Dolphyin: A Combinatorial Algorithm for Identifying 1-Dollo Phylogenies in Cancer

Daniel W. Feng
Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

Mohammed El-Kebir[1]
Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, USA
Cancer Center at Illinois, University of Illinois Urbana-Champaign, Urbana, IL, USA

Abstract

Several recent cancer phylogeny inference methods have used the k-Dollo evolutionary model for single-nucleotide variants. Specifically, in this problem one is given an $m \times n$ binary matrix B and seeks a rooted tree T with m leaves that correspond to the m rows of B, and each node of T is labeled by a binary state for each of the n characters subject to the restriction that each character is gained at most once (0-to-1 transition) and subsequently lost at most k times (1-to-0 transitions). The 1-Dollo variant, also known as the persistent perfect phylogeny where one is restricted to at most $k = 1$ losses per character, has been studied extensively, but its hardness remains an open question. Here, we prove that the 1-Dollo Linear Phylogeny (1DLP) problem, where we additionally require the resulting 1-Dollo phylogeny T to be linear, is equivalent to verifying whether the input matrix B adheres to the Consecutive Ones Property (C1P), which can be solved in polynomial time. Due to the equivalence, several known NP-hardness results for relevant variants of C1P carry over to 1DLP, including the minimization of false negatives (0-to-1 modifications to the input matrix B) or the allowance of 2 gains and 2 losses. We furthermore show how we can recursively decompose any, not necessarily linear, 1-Dollo phylogeny T into several 1-Dollo linear phylogenies, connected by matching branching points. We extend this characterization to matrices B that admit 1-Dollo phylogenies, giving necessary and sufficient conditions for the existence of a novel decomposition of B into several submatrices and corresponding branching points. This decomposition forms the basis of Dolphyin, a new exponential-time algorithm for inferring 1-Dollo phylogenies that efficiently leverages the determination of linear 1-Dollo phylogenies as a subroutine. Dolphyin can also be applied to input matrices B with false negatives. We demonstrate that Dolphyin is runtime-competitive with a previous integer linear programming based algorithm SPhyR on simulated datasets. We additionally analyze simulated datasets with false negative errors and find that in the median case, Dolphyin infers 1-Dollo phylogenies with inferred error rates at or below the ground truth rate. Finally, we apply Dolphyin to 99 acute myeloid leukemia single-cell sequencing datasets, finding that the majority of the cancers can be explained by 1-Dollo phylogenies with false negative error rates in line with the used sequencing technology.

Availability. Dolphyin is available at: `https://github.com/elkebir-group/Dolphyin`.

2012 ACM Subject Classification Applied computing → Computational genomics

Keywords and phrases Intra-tumor heterogeneity, persistent perfect phylogeny, consecutive ones property, combinatorics

Digital Object Identifier 10.4230/LIPIcs.WABI.2025.9

Supplementary Material
Software (Source Code): `https://github.com/elkebir-group/Dolphyin` [14]
archived at `swh:1:dir:61c0cd5f22da3a8e10cc6fdac70dd7d93ecb6be5`

Funding *Mohammed El-Kebir*: National Science Foundation grant CCF-2046488 (M.E-K.) – This project was conducted as part of Spring 2024, CS 598 MEB: Computational Cancer Genomics.

[1] Corresponding author

25th International Conference on Algorithms for Bioinformatics (WABI 2025).
Editors: Broňa Brejová and Rob Patro; Article No. 9; pp. 9:1–9:23
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

The clonal theory of cancer states that tumors are composed of heterogeneous clones, which are groups of cells with similar genotypes. Clones arise from an evolutionary process during which somatic mutations accumulate in cell populations [25]. By performing bulk or single-cell sequencing on these clones' DNA or RNA, scientists can identify common somatic mutations such as single-nucleotide variants (SNVs), copy number aberrations (CNAs), or structural variants (SVs). Then, algorithms attempt to infer phylogenies, which represent the evolution of the tumor, from sequencing data for important downstream analysis and clinical decision-making [30]. These phylogenetic inference algorithms utilize sequencing-technology specific error characteristics and constraints on how these somatic mutations accumulate, which constitute an evolutionary model specific to the somatic mutations of interest [29].

In this work, we focus on the presence or absence of SNVs in single-cell DNA sequencing data. More precisely, we are given single-cell data in the form of a matrix B where each row in the data is a taxon, representing a tumor cell, and each column is a single-nucleotide variant (SNV), hereafter referred to as a character. The entries in the data matrix would then be either 0 or 1, indicating the absence or presence of a mutation in a particular cell. We wish to find a phylogeny, i.e. a rooted, node-labeled tree T whose leaves represent the extant cells of the tumor, internal nodes represent ancestral tumor cells, and the root represents a normal cell [25], that explains this data. We would then need to assume an evolutionary model that constrains the phylogenies allowed on this data. Under the well-studied two-state perfect phylogeny model [1, 15, 20], for example, no character can be lost once gained in a path starting from the root of tree T. Detecting whether an assumed error-free binary data matrix allows a phylogeny under the perfect phylogeny model is solvable in polynomial time [1, 18].

However, the restrictiveness of the perfect phylogeny model of evolution has inspired investigation into a wide range of more generalized and biologically-plausible models [4, 13]. Many analyses have operated under the flexible k-Dollo model of evolution, under which any character may be gained exactly once but lost in the tumor's evolution at most k times [7, 9, 12, 28]. This flexibility affords the incorporation of common biological events, such as CNAs that may delete previously gained SNVs, and can thus be much more realistically versatile in biological analysis. The ∞-Dollo phylogeny inference [11, 26] and tree size-constrained versions of Dollo phylogeny inference [11] are known to be NP-hard. Additionally, the problem variant for Dollo inference where the total number of losses summed over all characters is minimized, rather than outright bounded per character, can be solved in polynomial time when the resulting phylogeny is clade-constrained [9].

An important subcase of the k-Dollo problem is the $k = 1$ subcase or 1-Dollo problem, also known as the persistent phylogeny problem (Figure 1). The 1-Dollo problem has been extensively studied, using various problem statements, for over 20 years [17]. For example, characterizations of the 1-Dollo problem have yielded an exact algorithm that solves the 1-Dollo problem in time polynomial to the number of taxa and exponential to the number of characters [2]. Other work has also developed Integer Linear Programming (ILP) solutions to the 1-Dollo problem and shown a connection between galled trees and 1-Dollo phylogenies [19]. Graph-based approaches, specifically the ability to manipulate colored graphs representative of data matrices using sequenced and specific graph operations, have additionally yielded polynomial-time algorithms for a restricted version of the 1-Dollo problem [3]. However, the complexity of the general 1-Dollo problem remains an open question [3].

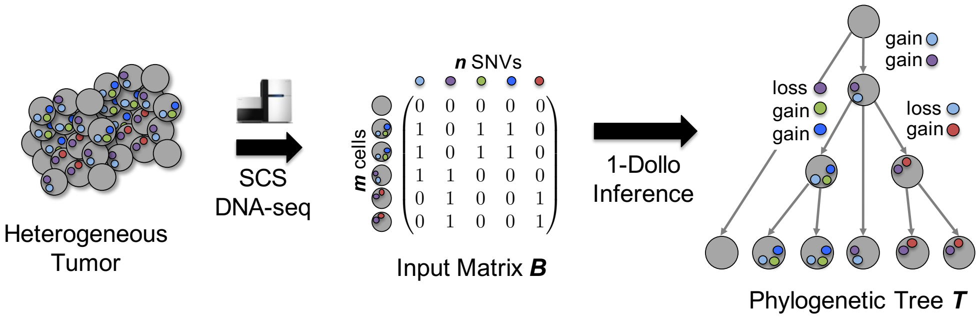

Figure 1 Tumor phylogeny estimation from single-cell sequencing (SCS) data under the 1-Dollo, or persistent phylogeny, model of evolution. Heterogeneous tumors are composed of distinct cellular populations with distinct complements of somatic mutations, including single-nucleotide variants (SNVs). During cancer progression, SNVs are frequently lost due to copy-number aberrations, but rarely introduced more than once. Here, single-cell sequencing of a tumor yields an input matrix B, whose m rows are taxa and n columns are SNVs. Under the 1-Dollo evolutionary model, each SNV can only be gained once and lost once. Our goal is to infer a satisfying phylogeny T under this model or demonstrate that one does not exist.

Separately, tumor phylogenies can either be linear – dash that is, phylogeny T has no nontrivial branching points in which cell evolution diverges – dash or branching otherwise, based on the tumor's selective pressures. In this work, we define "branching points" to entail internal nodes in T with more than one child that are also internal nodes. Many phylogenies on real data, while branching, have exhibited a disproportionately small number of branching points relative to the number of taxa [23, 28]. Such phylogenies plausibly arise when the tumor microenvironment exerts severe enough selective pressure to limit branching to a few, highly viable offshoot clones [10]. This observation motivates phylogenetic inference that is specialized for finding linear or near-linear phylogenies. For example, machine learning techniques have been used to determine if a tumor phylogeny is likely linear [27] and, in previous work, we showed that determining the minimum number of changes from 0 to 1 in a data matrix such that the altered matrix is then representative of even a linear perfect phylogeny is NP-hard [31]. However, as far we are aware, determining 1-Dollo phylogenies on data that are strictly linear has not yet been explicitly examined.

In this paper, our aims are first theoretical and second experimental. First, we draw an equivalence between determining if a data matrix B admits a 1-Dollo linear phylogeny and determining if B has the consecutive ones property, which is a known property verifiable in polynomial time [16]. We use this theoretical characterization of matrices admitting 1-Dollo linear phylogenies to discuss natural problem variants, such as determining 1-Dollo linear phylogenies with fixed character-state vectors for the root or terminating leaf, and show that determining the minimal number of false negative entries in a sequencing data input to allow such a phylogeny, contrastingly, is NP-hard. As a tree can be recursively decomposed around its branching points, we then use this linear subcase of the 1-Dollo problem to recursively characterize all matrices admitting any 1-Dollo phylogeny, regardless of branching, with a series of necessary and sufficient conditions. Second, we develop a combinatorial algorithm, Dolphyin (DOllo Linear PHYlogeny INference Method), that uses this theoretical characterization to practically determine 1-Dollo phylogenies on sequencing data. We show that Dolphyin, which relies on determining linear chains of taxa satisfying the 1-Dollo model of evolution and then recursing on remaining taxa, is runtime-competitive with SPhyR – dash an ILP-based method of inference for the k-Dollo problem. Additionally, we adapt Dolphyin to probabilistically correct for false-negative errors in sequencing. We apply Dolphyin to simulated datasets with false-negative errors and show that Dolphyin yields 1-Dollo phylogenies with inferred rates of error at or below the true rate of error. Finally, we apply Dolphyin to 99 real datasets of acute myeloid leukemia (AML) single cell sequencing data [23] and find that Dolphyin infers 1-Dollo phylogenies for 55 datasets with false negative error rates consistent with the used sequencing technology.

2 Problem Statement

Suppose we have sequenced m cells of a tumor and identified n single-nucleotide variants (SNVs). We are given a binary matrix $B \in \{0,1\}^{m \times n}$, whose m rows or *taxa* correspond to the sequenced cells and whose n columns or *characters* correspond to the SNVs. The tumor cells share a common evolutionary history, represented by a rooted and node-labeled tree T. Here, we require T to adhere to the k-Dollo evolutionary model [13], defined as follows.

▶ **Definition 1.** *A rooted, node-labeled tree T is a k-*Dollo phylogeny *for an $m \times n$* binary *matrix $B = [\mathbf{b}_1, \ldots, \mathbf{b}_m]^\top$* rooted at *$\mathbf{b}_0 = [b_{0,1}, \ldots, b_{0,n}]^\top$ provided*

(i) *each node v in T is labeled by a binary vector $\mathbf{b}_T(v) = [b(v,1), \ldots, b(v,n)]^\top$;*

(ii) *the root r of T is labeled by $\mathbf{b}_T(r) = \mathbf{b}_0$;*

(iii) *T has m leaves such that each taxon $t \in [m]$ corresponds to exactly one leaf $\sigma_T(t) = v$ in T with parent u such that $\mathbf{b}_T(v) = \mathbf{b}_T(u) = \mathbf{b}_t$;*

(iv) *for each character $c \in [n]$ where $b_{0,c} = 0$, there is at most one* gain edge *(u, v) such that $b_T(u,c) = 0$ and $b_T(v,c) = 1$, and at most k* loss edges *(u', v') such that $b_T(u', c) = 1$ and $b_T(v', c) = 0$;*

(v) *for each character $c \in [n]$ where $b_{0,c} = 1$, there is no gain edge and at most k loss edges.*

We omit the subscript T from node labeling $\mathbf{b}_T(v)$ and taxa mapping $\sigma_T(t)$ if it is clear from context that they apply to a particular tree T. In this paper, we restrict our attention to the common case where at most $k = 1$ loss per character is allowed and, unless otherwise stated, assume that the root must be labeled by all 0s, i.e. $\mathbf{b}(r) = \mathbf{b}_0 = \mathbf{0}$. We call such trees simply 1-Dollo phylogenies for B. Thus, we seek to solve the following problem.

▶ **Problem 1** (1-Dollo Phylogeny (1DP))**.** *Given binary matrix $B \in \{0,1\}^{m \times n}$, build a 1-Dollo phylogeny T for B or show that one does not exist.*

We note that the above problem is also known as the persistent phylogeny problem [2]. In addition to the above problem, we are also interested in the problem where T is required to be a 1-*Dollo linear phylogeny for B*.

▶ **Definition 2.** *A 1-Dollo phylogeny for a binary matrix $B \in \{0,1\}$ is* linear *if the removal of the m leaves of T corresponding to the m taxa yields a chain graph.*

▶ **Problem 2** (1-Dollo Linear Phylogeny (1DLP))**.** *Given binary matrix $B \in \{0,1\}^{m \times n}$, is there a 1-Dollo linear phylogeny T for B, and if so, build one.*

3 Combinatorial Characterization

We characterize the solution spaces of both 1DP and 1DLP, starting with the more restrictive problem, 1DLP, in Section 3.1. We then build on this result by discussing the complexity of common problem variants to 1DLP. Finally, we demonstrate in Section 3.2 that solutions to the 1DP problem can be recursively characterized in terms of itself and 1-Dollo linear phylogenies. Due to space constraints, we delegate the proofs of all lemmas and theorems to the supplement.

To simplify the exposition, we introduce the following definition of a *compact* phylogeny and lemma, stating that one can assume without loss of generality that all internal, non-root nodes of solutions T must either correspond to observed taxa or are branching points, and all internal edges must be either a gain or a loss edge for some character. Intuitively, any internal, non-root nodes of a given 1-Dollo phylogeny T' that do not correspond to one of these two cases can always be contracted to create a compact 1-Dollo phylogeny T.

▶ **Definition 3.** *A* 1*-Dollo phylogeny T with root r for a matrix $B \in \{0,1\}^{m\times n}$ is* compact *if*
(i) *each internal node $u \neq r$ of T either corresponds to an* observed taxon*, i.e. u has a leaf child v such that $\mathbf{b}_T(u) = \mathbf{b}_T(v)$; or is a* branching point*, i.e. u has two distinct outgoing edges (u, v_1), (u, v_2) such that $\mathbf{b}_T(u) \neq \mathbf{b}_T(v_1)$ and $\mathbf{b}_T(u) \neq \mathbf{b}_T(v_2)$; and*
(ii) *and every internal edge (u, v) is either a gain or loss edge for some character so that $\mathbf{b}_T(u) \neq \mathbf{b}_T(v)$.*

▶ **Lemma 4.** *For each* 1*-Dollo phylogeny T for a matrix B there exists a unique compact* 1*-Dollo phylogeny T' for matrix B obtained from T.*

3.1 1DLP and the Consecutive Ones Property

We show that the 1DLP problem is equivalent to determining whether the input binary matrix B satisfies the consecutive ones property, which is defined as follows.

▶ **Definition 5** (Ref. [16])**.** *An $m \times n$ binary matrix B has the* consecutive ones property *(C1P) if there exists a permutation $\pi : [m] \to [m]$ such that for each column c the* 1*s appear consecutively when permuting the rows of B according to π.*

To demonstrate this equivalence, we first propose the following construction of obtaining a tree T from a matrix B that is C1P with permutation π, illustrated in Figure 2.

▶ **Definition 6.** *The rooted, node-labeled tree $T(B, \pi)$ resulting from a binary matrix $B = [\mathbf{b}_1, \ldots, \mathbf{b}_m]^\top$ that is C1P with permutation $\pi : [m] \to [m]$ has (i) a root node $r = u_0$ labeled by $\mathbf{b}(r) = \mathbf{0}$, (ii) internal nodes $u_1, \ldots, u_m$ and leaves $v_1, \ldots, v_m$ labeled by $\mathbf{b}(u_t) = \mathbf{b}(v_t) = \mathbf{b}_{\pi(t)}$ for each taxon $t \in [m]$, (iii) edges (u_{t-1}, u_t) and (u_t, v_t) for each taxon $t \in [m]$ and taxon leaf labeling $\sigma(v_t) = t$ for each taxon $t \in [m]$.*

This construction leads us to the main theorem of this section.

▶ **Theorem 7.** *There exists a* 1*-Dollo linear phylogeny T for B if and only if B is C1P.*

As determining whether any binary matrix $B \in \{0,1\}^{m\times n}$ is C1P including determining the corresponding permutation $\pi : [m] \to [m]$ of rows is solvable in $O(mn)$ time using PQ trees [5], 1DLP is similarly solvable in $O(mn)$ time.

▶ **Corollary 8.** *1DLP is solvable in $O(mn)$ time.*

3.1.1 Rooted and Terminating Variants of 1DLP

A natural generalization of 1DLP is the Rooted 1-Dollo Linear Phylogeny (R1DLP) problem, where the root node r must be labeled by a given vector $\mathbf{b}_0 \in \{0,1\}^n$ not necessarily equal to $\mathbf{0}$.

▶ **Problem 3** (Rooted 1-Dollo Linear Phylogeny (R1DLP))**.** *Given binary matrix $B \in \{0,1\}^{m\times n}$ and binary vector $\mathbf{b}_0 \in \{0,1\}^n$, is there a* 1*-Dollo linear phylogeny T for B rooted at $\mathbf{b}_0$, and if so, build one.*

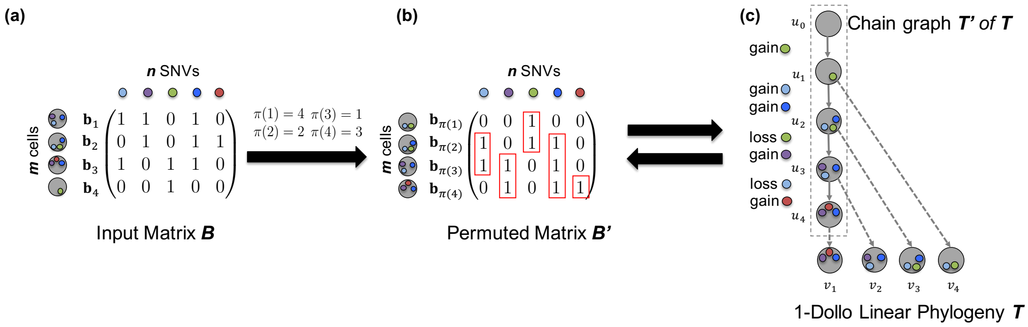

Figure 2 (a) An example matrix B demonstrating that the 1DLP problem is equivalent to the C1P problem. (b) Determining that B is C1P with permutation $\pi : [m] \to [m]$ yields permuted matrix B' such that the 1s are consecutive in each column. (c) This allows the construction of a 1-Dollo linear phylogeny T for B following Definition 6.

Clearly, the 1DLP problem is a special case of the 1RDLP problem where $\mathbf{b}_0 = \mathbf{0}$. In the following, we show that the 1RDLP problem can also be solved in $O(mn)$ time by extending matrix $B = [b_{t,c}]$ as follows (Figure 3).

▶ **Definition 9.** *The $(m+1) \times (n+1)$ binary matrix $B'(B, \mathbf{b}_0)$ resulting from $m \times n$ binary matrix $B = [b_{t,c}]$ and n-dimensional binary vector $\mathbf{b}_0 = [b_{0,c}]$ has entries $b'_{t,c}$ equal to*

$$b'_{t,c} = \begin{cases} b_{t,c}, & \text{if } t \in [m] \text{ and } c \in [n], \\ 1, & \text{if } t \in [m] \text{ and } c = n+1, \\ b_{0,c}, & \text{if } t = m+1 \text{ and } c \in [n], \\ 0, & \text{if } t = m+1 \text{ and } c = n+1. \end{cases} \tag{1}$$

▶ **Lemma 10.** *There exists a 1-Dollo linear phylogeny T for B rooted at $\mathbf{b}_0$ if and only if there exists a 1-Dollo linear phylogeny T' for $B'(B, \mathbf{b}_0)$.*

▶ **Corollary 11.** *R1DLP is solvable in $O(mn)$ time.*

A second generalization of 1DLP is the Rooted, Terminated 1-Dollo Linear Phylogeny (RT1DLP) problem where upon removal of the m leaves corresponding to the m taxa of B the root node r is labeled by $\mathbf{b}_0$ and the sink node s is labeled by $\mathbf{b}_*$ (Figure 3). More precisely, we have the following definition for such a constrained 1-Dollo linear phylogeny.

▶ **Definition 12.** *A 1-Dollo linear phylogeny T for $B \in \{0,1\}^{m \times n}$ rooted at $\mathbf{b}_0$* terminates *at $\mathbf{b}_*$ if removing the leaves $v_1, \ldots, v_m$ corresponding to the m taxa yields a chain graph terminating at node s such that $\mathbf{b}(s) = \mathbf{b}_*$.*

▶ **Problem 4** (Rooted, Terminated 1-Dollo Linear Phylogeny (RT1DLP))**.** *Given binary matrix $B \in \{0,1\}^{m \times n}$ and binary vectors $\mathbf{b}_0, \mathbf{b}_* \in \{0,1\}^n$, is there a 1-Dollo linear phylogeny T for B rooted at $\mathbf{b}_0$ and terminating at $\mathbf{b}_*$, and if so, build one.*

Again, R1DLP (Problem 3) is a special case of RT1DLP where $\mathbf{b}_* = \mathbf{0}$ and 1DLP (Problem 2) is a special case of RT1DLP where $\mathbf{b}_0 = \mathbf{b}_* = \mathbf{0}$. The 1RTDLP problem can be solved in $O(mn)$ time by a similar matrix extension to B as discussed regarding 1RDLP.

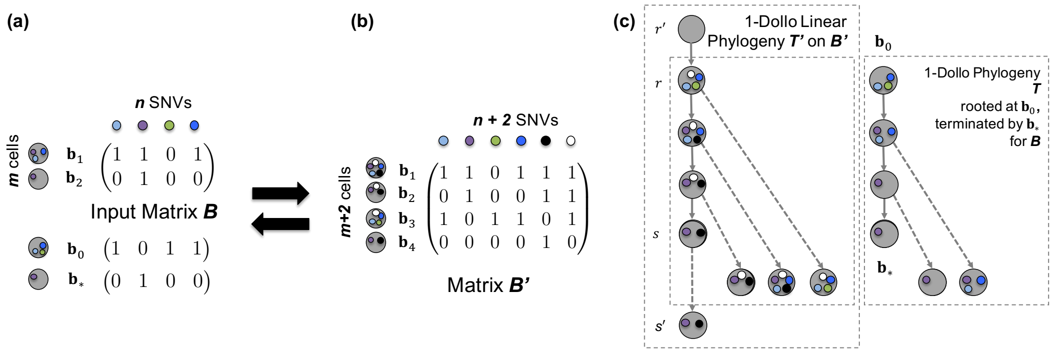
(a)
n SNVs
m cells
$\mathbf{b}_1$ (1 1 0 1)
$\mathbf{b}_2$ (0 1 0 0)
Input Matrix B
$\mathbf{b}_0$ (1 0 1 1)
$\mathbf{b}_*$ (0 1 0 0)
(b)
n + 2 SNVs
m+2 cells
$\mathbf{b}_1$ 1 1 0 1 1 1
$\mathbf{b}_2$ 0 1 0 0 1 1
$\mathbf{b}_3$ 1 0 1 1 0 1
$\mathbf{b}_4$ 0 0 0 0 1 0
Matrix B'
(c)
1-Dollo Linear Phylogeny T' on B'
r'
r
s
s'
$\mathbf{b}_0$
1-Dollo Phylogeny T rooted at $\mathbf{b}_0$, terminated by $\mathbf{b}_*$ for B
$\mathbf{b}_*$

Figure 3 RT1DLP can be solved with a transformation of 1DLP. (a) An example matrix B with root vector r and terminating state s. (b) Modified matrix B' derived from the transformation described in Definition 13 on B, r, and s. (c) 1-Dollo linear phylogeny T' for B', with corresponding 1-Dollo linear phylogeny T rooted at $\mathbf{b}_r$ and terminated at $\mathbf{b}_t$ for matrix B.

▶ **Definition 13.** *The $(m+2) \times (n+2)$ binary matrix $B'(B, \mathbf{b}_0, \mathbf{b}_*)$ resulting from $m \times n$ binary matrix $B = [b_{t,c}]$ and n-dimensional binary vectors $\mathbf{b}_0 = [b_{0,c}]$ and $\mathbf{b}_* = [b_{*,c}]$ has entries $b'_{t,c}$ equal to*

$$b'_{t,c} = \begin{cases} b_{0,c}, & \text{if } t = m+1 \text{ and } c \in [n], \\ 0, & \text{if } t = m+1 \text{ and } c = n+1, \\ 1, & \text{if } t = m+1 \text{ and } c = n+2, \\ b_{*,c}, & \text{if } t = m+2 \text{ and } c \in [n], \\ 0, & \text{if } t = m+2 \text{ and } c = n+2. \end{cases} \quad \begin{cases} 1, & \text{if } t = m+2 \text{ and } c = n+1, \\ b_{t,c}, & \text{if } t \in [m] \text{ and } c \in [n], \\ 1, & \text{if } t \in [m] \text{ and } c = n+1, \\ 1, & \text{if } t \in [m] \text{ and } c = n+2, \end{cases} \tag{2}$$

▶ **Lemma 14.** *There exists a 1-Dollo linear phylogeny T for B rooted at $\mathbf{b}_0$ and terminating at $\mathbf{b}_*$ if and only if there exists a 1-Dollo linear phylogeny T' for $B'(B, \mathbf{b}_0, \mathbf{b}_*)$.*

▶ **Corollary 15.** *RT1DLP is solvable in $O(mn)$ time.*

3.1.2 Additional Variants of 1DLP

The direct equivalence from 1DLP to the Consecutive Ones Property allows several known properties from the latter to apply to 1DLP. For example, allowing for false negatives, a typical phenomenon in single-cell DNA sequencing due to allelic dropout [24], yields the following problem.

▶ **Problem 5** (Minimum Error 1-Dollo Linear Phylogeny (ME1DLP)). *Given binary matrix $B \in \{0,1\}^{m \times n}$ for m cells and n SNVs, determine the minimum number of 0 to 1 replacements in B such that the resulting matrix B' has a 1-Dollo linear phylogeny T.*

As the equivalent problem of determining the minimum number of 0-to-1 matrix modifications of any binary matrix B to satisfy C1P is NP-hard [6], ME1DLP is also NP-hard.

▶ **Corollary 16.** *ME1DLP is NP-hard.*

In a similar vein, the 1DLP problem variant with at most two gains and at most two losses per character is equivalent to the C1P generalization $(2, \infty)$ observed in [8], which is NP-hard.

▶ **Corollary 17.** *Determining whether a binary matrix $B \in \{0,1\}^{m \times n}$ admits a 1-Dollo linear phylogeny with at most two gains and at most two losses per character is NP-hard.*

3.2 Recursive Characterization of 1DP

We begin by stating the rooted version of 1DP (Problem 1), which has no linear constraint.

▶ **Problem 6** (Rooted 1-Dollo Phylogeny (R1DP)). *Given binary matrix $B \in \{0,1\}^{m \times n}$ and binary vector $\mathbf{b}_0 \in \{0,1\}^n$, construct a 1-Dollo phylogeny T for B rooted at $\mathbf{b}_0$ or determine that one does not exist.*

In this section we will show how a 1DP problem instance $B \in \{0,1\}^{m \times n}$ can be recursively decomposed into smaller Rooted 1-Dollo Linear Phylogeny (R1DLP, Problem 3), Rooted, Terminated 1-Dollo Linear Phylogeny (RT1DLP, Problem 4) and Rooted 1-Dollo Phylogeny (R1DP, Problem 6) instances on submatrices of B.

To that end, we introduce the notation $B[X, C]$ to indicate the submatrix of B induced by rows/taxa $X \subseteq [m]$ and columns/characters $C \subseteq [n]$. Moreover, we define $\mathbf{b} \oslash C$ to be the restriction of binary vector $\mathbf{b} \in \{0,1\}^n$ to only characters $C \subseteq [n]$. For example, the restriction of $\mathbf{b} = [0,1,1,1,1]^\top$ to characters $C = \{1,3\}$ equals $\mathbf{b} \oslash C = [0,1]^\top$. We use the shorthand $T \oslash C$ to indicate that all node labels of T have been restricted to C, i.e. $\mathbf{b}(v) \oslash C$ for all nodes v of T.

3.2.1 Recursive Characterization of Rooted 1-Dollo Phylogenies

Like every tree structure, a 1-Dollo phylogeny T for a given binary matrix B can be recursively characterized. Key to the characterization are *branching points*, which are internal nodes with more than one non-leaf child. Let v_* be the first branching point encountered by a tree traversal on T starting from its root r labeled by $\mathbf{b}_0$. Let v_* be labeled by $\mathbf{b}_*$ and have $\ell > 1$ non-leaf children. If no such node exists then T is simply a 1-Dollo linear phylogeny for B, corresponding to the base case of the recurrence.

If there exists a branching point v_* then, on the tree traversal from r to v_*, we encounter a subset $X_0 \subseteq [m]$ of taxa as well as identify sets $C_0^+, C_0^- \subseteq [n]$ of characters that were gained or lost solely on this traversal, respectively. Note that a character first gained and then lost on this traversal is present in both C_0^+ and C_0^-. Also, note that C_0^- may not be a subset of C_0^+; for instance, there may be a character c that was previously gained such that $b_{0,c} = 1$ that is subsequently lost prior to the branching point v_*, leading to $c \notin C_0^+$ and $c \in C_0^-$. Let $C_0 = C_0^- \cup C_0^+$, and let v_* be labeled by binary vector $\mathbf{b}_*$. The encountered nodes on the traversal from r to v_* induce a subtree T_0 such that its restriction $T_0 \oslash C_0$ is precisely a 1-Dollo linear phylogeny for submatrix $B[X_0, C_0]$ rooted at $\mathbf{b}_0 \oslash C_0$ and terminating at $\mathbf{b}_* \oslash C_0$. To characterize the remainder of the tree, observe that performing a traversal of T starting at the i-th outgoing edge from v_* yields a tree T_i composed of taxa X_i, gained characters C_i^+ and lost characters C_i^-. Let $C_i = C_i^- \cup C_i^+$. Since decomposing a tree cannot add new unique edges or nodes to the sum of its parts, we have that $T_i \oslash C_i$ is precisely a 1-Dollo phylogeny for submatrix $B[X_i, C_i]$ rooted at $\mathbf{b}_* \oslash C_i$.

▶ **Lemma 18.** *For a given binary vector $\mathbf{b}_0 \in \{0,1\}^n$ and 1-Dollo phylogeny T for matrix $B \in \{0,1\}^{m \times n}$, let T_0 be the subtree of T obtained by traversing from the node v_0 labeled by $\mathbf{b}_0$ to a first branching point v_* with label $\mathbf{b}_*$, and let $T_1, \ldots, T_\ell$ be the subtrees of T obtained by traversing along each of the $\ell > 1$ outgoing edges from v_*. Let C_i^+, C_i^- and X_i be the gained characters, lost characters and observed taxa, respectively, in tree T_i where $i \in \{0, \ldots, \ell\}$. Let $C_i = C_i^- \cup C_i^+$ for all $i \in \{0, \ldots, \ell\}$. Then, the following conditions hold.*

(i) *Sets $C_0^+, \ldots, C_\ell^+$ are pairwise disjoint, sets $C_0^-, \ldots, C_\ell^-$ are pairwise disjoint, and sets $C_1, \ldots, C_\ell$ are pairwise disjoint.*

(ii) *Sets $X_0, \ldots, X_\ell$ are pairwise disjoint, and $X_0 \cup \ldots \cup X_\ell$ is the set of all taxa observed in the subtree of T rooted at v_0.*

(iii) *$C_i^+ \subseteq [n] \setminus C_0^-$ for all $i \in [\ell]$.*

(iv) *Tree $T_0 \oslash C_0$ is a 1-Dollo linear phylogeny for submatrix $B_0 = B[X_0, C_0]$ rooted at $\mathbf{b}_0 \oslash C_0$ terminating at $\mathbf{b}_* \oslash C_0$.*

(v) *For each $i \in [\ell]$, tree $T_i \oslash C_i$ is a 1-Dollo phylogeny for submatrix $B_i = B[X_i, C_i]$ rooted at $\mathbf{b}_* \oslash C_i$.*

Thus, at each branching point v_* of T with $\ell > 1$ non-leaf children, we obtain a single instance of Rooted, Terminated 1-Dollo Linear Phylogeny (RT1DLP, Problem 4) and ℓ instances of Rooted 1-Dollo Phylogeny (R1DP, Problem 6). Each of these ℓ instances can be further decomposed in a recursive fashion by identifying subsequent branching points.

3.2.2 Recursive decomposition of matrix B

The above recursive decomposition of a given 1-Dollo phylogeny T for matrix B rooted at some $\mathbf{b}_0$ yields trees $T_0, T_1, \ldots, T_\ell$ on submatrices $B_0, B_1, \ldots, B_\ell$, respectively. As a phylogeny T can be decomposed, therefore, matrix B can also be decomposed. However, it is far less apparent how to do so without prior knowledge of the 1-Dollo phylogeny T on B. Here, we describe how given B and $\mathbf{b}_0$, submatrices $B_0 = B[X_0, C_0], \ldots, B_\ell = B[X_\ell, C_\ell]$ can be inferred solely from the taxa set $X_0 \subseteq [m]$ and the two character sets $C_0^-, C_0^+ \subseteq [n]$. We begin by noting how $\mathbf{b}_0$ and character sets C_0^-, C_0^+ uniquely determine the label of a potential branching point, since there is a unique path from the root to the branching point containing gains C_0^+ and losses C_0^-. To that end, we define the following function.

▶ **Definition 19.** *Given binary vector $\mathbf{b}_0 = [b_{0,1}, \ldots, b_{0,n}]^\top$ and characters $C^-, C^+ \subseteq [n]$, the n-dimensional binary vector $\mathbf{b}(\mathbf{b}_0, C^+, C^-) = [b_1, \ldots, b_n]^\top$ consists of entries*

$$b_c = \begin{cases} 0, & \text{if } c \in C^-, \\ 1, & \text{if } c \in C^+ \setminus C^-, \\ b_{0,c}, & \text{otherwise.} \end{cases} \tag{3}$$

▶ **Lemma 20.** *Let T be a 1-Dollo phylogeny for matrix B. For any node v_0 labeled by $\mathbf{b}_0$ and descendant node v_* labeled by $\mathbf{b}_*$ it holds that $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C^+, C^-)$ where C^+ and C^- are the characters that are gained and lost, respectively, on the path from v_0 to v_*.*

Knowledge of sets X_0, C_0^+, and C_0^- immediately implies knowledge of $B_0 = B[X_0, C_0]$ since $C_0 = C_0^+ \cup C_0^-$. Additional knowledge of $\mathbf{b}_0$ allows us to infer the terminal label $\mathbf{b}_* \oslash X_0 = \mathbf{b}(\mathbf{b}_0, C_0^+, C_0^-) \oslash X_0$ of RT1DLP instance $(B_0, \mathbf{b}_0 \oslash X_0, \mathbf{b}_* \oslash X_0)$.

Thus, our only nontrivial goal is to infer submatrices $B_1 = B[X_1, C_1], \ldots, B_\ell = B[X_\ell, C_\ell]$ defined by taxa $X_1, \ldots, X_\ell$ and characters $C_1, \ldots, C_\ell$. We note that by the definition of a 1-Dollo phylogeny, only characters $c \in [n] \setminus C_0^-$ can be potentially gained or lost after the branching point labeled by $\mathbf{b}_*$, and each such c can only be gained and potentially lost in specifically one tree T_i. To detail whether any character c must be gained or lost in 1-Dollo phylogeny T_i on some proposed matrix B_i, we provide the following definition.

▶ **Definition 21.** *A character $c \in [n]$ is* variable *w.r.t. an $m \times n$ matrix $B = [b_{t,c}]$ and n-dimensional vector $\mathbf{b} = [b_c]$ if there exists a taxon $t \in [m]$ such that $b_{t,c} \neq b_c$.*

Our goal therefore translates into determining a partition $\{X_1, \ldots, X_\ell\}$ of $[m] \setminus X_0$ and partition $\{C_1, \ldots, C_\ell\}$ of characters $C_* = [n] \setminus C_0^-$ such that each character $c \in C_*$ is variable w.r.t. at most one submatrix B_i and $\mathbf{b}_*$ (where $i \in [\ell]$). To that end, we define the following matrix $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$.

▶ **Definition 22.** *The* $(m - |X_0|) \times |C_*|$ complement matrix $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$ *obtained from* $m \times n$ *binary matrix* $B = [b_{t,c}]$ *and* n*-dimensional vector* $\mathbf{b}_* = [b_{*,c}]$ *has entries*

$$\bar{b}_{t,c} = \begin{cases} b_{t,c}, & \text{if } b_{*,c} = 0, \\ 1 - b_{t,c}, & \text{if } b_{*,c} = 1, \end{cases} \tag{4}$$

where $t \in [m] \setminus X_0$ *and* $c \in C_*$*.*

▶ **Lemma 23.** *Let* $\bar{B}_i$ *be a submatrix of* $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$ *defined by characters* $C_i \subseteq C_*$ *and taxa* $X_i \subseteq [m] \setminus X_0$ *and let* $B_i = B[X_i, C_i]$*. Then, character* $c \in C_i$ *is variable w.r.t.* $(B_i, \mathbf{b}_*)$ *if and only if* $\bar{B}_i$ *contains a* 1 *in column* c*.*

Whether a character is variable with respect to vector $\mathbf{b}_*$ and submatrix B_i thus corresponds to whether the character column in $\bar{B}_i$ contains a 1. Since any character must be variable in at most one submatrix, our goal now finally equates to inferring an *block diagonal matrix decomposition* into block matrices $\bar{B}_1 = \bar{B}[X_1, C_1], \ldots, \bar{B}_\ell = \bar{B}[X_\ell, C_\ell]$ of $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$, i.e.

$$\bar{B} = \begin{pmatrix} \bar{B}_1 & 0 & \cdots & 0 \\ 0 & \bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{B}_\ell \end{pmatrix}. \tag{5}$$

▶ **Definition 24.** *Partition* $\{X_1, \ldots, X_\ell\}$ *of taxa* $[m] \setminus X_0$ *and partition* $\{C_1, \ldots, C_\ell\}$ *of characters* C_* *are a* block diagonal decomposition *of* $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$ *if, for every character* $c \in C_i$*, there exists no* $t \in X_j$ *for all* $j \neq i$ *such that* $\bar{b}_{t,c} = 1$*.*

This can be trivially achieved in $O(mn)$ time by the equivalent problem of, given a bipartite graph's adjacency matrix, determining its connected components. This equivalency also demonstrates that the block diagonal matrix decomposition of maximum size for any matrix is unique (excluding characters containing all values of 0 in $\bar{B}$, which can be trivially assigned to any C_i and have no restricting effect on determining a phylogeny). Thus, we assume we always infer block diagonal matrix decompositions of maximum size. We finally synthesize B_0, matrix complement $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$, and this block diagonal decomposition to formalize a 1*-Dollo matrix decomposition on* B *and* $\mathbf{b}_0$ *by* X_0*,* C_0^+*, and* C_0^-*.*

▶ **Definition 25.** *Given binary matrix* $B \in \{0,1\}^{m \times n}$ *and binary vector* $\mathbf{b}_0$*, the* 1-Dollo matrix decomposition of B *and* $\mathbf{b}_0$ *on* X_0*,* C_0^+*, and* C_0^- *is defined as the set of submatrices* $\{B_0, B_1, \ldots, B_\ell\}$ *such that* $B_0 = B[X_0, C_0^+ \cup C_0^-]$*, and* $B_1 = B[X_1, C_1], \ldots, B_\ell = B[X_\ell, C_\ell]$ *are each given by the block diagonal decomposition* $\{X_1, \ldots, X_\ell\}$ *and* $\{C_1, \ldots, C_\ell\}$ *of* $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$*, where* $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^+, C_0^-)$ *and* $C_* = [n] \setminus C_0^-$*, with maximum size* ℓ*.*

Therefore, given binary matrix $B \in \{0,1\}^{m \times n}$ and 1-Dollo phylogeny T for B rooted at $\mathbf{b}_0$, we have established a 1-Dollo matrix decomposition of B and $\mathbf{b}_0$ on known sets X_0, C_0^+, and C_0^- that yields $\{B_0, B_1, \ldots, B_\ell\}$. Of course, such a decomposition assumes that the values of X_0, C_0^+, and C_0^- are indeed correct; that is, that X_0, C_0^+, and C_0^- are constructed according to the aforementioned recursive characterization of T established in Section 3.2.1. Without knowledge of T, such values are not trivially known. We use the following definition and lemma to establish a necessary condition to X_0, C_0^+, and C_0^- being constructed according to this recursive characterization of some T, essentially dictating that every character lacking a gain or loss edge in T_0 cannot be variable across the taxa in T_0.

▶ **Definition 26.** *Sets X_0, C_0^+, C_0^- are* in agreement *with $B \in \{0,1\}^{m\times n}$ and $\mathbf{b}_0$ if all characters $c \in [n] \setminus (C_0^+ \cup C_0^-)$ are not variable w.r.t submatrix $B[X_0,[n]]$ and vector $\mathbf{b}_0$.*

We now arrive at the main theorem of this section, where we prove that this sole necessary condition, in tandem with the 1-Dollo matrix decomposition of binary matrix B and binary vector $\mathbf{b}_0$ by X_0, C_0^+, C_0^- into $\{B_0, B_1, \ldots, B_\ell\}$, is both necessary and sufficient to recursively characterize all B for which there exists a 1-Dollo phylogeny T rooted at some $\mathbf{b}_0$. Intuitively, we prove the forward direction by decomposing an existing 1-Dollo phylogeny T on B as shown in Section 3.2.1, deriving X_0, C_0^+, C_0^- directly. We then show that each subtree beneath the first-encountered branching point of T corresponds to some matrix resulting from the 1-Dollo matrix decomposition of B and $\mathbf{b}_0$ on these derived values. Conversely, we prove the reverse direction by beginning from a 1-Dollo matrix decomposition on some existing X_0, C_0^-, C_0^+ and directly constructing T from phylogenies on the decomposition's individual parts. As an aide, we show an example 1-Dollo phylogeny's recursive decomposition, paralleled by its data matrix's recursive 1-Dollo decomposition (Figure 4).

To precisely allow for the composition of phylogenies that each may be on distinct sets of characters, we introduce the notation $\mathbf{b} \oplus \mathbf{b}_0$ which, given a vector $\mathbf{b} \in \{0,1\}^{|C|}$ restricted to characters $C \subseteq [n]$ and vector $\mathbf{b}_0 \in \{0,1\}^n$ on all characters $[n]$, re-expands $\mathbf{b}$ to include all characters in $[n]$ by supplementing missing characters with values from $\mathbf{b}_0$. For example, given $\mathbf{b} = [0,1]^\top$ on characters $C = \{1,3\}$ and $\mathbf{b}_0 = [0,1,0,1,1]^\top$ on the full set $\{1,2,3,4,5\}$ of $n = 5$ characters, $\mathbf{b} \oplus \mathbf{b}_0$ yields $[0,1,1,1,1]^\top$. Then, the shorthand $T \oplus \mathbf{b}_0$ indicates the expansion of all node labels of phylogeny T, i.e. $\mathbf{b}(v) \oplus \mathbf{b}_0$ for all nodes v of T. We use this notation to state the following theorem.

▶ **Theorem 27.** *Given matrix $B \in \{0,1\}^{m\times n}$ and binary vector $\mathbf{b}_0$, there exists a 1-Dollo phylogeny T for B rooted at $\mathbf{b}_0$ if and only if there exists some set $X_0 \subseteq [m]$ of taxa and sets $C_0^-, C_0^+ \subseteq [n]$ of characters subject to the following conditions:*

1. *Sets X_0, C_0^+, C_0^- are in agreement with B and $\mathbf{b}_0$.*
2. *For the 1-Dollo matrix decomposition of B and $\mathbf{b}_0$ on X_0, C_0^+, C_0^- into submatrices $\{B_0 = B[X_0, C_0], \ldots, B_\ell = B[X_\ell, C_\ell]\}$, there exists a 1-Dollo linear phylogeny T_0 for B_0 rooted at $\mathbf{b}_0 \oslash C_0$ and terminating on $\mathbf{b}_* \oslash C_0$ and 1-Dollo phylogenies $T_1, \ldots, T_\ell$ for $B_1, \ldots, B_\ell$ rooted at $\mathbf{b}(\mathbf{b}_0, C_0^+, C_0^-) \oslash C_1, \ldots, \mathbf{b}(\mathbf{b}_0, C_0^+, C_0^-) \oslash C_\ell$, respectively.*

4 Methods

We introduce Dolphyin (DOllo Linear PHYlogeny INference), a combinatorial algorithm that uses the above, recursive, combinatorial characterization of rooted 1-Dollo phylogenies to solve R1DP altogether. Given any binary matrix B and binary vector $\mathbf{b}_0$, Dolphyin determines a 1-Dollo phylogeny on B rooted at $\mathbf{b}_0$ by exhaustively searching over all possible values of X_0, C_0^+, C_0^- and determining a set of values such that (i) X_0, C_0^+, C_0^- are in agreement with B and $\mathbf{b}_0$ and (ii) given the 1-Dollo matrix decomposition of X_0, C_0^+, C_0^- on B and $\mathbf{b}_0$ into submatrices $\{B_0, B_1, \ldots, B_\ell\}$, there exists a 1-Dollo linear phylogeny T_0 on B_0 rooted at $\mathbf{b}_0$ terminating on $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^-, C_0^+)$ and 1-Dollo phylogenies $T_1, \ldots, T_\ell$ on $B_1, \ldots, B_\ell$ each

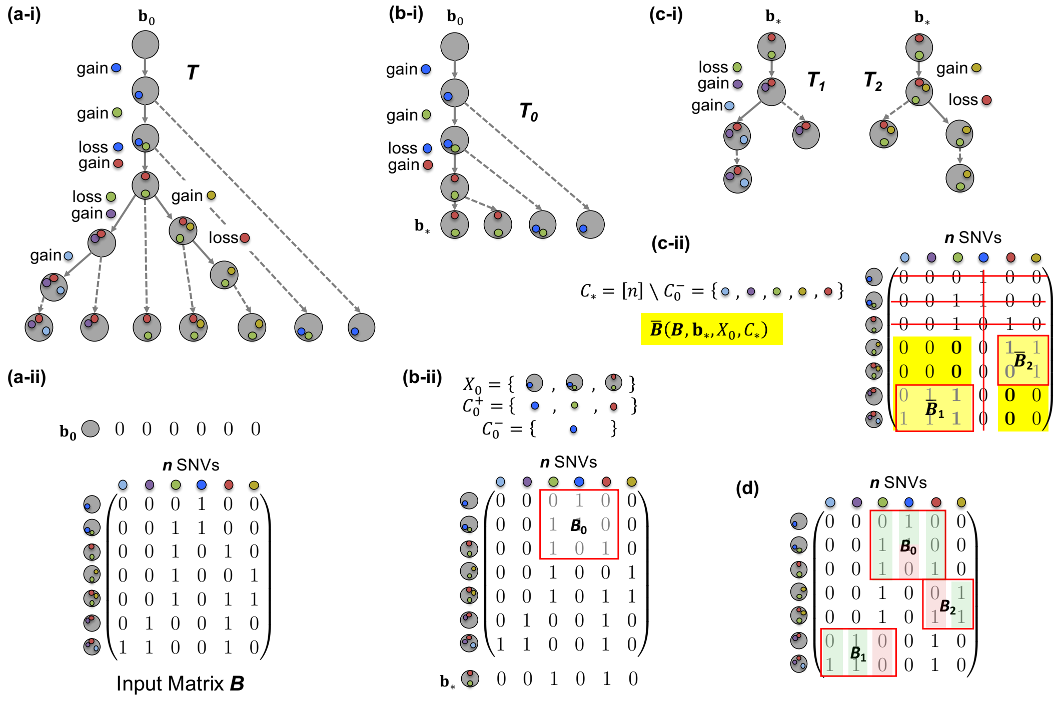

Figure 4 (a-i) An example 1-Dollo phylogeny T on (a-ii) matrix B, along with the decomposition of T into (b-i) 1-Dollo linear phylogeny T_0 and (c-i) 1-Dollo phylogenies T_1 and T_2. This recursive characterization of T is allowed by the corresponding 1-Dollo matrix decomposition of B by (b-ii) X_0, C_0^-, C_0^+ into B_0 and (c-ii) Matrix complement $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$ indicated in yellow, with bold entries representing inverted columns, and B_1 and B_2. T_0 is rooted at $\mathbf{b}_0 = \mathbf{0}$ and terminates at $\mathbf{b}_* = \mathbf{b}(C_0^-, C_0^+)$, and phylogenies T_1 and T_2 are rooted at $\mathbf{b}_*$. Critically, T_0, T_1, and T_2 split the gain and loss edges for each character $c \in [n]$ among themselves such that each character c has at most one gain edge and one loss edge among all of T_0, T_1, and T_2. The partition of gain and loss edges are visualized by (d) a color-coding of B by green and red, respectively.

rooted at $\mathbf{b}_*$. This 1-Dollo linear phylogeny T_0 can be inferred in $O(mn)$ time, and 1-Dollo phylogenies $T_1, \ldots, T_\ell$ are then inferred through recursive calls of Dolphyin on matrices $B_1, \ldots, B_\ell$. Dolphyin then returns the 1-Dollo phylogeny T rooted at $\mathbf{b}_0$ on B as the tree formed by the construction on $T_0, \ldots, T_\ell$ described in Theorem 27. This same Theorem 27 shows that this entire procedure is necessary and sufficient to solve any instance of R1DP. As an additional optimization, Dolphyin preprocesses all matrices by first removing all duplicate columns and rows, as well as trivial columns that contain 0, 1, or m 1s.

4.1 Heuristically determining candidate values of X_0, C_0^+, C_0^-

The exhaustive search over X_0, C_0^+, C_0^- as described above is necessary and sufficient to determine any existing 1-Dollo phylogeny. However, such a brute-force search over these sets can be computationally intensive, requiring roughly $O(2^m(2^n)^2) = O(2^{m+2n})$ time. Additionally, in practical R1DP instances, the vast majority of possible values of X_0, C_0^+, and C_0^- in agreement with B and $\mathbf{b}_0$ do not even yield a 1-Dollo linear phylogeny T_0 on B_0 rooted at $\mathbf{b}_0$ and terminating on $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^-, C_0^+)$.

Therefore, we enhance Dolphyin's performance with a practical heuristic that, prior to the fully exhaustive search, initially restricts the enumerated values of X_0, C_0^+, C_0^- in agreement with B and $\mathbf{b}_0$ to a subset of candidate values where such a T_0 has already been pre-determined to exist. Specifically, heuristic FindLinearChains constructs a set $\mathcal{T}_0$ of precomputed trees T_0 with corresponding values X_0, C_0^+, C_0^- such that for every element $[\mathbf{b}_0, T_0, X_0, C_0^+, C_0^-] \in \mathcal{T}_0$, T_0 is guaranteed to be a 1-Dollo phylogeny rooted at $\mathbf{b}_0$ and terminating on $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^-, C_0^+)$ on $B_0 = B[X_0, C_0^+ \cup C_0^-]$. Intuitively, FindLinearChains considers every taxon $t \in [m]$ individually and, assuming that taxon t is a branching point in T, attempt to pack as many taxa as possible into a linear phylogeny beginning with 0 and ending with t. Formally, we construct a directed acyclic graph over all taxa with source $\mathbf{b}_0$ such an edge between taxa exists if every character observed in $\mathbf{b}_t$ is not lost along the edge. Then, for every such path in this graph G_t beginning with $\mathbf{b}_0$, we consider all taxa X_0 in this path and check if such a 1-Dollo linear phylogeny indeed exists on these taxa across all characters. Critically, we record this set of taxa X_0, along with T_0, C_0^+, and C_0^-, if and only if a 1-Dollo linear phylogeny exists.

Even with the above heuristic, the worst-case running time of the initial recursive call in Dolphyin remains $O(2^{m+2n} \cdot mn)$, where the additional factor mn corresponds to checking whether the 1-Dollo decomposition of B by X^0, C_0^+, C_0^- yields a valid 1-Dollo linear phylogeny. Thus, the overall worst-case running time of Dolphyin is $\Omega(mn2^{m+2n})$.

4.2 Adapting Dolphyin to false negatives in data

When examining simulated datasets with false negative errors or real data, we modified Dolphyin to probabilistically employ error correction. Specifically, in every recursive call, Dolphyin randomly considers $p = 0.25$ pairs of taxa with replacement and, if the normalized Hamming distance over characters between both taxa is less than or equal to some value $0 \leq e \leq 1$, alter any character seen in one taxon to be present in both taxa. For each pair of taxa, one taxon was selected with uniform probability and the other was selected with probability inversely proportional to each row's prevalence in the dataset. To perform analysis on any given dataset with errors, we initially let $e = 0$, which equates to no error correction, and iteratively increased e by 0.2 until Dolphyin returned a solution within a 10 second time limit.

5 Results

Dolphyin, and its subsequent analysis and comparison to SPhyR, was implemented on an Apple 2.3 GHz 8-Core Intel Core i9 Macbook Pro in C++11. Dolphyin is available at: `https://github.com/elkebir-group/Dolphyin` with commit hash `fbf400f` used for the experiments in this paper.

5.1 Results on simulated data

We first used Dolphyin to analyze 540 simulated matrices of errorless, single-cell sequencing data with 1-Dollo ground truth phylogenies. These datasets had either $m \in \{25, 50, 100\}$ cells and $n \in \{25, 50, 100\}$ SNVs, with 90 datasets per combination of m and n, and were previously used to benchmark the ILP-based k-Dollo solver SPhyR (Figure 5a). Data was generated using `ms` [21], and full details of data generation and the data itself accessible from SPhyR's initial publication [12]. Dolphyin found and returned errorless 1-Dollo phylogenies for all 540 examined instances. We found that Dolphyin remained competitive with SPhyR in the majority of test cases across all input sizes, with identical median running times of 0.019 seconds for both methods across all instances. While Dolphyin slightly outperformed

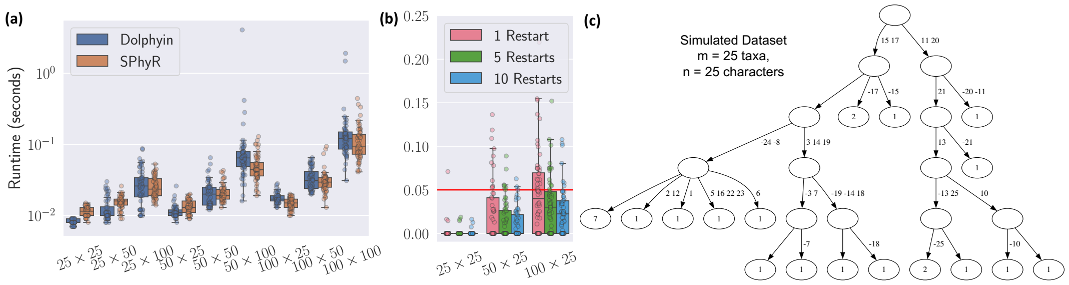

Figure 5 (a) Dolphyin's runtime on simulated, errorless datasets in comparison to that of ILP-based method SPhyR [12]. (b) The false negative error rate of data inferred by Dolphyin on simulated datasets randomly augmented with false negatives at a ground truth rate (red line) of 0.05. (c) An example phylogeny returned by Dolphyin on a dataset with false negatives and $m = n = 25$ taxa and characters. Nodes are annotated by the number of taxa present; edges are annotated by characters numbered 1 to 25. Loss edges are indicated by a minus sign.

SPhyR on the smaller input sizes (for 25 × 25 instances: 0.008 seconds for Dolphyin and 0.0115 seconds for SPhyR), SPhyR slightly outperformed Dolphyin on larger input sizes (for 100 × 100 instances: 0.122 seconds for Dolphyin and 0.0945 seconds for SPhyR).

To assay Dolphyin's ability to infer 1-Dollo phylogenies on datasets containing false negatives, we then augmented the 180 simulated datasets of sizes $m \in \{25, 50, 100\}$, $n = 25$ by flipping each 1 in these datasets to a 0 with a false-negative error rate of 0.05. We chose to augment matrices of these sizes because the number of characters $n = 25$ was most comparable to that of the real AML data we examined afterwards. Similarly, we chose an error rate of 0.05 in simulations because of its similarity to the estimated median false negative rate of 0.058 predicted under the Mission Bio Tapestri sequencing technology producing real AML data (allelic dropout rate of 5.8%) [23]. Since Dolphyin's false negative error correction is inherently probabilistic, we analyzed each dataset with 1, 5, or 10 restarts of Dolphyin. We report the lowest inferred false negative rate of all restarts (Figure 5b), which corresponds to the 1-Dollo phylogeny best fitting the data. In the median case, Dolphyin inferred phylogenies with a false negative rate at or below the ground truth rate used to generate the data ($m = 100$; 1 restart: median rate of 0.0399, 5 restarts: median rate of 0.0301, 10 restarts: median rate of 0.0225). Predictably, we found that increasing the number of restarts decreased the error rate of the best 1-Dollo phylogeny inferred. As an example, we provide a 1-Dollo phylogeny returned by Dolphyin in the analysis of a dataset with $m = 25$ taxa and $n = 25$ characters (Figure 5c). While Dolphyin is based on the characterization of 1-Dollo linear phylogenies, it clearly determines 1-Dollo phylogenies with branching.

5.2 Results on AML data

Having used Dolphyin to analyze simulated datasets both with and without false-negative sequencing errors, we then used Dolphyin to analyze 99 real sets of AML single-cell sequencing data [23] processed in a previous work [31] with 5 restarts per dataset. Prior to analysis, we removed all cells with unsequenced or unknown characters, yielding a mean of $m = 5460$ taxa, or cells, and $n = 4.42$ characters per dataset.

We show the error rates achieved on each dataset across all 5 restarts (Figure 6a), demonstrating a moderate level of consistency in inferred error rates between restarts (standard deviation between restarts, averaged over datasets of 0.0634). Taking the minimum over all 5 restarts, Dolphyin inferred phylogenies on the majority (55) of datasets with

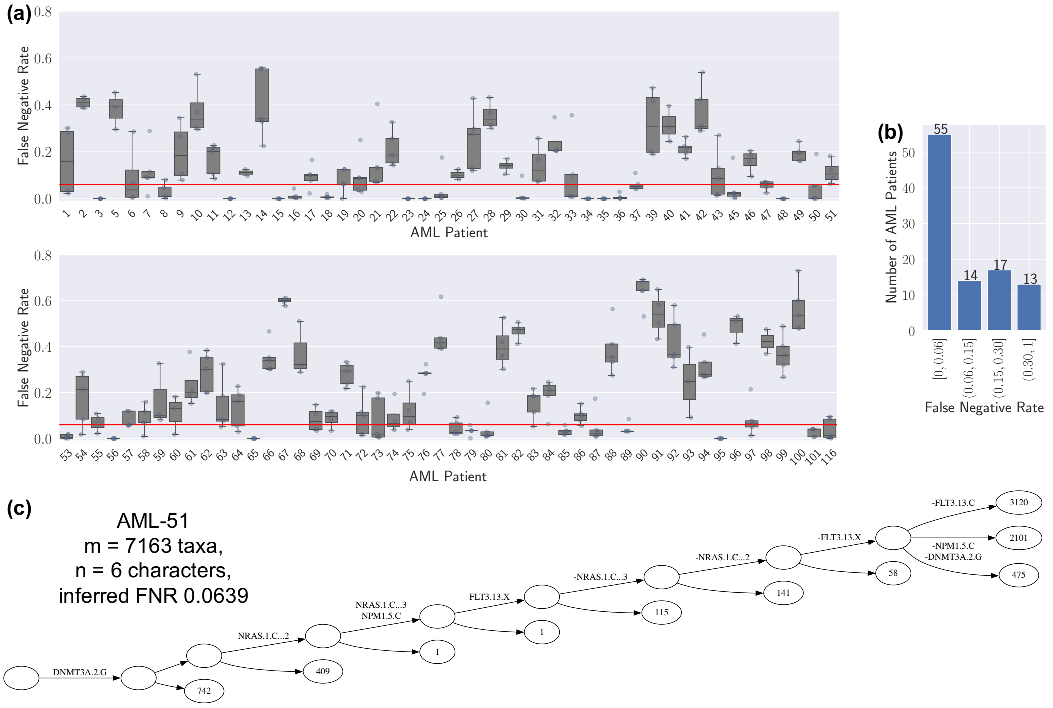

Figure 6 (a) The inferred false negative rates for all 99 examined AML datasets over 5 restarts per dataset, relative to the false negative rate expected of the sequencing technology (0.058, red line) [23]. (b) Among all restarts, slightly over half (55) of the AML datasets yielded phylogenies with a false negative rate lower than 0.06. (c) The phylogeny returned by Dolphyin on real dataset AML 51 [23] with $m = 7163$ taxa and $n = 6$ characters. Nodes are annotated by the number of taxa present; edges are annotated with characters, or SNVs. Loss edges are indicated by a minus sign.

an estimated false negative error rate at or below 0.06, approximately matching the false negative rate experimentally estimated in the datas' initial publication by Mission Bio Tapestri sequencing (median allelic dropout rate of 5.8%) [23] (Figure 6b). We speculate that for those datasets on which error rates much greater than 0.06 were inferred, there are three possibilities. Firstly, false positives in the real data, while less likely than false negatives (estimated false positive rate of 1% in initial work [23]), may be preventing Dolphyin from finding phylogenies with realistic error rates. Secondly and thirdly, more than $k = 1$ losses may be necessary to realistically explain these datasets, or simply more restarts of Dolphyin may be required to locate a realistic solution.

Finally, as a precise example of the 1-Dollo phylogenies inferred by Dolphyin, we supply the mutation-annotated phylogeny Dolphyin inferred on AML dataset 51 with size $m = 7163$, $n = 6$ with a false negative error rate of 0.0639 (Figure 6c). Dolphyin inferred that AML dataset 51 had a near-linear 1-Dollo phylogeny with a internal node of outdegree 3 furthest from the root.

6 Conclusion

This work examines the problem of inferring a 1-Dollo, or persistent phylogeny on single-cell sequencing DNA data for SNVs. We first examine the subcase in which our 1-Dollo phylogeny must be linear, and we prove an equivalence between whether a binary data matrix B admits a 1-Dollo linear phylogeny and whether B has the known consecutive ones property, which can be verified in polynomial time [16]. We also develop polynomial-time algorithms for natural extensions of the 1-Dollo linear subcase, such as the cases when we restrict our 1-Dollo phylogeny to be rooted at and/or terminate at some states. Using the linear subcase, we recursively characterize all binary matrices that admit a 1-Dollo phylogeny with a series of conditions that are provably necessary and sufficient. Unfortunately, determining whether a matrix B is 1-Dollo using this characterization takes exponential time, leaving the hardness of the 1-Dollo phylogeny problem open. In addition, we show that the problem of minimizing false negatives in data as to admit a 1-Dollo linear phylogeny is NP-hard.

We use these theoretical results to develop Dolphyin, a combinatorial algorithm that infers 1-Dollo phylogenies from sequencing data. Dolphyin directly leverages our above characterization of matrices admitting 1-Dollo phylogenies by identifying 1-Dollo linear phylogenies on subsets of taxa and recursing on remaining taxa and characters. Dolphyin also incorporates probabilistic error correction and thus can also be applied to data with false negative sequencing errors. We use Dolphyin to first analyze errorless, simulated datasets and show that Dolphyin is runtime competitive with SPhyR [12], a previous ILP-based approach for inferring k-Dollo phylogenies. We then apply Dolphyin to simulated datasets with false negative errors and demonstrate that Dolphyin, in the median case, infers 1-Dollo phylogenies with an inferred error rate at or below the ground truth rate. We finally apply Dolphyin to 99 real acute myeloid leukemia datasets [23] and find that Dolphyin infers 1-Dollo phylogenies on the majority of these datasets with an error rate at or below the previously, experimentally-estimated false negative error rate specific to the sequencing technology producing these datasets.

In future work, we may consider more advanced error correction schemes for more widely applying Dolphyin to existing datasets. We may also attempt to extend a similar, combinatorial and recursive framework to the k-Dollo model of evolution for $k > 1$, or models of evolution with more than one gain. However, we note that even determining a specifically linear and errorless phylogeny with two gains or losses, per character, is NP-hard [8]. Additionally, while we may consider problem extensions such as determining a maximal number of taxa or characters admitting 1-Dollo phylogenies in data, we note that the equivalence of 1DLP to the consecutive ones property makes several natural formulations NP-hard in even the linear and errorless cases [22]. Finally, while we argue that the first recursive call of Dolphyin is $O(2^{m+2n}mn)$ in the worst case, we would like to derive a more precise running time taking into account all recursive calls.

In summary, our work adds to the theoretical body of knowledge on the 1-Dollo, or persistent phylogeny, model of evolution and provides a practical algorithm for inferring phylogenies that leverages these theoretical results. We hope that this combinatorial approach will aid advances in determining the complexity of 1-Dollo problems and their variants.

References

1 Richa Agarwala and David Fernandez-Baca. A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM J. Comput.*, 23(6):1216–1224, December 1994. doi:10.1137/S0097539793244587.

2 Paola Bonizzoni, Chiara Braghin, Riccardo Dondi, and Gabriella Trucco. The binary perfect phylogeny with persistent characters. *Theoretical Computer Science*, 454:51–63, 2012. Formal and Natural Computing. doi:10.1016/j.tcs.2012.05.035.

3 Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, Raffaella Rizzi, and Gabriella Trucco. A colored graph approach to perfect phylogeny with persistent characters. *Theoretical Computer Science*, 658:60–73, 2017. Formal Languages and Automata: Models, Methods and Application In honour of the 70th birthday of Antonio Restivo. doi:10.1016/j.tcs.2016.08.015.

4 Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, and Gabriella Trucco. Explaining evolution via constrained persistent perfect phylogeny. *BMC Genomics*, 15, October 2014. doi:10.1186/1471-2164-15-S6-S10.

5 Kellogg S. Booth and George S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976. doi:10.1016/S0022-0000(76)80045-1.

6 Kellogg Speed Booth. *PQ-tree algorithms*. PhD thesis, University of California, Berkeley, 1975. AAI7615117.

7 Remco Bouckaert, Mareike Fischer, and Kristina Wicke. Combinatorial perspectives on dollo-k characters in phylogenetics. *Advances in Applied Mathematics*, 131:102252, 2021. doi:10.1016/j.aam.2021.102252.

8 Cedric Chauve, Jan Manuch, and Murray Patterson. Hardness results for the gapped consecutive-ones property. *Discrete Applied Mathematics*, 160, December 2009. doi:10.1016/j.dam.2012.03.019.

9 Junyan Dai, Tobias Rubel, Yunheng Han, and Erin Molloy. Dollo-cdp: a polynomial-time algorithm for the clade-constrained large dollo parsimony problem. *Algorithms for Molecular Biology*, 19, January 2024. doi:10.1186/s13015-023-00249-9.

10 Alexander Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):151–161, 2017. Evolutionary principles - heterogeneity in cancer? doi:10.1016/j.bbcan.2017.01.003.

11 William H.E. Day, David S. Johnson, and David Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81(1):33–42, 1986. doi:10.1016/0025-5564(86)90161-6.

12 Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, September 2018. doi:10.1093/bioinformatics/bty589.

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

14 Daniel W. Feng and Mohammed El-Kebir. Dolphyin. Software, swhId: swh:1:dir:61c0cd5f22da3a8e10cc6fdac70dd7d93ecb6be5 (visited on 2025-08-04). URL: https://github.com/elkebir-group/Dolphyin, doi:10.4230/artifacts.24317.

15 David Fernández-Baca. *The Perfect Phylogeny Problem*, pages 203–234. Springer US, Boston, MA, 2001. doi:10.1007/978-1-4613-0255-1_6.

16 Delbert Fulkerson and Oliver Gross. Incidence matrices and interval graphs. *Pacific Journal of Mathematics*, 15(3):835–855, September 1965. doi:10.2140/pjm.1965.15.835.

17 Leslie Ann Goldberg, Paul W. Goldberg, Cynthia A. Phillips, Elizabeth Sweedyk, and Tandy Warnow. Minimizing phylogenetic number to find good evolutionary trees. *Discrete Applied Mathematics*, 71(1):111–136, 1996. doi:10.1016/S0166-218X(96)00060-1.

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

19 Dan Gusfield. Persistent phylogeny: a galled-tree and integer linear programming approach. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, pages 443–451, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2808719.2808765.

20 Michel Habib and Juraj Stacho. Unique perfect phylogeny is NP-hard. In *Proceedings of the 22nd Annual Conference on Combinatorial Pattern Matching*, CPM'11, pages 132–146, Berlin, Heidelberg, 2011. Springer-Verlag. doi:10.1007/978-3-642-21458-5_13.

21 Richard R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, February 2002. doi:10.1093/bioinformatics/18.2.337.

22 Witold LipskiJr. Generalizations of the consecutive ones property and related np-complete problems1. *Fundamenta Informaticae*, 6(1):53–69, 1983. doi:10.3233/FI-1983-6104.

23 Kiyomi Morita, Feng Wang, Katharina Jahn, Tianyuan Hu, Tomoyuki Tanaka, Yuya Sasaki, Jack Kuipers, Sanam Loghavi, Sa A. Wang, Yuanqing Yan, Ken Furudate, Jairo Matthews, Latasha Little, Curtis Gumbs, Jianhua Zhang, Xingzhi Song, Erika Thompson, Keyur P. Patel, Carlos E. Bueso-Ramos, Courtney D. DiNardo, Farhad Ravandi, Elias Jabbour, Michael Andreeff, Jorge Cortes, Kapil Bhalla, Guillermo Garcia-Manero, Hagop Kantarjian, Marina Konopleva, Daisuke Nakada, Nicholas Navin, Niko Beerenwinkel, P. Andrew Futreal, and Koichi Takahashi. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature communications*, 11(1):5327–17, 2020.

24 Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, April 2011.

25 Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976. doi:10.1126/science.959840.

26 Teresa M. Przytycka, George B. Davis, Nan Song, and Dannie Durand. Graph theoretical insights into evolution of multidomain proteins. *Journal of computational biology : a journal of computational molecular cell biology*, 13 2:351–63, 2005. URL: https://api.semanticscholar.org/CorpusID:6639563.

27 Erfan Sadeqi Azer, Mohammad Haghir Ebrahimabadi, Salem Malikić, Roni Khardon, and S. Cenk Sahinalp. Tumor phylogeny topology inference via deep learning. *iScience*, 23(11):101655, 2020. doi:10.1016/j.isci.2020.101655.

28 Palash Sashittal, Haochen Zhang, Christine A. Iacobuzio-Donahue, and BenjaminJ. Raphael. Condor: tumor phylogeny inference with a copy-number constrained mutation loss model. *Genome Biology*, 24(1):272–23, 2023.

29 Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature reviews. Genetics*, 18(4):213–229, April 2017.

30 Doris Tabassum and Kornelia Polyak. Tumorigenesis: It takes a village. *Nature reviews. Cancer*, 15, July 2015. doi:10.1038/nrc3971.

31 Leah L. Weber and Mohammed El-Kebir. Phyolin: Identifying a Linear Perfect Phylogeny in Single-Cell DNA Sequencing Data of Tumors. In Carl Kingsford and Nadia Pisanti, editors, *20th International Workshop on Algorithms in Bioinformatics (WABI 2020)*, volume 172 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 5:1–5:14, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.WABI.2020.5.

A Proofs

Proof (Lemma 4). Let T be a 1-Dollo phylogeny for matrix B. Let U be the set of internal, non-root nodes u of T that do not correspond to observed taxa nor are branching points (Definition 3). If U is empty, then subsequently contracting all internal edges (u, v) such that $\mathbf{b}_T(u) = \mathbf{b}_T(v)$ yields the unique compact 1-Dollo phylogeny for matrix B obtained from T. If U is non-empty, consider any node $u \in U$ and let u' be the parent of u (which exists as $u \neq r$). Since u is an internal node, we have that u does not correspond to a taxon and that u is not a branching point. Therefore, u has exactly one child v. We obtain T' by removing the edges (u', u) and (u, v), removing the node u, and inserting the new edge

(u', v). Clearly, T' remains a 1-Dollo phylogeny for B. Removing all nodes in U (in any order) and subsequently contracting all internal edges (u, v) such that $\mathbf{b}_T(u) = \mathbf{b}_T(v)$ yields the unique compact 1-Dollo phylogeny for matrix B obtained from T. ◀

Proof (Theorem 7). (⇒) Let T be a 1-Dollo linear phylogeny for matrix $B \in \{0, 1\}^{m \times n}$. By Definition 1, we have that T has exactly m leaves $v_1, \ldots, v_m$ with incoming edges $(u_1, v_1), \ldots, (u_m, v_m)$ such that $\mathbf{b}_1 = \mathbf{b}(u_1) = \mathbf{b}(v_1), \ldots, \mathbf{b}_m = \mathbf{b}(u_m) = \mathbf{b}(v_m)$. Moreover, removal of these m leaves results in a linear, chain graph T', containing nodes $u_1, \ldots, u_m$ and root r. Performing a pre-order traversal of T' starting from r yields an ordering $u_{\pi(1)}, \ldots, u_{\pi(m)}$ of the nodes $u_1, \ldots, u_m$ and therefore the m taxa of B. Note that by Definition 1 each character is gained at most once, and if gained, lost at most once. Therefore, for each character $c \in [n]$, all taxa $t \in [m]$ such that $b_{t,c} = 1$ will appear consecutively in π. Hence, B is C1P as certified by permutation π.

(⇐) Let $B \in \{0, 1\}^{m \times n}$ be C1P. As such, there exists a permutation $\pi : [m] \to [m]$ such that for each column c there exists at most one row t such that $b_{\pi(t-1),c} = 0$ and $b_{\pi(t),c} = 1$ and at most one row t' such that $b_{\pi(t'-1),c} = 1$ and $b_{\pi(t'),c} = 0$. Let T be the tree obtained from $T(B, \pi)$ following Definition 6. We claim that T is a 1-Dollo phylogeny for B in line with Definition 1. Clearly, each node v of T is labeled by a binary vector $\mathbf{b}(v) \in \{0, 1\}^n$ (condition (i)), the root $r = u_0$ is labeled by $\mathbf{b}(r) = \mathbf{0}$ (condition (ii)) and each taxon $t \in [m]$ corresponds to a unique leaf $v_t = \sigma(t)$ with parent u_t such that $\mathbf{b}(v_t) = \mathbf{b}(u_t) = \mathbf{b}_{\pi(t)}$ (condition (iii)). It remains to show that there is at most one gain edge and at most one loss edge (condition (iv)) for each character $c \in [n]$. However, since the permutation π details at most one interval of consecutive taxa for each character c, it follows that taxa in T gain c through at most one gain edge and lose c through at most one subsequent loss edge. Condition (v) is similar satisfied with the further specification that there is no gain edge for any character c where $b_{0,c} = 1$. Hence, T is a 1-Dollo phylogeny for B. ◀

Proof (Lemma 10). (⇒) Let T be a 1-Dollo linear phylogeny for matrix B rooted at $\mathbf{b}_0$. Moreover let $B' = B'(B, \mathbf{b}_0)$ be the $(m + 1) \times (n + 1)$ matrix obtained from B and $\mathbf{b}_0$ following Definition 9. Given T, we will construct a 1-Dollo linear phylogeny T' for B'. Specifically, we construct T' from T by re-rooting on an appended node r' labeled by the n-dimensional vector $\mathbf{b}(r') = \mathbf{0}$ and adding the edge (r', r). We additionally add leaf v_0 labeled by $\mathbf{b}_0$ and include the edge (r, v_0). Then, we extend the n-dimensional binary vector $\mathbf{b}(v) = [b_{v,1}, \ldots, b_{v,n}]^\top$ for each node v of T' with an additional entry $b_{v,n+1}$ defined as

$$b_{v,n+1} = \begin{cases} 0, & \text{if } v \in \{r, r'\}, \\ 1, & \text{if } v \notin \{r, r'\}. \end{cases} \tag{6}$$

We claim that T' is a 1-Dollo linear phylogeny for B'. Clearly, T' is node-labeled and rooted at $\mathbf{0}$. By virtue of the fact that T is a 1-Dollo linear phylogeny for matrix $B \in \{0, 1\}^{m \times n}$ rooted at $\mathbf{b}_0$, every character $c \in [n]$ has at most one gain edge and one loss edge in T', and every original taxon $t \in [m]$ is present in T'. Additionally, character $c = n + 1$ has at most one gain edge outgoing from r and no loss edges, and the new taxon $m + 1$ correspond to the leaf $\sigma(m + 1) = v_r$. Finally, the construction of T' from T retains the linearity of T.

(⇐) Let T' be a 1-Dollo linear phylogeny for B' such that T' is compact. Following Theorem 7, let $\pi : [m] \to [m]$ be the permutation such that $T(B, \pi) = T'$.

Since T' is linear, the root node r' of T' must have at most one non-leaf child. If no such node exists then B trivially has $m = 0$ taxa and the tree T' consisting of a single node labeled by $\mathbf{b}_0$ is a 1-Dollo linear phylogeny for B rooted at $\mathbf{b}_0$.

We now focus on the case where B has $m > 0$ taxa. Let r be the non-leaf child of the root node r' of T'. Since $b'_{t,n+1} = 1$ for all $t \in [m]$ and $b'_{m+1,n+1} = 0$, it must be the case that either $\pi(m+1) = 1$ or $\pi(m+1) = m+1$. To see why observe that $1 < \pi(m+1) < m+1$ would imply that character $n+1$ would has two gain edges within T', violating the definition of a 1-Dollo linear phylogeny. We may assume without loss of generality on T' that $\pi(m+1) = 1$, since if $\pi(m+1) = m+1$, then for the permutation π^* such that π^* reverses π, tree $T^* = T(B, \pi^*)$ would be a 1-Dollo linear phylogeny for B such that $\pi^*(m+1) = 1$. Therefore, since $\pi(m+1) = 1$, it must be the case that r is labeled by $\mathbf{b}_0$.

Given T', we will now construct a 1-Dollo linear phylogeny T for matrix B rooted at $\mathbf{b}_0$. Let v_r be the leaf of T' whose parent is r. Specifically, we define T as the subtree of T' rooted at r that excludes the leaf v_r. We relabel each node v of T omitting the $n+1$th entry of its original label $\mathbf{b}_{T'}(v)$.

Since T contains all gain and loss edges present in T' precisely excluding edges (r', r) and (r, v_r), contains all taxa in T' precisely excluding leaf taxon v_r, and has root label $\mathbf{b}_0$, we have that T is a 1-Dollo linear phylogeny T for B rooted at $\mathbf{b}_0$. ◀

Proof (Lemma 14). ($\Rightarrow$) Let T be a 1-Dollo linear phylogeny for matrix B rooted at $\mathbf{b}_0$ and terminating at $\mathbf{b}_*$. Moreover let $B' = B'(B, \mathbf{b}_0, \mathbf{b}_*)$ be the $(m+2) \times (n+2)$ matrix obtained from B, $\mathbf{b}_0$ and $\mathbf{b}_*$ following Definition 13. Given T, we will construct a 1-Dollo linear phylogeny T' for B'. Let r be the root of T. Since T is linear, removal of the m leaves corresponding to the m taxa yields a directed chain graph. Let s be the sink node of this graph. Specifically, we construct T' from T by re-rooting on an appended node r' labeled by the n-dimensional vector $\mathbf{b}(r') = \mathbf{0}$ and adding the edge (r', r). We additionally add leaf v_0 labeled by $\mathbf{b}_0$ and include the edge (r, v_0). Next, we add a leaf s' labeled by $\mathbf{b}_*$ and include the edge (s, s'). Then, we extend the n-dimensional binary vector $\mathbf{b}(v) = [b_{v,1}, \ldots, b_{v,n}]^\top$ for each node v of T' with an two additional entries $b_{v,n+1}$ and $b_{v,n+2}$ defined as

$$b_{v,n+1} = \begin{cases} 0, & \text{if } v \in \{r, r'\}, \\ 1, & \text{if } v \notin \{r, r'\}, \end{cases} \quad \text{and} \quad b_{v,n+2} = \begin{cases} 0, & \text{if } v \in \{r, s, s'\}, \\ 1, & \text{if } v \notin \{r, s, s'\} \end{cases} \tag{7}$$

We claim that T' is a 1-Dollo linear phylogeny for B'. Clearly, T' is node-labeled and rooted at $\mathbf{0}$. By virtue of the fact that T is a 1-Dollo linear phylogeny for matrix $B \in \{0,1\}^{m \times n}$ rooted at $\mathbf{b}_0$, every character $c \in [n]$ has at most one gain edge and one loss edge in T', and every original taxon $t \in [m]$ is present in T'. Character $c = n+1$ has at most one gain edge outgoing from r and no loss edges, and the new taxon $m+1$ corresponds to the leaf $\sigma(m+1) = v_r$. Character $c = n+2$ has at most one gain edge (r', r) and at most one loss edge incoming to s, and the new taxon $m+2$ corresponds to the leaf $\sigma(m+2) = v_s$. Finally, the construction of T' from T retains the linearity of T.

($\Leftarrow$) Let T' be a 1-Dollo linear phylogeny for B' such that T' is compact. Following Theorem 7, let $\pi : [m] \to [m]$ be the permutation such that $T(B, \pi) = T'$. Since T' is linear, the root node r' of T' must have at most one non-leaf child. If no such node exists then B trivially has $m = 0$ taxa and the tree T' consisting of nodes $\{r, s\}$ with edge (r, s) such that r is labeled by $\mathbf{b}_0$ and s is labeled by $\mathbf{b}_*$ is a 1-Dollo linear phylogeny for B rooted at $\mathbf{b}_0$ terminating at $\mathbf{b}_*$.

We now focus on the case where B has $m > 0$ taxa. Let r be the non-leaf child of the root node r' of T', and let s be the sink node of the tree constructed by removing all leaves from T'. Since $b'_{t,n+1} = 1$ for all $t \in \{1, 2, \ldots, m, m+2\}$ and $b'_{m+1,n+1} = 0$, it must be the case that either $\pi(m+1) = 1$ or $\pi(m+1) = m+2$. To see why observe that $1 < \pi(m+1) < m+2$ would imply that character $n+1$ would has two gain edges within T', violating the definition of a 1-Dollo linear phylogeny. By the same logic, since $b'_{t,n+2} = 1$ for all $t \in \{1, 2, \ldots, m, m+1\}$ and $b'_{m+2,n+2} = 0$, it must be the case that either $\pi(m+2) = 1$ or $\pi(m+2) = m+2$. However, it is clear that $\pi(m+1) \neq \pi(m+2)$. Therefore, it must be the case that either (i) $\pi(m+1) = 1$ and $\pi(m+2) = m+2$, or (ii) $\pi(m+1) = m+2$ and $\pi(m+2) = 1$.

We may assume without loss of generality on T' that this first case holds, namely, that $\pi(m+1) = 1$ and $\pi(m+2) = m+2$. If $\pi(m+1) = m+2$ and $\pi(m+2) = 1$, then for the permutation π^* such that π^* reverses π, tree $T^* = T(B, \pi^*)$ would be a 1-Dollo linear phylogeny for B such that $\pi^*(m+1) = 1$. Since $\pi(m+1) = 1$ and $\pi(m+2) = m+2$, it must be the case that r is labeled by $\mathbf{b}_0$ and s is labeled by $\mathbf{b}_*$.

Given T', we will now construct a 1-Dollo linear phylogeny T for matrix B rooted at $\mathbf{b}_0$. Let v_r be the leaf of T' whose parent is r, and let s' be the leaf of T' whose parent is s. Specifically, we define T as the subtree of T' rooted at r that excludes the leaf v_r and leaf s'. We relabel each node v of T omitting the $n+1$th and $n+2$th entries of its original label $\mathbf{b}_{T'}(v)$. ◀

Proof (Lemma 18). We will prove the above conditions in the order they were proposed.

(i) Since T itself is a 1-Dollo phylogeny and characters can thus only be gained and lost once throughout all of T, it holds that $C_1^+, \ldots, C_\ell^+$ and $C_1^-, \ldots, C_\ell^-$ are each pairwise disjoint. Additionally, sets C_i^+ and C_j^- for all distinct $i, j \in [\ell]$ must be disjoint, since no character c can be gained in one subtree and lost in another subtree rooted at the same branching point v_*. Hence, $C_1, \ldots, C_\ell$ must additionally be pairwise disjoint.

(ii) Sets $X_0, \ldots, X_\ell$ are pairwise disjoint since, by definition, each taxon $t \in [m]$ is observed as a leaf exactly once in T, and these subsets are obtaining by a traversal on T. Additionally, their union must comprise the set of all taxa in the subtree of T rooted at v_0, since $X_0 \cup \ldots \cup X_\ell$ is simply a partition of all taxa rooted under v_0.

(iii) It holds that $C_i^+ \subseteq [n] \setminus C_0^-$, since previously-lost characters C_0^- cannot be regained in any T_i where $i \in [\ell]$.

(iv) By construction, $T_0 \oslash C_0$ is a rooted, node-labeled tree. Since this tree is formed precisely by the traversal from node v_0 to the first encountered branching point v_* of T without traversing any children of v_*, $T_0 \oslash C_0$ itself has no branching points and is thus linear. Since T_0 is rooted at $\mathbf{b}_0$, $T_0 \oslash C_0$ must be rooted at $\mathbf{b}_0 \oslash C_0$.

(v) By construction, $T_i \oslash C_i$ is a rooted, node-labeled tree. Since T_i is formed precisely by the traversal of T along the i-th outgoing edge from node v_0 labeled by $\mathbf{b}_0$, $T_i \oslash C_i$ is rooted at $\mathbf{b}_0 \oslash C_i$. ◀

Proof (Lemma 20). This follows directly from Definition 1. ◀

Proof (Lemma 23). ($\Rightarrow$) Given binary matrices $\bar{B}_i$ and B_i, consider some variable character $c \in C_i$ w.r.t. B_i and $\mathbf{b}_0$. Therefore, there is some taxon $t \in X_i$ such that $b_{t,c} \neq b_{0,c}$. We distinguish two cases.

1. If $b_{0,c} = 0$, then $b_{t,c} = 1$. Thus, $\bar{b}_{t,c} = 1$ by Definition 22.
2. If $b_{0,c} = 1$, then $b_{t,c} = 0$. Thus, $\bar{b}_{t,c} = 1$ by Definition 22.

($\Leftarrow$) Given binary matrices $\bar{B}_i$ and B_i, consider some character $c \in C_i$ such that C_i contains a 1 in column c, that is, there is some taxon $t \in X_i$ such that $\bar{b}_{t,c} = 1$. We distinguish two cases.

1. If $b_{0,c} = 0$, then $b_{t,c} = 1$. Thus, $b_{0,c} \neq b_{t,c}$, so c is variable w.r.t. B_i and $\mathbf{b}_0$.
2. If $b_{0,c} = 1$, then $b_{t,c} = 0$. Thus, $b_{0,c} \neq b_{t,c}$, so c is variable w.r.t. B_i and $\mathbf{b}_0$. ◀

Proof (Theorem 27). ($\Rightarrow$) Consider a compact 1-Dollo phylogeny T for matrix $B \in \{0,1\}^{m\times n}$ rooted at $\mathbf{b}_0$. Let T_0 be the subtree of T obtained by traversing from root node v_0 labeled by $\mathbf{b}_0$ to a first branching point v_* with label $\mathbf{b}_*$, and let C_0^+, C_0^- and X_0 be the gained characters, lost characters and observed taxa, respectively, in tree T_0. Since every character not gained or lost in T_0 can never change across the taxa in X_0, sets X_0, C_0^+, C_0^- must be in agreement with B and $\mathbf{b}_0$. So Condition 1 holds.

Let $\{B_0, B_1, \ldots, B_\ell\}$ be the 1-Dollo matrix decomposition of B and $\mathbf{b}_0$ on X_0, C_0^+, C_0^-. By Lemma 18, then, the tree induced on T by taxa set X_0 is a 1-Dollo linear phylogeny T_0 for B_0 rooted at $\mathbf{b}_0$ and terminating at $\mathbf{b}_*$.

We must finally show that there exist 1-Dollo phylogenies $T_1, \ldots, T_\ell$, for $B_1, \ldots, B_\ell$, respectively rooted at $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^-, C_0^+)$. However, consider the existing subtrees $T_1, \ldots, T_\ell$ of T obtained by traversing along each of the $\ell > 1$ outgoing edges from v_* indexed by $j \in [\ell]$, and let let C_j^+, C_j^- and X_j be the gained characters, lost characters and observed taxa, respectively, in each tree T_j. By Lemma 18, these trees are already exactly 1-Dollo phylogenies for $B[X_j, C_j]$ rooted at $\mathbf{b}_*$. So we will show that for every $i \in [\ell]$, there is some value of j, with existing tree T_j, such that $B_i = B[X_j, C_j]$. To do this, we will show outright that $\{B_1, \ldots, B_\ell\} = \{B[X_1, C_1], \ldots, B[X_\ell, C_\ell]\}$.

By Lemma 18, $\{C_1, \ldots, C_\ell\}$ are a partition of characters $C_* = [n] \setminus C_0^-$ and $\{X_0, \ldots, X_\ell\}$ are a partition of taxa $[m] \setminus X_0$. Since T is a 1-Dollo phylogeny for B, every character $c \in C_j$ is variable w.r.t. B_j and $\mathbf{b}_*$ and not variable with respect to $B_{j'}$ and $\mathbf{b}_*$ for $j' \in [\ell]$ such that $j \neq j'$. So by Lemma 23 and the definition of a complement matrix, $\{X_1, \ldots, X_\ell\}$ and $\{C_1, \ldots, C_\ell\}$ are precisely a block diagonal decomposition of $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$.

Additionally, $\{X_1, \ldots, X_\ell\}$ and $\{C_1, \ldots, C_\ell\}$ must be such a block diagonal matrix decomposition of maximum size. We will prove this by contradiction. If $\{X_1, \ldots, X_\ell\}$ and $\{C_1, \ldots, C_\ell\}$ was not on maximum size, it would be the case that for some $B[X_j, C_j]$, that there would exist two submatrices $B[X_j', C_j']$ and $B[X_j'', C_j'']$ such that X_j' and X_j'' partition X_j, C_j' and C_j'' partition C_j, all characters $c \in C_j'$ were not variable w.r.t. $B[X_j'', C_j'']$ and $\mathbf{b}_*$, and all characters $c \in C_j''$ were not variable w.r.t. $B[X_j'', C_j']$ and $\mathbf{b}_*$. This implies that there exists two subtrees T_j' and T_j'' of T_j such that T_j' contains all taxa in X_j' and all gain or loss edges of C_j', T_j'' contains all taxa in X_j'' and all gain or loss edges of C_j'', and T_j' and T_j'' are disjoint from each other. But then, this implies that edge (v_*, v_j) in T from v_* to the root node v_j of T_i is not a gain or loss edge for any character $c \in [n]$. Since T is compact, this cannot be the case.

By the definition of a 1-Dollo decomposition, the sets of taxa and characters defining $B_1, \ldots, B_\ell$ also comprise a block diagonal decomposition of $\bar{B}(B, \mathbf{b}_*, X_0, C_*)$ of maximum size. But the block diagonal matrix decomposition of maximum size for any matrix is unique, so it must be true that $\{B_1, \ldots, B_\ell\} = \{B[X_1, C_1], \ldots, B[X_\ell, C_\ell]\}$. So for every submatrix B_i, there exists a 1-Dollo phylogeny T_j for B_i rooted at $\mathbf{b}_*$. So Condition 2 holds.

($\Leftarrow$) Given binary matrix B and binary vector $\mathbf{b}_0$, let there exist $X_0 \subseteq [m]$ and $C_0^-, C_0^+ \subseteq [n]$ such that (i) X_0, C_0^-, and C_0^+ are in agreement with B and $\mathbf{b}_0$ and (ii) the 1-Dollo matrix decomposition of B and $\mathbf{b}_0$ on X_0, C_0^+, C_0^- yields $\{B_0 = B[X_0, C_0], B_1 = B[X_1, C_1], \ldots, B_\ell = B[X_\ell, C_\ell]\}$ such that there exists 1-Dollo linear phylogeny T_0 for B_0 rooted at $\mathbf{b}_0 \oslash C_0$ and terminating on $\mathbf{b}_* \oslash C_0$ and 1-Dollo phylogenies $T_1, \ldots, T_\ell$ for $B_1, \ldots, B_\ell$, respectively, rooted at $\mathbf{b}_* \oslash C_i$ for $\mathbf{b}_* = \mathbf{b}(\mathbf{b}_0, C_0^-, C_0^+)$.

We will construct T. Let node v_{*0} be the node of T_0 labeled by $\mathbf{b}_* \oslash C_0$, and let v_{*i} for all $i \in [\ell]$ be the root node of T_i labeled by $\mathbf{b}_* \oslash C_i$. Then, add edge (v_{*0}, v_{*i}) for all $i \in \ell$ to the composite of $T_0 \oplus \mathbf{b}_0, T_1 \oplus \mathbf{b}_*, \ldots, T_\ell \oplus \mathbf{b}_*$, and subsequently contract all such edges.

Each node of T is clearly labeled, and the root of T is labeled by $\mathbf{b}_0 = (\mathbf{b}_0 \oslash \mathbf{X}_0) \oplus \mathbf{b}_0$, so clearly T is rooted at $\mathbf{b}_0$. First, we prove that for every character c such that $b_{0,c} = 1$, there must be no gain edge on c in T. There is no gain edge for c in T prior to v_*, since T_0 has no gain edge for c by definition of rooted 1-Dollo linear phylogeny rooted at $\mathbf{b}_0$. To demonstrate that there is no gain edge for c in T after v_*, we differentiate two cases:

1. $b_{*,c} = 1$. Then, for all $i \in [\ell], T_i$ has no gain edge for c by definition of rooted 1-Dollo phylogeny rooted at $\mathbf{b}_*$. So there is no gain edge for c in T after encountering v_*, either, by construction of T.
2. $b_{*,c} = 0$. Then, it must be the case that $c \in C_0^-$. But since $C_i^+ \subseteq [n] \setminus C_0^-$, T_i has no gain edge for c for all $i \in [\ell]$. So there is no gain edge for c in T after encountering v_*, either, by construction of T.

Second, we prove that for every character c such that $b_{0,c} = 1$, there must be at most one loss edge on c in T. To demonstrate that there is at most one loss edge for c in T after encountering v_*, we differentiate two cases:

1. $b_{*,c} = 1$. Then, clearly c was not lost in T_0. By the definition of a block diagonal decomposition, we know that c must be variable w.r.t. B_i and b_* for at most one value of $i \in [\ell]$. So there must be at most one loss edge in $T_1, \ldots, T_\ell$.
2. $b_{*,c} = 0$. Then, c must have been lost in T_0. But since $C_i^+ \subseteq [n] \setminus C_0^-$, T_i has no gain edge, and thus no loss edge, on c for all $i \in [\ell]$.

Third, we prove that for every character c such that $b_{0,c} = 0$, there must be at most one gain and at most one loss edge on c in T. We prove three statements that together demonstrate this in full:

1. For any character c, there cannot be a gain edge in T_0 and a gain edge in T_i for $i \in [\ell]$. If there was, then $c \in C_0^-$. But since $C_i^+ \subseteq [n] \setminus C_0^-$, so T_i, cannot have a gain edge, and thus cannot have a loss edge, on c for all $i \in [\ell]$.
2. For any character c, there cannot be a loss edge in T_0 and a loss edge in T_i for $i \in [\ell]$. If there was, then $c \in C_0^-$. But since $C_i^+ \subseteq [n] \setminus C_0^-$, T_i cannot have a gain edge, and thus cannot have a loss edge, on c for all $i \in [\ell]$.
3. For any character c, there cannot be a gain edge in both T_i and T_j for distinct $i, j \in [\ell]$ such that $i \neq j$. This follows from the definition of a block matrix decomposition, since c must be variable w.r.t. B_i and b_* for at most one value of $i \in [\ell]$.

So T is a valid 1-Dollo phylogeny for B, and we are done. ◀