Extending the Variant Allele Frequency Factorization Problem (VAFFP) to Multi-Patient Data -
Peyton Wiecking, Virginia Tech (CS5124 Algorithms in Bioinformatics Final Project)

1. Introduction

Cancer is an evolutionary disease characterized by the accumulation of somatic mutations originating from a founder cell. As tumors grow, they diversify into genetically distinct subpopulations known as subclones [1]. Reconstructing the evolutionary history of these subclones provides critical insights into cancer diagnosis, prognosis, and the development of targeted therapies [1], [2]. One method for inferring tumor evolution is the Variant Allele Frequency Factorization Problem (VAFFP). VAFFP decomposes observed mutation frequencies from tumor samples into two components: a usage matrix describing the proportions of different subclones within each sample, and a binary mutation matrix representing the presence or absence of mutations in each subclone. A key constraint in VAFFP is the assumption of a perfect phylogeny, meaning that each mutation arises once and is never lost. This assumption ensures that tumor evolution follows a tree-like pattern of mutation accumulation [3]. While VAFFP has been effective for analyzing tumor evolution in individual patients, it does not model shared evolutionary patterns across multiple patients [3], [4]. Identifying conserved evolutionary trajectories, which are repeated mutation patterns or branching orders observed across patients, is important for discovering common drivers of cancer progression and for developing broadly applicable treatments [4]. This paper proposes an extension of the matrix factorization framework of VAFFP to model multi-patient datasets. The model infers a shared binary mutation matrix across patients while allowing for patient-specific usage matrices to capture individual clonal compositions. This extension aims to preserve the biological constraints of VAFFP (perfect phylogeny, sum, and ancestry conditions) while enabling the detection of conserved evolutionary structures across patients. Additionally, a comparative analysis of this proposed extension against related approaches, including TreeMHN [2], Sapling [5], and MASTRO [6], is presented. These models address aspects of tumor evolutionary modeling but differ in their assumptions, problem formulations, and ability to capture cross-patient structures.

2. Background

2.1 Tumor Evolution - Cancer arises through the accumulation of somatic mutations within an initially normal cell [1]. Over time, these mutations lead to the emergence of genetically distinct subpopulations of tumor cells known as subclones. This process, referred to as tumor evolution, results in significant intra-tumor heterogeneity, which is the coexistence of multiple genetically divergent subclones within a single tumor [2]. Understanding the evolutionary relationships among subclones provides invaluable insights into the progression of the disease and has critical implications for diagnosis, prognosis, and treatment selection [1]. Tumor phylogenies are typically represented as rooted trees, where each node corresponds to a subclone or a set of mutations, and edges indicate ancestral relationships. Accurate reconstruction of these trees allows researchers to infer the order of mutational events and the evolutionary dynamics underlying tumor evolution and development [2].

2.2 Variant Allele Frequency Factorization Problem (VAFFP) - VAFFP provides a mathematical framework for reconstructing tumor evolutionary histories from bulk sequencing data [3]. In VAFFP, the observed variant allele frequencies (VAFs) across multiple tumor samples are organized into a mutation frequency matrix $F \in R^{m \times n}$, where m denotes the number of samples and n denotes the number of observed somatic mutations. Each entry F_{ij} represents the measured frequency of mutation j in sample i. The goal of VAFFP is to decompose F into two matrices: a usage matrix $U \in R^{m \times k}$ and a binary mutation matrix $B \in \{0,1\}^{k \times n}$, where k is the number of inferred subclones [3]. Each entry U_{ik} indicates the proportion of subclone k present in sample i, while $B_{kj} = 1$ indicates that subclone k harbors mutation j, and $B_{kj} = 0$ otherwise. The matrix factorization approximates the observed mutation frequencies as:

$$F \approx U \times B$$

with each mutation frequency represented as a weighted sum of the mutational profiles of different subclones. Beyond simply decomposing F, VAFFP imposes additional biological constraints to ensure that the reconstructed clonal structures are consistent with tumor evolution models. Specifically, the inferred binary matrix B must represent a valid mutational history under a perfect phylogeny assumption. Informally, a perfect phylogeny asserts that each mutation arises exactly once during evolution and is inherited by all descendant subclones. A detailed discussion of perfect phylogeny and its mathematical properties follows in Section 2.3. In addition to enforcing perfect phylogeny, VAFFP requires that two consistency conditions are satisfied. The first condition is the Sum Condition, and it ensures that for each sample p and mutation j, the observed mutation frequency is the sum of the proportions of all subclones harboring that mutation:

$$F_{pj} = \sum_{k} U_{pk} B_{kj}$$

Second, the Ancestry Condition requires that for each sample p, and for any two mutations where mutation i is an ancestor of mutation j in the phylogeny, the observed frequency of i must be greater than or equal to that of j:

$$F_{pj} \geq F_{pk}$$

These conditions ensure that both the evolutionary relationships among mutations and the mixture proportions of subclones are reflected accurately in the reconstructed matrices. VAFFP has been proven to be computationally hard:

Theorem (El-Kebir et al., 2015): VAFFP is NP-complete [3].

Solving VAFFP exactly requires addressing a challenging combinatorial optimization problem, which is typically formulated as an integer linear program [3]. In practice, approximation algorithms and heuristic methods are often employed to accommodate noise, incomplete information, and computational complexity.

2.3 Perfect Phylogeny and its Properties - The concept of perfect phylogeny plays a central role in modeling tumor evolution, as it captures the biological assumption that somatic mutations arise once and are inherited by all descendant cells [1]. Under perfect phylogeny, the mutational history of a tumor can be represented as a rooted tree in which each mutation labels exactly one edge, and once a mutation appears, it persists in all descendants without being lost or reverted. Formally, a binary mutation matrix $B \in \{0,1\}^{k \times n}$, where k is the number of subclones and n is the number of mutations, is said to admit a perfect phylogeny if the inheritance relationships among mutations satisfy specific combinatorial properties [7]. One classic characterization of perfect phylogeny is the Containment Condition, which states that:

Theorem (Estabrook et al., 1975): A binary matrix B admits a perfect phylogeny if and only if, for every pair of mutations (i,j), the set of subclones harboring mutation i is either a subset of, a superset of, or disjoint from the set of subclones harboring mutation j [7].

In other words, for any two mutations, their occurrence across subclones must not conflict: either one mutation always appears alongside the other (containment), or they appear independently (disjoint sets).

This containment relationship naturally induces a tree-like structure on the set of mutations. An equivalent condition, often more convenient for algorithmic checks, is the Three-Gamete Condition, which states that:

Theorem (Gusfield et al., 1991): A binary matrix B admits a perfect phylogeny if and only if, for every pair of mutations (i,j), the rows of B restricted to columns i and j do not exhibit all three of the following binary pairs: (1,1),(1,0),(0,1) across different subclones [8].

Perfect phylogeny is critical in VAFFP because it ensures that the reconstructed subclonal structure reflects a biologically plausible model of tumor progression. By enforcing either the containment condition or the three-gamete condition on the binary matrix B, VAFFP ensures that the inferred mutation tree adheres to evolutionary principles consistent with clonal expansion, stability of mutations, and tumor heterogeneity [3].

2.4 Motivation for Multi-Patient Analysis - While VAFFP has been proven effective for reconstructing tumor evolution from single-patient sequencing data, it does not model shared evolutionary patterns across multiple patients. However, there is increasing biological evidence that certain mutational trajectories such as the sequential acquisition of key oncogenic drivers recur across tumors of the same cancer type [1], [2]. These conserved evolutionary trajectories may reflect common selective pressures in tumor development and progression. Detecting such conserved patterns has significant biological and clinical implications, including revealing universal mechanisms of tumorigenesis, identifying early biomarkers of disease progression, and uncovering potential therapeutic vulnerabilities shared across patients. Additionally, understanding the extent to which tumor evolution is deterministic versus stochastic across individuals remains a fundamental question in cancer biology. Traditional single-patient models, including VAFFP, operate independently for each patient and thus cannot identify trajectories that are recurrently conserved across individuals. Extending VAFFP to a multi-patient setting would enable the discovery of shared subclonal structures by jointly factorizing mutation frequency matrices from multiple patients. Such an approach could reveal conserved mutation orders or subclonal architectures while preserving biological constraints such as perfect phylogeny. This conceptual extension of VAFFP aims to capture conserved evolutionary structures across patients, providing new insights into tumor evolution at the population scale.

3. Problem Formulation

VAFFP models the mutation history of a tumor based on observed mutation frequencies from sequencing data. Given a mutation frequency matrix $F \in R^{m \times n}$, where *m* is the number of samples and *n* is the number of mutations, VAFFP's goal is to decompose F into a usage matrix $U \in R^{m \times k}$ and a binary matrix $B \in \{0,1\}^{k \times n}$ where U_{ik} represents the proportion of clone k in sample i, and $B_{kj} = 1$ if mutation j is present in clone k [3]. While VAFFP has been effective for reconstructing tumor evolution in individual patients, it does not capture evolutionary structures that may be conserved across multiple patients [3], [4]. Detecting such conserved patterns could provide insights into shared mechanisms of cancer progression and inform the development of specialized treatments. Therefore, the following problem statement is presented:

Problem 3.1 (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 $\mathrm{m}_i$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 $F^{(i)} \approx U^{(i)} \cdot B$ for all patients and that the matrices satisfy the perfect phylogeny, sum, and ancestry constraints.*

This formulation enables the discovery of conserved evolutionary trajectories across tumors, while maintaining biologically realistic mutational structures. All inferred matrices must satisfy the perfect phylogeny constraint, the sum condition, and the ancestry condition as described in Section 2.2 [7] [8].

4. Algorithmic Approach

4.1 Overview - The proposed framework extends VAFFP to jointly analyze multiple patients by inferring a shared binary mutation matrix B and patient-specific usage matrices $U^{(i)}$, subject to the same biological constraints in the single-patient setting. This enables the identification of conserved evolutionary trajectories across patients while maintaining interpretability. Although no solver is developed in this work, the formulation suggests a natural extension of existing integer linear programming strategies used for VAFFP. Additional constraints would ensure that the shared matrix B satisfies perfect phylogeny, sum, and ancestry consistency across all patients. This conceptual direction provides a basis for future implementation and evaluation.

4.2 Comparison to TreeMHN – TreeMHN's problem statement is as follows:

Problem 4.1 (TreeMHN). *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 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].

Both the TreeMHN model and the proposed extension of VAFFP aim to uncover conserved evolutionary structures across cancer patients by integrating data from multiple tumors. In both cases, the core objective is to move beyond individual tumor analysis and identify mutational patterns that recur across patients. However, the two models differ significantly in their input assumptions, biological constraints, and mathematical goals. TreeMHN operates on pre-inferred tumor trees and models the probabilistic influence of one mutation on another using a Mutual Hazard Network. In contrast, the proposed VAFFP extension works directly on raw mutation frequency matrices and infers a shared binary mutation matrix B along with patient-specific usage matrices $U^{(i)}$, subject to perfect phylogeny, sum, and ancestry constraints. TreeMHN does not explicitly enforce a perfect phylogeny constraint and allows for more flexible modeling, including mutation loss or recurrence and mutual exclusivity. The proposed approach, by contrast, enforces strict perfect phylogeny, assuming each mutation arises once and is inherited by all descendant subclones. Additionally, TreeMHN does not formally claim NP-hardness, while the original VAFFP formulation has been proven to be NP-complete [3]. These distinctions underscore that while both methods seek conserved evolutionary insights, the proposed formulation emphasizes biologically grounded, constraint-driven matrix modeling, in contrast to TreeMHN's probabilistic and tree-based formulation.

4.3 Comparison to Sapling – Sapling's problem statement is as follows:

Problem 4.2 (Sapling): *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].

Both Sapling and the proposed extension of VAFFP aim to reconstruct tumor evolutionary histories in a way that accounts for uncertainty and biological complexity. Both methods operate on bulk sequencing data and seek to explain the relationships among mutations observed in tumor samples. However, Sapling is fundamentally focused on summarizing the space of plausible evolutionary trees for a single patient. It begins with read-level data and generates a large number of trees consistent with the data above a likelihood threshold, which are then summarized into a small set of backbone trees. These trees aim to highlight the most frequent ancestral relationships, but no modeling is performed across patients. In contrast, the proposed extension of VAFFP is designed specifically to model multiple patients simultaneously by inferring a shared binary mutation matrix B and individual usage matrices $U^{(i)}$, subject to the biological constraints of perfect phylogeny, sum consistency, and ancestry consistency. Additionally, Sapling does not attempt to model conserved evolutionary trajectories across patients, nor does it enforce a perfect phylogeny constraint beyond what is implicitly assumed in the sampled tree space. The formulation underlying Sapling is proven to be NP-hard for both backbone tree inference and expansion [5], while VAFFP is known to be NP-complete [3]. Overall, while both methods address evolutionary reconstruction under biological constraints, the proposed VAFFP extension differs in scope (multi-patient vs. single-patient), input abstraction (frequency matrices vs. read counts), and mathematical formulation (matrix factorization vs. tree summarization).

4.4 Comparison to MASTRO – MASTRO's problem statement is as follows:

Problem 4.3 (MASTRO): *Given a multiset of rooted tumor phylogenies $D = \{T_1, \dots, T_n\}$, 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].

Both MASTRO and the proposed extension of VAFFP aim to discover evolutionary patterns that are shared across multiple tumors. In both approaches, the central motivation is to move beyond modeling individual tumors in isolation and instead identify commonalities in the evolutionary processes underlying different patients' cancers. However, MASTRO operates on a set of previously inferred tumor phylogenies and seeks to mine statistically significant mutational trajectories that recur across patients. It does not model tumor progression directly from raw mutation frequency data but instead treats trees as inputs and focuses on mining frequent mutation patterns using techniques adapted from itemset mining. In contrast, the proposed VAFFP extension begins with raw mutation frequency matrices $F^{(i)}$ and infers a shared binary matrix B alongside patient-specific usage matrices $U^{(i)}$, subject to perfect phylogeny, sum consistency, and ancestry consistency. Moreover, MASTRO does not enforce a perfect phylogeny constraint across tumors and allows for a much more flexible definition of evolutionary trajectories. It identifies partial orders among mutations without requiring tree-like structures or strict inheritance models. The Frequent Maximal Trajectories (FMT) problem that MASTRO solves is proven to be NP-hard, even when the recurrence threshold r equals the number of input tumors [6]. By contrast, the proposed VAFFP extension maintains strict biological realism through perfect phylogeny assumptions and provides a structured factorization framework that simultaneously models both patient-specific heterogeneity and cross-patient conservation.

Results

The conceptual extension of VAFFP proposed in this work offers a framework for uncovering conserved evolutionary trajectories across multiple cancer patients. By enforcing a shared binary mutation matrix B and patient-specific usage matrices $U^{(i)}$, the model captures conserved mutational patterns while preserving individual subclonal compositions under biologically realistic constraints. Unlike existing methods such as

TreeMHN, Sapling, and MASTRO, which operate on inferred trees or single-patient models, this approach directly models observed mutation frequency data via matrix factorization, preserving quantitative signal. This framework jointly models patient-specific heterogeneity and cross-patient conservation, addressing the biological question of identifying conserved mutation structures. It offers a principled foundation for future solver design and generalizations, such as relaxing the perfect phylogeny assumption or integrating longitudinal samples. Although this study does not present empirical results, future benchmarking against existing frameworks would further validate its utility. Ultimately, recovering shared clonal structures from raw data could advance applications such as identifying population-level therapeutic targets and characterizing evolutionary subtypes in cancer.

5. Conclusions

Understanding conserved evolutionary trajectories across cancer patients remains a fundamental challenge in computational oncology. While existing models such as TreeMHN, Sapling, and MASTRO address parts of this problem, they do not jointly model raw mutation frequency data across patients under perfect phylogeny constraints. This work proposes an extension of the Variant Allele Frequency Factorization Problem to fill this gap. The formulation infers a shared binary mutation matrix B and patient-specific usage matrices $U^{(i)}$, across multiple patients, subject to perfect phylogeny, sum, and ancestry constraints. Conceptually, the model detects conserved mutational structures while maintaining clonal heterogeneity unique to each individual. It retains biological interpretability by enforcing consistent evolutionary relationships. By identifying recurrent mutation trajectories, this framework may aid in the discovery of population-level therapeutic targets or early biomarkers of cancer progression. Although empirical validation remains future work, this formulation offers a principled foundation for recovering shared evolutionary pathways directly from bulk sequencing data.

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.