Journal of Mathematical Modeling
Vol. -, No. -, 2026, pp. -. Research Article

JMM

Integer linear programming for phylogenetic tree inference in multi-state mixtures

Razieh Enayatsefat[†], Mehri Bagherian[‡*], Hamidreza Vaziri[§]

[†]*Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran*
[‡]*Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran*
[§]*Department of Biology, Faculty of Sciences, University of Guilan, Rasht, Iran*

Email(s): raziyeh.enayat@gmail.com, mbagherian@guilan.ac.ir, vaziri@guilan.ac.ir

Abstract. The evolutionary history of somatic mutations within cells, which plays a key role in the development of cancer, can be represented using a phylogenetic tree. Studying the evolution of mutations is crucial for advancing cancer treatment. In this study, we address the problem of reconstructing phylogenetic trees from mixed populations of cells, modeled as the cladistic multi-state Perfect Phylogeny Mixture Deconvolution Problem (PPMDP). Given the frequency measurements of character states across bulk samples, the objective is to infer a tree that satisfies Infinite-Alleles Assumption (IAA). To this aim, we propose a Binary Linear Programming (BLP) formulation. The method incorporates both Single-Nucleotide Variations (SNVs) and Copy Number Variations (CNVs) as input data. We evaluate the approach using simulated and real data, and demonstrate that the BLP model reconstructs phylogenetic trees efficiently and accurately, showing high coincidence with the ground truth trees.

Keywords: Binary linear programming, multi-state perfect phylogeny, phylogenetic tree reconstruction, optimization, combinatorial algorithms.
AMS Subject Classification 90C10, 90C90:

1 Introduction

The evolution of cancer, is characterized by the accumulation of genetic alterations, including point mutations SNVs, CNVs, and various other structural changes in the genome. The SNVs involve genetic changes at a single genomic location. In contrast, genomes often contain multiple copies of certain genes, with these copy numbers are subject to alteration through duplication or deletion of genomic

*Corresponding author
Received: 28 March 2026/ Revised: 06 July 2026/ Accepted: 07 July 2026
DOI: 10.22124/jmm.2026.33204.3025

intervals. These mutations accumulate in the cell population over time [5]. Bulk tumor sequencing captures mutation information across a population of cells. Tumors are composed of a heterogeneous mixture of healthy and malignant cells, a phenomenon known as tumor heterogeneity. A **clone** refers to a subpopulation of cells that share an identical genomic profile. The CNVs are widespread in tumors and often overlap, making them difficult to analyze in isolation. As a result, tumor heterogeneity poses a significant challenge for accurate genomic interpretation [12, 16]. Accurate reconstruction of the phylogenetic tree is challenging due to the combinatorial nature of the problem and constraints such as the IAA, which requires that each mutation occurs at most once and is inherited by all descendants [16]. Two main approaches are commonly used for clonal reconstruction: probabilistic and combinatorial. On the other hand, these processes utilize data on either SNVs or CNVs [10, 15]. Reconstruction of phylogenetic trees from mixed cell populations can be formulated as the multi-state PPMDP. In this setting, the input consists of frequency measurements of character states across bulk samples, where each sample contains a mixture of clonal populations. Each clone carries distinct somatic mutations, including SNVs and CNVs [1, 5, 12]. Previous approaches proposed for solving multi-state PPMDP include enumerative combinatorial methods such as SPRUCE [18] and AncesTree [10], which generate compatibility graphs and identify valid trees. In fact, SPRUCE focuses on enumerating valid solutions and seeks to build a fully consistent structure. Our model introduces a BLP formulation for the multi-state PPMDP that directly captures the evolutionary constraints induced by both SNVs and CNVs. By encoding the Multi-State Ancestral Condition (MSAC) and the Multi-State Sum Condition (MSSC), it produces deterministic, reproducible, and mathematically rigorous solutions while integrating SNVs, CNVs, and cladistic constraints within a unified optimization framework. The model extracts the largest phylogenetically consistent structure from the data while identifying inconsistent events. Its output may be either a phylogenetic tree or a phylogenetic forest. When a forest is obtained, the largest connected component represents the dominant, well-supported phylogeny, whereas the remaining components correspond to poorly supported events, noisy observations, or mutations lacking sufficient evidence to be incorporated into the main evolutionary history. This behavior is consistent with real cancer sequencing data, where Variant Allele Frequency (VAF) measurements are often noisy, CNVs may be inaccurately estimated, and some mutations have low sequencing coverage. Rather than forcing every event into a single phylogeny, the proposed formulation preserves all available information by separating inconsistent events from the largest feasible evolutionary structure. This feature is advantageous for biological interpretation and downstream data analysis. Furthermore, the formulation is compatible with modern optimization solvers such as CPLEX and Gurobi, enabling efficient analysis of moderately sized datasets. Its modular design also facilitates the incorporation of additional data types and biological constraints, making it readily extensible to more complex settings, including multi-sample and longitudinal tumor phylogeny reconstruction. These characteristics make the proposed BLP formulation a flexible and effective framework for high-fidelity clonal reconstruction in cancer genomics [2, 3, 19]. The BLP model reconstructs phylogenetic trees representing the ancestral relationships between clones and is evaluated on both simulated and representative real datasets. Figure 1 illustrates the proposed pipeline. Program1 constructs the ancestry graph, and the resulting graph is subsequently used as the input to program2, which reconstructs the phylogenetic tree.

The rest of the paper is organized as follows. In Section 2, we discuss the PPMDP and in Section 3 we formulate it as a BLP. In Section 4, we implement BLP formulation on simulated and real data considering measurement errors to evaluate the designed method. We compare the results of BLP with the output of the SPRUCE algorithm. Finally, Section 5 is devoted to some concluding remarks and

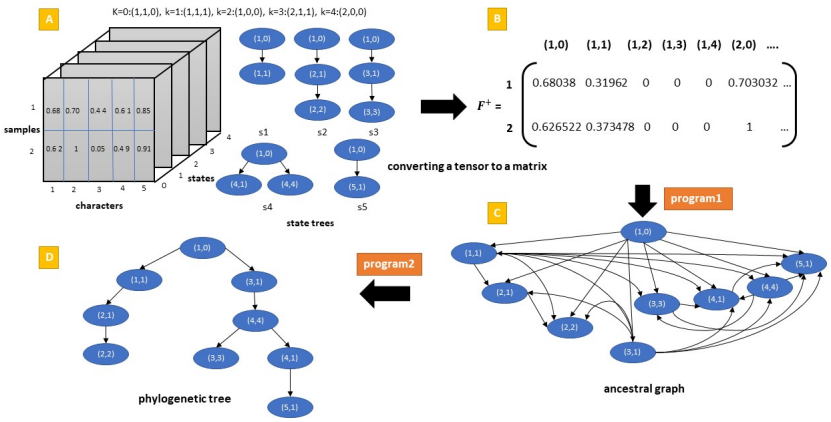

Figure 1: Overview of the proposed method
(A) Tensor $F_{m\times n\times k}$, where $m=2, n=5, k=5$, and state trees, we obtain the descendant set and, as a result, the cumulative frequencies of $f_p^+(D_{(c,i)})$, which are the components of the tensor F^+. (B) We convert tensor F^+ to matrix F^+. (C) We obtain the ancestral graph using the matrix F^+ and the MSAC. (D) Given tensor F and ancestral graph, we obtain the phylogenetic tree.

future research directions.

2 Character-based phylogeny mixture problems

A commonly used approach to illustrate the evolutionary trajectory of cancer is through phylogenetic trees. In such trees, the root represents the founder clone-corresponding to the earliest point in tumor evolution, while the leaves denote the present-day clones. These leaves correspond to advanced cancer cells, and the edges depict ancestral relationships among them [17]. In two-state phylogenetic reconstruction, where only SNVs are considered, taxa are typically described by binary characters, evolving under the Infinite Sites Assumption (ISA), which excludes homoplasy. The ISA posits that each genomic site can be mutated at most once during the evolutionary history of a cell population and that mutations, once acquired, are irreversibly inherited by all descendant cells without reversion or recurrence [8].

However, in the context of cancer, CNVs introduce more complex mutational patterns, where characters can exhibit more than two states. The multi-state PPMDP integrates both SNV and CNV data within a unified evolutionary framework. In this setting, taxa (i.e., the leaves of the phylogenetic tree) are characterized by multi-state characters that evolve under the IAA. The IAA postulates that each mutational event produces a novel allelic state that has not previously appeared in the population. As a result, no allelic state can arise independently more than once during the course of evolution. Under this assumption, a character may undergo multiple state transitions along the phylogenetic tree; however, transitions to any previously observed state (i.e., reversions or parallel recurrences of the same state) are prohibited. This constraint ensures that each allelic state corresponds to a unique mutational origin, thereby enforcing a perfect phylogenetic structure.

In the character-based mixture matrix, the input is not individual taxa or leaves but rather mixtures of

leaves. These mixtures are represented by a frequency matrix $F_{m\times n}$ in the two-state case and a frequency tensor $F_{m\times n\times k}$ in the multi-state case.

Definition 1 ([7]). *Matrix $F_{m\times n}$ is referred as a frequency matrix, if each component of F , denoted by f_{pi}, represents the observed VAF for mutation i in sample p such that $\sum_{i=1}^{n} f_{pi} \leq 1, f_{pi} \geq 0$.*

Definition 2 ([7]). *Matrix $U_{m\times n}$, is referred as a usage matrix if each component of it, denoted by u_{pi}, indicates the fraction of cells in sample p originating from clone v_i (note that $\sum_{i=1}^{n} u_{pi} \leq 1, u_{pi} \geq 0$).*

Definition 3 ([7]). *An n-clonal tree for a mutation set $[n] = \{1,\ldots,n\}$ is a rooted tree with n nodes, where each edge is labeled with a unique mutation, and no mutation emerges more than once.*

Under these assumptions, the ancestral relationships among clones in an evolving population can be modeled by a clonal tree, represented as a rooted directed tree. The root corresponds to the founding clone, while each internal or leaf node represents a descendant clone generated through the accumulation of one or more novel mutations. Edges in the clonal tree encode direct ancestral relationships between clones. Each edge is labeled by the mutation that differentiates the child clone from its parent. Consequently, the haplotype of a clone-defined as the set of mutations that uniquely characterizes it-is given by the collection of edge labels along the unique path from the root to the corresponding node. Leaf nodes represent clones observed at the terminal stage of the evolutionary process.

The clonal tree can be equivalently represented in matrix form. Let $B \in \{0,1\}^{n\times n}$ denote the binary mutation matrix, where the rows correspond to the n clones (i.e., tree nodes) and the columns correspond to the n mutation sites. The entry $b_{ij} = 1$ indicates that mutation j is present in clone i, whereas $b_{ij} = 0$ indicates its absence. Thus, each row of B encodes the haplotype associated with a specific clone in the clonal tree.

Proposition 1. *([3]) Matrix B (or tree T) yields matrix F if and only if there exists an usage matrix U such that $F = \frac{1}{2}U \times B$.*

In a two-state perfect phylogeny, the character c satisfies the ISA. Consequently, the following proposition is established.

Proposition 2. *([3]) If T yields F and $(c,1)$ is an ancestor of $(d,1)$ which is denoted as $(c,1) \prec_T (d,1)$, then all nodes in T with state 1 for character d must also have state 1 for character c.*

From Proposition 2, the following ancestral condition is implied [3].

Ancestral condition: If T yields F and $(c,1) \prec_T (d,1)$, then $f_{p,(c,1)} \geq f_{p,(d,1)}$ for all samples p and characters c and d. Under the ISA, $T_{(c,1)}$ (a subtree comprising all nodes with state 1 for character c) is equal to $\bar{T}_{(c,1)}$ (a subtree rooted at node $v_{(c,1)}$), which is expressed as

$$\bar{T}_{(c,1)} = v_{(c,1)} \cup \left(\cup_{(d,1)\in\delta(c,1)} \bar{T}_{(d,1)}\right), \tag{1}$$

where $\delta_{(c,1)}$ represents the children of $v_{(c,1)}$. From Equation (1), $f_{p,(c,1)} = \sum_{(d,1)\in T_{(c,1)}} u_{p,(d,1)}$, and the non-negativity of $u_{p,(c,1)}$, the sum condition follows.

Sum condition: If T yields F then $f_{p,(c,1)} \geq \sum_{(d,1)\in\delta_{(c,1)}} f_{p,(d,1)}$, for all samples p and characters c.

The ancestral condition is necessary, while the sum condition is both necessary and sufficient for solving the two-state perfect phylogeny mixture problem. The BLP model for this two-state problem has been proposed in [3].

The SPRUCE algorithm [4] addresses the multi-state PPMDP by integrating SNVs and CNVs data within a unified combinatorial framework. The model operates under the IAA, which formally requires that, for each character, every mutational event induces a novel state that has not previously appeared in the evolutionary history. Consequently, although multiple state transitions per character are allowed along a root-to-leaf path, transitions to any previously realized state are forbidden. This constraint enforces a directed acyclic structure consistent with perfect phylogeny. Let $[n] = \{1, \ldots, n\}$ denote the set of characters and let each character take values in a finite state set. The objective is to reconstruct a rooted tree T whose nodes correspond to clones and whose edge labels encode mutational state transitions, such that the observed frequency tensor F (derived from SNV and CNV measurements across samples) can be explained as a convex combination of clone genotypes consistent with T. The joint incorporation of SNV and CNV information introduces additional structural constraints linking mutation states across characters, thereby increasing the dimensionality of the feasible solution space and the combinatorial complexity of the reconstruction task. The multi-state PPMDP can therefore be formulated as the problem of determining a tree T and associated clone frequency variables that satisfy the IAA-induced phylogenetic constraints and reproduce the observed data F.

Problem 1 (PPMDP): Given an $m \times n \times k$ dimensional frequency tensor $F = [[f_{p,(c,i)}]]$, find an n,k-complete perfect phylogeny tree $T_{n,k}$ and an m,n,k-usage matrix $U = [u_{p,(c,i)}]$ such that $f_{p,(c,i)} = \sum_{(d,j)\in T_{(c,i)}} u_{p,(d,j)}$ holds for all character-state pairs (c,i) and all samples p.

Note that $f_{p,(c,i)}$ is the frequency of the character-state pair (c,i) in sample p, where $\sum_{i=0}^{k-1} f_{p,(c,i)} = 1, f_{p,(c,i)} \geq 0$, and $u_{p,(c,i)}$ represents the proportion of taxa $v_{(c,i)}$ in sample p, with $\sum_{c=1}^{n}\sum_{i=0}^{k-1} u_{p,(c,i)} = 1, u_{p,(c,i)} \geq 0$.

Definition 4. *An n,k-complete perfect phylogeny tree is a rooted tree with $n(k-1)+1$ nodes, where each of the $n(k-1)$ edges is labeled with a unique character-state pair of $[n] \times [k-1]$ where $[n] = \{1, \ldots, n\}$ no character-state pair emerges more than once.*

Proposition 3. *([4]) A tree T yields F if there exists a usage matrix U such that $f_{p,(c,i)} = \sum_{(d,j)\in T_{(c,i)}} u_{p,(d,j)}$.*

Under the IAA, each character c may attain multiple distinct states throughout evolution; however, every state is introduced exactly once in the phylogeny, and transitions to any previously realized state are prohibited. Let $T_{(c,i)}$ denote the subtree of the phylogenetic tree and comprises all nodes whose state for character c is i. $\bar{T}_{(c,i)}$ is defined as the disjoint union of the node $v_{(c,i)}$ and the subtrees rooted at each of its children. In general, $T_{(c,i)} \neq \bar{T}_{(c,i)}$, since the structure induced by state i depends not only on the node $v_{(c,i)}$ and its descendants, but also on its relative position with respect to other state nodes $v_{(c,j)}$, $j \neq i$ of the same character. Because multiple states of a character are organized along directed paths without recurrence, the subtrees $T_{(c,i)}$ and $T_{(c,j)}$ may exhibit hierarchical (ancestor-descendant) relationships or be disjoint, depending on their placement in the tree. To formally characterize the relationship between $T_{(c,i)}$ and $\bar{T}_{(c,i)}$, we introduce the notion of the descendant set of a character state, which captures all nodes reachable from $v_{(c,i)}$ via directed paths in the phylogenetic tree.

Definition 5. *Given a perfect phylogeny tree T, the descendant set for each character-state pair (c,i) is defined as $D_{(c,i)} = \{j | (c,i) \prec_T (c,j)\}$.*

Based on the definition of the descendant set, the descendant set $(c,0)$ (here, $(1,0)$ is considered), which represents the root of the tree T, is defined as $D_{(c,0)} = \{0,\ldots,k-1\}$, where k is the number of states that a character can take. Consequently, the descendant set establishes the relationship between $T_{(c,i)}$ and $\bar{T}_{(c,i)}$ as $\bar{T}_{(c,i)} = \cup_{l \in D_{(c,i)}} T_{(c,l)}$. Therefore, with the descendant set determined from T, the cumulative frequency of the descendant set for each character-state pair is calculated for each sample as

$$f_p^+(D_{(c,i)}) = \sum_{l \in D_{(c,i)}} f_{p,(c,l)}. \tag{2}$$

The components of the usage matrix are then derived from these cumulative frequencies as

$$u_{p,(c,i)} = f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}). \tag{3}$$

For T to yield F, $u_{p,(c,i)}$ must be non-negative. This condition is both necessary and sufficient to solve PPMDP. Consequently, the following theorem is proven.

Theorem 1. *(MSSC) ([4]) A complete perfect phylogeny tree T yields F if and only if*

$$f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \geq 0 \tag{4}$$

for all character-state pairs (c,i) and all samples p.

From Theorem 1, a necessary condition for the descendant sets $D_{(c,i)}$ and $D_{(d,j)}$ is derived.

Multi-state ancestral condition: Suppose (c,i) and (d,j) are distinct character-state pairs, and $D_{(c,i)}$, $D_{(d,j)} \subseteq \{0,\ldots,k-1\}$. A pair $(D_{(c,i)}, D_{(d,j)})$ is considered a valid descendant set pair if $f_p^+(D_{(c,i)}) - f_p^+(D_{(d,j)}) \geq 0$ for all samples of p. Furthermore, if $c = d$, then $D_{(c,i)} \subset D_{(d,j)}$.

The PPMDP can be reduced by assuming that, for each character $c \in [n]$ a corresponding state tree is given. This restricted formulation is referred to as the cladistic PPMDP, in which the evolutionary transitions among the states of each character are constrained to follow a predefined rooted tree structure. Formally, let S_c denote the state tree associated with character c. The feasible phylogenies are then required to be consistent with all state trees $\{S_c | c \in [n]\}$, thereby significantly reducing the combinatorial search space compared to the general multi-state PPMDP.

The cladistic PPMDP was previously addressed in [4] via the SPRUCE algorithm using a combinatorial reconstruction strategy. In this study we provide a formulation based on BLP, which models the phylogenetic consistency constraints and mixture deconvolution conditions within an optimization framework.

The state subtree specifies the relationship between character-state pairs that have the same character and different states. That is, in this way we can determine the generation set of each node. Therefore, from the tensor F and the state subtrees S_c, we obtain the tensor F^+ from (2). Then, by converting the tensor F^+ to the matrix F^+ and using the ancestral condition, the relationship between nodes in different trees is determined. In fact, the relationship between character-state pairs that do not have the same character is also determined, and thus the ancestral graph is obtained (program1). Then, we give the ancestral graph and the tensor F as input to our BLP model and obtain the desired phylogenetic tree (program2).

Definition 6. *$\mathscr{S} = \{S_c | c \in [n]\}$ denote the collection of state trees, where for each character c, the rooted tree S_c encodes the ancestral relationships among its states. The node set of S_c corresponds to the states of character c, and its directed edges represent immediate ancestor-descendant relations between states.*

For each character c and state i, define the descendant set $D_{(c,i)} = \{j | i \prec_{S_c} j\}$, where $\prec_{S_c}$ denotes the partial order induced by the rooted tree S_c. A rooted phylogeny T over the set of character-state pairs is said to be compatible with the state-tree collection $\mathscr{S}$ if, for every character c and states i, $(c,i) \prec_T (c,j)$ if and only if $i \prec_{S_c} j$. That is, the restriction of T to the states of each character c must preserve exactly the partial order defined by S_c.

Under this restriction, the multi-state ancestral graph reduces to a simple directed graph whose nodes are $v_{(c,i)}$, and whose edges $(v_{(c,i)}, v_{(d,j)})$ are labeled by the corresponding descendant sets $(D_{(c,i)}, D_{(d,j)})$. Moreover, the graph satisfies the ancestral consistency condition, namely that descendant relationships in T are consistent with the partial orders induced by all state trees in $\mathscr{S}$.

3 The BLP model

In this work, we develop a BLP model for the PPMDP. Proposition 3 shows that if there exists a set U inducing the decomposition $f_{p,(c,i)} = \sum_{(d,j)\in T_{(c,i)}} u_{p,(d,j)}$, then the resulting tree structure is consistent with the observed data. It is important to emphasize that $u_{p,(c,i)} = f_p^+(D_{(c,i)}) - \sum_{(d,j)\in\delta(c,i)} f^+(D_{(d,j)})$ is defined as an analytical device and is not part of the decision variables in the original optimization model.

Proposition 4. *([4]) A complete perfect phylogeny tree T yields F and is compatible with state trees $\mathscr{S}$ if and only if T is a spanning tree of the cladistic ancestry graph $G(F, \mathscr{S})$, where the MSSC is satisfied.*

In our formulation, we introduce binary decision variables $z_{(c,i),(d,j)} \in \{0,1\}$, for every ordered pair of distinct character–state nodes (c,i) and (d,j). The variable $z_{(c,i),(d,j)} = 1$ if and only if there exists a directed edge from $v_{(c,i)}$ to $v_{(d,j)}$ in the reconstructed phylogenetic tree T, representing a parent-child relationship; otherwise, $z_{(c,i),(d,j)} = 0$. The input data of the problem consist of the observed frequency tensor F and the collection of state trees $\mathscr{S} = \{S_c\}$. From the state trees, one can compute, for each node $v_{(c,i)}$, its descendant set $D_{(c,i)}$, and consequently determine the cumulative descendant frequencies in each sample by aggregating the frequencies over $D_{(c,i)}$. These quantities allow the construction of the associated ancestral graph, obtained by enforcing the ancestral consistency conditions induced by F.

Moreover, the state trees $\mathscr{S}$ uniquely determine the relative ancestral relationships among nodes corresponding to the same character but different states. Hence, for fixed c, the admissible edges between nodes $v_{(c,i)}$ and $v_{(c,j)}$ in the ancestral graph are completely specified by the partial order defined by S_c.

Inferring ancestral relationships: The ancestral relationships between nodes corresponding to different characters can be inferred from the cumulative frequencies computed from the state trees, together with the ancestral consistency condition.

For nodes corresponding to the same character, the state tree S_c uniquely determines the ancestral relationships among their states. Specifically, for each node $v_{(c,i)}$, its descendant set $D_{(c,i)}$ is defined as $D_{(c,i)} = \{j | i \prec_{S_c} j\}$. Using Equation (2), the cumulative frequency of the descendant set $f_p^+(D_{(c,i)})$ can be computed for each sample p. Once the cumulative frequencies are determined for all nodes, the ancestral condition can be applied to establish the parent-child relationships between nodes corresponding

to different characters. In particular, for two nodes $v_{(c,i)}$ and $v_{(d,j)}$ with $c \neq d$, the existence of a directed edge from $v_{(c,i)}$ to $v_{(d,j)}$ is inferred if and only if the cumulative frequencies are consistent with the ancestral condition across all samples. This ensures that the reconstructed multi-state phylogeny respects both the intra-character relationships from the state trees and the inter-character constraints dictated by the observed data.

In this manner, the ancestral graph is constructed. Let $a_{(c,i),(d,j)} \in \{0,1\}$ denote the binary parameter indicating the ancestral relationship between nodes $v_{(c,i)}$ and $v_{(d,j)}$. Specifically, $a_{(c,i),(d,j)} = 1$, if $v_{(c,i)}$ is an ancestor of $v_{(d,j)}$ and 0, otherwise. This parameter encodes the presence or absence of an ancestral relationship between the corresponding nodes in the multi-state phylogeny. If $c = d$, then the ancestral relationship is derived from the state tree. In a state tree S_c all vertices correspond to different states of character c. Therefore, whenever an edge $(v_{(c,i)}, v_{(c,j)}) \in S_c$ exists, state i of character c is an ancestor of state j, and consequently, $a_{(c,i),(c,j)} = 1$.

Definition 7. *Given an $m \times nk$ cumulative frequency matrix F^+, the ancestry graph $G = (V, A')$ is a directed graph with nodes $V = \{v_1, \ldots, v_{nk}\}$ corresponding to the set $\{v_{(1,0)}, v_{(1,1)}, \ldots, v_{(1,k-1)}, v_{(2,0)}, v_{(2,1)}, \ldots, v_{(2,k-1)}, \ldots, v_{(n,0)}, v_{(n,1)}, \ldots, v_{(n,k-1)}\}$ and arcs $A' = \{(v_{(c,i)}, v_{(c,j)}) | i \prec_{S_c} j, i \neq j, \forall S_c \in \mathscr{S}\} \cup \{(v_{(c,i)}, v_{(d,j)}) | f_p^+(D_{(c,i)}) \geq f_p^+(D_{(d,j)}), \forall c \neq d, \forall p \in [m] = \{1, \ldots, m\}\}$ where $A = \{(v_{(c,i)}, v_{(c,j)}) | i \prec_{S_c} j, \forall S_c \in \mathscr{S}\}$.*

Proposition 5. *Let $F \in \mathbb{R}_+^{m \times n \times k}$ is a matrix representing the frequencies of k states for n characters across m samples. Let $v_{(c,i)}$ denote the vertex associated with state i of character c, and let $D_{(c,i)}$ be its descendant set in the corresponding state tree S_c. The ancestral graph $G = (V, E)$ is constructed with vertex set $V = \{v_{(c,i)}\}$ and edge set $E = \{(v_{(c,i)}, v_{(d,j)})\}$ whenever the ancestry condition holds, namely when the cumulative frequency of $D_{(c,i)}$ dominates the cumulative frequency of $D_{(d,j)}$ in every sample. If all columns of the frequency matrix F^+ are distinct, this structure yields a directed acyclic graph (DAG), in which the edges encode all possible parent–child relations consistent with the observed frequencies and the constraints of the state trees. This DAG forms the basis of the BLP formulation of the cladistic PPMDP.*

Proof. By contradiction, suppose that there exists a directed cycle in the graph of the following form

$$(c_1, i_1) \to (c_2, i_2) \to \ldots \to (c_k, i_k) \to (c_1, i_1),$$

By the definition of an ancestral graph, each edge $(c_a, i_a) \to (c_b, i_b)$ implies that (c_a, i_a) is an ancestor of (c_b, i_b). Therefore, by MSAC, we have

$$f_p^+(D_{(c_a,i_a)}) \geq f_p^+(D_{(c_b,i_b)}), \quad \forall p.$$

That is, the cumulative frequency does not decrease along the edge. Applying this inequality to the entire circumference yields

$$f_p^+(D_{(c_1,i_1)}) \geq f_p^+(D_{(c_2,i_2)}) \geq \ldots \geq f_p^+(D_{(c_a,i_a)}).$$

Therefore, all these values must be identical.

$$f_p^+(D_{(c_1,i_1)}) = f_p^+(D_{(c_2,i_2)}) = \ldots = f_p^+(D_{(c_k,i_k)}).$$

But in the cladistic model, each edge represents a real ancestral relationship, and in a tree, for two distinct nodes, the relationship cannot be reciprocal. Perfect equality of frequencies for all samples is possible only when the two nodes are practically the same, not distinct. So the existence of a cycle is impossible. As a result, the ancestral relationship is a partial order, and any partial order can be represented as a DAG. □

If the columns of F^+ are not distinct, there must be an acyclic constraint in the model, but we can assume that identical frequency profiles are either absent or have been merged during preprocessing. This assumption is reasonable in practice, since exact equality of frequency vectors across all samples is uncommon in real sequencing datasets.

Theorem 2. *The multi-state PPMDP can be formulated as a BLP with the following constraints:*

1. *MSAC:*

$$\sum_{l\in D_{(c,i)}} a_{(c,i),(c,l)} f_{p,(c,l)} - \sum_{k\in D_{(d,j)}} a_{(d,j),(d,k)} f_{p,(d,k)} \geq 0, \qquad \forall v_{(c,i)}, v_{(d,j)}, \forall p.$$

2. *MSSC:*

$$\sum_{v_{(e,r)}\in\delta^-(v_{(c,i)})} \sum_{l\in D_{(c,i)}} a_{(c,i),(c,l)} z_{(e,r),(c,i)} f_{p,(c,l)} \geq$$
$$\sum_{v_{(d,j)}\in\delta^+(v_{(c,i)})} \sum_{k\in D_{(d,j)}} a_{(d,j),(d,k)} z_{(c,i),(d,j)} f_{p,(d,k)}, \qquad \forall v_{(c,i)}, p \in [m].$$

Here:

- $V = \{v_{(c,i)} | c \in [n], i \in [k]\}$ *is the set of nodes.*
- $z_{(c,i),(d,j)} \in \{0,1\}$ *is the BLP variable indicating the presence of edge* $(v_{(c,i)}, v_{(d,j)})$.
- $a_{(c,i),(d,j)} \in \{0,1\}$ *encodes ancestral relationships from the state trees.*
- $D_{(c,i)}$ *is the descendant set of node* $v_{(c,i)}$.
- $f_{p,(c,i)}$ *is the observed frequency of state i of character c in sample p, and* $f_p^+(D_{(c,i)}) = \sum_{l\in D_{(c,i)}} f_{p,(c,l)}$.
- *These constraints guarantee that the reconstructed tree is a valid multi-state perfect phylogeny, consistent with the ancestral relationships and the observed frequencies.*

Proof. To formulate the PPMDP as a BLP model, we introduce two sets, $\delta^-(v_{(c,i)})$ and $\delta^+(v_{(c,i)})$ as follows:

$$\delta^-(v_{(c,i)}) = \{v_{(d,j)} \in V | (v_{(d,j)}, v_{(c,i)}) \in A'\} \tag{5}$$

and

$$\delta^+(v_{(c,i)}) = \{v_{(d,j)} \in V | (v_{(c,i)}, v_{(d,j)}) \in A'\}. \tag{6}$$

Also, we define

$$D'_{(c,i)} = \{v_{(c,l)} | (v_{(c,i)}, v_{(c,l)}) \in A\}.$$

According to the MSAC and definition of cumulative frequency (2), we rewrite the MSAC as (7)

$$\sum_{l \in D_{(c,i)}} f_{p,(c,l)} - \sum_{k \in D_{(d,j)}} f_{p,(d,k)} \geq 0. \tag{7}$$

From (7), we have

$$\sum_{l \in D'_{(c,i)}} a_{(c,i),(c,l)} f_{p,(c,l)} - \sum_{k \in D'_{(d,j)}} a_{(d,j),(d,k)} f_{p,(d,k)} \geq 0, \qquad \forall v_{(c,i)}, v_{(d,j)}, \forall p. \tag{8}$$

Next, we rewrite the sum condition. We apply this condition to three nodes. To begin, we first express the sum condition for the nodes (e,r) and (c,i) , and then for the nodes (c,i) and (d,j). For nodes (e,r) and (c,i), we have

$$f_p^+(D_{(e,r)}) \geq \sum_{v_{(c,i)} \in \delta(v_{(e,r)})} f_p^+(D_{(c,i)}), \qquad \forall (e,r), p. \tag{9}$$

Additionally, for the nodes (c,i) and (d,j), we have

$$f_p^+(D_{(c,i)}) \geq \sum_{v_{(d,j)} \in \delta(v_{(c,i)})} f_p^+(D_{(d,j)}), \qquad \forall (c,i), p. \tag{10}$$

Equations (9) and (10) imply that

$$\begin{aligned}
&\sum_{v_{(c,i)} \in \delta(v_{(e,r)})} f_p^+(D_{(c,i)}) \geq \sum_{v_{(d,j)} \in \delta(v_{(c,i)})} f_p^+(D_{(d,j)}), \qquad \forall v_{(c,i)}, p, \\
&\sum_{v_{(c,i)} \in \delta(v_{(e,r)})} \sum_{l \in D_{(c,i)}} f_{p,(c,l)} \geq \sum_{v_{(d,j)} \in \delta(v_{(c,i)})} \sum_{k \in D_{(d,j)}} f_{p,(d,k)}, \\
&\sum_{v_{(e,r)} \in \delta^-(v_{(c,i)})} \sum_{l \in D'_{(c,i)}} a_{(c,i),(c,l)} z_{(e,r),(c,i)} f_{p,(c,l)} \geq \\
&\sum_{v_{(d,j)} \in \delta^+(v_{(c,i)})} \sum_{k \in D'_{(d,j)}} a_{(d,j),(d,k)} z_{(c,i),(d,j)} f_{p,(d,k)}, \qquad \forall v_{(c,i)}, p.
\end{aligned} \tag{11}$$

This enforces that the cumulative frequency of a node dominates the cumulative frequencies of its descendants along edges in the reconstructed tree. By combining these constraints for all nodes and samples, the BLP formulation guarantees that the resulting tree is:

- Directed and acyclic,
- Consistent with ancestral relationships, and
- Compatible with observed frequency data, thus providing a valid solution to the multi-state PP-MDP.

□

Therefore, the cladistic PPMDP is formulated as the following BLP model:

$$\max \sum_{(v_{(e,r)},v_{(c,i)})\in A'} z_{(e,r),(c,i)}$$

$$z_{(c,i),(d,j)} \leq \sum_{v_{(e,r)}\in\delta^-(v_{(c,i)})} z_{(e,r),(c,i)}, \qquad \forall (v_{(c,i)}, v_{(d,j)}) \in A', \tag{12}$$

$$\sum_{v_{(e,r)}\in\delta^-(v_{(c,i)})} z_{(e,r),(c,i)} \leq 1, \qquad \forall v_{(c,i)} \in V, \tag{13}$$

$$\sum_{v_{(e,r)}\in\delta^-(v_{(c,i)})} \sum_{l\in D'_{(c,i)}} a_{(c,i),(c,l)} z_{(e,r),(c,i)} f_{p,(c,l)} \geq$$

$$\sum_{v_{(d,j)}\in\delta^+(v_{(c,i)})} \sum_{k\in D'_{(d,j)}} a_{(d,j),(d,k)} z_{(c,i),(d,j)} f_{p,(d,k)}, \qquad \forall v_{(c,i)}, p \in [m], \tag{14}$$

$$z_{(e,r),(c,i)} \in \{0,1\}, \qquad \forall (v_{(e,r)}, v_{(c,i)}) \in A'. \tag{15}$$

Constraints (12) and (13) enforce connectedness in the reconstructed clonal tree T. Constraint (12) states that for every outgoing edge $(v_{(c,i)}, v_{(d,j)})$, there is at least one incoming edge $(v_{(e,r)}, v_{(c,i)})$ in T, and constraint (13) states that every node $v_{(c,i)}$ has at most one incoming edge $(v_{(e,r)}, v_{(c,i)})$. Constraint (14) enforces the MSSC, ensuring that the cumulative frequency of a node dominates that of its descendants, as defined previously. Since the ancestral graph is a DAG, no additional constraints are required to prevent cycles in the BLP formulation. Using an objective function that maximizes the number of edges in the clonal tree offers a powerful strategy for extracting the maximum amount of evolutionary information from genomic data such as SNVs and CNVs. This approach encourages the model to construct trees that include the largest number of mutational events possible, provided that they remain consistent with the data and biological constraints. Particularly when the data support a fine-grained separation of clones, this strategy can lead to more detailed reconstructions that reveal complex evolutionary trajectories. As a result, it enables a clearer interpretation of clonal relationships and mutation order. Moreover, this objective function may help uncover subclonal branches and low-frequency mutations that might be overlooked at simpler models [9, 14]. The proposed BLP formulation permits the reconstruction of either a single phylogenetic tree or, when required by the data, a forest consisting of multiple disconnected trees. A forest is produced when the available sequencing data cannot be consistently explained by a single evolutionary tree. Such situations may arise because of experimental noise, sequencing errors, missing observations, or other sources of uncertainty that prevent some mutation clusters from being placed unambiguously within the main phylogeny while satisfying all model constraints. Rather than forcing these clusters into uncertain evolutionary positions, the formulation represents them as separate connected components, yielding a forest. From a biological perspective, a forest should not be interpreted as evidence for multiple independent tumor origins. Instead, it reflects insufficient or uncertain information in the sequencing data to establish reliable evolutionary relationships among all inferred clones. Consequently, the disconnected components represent unresolved portions of the tumor evolutionary history caused by data uncertainty rather than distinct biological lineages. For all simulated and real datasets considered in this study, the reconstructed solutions were single phylogenetic trees, indicating that the available data were sufficient to connect all inferred clones into a unified evolutionary history. The forest formulation is nevertheless retained to improve the generality and robustness of the

proposed model, allowing it to accommodate datasets with higher levels of noise, sequencing artifacts, or incomplete information, where disconnected evolutionary components may naturally arise.

Theorem 3. *The BLP formulation (12)-(15) is correct in the sense that*

1. *If T is a complete perfect phylogenetic tree compatible with $\mathscr{S}$ that yields F, then there exist binary variable $z_{(c,i),(d,j)}$, satisfying (12)-(15).*

2. *If binary variables $z_{(c,i),(d,j)}$ satisfy (12)-(15), then the directed graph formed by edges with $z = 1$ is a rooted forest (a collection of trees) that satisfies the MSSC and yields F.*

Proof. 1. Suppose T is a complete perfect phylogenetic tree that is compatible with $\mathscr{S}$ and yields F. Define

$$z_{(c,i),(d,j)} = \begin{cases} 1, \text{If node } v_{(c,i)} \text{ is the parent of } v_{(d,j)}, \\ 0, \text{Otherwise.} \end{cases}$$

Since T is a rooted tree, every non-root vertex has exactly one parent. In fact, every vertex has at most one parent. So constraint (13) holds. Since every non-root vertex such as $v_{(c,i)}$ that appears as a parent must itself belong to the tree, $v_{(c,i)}$ is either a root or has a parent $v_{(e,r)}$ in T. Hence

$$z_{(c,i),(d,j)} \leq \sum_{v_{(e,r)} \in \delta^-(v_{(c,i)})} z_{(e,r),(c,i)}, \qquad \forall (v_{(c,i)}, v_{(d,j)}) \in A',$$

which satisfies constraint (12). Since T is compatible with $\mathscr{S}$ and yields F, the sum of the frequencies of the descendant set $D'_{(c,i)}$ must be at least equal to the total frequency assigned to the selected child descendant set. So constraint (14) holds. By Definition, $z \in \{0,1\}$ holds. So all constraints (12)-(15) hold.

2. Let the set of selected edges induced by the binary variables $z_{(c,i),(d,j)}$ be defined as $A'_a = \{(v_{(c,i)}, v_{(d,j)}) \in A' | z_{(c,i),(d,j)} = 1\}$. Let $T_a = (V_a, A'_a)$ denote the graph induced by these selected edges. We show that each T_a is a rooted tree. Consequently, $T = \{T_a\}$ is a rooted forest that satisfies MSSC and yields F. Constraint (13) states that every vertex has at most one parent. So the indegree of every vertex in T_a is at most 1. Constraint (12) states that whenever a vertex has a selected outgoing edge, that vertex must itself belong to a T_a and have a parent unless it is the root. By Proposition 5, the ancestry graph G is structurally acyclic, so every T_a in T is acyclic. Because $A'_a \subseteq A'$ (every subgraph of a DAG is a DAG). Every T_a is acyclic, and by Constraint (13) the indegree of every vertex is at most 1. So T_a is a rooted tree. Consequently, T is a rooted forest. Constraint (14) is exactly the algebraic encoding of MSSC. Since this constraint holds for every vertex and every sample, MSSC holds throughout the forest. The forest yields F because the ancestry graph G is constructed exactly from the observed frequency tensor of F. For each chosen edge, the coefficients of the descendant set $a_{(c,i),(c,l)}$ and the frequencies $f_{p,(c,l)}$ appear in Constraint (14) and guarantee that the frequencies assigned to the generations are consistent with MSSC. Hence the forest yields F. Thus, the ancestry graph induced by the $z_{(c,i),(d,j)}$ variables is a rooted forest yielding MSSC. □

The model yields a forest with the least number of connected components and a connected component with the most number of vertices. In fact, the goal is to find the T_a that has the largest number of vertices and is the largest spanning tree with V_a vertices, but at the same time shows the vertices that do not support

the phylogeny. Now if the output is a spanning tree on all vertices, the optimal solution corresponds to a complete perfect phylogenetic tree.

Formulation size: The size of the BLP formulation is polynomial of the form of $O(n^2k^2+mnk)$. In the ancestor graph, edges are directed from ancestor vertices to their descendants. When all vertices are mutually comparable, there exists a directed edge between every pair of comparable vertices. Consequently, the total number of edges is $n(k-1)(n(k-1)-1)/2$. Since the first set of constraints is defined for each edge, the total number of these constraints is $O(n^2k^2)$. The second set of constraints is defined for each vertex except for the root node, resulting in $n(k-1)$ constraints. Finally, the third set of constraints is defined for each sample and each vertex, yielding $mn(k-1)$ constraints. Therefore, the overall size of the BLP formulation is $O(n^2k^2+mnk)$. As a result, $O(n^2k^2)+O(nk)+O(mnk) \sim O(n^2k^2+mnk)$.

Complexity of the BLP: The proposed model is a BLP with $O(n(k-1))$ binary decision variables and a set of linear constraints. In the worst case, solving a BLP is equivalent to exploring all possible assignments of binary variables. Hence, the solution space has size $O(2^{|A'|})$ where $|A'| \sim O(n(k-1))$. Therefore, any exact algorithm (e.g., Branch-and-Bound or Branch-and-Cut) has exponential worst-case time complexity.

4 Numerical experiments

We have implemented PPMDP for the cladistic setting by formulating it as a BLP problem and solving it using CPLEX (version 22.1). Our evaluation encompasses 60 simulated datasets and two real-world dataset from a solid prostate cancer tumor [6]. The simulated datasets, described in [4], were generated by randomly constructing a tree with $n=5$ characters, followed by sampling $m \in \{2,5,10\}$ instances per tree. This process produced VAFs and copy number mixing proportions for each character across the sampled instances. For each value of m, 20 datasets (sim_0 through sim_{19}) were generated. Each dataset, such as sim_0, contains three versions: sim_{025}, sim_{055}, and sim_{0105}, corresponding to $n=5$ with 2, 5, and 10 samples, respectively. Altogether, this resulted in 60 distinct simulated instances.

In this study, the tensor $F = [[f_{p,(c,(x,y,z))}]]$ represents the frequency of the state (x,y,z) at locus c in sample p. Characters correspond to genomic positions whose states are represented by the triple (x,y,z), where x denotes the number of maternal copies, y the number of paternal copies, and z the number of mutant copies at that position. Each of these triples of states in $f_{p,(c,(x,y,z))}$ corresponds to state i in $f_{p,(c,i)}$. In Table 1, the states are listed for each of the simulated data instances.

4.1 Simulated data

In [4], twenty multi-state perfect phylogenetic trees with $n=5$ characters were generated using randomly constructed state trees, each allowing at most two copy number states per character, denoted by $x \leq 3$ and $y \leq 3$. For each simulated tree, a frequency tensor F containing $m \in \{2,5,10\}$ samples was generated. The VAF tensor is assumed to be nominal. SPRUCE is run on each simulated instance and the trees containing the largest subset of characters are obtained.

Using the frequency tensor and state trees, we construct the ancestral graph (Program1) based on Definition 7. By subsequently solving the specified BLP model, we derive a multi-state phylogenetic tree that captures the evolutionary relationships among subpopulations of cells within the tumor. We then

compare the simulated multi-state phylogenetic trees or ground truth trees with the phylogenetic trees obtained from our BLP model. Table 2 shows the execution time and memory occupied by Program1 and Program2 in seconds and megabytes, respectively, for each of the simulated data instances. All experiments were conducted on a computer running Windows 10 Pro, equipped with an Intel Core i7-7500U processor and 8GB of RAM.

Table 1: The states of each simulated data instance*

Data\States	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
sim_0	(1,1,0)	(1,1,1)	(1,0,0)	(2,1,1)	(2,2,0)	-	-	-	-	-
sim_1	(1,1,0)	(1,1,1)	(2,0,0)	(2,0,2)	-	-	-	-	-	-
sim_2	(1,1,0)	(1,1,1)	(2,0,2)	(2,1,1)	(3,2,1)	(1,0,0)	(2,2,2)	-	-	-
sim_3	(1,1,0)	(1,0,0)	(3,0,2)	(1,1,1)	(2,1,2)	(3,2,1)	-	-	-	-
sim_4	(1,1,0)	(1,0,0)	(1,0,1)	(1,1,1)	(3,1,1)	-	-	-	-	-
sim_5	(1,1,0)	(1,1,1)	(2,1,0)	(3,1,0)	(2,1,1)	-	-	-	-	-
sim_6	(1,1,0)	(3,3,2)	(1,1,1)	(2,1,2)	(1,0,0)	(3,0,0)	-	-	-	-
sim_7	(1,1,0)	(1,1,1)	(1,0,0)	(2,0,0)	(3,0,0)	-	-	-	-	-
sim_8	(1,1,0)	(1,1,1)	(2,1,0)	(2,2,2)	-	-	-	-	-	-
sim_9	(1,1,0)	(1,1,1)	(2,2,0)	(3,1,0)	(3,1,2)	(3,2,2)	-	-	-	-
sim_{10}	(1,1,0)	(1,1,1)	(2,1,0)	(3,1,1)	-	-	-	-	-	-
sim_{11}	(1,1,0)	(2,1,0)	(2,1,1)	(2,0,1)	(1,1,1)	(3,0,0)	(3,1,3)	(3,2,2)	-	-
sim_{12}	(1,1,0)	(1,1,1)	(1,0,0)	(2,0,0)	-	-	-	-	-	-
sim_{13}	(1,1,0)	(1,1,1)	(2,1,2)	(3,3,0)	(3,0,0)	(2,0,0)	(2,1,0)	(2,1,1)	(3,2,0)	-
sim_{14}	(1,1,0)	(1,1,1)	(3,1,3)	(3,0,0)	-	-	-	-	-	-
sim_{15}	(1,1,0)	(1,1,1)	(2,0,2)	(3,1,1)	-	-	-	-	-	-
sim_{16}	(1,1,0)	(1,1,1)	(1,0,1)	(3,2,0)	(2,1,1)	(3,2,2)	(2,0,1)	(3,3,2)	-	-
sim_{17}	(1,1,0)	(1,1,1)	(2,1,2)	(2,2,2)	(3,2,3)	(2,1,1)	-	-	-	-
sim_{18}	(1,1,0)	(1,1,1)	(2,1,2)	(1,0,0)	(1,0,1)	(2,1,1)	(2,0,1)	-	-	-
sim_{19}	(1,1,0)	(1,1,1)	(3,3,0)	(3,2,0)	(3,2,1)	(1,0,1)	(3,3,3)	(2,1,0)	(2,2,0)	(2,2,1)

The ground truth trees remain identical for fixed values of n and k, regardless of the number of samples $m \in \{2,5,10\}$. For example, in instance sim_0, where $n=5, k=5$, and $m \in \{2,5,10\}$, the trees (as shown in Figure 2) include five states: (1,1,0), (1,1,1), (1,0,0), (2,1,1), and (2,2,0), labeled with indices 0 through 4, respectively. These instance correspond to the same ground truth tree across different values of m. However, solving the same instances using the BLP model results in phylogenetic trees that may differ structurally.

Despite these differences, the phylogenetic trees produced by the BLP model exhibit a high degree of similarity to ground truth trees. For instance, in sim_4 (Figure 3), the parameters are identical to sim_0 ($n=5, k=5, m \in \{2,5,10\}$), but the set of states changes to (1,1,0), (1,0,0), (1,0,1), (1,1,1), and (3,1,1). As a result, the ground truth tree and the BLP-derived phylogenetic tree in sim_4 differ from those in sim_0. Moreover, the tensor F and the corresponding state trees in sim_4 also diverge from those in sim_0.

To quantitatively assess the similarity between ground truth trees and the trees generated by the BLP model, we define **Edge Accuracy** as the percentage of shared edges between the two trees. For each instance with fixed $n=5$ and $k=5$, and varying sample sizes $m \in \{2,5,10\}$, we calculated this coincidence to determine how closely the reconstructed tree from the BLP model matches the ground truth tree.

Our computational results indicate that, for the majority of instances, the BLP formulation reconstructs phylogenetic trees exhibiting a high degree of structural agreement with the ground truth trees.

* Five states are given for the sim_0 dataset. This means that all three datasets sim_{025}, sim_{055} and sim_{0105} which belong to the sim_0 dataset have these 5 states. This is true for the other datasets.

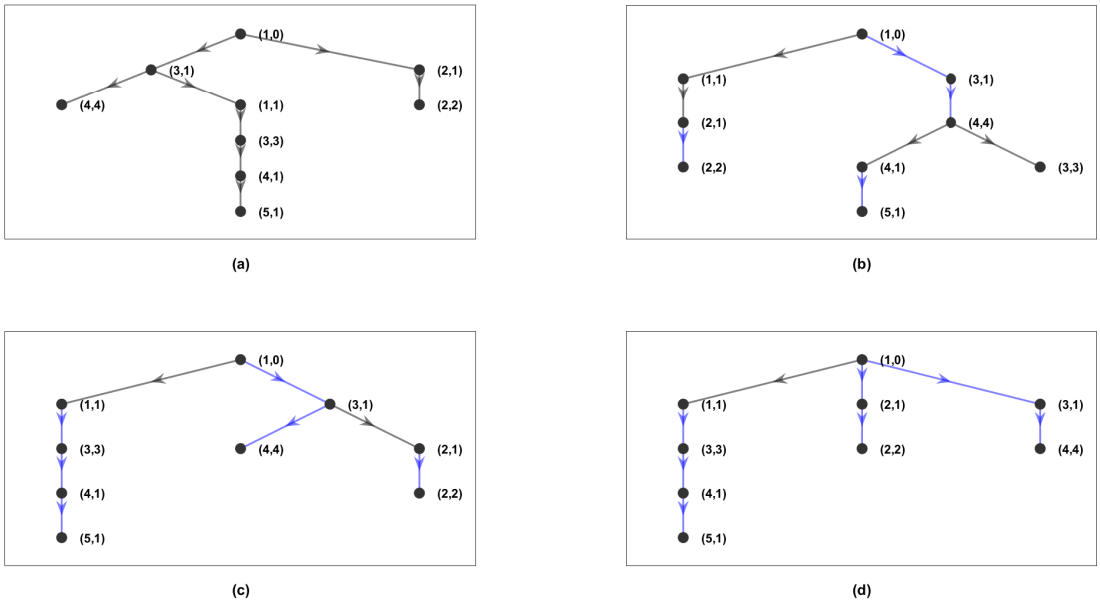

Figure 2: (a) Ground truth tree. (b) Multi-state perfect phylogenetic tree obtained from BLP model using sim_0 simulated data for $m = 2$, (c) for $m = 5$, (d) $m = 10$ and $k = 5$ on $n = 5$

.

Let TBLP and Ttruth denote the tree inferred by the BLP model and ground truth tree, respectively. Let Edge Accuracy(TBLP,Ttruth) measure the percentage of shared structural features. For most simulated datasets, Edge Accuracy(TBLP,Ttruth)≥ 0.8, indicating strong topological similarity. As shown in Table 3, for dataset sim_8 and sim_{15}, the Edge Accuracy over all values $m \in \{2,5,10\}$ exceeded 80%, in several instances, TBLP and Ttruth, i.e., the reconstructed trees were isomorphic. In contrast, for instances such as sim_4, where the induced character-state configurations generate more divergent or ambiguous topologies, the Edge Accuracy values were significantly lower. Formally, Edge Accuracy (TBLP,Ttruth)≤ 0.8, reflecting increased combinatorial complexity in the feasible tree space. This variability demonstrates that the BLP model is sensitive to the heterogeneity of the input tensor F, particularly when the underlying evolutionary structure is dense, weakly separated, or topologically ambiguous results are obtained.

Overall, although the BLP approach does not necessarily recover the ground truth tree in every instance exactly, it typically produces a tree that is structurally close to it, especially when the evolutionary signal encoded in F is sufficiently strong to constrain the feasible region of the optimization problem.

Impact of sample size on tree edge accuracy: In general, the Edge Accuracy increases with increasing sample size. However, this trend is not strictly monotonic and can fluctuate slightly depending on the components of tensor F and the structure of the state trees that define the ancestral graph. For instance, in sim_2 with $n = 5, k = 7$, and $m \in \{2,5,10\}$, the Edge Accuracy increases consistently with the number of samples, as illustrated in Figure 4. The seven states in this instance, indexed from 0 to 6, are: (1,1,0), (1,1,1), (2,0,2), (2,1,1), (3,2,1), (1,0,0), and (2,2,2). A similar increasing trend in Edge Accuracy is observed in several other instances, including $sim_0, sim_1, sim_4, sim_7, sim_8, sim_9, sim_{12}, sim_{14}, sim_{15}$, and sim_{18}. The Edge Accuracy values are reported in Table 3. In some instances, such as sim_3 ($m = 5$),

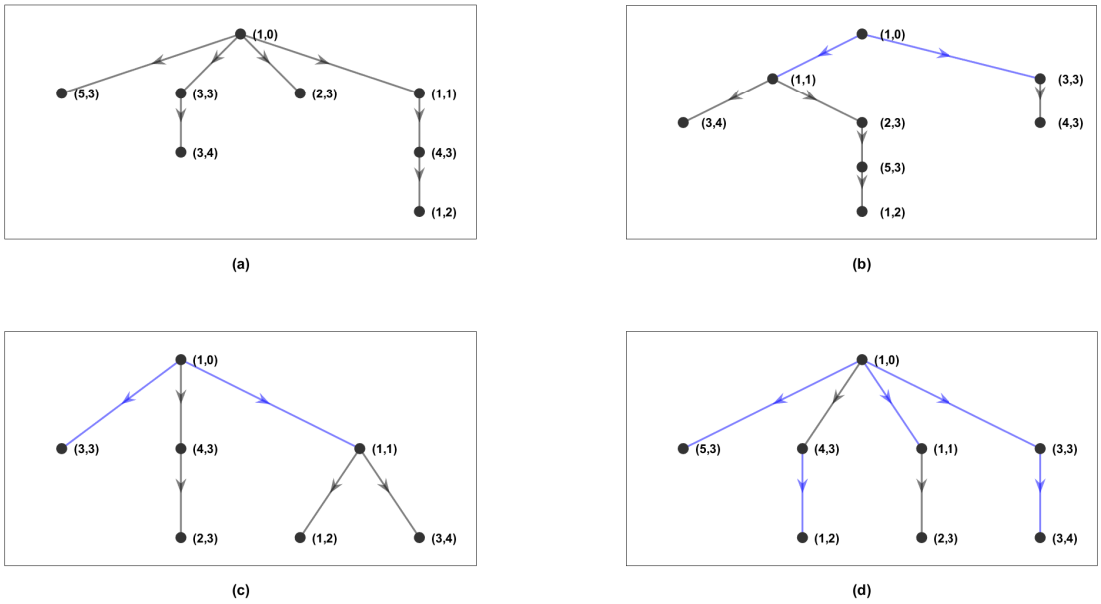

Figure 3: (a) Ground truth tree. (b) Multi-state perfect phylogenetic tree obtained from BLP model using sim_4 simulated data for $m = 2$, (c) for $m = 5$, (d) $m = 10$ and $k = 5$ on $n = 5$

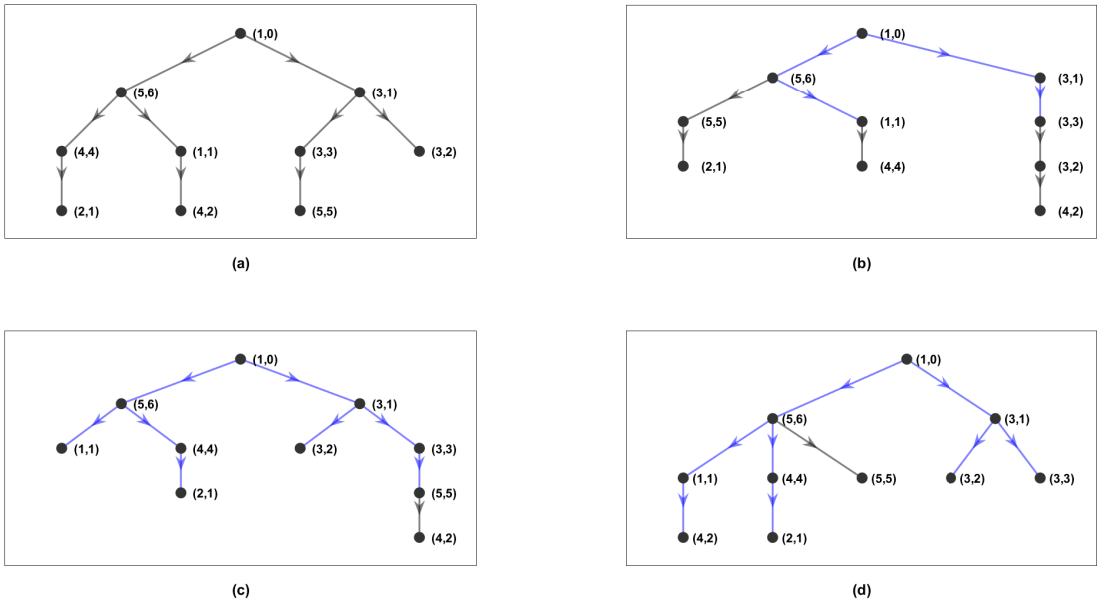

Figure 4: (a) Ground truth tree. (b) Multi-state perfect phylogenetic tree obtained from BLP model using sim_2 simulated data for $m = 2$, (c) for $m = 5$, (d) $m = 10$ and $k = 7$ on $n = 5$

sim_7 ($m = 10$), sim_{11} ($m = 5$), sim_{14}, sim_{15} ($m = 5, 10$) the BLP inferred tree achieved an Edge Accuracy of 100% with respect to the ground truth tree, indicating complete agreement in the edge sets. While

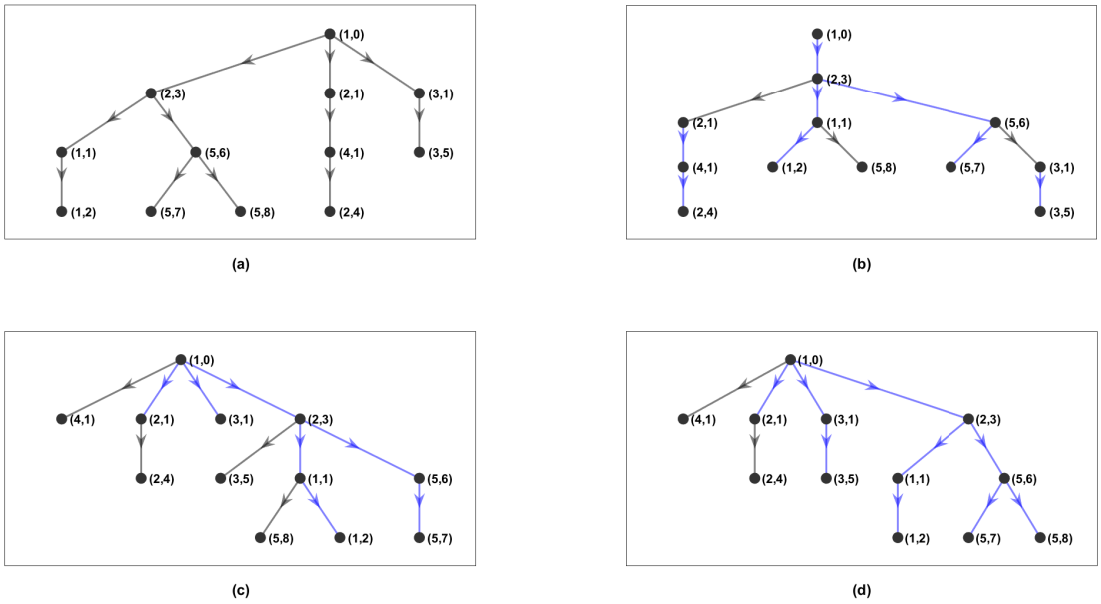

Figure 5: (a) Ground truth tree. (b) Multi-state perfect phylogenetic tree obtained from BLP model using sim_{13} simulated data for $m = 2$, (c) for $m = 5$, (d) $m = 10$ and $k = 9$ on $n = 5$.

most instances show an increasing trend, there are occasional fluctuations, such as sim_3, sim_5 and sim_{11}. Nonetheless, the overall pattern indicates improved Edge Accuracy with more samples. Additional examples also support this observation, such as sim_6, sim_{10}, sim_{13} and sim_{16}. This trend is illustrated in Figure 5 for the sim_{13} dataset.

These results collectively suggest that increasing the number of samples enhances the structural similarity between trees produced by the BLP model and ground truth trees, although certain complexities in the evolutionary structure (as reflected in tensor F and state trees) may introduce minor inconsistencies. In sim_{17} ($k = 6$), the Edge Accuracy for $m = 2$ is 66%, $m = 5$ is 44%, and $m = 10$ is 77%. In sim_{19} the Edge Accuracy for each m on $n = 5$ characters and $k = 10$ is 36%, 54%, and 72%, respectively. For all datasets, as k decreases, the Edge Accuracy increases as m increases.

4.2 Real data

In real data, given the tensor F and the ancestral graph as inputs, we obtain the multi-state perfect phylogenetic tree resulting from the BLP model (Program2). In Program1, we obtain this ancestral graph using the tensor F and the state tree that we give as input to the program. Table 4 shows the real data states.

4.2.1 Comparison of BLP model with SPRUCE algorithm using real data

In this case, measurement errors are taken into account. Accordingly, two datasets (file-A22.666 and file-A22.1136), each containing 26 characters (see Table 5), are considered. Although both datasets share the same set of character states (as shown in Table 4), the ordering of characters differs between

Table 2: Execution time (s represents seconds) and memory (M represents megabytes) occupied by Program1 (Program1 is a program that returns the ancestor graph) and Program2 (Program2 is a program that obtains a phylogenetic tree) with simulated data.

	Program1		Program2			Program1		Program2	
Datasets	Time(s)	Memory(M)	Time	Memory	Datasets	Time	Memory	Time	Memory
sim_{025}	1.84	2.35	3.77	2.71	sim_{1025}	1.79	3.82	1.66	3.45
sim_{055}	2.08	2.35	3.57	4.03	sim_{1055}	1.83	4.39	1.94	2.28
sim_{0105}	1.89	1.96	3.15	4.31	sim_{10105}	1.83	3.42	1.89	2.52
sim_{125}	1.86	2.36	1.87	4.36	sim_{1125}	1.79	2.32	2.42	3.04
sim_{155}	1.94	4.04	2.53	3.59	sim_{1155}	1.92	2.08	2.21	4.40
sim_{1105}	1.81	2.07	1.87	2.80	sim_{11105}	2.00	3.55	1.87	3.95
sim_{225}	1.81	2.57	2.50	5.25	sim_{1225}	1.73	2.22	1.96	2.64
sim_{255}	1.95	2.76	2.29	2.70	sim_{1255}	1.76	2.69	1.91	3.18
sim_{2105}	1.96	2.70	2.83	2.98	sim_{12105}	1.71	4.06	1.92	5.18
sim_{325}	1.85	2.25	2.15	4.57	sim_{1325}	1.80	2.51	2.27	5.81
sim_{355}	1.94	3.78	2.07	2.72	sim_{1355}	2.02	2.28	2.11	5.33
sim_{3105}	1.92	2.58	2.00	4.40	sim_{13105}	1.93	4.06	2.56	4.55
sim_{425}	1.72	4.41	2.00	4.98	sim_{1425}	1.62	2.24	1.94	3.92
sim_{455}	1.89	4.50	2.08	3.95	sim_{1455}	1.95	3.95	1.88	2.52
sim_{4105}	1.84	4.19	2.03	3.60	sim_{14105}	1.71	2.5	1.97	2.66
sim_{525}	1.76	2.23	1.84	3.86	sim_{1525}	1.72	2.36	2.06	3.88
sim_{555}	1.93	3.60	1.91	3.82	sim_{1555}	1.84	2.16	1.93	4.09
sim_{5105}	1.79	2.02	1.91	2.46	sim_{15105}	1.82	2.32	1.93	3.85
sim_{625}	1.78	2.18	2.00	4.79	sim_{1625}	1.78	2.66	2.88	3.48
sim_{655}	1.88	2.78	2.00	4.74	sim_{1655}	1.88	2.50	2.01	4.23
sim_{6105}	1.83	2.13	2.19	4.68	sim_{16105}	1.96	3.62	2.78	2.77
sim_{725}	1.79	2.52	2.13	2.58	sim_{1725}	1.74	2.32	2.02	4.53
sim_{755}	1.90	2.12	3.70	3.85	sim_{1755}	1.73	2.47	2.28	2.98
sim_{7105}	1.80	2.23	2.10	3.94	sim_{17105}	1.83	1.99	2.04	3.10
sim_{825}	1.71	2.86	1.94	4.32	sim_{1825}	1.82	2.16	2.11	2.86
sim_{855}	1.83	2.37	1.97	4.21	sim_{1855}	1.90	3.86	2.74	3.27
sim_{8105}	1.79	2.25	1.98	3.48	sim_{18105}	1.82	2.37	2.49	4.62
sim_{925}	1.78	2.54	2.25	2.81	sim_{1925}	1.87	3.85	2.61	5.04
sim_{955}	1.67	2.34	2.02	4.37	sim_{1955}	2.01	2.66	2.36	4.99
sim_{9105}	1.88	2.65	2.52	3.93	sim_{19105}	2.06	3.90	2.36	5.00

them, and their corresponding state trees are also distinct. For each dataset, the tensor F together with the associated state trees are provided as input to program1. the resulting ancestral graph is then passed to program2, which yields a phylogenetic tree satisfying both the ancestry and sum conditions. This tree represents the inferred evolutionary relationships among clones within a subpopulation of tumor cells. For real tumor sequencing datasets, the true phylogenetic history is generally unavailable, making direct validation against a known ground-truth tree impossible. Consequently, the quality of the reconstructed phylogenies is evaluated primarily based on their biological plausibility and consistency with established models of tumor evolution. In prostate cancer, tumor progression is widely understood to occur through the sequential accumulation of somatic mutations, leading to a hierarchical organization of ancestral and descendant clones. The phylogenetic trees reconstructed by both the proposed BLP formulation and SPRUCE are consistent with this general model of clonal evolution. Therefore, differences in their branching patterns should be interpreted as alternative yet biologically plausible evolutionary histories that are supported by the available sequencing data, rather than as deviations from an unknown ground-truth phylogeny. Figure 6(a) presents the phylogenetic tree obtained for dataset file-A22.666 using program2 (BLP), while Figure 6(b) shows the corresponding tree generated by the SPRUCE algo-

Table 3: Edge accuracy of each dataset for $n = 5$, $m \in \{2,5,10\}$ and different k

Datasets	Edge Accuracy (%)	k
sim_0	50, 75, 87	5
sim_1	66,66,83	4
sim_2	44,88,88	7
sim_3	42,100,85	6
sim_4	28,28,71	5
sim_5	37,75,62	5
sim_6	50,25,75	6
sim_7	37,62,100	5
sim_8	83,83,83	4
sim_9	62,75,87	6
sim_{10}	62,87,50	4
sim_{11}	0,100,85	8
sim_{12}	37,87,100	4
sim_{13}	72,63,81	9
sim_{14}	71,100,100	4
sim_{15}	83,100,100	4
sim_{16}	50,90,70	8
sim_{17}	66,44,77	6
sim_{18}	25,37,62	7
sim_{19}	36,54,72	10

Table 4: Real data states related to solid prostate cancer

states\Data	solid prostate cancer
$i = 0$	(1,1,0)
$i = 1$	(1,1,1)
$i = 2$	(2,1,0)
$i = 3$	(1,0,0)
$i = 4$	(1,0,1)
$i = 5$	(2,1,1)
$i = 6$	(3,1,1)
$i = 7$	(2,2,0)
$i = 8$	(3,0,0)
$i = 9$	(3,1,0)
$i = 10$	(3,2,0)
$i = 11$	(4,0,0)
$i = 12$	(2,2,1)
$i = 13$	(2,2,2)
$i = 14$	(2,0,2)
$i = 15$	(3,0,3)
$i = 16$	(2,1,2)
$i = 17$	(2,0,0)
$i = 18$	(3,2,1)
$i = 19$	(3,3,0)
$i = 20$	(3,0,1)
$i = 21$	(4,2,0)
$i = 22$	(3,1,2)
$i = 23$	(2,0,1)
$i = 24$	(3,3,1)
$i = 25$	(3,3,2)

rithm. Each node is labeled by a character-state pair, and colors correspond to clones reported in [6], as summarized in Table 6. In this context, a mutation occurring at a specific genomic location is assigned to a clone due to coincident occurrence, where a clone is defined as a subpopulation of cells sharing an

identical genome. For this dataset, the execution time and memory usage of program1 are 36.30 s and 23.3 MB, respectively, while program2 requires 49.14 s and 177.25 MB. Similarly, Figure 7(a) illustrates the phylogenetic tree for dataset file-A22.1136 obtained using program2, and Figure 7(b) presents the corresponding tree generated by the SPRUCE algorithm. For this dataset, program1 requires 35.13 s and 22.54 MB of memory, whereas program2 requires 41.79 s and 171.78 MB. In Figure 6, the tree resulting from BLP shows the evolutionary relationships in more detail. The branch separation is greater. As a result, more information is included in the model, but not all of it is necessarily definitive. The model resulting from SPRUCE is more conservative in the case of ambiguous data. In Figure 7, the SPRUCE tree shows the evolutionary relationships more precisely. The tree resulting from BLP shows less information.

Table 5: Real data characters related to solid prostate cancer

characters\Data	file-A22.666	file-A22.1136
$i = 1$	FAM184B	GPR98
$i = 2$	PPFIA1	SLC26A2
$i = 3$	SLC26A2	TBL1XR1\CPNE4
$i = 4$	TBL1XR1\CPNE4	ZNF420
$i = 5$	ZNF420	ELMOD2
$i = 6$	ELMOD2	FAM184B
$i = 7$	C2orf70	C2orf70
$i = 8$	ADRA1B	ADRA1B
$i = 9$	ZBTB20	ZBTB20
$i = 10$	WHSC2	PPFIA1
$i = 11$	PRDM5	PRDM5
$i = 12$	ENSG00000214793	ENSG00000214793
$i = 13$	SCN4A	WHSC2
$i = 14$	TFB1M	SCN4A
$i = 15$	ZFP28\ZNF527\GPR4\NAPSB\C2orf16\XRCC5	TFB1M
$i = 16$	EPB41L1	ZFP28\ZNF527\GPR4\NAPSB\C2orf16\XRCC5
$i = 17$	ZNF579	EPB41L1
$i = 18$	GALNS	ZNF579
$i = 19$	HEATR6	GALNS
$i = 20$	PTPRT	HEATR6
$i = 21$	FREM3\APLF	PTPRT
$i = 22$	OR8B8	FREM3\APLF
$i = 23$	RB1	OR8B8
$i = 24$	DNAH5	RB1
$i = 25$	CDHR5	DNAH5
$i = 26$	ACE	ACE

The differences observed between the reconstructed phylogenies in Figures 6 and 7 reflect the distinct evolutionary characteristics of the corresponding datasets rather than the overall superiority of one reconstruction method over another. Different mutation frequency distributions, VAF patterns, and clonal evolutionary structures may lead to different levels of phylogenetic resolution. For the file-A22.666 dataset, the proposed BLP formulation reconstructs additional branching events, resulting in a more detailed representation of the inferred clonal architecture. In contrast, for the file-A22.1136 dataset, SPRUCE reconstructs parent–child relationships that are more consistent with the inferred evolutionary history, whereas the BLP formulation produces a more compact phylogeny. These differences illustrate that cancer phylogeny reconstruction is inherently data-dependent, and multiple biologically plausible evolutionary histories may be consistent with the observed sequencing data. To provide a quantitative comparison, the reconstructed phylogenies are evaluated using the number of shared parent–child relationships (shared edges). This criterion was selected because, in cancer phylogeny, preserving ancestor–descendant relationships is generally more biologically meaningful than topology-only measures such as the Robinson–Foulds distance, which compares tree topology without considering mutation placement or evolutionary direction. Accordingly, the comparison focuses on the recovery of biologically meaningful evolutionary relationships rather than topological similarity alone. Overall, neither

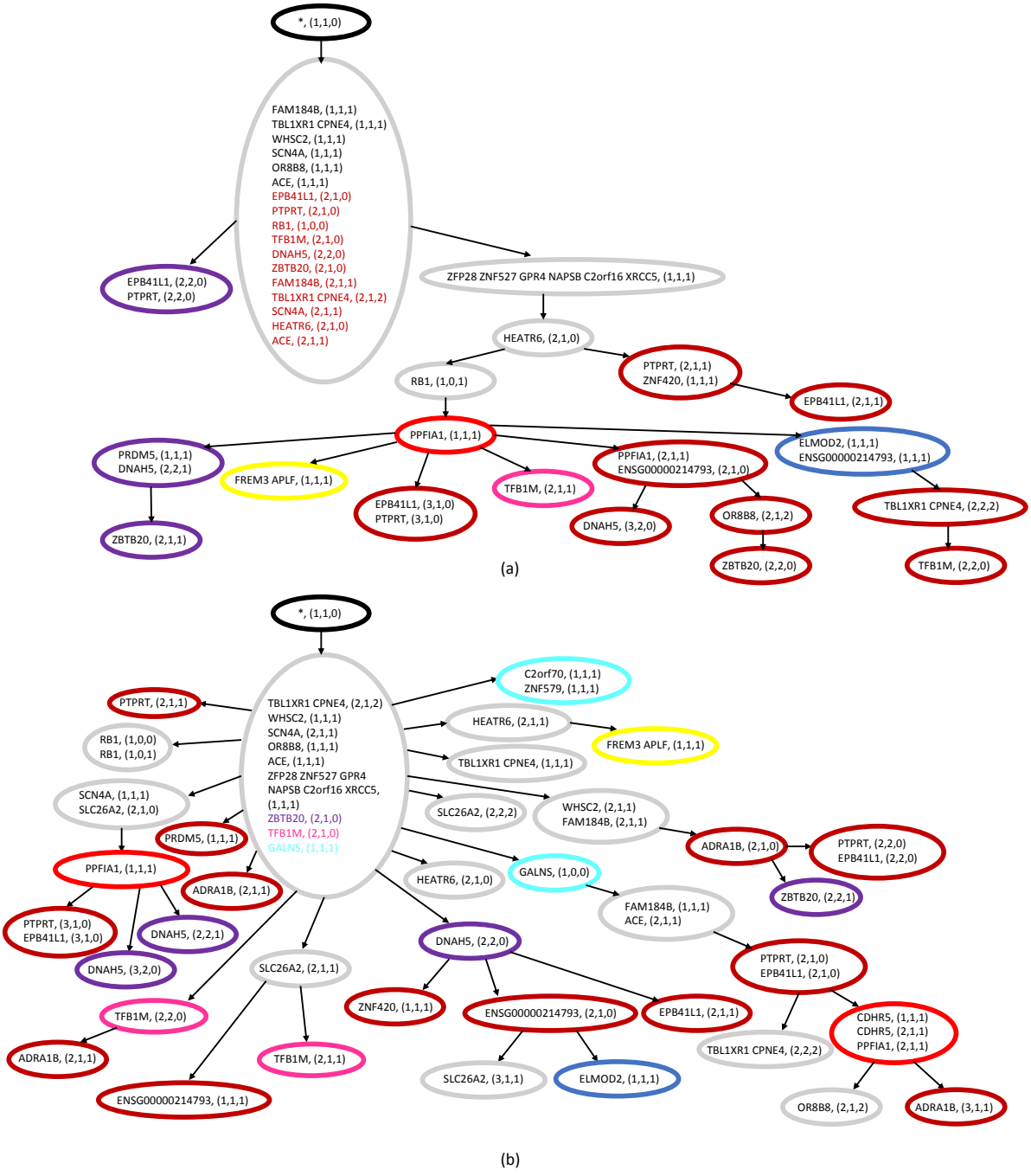

Figure 6: Multi-state perfect phylogenetic tree obtained from (a) SPRUCE algorithm and (b) BLP.

method is consistently superior across all datasets. The proposed BLP formulation is capable of recovering more detailed clonal architectures when the available data support multiple branching events, whereas SPRUCE may reconstruct parent–child relationships that are more consistent for datasets ex-

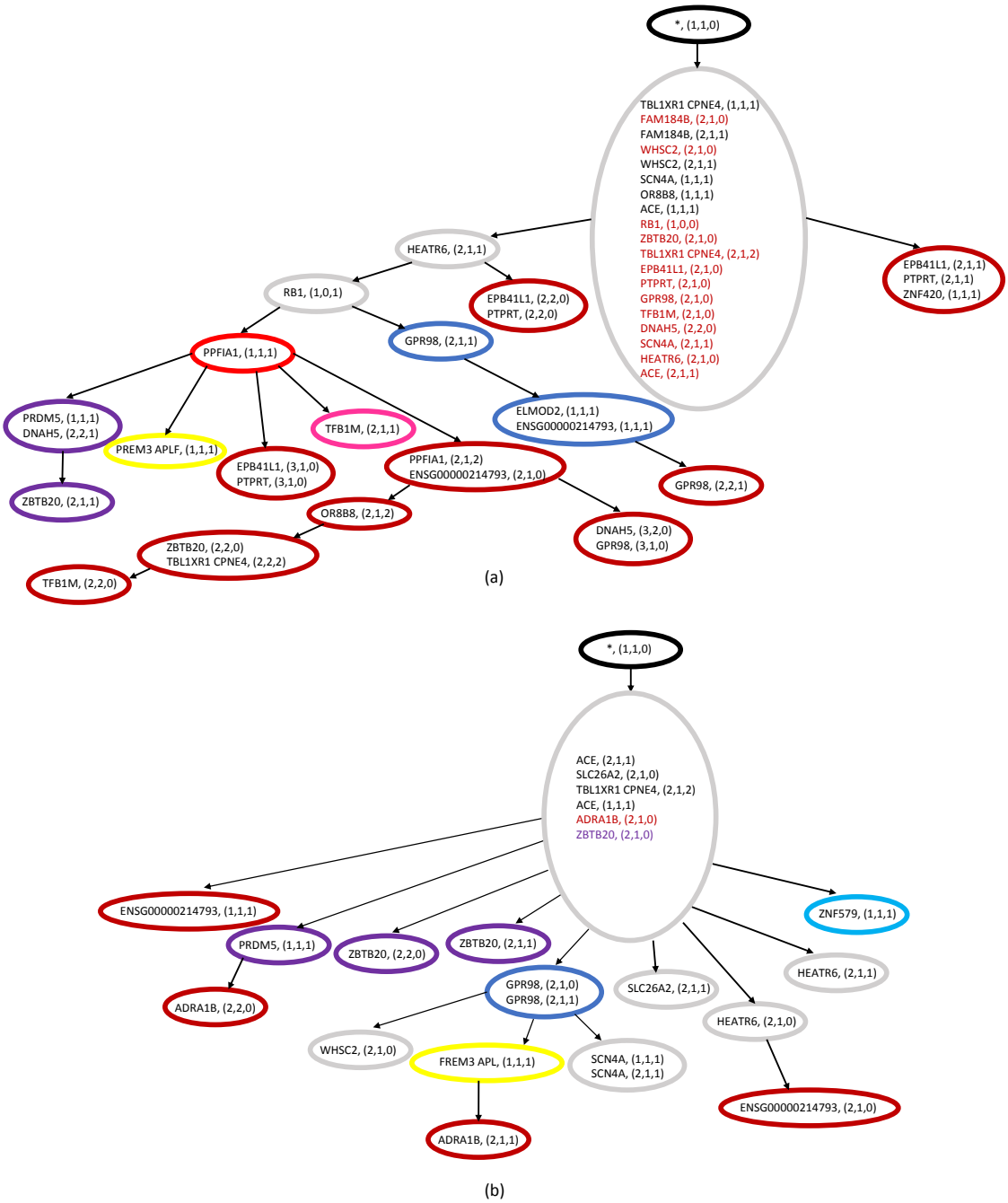

Figure 7: Multi-state perfect phylogenetic tree obtained from (a) SPRUCE algorithm and (b) BLP.

hibiting simpler evolutionary patterns. Therefore, the relative performance of the two methods depends on the characteristics of the input data, including the mutation distribution, VAF patterns, and the complexity of the underlying clonal evolution.

Table 6: Colors correspond to clones

characters	colors
ACE	gray
C2orf70	cyan
DNAH5	violet
ELMOD2	dodgerblue
EPB41L1	brown
FAM184B	gray
FREM3 \ APLF	yellow
GALNS	cyan
GPR98	dodgerblue
HEATR6	gray
OR8B8	gray
PPFIA1	red
PRDM5	violet
PTPRT	brown
RB1	gray
SCN4A	gray
SLC26A2	gray
TBL1XR1 CPNE4	gray
TFB1M	pink
WHSC2	gray
ZBTB20	violet
ZFP28 \ ZNF527 \ GPR4 \ NAPSB \ C2orf16 \ XRCC5	gray
ZNF420	brown
ZNF579	cyan
CDHR5	red
ADRA1B	brown
ENSG00000214793	brown

5 Conclusion

In this study, we developed a BLP formulation for the cladistic PPMDP. The proposed model integrates the tensor representation F and the associated state trees to infer phylogenetic structures in a systematic optimization framework. For simulated datasets, the ancestral graph is first derived from F and the state trees, and the optimization model is subsequently solved. For real datasets, the tensor F, state trees, and ancestral graph are jointly utilized as inputs to the model. The resulting formulation is a linear programming model that can be solved efficiently using CPLEX with low computational overhead. Experimental results demonstrate that the phylogenetic trees obtained from the proposed BLP approach exhibit a high level of agreement with ground truth trees, confirming the effectiveness and reliability of the model in capturing evolutionary relationships among tumor subclones. Future research directions include incorporating measurement errors arising from sequencing technologies into the cladistic PPMDP framework to enhance robustness under realistic conditions. Despite its encouraging performance, the proposed BLP formulation has several limitations. First, the model relies on the IAA and the state-tree framework to define the admissible evolutionary transitions. Although these assumptions are widely adopted in cancer phylogeny reconstruction, they may not fully capture the complexity of all tumor evolutionary processes. Second, while the computational results demonstrate that the proposed formulation can be solved efficiently for the benchmark datasets considered in this study, the size of the optimization model increases with the numbers of mutation clusters, samples, and mutation states. Consequently, very large datasets may require greater computational resources and longer solution times. Several directions may

be pursued to extend the proposed framework. Future work includes relaxing the underlying evolutionary assumptions, jointly inferring state trees and tumor phylogenies, improving scalability through decomposition techniques and more efficient optimization algorithms, integrating additional biological information and multi-omics sequencing data, and extending the formulation to accommodate more complex models of tumor evolution. These extensions have the potential to improve both the biological realism and the applicability of the proposed approach to large-scale cancer phylogeny reconstruction problems. Another important extension is the general (non-cladistic) version of the problem, in which state trees are unknown and must be inferred. Finally, the problem can be further extended to uncertainty settings by developing robust optimization models based on interval uncertainty sets, such as the Bertsimas and Sim framework.

Acknowledgements:

The authors appreciate the anonymous referees for their helpful comments which lead to this improved version of the paper.

Conflict of interest

The authors have not reported any potential conflicts of interest.

Availability of data and materials:

The raw data supporting the findings of this study are available in Github at https://github.com/Razieh-E/BLP-cladisticPPMDP.git

Author contribution:

All authors were involved in designing the procedure and drafting the manuscript. R.E. conducted the implementation of codes.

References

[1] K. Chen, J.C. Moravec, A. Gavryushkin, D. Welch, A.J. Drummond, *Accounting for errors in data improves divergence time estimates in single-cell*, Mol. Biol. Evol. **39(8)** (2022) msac143.

[2] S. Ciccolella, M.D. Patterson, B.J. Raphael, M. El-Kebir, *Inferring cancer progression with mutation losses from single-cell data using ILP and cladistic constraints*, Bioinformatics **36** (2020) i671–i679.

[3] M. El-Kebir, L. Oesper, H. Acheson-Field, B.J. Raphael, *Reconstruction of clonal trees and tumor composition from multi-sample sequencing data*, Bioinformatics **31** (2015) i62–i70.

[4] M. El-Kebir, G. Satas, L. Oesper, B.J. Raphael, *Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures*, Cell Syst. **3** (2016) 43–53.

[5] T. Funnell, C.H. O'Flanagan, M.J. Williams, A. McPherson, S. McKinney, F. Kabeer, H. Lee, S. Salehi, I. Vázquez-García, H. Shi, E. Leventhal, T. Masud, P. Eirew, D. Yap, A.W. Zhang, J.L.P. Lim, B. Wang, J. Brimhall, J. Biele, J. Ting, V. Au, M. Van Vliet, Y. F. Liu, S. Beatty, D. Lai, J. Pham, D. Grewal, D. Abrams, E. Havasov, S. Leung, V. Bojilova, R.A. Moore, N. Rusk, F. Uhlitz, N. Ceglia, A.C. Weiner, E. Zaikova, J.M. Douglas, D. Zamarin, B. Weigelt, S.H. Kim, A. Da Cruz Paula, J.S. Reis-Filho, S.D. Martin, Y. Li, H. Xu, T R. de Algara, S.R. Lee, V.C. Llanos, D.G. Huntsman, J N. McAlpine, S.P. Shah, S. Aparicio, *Single-cell genomic variation induced by mutational processes in cancer*, Nature **612** (2022) 106–115.

[6] G. Gundem, P. Van Loo, B. Kremeyer, L.B. Alexandrov, J.M. C. Tubio, E. Papaemmanuil, D.S. Brewer, H.M. L. Kallio, G. Högnäs, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O'Meara, K.J. Dawson, W. Isaacs, M.R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, D.E. Neal, C.S. Cooper, R.A. Eeles, T. Visakorpi, P.J. Campbell, U. McDermott, D.C. Wedge, G.S. Bova, *The evolutionary history of lethal metastatic prostate cancer*, Nature **520** (2015) 353–357.

[7] D. Gusfield, *Efficient algorithms for inferring evolutionary trees*, Networks **21** (1991) 19–28.

[8] M. Kimura, *The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations*, Genetics **61** (1969) 893–903.

[9] S. Malikic, K. Jahn, J. Kuipers, S.C. Sahinalp, N. Beerenwinkel, *Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data*, Nat Commun. **10** (2019) 2750.

[10] M. Markowska, T. cakala, B. Miasojedow, B. Aybey, D. Juraeva, J. Mazur, E. Ross, E. Staub, E. Szczurek, *CONET: copy number event tree model of evolutionary tumor history for single-cell data*, Genome Biol. **23** (2022) 128.

[11] C.A. Miller, B.S. White, N.D. Dees, M. Griffith, J.S. Welch, O.L. Griffith, R. Vij, M.H. Tomasson, T.A. Graubert, M. . Walter, M.J. Ellis, W. Schierding, J.F. DiPersio, T.J. Ley, E.R. Mardis, R.K. Wilson, L. Ding, *SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution*, PLoS Comput. Biol. **10** (2014) e1003665.

[12] M.A. Myers, B.J. Arnold, V. Bansal, M. Balaban, K.M. Mullen, S. Zaccaria, B.J. Raphael, *HATCHet2: clone- and haplotype-specific copy number inference from bulk tumor sequencing data*, Genome Biol. **25** (2024) 130.

[13] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, S.P. Shah, *PyClone: statistical inference of clonal population structure in cancer*, Nat. Methods **11(4)** (2014) 396-8.

[14] A.E. Sadeqi, M. El-Kebir, *Enumerating clonal trees*, Bioinformatics **38** (2022) i143–i151.

[15] P. Sashittal, H. Zhang, C.A. Iacobuzio-Donahue, B.J. Raphael, *ConDoR: tumor phylogeny inference with a copy-number constrained mutation loss model*, Genome Biol. **24** (2023) 272.

[16] P. Shahrouzi, F. Forouz, A. Mathelier, V.N. Kristensen, P.H.G. Duijf, *Copy number alterations: a catastrophic orchestration of the breast cancer*, Trends Mol. Med. **30** (2024) 750–764.

[17] C.D. Steele, A. Abbasi, S.M.A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, A. McLatchie, M. Lechner, N. Light, A. Shlien, D. Malkin, A. Feber, P. Proszek, T. Lesluyes, F. Mertens, A. Flanagan, M. Tarabichi, P. Van Loo, L. Alexandrov, N. Pillay, *Signatures of copy number alterations in human cancer*, Nature **606** (2022) 984–991.

[18] S. Zaccaria, M. El-Kebir, G.W. Klau, B.J. Raphael, *The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data*, In: Sahinalp, S. (eds) Research in Computational Molecular Biology. RECOMB Lecture Notes in Computer Science, Springer, 10229 (2017) 318–335.

[19] S. Zaccaria, B.J. Raphael, *Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data*, Nat. Commun. **11** (2020) 4301.