Beyond Perfect Phylogeny: Multisample Phylogeny Reconstruction via ILP

Paola Bonizzoni
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Viale Sarca 336, 20126
Milano
bonizzoni@disco.unimib.it

Simone Ciccolella
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Viale Sarca 336, 20126
Milano
s.ciccolella@campus.unimib.it

Gianluca Della Vedova
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Viale Sarca 336, 20126
Milano
gianluca.dellavedova@unimib.it

Mauricio Soto
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Viale Sarca 336, 20126
Milano
mauricio.sotogomez@unimib.it

ABSTRACT

Most of the evolutionary history reconstruction approaches are based on the infinite site assumption which is underlying the Perfect Phylogeny model. This is one of the most used models in cancer genomics. Recent results gives a strong evidence that recurrent and back mutations are present in the evolutionary history of tumors [19], thus showing that more general models then the Perfect phylogeny are required. To address this problem we propose a framework based on the notion of Incomplete Perfect Phylogeny. Our framework incorporates losing and gaining mutations, hence including the Dollo and the Camin-Sokal models, and is described with an Integer Linear Programming (ILP) formulation. Our approach generalizes the notion of persistent phylogeny [1] and the ILP approach [14, 15] proposed to solve the corresponding phylogeny reconstruction problem on character data.

The final goal of our paper is to integrate our approach into an ILP formulation of the problem of reconstructing trees on mixed populations, where the input data consists of the fraction of cells in a set of samples that have a certain mutation. This is a fundamental problem in cancer genomics, where the goal is to study the evolutionary history of a tumor. An experimental analysis shows that our ILP approach is able to explain data that do not fit the perfect phylogeny assumption, thereby allowing (1) multiple losses and gains of mutations, and (2) a number of subpopulations that is smaller than the number of input mutations.

CCS CONCEPTS

• **Applied computing** → *Bioinformatics*;

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA

ACM ISBN 978-1-4503-4722-8/17/08…$15.00
https://doi.org/10.1145/3107411.3107441

KEYWORDS

phylogeny; clonal reconstruction

ACM Reference format:
Paola Bonizzoni, Simone Ciccolella, Gianluca Della Vedova, and Mauricio Soto. 2017. Beyond Perfect Phylogeny: Multisample Phylogeny Reconstruction via ILP. In *Proceedings of ACM-BCB '17, Boston, MA, USA, August 20-23, 2017,* 10 pages.
https://doi.org/10.1145/3107411.3107441

1 INTRODUCTION

Character-based phylogeny reconstruction is one of the fundamental problems in Bioinformatics, with a large literature [10, 13, 25, 26] focusing on a simple assumption: the input data consists of a set of species (or individuals) for which we know the set of characters that it possesses. In this case, the goal is to compute a phylogeny that explains the set of input species and characters, where each edge of the phylogeny allows characters gains and losses. Character-based phylogenies play a crucial role in modeling the evolution in cancer genomics. Cancer is an uncontrolled evolutionary process of somatic mutations of tumor cells from a single founder cell [11] creating a diverse set of subpopulations [6, 20, 27], each originated from a single *clone*: each clone (and each subpopulation) has a distinctive set of mutations. From this point of view, a tumor progression is a phylogeny where clones and mutations have the same role as species and mutations in the classical phylogeny reconstruction setting as characters.

To fall within the classical framework we would need to obtain data directly from a cell. Unfortunately, single cell sequencing is not cheap [21] and is prone to errors, therefore we have to study *samples* comprising lots of cells belonging to an unknown set of subpopulations. This adds a new complication, since for each sample we know the (approximate) fraction of cells that have a given somatic mutation. More precisely, each read extracted from the sample is mapped against the reference genome, therefore we obtain the mutations of each read. Since reads can be extracted from repeated regions of the genome and the coverage of the reads is not uniform throughout the genome or the cells of the sample, the

fraction of reads that have a mutation is only an approximation of the fraction of cells of the sample that have that mutation. In other words, the observed frequencies are an estimate of the actual frequencies of the mutation.

The above argument leads to define a computational problem called *variant allele frequency factorization problem* (VAFFP) [7, 8, 16], where the input is the observed frequencies of the mutation in each sample and the desired output is a phylogeny representing the tumoral evolution, as well as the composition of each sample in terms of the subpopulations or clones. The literature has mainly focused on the binary perfect phylogenies [7], where samples contain mixtures of two-state characters, i.e. where each character/locus is either mutated or not and mutation can be gained only once and never lost in the entire history of the tumor. A possible generalization (that we do not explore in this paper) to the multi-state perfect phylogeny has been recently proposed in order to take into account the effect of copy number aberrations on alleles [8]. Then characters that can assume different states (multi states), but as in the binary case changes to the same state occur only once. This restriction — known as the infinite site assumption — allows to obtain efficient algorithms, but most recent studies refutes it [19] and state that more complex models are needed to describe the tumor evolution. In this paper we describe some efficient approaches that overcome this limitation and allow to reconstruct phylogenies that are more general than perfect phylogeny and are able to capture a likely evolutionary history of the tumor studied.

We will focus on three main character-based models: the Persistent Phylogeny [1] (where each character can be gained once and lost at most once), Camin-Sokal [5] (where each character can be gained several times, but never lost), and Dollo [9] (where each character can be gained at most once, but lost several times). We denote by Camin-Sokal(k) the restriction of the Camin-Sokal model where each character can be gained at most k times in the entire tree. Moreover, we denote by Dollo(k) the restriction of the Dollo model where each character can be lost at most k times in the entire tree. Clearly, the Persistent Phylogeny [1] corresponds to the Dollo(1) model which has been recently investigated in several works aiming to develop efficient solutions for the model [3, 4, 14] since its use is motivated also in other contexts [2, 23]. In particular, in [1] it is proved that the Persistent Phylogeny Problem over a binary matrix M can be formulated as finding a special completion of an extended matrix M_e that is a Perfect Phylogeny. Based on this characterization, an ILP formulation for the Persistent Phylogeny has been developed in [14]. In [7] the approach used to solve the VAFFP problem is a combination of an integer linear programming (ILP) formulation and a clever approach to compute the set of relevant phylogenies, based on the notion of ancestry graph. Since the last component is tightly coupled with the fact that perfect phylogenies have as many species as characters, it is not immediate to extend the approach of [7] to more general models. Starting, from this ILP formulation and the main characterization in [1], we developed a novel approach to the VAFFP problem that is entirely based on ILP and allows to take into account the three evolutionary models presented above. We have experimented our ILP approach on simulated and real data to test whether allowing the models to violate the infinite site assumption leads to better solutions, even when the Perfect Phylogeny model is able to explain the input data. Indeed, our experiments show that the Persistent phylogeny may explain the input data better than the Perfect Phylogeny while requiring a number of clones that is smaller than the number of mutations. Finally, the inferred tree from real data on a Leukemia tumor CLL077 reveals the losses of a mutation, though being the tree mostly consistent with the one reconstructed by other known methods [18].

2 PRELIMINARIES

The character-based phylogeny reconstruction problems we study in this paper are constrained version of the general Incomplete Directed Perfect Phylogeny (IDP) [22], where the input is an $n \times m$ matrix $M_?$, where $M_?(i, j) \in \{0, 1, ?\}$ represents the absence, presence or uncertainty of a character j in the species i respectively. A solution consists of changing each ? into 0 or 1 obtaining a new binary matrix M_s that has a directed perfect phylogeny (since all phylogenies of this paper are directed, we will skip this fact). The unconstrained IDP problem has a $O(mn \log^2(m + n))$-time algorithm [22].

A binary matrix M_s has a perfect phylogeny if and only if there are not two columns containing all the pairs $(0, 0)$, $(0, 1)$, $(1, 0)$ — two columns containing all those pairs are called incompatible or conflicting. The problem of determining if a binary matrix has a perfect phylogeny, and to compute such perfect phylogeny if possible, has a linear-time algorithm [12, 13]. Moreover, there exists an ILP formulation for determining if a binary matrix has a perfect phylogeny [15]. Since finding a perfect phylogeny is easy, the main difficulty in solving the IDP problem consists of determining if each ? must be replaced with a 0 or a 1.

In [1] the Persistent Perfect Phylogeny has been restated as a constrained IDP problem. More precisely, given a binary matrix M, they solve the IDP problem on an extended matrix M_e where each entry $M[s, c]$ is replaced by two entries $M_e[s, c^+]$ and $M_e[s, c^-]$ as follows: if $M[s, c] = 1$ then $M_e[s, c^+] = 1$ and $M_e[s, c^-] = 0$; if $M[s, c] = 0$ then $M_e[s, c^+] = M_e[s, c^-] =?$. The constraint is that, for each pair $(M_e[s, c^+], M_e[s, c^-])$ of ? entries, the corresponding entries in the matrix M_s must be the same, that is $M_e[s, c^+] = M_e[s, c^-]$. This additional constraints make impossible to use the algorithm of [22], but some algorithms have been developed [3, 4], including one based on an ILP formulation [14]. The latter algorithm can be trivially extended to compute the perfect phylogeny with the fewest edges.

A $p \times m$ *frequency* matrix F, contains the frequencies of the mutation in a set of samples. More precisely, each entry $F[t, j]$ indicates the proportion of cells in sample t with the mutation j. A $p \times n$ *usage* matrix U, contains the mixture of cells in each sample. More precisely, each entry $U[t, i]$ is the proportion of the cells in the sample t belonging to the subpopulation i. Finally, the $n \times m$ (clonal) matrix M contains which subpopulation has a given mutation. An evolution model $\mathcal{M}$ consists of a set of constraints that a phylogeny T realizing the clonal matrix M must obey. For example, when the evolution model is the persistent phylogeny, then the phylogeny T cannot have two edges corresponding to two gains or two losses of the same character. The $\mathcal{P}$-VAFF problem can be formally defined as follows.

Definition 2.1. Given a $p \times m$ frequency matrix F, a number of clones n, and an evolution model $\mathcal{P}$, the $\mathcal{P}$-VAFFP (short for $\mathcal{P}$-Variant Allele Frequency Factorization Problem) asks for an $p \times n$ usage matrix U and an $n \times m$ clonal matrix M such that (1) $F = \frac{1}{2}UM$, and (2) M admits a phylogeny under the model $\mathcal{P}$.

The $1/2$ factor in the definition is a technical consequence of the fact that the healthy (wild type) cell subpopulation exists, but is not one of the clones of M, and human beings are diploid, that is they have two copies of each chromosome.

We decouple the $\mathcal{P}$-VAFFP problems into two different problems: computing M and U. In fact, once we have computed a clonal matrix M, the problem of finding a composition of samples, i.e. a usage matrix U, compatible with M consists of finding a matrix U such that $\sum_{i=1}^{n} U(t,i)M(i,j) = F(t,j)$ and $\sum_{i=1}^{n} U(t,i) \leq 1 \; \forall t, j$.

In our setting, clonal matrix M is an unknown variable generating two main issues. First, the restrictions regarding the sample proportions become non linear. Second, we must ensure that clones can be represented under parsimony rule $\mathcal{P}$. The first difficulty can be easily resolved since the product of two $\{0,1\}$ variables can be expressed as a set of linear constraints. We detail this technique in section 4. Second issue is a more daunting task since involves the recognition problem for general parsimony rules. We deal with this issue in section 3 by proposing an ILP formulation for the reconstruction problem.

3 THE $\mathcal{P}$ PHYLOGENY RECONSTRUCTION PROBLEM

This section is devoted to the development of an ILP formulation for the following decision problem:

$\mathcal{P}$ Phylogeny Reconstruction Problem. Given a character-based phylogeny model $\mathcal{P}$ and a binary matrix M, decide if M admits a representation in the phylogeny model $\mathcal{P}$.

In this paper we focus on Dollo(k) and Camin-Sokal(k) models.

Our strategy is based on the approach discussed by Gusfield in [14] for the Persistent Phylogeny reconstruction problem. The formulation presented by Gusfield is founded on two main results:

(1) Every instance of the Persistent Perfect Phylogeny reconstruction problem consisting of a binary matrix M can be reduced to an instance M_e, consisting of a matrix over alphabet $\{0, 1, ?\}$, called *extended matrix*, of an equivalent Incomplete Directed Perfect Phylogeny problem with additional constraints [1].
(2) The Incomplete Directed Phylogeny problem can be stated as an ILP problem by minimizing the number of conflicts between characters [15].

We will extend the aforementioned results to solve the Dollo(k) (Camin-Sokal(k)) Phylogeny Reconstruction Problem by generalizing the construction of the extended matrix proposed in [1] to those more general models. Additionally, we extend the ILP formulation presented in [15] in order to incorporate constraints imposed by our generalization of (1). The later result will allow us to develop an ILP formulation for finding a Dollo(k) or Camin-Sokal(k) tree representation for the input matrix M of the $\mathcal{P}$ Phylogeny Reconstruction Problem.

3.1 Construction of the Extended Matrices

Definition 3.1. Given an incomplete binary matrix M and a set $\mathcal{R} = \{F_i(M) \leq 0\}_{i \in [1,r]}$ of r constraints, the *Modified Incomplete Directed Perfect Phylogeny for the set $\mathcal{R}$*, denoted by MIDPP($M, \mathcal{R}$), asks for finding a completion of matrix M which admits a Perfect Phylogeny and satisfies all constraints in $\mathcal{R}$

Is easy to see that if every constraint in $\mathcal{R}$ can be expressed as a linear constraint in terms of the matrix entries, then the problem MIDPP($M, \mathcal{R}$) admits a ILP formulation. The formulation can be obtained by simply adding the set of (linear) constraint $\mathcal{R}$ to the basic ILP formulation presented in [15].

Therefore, in the following we show the construction of the extended matrices and the set of constraints for Dollo(k) and Camin-Sokal(k) models.

Extended Matrix and constraints for Dollo(k). Let M be a binary matrix with n species and m characters. The extended matrix $M_{e,D(k)}$ for the Dollo(k) model is defined as follows:

- $M_{e,D(k)}$ has n rows and $m \times (k+1)$ columns, where each character j of matrix M is associated to $k+1$ columns in $M_{e,D(k)}$ denoted by $j_0, j_1, \ldots, j_k$.
- If $M(i,j) = 1$ then $M_{e,D(k)}(i, j_0) = 1$ and $M_{e,D(k)}(i, j_l) = 0$, $l \in [1,k]$.
- If $M(i,j) = 0$ then $M_{e,D(k)}(i, j_l) =?$ for each $l \in [0,k]$.

For a character j, the column j_0 represents the acquisition of character j while each of the following k columns represents a possible loss of the gained character. In every feasible solution, a character can be gained/lost at most once along any path from the root, therefore, if $M(i,j) = 0$ it must hold that $\sum_{1 \leq l \leq k} M_{e,D(k)}(i, j_l) = M_{e,D(k)}(i, j_0)$. Let us define the following set of constraints for the matrix $M_{e,D(k)}$:

$$\mathcal{R}_{D(k)} = \left\{ \sum_{1 \leq l \leq k} M_{e,D(k)}(i, j_l) = M_{e,D(k)}(i, j_0), (i,j) : M(i,j) = 0 \right\} \tag{1}$$

Extended Matrix and constraints for Camin-Sokal(k). Let M be a binary matrix with n species and m characters. The extended matrix $M_{e,CS(k)}$ for the Camin-Sokal(k) model is defined as follows:

- $M_{e,CS(k)}$ has n rows and $m \times k$ columns; each character j of matrix M is associated to k columns in $M_{e,CS(k)}$ denoted by $j_1, \ldots, j_k$.
- If $M(i,j) = 0$ then $M_{e,CS(k)}(i, j_l) = 0$, $l \in [1,k]$.
- If $M(i,j) = 1$ then $M_{e,CS(k)}(i, j_l) =?, l \in [1,k]$.

Every group of columns $j_1, \ldots, j_k$ represent the possible gain of a character in the resulting phylogenetic tree. In every feasible solution, a character can be gained at most once on any path from the root to a leaf, therefore we define the set following set of constrains for the extended matrix $M_{e,CS(k)}$:

$$\mathcal{R}_{CS(k)} = \left\{ \sum_{1 \leq l \leq k} M_{e,CS(k)}(i, j_l) = 1, (i,j) : M(i,j) = 1 \right\} \tag{2}$$

Thus, the Dollo(k) ((Camin-Sokal(k)) Phylogeny Reconstruction Problem is equivalent to solve the MIDPP problem over an extended

matrix $M_{e,D(k)}$ and constraints $R_{D(k)}$, as stated in the following Theorem that generalizes the main Theorem in [1].

Theorem 3.2. *Given a binary matrix M, then M admits a phylogeny under the Dollo(k) model if and only if the decision problem MIDPP($M_{e,D(k)}, \mathcal{R}_{D(k)}$) admits a solution.*

Proof. ($\Rightarrow$) Let M a matrix, and let T be a Dollo(k) phylogeny representing M. For each character j we relabel T as follows: the edge labeled as j^+ will have label j_0 while edges labeled with j^- are relabeled arbitrarily from the set $\{j_1, \ldots, j_k\}$ (no two edges have the same label). Let $\hat{T}$ be the tree obtained from M after relabeling. It is immediate to notice that $\hat{T}$ is a perfect phylogeny.

Let $\hat{M}$ be the matrix associated with $\hat{T}$. Let s be a generic species, and let c be a character. If s has the character j in M, then the path from s to the root of $\hat{T}$ traverses the edge j_0 but not any of the edges j_h with $h > 0$, that is $\hat{M}(i, j_h) = 0$ for $h > 0$ and $\hat{M}(i, j_0) = 1$. If s does not have the character j in M, then either the path from s to the root of $\hat{T}$ does not traverse the edge j_0, or it traverses the edge j_0 and one of the edges j_h with $h > 0$. In both cases $\sum_{h=1}^{k} \hat{M}(i, j_h) = \hat{M}(i, j_0)$, hence $\hat{M}$ satisfies (1).

($\Leftarrow$) Conversely, let M be a binary matrix whose corresponding instance of MIDPP($M_{e,D(k)}, \mathcal{R}_{D(k)}$) is solved by the extended matrix $\hat{M}$. We will proof that M admits a Dollo(k) representation.

Let $\hat{T}$ be a perfect phylogeny solving $\hat{M}$. Let T be the phylogeny obtained from $\hat{T}$ by replacing each label j_0 with j_+ and any j_h with $h > 0$ with j^-. Just as for the first part of the proof, since the assignment $\hat{M}$ satisfies (1), for each species s and each character j of M, one of the following three cases holds: (1) $\hat{M}(i, j_h) = 0$ for $h > 0$ and $\hat{M}(i, j_0) = 1$, (2) $\sum_{h=1}^{k} \hat{M}(i, j_h) = \hat{M}(i, j_0) = 0$, (3) $\sum_{h=1}^{k} \hat{M}(i, j_h) = \hat{M}(i, j_0) = 1$. The same argument of the first part of the proof shows that in case (1) the path from s to the root of $\hat{T}$ traverses the edge j_0 but not any of the edges j_h with $h > 0$: hence the path from s to the root of T traverses the edge j^+ but not any of the edges j^-. In case (2), the path from s to the root of T does not traverse the edge j^+, while in case (3) the path from s to the root of T traverses the edge j^+ and of the edges j^-. We conclude that the tree T is a Dollo(k) phylogeny for the matrix M. □

Notice that a similar proof shows that the binary matrix M admits a phylogeny under the Camin-Sokal(k) model if and only if the decision problem (MIDPP($M_{e,CS(k)}, \mathcal{R}_{CS(k)}$)) has a solution. Moreover, it is quite natural to observe that every Dollo(k) (Camin-Sokal(k)) phylogeny for the matrix M is uniquely associated to a completion of the extended matrix which is instance to the problem MIDPP.

Theorem 3.2 allows us to reduce the $\mathcal{P}$ reconstruction problem to an ILP problem. As it was already mentioned, our formulation is based on the one presented in [15] for the IDP problem which we detail in the next section.

3.2 ILP formulation for the MIDPP

In this section we revisit the formulation proposed in [15] for the IDP problem.

The input of the problem is an incomplete matrix M. The goal is to decide if there exists a completion of the unknown entries of M in order to obtain a (complete) matrix admitting a Perfect Phylogeny. The main strategy is the minimization of the conflicts between pairs of characters. In virtue of the Perfect Phylogeny Theorem, the IDP problem will have a solution if and only if the optimal value of the problem is zero.

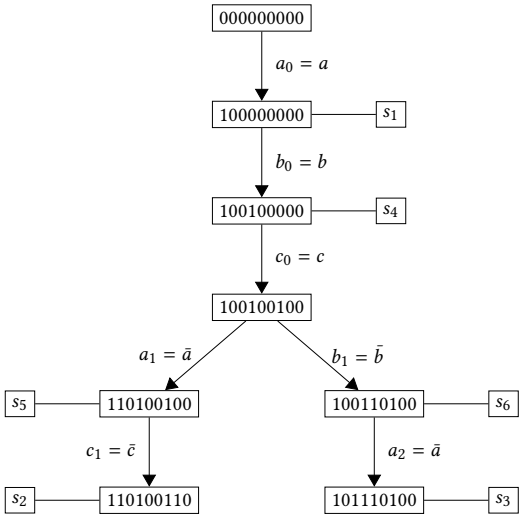
In what follows, we briefly explain the key elements of the ILP and we discuss the required changes for its use in the MIDPP problem.

Variables. We define a binary variable $Y(i, j)$ for each unknown position of M. With abuse of notation, $Y(i, j)$ will be a constant for every known position of the matrix of value $M(i, j)$. Since the objective is to determine if two columns are in conflict, for every pair of columns p, q we define a binary variable $C(p, q)$ that indicates the existence of a conflict between these two columns. To establish if two columns are in conflict, binary variables $B(p, q, a, b)$ are defined

	a	b	c
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	0
5	0	1	1
6	1	0	1

	a_0	a_1	a_2	b_0	b_1	b_2	c_0	c_1	c_2
1	1	0	0	?	?	?	?	?	?
2	?	?	?	1	0	0	?	?	?
3	?	?	?	?	?	?	1	0	0
4	1	0	0	1	0	0	?	?	?
5	?	?	?	1	0	0	1	0	0
6	1	0	0	?	?	?	1	0	0

Figure 1: Input matrix M (left) and the corresponding Dollo(2) extended matrix (right)



	a_0	a_1	a_2	b_0	b_1	b_2	c_0	c_1	c_2
1	1	0	0	0	0	0	0	0	0
2	1	1	0	1	0	0	1	1	0
3	1	0	1	1	1	0	1	0	0
4	1	0	0	1	0	0	0	0	0
5	1	1	0	1	0	0	1	0	0
6	1	0	0	1	1	0	1	0	0

Figure 2: A phylogeny for the input matrix M of Figure 1 (see Theorem 3.2) and the corresponding completion for the MIDPP($M_{e,D(k)}, \mathcal{R}_{D(k)}$).

for each pair of columns (p, q) and for each possible pair of values $(a, b) \in \{0, 1\}^2$. These variables indicate if for the (ordered) pair of columns (p, q) there exists a row i where $Y(i, p) = a$ and $Y(i, q) = b$. Similarly to the variable $Y(i, j)$, if for a pair (p, q) it already exits among known entries of the matrix a row where $Y(i, p) = a$ and $Y(j, q) = b$ then $B(p, q, a, b)$ will be considered as a constant of value 1.

Inequalities. For every pair of columns (p, q), every binary pair $(a, b) \in \{0, 1\}^2$ and every i, the following set of inequalities

$$B(p, q, a, b) \geq 1 - [a + (-1)^a \cdot Y(i, p)] - [b + (-1)^b \cdot Y(i, q)] \quad (3)$$

force the variable $B(p, q, a, b)$ to be 1 if and only if columns p, q exhibit the pair (a, b) in some rows. On the other hand, the following inequalities forces variables $C(p, q)$ to be 1 when characters p and q are in conflict.

$$C(p, q) \geq B(p, q, 0, 1) + B(p, q, 1, 0) + B(p, q, 1, 1) - 2 \quad (4)$$

Objective Function. Since we aim to minimize the number of conflicts, the objective function is defined as: $\min \sum_{(p,q)} C(p, q)$.

This formulation can be extended to solve the $\mathcal{P}$ phylogeny reconstruction problem under different $\mathcal{P}$ models Section 3 in order to find minimal conflict completions. Nevertheless, in the reconstruction problem we are mainly concern about feasible solutions with no conflicts. For this reason, we consider an alternative version of constraint (4):

$$B(p, q, 0, 1) + B(p, q, 1, 0) + B(p, q, 1, 1) \leq 2 \quad (5)$$

Hence, we can state the Dollo(k) (Camin-Sokal(k)) phylogeny reconstruction problem as any feasible solution of the set of constraints (3), (5) and $\mathcal{R}_{D(k)}$ ($\mathcal{R}_{CS(k)}$).

Additionally, since the number of total negated characters corresponds to $N(M) = \sum_{j=1}^{m} \sum_{l=1}^{k} B(j_0, j_l, 1, 1)$, we can state the problem of finding a parsimony representation with the minimum number of conflicts by considering the solution of the following minimization problem: $\min \sum_{(p,q)} C(p, q)$ s.t. (3), and (4).

4 THE CLONAL RECONSTRUCTION PROBLEM

In this section we present an ILP formulation for the $\mathcal{P}$-VAFFP. Let us recall that a $\mathcal{P}$-VAFFP instance is a $p \times m$ frequency matrix F, a number of clones n and a model $\mathcal{M}$. The objective is to find matrices U and M, defining clone populations and the configuration of different samples, such that $F = \frac{1}{2}UM$ where matrix M admits a phylogeny satisfying rule $\mathcal{P}$.

Variables. Variables are the entries of usage, clonal and extended matrices which we denote by U, M and M_e respectively. Extended matrix is constructed according to Section 3.1 based on the considered phylogeny model.

The formulation is divided in two sets of inequalities. The first part of the formulation deals with the feasibility of sample proportions for a set of clones. The second set of inequalities ensures that clones can be explained by an evolutionary history following a specific phylogeny model using the extended matrix formulation proposed in Theorem 3.2.

On the one hand, the relation between the proportion of the somatic mutations in the samples and the proportion of clones in the samples can be described as:

$$\frac{1}{2} \sum_{i=1}^{n} U(t, i)M(i, j) = F(t, j). \quad (6)$$

That is, for each sample $t \in [1, p]$, clone $i \in [1, n]$ and mutation $j \in [i, m]$ the sum of the proportions of clones present in the sample t which present mutation j must be equal to $F(i, j)$. Since each sample row t in matrix U describe a sample composition then it holds:

$$\sum_{i=1}^{n} U(t, i) \leq 1. \quad (7)$$

The (non linear) product of Equation 6 can be expressed in a linear form by introducing a set of auxiliary binary variables $X(t, i, j)$ as follows:

$$\sum_{i=1}^{n} X(t, i, j) = F(t, j), \quad X(t, i, j) \geq 0, \quad X(t, i, j) \leq M(i, j), \quad X(t, i, j) \leq U(t, i), \quad X(t, i, j) \geq U(t, i) + M(i, j) - 1 \quad (8)$$

On the other hand, we must guarantee that the clonal matrix M admits a phylogeny under the $\mathcal{P}$ model. As it was discussed in Section 3, it is possible to state the $\mathcal{P}$ phylogeny reconstruction problem as a solution for an IDP problem on the corresponding extended matrix (Theorem 3.2).

Notice that the set of constraints $\mathcal{R}_{D(k)}$ and $\mathcal{R}_{CS(k)}$ are defined for a known incomplete matrix M_e, while we have an incomplete matrix that is only described by a set of variables $M(i, j)$. Therefore we have to modify the set of constraints in order to overcome this problem. Namely, for each clone i and mutation j, we define the following set of constraints on the variables Y:

$$M_e(i, j_0) - \sum_{1 \leq l \leq k} M_e(i, j_l) = M(i, j) \text{ for the Dollo}(k) \text{ model, and} \quad (9)$$

$$\sum_{1 \leq l \leq k} M_e(i, j_l) = M(i, j) \text{ for the Camin-Sokal}(k) \text{ model.} \quad (10)$$

Thus, the $\mathcal{P}$-VAFFP corresponds to finding a feasible solution with the linear constraints (7), (8), (9), (3), (5) for the Dollo(k) model, and (7), (8), (10), (3), (5) for the Camin-Sokal(k) model.

Finally, let m_e the number of columns in the matrix M_e, our formulation has $O(nm_e + m_e^2 + mpn)$ variables and $O(m_e^2 + npm)$ constraints.

4.1 Clonal Reconstruction admitting errors

Since frequency matrices are obtained experimentally, they can only contain approximation of the actual frequencies. Hence, we have to incorporate frequency errors in the formulation of the problem, most notably in the construction of the usage matrix U. Let us define the $p \times m$ matrix E, where $E(t, j)$ represents the error of the frequency $F(t, j)$.

The set of constraints bounding the difference between input frequencies and the frequencies associated to our proposed solution (that is, corresponding to the product UM) is:

$$-E(t, j) \leq \sum_{i=1}^{n} X(t, i, j) - F(t, j) \leq E(t, j)$$

Since now our goal is to minimize the overall error introduced in the reconstruction, the objective function is:

$$\min \sum_{(t,j)\in[1,p]\times[1,m]} E(t,j)$$

5 EXPERIMENTAL RESULTS

We implemented our approach with a Python program called `gppf` that receives a frequency matrix M and outputs the corresponding ILP in a format that is then fed to Gurobi 6.5.2. The program `gppf` is available at https://github.com/AlgoLab/gppf. All experiments have been performed on an Ubuntu 14.04 server with four 8-core Intel Xeon E5-4610v2 2.30GHz CPUs (hyperthreading was enabled for a total of 16 threads per processor). The proposed ILP formulations have been experimented to test how the proposed general models (Persistent Phylogeny, Dollo(k) and Camin-Sokal(k)) fit the input data w.r.t. the Perfect Phylogeny model. For this purpose we have tested the formulation in both simulated and real data. The size of the instances are typical for real data applications such as liquid cancer and in particular Leukemia for which we show here the inferred (persistent) tree for a real instance (CLL077). For our experiments we have generated different sets of frequency matrices F as follows:

(1) generate a clonal $n \times m$ matrix M by using the tool developed in [17], called `ms`, obtaining a Perfect Phylogeny on n clones and m mutations.
(2) In order to avoid obtaining only matrices admitting a Perfect Phylogeny, some values of the original are modified randomly from 0 to 1.
(3) Generate a usage matrix U of dimensions $p \times n$ assigning to each clone a proportion in each sample. Those values are chosen randomly, following a Dirichlet distribution. The number p of samples is decided a priori and n is the number of clones imposed by M.
(4) Multiply U and M to generate a $p \times m$ frequency matrix F.

Generated matrices are used as input of our implementation for different formulation settings. Finally, we measure the quality of the predictions for each phylogeny model, by considering the total error. We remark that we do not compare the predicted clonal matrix M with the original, since different models can generate diverse clonal evolution trees.

The parameters of the implementation are the maximum number of clones that a solution can use (expressed as the percentage of the number of mutations), the maximum time permitted for each execution, the maximum number of persistent character allowed, and a parameter k associated to the model Dollo(k) and the Camin-Sokal(k) in the formulation. Moreover, we have introduced a timeout on the running time, since the generated ILP problem is often large and its resolution could require a considerable amount of time. Nevertheless, imposing a timeout allows the ILP solver to compute a solution with a small total error.

We evaluate the obtained solutions according to the following measure:

$$\text{Error}_F(\overline{F}) = \frac{\|F - \overline{F}\|}{\|F\|}$$

Where F is the input frequency matrix, $\overline{F}$ is the frequency matrix inferred by the solution, and $\|A\| = [\sum_{ij}|a_{ij}|^2]^{1/2}$ is the Frobenius norm. This metric give us the ratio between the total error and the optimal value, therefore it is not too dependent on the actual values.

Previous works focused on Perfect Phylogeny as the evolutionary model, thereby restricting the attention to a number of clones equal to the number of mutations. Since more general evolutionary models are allowed, the number of clones might be different, the user can provide the maximum number of clones. We have investigated the effect of choosing different values for such pameter. More precisely, we have explored bounding the number of clones to be at most 100%, 80%, 60% and 40% of the maximum number of clones in the instance, which is a representative set of values. We recall that these values are upper bounds, while the actual number of clones used in the actual solution might be smaller.

5.1 Simulated Data

For the simulated data, we have generated two different data sets:

Exp. 1 contains 100 frequency matrices composed of 6 samples and 10 mutations. Matrices are generated from a 20×20 clonal matrix M. The phylogenetic models tested in this set are: Perfect, Persistent, Dollo(2) and Camin-Sokal(2).

Exp. 2 contains 10 frequency matrices with 12 samples and 25 mutations, generated by a 25×50 clonal matrix M, The models tested in this set are: Perfect, Persistent, Dollo(4) and Camin-Sokal(4).

Figure 3 shows how the error of each solution varies as a function of the running time for both experiments. Analyzing those plots, we note that the total error rapidly stabilizes. Therefore we have decided to set a time limit for the running time equal to 5 minutes for Exp. 1 and 2 hours for Exp.2, since allowing a large time limit results in only marginal improvements of the quality of the solutions computed.

In almost all cases the solver uses the entire available time to find a solution. The only exception is the Perfect Phylogeny model when maximum number of clones is 40% of the mutations. Moreover, we notice that the computed solutions usually use fewer clones than maximum allowed.

Figures 4 and 5 shows the total error of the solutions obtained under different phylogenetic models and different upper bounds on the number of clones for the set Exp. 1 and Exp. 2 respectively. Additionally, Tables 1 and 2 exhibit the number of instances for which the general phylogeny models outperform the Perfect Phylogeny model on both datasets.

As we can see from Figure 4, the obtained solutions have average error smaller than the 15% of the input matrix norm. As expected, we observe that the error of all models increases as we decrease the number of maximum clones allowed. The increase in the total error is larger for the Perfect Phylogeny model, since it is not sufficient to correctly explain the instances for a small number of clones. In fact, Table 2 shows that, in almost all instances, a general phylogeny model outperform the results of the Perfect Phylogeny solution.

In Figure 5 and Table 2 we see that Dollo(4) and Camin-Sokal(4) have worse total error than more restricted models. This effect is due to the much larger size of the ILP formulation for Dollo(4) and

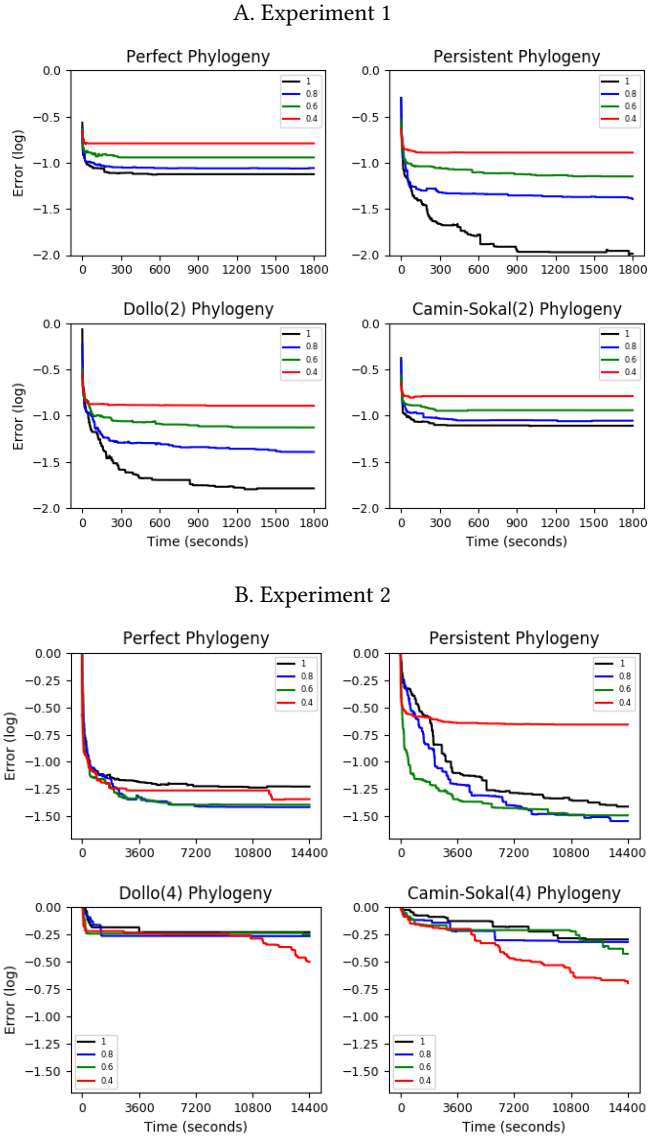


Figure 3: Mean error evolution of solutions in terms of time.

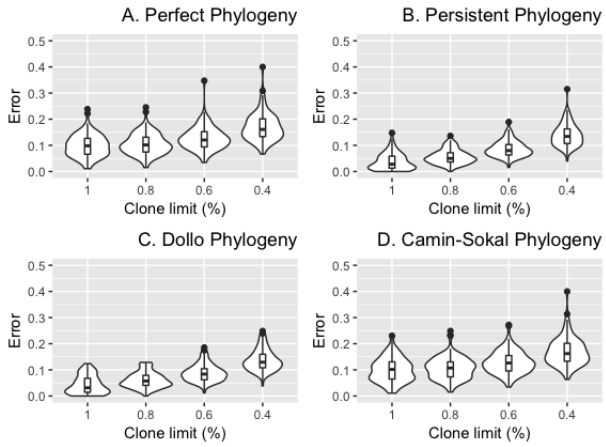


Figure 4: Error for the solutions obtained from Exp. 1 under different evolutionary models

Total error	Persistent	Dollo(2)	Camin-Sokal(2)
Clone limit 100%			
≤ PPE	100/100	94/100	47/100
≤ 90% PPE	99/100	92/100	18/100
≤ 80% PPE	97/100	86/100	10/100
≤ 50% PPE	74/100	66/100	0/100
Clone limit 80%			
≤ PPE	100/100	97/100	53/100
≤ 90% PPE	99/100	91/100	11/100
≤ 80% PPE	89/100	84/100	4/100
≤ 50% PPE	44/100	39/100	0/100
Clone limit 60%			
≤ PPE	98/100	92/100	51/100
≤ 90% PPE	91/100	83/100	2/100
≤ 80% PPE	74/100	62/100	1/100
≤ 50% PPE	13/100	17/100	0/100
Clone limit 40%			
≤ PPE	90/100	93/100	79/100
≤ 90% PPE	70/100	74/100	0/100
≤ 80% PPE	43/100	43/100	0/100
≤ 50% PPE	0/100	0/100	0/100

Table 1: Comparison between evolution models on Exp. 1. Each entry contains the number of instances (out of 100) where the formulations based on the Persistent Phylogeny, Dollo(2), Camin-Sokal(2) models obtain a total error that is smaller than the one obtained with the Perfect Phylogeny model.

Camin-Sokal(4), which does not allow to find a proper solution within the allotted timeout (Figure 3). Nevertheless, we note that Persistent model obtains better results than the Perfect Phylogeny, especially when the allowed number of clones is small. The experimentations and considerations exposed here required a total of 105 CPU hour for Exp. 1 and 120 CPU hour for Exp. 2.

5.2 Real Data

We also tested our models on real cancer data, in particular Leukemia data from [24] because liquid tumors seem to have the fewest somatic mutations, therefore we are able to calculate an optimal solution in a reasonable amount of time. We have been able to compute a clonal evolution for the CLL patient 077[7, 18] under the Persistent Phylogeny model in approximately 3 days of computation.

We cannot directly compare the persistent tree we inferred (Figure 6.A) with AncesTree [7] (Figure 6.C), because the latter infers only seven of the 16 mutations present in the sample. In order to perform such comparison we had to restrict the instance to contain

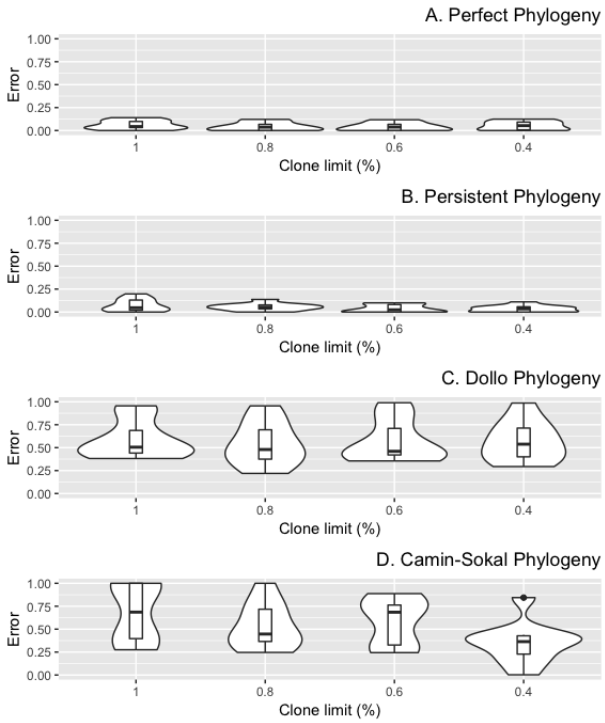


Figure 5: Errors for the solution founded for instances in Exp. 2 for different phylogenetic models.

Total error	Persistent	Dollo(4)	Camin-Sokal(4)
Clone limit 100%			
≤ PPE	3/10	0/10	0/10
≤ 90% PPE	3/10	0/10	0/10
≤ 80% PPE	3/10	0/10	0/10
≤ 50% PPE	2/10	0/10	0/10
Clone limit 80%			
≤ PPE	5/10	0/10	0/10
≤ 90% PPE	3/10	0/10	0/10
≤ 80% PPE	3/10	0/10	0/10
≤ 50% PPE	0/10	0/10	0/10
Clone limit 60%			
≤ PPE	6/10	0/10	0/10
≤ 90% PPE	5/10	0/10	0/10
≤ 80% PPE	4/10	0/10	0/10
≤ 50% PPE	4/10	0/10	0/10
Clone limit 40%			
≤ PPE	7/10	0/10	1/10
≤ 90% PPE	5/10	0/10	1/10
≤ 80% PPE	5/10	0/10	1/10
≤ 50% PPE	4/10	0/10	1/10

Table 2: Comparison between evolution models on Exp. 2. Each entry contains the number of instances (out of 10) where the formulations based on the Persistent Phylogeny, Dollo(4), Camin-Sokal(4) models obtain a total error that is smaller than the one obtained with the Perfect Phylogeny model.

only the mutations that are also in the solution computed by AncesTree. The output is presented in Figure 6.D and shares several structural similarities with the AncesTree solution. Moreover, we would also like to point out that our solution for the restricted instance (Figure 6.D) has zero errors (and is therefore optimal).

We have compared our predictions with those of PhyloSub [18] (Figure 6.B). PhyloSub compacts different mutations in the same clone while we infer a tree in which each mutation correspond to a vertex. The differences between the two are that the cluster of mutations containing NAMPTL, PLAG2616, SLC12A1, BCL2L13 and GPR158 occurs before mutation EXOC6B while in our model the latter occurs first. The main differences are that the clone containing mutation COL24A1 in our model contains also mutation EXOC6B while it doesn't in PhyloSub, on contrary in our model the clone containing LRRPC16A does not contain NAMPTL and in PhyloSub does. The most interesting fact regards the clone that contains NOD1: in PhyloSub it lies on a branch that does not contain EXOC6B, while we have predicted a phylogeny where such clone occurs after the acquisition and the loss of EXOC6B.

From the experiment on CLL077, we can see that our prediction is similar to that obtained with the other available tools, but we incorporated in the evolutionary history two mutations losses, that is our evolutionary history is not consistent with the limitations of the Perfect Phylogeny model.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a ILP formulation of the problem of reconstructing the evolutionary history of tumors, where the evolutionary tree is character-based and can violate the infinite site assumption of the Perfect Phylogeny model. First, we have proposed an ILP framework for the Dollo(k) and Camin-Sokal(k) models — k is a bound on the number of losses and gains of each mutation. Then we have shown how to extend it for solving the Variant Allele Frequency Factorization Problem under those evolution models. We have performed an experimental analysis on simulated and real data which shows that the Persistent and Dollo(k) model for $k > 1$, allow to obtain phylogenies whose predicted frequencies are closer to the true frequencies than those obtained via the Perfect Phylogeny model. Mainly, our approach achieves good performances when the input data do not fit the Perfect Phylogeny model, but the number of recurrent or back mutations is relatively small. Moreover, our approach allows to relax the constraint that then number of clones is equal to the number of distinct mutations. These promising results have been obtained with ILP formulations that have not been optimized for efficiency.

Future research will be devoted to further investigate our approach on larger instances (more samples and mutations): this will require to improve the computational efficiency of the ILP formulation or adopting some combinatorial strategies to govern the introduction of a small number of mutation losses and gains in the solution, as our initial analysis has not investigated the scalability

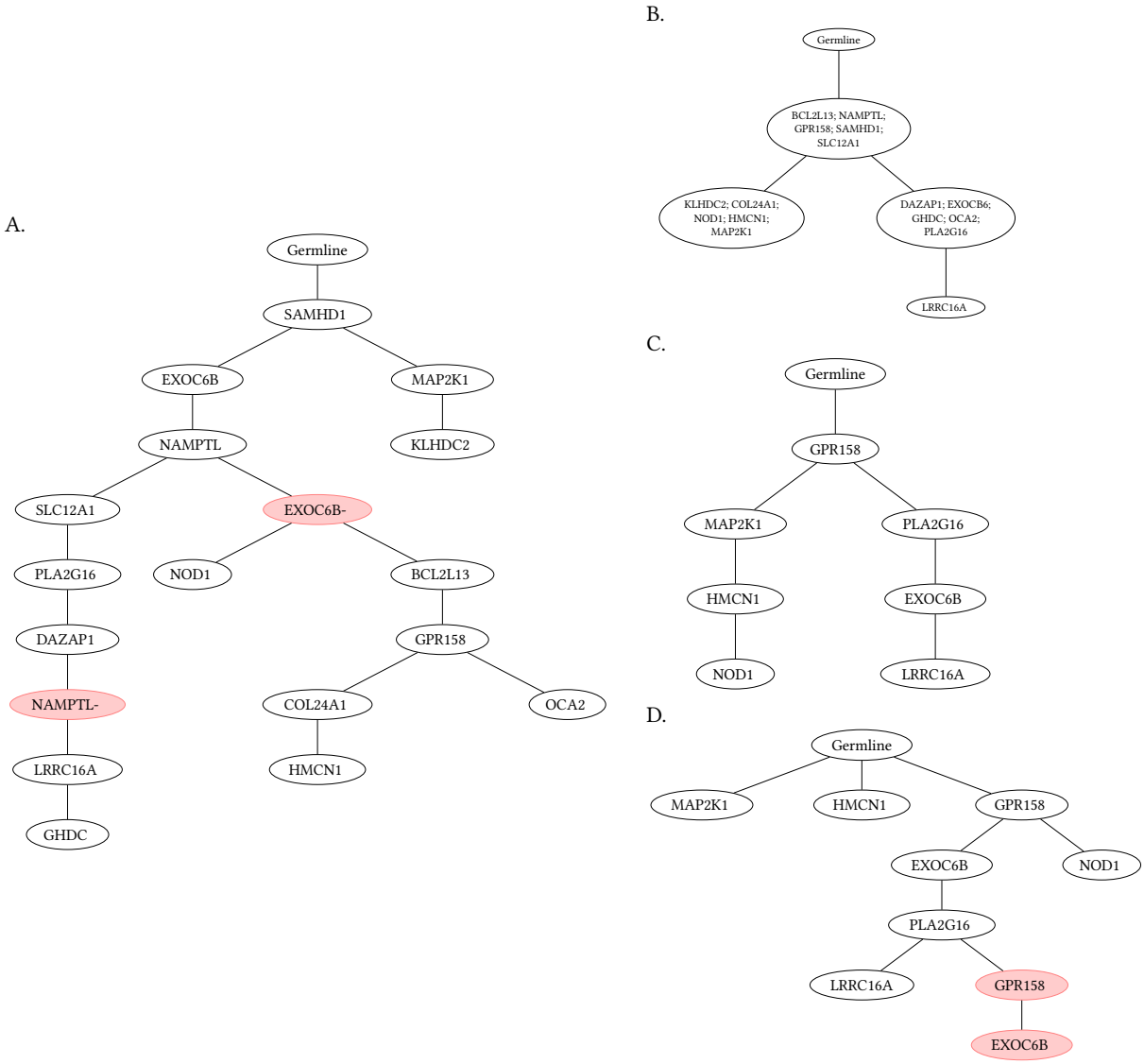


Figure 6: Persistent inferred tree for tumor CLL077 (A) Tree inferred by `gppf`, the red circular vertices are the losses of a mutation, therefore we have two clonal expansion where mutations *EXOC6B* and *NAMPTL* are lost during the clonal evolutionary history. B. shows the tree inferred by PhyloSub, while C. shows the result of AncesTree. D. shows the solution under the Persistent model for the restricted instance presented in [7] for the AncesTree algorithm.

of the approach. Moreover, we would like to assess the biological soundness of the solutions provided by our approach.

ACKNOWLEDGMENTS

The authors would like to thank Simone Zaccaria for the useful discussion on the VAFFP and ILP formulation. We also wish to thank the reviewers for their valuable comments allowing us to improve the clarity and content of the paper.

We acknowledge the support of the MIUR PRIN 2010–2011 grant "Automi e Linguaggi Formali: Aspetti Matematici e Applicativi" code 2010LYA9RH, of the Cariplo Foundation grant 2013–0955 (Modulation of anti cancer immune response by regulatory non-coding RNAs), of the FA grants 2013-ATE-0281, 2014-ATE-0382, and 2015-ATE-0113.

REFERENCES

[1] Paola Bonizzoni, Chiara Braghin, Riccardo Dondi, and Gabriella Trucco. 2012. The binary perfect phylogeny with persistent characters. *Theor. Comput. Sci.* 454 (2012), 51–63. https://doi.org/10.1016/j.tcs.2012.05.035

[2] Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, Riccardo Dondi, and Teresa M. Przytycka. 2014. When and How the Perfect Phylogeny Model Explains Evolution. In *Discrete and Topological Models in Molecular Biology*, Nataša Jonoska and Masahico Saito (Eds.). Springer Berlin Heidelberg, Berlin, Germany, 67–83.

[3] Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, Raffaella Rizzi, and Gabriella Trucco. 2017. A colored graph approach to perfect phylogeny with persistent characters. *Theoretical Computer Science* 658 (2017), 60–73. https://doi.org/10.1016/j.tcs.2016.08.015

[4] Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, and Gabriella Trucco. 2014. Explaining evolution via constrained persistent perfect phylogeny. *BMC Genomics* 15, Suppl 6 (Oct. 2014), S10. https://doi.org/10.1186/1471-2164-15-S6-S10

[5] Joseph H. Camin and Robert R. Sokal. 1965. A Method for Deducing Branching Sequences in Phylogeny. *Evolution* 19, 3 (Sep 1965), 311–326. https://doi.org/10.2307/2406441

[6] Li Ding, Benjamin J. Raphael, Feng Chen, and Michael C. Wendl. 2013. Advances for Studying Clonal Evolution in Cancer. *Cancer Letters* 340, 2 (2013), 212–219. https://doi.org/10.1016/j.canlet.2012.12.028

[7] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31, 12 (2015), 62–70. https://doi.org/10.1093/bioinformatics/btv261

[8] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael. 2016. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems* 3, 1 (2016), 43–53. https://doi.org/10.1016/j.cels.2016.07.004

[9] J. S. Farris. 1977. Phylogenetic Analysis Under Dollo's Law. *Systematic Biology* 26, 1 (Mar 1977), 77–88. https://doi.org/10.1093/sysbio/26.1.77

[10] Joseph Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA (USA).

[11] Mel Greaves and Carlo C. Maley. 2012. Clonal Evolution in Cancer. *Nature* 481, 7381 (2012), 306–313. https://doi.org/10.1038/nature10762

[12] Dan Gusfield. 1991. Efficient algorithms for inferring evolutionary trees. *Networks* 21, 1 (1991), 19–28. https://doi.org/10.1002/net.3230210104

[13] Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

[14] Dan Gusfield. 2015. Persistent phylogeny: a galled-tree and integer linear programming approach. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, 443–451.

[15] Dan Gusfield, Yelena Frid, and Dan Brown. 2007. Integer Programming Formulations and Computations Solving Phylogenetic and Population Genetic Problems with Missing or Genotypic Data. In *Computing and Combinatorics: 13th Annual International Conference, COCOON 2007, Banff, Canada, July 16-19, 2007. Proceedings*, Guohui Lin (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 51–64. https://doi.org/10.1007/978-3-540-73545-8_8

[16] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J. Raphael. 2014. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* 30, 12 (June 2014), i78–i86. https://doi.org/10.1093/bioinformatics/btu284

[17] Richard R. Hudson. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 2 (2002), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

[18] Wei Jiao, Shankar Vembu, Amit G. Deshwar, Lincoln Stein, and Quaid Morris. 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* 15, 1 (2014), 35. https://doi.org/10.1186/1471-2105-15-35

[19] Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. 2016. A statistical test on single-cell data reveals widespread recurrent mutations in tumor evolution. *bioRxiv* (Dec. 2016), 094722. https://doi.org/10.1101/094722

[20] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. 2013. Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes. *Nature* 499, 7457 (2013), 214–218. https://doi.org/10.1038/nature12213

[21] Nicholas E Navin. 2014. Cancer Genomics: One Cell At a Time. *Genome Biology* 15, 8 (2014), 452. https://doi.org/10.1186/s13059-014-0452-9

[22] Itsik Pe'er, Tal Pupko, Ron Shamir, and Roded Sharan. 2004. Incomplete Directed Perfect Phylogeny. *SIAM J. Comput.* 33, 3 (Jan 2004), 590–607. https://doi.org/10.1137/s0097539702406510

[23] Teresa Przytycka, George Davis, Nan Song, and Dannie Durand. 2006. Graph Theoretical Insights into Dollo Parsimony and Evolution of Multidomain Proteins. *Journal of Computational Biology* 13(2) (2006), 351–363.

[24] Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M. Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, Zoya Kingsbury, Shujun Luo, David McBride, Lisa Murray, Toshi Menju, Adele Timbs, Mark Ross, Jenny Taylor, and David Bentley. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 120, 20 (2012), 4191–4196. https://doi.org/10.1182/blood-2012-05-433540

[25] C. Semple and M. Steel. 2003. *Phylogenetics*. Oxford University Press, USA.

[26] M. A. Steel. 2016. *Phylogeny: discrete and random processes in evolution*. Number 89 in CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia.

[27] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. 2013. Cancer Genome Landscapes. *Science* 339, 6127 (2013), 1546–1558. https://doi.org/10.1126/science.1235122