Bioinformatics, 36(7), 2020, 2090–2097
doi: 10.1093/bioinformatics/btz869
Advance Access Publication Date: 21 November 2019
Original Paper

OXFORD

Phylogenetics

Distance measures for tumor evolutionary trees

Zach DiNardo[1,†], Kiran Tomlinson[1,2,†], Anna Ritz[3] and Layla Oesper[1,*]

[1]Department of Computer Science, Carleton College, Northfield, MN 55057, USA, [2]Department of Computer Science, Cornell University, Ithaca, NY 14853, USA and [3]Department of Biology, Reed College, Portland, OR 97202, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Yann Ponty

Received on May 21, 2019; revised on September 4, 2019; editorial decision on November 15, 2019; accepted on November 19, 2019

Abstract

Motivation: There has been recent increased interest in using algorithmic methods to infer the evolutionary tree underlying the developmental history of a tumor. Quantitative measures that compare such trees are vital to a number of different applications including benchmarking tree inference methods and evaluating common inheritance patterns across patients. However, few appropriate distance measures exist, and those that do have low resolution for differentiating trees or do not fully account for the complex relationship between tree topology and the inheritance of the mutations labeling that topology.

Results: Here, we present two novel distance measures, ***C***ommon ***A***ncestor ***Set*** distance (CASet) and ***D***istinctly ***I***nherited ***S***et ***C***omparison distance (DISC), that are specifically designed to account for the subclonal mutation inheritance patterns characteristic of tumor evolutionary trees. We apply CASet and DISC to multiple simulated datasets and two breast cancer datasets and show that our distance measures allow for more nuanced and accurate delineation between tumor evolutionary trees than existing distance measures.

Availability and implementation: Implementations of CASet and DISC are freely available at: https://bitbucket.org/oesperlab/stereodist.

Contact: loesper@carleton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The clonal theory of cancer (Nowell, 1976) states that a tumor is the result of an evolutionary process. The history of somatic mutation acquisition during this process is often represented using a rooted tree structure. The root represents the *founding cell* and every other vertex represents a distinct *clone*, or population of tumor cells sharing a set of mutations, that existed at some point during the tumor's evolution. Directed edges represent direct ancestral relationships between populations. Vertices, or sometimes edges incident to those vertices, are labeled with the set of mutations that first appeared in that clone. In the most general case, these trees are referred to as *clonal* trees (El-Kebir *et al.*, 2015). When exactly one mutation labels each vertex, they are instead called *mutation* trees (Kim and Simon, 2014). Mutation trees represent the highest resolution possible for viewing the complete temporal history of the tumor, barring trees representing ancestry at the cellular level.

In recent years, a number of methods have been developed to infer either clonal or mutation trees from single nucleotide variants in bulk sequencing data (El-Kebir *et al.*, 2015; Husić *et al.*, 2018; Jiao *et al.*, 2014; Malikic *et al.*, 2015; Niknafs *et al.*, 2015; Popic *et al.*, 2015; Satas and Raphael, 2017; Toosi *et al.*, 2017) and single cell sequencing data (El-Kebir, 2018; Jahn *et al.*, 2016; Miura *et al.*, 2018; Ross and Markowetz, 2016; Zafar *et al.*, 2017). See Schwartz and Schäffer (2017) for a more complete listing of such methods. The goal of tree inference is to gain a better understanding of tumor development. This understanding may reveal insights about the mutations that drive a tumor's growth (Jolly and Van Loo, 2018; Raphael *et al.*, 2014) and may be targeted for patient treatment (Amirouchene-Angelozzi *et al.*, 2017; Blakely, 2017).

Due to the ongoing development of tumor evolution inference methods, the similarity of two potential tumor histories often needs to be quantified. First, new methods need to be benchmarked against other methods or against a ground truth tree, and the ad hoc measures that have typically been used in these situations (El-Kebir *et al.*, 2015; Malikic *et al.*, 2015; Popic *et al.*, 2015) have not been rigorously studied. Second, some methods themselves rely on the use of a distance measure when inferring a tumor's evolutionary history. For example, the GraPhyC method (Govek *et al.*, 2018) uses a distance measure to create a consensus tumor history from several input histories. Third, some Bayesian methods (Jahn *et al.*, 2016; Jiang *et al.*, 2016; Ross and Markowetz, 2016) attempt to account for uncertainty in the inferred tree structure either through sampling procedures or inference of unobserved clones. These methods could benefit from analysis of the similarity of the trees considered. Lastly, there have been growing questions about the structure of the space

of possible evolutionary histories consistent with the underlying sequence data (El-Kebir *et al.*, 2016; Pradhan and El-Kebir, 2018; Tomlinson and Oesper, 2018) and how tumor evolutionary histories across patients can be used to identify patterns of tumor evolution (Matsui *et al.*, 2017). Further analysis of these questions would be aided by distance measures tuned to the key features of tumor evolution histories.

For species trees, there are several well-known distance measures such as the method of Robinson and Foulds (1981) and triplet distance (Critchlow *et al.*, 1996), among others. However, there are substantial differences between species and cancer trees. One such difference is that the labels in tumor evolutionary trees represent mutations rather than species. Moreover, all vertices in these trees are labeled with mutations, in contrast to phylogenetic trees where only the leaves are labeled with species. So while methods designed to compare species trees may assume that both trees have identical sets of leaf labels, the same is not true for tumor evolutionary trees.

Various distance measures also exist for labeled trees where each node in the tree has a single label from a finite set (for a review of such methods, see Bille, 2005). However, nodes in tumor evolutionary trees may contain multiple labels indicating mutations whose order of appearance cannot be readily identified (Govek *et al.*, 2018; Karpov *et al.*, 2018). Furthermore, the mutations in a tumor evolutionary tree are inherited by all descendant tumor populations, creating a complex underlying relationship between topology and mutation labeling. Distance measures for labeled trees cannot handle multiple labels, nor do they consider this complex pattern of mutation inheritance. Thus, there is a need for distance measures specifically tailored to the intricacies of tumor evolutionary trees.

Despite this need for tumor tree distance measures, a limited number of such measures have been rigorously developed and comprehensively evaluated. Several simple distance measures on clonal trees described by Govek *et al.* (2018) generalized earlier ad hoc approaches that relied on the existence of a ground truth tree (El-Kebir *et al.*, 2015; Malikic *et al.*, 2015; Popic *et al.*, 2015), but were not the focus of that work, and their effectiveness was not analyzed in depth. Additionally, each of these distance measures focused on a single aspect of similarity between trees, but failed to look at how these aspects affected the global structure of the constituent trees. Another recently proposed approach called MLTED uses an edit distance-based measure focused on handling multi-labeled nodes within clonal trees to count the minimum number of moves to convert both trees into a specific common tree (Karpov *et al.*, 2018). This distance allows trees observed at different levels of resolution to be considered identical (e.g. a clonal tree and its expanded mutation tree). While this may be desirable when comparing trees reconstructed using different resolution datasets, it may not work well for benchmarking new algorithms. In this use case, a distance measure is used to compare trees inferred by different methods against the true underlying tree. A method that is able to resolve more specific ancestral relationships should be considered a closer match to the underlying tree than one that simply groups descendants together without specifying an order.

In this work, we formalize the definition of a *tumor evolution distance measure* by precisely defining the input and output of such a function and describing what features of such a measure are desirable in the context of tumor evolution. We then describe two novel tumor evolution distance measures, Common *A*ncestor *Set* distance (CASet) and *D*istinctly *I*nherited *S*et Comparison distance (DISC), which are specifically designed to account for structure of mutation inheritance by subsequent tumor populations. After defining these measures in a simple case, we extend them to trees that do not share the same set of mutations. We apply our distance measures to multiple simulated datasets and two breast cancer datasets. We find that CASet and DISC allow for more precision and are better able to identify groups of similar trees in a clustering scenario than existing distance measures. In testing this application, we find that CASet performs especially well in determining the clustering of sets of dissimilar trees while DISC is able to distinguish between relatively similar trees with very high granularity.

2 Materials and methods

2.1 Tumor evolutionary trees

We first make two common assumptions about tumor evolution. The first is the *infinite sites assumption* (ISA) which states that no mutation occurs more than once during a tumor's history, and that once gained, a mutation is never lost. This has been a common assumption made by many methods that infer tumor evolutionary histories (El-Kebir *et al.*, 2015; Husić *et al.*, 2018, among others; Jiao *et al.*, 2014; Malikic *et al.*, 2015; Popic *et al.*, 2015). While some recent phylogeny inference methods do allow for minor violations of the ISA (e.g. Bonizzoni *et al.*, 2018; Marass *et al.*, 2016), our work here will be most widely applicable if we assume the ISA. The second assumption is that all tumor cells are descended from a single founding tumor cell, and hence the tumor's evolution can be described as *monoclonal*. This assumption is non-essential to our approaches and can easily be dropped by rooting evolution trees with healthy cells instead of founding tumor cells. Nonetheless, we make the monoclonal assumption to simplify our definitions. We now formally describe the evolutionary history of a monoclonal tumor adhering to the ISA as a clonal tree.

A *clonal tree* is a rooted, directed tree T in which: (i) each vertex in the tree is labeled by one or more mutations and (ii) no mutation appears more than once. Here, mutation may mean any type of genomic variant (e.g. SNV, CNA etc.). We also define $\mathcal{T}$ to be the set of all clonal trees. Given a tree $T \in \mathcal{T}$, we define $M(T)$ to be the set of all mutations (i.e. vertex labels) in T. In this representation, every vertex represents a distinct tumor clone (or population) that existed at some point during the tumor's evolution. The mutation labels indicate the clone in which the mutation first appeared. Thus, the complete set of mutations that exist in any particular clone, represented by vertex v, is the set of mutations that label all vertices on the path from the root to vertex v.

We sometimes wish to restrict our attention to a predetermined set of mutations, so we also define m-clonal trees for this purpose. An *m-clonal tree* is a clonal tree T with $M(T) = \{1, \ldots, m\}$. We emphasize that this definition uses the variable m to refer to the number of mutations rather than the number of clones in the tree. We also define $\mathcal{T}_m$ to be the set of all m-clonal trees that share the same mutation set $[m] = \{1, 2, \ldots, m\}$.

2.2 Tumor evolution distances

A *tumor evolution distance measure* is a function $d : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}^{\geq 0}$ for which a value of $d(\cdot, \cdot)$ that is close to 0 indicates that the two input trees are very similar and progressively larger values of $d(\cdot, \cdot)$ indicate the clonal trees are more dissimilar. A tumor evolution distance measure must give a quantitative evaluation of how different two tumor histories are from each other, but how best to define 'different' is not immediately obvious. There are two main aspects of tumor evolutionary trees that should contribute to a distance measure: (i) the topology of the tree and (ii) the labels present in the vertices of the trees. Topology can be separated from labeling by simply ignoring all labels, and the labels can be separated from topology by considering only the set(s) of labels that appear or appear together. Thus, simple distance measures could certainly consider each of these aspects separately. However, these attributes are inherently intertwined.

Since mutation labelings indicate in which vertex a mutation was first acquired, all descendants of that vertex also inherit that mutation. A difference in a vertex with many descendants should then contribute more to a distance measure than one in a vertex with few descendants, since it affects more clonal populations. Thus, a tumor evolution distance measure that simply counts the differences between trees (often referred to as a tree edit distance) does not fully address the impact any given label change may have. A distance measure should assign different weights to disagreements in different locations in order to appropriately address the relationship between topology and mutation labeling. These observations form the basis for the distance measures presented in the following section.

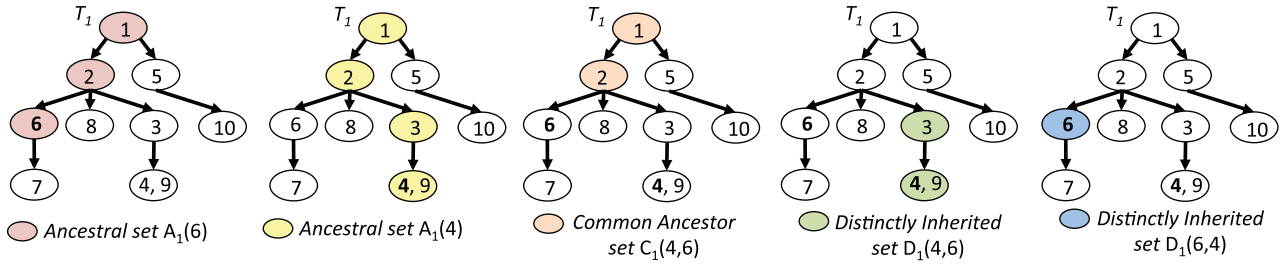

Fig. 1. Examples of ancestral sets, common ancestor sets and distinctly inherited sets on one tree

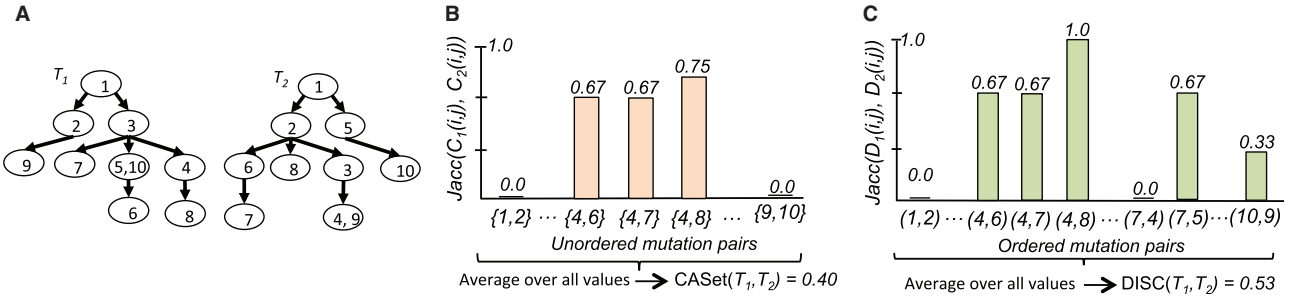

Fig. 2. (**A**) A pair of 10-clonal trees. (**B**) Example of CASet distance applied to the 10-clonal trees in (A). (**C**) Example of DISC distance applied to the 10-clonal trees in (A)

2.3 Two new distance measures on *m*-clonal trees

2.3.1 Notation

Suppose $T \in \mathcal{T}$ is a clonal tree. Given a mutation $i \in M(T)$, we denote the vertex in T that is labeled with mutation i as v_i. We also define the *ancestral set* $A(i)$ as the set of mutations that label the path from the root vertex r to v_i in T. Thus, $A(i)$ gives the set of mutations that exist in the clone represented by vertex v_i. Given $i, j \in M(T)$, we define the *common ancestor set* $C(i, j)$ to be $A(i) \cap A(j)$. That is, $C(i, j)$ is the set of mutations that are ancestral to both mutations i and j. Given $i, j \in M(T_k)$, we also define the *distinctly inherited set* $D(i, j)$ to be $A(i) \setminus A(j)$. That is, $D(i, j)$ is the set of mutations that are ancestral to mutation i but not mutation j in T. Note that under this definition it is almost always the case that $D(i, j) \neq D(j, i)$, although both are empty when $i = j$. When we have more than one tree, we use subscripts to distinguish between them. For instance, $A_k(i), C_k(i, j)$ and $D_k(i, j)$ all refer to T_k. See Figure 1, e.g. of ancestral sets, common ancestor sets and distinctly inherited sets. Given two sets of mutations A and B, we note that the *Jaccard distance* between them is defined as $\text{Jacc}(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$ and $\text{Jacc}(\varnothing, \varnothing) = 0$.

2.3.2 Common ancestor set distance

Given two m-clonal trees $T_k, T_\ell \in \mathcal{T}_m$, we define a new tumor evolutionary tree distance measure called CASet that computes a distance between T_k and T_ℓ (Fig. 2). Informally, CASet distance is the average Jaccard distance between all corresponding common ancestor sets in T_k and T_ℓ. Formally,

$$\text{CASet}(T_k, T_\ell) = \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i, j), C_\ell(i, j)). \quad (1)$$

Observation 2.1 *The running time to compute* CASet *is* $O(m^2)$.

Observation 2.2 *CASet distance is a metric on* $\mathcal{T}_m$.

Full proofs of Observations 2.1 and 2.2 are in Supplementary Appendix.

2.3.3 Distinctly inherited set comparison

CASet distance compares the common ancestor sets of all pairs of mutations, which emphasizes differences close to the root. However, we might also want to emphasize differences in more recently acquired mutations. Given two m-clonal trees $T_k, T_\ell \in \mathcal{T}_m$, we define a new tumor evolution distance measure between T_k and T_ℓ called DISC that accounts for mutation differences in the more recent tumor clones (Fig. 2). Informally, DISC distance is the average Jaccard distance between all corresponding distinctly inherited ancestor sets in T_k and T_ℓ. Formally,

$$\text{DISC}(T_k, T_\ell) = \frac{1}{m(m-1)} \sum_{\substack{(i,j) \in [m]^2 \\ i \neq j}} \text{Jacc}(D_k(i, j), D_\ell(i, j)). \quad (2)$$

Note that this range of summation is different from CASet, which only considers unordered pairs $\{i, j\}$.

Observation 2.3 *The running time to compute* DISC *is* $O(m^3)$.

Observation 2.4 *DISC distance is a metric on* $\mathcal{T}_m$.

Full proofs of Observations 2.3 and 2.4 are similar to the CASet proofs and also appear in the Supplementary Appendix. The difference in runtime for CASet and DISC is related to the fact that there are fewer than m distinct common ancestor sets in an m-clonal tree, but the same is not true for distinctly inherited sets.

2.4 Extending CASet and DISC to clonal trees

Thus far, we have assumed that any two tumor evolutionary trees to be compared have the exact same set of mutation labels (i.e. that they are both m-clonal). However, there are many scenarios in which this may not be the case. For instance, some methods such as El-Kebir *et al.* (2015) may not use all available mutations when creating a tumor evolutionary history tree. Furthermore, different trees may be reconstructed from different data types for the same tumor (e.g. single cell and bulk sequencing) that do not share the same set of observed mutations. In this section, we present two extensions to both of our distance measures that allow for the comparison of clonal trees with different sets of mutation labels.

2.4.1 Intersection of mutation sets

In the first extension to clonal trees, we consider the intersection of the mutation sets for the input trees. This allows us to compute a distance between two trees by only considering pairs of mutations that the two trees share. Let $I_{k,\ell} = M(T_k) \cap M(T_\ell)$ be the intersection

of the sets of mutations labeling T_k and T_ℓ. Thus, we can modify both CASet and DISC distances as follows:

$$\text{CASet}_\cap(T_k, T_\ell) = \frac{1}{\binom{|I_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq I_{k,\ell}} \text{Jacc}\left(C_k(i,j), C_\ell(i,j)\right)$$

and

$$\text{DISC}_\cap(T_k, T_\ell) = \frac{1}{I_{k,\ell}(I_{k,\ell}-1)} \sum_{\substack{(i,j) \in I_{k,\ell}{}^2 \\ i \neq j}} \text{Jacc}\left(D_k(i,j), D_\ell(i,j)\right).$$

Note that the actual sets compared using the Jaccard distance may themselves contain mutations that exist in only one of the trees. Thus, these distances are not the same as removing all non-shared mutations from the trees, contracting the tree topology and computing the original version of the distances. See the Supplementary Appendix for analyses of these measures with regard to metric properties.

This variation is most useful when differences in mutation labelings between trees should not strongly contribute to their distance value. For example, this method may be useful when trying to identify common patterns of evolution across different patients. A degenerate case arises if this approach is applied to two trees with disjoint mutation sets, resulting in both $\text{CASet}_\cap$ and $\text{DISC}_\cap$ being 0.

2.4.2 Union of mutation sets

In the second extension to clonal trees, we consider the union of the mutation sets for the input trees. To do so, we need to address how to handle mutations that exist in only one tree. Let $U_{k,\ell} = M(T_k) \cup M(T_\ell)$ be the union of the sets of mutations labeling T_k and T_ℓ. If $i \notin M(T_k)$, then we define $A_k(i) = \varnothing$. Thus, we can modify both CASet and DISC distances as follows:

$$\text{CASet}_\cup(T_k, T_\ell) = \frac{1}{\binom{|U_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq U_{k,\ell}} \text{Jacc}\ \left(C_k(i,j), C_\ell(i,j)\right)$$

and

$$\text{DISC}_\cup(T_k, T_\ell) = \frac{1}{U_{k,\ell}(U_{k,\ell}-1)} \sum_{\substack{(i,j) \in U_{k,\ell}{}^2 \\ i \neq j}} \text{Jacc}\left(D_k(i,j), D_\ell(i,j)\right).$$

This variation allows differences in the sets of mutation labels to contribute to the distance computed between two trees. Thus, this variation may be most useful for comparing tumor evolutionary trees generated by different data types, across samples taken at different times, or even across patients. Note that because $\text{Jacc}(X, \varnothing) = 1$ if $X \neq \varnothing$, the distance between trees with disjoint labels is 0. In the Supplementary Appendix, we describe a formula that relates $\text{CASet}_\cup$ and $\text{CASet}_\cap$, allowing for computation of $\text{CASet}_\cup$ with fewer operations, and examine the metric properties of both of these measures.

3 Results

We analyze and compare CASet and DISC to existing distance measures on simulated datasets and two real datasets. We find that our methods allow for more granularity in comparing tumor evolutionary trees and outperform other methods when used in a clustering context.

3.1 Edit location and tree structure effects

We first evaluate how specific labeling and topology differences between trees affect CASet and DISC and then compare the effects for other distance measures. To measure the effect caused by labeling differences, we constructed a seven-node complete binary tree (the 'base tree') and then created a set of trees consisting of all possible pairwise label swaps of the base tree. Figure 3A shows the computed distance between the base tree and all other trees for CASet, DISC, MLTED (Karpov *et al.*, 2018) and Ancestor–Descendant (A–D) (Govek *et al.*, 2018) distances. See the Supplementary Appendix for results including other distances from Govek *et al.* (2018). We note that both CASet and DISC are able to completely distinguish all classes of mutation swaps based on the location of the swapped labels in the tree. In contrast, A–D is unable to distinguish some of these classes and MLTED computes the same distance for almost all classes. We do see a loose trend that swaps that include mutations labeling the leaves of the tree result in smaller distances than swaps that include mutations that label nodes closer to the root of the tree for both CASet and DISC. However, CASet and DISC do have different relative orderings of some classes of label swaps compared with the base tree. CASet most strongly penalizes swaps that include the label on the root of the tree and has a large corresponding jump in computed distance when comparing these trees to the base tree. On the other hand, DISC considers the tree that swaps the two children of the root to have the largest distance from the base tree. We see similar trends when we consider other base trees (Supplementary Appendix).

We also perform an analogous experiment to assess the effect of topology rather than labeling. Specifically, we constructed a 10-node linear tree as the base tree and constructed a set of comparison trees by moving the leaf node of the base tree to every level of the tree (Fig. 3B). We find that CASet, DISC and A–D all consider placement of the moved node closer to the root to have a larger distance from the base tree. This is as we might expect, since these represent older evolutionary differences. While the decrease in A–D is linear as the node is placed deeper, the curves for CASet and DISC are concave up and down, respectively. Intuitively, this means that CASet more heavily prioritizes topological changes higher up in the tree while assigning a lower weight to changes near the leaves. In contrast, DISC places higher relative weight on topological differences closer to the leaves.

Having seen that the location of differences between trees affect our measures in accordance with evolutionary context, we also consider how global properties of tree structure affect our measures. To this end, we created a base 15-node complete binary tree and measured its distance to 5000 randomly generated 15-clonal trees. We compare the height, maximum branching factor and balance (as measured by total cophenetic index; Mir *et al.*, 2013) of the random trees to their distances from the base tree (full results in the Supplementary Appendix). We find that trees taller than the base tree tend to have higher CASet, DISC and A–D distances, but do not observe this effect in MLTED. On the other hand, we find that a higher maximum branching factor than the base tree correlates with decreased DISC and A–D distance, while the opposite is true for MLTED and CASet does not show a clear correlation with branching factor. Unbalanced trees tend to have higher CASet and A–D distances, but balance does not have a strong effect on DISC or MLTED.

3.2 Clustering application

Previous work has shown that many different trees can be consistent with data from a single patient (Pradhan and El-Kebir, 2018; Tomlinson and Oesper, 2018). Clustering these trees is a compelling use case for tumor evolution distance measures as it has the potential to reveal structure in the space of compatible trees of a single dataset. Clustering trees inferred from different patients can also be used to identify shared evolutionary patterns.

3.2.1 Dataset generation

We created two different simulated datasets for our clustering analysis. In the first dataset, we manually constructed 5 base clonal trees each with 15 mutations but different topologies and labelings (see Supplementary Appendix), and generated 5 variants of each base tree for a total of 25 trees. The second dataset was generated using the OncoLib (Pradhan and El-Kebir, 2018) tree generation tool, which simulates tumor evolutionary trees and read count data. We took the read count data from 5 OncoLib simulations (each with

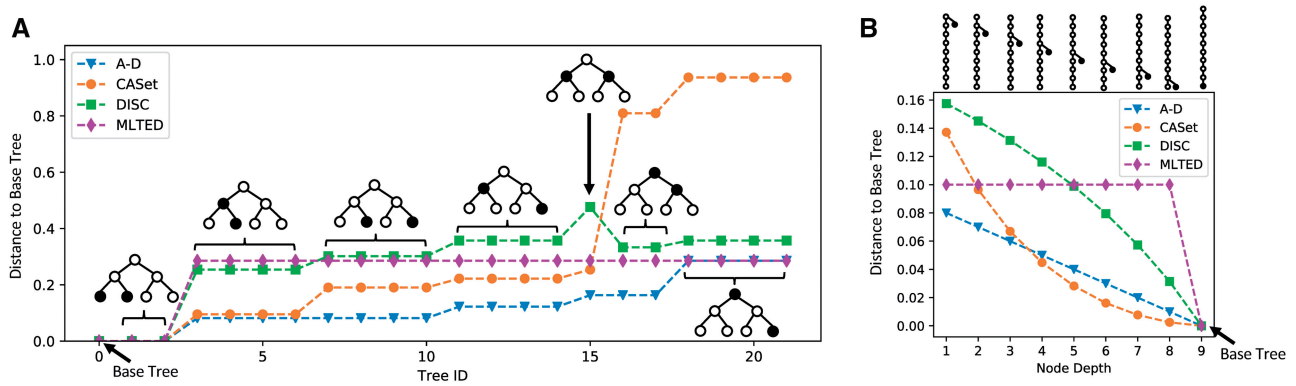

Fig. 3. (**A**) Effect of label swaps when comparing a tree to the base tree. Each point shows the distance between a tree with a single pair of label swaps and the base tree. Trees are ordered according to increasing CASet distance. The tree annotations show the type of label swap to which each data point corresponds. (**B**) Effect of a single topological changes to a tree when comparing a tree to the base tree. Node depth refers the height to which the leaf of the base tree was moved in the modified tree. The trees above the plot illustrate the movement of the leaf. In both (A) and (B), A–D distance is normalized by m^2 to place it on the same scale as the other distances.

3 sequenced samples and 10–20 mutations) and reconstructed all clonal trees that were consistent with the simulated data using the approach of Tomlinson and Oesper (2018). Each of the five OncoLib simulations yielded a 'tree family' of between 50 and 10 000 clonal trees that were consistent with the simulated sequencing data, of which we randomly sampled 50 from each family to create a dataset of 250 trees from 5 families (labeled A–E). All trees within a family have the same set of mutations, but mutation sets differ across families. The true underlying clonal trees produced by OncoLib are in the Supplementary Appendix. For an analysis of the correlation between CASet and DISC for the trees in these datasets, see the Supplementary Appendix.

3.2.2 Clustering clonal trees

We perform hierarchical clustering with average linkage on both the manual and OncoLib datasets. On the manual dataset, we compare our distance measures to the following other methods: MLTED (Karpov *et al.*, 2018), the four distance measures, A–D, parent–child, clonal and path distances, described in Govek *et al.* (2018) (see Supplementary Appendix for details on normalizations applied), and triplet distance (Critchlow *et al.*, 1996) (a modified version of a distance designed for phylogenetic trees, described in the Supplementary Appendix). For each method, we compute the average silhouette value (Rousseeuw, 1987) for different cuts of the resulting tree and use the cut with the highest such value to produce a clustering of the data. We find that several of the methods (including CASet and DISC) produce the correct number of clusters (five), and that CASet has the highest average silhouette score (0.85) for this cut (Supplementary Appendix).

We perform a similar analysis on the OncoLib dataset, but only compare $\text{CASet}_\cup$, $\text{CASet}_\cap$, $\text{DISC}_\cup$, $\text{DISC}_\cap$, MLTED and A–D distance. We do not test parent–child or clonal distance, since they performed poorly on the simpler manual dataset. We also do not apply triplet distance, since it does not have a natural extension to clonal trees with different mutation sets, as are present in the different OncoLib tree families. When the hierarchical clustering is cut at five clusters, all six distance measures correctly clustered the trees, but they did so with varying degrees of tightness (Fig. 4A). Furthermore, if we consider the cut with the best silhouette score, all methods identified as the optimal hierarchical clustering cut except for $\text{CASet}_\cap$, which had an optimal cut at three clusters (Fig. 4B). $\text{CASet}_\cup$ performed best in distinguishing trees belonging to different datasets, with an average silhouette score of 0.81 over the five tree families. While it performs worse at separating different families, $\text{CASet}_\cap$ identifies that the pairs A, B and D, E have strong agreement about ancestral relationships among their shared mutations. This highlights the different useful features of $\text{CASet}_\cup$ and $\text{CASet}_\cap$.

3.2.3 Intra-family clustering structure

Figure 4 shows that tree family E from the OncoLib dataset might have internal structure, so we took a closer look at the

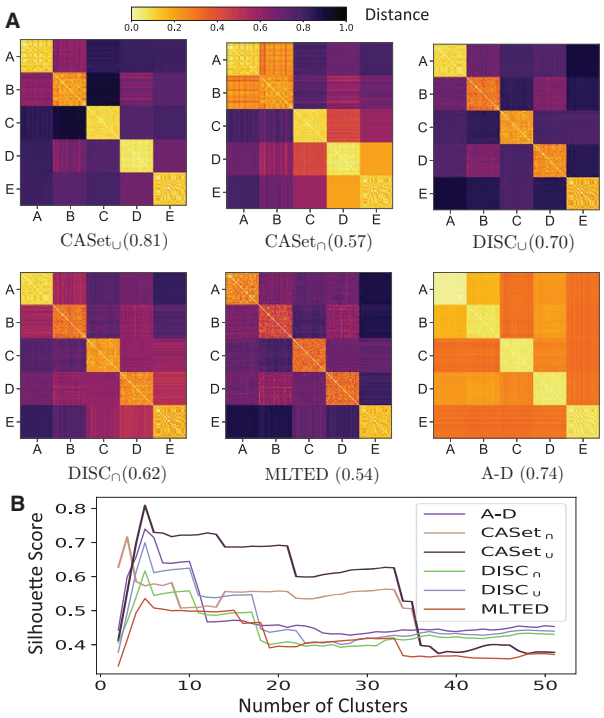

Fig. 4. (**A**) Inter-dataset distance heatmaps of five tree families in the OncoLib dataset. The color of each cell represents the distance between two trees. Average silhouette scores for five clusters are displayed in parentheses to quantify clustering tightness and separation. We note that we normalized the A–D distance by dividing by the total number of possible ancestral relationships, so that all distances were in the interval [0, 1]. While this results in A–D distances that are all relatively small (<0.4), the families still exhibit clear clustering. (**B**) Silhouette scores of clustering the 250 trees in the OncoLib dataset with different distance measures

internal clustering structure for this tree family (Fig. 5). In this dataset, CASet, DISC and MLTED all identify two primary clusters of trees, but CASet and DISC are both able to resolve more complex substructure. In particular, CASet distinguishes between eight strongly defined subfamilies, with an average silhouette score of 0.94. The same eight subfamilies are visible in the DISC heatmap along with a finer-grained resolution within each of these clusters. In contrast, the MLTED heatmap shows less resolution within the two primary categories. None of the other four tree families had internal structure as well-defined as family E, but for a similar analysis of tree family A, see the Supplementary Appendix.

3.3 Results on real datasets

We apply our distance measures to two different breast cancer datasets: Wang *et al.* (2014) and Eirew *et al.* (2015). We first apply CASet and DISC to the three potential tumor evolutionary histories reported in Karpov *et al.* (2018) recovered using three different methods, PHiSCS (Malikic *et al.*, 2018), SciFit (Zafar *et al.*, 2017) and SCITE (Jahn *et al.*, 2016) applied to single cell sequencing data from a triple negative breast cancer patient (Wang *et al.*, 2014) (Fig. 6). Since CASet and DISC are designed to evaluate the topology of trees in addition to their labels and inheritance, both measures are able to provide more granular information about the similarity of the trees making it possible to conclude that T_2 is more similar to T_1 than it is to T_3. This is in contrast to MLTED (Karpov *et al.*, 2018), which considers these pairs of trees to have the same similarity, and, furthermore, computes the distance between T_1 and T_3 as 0 since they can be clonally expanded to match. While it may be useful to evaluate the similarity of trees with regard to such clonal expansion, a tumor evolutionary tree is categorized in part by subclonal populations that represent the evolutionary patterns of the tumor. Ignoring these does not fully take into account the information represented in a tumor evolutionary tree.

We also emphasize the importance of using quantitative measures when comparing reconstructed trees. Husić *et al.* (2018) introduced a new tree reconstruction method MIPUP, and when they compare their results to those of another method, LICHeE (Popic *et al.*, 2015), on breast cancer xenoengraftment data sample SA501 (Eirew *et al.*, 2015), they use only qualitative analysis to claim their method produces phylogenies closer to those in the original publication. We assess this claim quantitatively by running CASet, DISC, MLTED and A–D on the SA501 tree from Eirew *et al.* (2015) and the corresponding trees reconstructed by MIPUP and LICHeE, as reported by Husić *et al.* (2018) (Table 1; Supplementary Appendix). We find that most of the distances measures reported that in fact the LICHeE tree is more similar to the phylogeny proposed by Eirew *et al.* (2015) than the MIPUP tree. The only measures to report the opposite are MLTED (by only a small margin) and $\text{DISC}_\cup$. The MIPUP tree contains many mutations not in the tree proposed by Eirew *et al.* (2015) that are clustered together in the tree. As a result, all pairs of these mutations have an empty distinctly inherited set, and therefore act to artificially lower the $\text{DISC}_\cup$ distance. Thus we can see that when considering ancestral patterns of inheritance, the LICHeE tree may in fact better match the original tree.

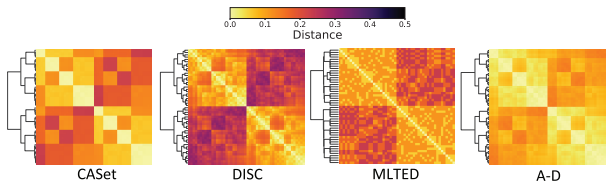

Fig. 5. Hierarchical clustering of tree family E shown with corresponding heatmap for pairwise distances. Note that the colormap range has been reduced to provide more contrast

4 Discussion

In this work, we argue that distance measures designed for tumor evolutionary trees are needed for assessing phylogeny inference methods and for exploring the relationships between sets of evolutionary trees. To this end, we introduce two new tumor evolution distance measures, CASet and DISC. By comparing the common ancestors of all mutation pairs, CASet incorporates differences in both mutation labeling and tree topology. In particular, CASet uses the number of clones that inherit common mutations when weighting the effect of mutation labeling differences between trees. In contrast, DISC pays special attention to the set of mutations that distinguish clones from each other, placing comparatively more emphasis on recently acquired mutations. We extend both distance measures to apply clonal trees with different sets of mutations.

We demonstrate the differences between CASet and DISC on simulated data and use a clustering application to show that that CASet is better able to distinguish groups of trees than existing distance measures. Moreover, we find that both CASet and DISC can identify complex clustering structure in a space of trees that is missed by other distance measures.

Using a breast cancer dataset, we show that CASet and DISC are able to differentiate between trees with differing clonal makeups, demonstrating their topological acuity. In addition, we use CASet and DISC to assess trees reconstructed by MIPUP (Husić *et al.*, 2018) and LICHeE (Popic *et al.*, 2015) from a breast cancer xenograft dataset (Eirew *et al.*, 2015). Our results suggest that the MIPUP tree may not more closely resemble the original

Table 1. Pairwise distances between the base tree reported in Eirew *et al.* (2015) for sample SA501, and the trees inferred by LICHeE and MIPUP as reported in Husić *et al.* (2018)

Distance measure	MIPUP to base tree	LICHeE to base tree
$\text{CASet}_\cup$	0.88	**0.74**
$\text{CASet}_\cap$	0.84	**0.78**
$\text{DISC}_\cup$	**0.40**	0.60
$\text{DISC}_\cap$	0.40	**0.38**
MLTED	**0.80**	0.81
A–D	0.70	**0.46**

Note: Bold text indicates which pair of trees are reported to be more similar.

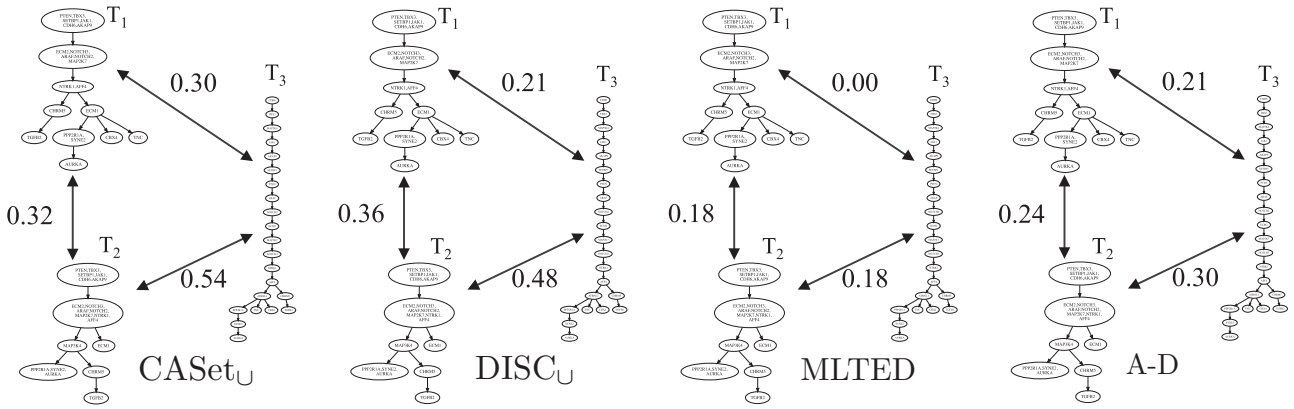

Fig. 6. Tumor evolutionary trees inferred by PHiSCS (T_1), SciFit (T_2) and SCITE (T_3) from a triple negative breast cancer patient and corresponding pairwise distances. See the Supplementary Appendix for full sized images of the trees

hypothesized tree than the LICHeE tree, as is suggested in Husić *et al.* (2018) based solely on qualitative analysis.

Future work is needed to determine the benefits of using CASet or DISC in methods such as GraPhyC (Govek *et al.*, 2018) that rely explicitly on distance measures. Our findings on the internal clustering structure of the OncoLib tree families also invite further investigation. It remains to be seen whether the same kind of structure is found in real cancer data—either within sets of trees consistent with data from a single patient or across sets of patients with the same type of cancer. If so, this structure could provide insight into improved tumor phylogeny inference methods or common mutational patterns across patients.

A number of different methodological extensions may make these distance measures more widely useful. As presented here, both CASet and DISC may allow for violations of the ISA in the form of mutation deletions by simply using a unique label for each deletion. However, future work is needed to allow for more complex violations of the ISA. Also, as currently implemented these distance measures treat all mutations equally. However, a user may wish to weight mutations differently, e.g. weighting mutations to a known driver gene more heavily. Our method could be extended to take a user defined weight for each mutation and then weighting each computed Jaccard distance relative to the sum of the weights of the mutations considered when computing the sets being compared. We also could explore the use of distance measures other than the Jaccard distance in our approach. Specifically, different evolutionary models that account for the difference in transversion rates may be useful when trees only contain SNVs.

Our analysis indicates that CASet more strongly penalizes edits that have the biggest evolutionary impact, such as those including the root, than other distance measures including DISC. Therefore, we recommend that CASet be used for benchmarking new inference methods since mistakes in inference near the root should be much more costly than mistakes near the leaves. As far as clustering applications, our analysis indicates that CASet may be most useful when a user expects there are sets of trees that are very different from each other whereas DISC should be used to achieve more granularity when clustering sets of relatively similar trees. However, a distance measure that integrates these disparate properties would be very useful for clustering applications. Thus, future work will include development of a method that, rather than using common ancestor and distinctly inherited sets separately, uses a partition of all mutations in each tree. Specifically, each pair of mutations in a tree defines such a partition over all mutations in the tree where components in this partition correspond to the associated commonly and distinctly inherited sets for the mutations being considered. Then, an approach such as the Rand Index could be used to compare these partitions.

For applications outside of clustering or benchmarking, we encourage users to consider what types of differences between trees they wish to prioritize. Users then should use our experimental results that show how labeling and topology affect the distance measures as a guide for choosing which distance measure is most appropriate for their application. For instance, if differences in labeling close to the root should be emphasized, CASet would be the appropriate choice. On the other hand, DISC would be better suited to comparing shallow trees whose differences occur near the leaves.

Acknowledgements

We thank Jack Kupiers for useful conversations related to this work.

Funding

This work has been supported by National Science Foundation award IIS-1657380, Elledge, Eugster and Class of '49 Fellowships from Carleton College.

Conflict of Interest: none declared.

References

Amirouchene-Angelozzi,N. *et al.* (2017) Tumor evolution as a therapeutic target. *Cancer Discov.*, **7**, 805–817.

Bille,P. (2005) A survey on tree edit distance and related problems. *Theor. Comput. Sci.*, **337**, 217–239.

Blakely,C.M. (2017) Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat. Genet.*, **49**, 1693–1704.

Bonizzoni,P. *et al.* (2018) Does relaxing the infinite sites assumption give better tumor phylogenies? An ILP-based comparative approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. **16**, 1410–1423.

Critchlow,D.E. *et al.* (1996) The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.*, **45**, 323–334.

Eirew,P. *et al.* (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, **518**, 422.

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

El-Kebir,M. *et al.* (2015) Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, **31**, i62–i70.

El-Kebir,M. *et al.* (2016) Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.*, **3**, 43–53.

Govek,K. *et al.* (2018) A consensus approach to infer tumor evolutionary histories. In: *2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB'18*, pp. 63–72. ACM, New York, NY, USA.

Husić,E. *et al.* (2018) MIPUP: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP. *Bioinformatics*, **35**, 769–777.

Jahn,K. *et al.* (2016) Tree inference for single-cell data. *Genome Biol.*, **17**, 86.

Jiang,Y. *et al.* (2016) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. USA*, **113**, E5528–37.

Jiao,W. *et al.* (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, **15**, 35.

Jolly,C. and Van Loo,P. (2018) Timing somatic events in the evolution of cancer. *Genome Biol.*, **19**, 95.

Karpov,N. *et al.* (2018) A multi-labeled tree edit distance for comparing "clonal trees" of tumor progression. In: Parida,L. and Ukkonen,E. (eds) *18th International Workshop on Algorithms in Bioinformatics (WABI 2018), Volume 113 of Leibniz International Proceedings in Informatics (LIPIcs)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 22:1–22:19.

Kim,K.I. and Simon,R. (2014) Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, **15**, 27.

Malikic,S. *et al.* (2018) PhISCS—a combinatorial approach for sub-perfect tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data. *bioRxiv*.

Malikic,S. *et al.* (2015) Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, **31**, 1349–1356.

Marass,F. *et al.* (2016) A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.*, **10**, 2377–2404.

Matsui,Y. *et al.* (2017) phyc: clustering cancer evolutionary trees. *PLoS Comput. Biol.*, **13**, e1005509.

Mir,A. *et al.* (2013) A new balance index for phylogenetic trees. *Math. Biosci.*, **241**, 125–136.

Miura,S. *et al.* (2018) Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics*, **34**, i917–i926.

Niknafs,N. *et al.* (2015) Subclonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput. Biol.*, **11**, e1004416.

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

Popic,V. *et al.* (2015) Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.*, **16**, 91.

Pradhan,D. and El-Kebir,M. (2018) On the non-uniqueness of solutions to the perfect phylogeny mixture problem. In: *RECOMB International Conference on Comparative Genomics*, RECOMB-CG'18, pp. 277–293. Springer, Berlin.

Raphael,B.J. *et al.* (2014) Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.*, **6**, 5.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Ross,E.M. and Markowetz,F. (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, **17**, 69.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Satas,G. and Raphael,B.J. (2017) Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, **33**, i152–i160.

Schwartz,R. and Schäffer,A.A. (2017) The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, **18**, 213–229.

Tomlinson,K. and Oesper,L. (2018) Examining tumor phylogeny inference in noisy sequencing data. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM'18, pp. 36–43. IEEE.

Toosi,H. *et al.* (2017) BAMSE: Bayesian model selection for tumor phylogeny inference among multiple tumor samples. In: *2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pp. 1–1. IEEE, Orlando, FL.

Wang,Y. *et al.* (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, **512**, 155–160.

Zafar,H. *et al.* (2017) SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.*, **18**, 178.