Bioinformatics, 37(2), 2021, 178–184
doi: 10.1093/bioinformatics/btaa676
Advance Access Publication Date: 30 July 2020
Original Paper

OXFORD

Phylogenetics

Triplet-based similarity score for fully multilabeled trees with poly-occurring labels

Simone Ciccolella * Giulia Bernardini, Luca Denti , Paola Bonizzoni, Marco Previtali and Gianluca Della Vedova

Department of Informatics, Systems and Communication, University of Milano-Biccoca, Milan 20126, Italy

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson
Received on April 29, 2020; revised on June 29, 2020; editorial decision on July 18, 2020; accepted on July 22, 2020

Abstract

Motivation: The latest advances in cancer sequencing, and the availability of a wide range of methods to infer the evolutionary history of tumors, have made it important to evaluate, reconcile and cluster different tumor phylogenies. Recently, several notions of distance or similarities have been proposed in the literature, but none of them has emerged as the golden standard. Moreover, none of the known similarity measures is able to manage mutations occurring multiple times in the tree, a circumstance often occurring in real cases.

Results: To overcome these limitations, in this article, we propose MP3, the first similarity measure for tumor phylogenies able to effectively manage cases where multiple mutations can occur at the same time and mutations can occur multiple times. Moreover, a comparison of MP3 with other measures shows that it is able to classify correctly similar and dissimilar trees, both on simulated and on real data.

Availability and implementation: An open source implementation of MP3 is publicly available at https://github.com/AlgoLab/mp3treesim.

Contact: simone.ciccolella@unimib.it

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

1 Introduction

Recent methods to accurately infer the clonal evolution and progression of cancer have made it possible to develop targeted therapies for treating the disease. As discussed in several studies (Morrissy *et al.*, 2016; Wang *et al.*, 2016), understanding the history of accumulation and the prevalence of somatic mutations during cancer progression is a fundamental step to devise these treatment strategies.

Given the importance of the task, a multitude of methods for cancer phylogeny reconstruction have been developed over the years. The increasing number of tools created has been encouraged by the diversity of data available; for instance, we are witnessing a shift from bulk sequencing data (Bonizzoni *et al.*, 2017, Bonizzoni et al., 2019, ; Hajirasouliha *et al.*, 2014; Hajirasouliha and Raphael, 2014; Yuan *et al.*, 2015) toward single-cell data (Ciccolella *et al.*, 2018a,b; El-Kebir, 2018; Jahn *et al.*, 2016; Zafar *et al.*, 2017) and hybrid approaches (Malikic *et al.*, 2019a,b).

Having many different tools accomplishing the same task requires solid methods to compare their results. In contrast with classical phylogenetic trees, whose leaves, and only leaves, are labeled (with the species they represent), the trees that model tumor phylogenies are *fully labeled*, i.e. they also have labels (corresponding to the mutations) on the internal nodes. While there is a wide range of measures to compare leaf-labeled trees in the literature, *ad hoc* methods for tumor phylogenies are starting to appear in the past few years (DiNardo *et al.*, 2020; Bernardini *et al.*, 2019, Bernardini et al. 2020; Govek et al. 2018; Karpov *et al.*, 2019); in particular, a detailed study of some notions of distance (DiNardo *et al.*, 2020) has introduced two new measures complementing some more established definitions used in various cancer inference studies (Ciccolella *et al.*, 2018a,b). Those new measures are more nuanced, to capture some aspects of the mutation inheritance process, while still being very efficient to compute. A common trait of all the latter distances is their reliance on the analysis of *pairs* of nodes.

On the other hand, some of the most widely used distances on classical phylogenies are based on rooted triples (Aho *et al.*, 1981; Brodal *et al.*, 2013; Dobson, 1975) (for rooted phylogenies) or quartets (Dudek and Gawrychowski, 2019) (for unrooted phylogenies) of labeled leaves. Although such metrics have major limitations for our purposes, as they do not apply directly to fully labeled trees, they also have some desirable properties that we would like to transfer in our setting. Specifically, this kind of metric captures well the differences in the topology of the trees; a feature that, to the best of our knowledge, lacks in most of the existing methods for tumor phylogenies. Therefore, we expect a triplet-based measure to provide additional insights on the different evolutionary histories, when applied to cancer progression.

In this article, we generalize the notion of rooted triples similarity for classical phylogenies to tumor phylogenies. Moreover, we

further extend this to multilabeled trees (i.e. where each node is labeled by a set of labels) and poly-occurring labels (i.e. each label can be assigned to more than one node). The latter feature is needed since recent studies (Brown *et al.*, 2017; Kuipers *et al.*, 2017) suggest widespread recurrence and loss of mutations, and more and more methods designed to infer tumor phylogenies considering such a possibility are starting to appear (Ciccolella *et al.*, 2018a,b; El-Kebir, 2018). In a phylogenetic tree, a mutation loss is represented by a special character in the label, such as a minus sign: the design of our measure allows to handle such evolutionary events effectively, as they uniquely correspond to their label like any other kind of mutation.

Through an extensive experimental analysis, we show that our novel measure is able to overcome the limitations in the existing literature and to provide a better alternative to both the direct comparison of evolutionary histories and the application to established clustering techniques, following the approach of DiNardo *et al.* (2020). Such a performing measure can also be incorporated in recent works (Aguse *et al.*, 2019; Govek *et al.*, 2018) designed to cluster and build consensus across multiple cancer progressions.

2 Materials and methods

A classical phylogenetic tree is a rooted, unordered and leaf-labeled tree. The set of all the labels occurring in T is denoted by $\lambda(T)$, and a function $\mathtt{N}(\cdot)$ maps each element of $\lambda(T)$ to a leaf of T. We denote with $\mathtt{LCA}(u,v)$ the Lowest Common Ancestor of nodes u and v. Given three leaves $u,v,z \in V_T$, the *minimal tree topology* they induce on T, denoted as $\mathtt{MTT}_T(u,v,z)$, is the smallest subtree of T that includes the nodes $V_T^{u,v,z} = \{u,v,z\} \cup \mathtt{LCA}(u,v) \cup \mathtt{LCA}(v,z) \cup \mathtt{LCA}(u,z)$, and where all the nodes with degree 2 not in $V_T^{u,v,z}$ are contracted.

The rooted triplet distance measures the dissimilarity between two leaf-labeled trees with identical labels. It is given by the number of rooted triplets that induce different minimal topologies (Fig. 1) in the two trees over the total number of triplets (Jansson and Rajaby, 2017). As tumor progression trees are fully labeled, such metric cannot be directly applied: in this section, we propose a novel similarity measure, inspired by the triplet distance, specifically designed for these more general trees.

2.1 Extension to fully labeled trees and multilabeled trees

A tree T on a set V_T of n nodes is *fully labeled* by a set $\lambda(T)$ of labels if there is a bijection $\mathtt{N} : \lambda(T) \to V_T$. The definition of minimal topology of three leaves can be trivially extended to the minimal topology of three nodes: we next show that there are only five possible configurations (see Fig. 2).

Lemma 1. *Given nodes $u,v,z \in V_T$, there exist only five possible configurations for* $\mathtt{MTT}_T(u,v,z)$.

Proof. We start by dividing two possible cases: (i) $\mathtt{LCA}(u,v) = \mathtt{LCA}(v,z) = \mathtt{LCA}(u,z)$, or (ii) just two LCAs are the same, say $\mathtt{LCA}(v,z) = \mathtt{LCA}(u,z)$. There are no other possibilities, as $\mathtt{LCA}(u,v) \neq \mathtt{LCA}(v,z) \neq \mathtt{LCA}(u,z)$ is impossible: indeed, suppose without loss of generality that $\mathtt{LCA}(u,v)$ is a descendant of $\mathtt{LCA}(u,z)$, $\mathtt{LCA}(u,v) \neq \mathtt{LCA}(u,z)$: they cannot be unrelated, as by definition they are both ancestors of u. $\mathtt{LCA}(u,z)$ is thus a common ancestor for v and z. Suppose toward a contradiction that $\mathtt{LCA}(v,z) \neq \mathtt{LCA}(u,z)$, thus it is a descendant of $\mathtt{LCA}(u,z)$ and an ancestor of $\mathtt{LCA}(u,v)$. But then it is an ancestor of both u an z and it is lower than $\mathtt{LCA}(u,z)$, a contradiction.

Case (i) has two subcases: either $\mathtt{LCA}(u,v) \in \{u,v,z\}$, corresponding to the rightmost configuration in Figure 2, or $\mathtt{LCA}(u,v) \notin \{u,v,z\}$, corresponding to the second configuration from the left. Case (ii) has three subcases: either both the distinct LCAs are in $\{u,v,z\}$, or none of the two is, or finally one is in $\{u,v,z\}$ and the other is not. The first subcase corresponds to the leftmost configuration in Figure 2, the second subcase to the fourth configuration from the left. For the third subcase, either the external LCA is an ancestor of all of the three $\{u,v,z\}$, corresponding to the third configuration, or it is an ancestor of two nodes and a descendant of the third one, say u. In the latter case, though, the external node would be the only child of u, and thus would be contracted by definition of $\mathtt{MTT}_T(u,v,z)$, leading again to the rightmost configuration of Figure 2.

In the case of fully labeled trees, the definition of LCA of two nodes and MTT of three nodes can trivially be extended to the LCA of two labels and the MTT of three labels, as there is a one-to-one correspondence between nodes and labels. From now on, for ease of presentation, given two nodes u and v and their respective labels a and b, we will use $\mathtt{LCA}(u,v)$ or $\mathtt{LCA}(a,b)$ interchangeably. When modeling tumor progression, though, to have a bijection between nodes and labels (i.e. mutations) is quite a strong assumption, as multiple mutations often appear at the same time in the evolutionary history of cancer. We thus relax our assumptions and consider *multilabeled* instead of fully labeled trees.

A rooted, unordered tree T is multilabeled if there exists a surjective function $\mathtt{N} : \lambda(T) \to V_T$ that labels each node of T with a set of labels from $\lambda(T)$: note that, in this model, each label is assigned to one and only one node of T. We extend the definition of lowest common ancestor of two labels for a multilabeled tree as follows: if $a \in \lambda(T)$ and $b \in \lambda(T)$ label the same node u, then $\mathtt{LCA}(a,b) = u$; if they label two distinct nodes u, v, then $\mathtt{LCA}(a,b) = \mathtt{LCA}(u,v)$. This allows us to straightforwardly extend the definition of minimal tree topology of three labels for multilabeled trees. There are only four possible additional configurations for the minimal tree topology of multilabeled trees, shown in Figure 3: a proof can be found in the Supplementary Materials.

Lemma 2. *Given T multilabeled and $a,b,c \in \lambda(T)$, there exist nine configurations for* $\mathtt{MTT}_T(a,b,c)$.

2.2 Extension to poly-occurring labels

We further extend our model of tumor phylogeny by allowing the same label of $\lambda(T)$ to be assigned to multiple nodes of T. An element of $\lambda(T)$ that labels more than one node of T is said to be a *poly-occurring* label. To the best of our knowledge, none of the existing tools is able to handle poly-occurring labels: indeed, although some of them accept input trees with poly-occurring labels, they simply disregard the multiple occurrences of a same label.

Since it is often the case where the inferred evolutionary history involves the appearance of the same mutation in multiple events, a meaningful comparison between tumor phylogenies cannot

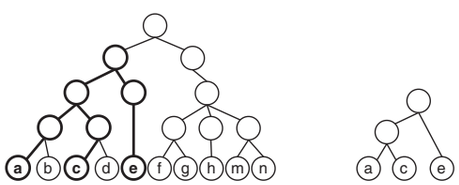

Fig. 1. Rooted triplet on labels (a, c and e). (*Left*) Tree T where the smallest subtree that contains all three labels is highlighted. (*Right*) The minimal topology induced by (a, c and e)

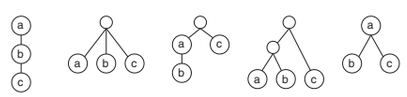

Fig. 2. The five possible configurations for the minimal tree topology induced by an unordered set of three labels

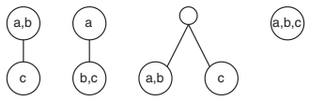

Fig. 3. The four additional possible configurations for the minimal tree topology of multilabeled trees induced by an unordered set three labels

overlook such a phenomenon. To consider poly-occurring labels in our similarity measure, we extend the definition of minimal tree topology. First, note that if a label occurs multiple times in the tree, then $\mathtt{N}$ maps each label to one or more nodes in V_T. Then, we define the minimal tree topology of poly-occurring labels a, b, c, denoted by M, as follows, where $\sqcup$ indicates the multiset union:

$$\mathtt{M}_T(a,b,c) = \bigsqcup_{u\in \mathtt{N}(a), v\in \mathtt{N}(b), z\in \mathtt{N}(c)} \mathtt{MTT}_T(u,v,z)$$

In other words, the minimal tree topology of three labels is the multiset of all the minimal tree topologies of the nodes where a, b and c appear. We remark that in this setting $\mathtt{M}_T$ is a multiset of configurations, thus the same configuration may appear multiple times in $\mathtt{M}_T$.

2.3 Similarity measure between trees

We are now able to define a similarity measure between fully labeled trees with poly-occurring labels. Let S be a multiset and let $|S|$ be its cardinality. We define the number of shared configurations of labels a, b, c between two trees T_1 and T_2 as $N(a,b,c) = |\mathtt{M}_{T_1}(a,b,c) \sqcap \mathtt{M}_{T_2}(a,b,c)|$, i.e. the cardinality of the multiset intersection, and the maximum number of configurations of the triplet in the trees as $D(a,b,c) = max\{|\mathtt{M}_{T_1}(a,b,c)|, |\mathtt{M}_{T_2}(a,b,c)|\}$.

Based on these two values, we define multiple variations of the multipoly-occurring labels triplet-based (MP3) similarity measure that we will later combine into a single score. We define $\mathtt{MP3}_\cap$ as the similarity computed between triplets of labels shared by the two trees:

$$\mathtt{MP3}_\cap = \frac{\sum_{(a,b,c)\in I} N(a,b,c)}{\sum_{(a,b,c)\in I} D(a,b,c)} \tag{1}$$

where I is the set of triples in $\lambda(T_1) \cap \lambda(T_2)$. Due to the nature of only considering the subset of labels that appears in both trees, $\mathtt{MP3}_\cap$ is a conservative measure, therefore, we present a variation that considers all possible configurations in both trees, thus having a wider view:

$$\mathtt{MP3}_\cup = \frac{\sum_{(a,b,c)\in J} N(a,b,c)}{\sum_{(a,b,c)\in J} D(a,b,c)} \tag{2}$$

where J is the set of triples in $\lambda(T_1) \cup \lambda(T_2)$. Different from $\mathtt{MP3}_\cap$, $\mathtt{MP3}_\cup$ weighs also the labels that appear only in one of the trees. Note that, for every pair of trees, $\mathtt{MP3}_\cup \leq \mathtt{MP3}_\cap$, as the numerator remains identical in both, while the denominator of $\mathtt{MP3}_\cup$ has all the elements in $\mathtt{MP3}_\cap$ with the addition of the values of D for the triples present only in one of the input trees.

Although $\mathtt{MP3}_\cap$ and $\mathtt{MP3}_\cup$ are closely related, they provide two different views of a tumor phylogeny. Indeed, on one hand, $\mathtt{MP3}_\cap$ measures how similar the shared history of two tumor phylogenies is, i.e. it provides an idea of how well the two progressions can be reduced to the same subsequence of common mutations. On the other hand, $\mathtt{MP3}_\cup$ measures how similar the whole histories of the two evolutions are, i.e. it considers the impact of mutations acquired (or lost) in only one progression.

Since the previous measures capture different aspects of the progressions, we want to combine them into a single, usable and powerful similarity measure that couples the strengths of both. The most intuitive method is to simply use a mean. We opted for the geometric mean: $\mathtt{MP3}_G = \sqrt{\mathtt{MP3}_\cap \cdot \mathtt{MP3}_\cup}$.

This function is not completely satisfactory, as a uniform function of 1 and 2 is not able to comprehensively capture the nuances in the input trees. Therefore, we developed a weighted mean with an intentional bias toward $\mathtt{MP3}_\cap$ to catch inner similarities in different trees. Such combination then tends to be closer to $\mathtt{MP3}_\cap$ when the trees are similar while moving toward $\mathtt{MP3}_\cup$ as the trees are less similar:

$$\mathtt{MP3}_\sigma = \mathtt{MP3}_\cup + \sigma(\mathtt{MP3}_\cap) \cdot min\{\mathtt{MP3}_\cap - \mathtt{MP3}_\cup, \mathtt{MP3}_\cup\},$$

where $\sigma(x) = (1 + e^{-\mu(x-\frac{1}{2})})^{-1}$ is the classic sigmoid function centered in 1/2 and μ is used to adjust the slopeness of the curve; we set $\mu = 10$ in our experimentation. In addition, the sigmoid polarizes the values close to 1/2, thus helping to decide whether they are closer to 1 or 0, therefore, moving the final score closer to $\mathtt{MP3}_\cap$ or $\mathtt{MP3}_\cup$.

While all four measures are available in our implementation, we decided to use $\mathtt{MP3}_\sigma$ as default measure and is denoted simply as MP3. An experimental comparison of all four measures is shown in the Supplementary Materials.

3 Results

3.1 Simulated data

To perform our experiments, we follow an approach similar to the one performed in the study by DiNardo *et al.* (2020). We start from a base tree on which we apply a series of perturbations selected from: label swapping, label removal, label duplication, node swapping and node removal. Both the perturbations and the nodes and labels on which they are applied are chosen at random: our procedure allows to select a user-specified total number of actions and a probability vector that will be used to select the perturbations from the previous list.

For the measure comparison experiments, we generated 30 perturbations from each of the 5 base trees, for a total of 150 trees. For the clustering evaluation, 3 base trees are entirely different from each other, and another 2 are perturbations of two of the others, to simulate similar subfamilies of the same tumor type: we perform a total of 10 perturbations on such 5 trees. More details on the perturbation parameters will be described in each section, while the entire configuration is available and reproducible at https://github.com/AlgoLab/mp3treesim_supp.

3.2 Measures comparison

We compared MP3 against all the different versions of DISC and CASet from the study by DiNardo *et al.* (2020) and MLTD (Karpov *et al.*, 2019). While MP3 and MLTD provide similarity scores, DISC and CASet compute a dissimilarity score, which we convert into a similarity measure by simply subtracting their value from 1.

3.2.1 Effect of changes in the tree topology

A key feature a measure on tumor phylogenies should have is to discern changes at different tree depths; indeed, a change close to the root should be more impactful than a change toward the leaves. Such a behavior is fundamental, as driver mutations are often acquired early in the evolutionary history, while less important passenger mutations usually happen at later stages: to mistake the two types of mutations should therefore have a high impact on a good similarity measure.

To estimate this effect on all the measures, we start from a linear base tree [T_0 in Fig. 4 (*left*)]; we then raise its only leaf one level at the time and compute its similarity to the base tree, expecting a drop in similarity as the leaf raises to the root, similarly to experiment proposed in the study by DiNardo *et al.* (2020). Figure 4 (*left*) clearly displays such effect for MP3, showing that it has the highest

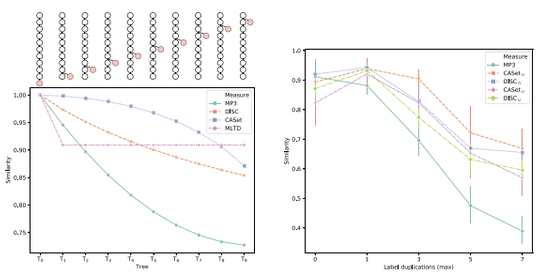

Fig. 4. (*Left*) Effect of a node (highlighted in red) that ascends from leaf to child of the root, T_0 is the base tree to which the others are compared. (*Right*) Effect of label duplication on the similarity scores. Similarities are the average of 15 trees generated from the same base with the specified maximum number of duplications. MLTD was excluded since it failed to run on instances with poly-occurring labels

similarity decrease among all measures; DISC and CASet also have similar trends, but to a lower extent. Since the set of labels is the same for all trees, there is no difference between union and intersection versions of DISC and CASet. Contrarily, as already observed in the study by DiNardo *et al.* (2020), MLTD plateaus after the first change.

Another interesting aspect to investigate is how the presence of poly-occurring labels influences the similarity scores, as the more sophisticated the inference tools get, the more is common to have tumor phylogenies with multiple acquisitions or losses of the same mutation. To evaluate this aspect, we started from a multilabeled base tree with all labels occurring only once. We then created 15 perturbed trees for 5 different configurations. In the first one [on the abscissa 0 in Figure 4(*right*)], we allowed one operation excluding label duplication; for the others, we allowed a total of 1, 3, 5 and 7 operations with much higher chance of selecting a label duplication. Since perturbations occur randomly, we are only sure that at most the specified number of duplication occurred, and not necessarily to the same label.

Figure 4(*right*) shows that $\text{CASet}_{\cap}$, $\text{CASet}_{\cup}$, $\text{DISC}_{\cap}$ and $\text{DISC}_{\cup}$ have similar trends in this setting, MP3 being the only one that differs. In particular, the other measures assign a higher similarity score to the second configuration than to the first one, despite they are both obtained with one perturbing operation, allowing label duplication only in the second one. MP3 is the only measure that positively displays a monotonic decrease in similarity as the number of poly-occurring labels increases, being markedly steeper than the others. We believe that a larger steepness will be more informative than a plateauing curve, since while being true that after many of poly-occurrences no more information is gained, all the duplications will inevitably add more and more noise to the tree. Since MLTD assumes that every label appears only once, it failed to run on this experiment and was therefore excluded.

3.2.2 Results on simulated data

To analyze the differences between all measures, we designed two experimental settings: from 5 different base trees (available in the Supplementary Materials), we generated 30 perturbations for each class and computed similarities scores between all the 150 resulting trees. In the first configuration, we allowed a total of three operations excluding label duplications, while in the second one we allowed them. All the parameters and the different probabilities used for applying perturbations are available at our Supplementary Repository https://github.com/AlgoLab/mp3treesim_supp.

Results for the first configuration are shown in Figure 5. The heatmaps (*left*) show that MP3 discerns the best between the trees in the same class (main diagonal) and the others: the results of $\text{DISC}_{\cup}$ are really close to ours, but there is a more noticeable noise outside the main diagonal. $\text{DISC}_{\cap}$ and $\text{CASet}_{\cup}$ present even more noise than the others, but are still mostly able to distinguish the different classes; $\text{CASet}_{\cap}$ seems to struggle the most on this setting, while MLTD displays high values of similarities for every couple of trees, but it is still able to differentiate between the bases.

The boxplots in Figure 5(*top-right*) show the same result quantitatively: the crucial feature is to correctly distinguish the different classes. The values represent the distribution of the similarities between the trees in the same class (intrasimilarity) and in different classes (intersimilarity). MP3 differentiates better between intra- and intersimilarity, exhibiting the most compact distribution for the intersimilarities scores, while being a little more dispersed on the intrasimilarity due to the action of the sigmoid, that pulls apart the values around 1/2. Similarly to the previous case, $\text{DISC}_{\cup}$, $\text{CASet}_{\cup}$ and MLTD show similar trends, while $\text{CASet}_{\cap}$ displays the most overlapping distributions.

Finally, in Figure 5(*bottom-right*), we computed a silhouette score from the data using a hierarchical linkage clustering with cuts from 2 to 15 to simulate a clustering scenario. Once again, MP3 performs the best expressing the maximum value for 5 cuts, being the five classes. $\text{DISC}_{\cap}$, $\text{DISC}_{\cup}$ also show the largest value at the same cut. MLTD was excluded from the plot since it scored values close to −1 for every cut, thus causing the figure to be hard to interpret.

In the second experimental setting, we introduced poly-occurring labels to the simulation. Figure 6 exhibits results very similar to the previous ones. The main difference is that in the silhouette score (*bottom-right*) MP3, while still having its maximum value in correspondence of five cuts, is slightly lower than the other measures. On this experiment MLTD, not allowing poly-occurring labels, failed to compute the score in most of the instances, shown in gray in the heatmaps (*left*); it was excluded from the other plots given the high amount of failed runs. On the other hand, CASet and DISC accept input trees with poly-occurring labels, but they

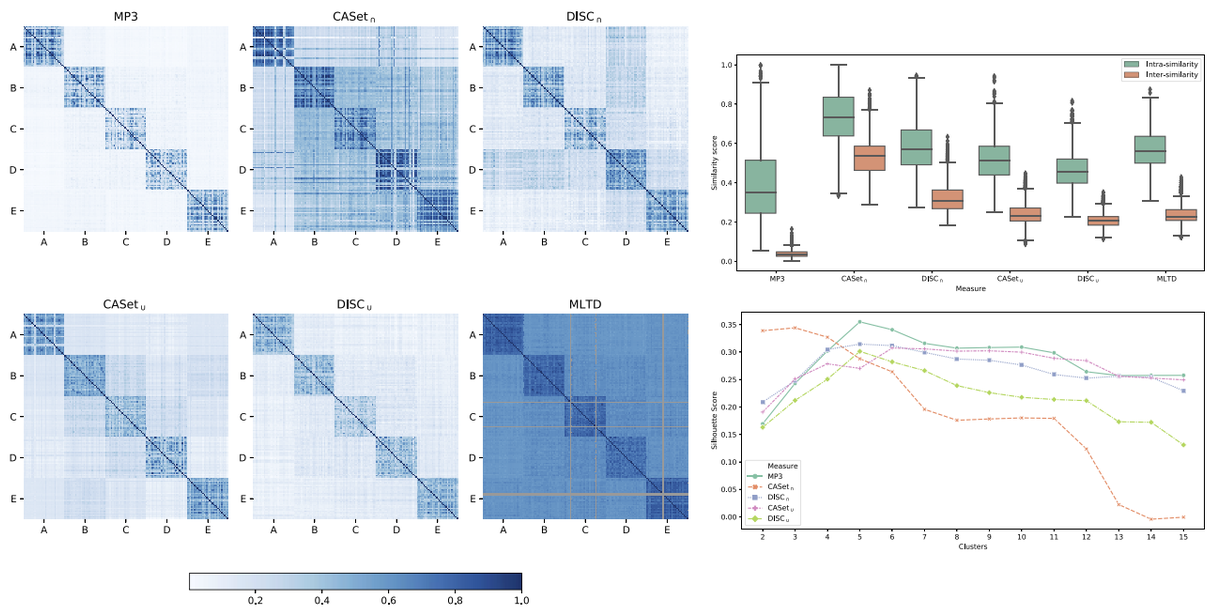

Fig. 5. Results for the first experimental configuration: (*Left*) Heatmaps displaying the scores between all-pairs 150 simulate trees. (*Top-Right*) Distribution of the similarities between the trees in the same class (intrasimilarity) and in different classes (intersimilarity). (*Bottom-Right*) Silhouette score computed using a hierarchical linkage clustering with cuts from 2 to 15

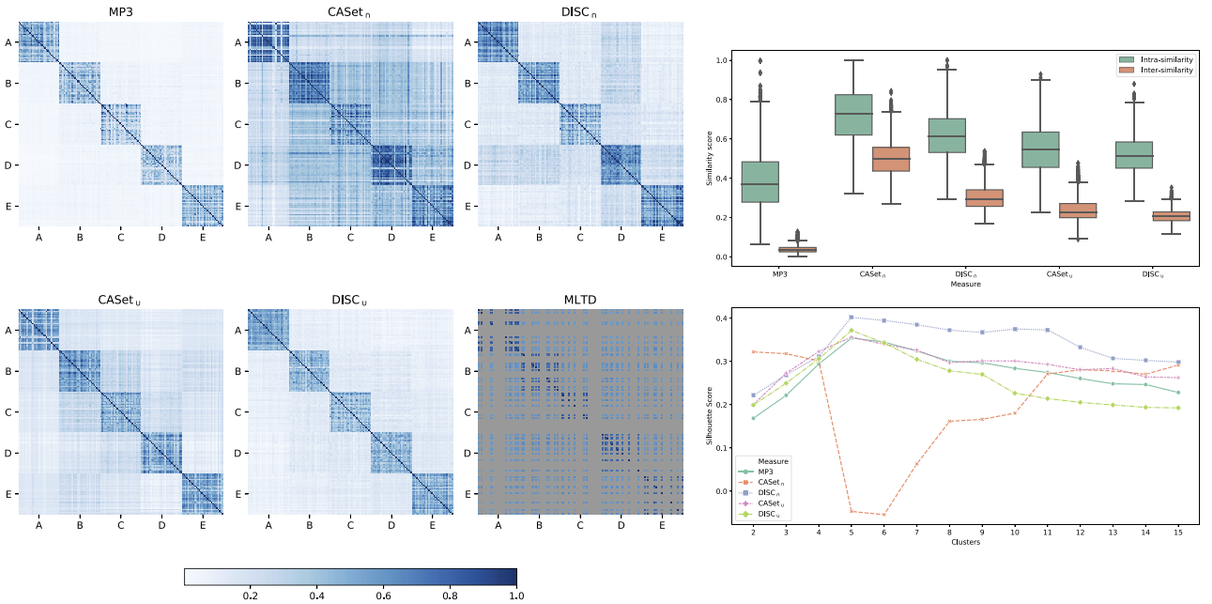

Fig. 6. Results for the second experimental configuration: (*Left*) Heatmaps displaying the scores between all the 150 simulate trees. (*Top-Right*) Distribution of the similarities between the trees in the same class (intrasimilarity) and in different classes (intersimilarity). (*Bottom-Right*) Silhouette score computed using a hierarchical linkage clustering with cuts from 2 to 15

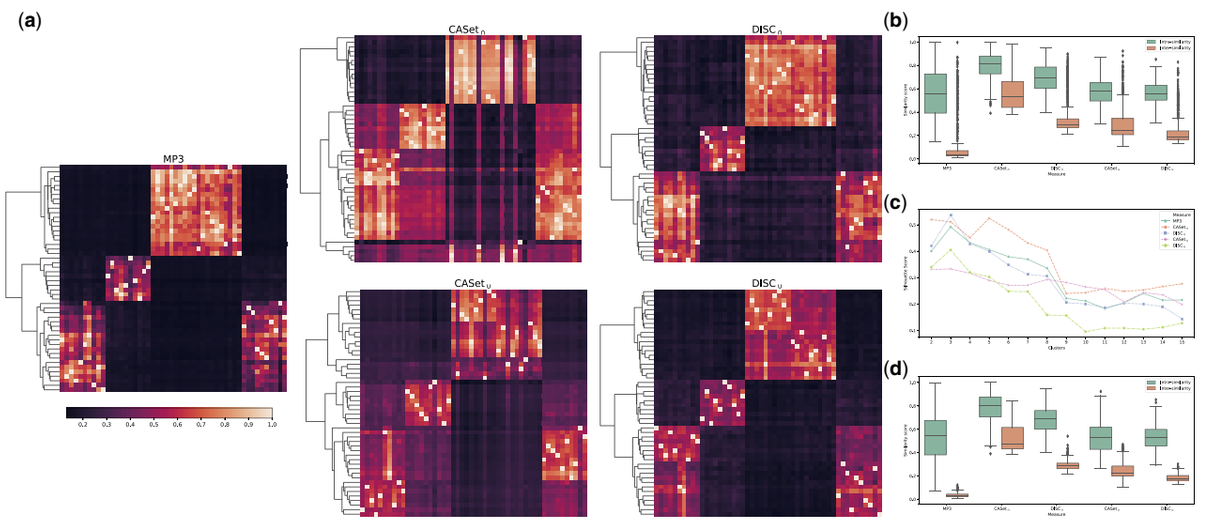

Fig. 7. Results for the clustering experiment: (**a**) Clustermaps of the 50 simulated trees computed using hierarchical linkage clustering. (**b**) Distribution of the similarities between the trees in the same class (intrasimilarity) and in different classes (intersimilarity) for the 5 classes. The high number of outliers for all methods is due to the high similarity of the two subclasses. (**c**) Silhouette score computed using a hierarchical linkage clustering with cuts from 2 to 15. (**d**) Distribution of the similarities between the trees in the same class (intrasimilarity) and in different classes (intersimilarity) for the three main classes, remapping the subclasses to the original corresponding base. MLTD was excluded from this experiment because it failed to run on most instances due to the presence of poly-occurring labels

disregard the multiple occurrences of a same label, considering only one occurrence for each label in the computation.

3.3 Application to clustering of trees

A very important application of a tree similarity measure is clustering, e.g. to classify cancer type of patients by the similarity of their inferred phylogenies. This is of crucial interest for the development of precision therapies based on the topological structure and the evolution of mutations. Since to curate such classifications manually would be unfeasible as the size and the number of mutations increases, a good measure to use in conjunction with a clustering method is necessary.

To evaluate a similar scenario, we started from three different bases, then perturbing two of such trees chosen at random; these new trees are then considered as additional base trees. Given this 5 bases, we created a total of 10 perturbed trees from each class. The goal was to simulate an experiment with three separate classes, with two of them further split in two subclasses, to obtain subtypes of the same cancer families. The five resulting bases are available in the Supplementary Materials, and the parameters used for the simulations are in our Supplementary Repository.

Results for the clustering experiment are reported in Figure 7; Figure 7(a) shows the clustermaps computed using hierarchical linkage clustering. MP3, $DISC_{\cap}$ and $DISC_{\cup}$ correctly cluster the three main families as well as the two subfamilies, while both versions of CASet struggle the most in this experiment. Figure 7(b) displays the distribution of intra- and intersimilarity between the five bases; MP3 has the most compact intersimilarity distribution and is the only method that completely separates intra- and interdistributions. The high number of outliers for all methods is due to the high similarity of the two subclasses. To confirm this hypothesis, we computed the same distributions only for the three main classes, remapping the subclasses to the original corresponding base class in Figure 7(d), where we note that the number of outliers is significantly reduced. Finally, Figure 7(c) shows the silhouette scores for the dataset; all measures express a higher score with three cuts, suggesting that the two subclasses are very similar to the two main bases they are derived from. The scores are very similar for all measures, with $DISC_{\cup}$ having a higher value with three cuts and MP3 having a slightly higher with five clusters. $CASet_{\cap}$ is the only method that have a much higher score in five, however, as shown in Figure 7(a), the five clusters it reports are not the correct ones. MLTD was excluded from this experiment because it failed to run on most instances due to poly-occurring labels.

3.4 Application to real dataset

To further evaluate our similarity measure, we applied it to two publicly available real datasets: breast cancer xenoengraftment in immunodeficient mice (Eirew *et al.*, 2015) and ultradeep-sequencing of clear cell renal-cell carcinoma (Gerlinger *et al.*, 2014). Both datasets were previously considered for analyses by the two cancer phylogeny reconstruction methods LICHeE (Popic *et al.*, 2015) and MIPUP (Husić *et al.*, 2019). Data from the study by Eirew *et al.* (2015) was also used in the study by DiNardo *et al.* (2020) for evaluation. An interesting feature of the data in the study by Gerlinger *et al.* (2014) is that most samples in the study present poly-occurring labels, suggesting recurrent mutations at different evolutionary stages. We recall that DISC and CASet compute dissimilarity scores, that we convert into a similarity measure subtracting their value from 1. All the analyzed trees are available in the Supplementary Materials.

To evaluate the effectiveness of the measures in real scenarios, we selected the manually curated trees, published in the corresponding original sequencing studies, for case SA501 from the study by Eirew *et al.* (2015) and for patient RMH002 from the study by Gerlinger *et al.* (2014). We then computed similarities between these reference trees and the ones inferred by LICHeE and MIPUP, as reported in the study by Husić *et al.* (2019).

The reference RMH002 is very similar to the evolutions inferred by LICHeE and MIPUP, thus most of the measures agree on a high similarity score, as reported in Figure 8(*left*), with the exception of $CASet_{\cup}$. The scores computed by MP3 are higher than the others, possibly because it is the only method to correctly identify and process poly-occurring labels in the reference trees, due to the discovered recurring mutations. Differently from the previous analysis, the measures disagree considerably for SA501, as depicted in Figure 8(*center*). Indeed, MP3 reports a similarity value close to 0, suggesting that the considered trees are quite different, whereas the other measures report a higher similarity, especially DISC scoring up to 60% similarity.

To thoroughly investigate this behavior, we defined some naïve approaches used as a proxy to analyze some basic aspects of the trees, such as the count of pairs of labels appearing in the same node in both trees. Even with such a naïve measure, the reference tree for SA501 from the study by Eirew *et al.* (2015) and the trees inferred by MIPUP and LICHeE disagree considerably. The base tree contains only 50 labels, whereas the trees inferred by LICHeE and MIPUP contain 95 and 158 labels, respectively; of these, the reference shares a total of 24 label with LICHeE and 37 with MIPUP. Most importantly, only 54 out of 1759 pairs of labels appear in the same node both in the reference and LICHeE and 124 out of 8424 in MIPUP. Such evaluations, albeit very simplistic, suggest that the trees are indeed dissimilar and thus a lower score, as provided by MP3, is more reasonable than a high value of similarity.

To better understand this phenomenon, we created the edge case of a single-node tree with all the 158 labels from MIPUP, and compared it against the reference SA501. The resulting values in Figure 8(*right*) show a high similarity score for DISC with values up to 69%, with CASet and MLTD being less influenced by this aspect with scores up to 11 and 20%. On the other hand, MP3 clearly defines the trees as extremely dissimilar, with a score of 0.04%. Such results for trees that are clearly extremely different show a strong bias for DISC toward high similarity values.

	LICHeE	MIPUP
MP3	0.997	0.897
$CASet_{\cap}$	0.805	0.779
$DISC_{\cap}$	0.930	0.876
$CASet_{\cup}$	0.569	0.551
$DISC_{\cup}$	0.764	0.725
MLTD	0.842	0.807

	LICHeE	MIPUP
MP3	0.017	0.004
$CASet_{\cap}$	0.139	0.111
$DISC_{\cap}$	0.627	0.624
$CASet_{\cup}$	0.260	0.113
$DISC_{\cup}$	0.405	0.610
MLTD	0.182	0.205

	Edge case
MP3	0.0004
$CASet_{\cap}$	0.0927
$DISC_{\cap}$	0.5571
$CASet_{\cup}$	0.1120
$DISC_{\cup}$	0.6933
MLTD	0.2046

Fig. 8. (*Left*) Similarities between the manually curated tree reported by Gerlinger *et al.* (2014) for patient RMH002 and the trees inferred by LICHeE and MIPUP. (*Center*) Similarities between the manually curated tree reported by Eirew *et al.* (2015) for sample SA501 and the trees inferred by LICHeE and MIPUP. (*Right*) Similarities between the manually curated tree reported by Eirew *et al.* (2015) for sample SA501 and the edge case with all mutations appearing in a single node

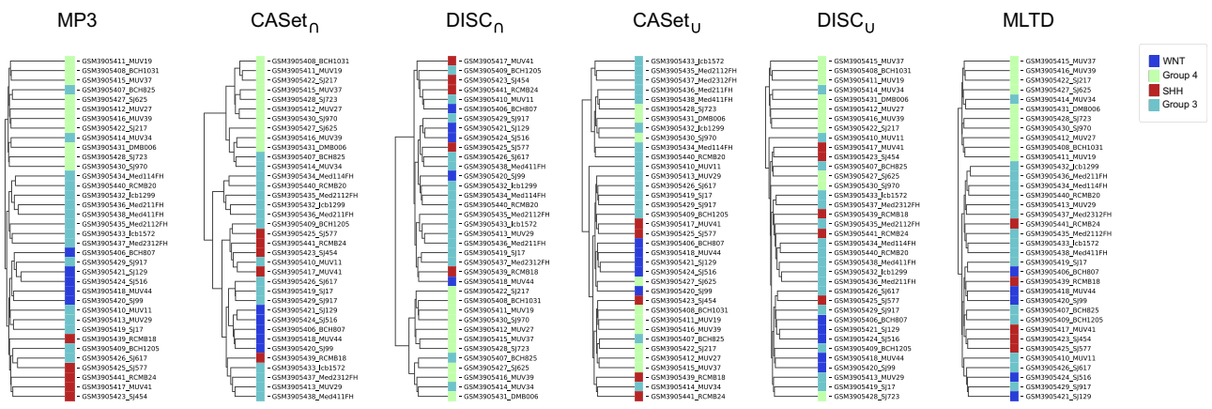

Fig. 9. Results for the clustering experiment: Hierarchical clustering obtained from 36 medulloblastoma patients (Hovestadt *et al.*, 2019). The trees were computed from the available scRNA-seq datasets using the inference tool SCITE (Jahn *et al.*, 2016). The colors indicate the true tumor subtype of each patient

3.5 Application to clustering of patients

As a final evaluation of the measures, we computed a clustering of 36 medulloblastoma patients from the study by Hovestadt *et al.* (2019); the patients are classified according to four different subtypes of tumor. From the available scRNA-seq data, we inferred the cancer phylogeny of each patient using SCITE (Jahn *et al.*, 2016); we then computed the similarities between all the inferred trees and used them to perform a hierarchical clustering.

Figure 9 displays the clustering results for all the measures; in particular, using MP3 is possible to distinctively group the patients in their relative subtypes with only a few mismatched trees. A similar result is achieved by $CASet_{\cap}$, while the other measures tend to cluster together subtypes SHH and WNT, without a clear distinction between them.

4 Discussion

We identified two major limitations in the existing methods to compare tumor phylogenies: first, they are not sensitive enough to detect even major differences in the topology of the trees, as we demonstrated with *ad hoc* experiments. Second, they are not able to meaningfully compare trees where the same label is assigned to more than one node.

We addressed the latter by representing tumor phylogenies as multilabeled trees with poly-occurring labels. Such model is best suited to cancer progression than the ones previously adopted, as it allows the same mutation to appear in multiple evolutionary events, a circumstance often occurring in real applications. Being inspired by the triplet distance for classical phylogenies, our new similarity measure correctly detects differences in the topology of the trees.

Our experiments show that our method performs very well both on synthetic and real data and, unlike the other existing tools, it is able to detect differences regarding poly-occurring labels and it suitably distinguish trees with different topologies. Moreover, when applied to hierarchical clustering, it outperforms every other method.

Acknowledgements

The authors acknowledge Iman Hajirasouliha, Victoria Popic and Murray Patterson for many illuminating discussions on tumor phylogenies.

Funding

This project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [872539].

Conflict of Interest: none declared.

References

Aguse,N. *et al.* (2019) Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. *Bioinformatics*, **35**, i408–i416.

Aho,A.V. *et al.* (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, **10**, 405–421.

Bernardini,G. *et al.* (2019) A rearrangement distance for fully-labelled trees. In: *30th Annual Symposium on Combinatorial Pattern Matching (CPM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. Dagsthl, Germany, pp. 28:1-28:15

Bernardini,G. *et al.* (2020) On two measures of distance between fully-labelled trees. In: Gørtz,I.L. and Weimann,O. (eds.) *31st Annual Symposium on Combinatorial Pattern Matching (CPM 2020)*, volume 161 of *Leibniz International Proceedings in Informatics (LIPIcs)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, pp. 6:1–6:16.

Bonizzoni,P. *et al.*. (2019) Does Relaxing the Infinite Sites Assumption Give Better Tumor Phylogenies? An ILP-Based Comparative Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics, **16**, 1410–1423. 10.1109/TCBB.2018.2865729

Bonizzoni,P. *et al.* (2017) Beyond perfect phylogeny: multisample phylogeny reconstruction via ilp. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Association for Computing Machinery, New York, NY, USA, pp. 1–10.

Brodal,G.S. *et al.* (2013) Efficient algorithms for computing the triplet and quartet distance between trees of arbitrary degree. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, New Orleans, LA, USA, pp. 1814–1832.

Brown,D. *et al.* (2017) Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat. Commun.*, **8**, 14944.

Ciccolella,S. *et al.* (2018a) GPPS: an ilp-based approach for inferring cancer progression with mutation losses from single cell data. In: *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, Las Vegas, NV, 2018, pp. 1–1.

Ciccolella,S. *et al.* (2018b) Inferring cancer progression from single-cell sequencing while allowing mutation losses. *bioRxiv*.

DiNardo,Z. *et al.* (2020) Distance measures for tumor evolutionary trees. *Bioinformatics*, **36**, 2090–2097.

Dobson,A.J. (1975) Comparing the shapes of trees. In: *Combinatorial Mathematics III*. Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, pp. 95–100.

Dudek,B. and Gawrychowski,P. (2019) Computing quartet distance is equivalent to counting 4-cycles. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, Association for Computing Machinery, New York, NY, USA, pp 733–743.

Eirew,P. *et al.* (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, **518**, 422–426.

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

Gerlinger,M. *et al.* (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.*, **46**, 225–233.

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

Hajirasouliha,I. and Raphael,B.J. (2014) *Reconstructing Mutational History in Multiply Sampled Tumors Using Perfect Phylogeny Mixtures*. Lecture Notes in Computer Science. Springer Nature, Berlin, Heidelberg, Germany, pp. 354–367.

Hajirasouliha,I. *et al.* (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, **30**, i78–i86.

Hovestadt,V. *et al.* (2019) Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature*, **572**, 74–79.

Husić,E. *et al.* (2019) 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.

Jansson,J. and Rajaby,R. (2017) A more practical algorithm for the rooted triplet distance. *J. Comput. Biol.*, **24**, 106–126.

Karpov,N. *et al.* (2019) A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression. *Algorithms Mol. Biol.*, **14**, 17.

Kuipers,J. *et al.* (2017) Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.*, **27**, 1885–1894.

Malikic,S. *et al.* (2019a) Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.*, **10**, 2750.

Malikic,S. *et al.* (2019b) Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Res.*, **29**, 1860–1877.

Morrissy,A.S. *et al.* (2016) Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, **529**, 351– 357.

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

Wang,J. *et al.* (2016) Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, **48**, 768– 776.

Yuan,K. *et al.* (2015) Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.*, **16**, 36.

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