Generalized Matching Distance: Tumor Phylogeny Comparison Beyond the Infinite Sites Assumption

Quoc Nguyen
nguyenq2@carleton.edu
Carleton College
Northfield, MN, USA

Layla Oesper
loesper@carleton.edu
Carleton College
Northfield, MN, USA

ABSTRACT

As the field of tumor phylogenomics matures, numerous methods have been developed to infer tumor phylogenies from many types of sequencing data. The tumor phylogenies being inferred have transitioned from abiding strictly to the Infinite Sites Assumption (ISA), which says that mutations are gained once and never lost, to more relaxed and more biologically accurate models such as Camin-Sokal or k-Dollo models which allow mutations to be either gained or lost multiple times, respectively. As the tumor phylogenies being inferred have become more attuned to the underlying biology of cancer, methods of comparing, or computing distances, between these phylogenies have not yet caught up. In order to address this discrepancy, we propose the Generalized Matching Distance (GMD) Problem which allows for ISA distance measures to be applied to non-ISA phylogenies after a particular type of transformation. We provide a simple, but effective solution for exactly solving the GMD Problem which is often efficient enough for many tumor phylogenies. We also provide a heuristic approach to solving the GMD Problem for instances where our exact solution is not appropriate. In our simulated experiments, we show that by using our approach to solve the GMD Problem we can effectively use ISA tumor distance measures to compare phylogenies with parallel mutations (those that are gained multiple times). Additionally, we show that our heuristic approach works well on a subset of phylogenies under the Camin-Sokal and k-Dollo models. Finally, we apply our method for solving the GMD to three tumor phylogenies generated from a colorectal cancer patient. The data for our experiments and the code for using GMD is available at: https://bitbucket.org/oesperlab/gmd/src/master/

CCS CONCEPTS

• **Applied computing** → *Molecular evolution*; *Computational genomics.*

KEYWORDS

cancer, phylogeny, infinite sites assumption, distance measure

ACM Reference Format:
Quoc Nguyen and Layla Oesper. 2023. Generalized Matching Distance: Tumor Phylogeny Comparison Beyond the Infinite Sites Assumption. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3584371.3612970

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3612970

1 INTRODUCTION

The clonal theory of cancer [21] describes how tumors grow as the result of an evolutionary process. As descendants of the original founder cell acquire new somatic mutations, the resulting cell populations may proliferate even more quickly, leading to a tumor that is a heterogeneous collection of different cell populations. There has been a strong interest in the computational cancer research community to design computational methods that are able to reconstruct the evolutionary history of an individual's cancer as a particular type of rooted tree [3]. The vertices in this tree represent distinct populations of cells, with unique complements of somatic mutations, that exist or existed at some point during the tumor's evolution. The edges in the tree represent ancestral relationships between those cell populations. Being able to accurately identify such a tree has potential implications for both improving a general understanding of how tumors evolve and how a particular patient might be best treated [2, 20, 22].

In recent years there have been many new methods developed to infer the tree encoding a tumor's evolution from various types of sequencing data. See [11, 25, 28] for good reviews of such approaches and the challenges they face. As the number of methods for inferring a tumor's evolutionary history from DNA sequencing data have proliferated, there has been a recent interest in how to appropriately compare, or compute a distance, between two such trees. In addition to being an interesting problem in its own right, there are certain applications that would benefit immensely from such distance measures. In particular, such distance measures are essential for benchmarking the performance of novel phylogenetic inference methods on simulated data as their output needs to be compared to known ground truth trees [6].

Following the trends that appeared in the tumor phylogenetic tree inference space, the first distance measures developed for tumor evolutionary trees assumed the input trees adhered to the *infinite sites assumption (ISA)* which states that any mutation is only gained once and never lost [16]. Early such distance measures include parent-child and ancestor-descendant distances [9, 10] which capture information on the number of parent-child (ancestor-descendant) relationships in one tree but not the other. Since then, more distance measures designed specifically to capture features important in the development of cancer have been developed, but still assume the input trees adhere to the ISA. This includes CASet and DISC [6] which both aim to capture aspects of how mutations

are inherited in subsequent populations when computing the distance between two trees. This also includes Bourque distances [12], which generalize the Robinson-Foulds distance [23] from traditional phylogenetics to be applicable to tumor evolution trees. Finally, another such related distance measure that assumes ISA trees is MLTED/MLTD [13, 14] which in contrast to other distances, considers two trees that could represent the same mutational history, but at different levels of resolution, as identical.

In response to recent studies that suggest that the ISA is not appropriate in many cancers [18], the number of methods that can infer tumor evolution trees that don't adhere to the ISA has been increasing. These methods often utilize a different model of tumor evolution such as the k-Dollo model [7] which allows each mutations to be lost up to k times, or the Camin-Sokal model [4] which allows each mutation to be gained multiple times. Some recent such inference methods include SiCloneFit [27], SCARLET [24], MEDICC2 [15], recent pre-print FiMO [1], and many others. Correspondingly, there is starting to be an interest in the design of distance measures that can handle input trees that don't adhere to the ISA. However, to our knowledge, the only existing tumor tree distance measure with this capability is MP3 [5] which generalizes the classical phylogenetic concept of using rooted triplets for determining similarity and applies it to tumor phylogenetics. This method is also designed to be able to handle multiply occurring/parallel mutations or losses of mutations. However, given the number of existing distance measures that capture different features of tumor evolution, but rely on the ISA, it would be advantageous to have a way to use these measures on non-ISA trees, rather than wait for new distance methods to be designed.

In this paper, we address the need for a diverse set of tumor distance measures that don't assume the ISA. We introduce a framework that enables all non-ISA tumor phylogenetic trees to be converted into ISA trees without loss of information. Existing ISA dependent distance measures can then be applied to these transformed trees. Specifically, we propose the Generalized Matching Distance (GMD) Problem and describe both an exact algorithm and a heuristic approach to solving it. Our approaches can be applicable for any input trees that allow for mutiple gains and/or losses of mutations. On simulated data we demonstrate the effectiveness of our exact approach for solving the GMD when applied to trees with parallel mutations (those that are gained multiple times). We also demonstrate the effectiveness of our heuristic approach when input trees contain either parallel mutations or losses of mutations. Finally, on both simulated data and a real colorectal cancer data set [19] we demonstrate the ability of our approaches to effectively measure the distance between trees while maintaining important properties of the original ISA dependent distance measures.

2 METHODS

2.1 Tumor Phylogenies

We consider a tumor that contains m mutations. We won't distinguish what types of genomic alterations these may be (SNV, CNA, etc.). We model the presence or absence of a mutation as a *binary character* where 1 represents the presence of the mutation and 0 indicates its absence. Thus any cell in the tumor may be described using a binary *mutation vector* $\mathbf{b} \in \{0,1\}^m$ whose i^{th} entry, $b(i)$, indicates the state of mutation i in the cell. A *clone* is a collection of cells with identical mutation vectors. We can now describe the history of a tumor as a *tumor phylogeny* T where vertices represent clones that either currently or previously existed at some point during the tumor's evolution and directed edges represent the direct ancestral relationships between those clones. We note that inherently all edges are directed away from the root.

Definition 2.1. A *tumor phylogeny* T is a rooted tree with the following conditions:

(1) Each vertex v is labeled with a binary mutation vector $\mathbf{b}_v \in \{0,1\}^m$ indicating the mutations present in that clone.
(2) Tumors evolve from a healthy cell (without mutations), so the root r is labeled with the vector $\mathbf{b}_r = [0,\ldots,0]^T$.
(3) Any two vertices connected by an edge must have binary mutation vectors that differ in at least one place. That is, if (v,w) is an edge in T then there must exist some $i \in \{1,\ldots,m\}$ such that $b_v(i) \neq b_w(i)$.
(4) All children of any vertex v must have unique mutation vectors. That is, if vertices v and w are siblings, then $\mathbf{b}_v \neq \mathbf{b}_w$.
(5) All m mutations appear at least once in T. That is, for all $i \in \{1,\ldots,m\}$ there exists some vertex v where $b_v(i) = 1$.

We define an edge (v,w) in T as a *gain edge* for mutation i if $b_v(i) = 0$ and $b_w(i) = 1$. Similarly, we define an edge (v,w) in T as a *loss edge* for mutation i if $b_v(i) = 1$ and $b_w(i) = 0$. Note, with these definitions a single edge (v,w) may be a gain edge for mutation i but also is a loss edge for mutation j. A tumor phylogeny T can be given one of two different categorizations depending on the number of distinct mutations that are gained or lost on the edges in T. Specifically, T is called a *mutation phylogeny* if for any e in T there exists exactly one mutation i such that e is either a gain or loss edge for i. Otherwise, there must exist some edge e in T where more than one mutation is either gained or lost. In that case, we refer to the tumor phylogeny T as a *clonal phylogeny*. Intuitively, a clonal phylogeny is when some mutations cannot be ordered and are instead clustered together. Finally, we note that we may use the terms phylogeny and tree interchangeably.

2.2 Models of Tumor Evolution

The mutational history of a tumor is generally not as permissive as our definition of a tumor phylogeny. Instead, we often need to apply a model of evolution that further constricts how mutations are gained or lost. We now can identify two types of tumor phylogenies that adhere to two different existing models of evolution.

The *k-Dollo model* allows for each mutation to be gained exactly once but lost up to k-times [7]. Formally, we now define a *k-Dollo phylogeny* as follows.

Definition 2.2. A *k-Dollo phylogeny* is a tumor phylogeny T with the following additional restrictions:

(1) T contains exactly one gain edge for all mutations $i \in \{1,\ldots,m\}$.
(2) For all mutations $i \in \{1,\ldots,m\}$, T contains at most k loss edges.

We note that a 0-Dollo phylogeny represents a special case called the *Infinite Sites Assumption* (ISA) [16] where mutations are gained but never lost. The ISA model has been used extensively in the field

of tumor evolution as it provides helpful constraints for inferring tumor phylogenies (e.g., [8]). In recent years there has been a growing interest in dropping the ISA assumption [18], and in particular the k-Dollo model has been shown to be a useful alternative.

The *Camin-Sokal model* allows for mutations to be gained any number of times, but never lost [4]. This model has also been shown to be a useful alternative to the more restrictive ISA [16]. Let *k-Camin-Sokal* denote the restriction of the Camin-Sokal model where each mutation can be gained at most k times. Formally, we now define a *k-Camin-Sokal* phylogeny as follows.

Definition 2.3. A *k-Camin-Sokal phylogeny* is a tumor phylogeny T with the following additional restrictions:

(1) For all mutations $i \in \{1, \ldots, m\}$, T contains between 1 and k gain edges.
(2) T contains no loss edges for any mutation $i \in \{1, \ldots, m\}$.

2.3 Distance Measures on Tumor Phylogenies

A *distance measure* on tumor phylogenies is a function that takes in two tumor phylogenies T_1 and T_2 and returns a non-negative real valued number that indicates how dissimilar they are. The larger the value, the more dissimilar the trees are to each other; the closer the value is to 0, the more similar they are. Note that a distance measure does not need to be a distance metric (i.e. observe the triangle inequality, symmetry, etc.), although some distance measures such as CASet and DISC are distance metrics in certain contexts. Most existing distance measures that are designed for tumor phylogenies assume that the input phylogenies adhere to the ISA. In this section, we will build a framework that allows for distance measures that assume the ISA to be applied to *any* tumor phylogeny. We will first show how distance measures that assume the ISA can be applied to 1-Dollo phylogenies. Then, we will generalize that approach.

2.3.1 Distance Measures applied to 1-Dollo Phylogenies. First we will describe a transformation process to turn a 1-Dollo phylogeny into a 0-Dollo phylogeny, which is the same as a tumor phylogeny adhering to the ISA–sometimes also called a perfect phylogeny. Consider a 1-Dollo phylogeny T with m mutations and n of those mutations have a single loss. We will show how to convert T into a 0-Dollo phylogeny T' on $m+n$ mutations. The only real difference between T and T' is how the mutation vectors that label the vertices in T' are constructed, the phylogenies themselves have the same topology. So, we describe only how to construct the mutation vectors for T'. The first m indices in each mutation vector $\mathbf{b'}_v$ in T' correspond to the gains of the m characters in T. The last n indices correspond to losses of these characters (if present), effectively representing each loss state as a new character. We can construct such a phylogeny T' using the following three steps. (1) For each vertex v, directly copy over all entries from $\mathbf{b}_v$ to the first m indices in $\mathbf{b'}_v$ and set $b'_v(i) = 0$ for the remaining n indices. This means that all mutation gains are encoded in the same way as in the original phylogenies. (2) Iterate through all loss edges in T. For the j^{th} loss edge considered (v, w) in T where the loss edge corresponds to character i, set $b'_w(i) = 1$ and $b'_w(m+j) = 1$. So, instead of encoding losses as a change of a single mutation from present to absent, we instead encode it as the gain of a new 'loss' mutation while keeping the original mutation as present as well. (3) For any vertex x such that w is an ancestor of x in T, also make the following updates: $b'_x(i) = 1$ and $b'_x(m+j) = 1$. This procedure allows the original mutation and the newly created 'loss' mutation to be inherited by a descendant populations.

With this transformation, we can now describe a process for applying distance measures designed for ISA phylogenies to 1-Dollo phylogenies. Given two 1-Dollo phylogenies T_1 and T_2, convert them into 0-Dollo phylogenies T'_1 and T'_2 using the transformation described above. Now, any distance measure that assumes the ISA can be applied directly to T'_1 and T'_2 as these encode all the information from the original phylogenies.

2.3.2 Distance Measures applied to Generalized Tumor Phylogenies. We now describe how the transformation approach described above for 1-Dollo phylogenies can be generalized for any tumor phylogeny, but especially k-Dollo phylogenies or k-Camin-Sokal phylogenies. While for 1-Dollo phylogenies there was exactly one gain and at most one loss for each mutation, we may now have multiple gains and losses for each mutation. To apply the same transformation, we need to match losses and gains of the same mutations between tumor phylogenies T_1 and T_2. Mathematically, we may view such a pairing as a matching of a bipartite graph $G(T_1, T_2)$, whose vertices and edges encode allowed matches between gains and losses of mutations. To formalize this, we start by defining the matching graph $G(T_1, T_2)$ obtained from T_1 and T_2.

Definition 2.4. The *matching graph* of two tumor phylogenies T_1 and T_2 is a bipartite graph $G(T_1, T_2) = (A \cup B, E)$ whose vertices A (B) correspond to gains and losses of mutations in T_1 (T_2), and whose edge set E is composed of edges (a, b) such that a and b correspond to either two gains or two losses of the same mutation, one in each tree.

Recall that a matching M in a bipartite graph is a subset of edges such that no two edges in M are incident to the same vertex. Intuitively, a matching of the matching graph $G(T_1, T_2)$ of tumor phylogenies T_1 and T_2 describes how to match the gains and losses of the same mutation between the two phylogenies. Given a matching $M = \{(a_1, b_1), \ldots, (a_{|M|}, b_{|M|})\}$ of $G(T_1, T_2)$, let $A^- = \{a_1, \ldots, a_{|A^-|}\} \subseteq A$ and $B^- = \{b_1, \ldots, b_{|B^-|}\} \subseteq B$ be the subsets of unmatched vertices (mutations) of $G(T_1, T_2)$, where A^- indicates the unmatched mutations from T_1 and B^- the unmatched mutations from T_2. From M, A^- and B^-, we obtain the corresponding 0-Dollo (ISA) phylogenies T'_1 and T'_2 as follows. Similar to the 1-Dollo case, the topologies of the transformed phylogenies are identical, so we need only to describe how the mutation vectors for T'_1 and T'_2 are created. Each vertex in T'_1 and T'_2 are labeled by mutation vectors $\mathbf{b'}_1$ and $\mathbf{b'}_2$, respectively, of size $|M| + |A^-| + |B^-|$. The first $|M|$ indices correspond to matched gains/losses of mutations between T_1 and T_2, followed by $|A^-|$ indices corresponding to unmatched gains/losses of mutations in T_1 and then $|B^-|$ indices corresponding to unmatched gains/losses in T_2.

In $\mathbf{b'}_1$ and $\mathbf{b'}_2$ mutation gains are largely filled out in the same way as in the 0-Dollo case. That is, a 1 entry at index i indicates the gain of mutation i, and this entry persists in the mutation vector of all descendant vertices. The one difference is that a new mutation vector index j exists for each novel gain of a single mutation (e.g.,

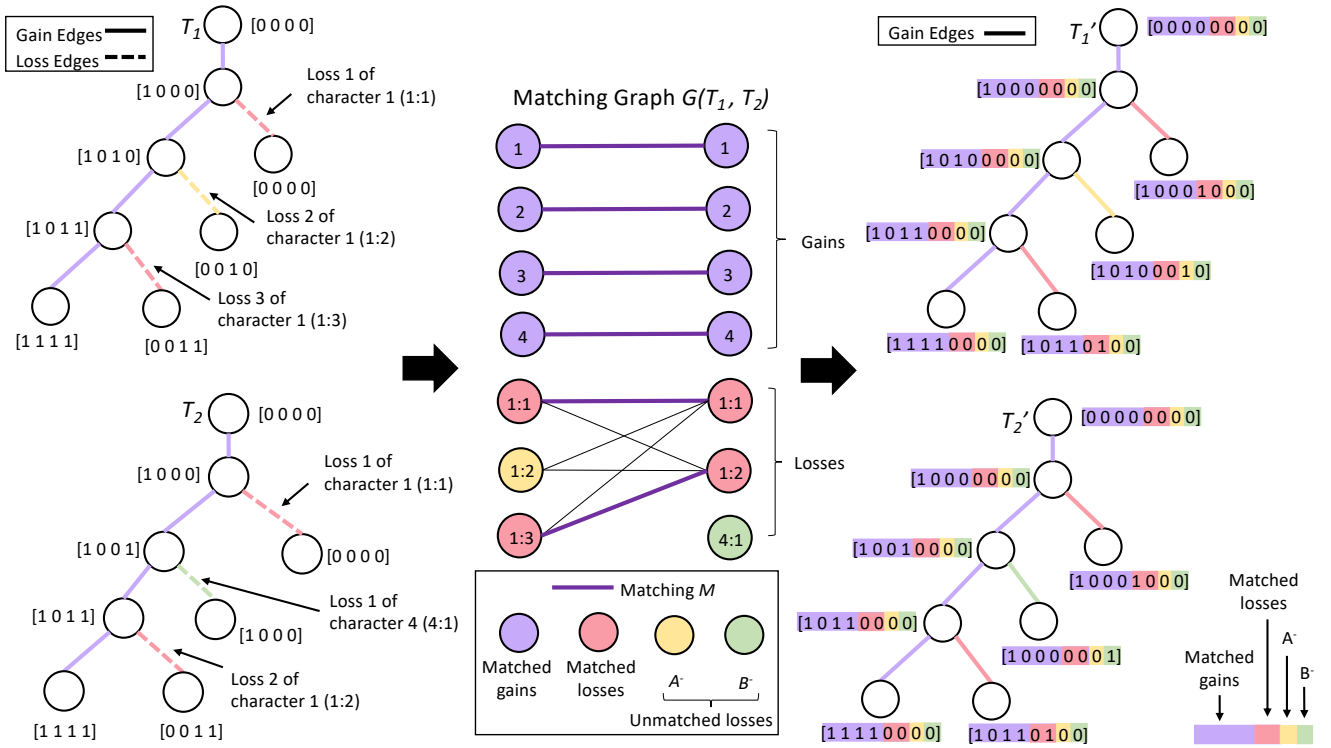

Figure 1: Phylogenies T_1 and T_2 (left) are both 3-Dollo Phylogenies. Loss edges are indicated by a dashed line and gain edges by a solid line. The corresponding matching graph $G(T_1, T_2)$is also shown (center), with node colors that match the respective gain/loss edges in the original phylogenies. The matching indicated by the dark purple edges in G can be used to transform T_1 and T_2 into T_1' and T_2', two 0-Dollo Phylogenies (right) by introducing a new index into their corresponding mutation vectors $\mathbf{b'}_1$ and $\mathbf{b'}_2$ for every time a mutation is lost in either phylogeny. Each loss is now encoded as a gain of the mutation at the newly added index.

SNV/CNA). Losses are encoded by introducing a new mutation index in the mutation vector instead of having the original mutation revert from a 1 to a 0. Figure 1 shows a complete example for two 3-Dollo phylogenies including their matching graph and the resulting 0-Dollo phylogenies for the specified matching.

We can now apply any distance measure *dist* designed for 0-Dollo (ISA) Phylogenies to T_1' and T_2'. Thus, a matching M of $G(T_1, T_2)$, the matching graph for the original phylogenies, induces a distance of $dist(T_1', T_2')$ on the transformed phylogenies. This leads to the following optimization problem.

Problem 2.1. ***Generalized Matching Distance (GMD) Problem:*** *Given tumor phylogenies T_1 and T_2 and a distance measure* dist *under the the ISA, find a maximum cardinality matching M of the matching graph $G(T_1, T_2)$ such that the resulting 0-Dollo phylogenies T_1' and T_2' have minimum distance* $\text{dist}(T_1', T_2')$.

2.3.3 An Exact Algorithm to Solve the GMD. The structure of the matching graph $G(T_1, T_2)$ makes finding the exact solution to the GMD fairly straightforward, although potentially computationally costly. We propose the following method: (1) Create the matching graph G(T_1,T_2); (2) Enumerate all maximum cardinality matchings M in the graph; (3) For each matching, compute the transformed phylogenies T_1' and T_2' and compute dist(T_1', T_2'); and (4) Return the matching M* that produces the smallest such distance. Note that while the number of matchings that need to be checked has a factorial growth rate, in practice we expect the number of matchings to often remain relatively small.

To enumerate all maximum cardinality matchings we first observe that all connected components in $G(T_1, T_2)$ consist of vertices labeled entirely by gains or losses of a single mutation. Therefore, if we can describe how to enumerate all maximum cardinality matchings for a single connected component, then that approach can be generalized for all maximum cardinality matchings across the whole graph by combining matchings for all connected components. Furthermore, each such connected component is always a complete bipartite sub-graph of $G(T_1, T_2)$ (containing all possible edges between the two sets of vertices). Consider the connected sub-graph $S = (A_i \cup B_i, E_i)$ of $G(T_1, T_2)$ for a gain (or loss) of mutation i containing $|A_i|$ vertices from A and $|B_i|$ vertices from B. Without loss of generality, assume that $|A_i| \leq |B_i|$. This connected component is a particular graph often denoted as $K_{|A_i|,|B_i|}$ and whose maximum cardinality matching will be of size $|A_i|$. For example, in Figure 1, the connected component for the losses of mutation 1 consists of 3 vertices from A (labeled 1:1, 1:2, 1:3) and 2 vertices from B (labeled

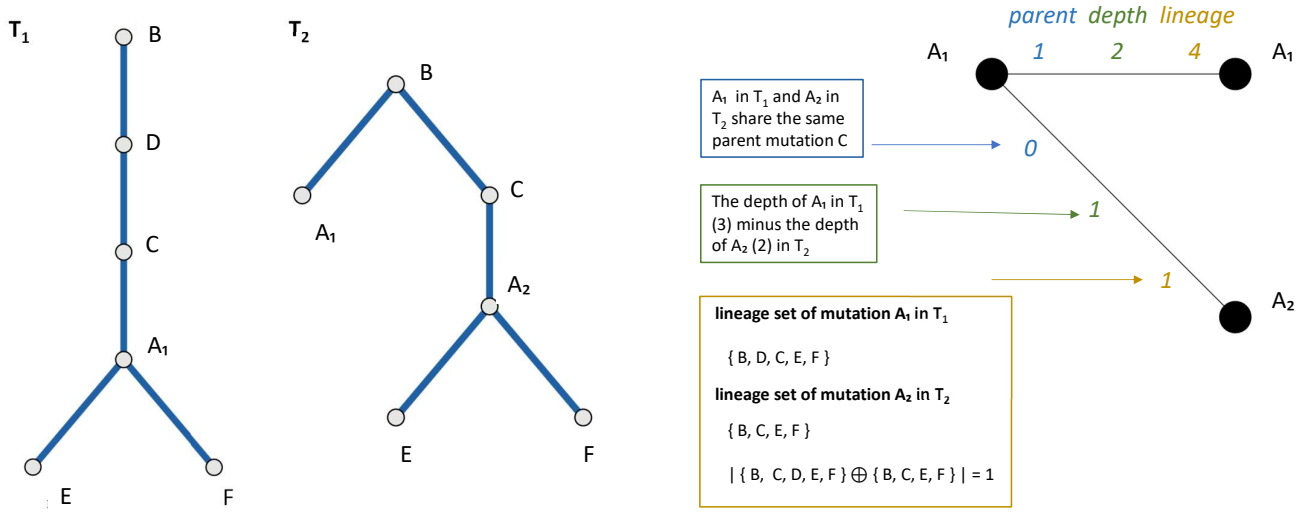

Figure 2: A toy example comparing two phylogenies to demonstrate the weighting schemes used as part of the heuristic approach. (Left) Two mutation phylogenies T_1 and T_2, with T_2 containing parallel mutations of A. (Right) A small portion of the matching graph, specifically the portion of the graph that matches mutation A, is shown. *parent* is shown in blue, *depth* is shown in green, and *lineage* is shown in yellow. Brief descriptions of the calculation of the weighting schemes are attached to the matching graph.

1:1, 1:2) and contains all 6 possible edges between these vertices. The maximum cardinality matching in this component has size 2 and there are 6 such matchings. Algorithms such as [26] exist for enumerating maximum cardinality matchings in bipartite graphs. Furthermore, the fact that S is a complete bipartite graph makes enumerating these matchings an easy two step process: (1) Choose all sets of size $|A_i|$ vertices from the set of vertices B_i ; (2) Consider all possible ways of connecting these vertices to the vertices in A_i.

In cases where the matching graph $G(T_1, T_2)$ has relatively few maximum cardinality matchings (when there are few mutations with relatively small numbers of multiple gains or losses), it can be computationally feasible to simply check all such matchings. We will refer to this as the *enumerative* approach to solving the GMD. Alternatively, we can take a heuristic approach to quickly pick a good max-sized matching, which we describe in the following section.

2.3.4 Minimum-weight matchings: a heuristic approach. In the GMD problem, the matching M on the matching graph $G(T_1, T_2)$ is the component that identifies how to map gains and losses in T_1 to those in T_2. The optimal such mapping will depend on the distance measure *dist* being used. In place of an enumerative, exact solution that may be computationally expensive, we propose a heuristic approach. Specifically, we propose the following procedure: (1) Assign weights $f(a, b)$ to edges (a, b) in $G(T_1, T_2)$ based on features in both T_1 and T_2; (2) Find a minimum-weight, maximum cardinality matching using the Hungarian algorithm [17]; (3) Use this matching to perform the transformation to 0-Dollo phylogenies rather than exhaustively check all possible matchings. The details of the exact distance measure *dist* used may also be helpful for picking an useful weighting scheme for the edges in the matching graph. In particular, we explore the following three weighting schemes.

depth – the weight $f(a, b)$ of an edge (a, b) in $G(T_1, T_2)$ is set equal to the absolute value of the difference between the depth of a in T_1 and b in T_2.

parent – the weight $f(a, b)$ of an edge (a, b) in $G(T_1, T_2)$ is set equal to 0 if a and b share the same parent mutation(s), 1 otherwise.

lineage – the weight $f(a, b)$ of an edge (a, b) in $G(T_1, T_2)$ is set equal to the cardinality of the symmetric difference between the lineage sets of a and b. We define the *lineage set* of a mutation as the union of all of its ancestor mutations and all of its descendant mutations. Figure 2 provides a visual for understanding these weighting schemes.

3 RESULTS

We apply and analyze our proposed approaches on both simulated and real data.

3.1 Results on Simulated Data

On simulated data sets we evaluate several aspects of both the enumerative and heuristic approaches to solving the GMD. Specifically, we evaluate: (1) If brute-force GMD enables ISA distance measures to appropriately penalize differences between phylogenies with parallel mutations; and (2) How well the proposed heuristic approach for the GMD works when applied with different edge weighting schemes and distance measures. This analysis includes both parallel mutations and mutation losses.

3.1.1 Data Simulation. Our general simulation procedure for each of the experiments described below was to use a recursive approach

to enumerate all trees in a specified space (e.g., k-dollo trees with 8 mutations and 3 losses of one of those mutations) and then to randomly sample a subset of these trees to use in each experiment. We note that such trees are inherently biologically feasible as they do not allow a mutation to be lost before it is gained. The specific details for simulating data for each experiment are outlined in the corresponding sections below.

3.1.2 Examining the Effects of Parallel Mutations. Similar to Ciccolella et al. [5], we wanted to investigate the effect on the distance evaluation between clonal phylogenies as the number of parallel mutations increases. We would expect that a phylogeny with more occurrences of a mutation when compared with a phylogeny with less occurrences of that same mutation should be deemed further away from each other than two phylogenies with more similar numbers of mutation occurrences. In order to set up an experiment to support this investigation we first create 4 data sets containing all possible mutation phylogenies with 8 mutations and 1, 2, 3, or 4 gains of mutation 'A'. Note, these are a specific subset of k-Camin-Sokal phylogenies. We then created a base set of trees (called Group 1) by randomly sampling 50 trees from the data set containing 1 gain of mutation 'A'. We also created 4 test sets (Groups 2-5) by then randomly sampling 50 trees from the sets with 1-4 gains of mutation 'A'. For each group of trees we converted them from mutation phylogenies to clonal phylogenies by randomly selecting 2 pairs of connected nodes in the mutation phylogeny to collapse to simulate mutations whose order cannot be ascertained. We do this collapsing since real data is much more likely to be a clonal phylogeny with gains or losses of different mutations grouped on single vertices rather than the idealized mutation phylogeny where each new gain or loss appears on its own vertex. We then conducted pairwise comparisons between clonal phylogenies in Group 1 with Groups 2-5. Figure 3 shows the average distance between each test set and the base set for various different distance measures.

All methods that included the enumerative GMD transformation plus an ISA dependent distance measure (CASet [6], DISC [6] and MLTED [14]) show the desired property of monotonically increasing distances as the number of parallel mutations of mutation 'A' increased. The MP3 method [5], which was designed to handle parallel mutations, also shows the same monotonically increasing property. While neither CASet nor DISC was designed to handle parallel mutations, neither program throws an error when run with such data (without the GMD transformation). Specifically, the implementation of these methods utilize set data-structures rather than multisets for mutations and thus only one mutation gain is considered whenever parallel mutations are present within the phylogeny. This explains the relatively flat slope for these results. However, we intentionally include these results here to better demonstrate the exact impact of our approach. We also note that when phylogenies being compared do not have identical sets of mutations, both CASet and DISC have the option to either use the union or intersection of those sets. We only include here results using the union option because the intersection option effectively removes all signals from the parallel mutations as only a single gain of that mutation can be included in the intersection set. As a result, CASet and DISC when applied without GMD behave almost identically for both union and intersection. MLTED, without GMD, returns an error on phylogenies with parallel mutations.

When Ciccolella et al. [5] proposed MP3 (and performed an experiment similar to this one), they suggested that it is better to have a steeper curve when comparing phlogenies with differing numbers of occurrences of a mutation. Following this, CASet (union) + GMD , DISC (union) + GMD, and MLTED + GMD all have a steeper curve than MP3 and therefore penalize differences across mutation sets more than MP3. Among these distance measures, MLTED + GMD has the steepest curve with a total change in distance of 0.15.

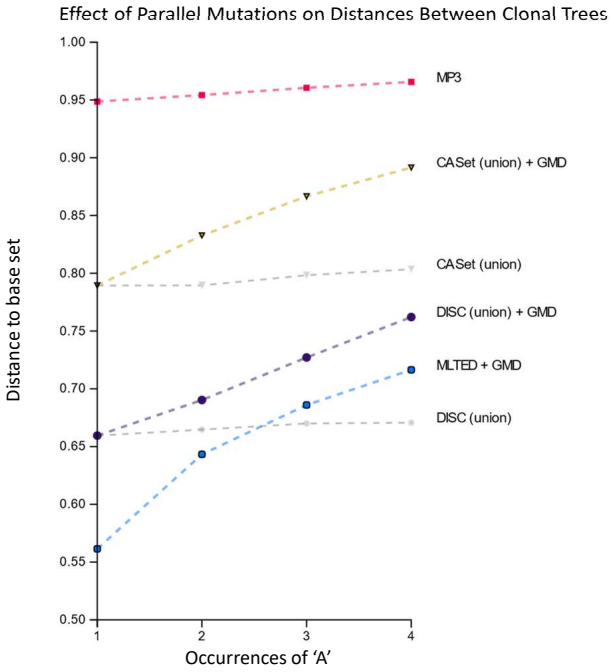

Figure 3: The results of an experiment to evaluate how various distance measures with and without enumerative GMD vary in their evaluation of clonal phylogenies. Specifically, the evaluated phylogenies are a subset of k-Camin-Sokal phylogenies where the only mutation duplicated is mutation 'A' and k is 1, 2, 3, and 4. We evaluate distances using MP3, CASet (union) + GMD, CASet (union), DISC (union) + GMD, MLTED + GMD, and DISC (union).

3.1.3 Performance of Heuristic Approach. We also analyzed the performance of our proposed heuristic approach in approximating solutions to the GMD Problem. Specifically, we explored which weighting schemes (*depth*, *lineage*, *parent*) paired best with different ISA distance measures (parent-child (PC) [10], ancestor-descendant (AD) [10], MLTED [14], CASet [6], DISC [6]). We first create a data set containing all possible mutation phylogenies with 5 mutations and 2 losses of mutation 'A'. Note, these are a specific subset of k-Dollo phylogenies. We then randomly sampled 100 mutation phylogenies from the data set and ran pairwise comparisons between these 100 mutation phylogenies using all combinations of weighting schemes (for our heuristic approach) and different ISA distance

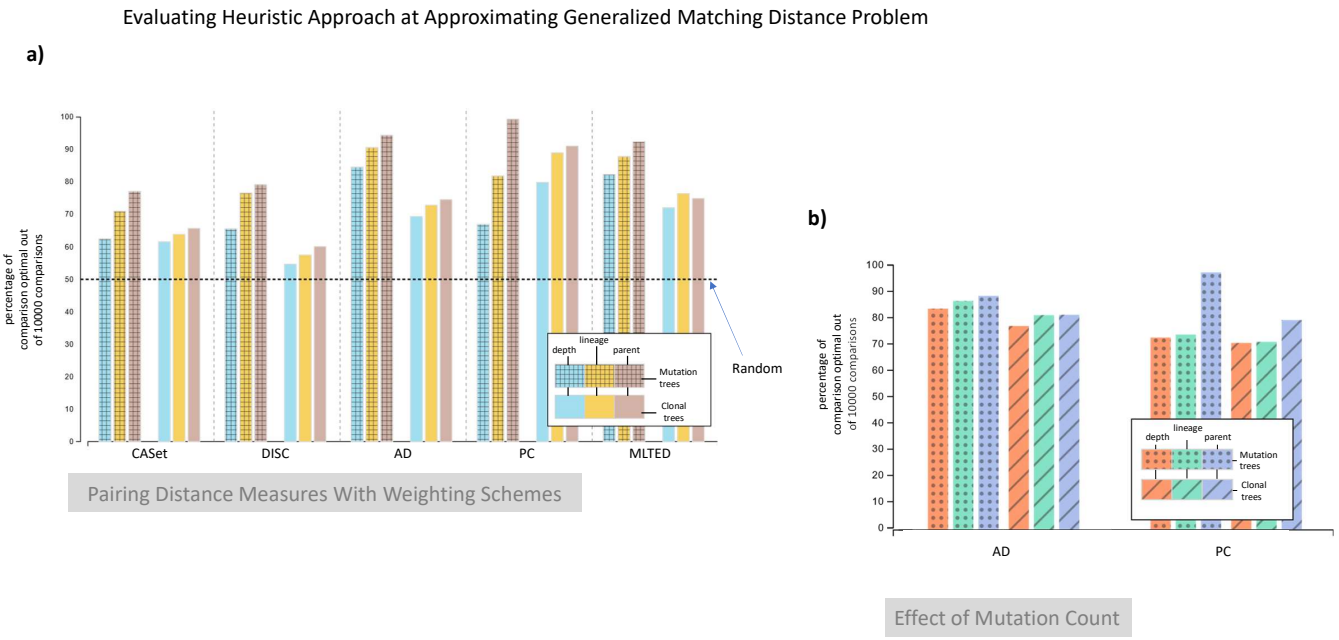

Figure 4: The results of experiments to evaluate various factors that could effect the performance of our heuristic approach at approximating solutions to the GMD Problem. We paired various ISA distance measures (CASet, DISC, PC, AD, and MLTED) with our proposed weighting schemes (*depth*, *lineage*, *parent*). a) Fraction of trials where the heuristic approach achieved the optimal solution on phylogenies with 5 mutations and 2 losses of mutation 'A'. Results are also shown for both mutation phylogenies and clonal phylogenies that resulted from collapsing nodes in the mutation phylogenies. Random indicates the probability of selecting an optimal matching randomly from all possible matchings. b) Fraction of trials where the heuristic approach achieved the optimal solution for a data set with 8 mutations and 2 gains of mutation 'A'. Results are shown for both mutation phylogenies and clonal phylogenies obtained by collapsing nodes in the mutation phylogenies.

measures. We then compare these results to using the enumerative approach for solving the GMD problem for all distance measures. This allows us to capture what fraction of trials optimally solve the GMD problem for each combination of weighting scheme and ISA distance measure. We also perform this same experiment after using the same collapsing approach described in the previous section to turn all mutation phylogenies into clonal phylogenies. Figure 4a shows the complete results from this experiment. We note we also performed this same experiment but with phylogenies with 8 mutations and 2 losses of a single mutation and found the results to be similar.

Across all distance measures and heuristic weighting schemes there is a decrease in performance when applied to clonal phylogenies in comparison to application to the corresponding mutation phylogenies. This could be because the clonal phylogenies were shorter, providing less information to compute the weighting schemes-especially in the case of the *lineage* weighting scheme. For mutation phylogenies, the heuristic approach is almost always optimal when using PC distance with the*parent* weighting scheme. Specifically, for 5 mutations and 2 losses of 'A', it finds the optimal solution 99.32% of the time. AD distance with the *parent* weighting scheme and MLTED distance with the *parent* weighting scheme also perform well, and identify an optimal solution in 94.30% and 92.40% of trials. While intuitively simple, the *parent* weighting scheme performed the best across all the distance measures for both clonal and mutation phylogenies except for MLTED in which *lineage* achieved better performance for clonal phylogenies. All pairings of ISA distance measures and weighting schemes performed better than randomly selecting any maximum cardinality matching.

We also performed a similar analysis for phylogenies with 5,6,7 and 8 mutations and 2 gains of mutation 'A' (a subset of k-Camin-Sokal phylogenies). In this case we used only the distance measures that performed the best in the previous experiment, parent-child and ancestor descendant. We saw little difference in the results for the different number of unique mutations in each tree, so Figure 4b shows only the results for 8 mutations with 2 gains of mutation 'A'. Despite the fact that the number of unique mutations was increased and that the phylogenies in this experiment contained multiple gains rather than multiple losses, many of the patterns we saw in the previous experiment persist. PC distance and the *parent* weighting scheme still perform the best with the combination of these obtaining an optimal solution in 98.10% of trials on mutation phylogenies and 79.89% of trials on clonal phylogenies. Similar to the previous experiment, we also see a performance decrease between clonal phylogenies and mutation phylogenies.

3.2 Results on Real Data

We also compare three phylogenies that were inferred for a colorectal cancer patient CRC2 from Leung et al. [19]. The original data set consists of targeted single-cell sequencing of 182 cells from a primary colon tumor and a liver metastasis and a 1000-cancer gene panel used as the target region for sequencing. The phylogenies were inferred by three different methods in three separate papers–SCARLET [24], SiCloneFit [27], and FiMO [1] (a pre-print paper). The phylogenies and computed distances for CASet (union) + GMD, DISC (union) + GMD, MLTED + GMD and MP3 (as these were the only distance measures to accurately penalize increasing differences in mutation occurrences in the simulated experiments) are shown in Figure 5. We ran the enumerative variant of GMD as the number of gains/losses were small, and it finished virtually instantaneously. Specifically, we ran the code on a Dell Poweredge R540 server with 28 cores and 384 GB of RAM. Note, the SCARLET tree is the only one containing back mutations (losses) whereas the SiCloneFit and FiMO phylogenies contain parallel mutations. Furthermore, the SCARLET tree explicitly orders many mutations that the other methods simply group together. Thus, we may reasonably expect the SCARLET tree to be more dissimilar to the other two phylogenies for most distances. The results of CASet (union) + GMD, DISC (union) + GMD, MLTED + GMD and MP3 agree with that assumption. We note that the MP3 analysis is more extreme in both the similarity of the SiCloneFit and FiMO phylogenies and the dissimilarity of the SCARLET tree to the others. This resembles the situation in Figure 3 with MP3 yielding a larger dissimiliarity in its evaluation than the other distance measures. On the other hand, while MLTED + GMD still evaluates the SiCloneFit and FiMO trees as most similar, it does not identify the SCARLET tree to be as dissimilar from them as the other methods do. This is consistent with the intended behavior of MLTED, which is different than the other distance measures. Specifically, MLTED was designed to evaluate phylogenies at different resolutions that could represent the same underlying tumor evolutionary history as similar. Thus, the expanded nature of many mutations in the SCARLET tree should contribute less to the total distance to the other two phylogenies when using the MLTED distance, which is exactly what we see. Thus, our results here suggest the ability of the GMD approach to maintain the desired properties of the distance measures that it is used with.

4 CONCLUSION AND FUTURE WORK

There are many existing distance measures to compare tumor phylogenies that abide by the Infinite Sites Assumption (ISA). However, the field of tumor phylogenomics is gradually transitioning to models of tumor evolution beyond the ISA such as the k-Dollo and Camin-Sokal models in order to better represent the realities of tumor evolution. In order to leverage already existing ISA tumor distance measures to evaluate tumor phylogenies inferred under these more relaxed models, we propose the Generalized Matching Distance (GMD) Problem. We both provide an enumerative approach to solving GMD (which is often very practical to use), and also propose a heuristic to solve the GMD that utilizes various weighting schemes (*depth*, *lineage*, *parent*) to identify a single matching that is likely to produce a good result. We have shown that without GMD, some existing ISA distance measures are incapable of correctly penalizing differences in occurrences of parallel mutations; in some cases, these distance measures are even unusable without first using GMD. We also showed that the heuristic approach we proposed performs well in some restricted cases of tumor phylogenies; parent-child distance combined with the *parent* weighting schemes even generally produces an optimal solution to the GMD in the case of small mutation phylogenies. Finally, we applied our GMD approach along with several ISA distance measures to a real colorectal cancer dataset. We found that our approach allowed the distance measures to retain their originally designed properties such as the ability of the MLTED distance measure to consider phylogenies at different resolutions as similar if they could represent the same evolutionary history.

Yet, there is much future work that could be done. For one, more weighting schemes can be developed in addition to the three that were provided in this paper (*depth*, *lineage*, *parent*). Specifically, while *parent* maps quite well to parent-child distance and some sort of intuition guides the pairing of *lineage* and ancestor-descendant distance, we haven't extensively explored weights that would work especially well for MLTED, which relies on edit distance rather than sets of mutations. In addition, experiments on larger phylogenies containing greater complexity like the gain and loss of multiple different mutations could provide more comprehensive information on the optimality of the heuristic approach. Finally, more extensive comparison between all the distance measures and their GMD extensions could help researchers in the field better determine which distance measures to use for their use case.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation (NSF) award CAREER-IIS-2046011 and a Large Faculty Development Endowment grant from Carleton College.

For useful conversations in coming up with the initial ideas for this project, we'd like to thank Mohammed El-Kebir. We would also like to thank Conor Babcock O'Neill, Cecilia Ehrlichman, Matthew Smith-Erb, and Cathy Guang for their support in refining the code. Finally, we thank Anwesha Mukherji and Jayti Arora for their feedback as it pertains to polishing the visualizations found in the results.

REFERENCES

[1] Avesh Kumar Agrawal and Hamim Zafar. 2022. FiMO: Inferring the Temporal Order of Mutations on Clonal Phylogeny under Finite-sites Models. *bioRxiv* (2022), 2022–01.
[2] Nabil Amirouchene-Angelozzi, Charles Swanton, and Alberto Bardelli. 2017. Tumor Evolution as a Therapeutic TargetThe Impact of Tumor Evolution in Precision Medicine. *Cancer discovery* 7, 8 (2017), 805–817.
[3] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. 2015. Cancer evolution: mathematical models and computational inference. *Systematic biology* 64, 1 (2015), e1–e25.
[4] Joseph H Camin and Robert R Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution* (1965), 311–326.
[5] Simone Ciccolella, Giulia Bernardini, Luca Denti, Paola Bonizzoni, Marco Previtali, and Gianluca Della Vedova. 2021. Triplet-based similarity score for fully multilabeled trees with poly-occurring labels. *Bioinformatics* 37, 2 (2021), 178–184.
[6] Zach DiNardo, Kiran Tomlinson, Anna Ritz, and Layla Oesper. 2020. Distance measures for tumor evolutionary trees. *Bioinformatics* 36, 7 (2020), 2090–2097.
[7] Louis Dollo. 1893. The laws of evolution. *Bull. Soc. Bel. Geol. Paleontol* 7 (1893), 164–166.

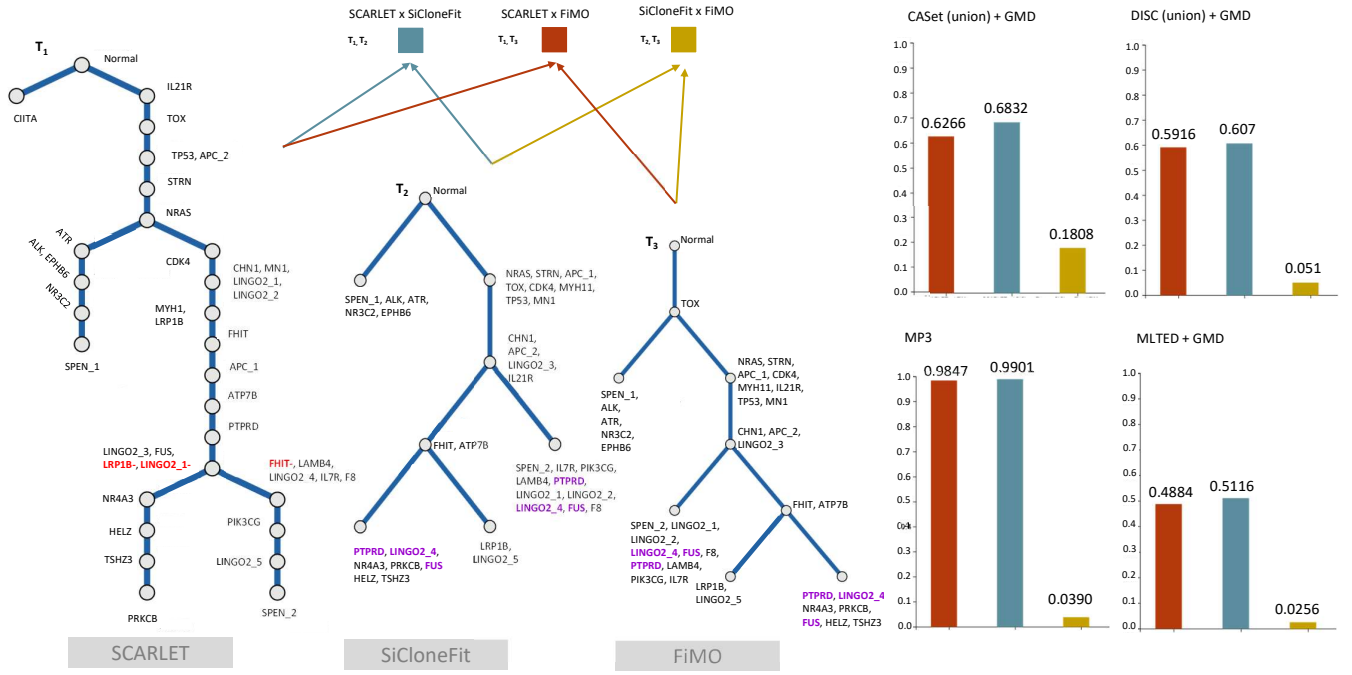

Figure 5: Pairwise comparisons between three phylogenies representing a patient with colorectal cancer. (Left) T_1 is the tree inferred by SCARLET, containing losses (indicated in red). T_2 is the tree inferred by SiCloneFit, containing parallel mutations (indicated in purple). T_3 was inferred by FiMO, also containing parallel mutations. (Right) Pairwise distance comparisons between the phylogenies on the left using CASet (union) + GMD, DISC(union) + GMD, MLTED + GMD, and MP3.

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

[9] Kiya Govek, Camden Sikes, and Layla Oesper. 2018. A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 Acm international conference on bioinformatics, computational biology, and health informatics*. 63–72.

[10] Kiya Govek, Camden Sikes, Yangqiaoyu Zhou, and Layla Oesper. 2020. Graphyc: Using consensus to infer tumor evolution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 1 (2020), 465–478.

[11] Philippe Gui and Trever G Bivona. 2022. Evolution of metastasis: New tools and insights. *Trends in Cancer* 8, 2 (2022), 98–109.

[12] Katharina Jahn, Niko Beerenwinkel, and Louxin Zhang. 2021. The Bourque distances for mutation trees of cancers. *Algorithms for Molecular Biology* 16, 1 (2021), 1–15.

[13] Nikolai Karpov, Salem Malikic, Md Rahman, S Cenk Sahinalp, et al. 2018. A multi-labeled tree edit distance for comparing" clonal trees" of tumor progression. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[14] Nikolai Karpov, Salem Malikic, Md Rahman, S Cenk Sahinalp, et al. 2019. A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression. *Algorithms for Molecular Biology* 14, 1 (2019), 1–18.

[15] Tom L Kaufmann, Marina Petkovic, Thomas BK Watkins, Emma C Colliver, Sofya Laskina, Nisha Thapa, Darlan C Minussi, Nicholas Navin, Charles Swanton, Peter Van Loo, et al. 2022. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome biology* 23, 1 (2022), 241.

[16] Motoo Kimura. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 4 (1969), 893.

[17] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

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

[19] Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E Navin. 2017. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research* 27, 8 (2017), 1287–1299.

[20] Nicholas McGranahan and Charles Swanton. 2017. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 168, 4 (2017), 613–628.

[21] Peter C Nowell. 1976. The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science* 194, 4260 (1976), 23–28.

[22] Katherine L Pogrebniak and Christina Curtis. 2018. Harnessing tumor evolution to circumvent resistance. *Trends in Genetics* 34, 8 (2018), 639–651.

[23] David F Robinson and Leslie R Foulds. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53, 1-2 (1981), 131–147.

[24] Gryte Satas, Simone Zaccaria, Geoffrey Mon, and Benjamin J Raphael. 2020. SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell systems* 10, 4 (2020), 323–332.

[25] Russell Schwartz and Alejandro A Schäffer. 2017. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* 18, 4 (2017), 213–229.

[26] Takeaki Uno. 1997. Algorithms for enumerating all perfect, maximum and maximal matchings in bipartite graphs. In *Algorithms and Computation: 8th International Symposium, ISAAC'97 Singapore, December 17–19, 1997 Proceedings 8*. Springer, 92–101.

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

[28] Hamim Zafar, Nicholas Navin, Luay Nakhleh, and Ken Chen. 2018. Computational approaches for inferring tumor evolution from single-cell genomic data. *Current Opinion in Systems Biology* 7 (2018), 16–25.