree is expanded to include all observed mutations [5].

Simple Objective(s):

Backbone Tree Inference: Find a minimal set of backbone trees summarizing the space of high-likelihood trees above threshold ρ

Backbone Tree Expansion: Extend a backbone tree to include all observed mutations [5].

Sapling – Visual Diagram [5]

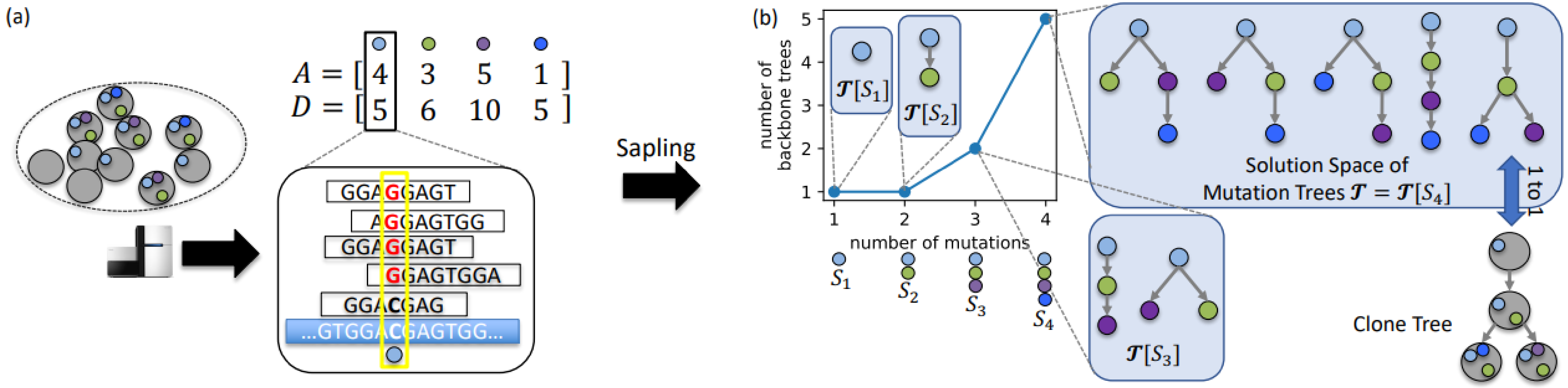


Fig. 1. Overview of Sapling. (a) Bulk DNA sequencing, alignment and SNV calling results in matrices A and D of variant and total read counts of n SNVs in m samples. (b) Sapling is a heuristic for the Backbone Tree Inference from Reads problem, returning a small set of backbone trees for a given number ℓ of mutations. Here, with $\ell = 3$ mutations, the solution space $\mathcal{T}$ of 5 mutation trees can be summarized with two backbone trees $\mathcal{T}[S_3]$.

Sapling vs Multi-Patient VAFFP

Aspect	Sapling [5]	Multi-Patient VAFFP
Input	Read count matrices: $A, D \in \mathbb{N}^{m \times n}$, (variant and total reads) from a single patient	Raw mutation frequency matrices $F^{(i)} \in R^{m_i \times n}$ from multiple patients
Model	Explores the space of phylogenetic trees consistent with read-level data	Jointly factorizes patient matrices using a shared *B* and per-patient $U^{(i)}$
Output	Set of backbone trees summarizing high-likelihood mutation orders	Shared binary mutation matrix *B*; usage matrices $U^{(i)}$
Constraints	Tree structure enforced locally by thresholded likelihood; no global constraint	Enforces perfect phylogeny, sum condition, and ancestry condition [3, 7]
Complexity	Proven NP-hard for both backbone tree inference and expansion	NP-complete [3]
Data Layer	Operates directly on read-level data (raw sequencing input)	Operates directly on raw sequencing data
Scope	Designed for single-patient tumor reconstruction under tree uncertainty	Designed for multi-patient modeling of conserved evolutionary trajectories

MASTRO – Introduction

- MASTRO (MAximal tumor treeS TRajectOries) identifies mutation trajectories that are frequently observed across multiple tumor phylogenies [6]
- It takes as input a set of previously inferred rooted mutation trees, one per patient
- A mutation trajectory is defined as a partially ordered set of mutations consistent with a subset of trees
- MASTRO outputs maximal trajectories: sets of mutations that appear together in a certain order in at least r input trees
- Focuses on mining recurrence patterns, rather than reconstructing trees or subclonal compositions.

MASTRO – Problem Statement

Given a multiset of rooted tumor phylogenies $D = \{T_1, \ldots, T_n\}$*, where each tree* T_i *has nodes labeled by mutations, and a minimum support threshold* $r \in \mathbb{N}$*, the goal of MASTRO is to identify all maximal mutation trajectories that are observed in at least* r *tumor trees. A maximal trajectory is defined as a set of mutations partially ordered by ancestral relationships, consistent with the input trees, and not properly contained within any other trajectory satisfying the support threshold* [6].

Simple Objective:

To identify all maximal mutation trajectories observed in at least r tumor trees [6]

MASTRO - Visual Diagram [6]

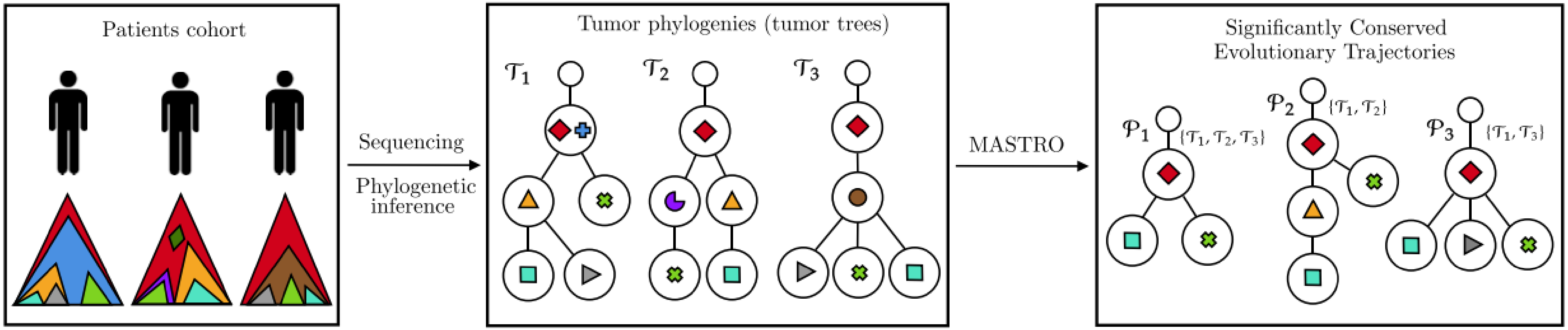

Patients cohort
Sequencing
Phylogenetic inference
Tumor phylogenies (tumor trees)
$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_3$
MASTRO
Significantly Conserved Evolutionary Trajectories
$\mathcal{P}_1$
$\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$
$\mathcal{P}_2$
$\{\mathcal{T}_1, \mathcal{T}_2\}$
$\mathcal{P}_3$
$\{\mathcal{T}_1, \mathcal{T}_3\}$

Fig. 1. High-level description of MASTRO. Leveraging sequencing data from a cohort of patients and phylogenetic inference algorithms, it is possible to infer the tumor trees describing the clonal evolution of the tumors as the accumulation of different alterations (different alterations correspond to different colored shapes within the nodes of the trees). MASTRO identifies conserved evolutionary trajectories, describing complex interactions among alterations, that are frequently observed in the tumor trees. MASTRO assesses their statistical significance and provides sound control of false discoveries. In the example, we show three trajectories ($\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$) observed in at least two tumor trees (shown to the right of the root nodes).

MASTRO vs Multi-Patient VAFFP

Aspect	MASTRO [6]	Multi-Patient VAFFP
Input	A multiset of pre-inferred tumor phylogenies: $\{T_1, \dots, T_n\}$, one tree per patient	Raw mutation frequency matrices $F^{(i)} \in R^{m_i \times n}$ from multiple patients
Model	Mines frequently recurring mutation trajectories from input trees using support threshold *r*	Jointly factorizes patient matrices using a shared *B* and per-patient $U^{(i)}$
Output	Set of maximal mutation trajectories, partially ordered by ancestry	Shared binary mutation matrix *B*; usage matrices $U^{(i)}$
Constraints	Allows flexible trajectory structures; no perfect phylogeny enforced globally	Enforces perfect phylogeny, sum condition, and ancestry condition [3, 7]
Complexity	Proven NP-hard, even when the recurrence threshold equals the number of input trees	NP-complete [3]
Data Layer	Operates after tree reconstruction, using inferred evolutionary trees	Operates directly on raw sequencing data
Scope	Designed for identifying frequent substructures across patients' trees	Designed for multi-patient modeling of conserved evolutionary trajectories

Aspect	TreeMHN [2]	Sapling [5]	MASTRO [6]	Multi-Patient VAFFP
Input	Inferred tumor trees (one per patient)	Read count matrices (variant + total reads)	Inferred tumor trees (one per patient)	Mutation frequency matrices $F^{(i)} \in R^{m_i \times n}$
Model	Learns mutation dependencies via Mutual Hazard Network	Samples and summarizes tree space	Mines frequently recurring mutation trajectories	Joint matrix factorization with shared B, per-patient $U^{(i)}$
Output	Rate matrix Θ (mutation influence)	Set of backbone trees (single patient)	Maximal partial-order mutation trajectories	Shared mutation matrix B, usage matrices $U^{(i)}$
Constraints	No perfect phylogeny; supports mutual exclusivity	Local tree constraints only; no global structure enforced	No global constraints; allows flexible orderings	Enforces perfect phylogeny, sum, and ancestry conditions [3, 7]
Complexity	Not stated	NP-hard	NP-hard	NP-complete [3]
Data Layer	Operates on inferred trees	Operates on read-level data	Operates on inferred trees	Operates on mutation frequency matrices
Scope	Population-level mutation trends	Single-patient tree reconstruction with uncertainty	Multi-patient mutation pattern mining	Multi-patient modeling of conserved evolutionary trajectories

What Multi-Patient VAFFP Achieves (Results)

- Identifies conserved evolutionary trajectories shared across patients
- Maintains biological realism via perfect phylogeny, sum, and ancestry constraints
- Works directly on raw mutation frequency data, no tree inference required
- Captures both shared patterns and patient-specific subclonal diversity
- Offers a foundation for explaining repeatable cancer progression paths
- Conceptually distinct from TreeMHN, Sapling, and MASTRO

Conclusion

- Presented a conceptual extension of the Variant Allele Frequency Factorization Problem (VAFFP) to the multi-patient setting
- Proposed a model that jointly infers:
  - A shared binary mutation matrix across patients
  - Patient-specific subclone usage matrices
- Enforces key biological constraints: perfect phylogeny, sum, and ancestry
- Provides a theoretical framework for discovering conserved evolutionary trajectories across cancer patients
- Offers a principled alternative to models like TreeMHN, Sapling, and MASTRO, which do not integrate raw data across patients under evolutionary constraints
- Lays groundwork for future solver development and empirical benchmarking on real tumor sequencing data

Image References

- Slide 3 (Background): MedicalXpress
- Slide 5 (VAFFP Example): lecture12_bluksequencing2_annotated.pdf
- Slide 6 (Motivation): NIH – National Cancer Institute
- Slide 9 (Extended VAFFP Example): Modified from Slide 5
- Slide 13 (TreeMHN Diagram): [2]
- Slide 17 (Sapling Diagram): [5]
- Slide 21 (MASTRO Diagram): [6]
- Slides 14, 18, 22, and 23 (Comparison Charts): Developed by Presenter
- Slide 25 (Questions): Microsoft PowerPoint Stock Image

References

[1] P. C. Nowell, "The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression.," Science, vol. 194, no. 4260, pp. 23-28, 1 October 1976.

[2] X. G. Luo, J. Kuipers and N. Beerenwinkel, "Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees," Nature Communications, no. 1, pp. 1-14, 21 June 2023.

[3] M. El-Kebir, L. Oesper, H. Acheson-Field and B. J. Raphael, "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," Bioinformatics, vol. 12, no. i62–i70, p. 31, 10 June 2015.

[4] M. A. Myers, G. Satas and B. J. Raphael, "CALDER: Inferring phylogenetic trees from longitudinal tumor samples," Cell Systems, vol. 8, no. 6, pp. 514-522, 26 June 2019.

[5] Y. Qi and M. El-Kebir, "Sapling: Inferring and Summarizing Tumor Phylogenies from Bulk Data using Backbone Trees," bioRxiv (Preprint), pp. 1-7, 2022.

[6] L. Pellegrina and F. Vandin, "Discovering significant evolutionary trajectories in cancer phylogenies," Bioinformatics, vol. 38, no. Supplement 1, p. ii49–ii55, 16 September 2022.

[7] G. F. Estabrook, F. R. McMorris and C. A. Meacham, "Comparison of undirected phylogenetic trees based on subtrees," Mathematical Biosciences, vol. 23, no. 3, pp. 263-276, 1975.

[8] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," Networks, vol. 21, no. 1, pp. 19-28, 15 January 1991.

Questions