Phyolin: Identifying a Linear Perfect Phylogeny in Single-Cell DNA Sequencing Data of Tumors

Leah L. Weber
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL
leahlw2@illinois.edu

Mohammed El-Kebir[1]
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL
melkebir@illinois.edu

Abstract

Cancer arises from an evolutionary process where somatic mutations occur and eventually give rise to clonal expansions. Modeling this evolutionary process as a phylogeny is useful for treatment decision-making as well as understanding evolutionary patterns across patients and cancer types. However, cancer phylogeny inference from single-cell DNA sequencing data of tumors is challenging due to limitations with sequencing technology and the complexity of the resulting problem. Therefore, as a first step some value might be obtained from correctly classifying the evolutionary process as either linear or branched. The biological implications of these two high-level patterns are different and understanding what cancer types and which patients have each of these trajectories could provide useful insight for both clinicians and researchers. Here, we introduce the Linear Perfect Phylogeny Flipping Problem as a means of testing a null model that the tree topology is linear and show that it is NP-hard. We develop Phyolin and, through both *in silico* experiments and real data application, show that it is an accurate, easy to use and a reasonably fast method for classifying an evolutionary trajectory as linear or branched.

2012 ACM Subject Classification Applied computing → Molecular evolution

Keywords and phrases Constraint programming, intra-tumor heterogeneity, combinatorial optimization

Digital Object Identifier 10.4230/LIPIcs.WABI.2020.5

Supplementary Material Code and data are available at `https://github.com/elkebir-group/phyolin`.

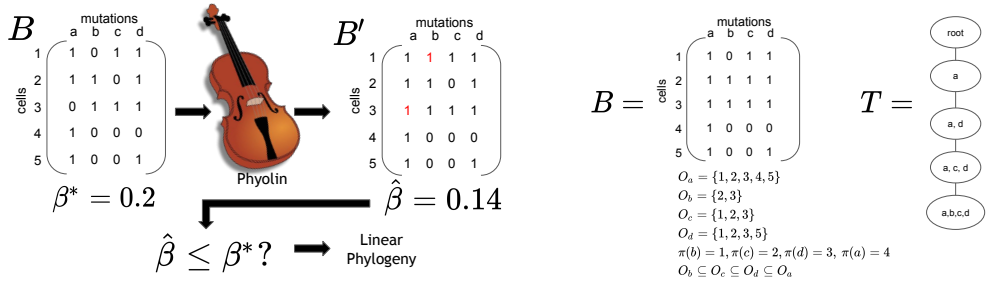
Funding *Mohammed El-Kebir*: This work was supported by the National Science Foundation under award number CCF 1850502.

Acknowledgements This work was a project in the course CS598MEB (Computational Cancer Genomics, Spring 2020) at UIUC. We thank the students in this course for their valuable feedback.

1 Introduction

The clonal theory of cancer states that tumors arise from the accumulation of somatic mutations in a population of cells [18]. This process leads to a tumor comprised of heterogeneous clones – groups of cells with similar genotypes – or what is commonly referred to as intra-tumor heterogeneity. By performing bulk and/or single-cell DNA sequencing of a heterogeneous tumor biopsy, researchers and clinicians may infer reasonable models of this evolutionary process for important downstream analysis and clinical decision-making. Specifically, the evolution of a tumor is represented by a phylogeny, i.e. a rooted tree where the leaves of the tree represent the extant cells of the tumor, internal vertices represent

[1] Corresponding author



(a) **(b)**

Figure 1 Phyolin identifying a linear perfect phylogeny in single-cell DNA sequencing data (a) A graphical depiction of Phyolin identifying a linear perfect phylogeny in single-cell DNA sequencing data when given a binary matrix B and a false negative rate β^*. (b) An example of error free single-cell data that represents a linear perfect phylogeny and the equivalent clonal tree representation.

ancestral tumor cells, and the root represents a normal cell. Due to trade-offs between the two data types, techniques for phylogeny inference have been developed for both bulk-sequencing and single-cell data individually [4–6, 13, 20] as well as combined for joint inference [15, 16, 25]. Bulk-sequencing data is less costly than single-cell data but results in a set of plausible phylogenies making it difficult to uniquely determine the true evolutionary history of a tumor [7, 19]. Conversely, single-cell data allows high resolution of the evolutionary process but is subject to high rates of sequencing errors and is more expensive than bulk-sequencing. In particular, single-cell sequencing has a high false negative rate, as much as 40% [8], implying that actual mutations present in a cell might not be indicated correctly in the resulting data. Doublets, where multiple cells are simultaneously sequenced as a single cell, are also a unique challenge of single-cell data.

One important open question is whether certain types or subtypes of cancers follow specific evolutionary patterns. Since tumors are typically biopsied at only a single point in time for reasons related to patient care, there does not yet exist sufficient longitudinal data to fully answer this question. However, it is believed that there are four high-level categories of tumor evolution: linear evolution, branching evolution, neutral evolution and punctuated evolution [3]. Of these four types, the simplest is linear evolution and will be the focus of this work. Linear evolution results when subsequent driver mutations develop a strong selective advantage and outcompete other clones during a clonal expansion [3]. By contrast, in branching evolution, a clone can diverge into separate lineages resulting in distinct branches and a tree-like model of evolution.

A useful first step in gaining insight into the evolutionary patterns of different cancer types is to determine the likelihood of each evolutionary process under available sequencing data. Suppose, we are given single-cell data in the form of a matrix where each row in the data is a cell and each column is a single-nucleotide variant (SNV), hereafter referred to as mutation. The entries in the data would then be either 1 or 0 indicating the presence of a mutation in a particular cell. Suppose also that we assume a null model of linear evolution and are given a false negative rate for the technology under which the single-cell sequencing was performed. We could then determine the minimum number of changes from 0 to 1, indicating the entry was a false negative, such that the data is representative of a linear perfect phylogeny. This would then provide an estimate of the false negative rate which could be compared with the expected false negative rate.

Azer et al. [1] utilized a deep learning approach to decide if single-cell data indicates whether a tumor followed a linear or branched evolutionary process. Although their method is fast at prediction time and performs well on simulated data, it has not yet been proved whether the problem of identifying the minimum number of flips to obtain a linear perfect phylogeny is NP-hard. Additionally, the neural networks are trained on inputs of a fixed size and while padding could be used in predicting an input smaller than the fixed size [1], a new network would have to be trained if the input size is larger than the trained network. This drawback significantly reduces the speed advantage of such an approach as training of neural networks is both a time consuming and intensive process.

Here, we prove that the problem of determining the minimum number of flips from 0 to 1 in single-cell data in order for the data to represent a linear perfect phylogeny under the infinite sites model is NP-hard. Therefore, we develop a method called Phyolin that makes use of constraint programming to find the minimum number of flips required to represent a linear perfect phylogeny. The outputted number of flips from Phyolin is then used to compute the estimated false negative rate to assess the plausibility of a linear evolutionary pattern (Figure 1(a)). We evaluate the performance of Phyolin on both simulated and real datasets, demonstrating that our method is an accurate and reasonably fast method for classifying an evolutionary trajectory as linear or branched.

2 Problem Statement

Let n be the number of single cells sequenced and m be the number of unique mutations present in the n cells. When a single cell is sequenced, assuming no errors, the group of mutations present in that cell form a clone of the tumor. Under the infinite sites assumption (ISA), where each mutation i is gained exactly once and never subsequently lost, each sequenced cell corresponds to a leaf and we may infer a two-state perfect phylogeny using a polynomial time algorithm [12] where the binary character states encode the presence of mutation i in a cell j. We may equivalently represent a perfect phylogeny T as a binary matrix $B \in \{0,1\}^{n \times m}$ where $b_{ij} = 1$ if cell i harbors mutation j and 0 otherwise, for all $i \in [n]$ and $j \in [m]$. We provide the following definition [11] for convenience:

▶ **Definition 1.** *Given an n by m binary-character matrix B for n cells and m mutations, a* perfect phylogeny for B *is a rooted tree T with exactly n leaves provided that:*

1. *Each of the n cells labels exactly one leaf of T.*
2. *Each of the m mutations labels exactly one edge of T.*
3. *For any cell p, the mutations that label the edges along the unique path from the corresponding leaf to the root specify all of the mutations of p whose state is one.*

Next, we formalize the notation of a set of cells that contain a mutation as the *one-state*.

▶ **Definition 2.** *The* one-state O_j *of mutation j is the set of single cells i where $b_{ij} = 1$.*

A perfect phylogeny T either depicts linear evolution or branched evolution. Intuitively, a matrix B represents linear evolution if there exists a total order of the set of *one-states* with respect to the subset relation. Otherwise, we say perfect phylogeny T represents branched evolution. Also, we note that perfect phylogeny T is not necessarily bifurcating.

Utilizing the collection of *one-states* for all m mutations, we determine if a given binary matrix B represents a linear perfect phylogeny as follows (Figure 1(b)):

▶ **Definition 3.** *A binary matrix $B \in \{0,1\}^{n \times m}$ represents a* linear perfect phylogeny *if there exists a permutation $\pi : [m] \to [m]$ such that $O_{\pi(1)} \subseteq O_{\pi(2)} \subseteq \ldots \subseteq O_{\pi(|m|)}$.*

However, single-cell sequencing is not error free and matrix B can fail to represent a linear perfect phylogeny even when it is representative of the true evolutionary process. False negatives, where a mutation that is present is not indicated as such, are particularly problematic with rates of up to 0.4 [8]. False positives, where absent mutations are indicated as present, are less of an issue in practice with rates less than 0.0005 for typically used whole-genome amplification strategies [9]. Given that false negatives are particularly prevalent, we would like to know how many false negatives would have had to occur in order for the inferred perfect phylogeny under the ISA to have a linear structure? This leads to the following problem statement.

▶ **Problem 1** (Linear Perfect Phylogeny Flipping Problem (LPPFP)). *Given a matrix $B \in \{0,1\}^{n \times m}$, find the minimum number of bit flips from 0 to 1 such that B represents a linear perfect phylogeny.*

Under the null model, the true phylogeny is linear and thus the input data B must represent a linear perfect phylogeny. Therefore, any implied branching in matrix B is interpreted as a false negative and must be corrected through flipping to represent the presumed linear perfect phylogeny. It is important to note that under this null hypothesis a trivial solution always exists for any input. In the worst case, every 0 can be flipped to a 1. This results in a binary matrix with all values equal to 1, suggesting a linear perfect phylogeny with a single clone harboring all of the mutations. By seeking the solution that requires the fewest number of flips, we maximize the likelihood of the null model given the observed data, assuming an estimated false negative rate $\beta^* \leq 0.5$ of the sequencing technology. Upon obtaining the solution to the LPPFP, we reverse the problem and compute the implied false negative rate $\hat{\beta}$ that resulted from flipping in order to assess the plausibility of our null model. Then given the following null hypothesis, $H_0 : \hat{\beta} \leq \beta^*$, rejection of H_0 is equivalent to concluding that a linear perfect phylogeny is not plausible.

3 Complexity

Following [2], we will prove that LPPFP is NP-hard by a reduction from the chain graph insertion problem, a known NP-complete problem [24].

▶ **Theorem 4.** *LPPFP is NP-hard.*

Proof. We prove that LPPFP is NP-hard by considering a decision version k-LPPFP asking whether there exist k bit flips in input matrix B from 0 to 1 yielding a linear perfect phylogeny. We claim that k-LPPFP is NP-complete by reduction from the chain graph insertion problem.

We begin by stating the definition of a chain graph and introduce the chain graph insertion problem.

▶ **Definition 5.** *A bipartite graph $G = (X \cup Y, E)$ is a* chain graph *if there exists a permutation $\phi : \{1, \ldots, |Y|\} \to Y$ such that $\eta(\phi(1)) \subseteq \eta(\phi(2)) \subseteq \ldots \subseteq \eta(\phi(|Y|)$ where $\eta(v) = \{w \in X : (v,w) \in E\}$ is the set of adjacent nodes of v.*

▶ **Problem 2** (Chain Graph Insertion Problem (CG-IP) [24]). *Given a bipartite graph $G = (X \cup Y, E)$ and integer k, does there exist a chain graph $G' = (X \cup Y, E')$ such that $E \subseteq E'$ and $|E| + k = |E'|$?*

k-LPPFP $\in$ NP because given a certificate (set of k flips from 0 to 1) to k-LPPFP, we could order the columns by increasing cardinality of the resulting one-state sets and then check if that permutation satisfies the definition of a linear perfect phylogeny. We will now show that CG-IP $\leq_p$ k-LPPFP.

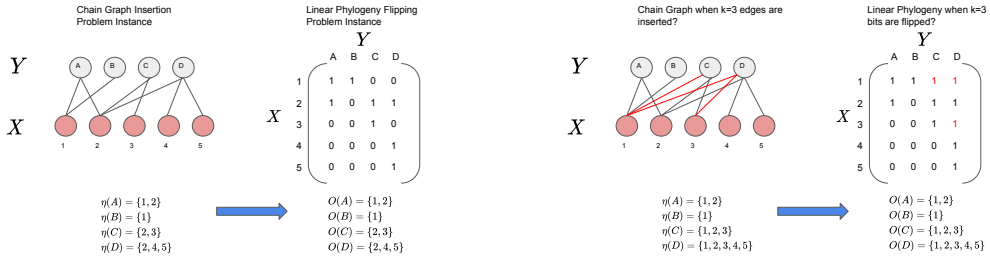


Figure 2 Chain graph insertion problem reduction (a) Polynomial time reduction of the CG-IP to LPPFP. (b) An equivalent solution of the CG-IP and LPPFP when $k = 3$.

Starting from an instance $(G = (X \cup Y,E),k)$ of CG-IP, we construct an instance B of k-LPPFP in the following manner. Binary matrix B has $|X|$ rows and $|Y|$ columns, and its entries are directly obtained from the edge set E of G: For each $v \in Y$, we set $b_{iv} = 1$ if i is a neighbor of v for all $i \in X$ and let $b_{iv} = 0$ otherwise. This can be done in polynomial time. Figure 2(a) demonstrates the polynomial time reduction. We claim that k edge insertions suffice to obtain a chain graph from G if and only if B represents a linear perfect phylogeny when k bits are flipped from 0 to 1.

($\Longrightarrow$) Suppose $(G,k) \in$ CG-IP. Then there exists an edge set $D \subseteq \{(u,v) : u \in X, v \in Y\} - E$ such that $|D| = k$ and $H = (X \cup Y, E \cup D)$ a chain graph. Then by definition of chain graph, there exists an permutation $\phi : \{1 \dots |Y|\} \to Y$ such that $\eta(\phi(1)) \subseteq \eta(\phi(2)) \subseteq \dots \subseteq \eta(\phi(|Y|))$. It is easy to see that by construction, $\eta(v) = O_v$, for all $v \in Y$. Since ϕ exists, a permutation π of the one-states also exists. Therefore, B represents a linear perfect phylogeny.

($\Longleftarrow$) Suppose $(B,k) \in k$-LPPFP. Then there exists a set F of positions (i,j) where $b_{ij} = 0$ to $b_{ij} = 1$ such that $|F| = k$. Let B^* be the resulting matrix after each flip at position $(i,j) \in F$ is made. Then B^* represents a linear perfect phylogeny and there exists a permutation $\pi : \{1 \dots m\} \to [m]$ such that $O_{\pi(1)} \subseteq O_{\pi(2)} \subseteq \dots \subseteq O_{\pi(|m|)}$. Using the equivalence between one-states and neighbors, $H = (X \cup Y,E)$ can be constructed from B^* in the following manner. First, create the set $X = [n]$ from the rows of B^* and the set $Y = [m]$ from the columns of B^*. Then, create the set E of edges as $\{(x,y) \in X \times Y \mid b^*_{xy} = 1\}$. By construction, $H = (X \cup Y, E)$ is a chain graph. ◀

4 Method

4.1 Model

To solve the LPPFP, we formulate a constraint optimization problem (COP) [21]. A COP is a constraint satisfaction problem (CSP) with an objective function that specifies which feasible solutions are preferred based on an optimization criteria. A CSP is defined by a tuple $(\mathcal{X},\mathcal{D},\mathcal{C})$, where $\mathcal{X} = \{x_1, \dots, x_n\}$ is the set of decision variables, $\mathcal{D} = \{d_1, \dots d_n\}$ is the set of domains for $\mathcal{X}$ and $\mathcal{C}$ is a set of constraints that must be satisfied. A solution $a \in \mathcal{A}(\mathcal{X},\mathcal{D},\mathcal{C})$ to a CSP is an assignment of values $\{x_1 \mapsto v_1, \dots, x_n \mapsto v_n\}$ such that $v_i \in d_i$ for all $i \in [n]$ and all constraints C are satisfied. To facilitate an objective function $f : \mathcal{A}(\mathcal{X},\mathcal{D},\mathcal{C}) \to \mathbb{R}$, an initial assignment $\hat{a} \in \mathcal{A}(\mathcal{X},\mathcal{D},\mathcal{C})$ is found. Then, a preference constraint is added to C, such that $f(a) \leq f(\hat{a})$ for a minimization problem or $f(a) \geq f(\hat{a})$ for a maximization

problem. The search is continued and the preference constraint updated each time a feasible assignment $\hat{a}$ is found until no more feasible assignments exist. When this occurs, the assignment $\hat{a}$ is returned and $f(\hat{a})$ is the objective value. We note that in problems where multiple optimal solutions exist, it is possible to return all such valid solutions. But even though multiple solutions may exist, our focus is on assessing the plausibility of the null hypothesis. Therefore, it is sufficient to consider any optimal solution even if the respective assignments yield different linear perfect phylogenies.

First, we describe the set $\mathcal{X}$ of decision variables and the associated domains $\mathcal{D}$ used in the formulation. The set $\mathcal{X}$ contains the variables $\mathbf{x}$ and $\mathbf{c}$. Intuitively, the values taken by $\mathbf{x}$ represent a binary matrix B' used to represent a linear perfect phylogeny after flipping. More formally, given a set n of cells and a set m of mutations, let $x_{ij} = 1$ if cell i has mutation j in the linear perfect phylogeny B' after flipping and 0 otherwise for each cell $i \in [n]$, and mutation $j \in [m]$. Then, $\mathcal{D}(x_{ij}) = \{0,1\}$, for all $i \in [n], j \in [m]$. Note that a decision variable is defined for every entry in B, even though only flips from 0 to 1 are allowed. This is to facilitate future handling of false positives. The variables $\mathbf{c}$, are used to define a permutation of the columns in B', such that after flipping is completed, B' will adhere to the definition of a linear perfect phylogeny. Recall that in order to represent a linear perfect phylogeny, there must exist permutation $\pi : [m] \to [m]$ such that $O_{\pi(1)} \subseteq O_{\pi(2)} \subseteq \ldots \subseteq O_{\pi(m)}$. Let $c_j = \pi(j)$ for all $j \in [m]$. Then $\mathcal{D}(c_j) = [m]$ for all $j \in [m]$.

Since, our goal is to find the linear perfect phylogeny that requires as few flips as possible, that is minimizing the number of false negatives we infer, we define an objective function that minimizes the number of flips from 0 to 1,

$$\min \sum_{i \in [n]} \sum_{j \in [m]} (x_{ij} - b_{ij}). \tag{1}$$

The set $\mathcal{C}$ ensures that the outputted binary matrix B' meets the conditions of representing a linear perfect phylogeny and also that only flips from 0 to 1 can be made. The set $\mathcal{C}$ consists of three constraints,

$$\textsc{alldifferent}(c), \tag{2}$$

$$(b_{ij} = 1) \Rightarrow (x_{ij} = 1) \qquad \forall i \in [n], \forall j \in [m], \tag{3}$$

$$(c_k < c_j) \Rightarrow (x_{ij} \leq x_{ik}) \qquad \forall k,j \in [m], \forall i \in [n]. \tag{4}$$

Equation (2) is a global constraint that ensures that every mutation is assigned a unique ordering in the permutation. Equation (3) ensures that all entries in B, where $b_{ij} = 1$ remain 1 in the linear perfect phylogeny B'. That is, they cannot be flipped from 1 to 0. Finally, Equation (4) ensures the defining property of a linear perfect phylogeny is met by ensuring that if $\pi(k)$ is less than $\pi(j)$ then it must hold that $O_k \subseteq O_j$ for all $k,j \in [m]$.

We implemented Phyolin in C++ utilizing IBM ILOG CP OPTIMIZER[2]. Phyolin is publicly available at `https://github.com/elkebir-group/phyolin`.

4.2 Null Hypothesis

The formulation of the method assumes a null hypothesis that the phylogeny is linear. The output of Phyolin is an estimate of the false negative rate $\hat{\beta}$ under this hypothesis such that

$$\hat{\beta} = \frac{\sum_{i \in [n]} \sum_{j \in [m]} \mathbb{1}(b'_{ij} = 1, b_{ij} = 0)}{\sum_{i \in [n]} \sum_{j \in [m]} b'_{ij}}, \tag{5}$$

[2] `https://www.ibm.com/analytics/cplex-cp-optimizer`

where $B' = [b'_{ij}]$ is the value of the decision variables $\mathbf{x}$ obtained from the solution of Phyolin and $B = [b_{ij}]$ is the input matrix.

Given some threshold β^*, which could be based on knowledge of the system estimated false negative rate or alternatively conservatively set at 0.4 [8], then we can reject the null hypothesis that the phylogeny is linear whenever $\hat{\beta} > \beta^*$.

5 Results

In order to evaluate Phyolin, we perform *in silico* experiments as well as run Phyolin on real data. First, we seek to evaluate the performance of Phyolin when the simulated data closely approximates a recently published high throughput single-cell DNA sequencing study of an acute myeloid leukemia (AML) cohort [17] (Section 5.1). Section 5.2 describes the application of Phyolin to patients with childhood acute lymphoblastic leukemia [10]. All experiments were conducted on a server with Intel Xeon Gold 5120 dual CPUs with 14 cores each at 2.20GHz and 512 GB RAM.

5.1 Simulations Approximating an Acute Myeloid Leukemia Cohort

In a recent study, Morita et al. [17] performed high-throughput targeted droplet microfluidic DNA single-cell sequencing on a cohort of 77 patients with acute myeloid leukemia (AML) and inferred the evolutionary tree of each patient using SCITE [13]. Utilizing high-throughput sequencing resulted in a median of 7,584 cells sequenced per patient [17]. AML is a cancer type where both linear and branching evolutionary patterns are suspected [17]. As a result, the set of trees published [17] were a mix of linear and branched patterns.

Unfortunately, the single-cell data is not yet publicly available but the estimated clonal prevalence of each clone in the inferred publish tree was included in the supplementary material along with the estimated per patient false negative rate for the sequencing technology. The clonal prevalence of clone i is the number of cells in the sample mapped to clone i divided by the total number of cells in the sample. We utilize this published data to evaluate Phyolin on a scenario that closely aligns to a highly realistic scenario. Therefore, we utilize the published clonal prevalence rates for a subset of 12 patients in the cohort in order to simulate the total number of single-cells sequenced at false negative rate $\beta = 0.05$. This value of β is to approximate the average system false negative rate for the sequencing technology used in [17]. The subset contains six patients with linear trees and mutations ranging from 3-5 and six patients with branched trees and mutations ranging from 3-7. A total of 10 replications were performed per simulated patient. Table 1 shows a summary of the patients selected for inclusion in the simulation study.

We set an upper limit on runtime of Phyolin at 500 seconds with 80% of the replications returning an optimal solution in under the time limit. We chose the 500 second time limit to facilitate timely analysis of the input data. Figure 3(a) shows the distribution of runtime by the simulated evolutionary pattern. Linear patterns resulted in a median runtime of 33 seconds (IQR: 8-82 seconds) and branched patterns resulted in a median of 156 seconds (IQR: 117-502 seconds). The median input size (cells $\times$ mutations) of the linear patterns was 24,435 and 28,775 for branched patterns. The largest input was for AML-74 with 9,279 cells by 5 mutations and no replications completed within the time limit. Only 1 replication with a linear pattern did not complete within the time limit. This implies that optimal solutions are found much faster when the true pattern is linear.

Figure 3(b) compares the distribution of the estimated false negative rate $\hat{\beta}$ for the patients with linear versus branched published trees over all 10 replications. The simulated system false negative rate was 5% and is shown as a dashed line where relevant. From

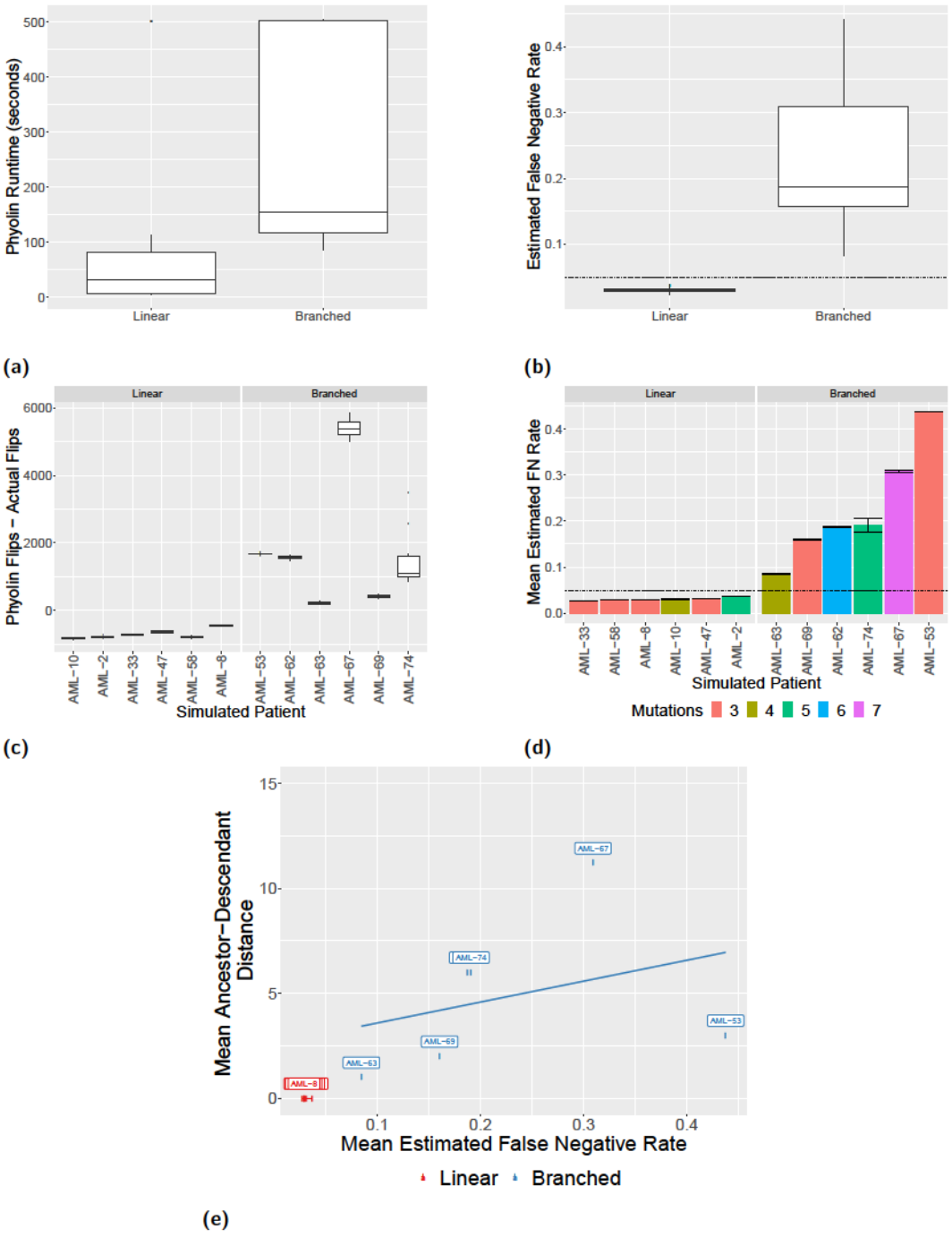


Figure 3 Phyolin results of the *in silico* experiments on a simulated cohort of patients with AML (a) A comparison of Phyolin runtimes in seconds between instances with different evolutionary patterns. (b) A comparison of the distribution of $\hat{\beta}$ between simulated linear and branched topologies over 10 replications. (c) The distribution of the difference between the number of flips performed by Phyolin and the simulated flips per patient. (d) The mean value of $\hat{\beta}$ for each patient along with the standard error. (e) Relationship between estimated false negative rate and the ancestor-descendant distance. Each point represents the mean value over 10 replications and is labeled by the numerical patient identifier. A linear trend line is shown with a 95% confidence interval.

Table 1 Simulation study based on characteristics of a published AML cohort [17] Shown is the patient identifier, the published evolutionary pattern of the tree, the number of mutations, the total cells sequenced [17], the median number of false negatives over 10 replications, Phyolin estimated number of false negatives over 10 replications, the median $\hat{\beta}$ over 10 replications, and the median probability of a linear perfect phylogeny as determined by the comparison deep learning method [1].

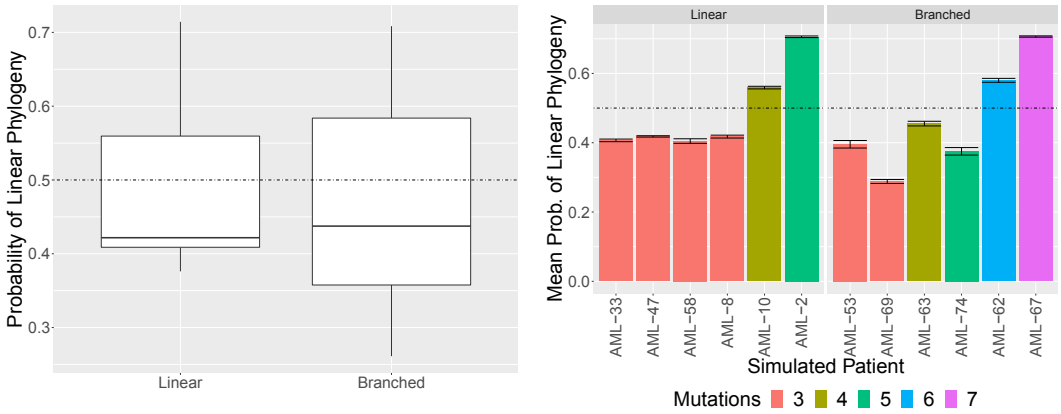
patient	pattern	m	n [17]	median flips	Phyolin median flips	median $\hat{\beta}$	med prob.
AML-2	Linear	5	7931	1826	1039	0.037	0.70
AML-8	Linear	3	4675	759	294	0.029	0.42
AML-10	Linear	4	8729	1427	584	0.037	0.56
AML-33	Linear	3	8120	1091	350	0.027	0.41
AML-47	Linear	3	6491	1135	488	0.032	0.42
AML-58	Linear	3	8170	1280	472	0.029	0.40
AML-53	Branched	3	8013	544	2220	0.44	0.39
AML-62	Branched	6	4027	726.5	2299	0.19	0.58
AML-63	Branched	4	8347	1238.5	1432	0.084	0.44
AML-67	Branched	7	6024	1061.5	6440	0.31	0.71
AML-69	Branched	3	7462	651.5	2122	0.16	0.29
AML-74	Branched	5	9279	1020	294	0.17	0.38

Figure 3(b), we note a significant difference in the distributions between linear and branched instances. Additionally, the median of the linear perfect phylogeny patients is 0.03 (IQR: 0.028-0.032) and every linear replication is less than $\beta^* = 0.05$ while the median of the branched perfect phylogeny patients is 0.19 (IQR: 0.16-0.31) and every branched replication is greater than $\beta^* = 0.05$.

Figure 3(c) compares the difference between the number of flips performed by Phyolin and the actual number of simulated false negatives. Interestingly, Phyolin performs fewer flips than simulated false negatives for all linear patients and performs a much higher number of flips than simulated for branched patterns. This underestimation can be attributed to the fact that not every false negative implies that branching occurs and so Phyolin only needs to flip those that do. Thus, the difference between the number of Phyolin flips and the number of simulated flips may be negative. The fact that Phyolin performs more flips than was actually simulated in the branched cases aligns with the original intuition for its design as we not only need to correct false negatives but also true negatives in order to force the pattern to be linear.

Figure 3(d) shows the mean $\hat{\beta}$ and standard error for each simulated patient over the 10 replications. The number of mutations are also shown in order to investigate if increasing number of mutations increases $\hat{\beta}$. Although there appears to be some effect when increasing the number of mutations, it does not strictly hold. However, utilizing a strict threshold of $\beta^* = 0.05$ results in perfect classification of the topology for all patients and all replications.

Since the number of mutations does not necessarily impact the estimated false negative rate $\hat{\beta}$, another consideration is whether or not $\hat{\beta}$ increases with the amount of branching. To this end, we compare the ancestor-descendant distance between the simulated true tree B^* and the inferred linear tree B'. A mutation x is an *ancestor* of mutation y if x occurs on the path from the root to y, in which case y is said to be a descendant of x. Ancestor-descendant (AD) distance is defined as the size of the symmetric difference between the sets of ordered pairs of characters, or ancestor-descendant pairs, introduced on distinct edges of perfect phylogenies B^* and B'. A higher AD distance implies a greater degree of branching in



(a) **(b)**

Figure 4 Results of the Deep Learning approach [1] applied to the *in silico* experiments on a simulated cohort of patients with AML (a) A comparison of the distribution of the probability of a linear perfect phylogeny between simulated linear and branched topologies over 10 replications. (d) The mean value of the probability of a linear perfect phylogeny for each patient along with the standard error. A horizontal line indicates the threshold probability (0.05) used to classify an input as linear.

the true tree. For example AML-63 has only one branch and AML-67 has three distinct branching events. Figure 3(e) shows the relationship between the mean estimated false negative rate $\hat{\beta}$ and the mean AD distance per simulated patient over 10 replications. The AD distance is 0 for all patients with a simulated linear perfect phylogeny. This means that Phyolin correctly infers the true tree when it is linear. Also, there is evidence of a correlation between the estimated false negative rate $\hat{\beta}$ and the AD distance.

Azer et al.'s [1] deep learning method for classifying topology is the most similar method for comparison with Phyolin. Therefore, we retrained this deep neural network to support our input size of 9300 cells and 7 mutations. We used a default hidden layer size of 100 and drop-out rate of 0.9, 5000 training examples and 500 epochs. We did not modify any other hyperparameters. The input size was selected so that only one network needed to be trained for all *in silico* experiments and we used padding for any instances where the $n < 9300$ or $m < 7$. After 200 epochs the best validation accuracy was 64.1% and completed in 2168 seconds (36.1 minutes). After 500 epochs, the best validation accuracy was 64.8% and completed in 3997 seconds (66.6 minutes). This suggests that further learning was unlikely. We report the probability that the phylogeny is linear on the same simulation instances when evaluated with the trained model.

Table 1 shows the median probability that the phylogeny is linear over the 10 replications per simulated patient. We use a cutoff of 0.5 as the threshold for classifying a topology as linear. Figure 4(a) shows the distribution of the probabilities over all patient replications by ground truth topology. Classification accuracy was 100% for 6 of the 12 simulated patients (Linear: AML-2, AML-10, Branched: AML-53, AML-63, AML-69 and AML-74) and 0% for the remaining 6 simulated patients. Figure 4(b) shows mean estimated probability per patient and standard error for the 10 replications. The predicted probability tends to increase as the number of mutations increases.

In summary, the simulated AML cohort results show that, in contrast to the Deep Learning approach [1], Phyolin correctly and quickly classifies large instances as linear with a strict threshold β^* set at the system estimated false negative rate. Furthermore, as the

Table 2 Summary of Phyolin analysis of two patients with ALL Shown is the patient identifier, the number of cells sequenced [10], the number of mutations, the number of flips performed by Phyolin, the estimated false negative rate $\hat{\beta}$ and the false negative rate threshold β^* that was estimated for the sequencing technology [14].

patient	cells sequenced [10]	mutations	Phyolin flips	$\hat{\beta}$	β^* [10]
Patient 2	115	16	403	0.36	0.18
Patient 6	146	10	191	0.15	0.18

amount of branching increases, the estimated $\hat{\beta}$ tends to increase. Thus the greater the difference between $\hat{\beta}$ and β^*, the more confident we can be in rejecting the null hypothesis that the phylogeny is linear.

5.2 Real Data of Childhood Acute Lympoblastic Leukemia Patients

Gawad et al. [10] performed single-cell DNA sequencing on a cohort of six patients with childhood acute lymphoblastic leukemia (ALL). As a subtype of leukemia, ALL is also postulated to follow both linear and branched trajectories [22]. We evaluate Phyolin on two of the six patients in this cohort: Patient 2 and Patient 6. The input size for Patient 2 was 115 cells by 16 mutations. Two independent, previous analyses of the sequencing data of Patient 2 suggested a branched topology [15, 23].

In addition, we consider Patient 6 because this patient was analyzed by the deep learning method [1]. In another line of work, Kuipers et al. [14] investigated the validity of the ISA in single-cell data within the ALL dataset [10]. Using their method, they compared the likelihood of the data under both a finite sites and infinite sites model via a Bayes Factor and determined with high probability that Patient 6 suffered a loss of the *SUSD2* mutation. They used SCITE [13] to infer trees from the single-cell data under both an infinite sites and finite sites model. These trees are show in Figure 5. The tree inferred under the ISA is linear (Figure 5(a)) while the tree inferred under the finite sites model is branched due to the loss of *SUSD2* (Figure 5(b)). The input size for Patient 6 was 146 cells by 10 mutations. Table 2 summarizes results obtained by Phyolin.

For Patient 2, Phyolin estimated a false negative rate of 0.36, which is much greater than the rate of 0.18 estimated in [10]. Taking $\beta^* = 0.18$ implies rejecting the null hypothesis of a linear perfect phylogeny. This concurs with the branched trees published in [15, 23]. Phyolin utilized the 500 second time limit to complete its run for Patient 2 despite the small input size.

For Patient 6, Phyolin estimated a false negative rate of 0.15. The published false negative rate β^* in [10] was 0.18. This implies that we cannot reject the null hypothesis of a perfect linear perfect phylogeny. Indeed, the linear tree output by Phyolin does concur with Figure 5(a). The comparison deep learning approach [1] also concluded that the phylogeny was linear with probability 0.79. It is important to note that allowing for mutation loss of *SUSD2*, in line with [14], will lead to a smaller false negative rate $\hat{\beta}$. This suggests that incorporating mutation loss into Phyolin is an important area for future study.

6 Discussion

In this work, we introduced the Linear Perfect Phylogeny Flipping Problem and showed that it is NP-hard. To solve the LPPFP, we developed a method named Phyolin that takes as input a binary matrix of single-cell DNA sequencing data and then identifies a linear

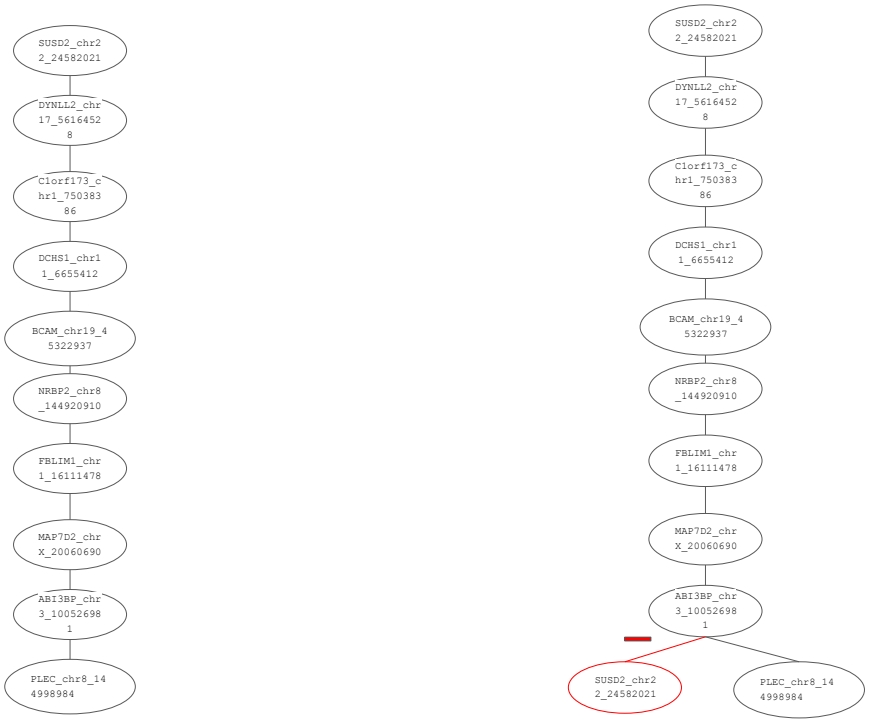


Figure 5 Patient 6 candidate trees Two possible trees published in [14]. (a) Tree adheres to the ISA, (b) Tree with a branched topology indicating a mutation loss of *SUSD2*.

perfect phylogeny in the data by assuming that any implied branching are actually false negatives. It returns an estimate of the false negative rate under the null hypothesis that the perfect phylogeny is linear and allows the user to compare this estimate to a false negative threshold at which the null hypothesis of a linear perfect phylogeny can subsequently be accepted or rejected. We tested Phyolin on both simulated data and real data and showed that it is more accurate than a recent deep learning method [1]. In conclusion, Phyolin is a reliable, easy to use and fast method to assess the likelihood of a linear evolution before more complex reconstruction methods are utilized.

There are several future research directions. First, Phyolin lacks an absolute criterion or threshold for rejecting the null hypothesis that the phylogeny is linear. To address this, we plan to modify Phyolin so that instead of trying to solve the problem to optimality, a user could input a false negative rate at which he or she would fail to reject the null hypothesis of a linear perfect phylogeny if any solution exists below the supplied threshold. This would allow Phyolin to explicitly solve a constraint satisfaction problem and likely reduce the runtime.

Second, Phyolin can be modified to allow false positives, which means allowing flips from 1 to 0. However, before that modification is made, more robust *in silico* experiments should be conducted with simulated false positives and doublets. Although false positives are rare, it is possible that a single or a few critically positioned false positives requires excessive inference of false negatives in order to represent a linear perfect phylogeny. High doublet rates could potentially result in low estimates of false negative when the true phylogeny is branched. Therefore, a constraint could also be incorporated to ignore up to the inputted number of doublets.

Third, Phyolin could easily incorporate mutation and cell clustering through additional constraints when supplied with a number of cell clusters and/or mutation clusters. A search could be performed to find the optimal number of clusters such that the likelihood of the data of is maximized. Fourth, even when the phylogeny is branched, the trunk of the tree may be linear or there might be a long branch with linear evolution within that branch. This means that a subset of the mutations form a linear perfect phylogeny. A future direction is to explore if Phyolin can identify a subset of mutations that are likely to be truncal or form a long branch of the tree, thus potentially providing fast partial inference of the tree.

Finally, given the results of ALL Patient 6, exploring evolutionary models that allow ISA violations, such as mutation loss, is an exciting direction for future study. In particular, modeling Patient 6 as a 1-Dollo phylogeny, where each mutation is gained only once and subsequently lost at most once, could potentially be achieved by replicating each column once and then using Phyolin. If the two columns representing the same mutation are distinct in the inferred linear perfect phylogeny, then that implies that the mutation was lost once. The plausibility of a linear perfect phylogeny under both ISA and under a 1-Dollo model could be compared [5].

References

1 Erfan Sadeqi Azer, Mohammad Haghir Ebrahimabadi, Salem Malikić, Roni Khardon, and S Cenk Sahinalp. Tumor Phylogeny Topology Inference via Deep Learning. *bioRxiv*, 2020. doi:10.1101/2020.02.07.938852.

2 Duhong Chen, Oliver Eulenstein, David Fernandez-Baca, and Michael Sanderson. Minimum-flip supertrees: complexity and algorithms. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(2):165–173, 2006.

3 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.

4 Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015.

5 Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.

6 Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

7 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, 2016.

8 Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang, X Sunney Xie, and Yanyi Huang. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences*, 112(38):11923–11928, 2015.

9 Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang, X Sunney Xie, and Yanyi Huang. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 112(38):11923–11928, September 2015.

10 Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, 111(50):17947–17952, 2014.

11 Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997. doi:10.1017/CBO9780511574931.

12 Dan Gusfield. *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT press, 2014.

13 Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

14 Jack Kuipers, Katharina Jahn, Benjamin J Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 27(11):1885–1894, 2017.

15 Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):1–12, 2019.

16 Salem Malikic, Farid Rashidi Mehrabadi, Simone Ciccolella, Md Khaledur Rahman, Camir Ricketts, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S Cenk Sahinalp. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome research*, 29(11):1860–1877, 2019.

17 Kiyomi Morita, Feng Wang, Katharina Jahn, Jack Kuipers, Yuanqing Yan, Jairo Matthews, Latasha Little, Curtis Gumbs, Shujuan Chen, Jianhua Zhang, et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *bioRxiv*, 2020.

18 Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

19 Yuanyuan Qi, Dikshant Pradhan, and Mohammed El-Kebir. Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors. *Algorithms for Molecular Biology*, 14(1):19, 2019.

20 Edith M Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.

21 Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach Third Edition*. Pearson, 2010.

22 Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood, The Journal of the American Society of Hematology*, 120(20):4191–4196, 2012.

23 Leah Weber, Nuraini Aguse, Nicholas Chia, and Mohammed El-Kebir. PhyDOSE: Design of follow-up single-cell sequencing experiments of tumors. *BioRxiv*, 2020.

24 Mihalis Yannakakis. Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic Discrete Methods*, 2(1):77–79, 1981.

25 Hamim Zafar, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, 29(11):1847–1859, 2019.