The Bourque Distances for Mutation Trees of Cancers

Katharina Jahn
Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Niko Beerenwinkel
Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Louxin Zhang
Department of Mathematics and Computational Biology Program, National University of Singapore, Singapore
matzlx@nus.edu.sg

Abstract

Mutation trees are rooted trees of arbitrary node degree in which each node is labeled with a mutation set. These trees, also referred to as clonal trees, are used in computational oncology to represent the mutational history of tumours. Classical tree metrics such as the popular Robinson–Foulds distance are of limited use for the comparison of mutation trees. One reason is that mutation trees inferred with different methods or for different patients often contain different sets of mutation labels. Here, we generalize the Robinson–Foulds distance into a set of distance metrics called Bourque distances for comparing mutation trees. A connection between the Robinson–Foulds distance and the nearest neighbor interchange distance is also presented.

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases mutation trees, clonal trees, tree distance, phylogenetic trees, tree metric, Robinson–Foulds distance, Bourque distance

Digital Object Identifier 10.4230/LIPIcs.WABI.2020.14

Funding LX Zhang was supported by Singapore Ministry of Education Academic Research Fund Tier-1 [grant R-146-000-238-114].

1 Introduction

Trees have been used in biology to model the evolution of species, genes and cancer cells [15, 32, 39]; to represent the secondary structures of RNA molecules and to classify cell types, to name just a few uses [23, 37]. A fundamental issue arising from these applications of trees is how to quantitatively compare tree models that are inferred by different methods or from different data. A number of tree metrics have been proposed for comparisons, including the Robinson–Foulds (RF) [3, 35, 36], nearest-neighbor interchange (NNI) [31, 35] and triple(t) distances [7] for phylogenetic trees; gene duplication, gene loss and reconciliation costs [17, 27] for gene and species trees; and the tree-edit distances [40, 37, 43] for tree models of secondary RNA structures, etc. [2, 21, 26, 33, 41].

With advances in next-generation sequencing and single-cell sequencing technologies, a large amount of genomic data is now available for identifying tumour subclones and inferring their evolutionary relationships. The most common representation of these relationships are mutation trees, also known as clonal trees, which encode the (partial) temporal order in which mutations were acquired. Formally, a mutation tree on a finite set of mutations Γ is a rooted tree T with k nodes and a partition of Γ into k disjoint non-empty parts P_i so that each P_i is assigned as the label of a node of T [16, 32]. A large number of computational approaches

20th International Workshop on Algorithms in Bioinformatics (WABI 2020).
Editors: Carl Kingsford and Nadia Pisanti; Article No. 14; pp. 14:1–14:22
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

for reconstructing mutation trees from bulk sequencing data [9, 11, 14, 29, 34], single-cell sequencing data [6, 13, 19, 42], or a combination of both [30, 28] have been developed over the last years. Unlike phylogenetic trees, mutation trees inferred with these methods will not only differ in their topology but may also be defined on different sets of mutations. The latter happens in the comparison of methods using different data (e. g. single-cell vs. bulk) or divergent criteria for mutation calling. For that reason, classical tree distance measures are not immediately applicable to mutation trees. Instead novel measures have recently been developed [1, 4, 5, 10, 18, 20], but no standard approach for mutation tree comparison has yet emerged. Instead, shortcomings of some of these measures such as the inability to resolve major differences between trees have recently been demonstrated [5]. Additionally, computing the distances between two mutation trees takes at east quadratic time for each of these measures.

Here, we generalize the Robinson-Foulds metric, a classic distance measure for unrooted trees, for the comparison of mutation trees. This metric is based on the so-called (edge) contraction and decontraction operations introduced by Bourque for leaf-labeled unrooted trees in a study of Steiner trees [3]. A contraction on an edge $(u.v)$ of a tree T is an operation that transforms T into a new tree by shrinking (u, v) into a single node. The decontraction operation is the reverse of contraction. Robinson and Foulds independently adopted the contraction and decontraction to define a metric of unrooted labeled trees, where there is a finite set S and a partition of S into disjoint parts (some of which may be empty) so that nodes with a degree of at most 2 are each labeled with a unique non-empty part, and nodes with a degree of at least 3 are labeled with either a unique non-empty part or an empty part. They defined a metric, now called the Robinson-Foulds (RF) distance, in which the distance between two unrooted labeled trees is the minimum number of contraction or decontraction operations that are necessary to transform one into another [36]. The RF distance is equal to the number of edge-induced partitions that are not shared between the two trees and thus is computable in linear time [8].

Although the RF distance is popular in phylogenetic analysis, it is not robust when applied to the comparison of mutation trees with different sets of mutations, as it is simply equal to the total number of edges in the trees and thus fails to capture any topological similarity between the trees.

In this paper, by generalizing the RF distance, we propose a collection of distance measures to measure the topological dissimilarity between unrooted (resp. rooted) labeled trees with different label sets. We also apply these measures to simulated and real tumour mutation trees. To set our distances apart from another recently introduced generalised RF distance that is based on a node flip operation [4], we refer to our generalisations as *Bourque distances*, as they are closely related to the edge contraction and decontraction operations introduced by Bourque for leaf-labeled unrooted trees [3]. They are also shown to be related to the NNI distance [35]. Unlike previous measures proposed for the comparison of mutation trees, the basic version of the Bourque distance can be computed in linear time.

The rest of this paper is divided into seven sections. Section 2 introduces basic concepts and the notation that will be used. In Section 3, we present a connection between the NNI distance and the RF distance for both phylogenetic and arbitrary trees that are unrooted and labeled. In Section 4, we generalize the RF distance into the Bourque distances for unrooted labeled trees. In Section 5, we define the Bourque distances for mutation trees. In Section 6, we examine the relationships among the distance measures proposed in [10, 20, 5] and the Bourque distances on rooted 7-node trees and on random rooted trees with 30 nodes. In Section 7, we computed the Bourque distances on two sets of mutation trees. Section 8 concludes the study with a few remarks.

2 Concepts and Notation

A (unrooted) *tree* is an acyclic graph. A *rooted tree* is a directed tree with a designated root node ρ in which the edges are oriented away from ρ. There is a unique directed path from ρ to every other node.

For a tree or rooted tree T, the nodes, leaves and edges are denoted $V(T)$, $\text{Leaf}(T)$ and $E(T)$, respectively. Let $u \in V(T)$. The *degree* of u is the number of edges incident to it, where edge orientation is ignored if T is rooted. In a rooted tree, non-root nodes with a degree of one are called the *leaves*; non-leaf nodes are called *internal* nodes. One or more edges may leave an internal node, but exactly one edge enters every node that is not the root. An *internal edge* is an edge between two internal nodes.

Let T be a rooted tree and $u, v \in V(T)$. The node v is called a *child* of u and u is called the *parent* of v if $(u, v) \in E(T)$. The node v is a *descendant* of u and u is an *ancestor* of v if the unique path from the tree root to v contains u. We use $C_T(u)$, $A_T(u)$ and $D_T(u)$ to denote the set of all children, ancestors and descendants of u in T, respectively. Note that u is in neither $A_T(u)$ nor $D_T(u)$.

A *star tree* is a tree that contains only one non-leaf node, which is called the *center* of the tree. A *line tree* is a tree in which every internal node is of degree 2. A rooted line tree is a line tree whose root is of degree 1.

A tree is *binary* if every internal node is of degree 3. A rooted tree is *binary* if the root is of degree 2 and every other internal node is of degree 3. A *caterpillar tree* is a binary tree in which each internal node is adjacent to one or two leaves.

Let X be a finite set. A (rooted) *phylogenetic tree* on X is a binary (rooted) tree where the leaves are uniquely labeled with the elements of X, the taxon set. A (rooted) phylogenetic tree T on X is *labeled* if there is a set I that is disjoint from X and a labeling function $\ell : V(T) \setminus \text{Leaf}(T) \to I$ such that each u of $V(T) \setminus \text{Leaf}(T)$ is labeled with $\ell(u)$. If ℓ is a one-to-one function, T is said to be uniquely labeled. In a labeled phylogenetic tree, the label set for the internal nodes and the taxon set for the leaves are distinct and thus are not interchangeable.

A tree or rooted tree T with n nodes is *labeled* if there is a finite set M and a labeling function $\ell : V(T) \to 2^M$ satisfying $\cup_{v \in V(T)} \ell(v) = M$ and $\ell(v) \neq \emptyset$ for $v \in V(T)$ so that $f(v)$ is assigned as the label of v, where 2^M denotes the collection of subsets of M. Furthermore, if $\ell(v)$ contains exactly one element for each node v, we say T is *1-labeled* with L. Here, M is called the *label set* of T.

A *mutation tree* on a set M of mutations is a rooted labeled tree that has M as the label set, where the labels of different nodes are disjoint.

3 Metrics for labeled trees

For convenience, we will introduce new metrics on the space of 1-labeled trees and then generalize them to the space of mutation trees later.

3.1 Nearest neighbor interchanges on labeled phylogenetic trees

The NNI operation (Fig. 1A) and NNI distance were originally introduced for binary phylogenetic trees [35]. It is known that any binary phylogenetic tree can be transformed into another in $n \log n + 2n - 4$ NNIs at most [24]. The NNI operation for rooted phylogenetic trees is given in Fig. 1B. Since the NNI operation does never interchange the labels of internal nodes and of leaves, Proposition 1 is simple, but as far as we know, it has never appeared in literature.

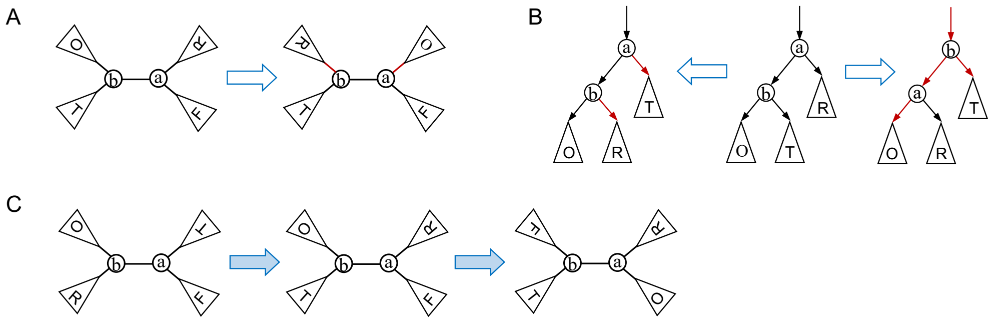

Figure 1 Illustration of the NNI operation on phylogenetic trees. (A) In a phylogenetic tree, an NNI operation on an internal edge (a, b) first selects two edges (a, x) and (b, y) that are, respectively, incident to a and b such that $(a, x) \neq (a, b) \neq (y, b)$; it then rewires them to the opposite end so that (a, y) and (b, x) are the two edges in the resulting tree (red). Since a and b are labeled differently, a unrooted tree can be transformed into one of four possible trees in one NNI. (B) In a rooted phylogenetic tree T, an NNI operation on an internal edge (a, b) (where b is a child of a) transforms T by either (i) selecting two edges (a, x) and (b, y) that leave from a and b, respectively, and replacing them with (a, y) and (b, x) (left), where $x \neq b$, or (ii) selecting an edge (b, y) leaving from b and replacing the unique edge (z, a) that enters a, (a, b) and (b, y) with (z, b), (b, a) and (a, y) (right), respectively. A rooted tree can be transformed into four different trees in one NNI. (C) Illustration of the interchange of two labels of the ends of an internal edge in two NNIs in an 1-labeled phylogenetic tree.

▶ **Proposition 1.** *In the space of binary (resp. rooted) phylogenetic trees where the internal nodes are 1-labeled, any tree can be transformed into another.*

Proof. This follows from the fact that two NNIs on an internal edge (a, b) are enough to exchange the labels of a and b (Fig. 1C). A similar fact is also true for binary rooted phylogenetic trees. ◀

3.2 Generalized NNI on 1-labeled trees

An arbitrary tree with n nodes can have 1 to $n-2$ internal nodes of degree ≥ 2. To transform a tree into any other of the same size with the same label set, we define the generalized NNI (gNNI) operation as follows.

▶ **Definition 2.** *Let T be a 1-labeled tree and $e = (a, b) \in E(T)$. A gNNI on e is an operation that transforms T into a new tree S by (i) selecting a subset C_a and a subset C_b of the edges that are, respectively, incident to a and b such that $e \notin C_a \cup C_b$ and then (ii) replacing each edge (a, x) of C_a with (b, x) and each edge (b, y) of C_b with (a, y).*

The gNNI operation is illustrated in Fig. 2. Note that if we apply a gNNI operation on an edge $e = (a, b)$ to reconnect all the children of a to b while keeping the children of b unmoved, a will become a leaf adjacent to b in the resulting tree. Another difference between gNNI and NNI is that gNNI can be applied to any edge, whereas NNI can only be applied on an internal edge.

Let L be a set of n elements. The gNNI graph $G_{\text{gnni}}(L)$ is defined as a graph in which the nodes are all 1-labeled trees with nodes labeled with L and two trees are connected by an edge if the two trees are one gNNI apart. The diameter of $G_{\text{gnni}}(L)$ is written as $D(G_{\text{gnni}}(L))$. The distance between two trees T' and T'' in the graph is called the *gNNI distance* between them, written as $d_{\text{gnni}}(T', T'')$.

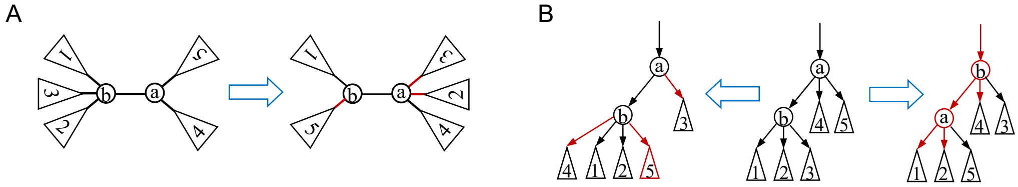

Figure 2 An illustration of the gNNI operation on a labeled tree (A) or a rooted labeled tree (B). A. A gNNI operation on an edge (a, b) interchanges one or more children of a with an arbitrary number of children of b. B. A gNNI operation on an edge (a, b) (where b is the child of a) not only rewires the selected edges leaving a and b (left), but also rewires the unique edge entering a and b simultaneously if necessary (right).

▶ **Proposition 3.** *Let L be a set of n elements. The graph $G_{\text{gnni}}(L)$ has the following properties:*

- $|V(G_{\text{gnni}}(L))| = n^{n-2}$*;*
- $G_{\text{gnni}}(L)$ *is connected;*
- $n - 2 \leq D(G_{\text{gnni}}(L)) \leq 2n - 4$

Proof. The first property is the Cayley formula on the count of 1-labeled trees with n nodes. The second property is a consequence of the third. We prove the third property as follows.

Let $T_1, T_2 \in V(G_{\text{gnni}}(L))$. Let r_1 and r_2 be the two nodes of T_1 and T_2, respectively, that have the same label. Each n-node tree has at least two leaves and therefore $n - 2$ internal nodes at most. By applying a gNNI operation on an edge (r_1, u), we can reconnect all the subtrees that each contain exactly one neighbor of u to r_1, producing a tree in which u becomes a leaf adjacent to r_1. By continuing to apply the gNNI operation on the edges between r_1 and its non-leaf neighbors, we can transform T_1 into the star tree centered at r_1 in $n - 2$ gNNIs at most. In reverse, we can transform the star tree centered at r_2 into T_2 in $n - 2$ gNNIs at most. By combining these two transformations, we transform T_1 into T_2 by using $2n - 4$ gNNIs at most. This proves the upper bound of the third property.

Let S be a line tree where the leaves are labeled with a and b and let T be a 1-labeled star tree centered at the node of the label a. The distances between a and b are $(n - 1)$ and 1 in S and T, respectively. It takes at least $(n - 2)$ gNNIs to transform S to T, as each gNNI can only decrease the distance between a and b by 1. This proves the lower bound of the third property. ◀

Let T be a tree in $G_{\text{gnni}}(L)$. We use $d(u, v)$ to denote the number of edges in the unique path between u and v in T. Any edge $(u, v) \in E(T)$ induces a two-part partition $P(e) = \{P_u, P_v\}$ of L, where $P_u = \{\ell(x) \mid d(x, u) < d(x, v)\}$, which contains u, and $P_v = \{\ell(y) \mid d(y, v) < d(y, u)\}$, which contains v. Let us define $\mathcal{P}(T) = \{P(e) \mid e \in E(T)\}$.

▶ **Proposition 4.** *For any two 1-labeled trees S, T of $G_{\text{gnni}}(L)$,*

$$\frac{1}{2}|\mathcal{P}(S)\Delta\mathcal{P}(T)| \leq d_{\text{gnni}}(S, T) < |\mathcal{P}(S)\Delta\mathcal{P}(T)|,$$

where Δ is the set symmetric difference operator.

Proof. Let S and T have n nodes in the tree space. The first inequality is derived from the following two facts:

- $\mathcal{P}(S) \setminus \mathcal{P}(T)$ contains exactly one partition $P(e)$ if T is obtained from S by applying a gNNI on any $e \in E(S)$;
- $A\Delta B \subseteq (A\Delta C) \cup (C\Delta B)$ for any three sets.

To prove the upper bound, we let $m = |\mathcal{P}(S) \cap \mathcal{P}(T)|$ and let

$$\mathcal{P}(S) \cap \mathcal{P}(T) = \{P(e'_1), P(e'_2), \cdots P(e'_m)\} = \{P(e''_1), P(e''_2), \cdots P(e''_m)\},$$

where $e'_i \in E(S), e''_i \in E(T)$ such that $P(e'_i) = P(e''_i)$ for each i. $S - \{e'_i | 1 \leq i \leq m\}$ is the disjoint union of $m+1$ subtrees S_j $(0 \leq j \leq m)$; similarly, $T - \{e''_i | 1 \leq i \leq m\}$ is the disjoint union of $m+1$ subtrees T_i $(0 \leq i \leq m)$. Additionally, for each $0 \leq j \leq m$, a unique index $k(j)$ exists such that S_j and $T_{k(j)}$ contain the same number (say o_i) of nodes. Note that

$$|\mathcal{P}(S) \Delta \mathcal{P}(T)| + 2m = |E(S)| + |E(T)| = 2n - 2. \tag{1}$$

There are three possible cases for each pair of subtrees S_j and $T_{k(j)}$. First, if $o_j = 1$, we do not need to do any local adjustments of S_j to transform S to T.

If both S_j and $T_{k(j)}$ contain two nodes u and v, (u, v) is then the only edge of S_j and $T_{k(j)}$. This implies that the two nodes are the ends of different edges of $\mathcal{P}(S) \cap \mathcal{P}(T)$ in S and T, and thus we need one gNNI to switch these two nodes in S so that they are incident to the same edges as in T after the operation.

If both S_j and $T_{k(j)}$ contain o_j (≥ 3) nodes, we select an internal node s of S_j and a node t of $T_{k(j)}$ such that s and t have the same label. By continuing to apply, at most, $o_j - 3$ gNNIs on the edges incident to s, we can transform S_j into a star tree C centered on s, as s is an internal node. Similarly, by applying $o_j - 2$ gNNIs at most, we can transform C into $T_{k(j)}$. Taken together, the two transformations give a transformation from S_j into $T_{k(j)}$ consisting of at most $2o_j - 5$ gNNIs at most.

Let m_i be the number of subtrees S_j such that $|S_j| = i$ for $i = 1, 2$ and let m_3 be the number of subtrees S_j such that $|S_j| \geq 3$. We have that $m_1 + m_2 + m_3 = m + 1$ and there are $n - m_1 - 2m_2$ nodes in the union of all subtrees S_j in Case 3. By combining all the transformations from S_j to $T_{k(j)}$, we can transform S to T in c gNNIs at most, where:

$$\begin{aligned} c &= 0 + m_2 + [2(n - m_1 - 2m_2) - 5m_3] \\ &= 2n - 2m_1 - 3m_2 - 3m_3 - 2m_3 \\ &= 2n - 2m_1 - 3m_2 - 3m_3 - 2(m + 1 - m_1 - m_2) \\ &= 2n - 2m - 2 - m_2 - 3m_3 \end{aligned}$$

Since $m_2 \geq 0$ and $m_3 \geq 0$, by Eqn. (1), $c \leq 2n - 2m - 2 = |\mathcal{P}(S) \Delta \mathcal{P}(T)|$. ◀

3.3 The RF distance

Let S and T be two 1-labeled trees. $|\mathcal{P}(S) \Delta \mathcal{P}(T)|$ is called the *RF distance* between S and T, denoted $\mathrm{RF}(S, T)$. For example, in the left tree given in Fig. 3A, the edge $(2, 4)$ (bold) induces the two-part partition $\{\{1, 2, 3\}, \{4, 5, 6, 7, 8\}\}$; the edge $(7, 8)$ (bold) induces $\{\{7\}, \{1, 2, 3, 4, 5, 6, 8\}\}$. These two partitions are not equal to any edge-induced partition in the right tree. Similarly, we have that the two-part partitions induced by the edges $(2, 4)$ and $(7, 8)$ in the right tree are not found in the left tree. One can also verify that the other five edge-induced partitions in both trees are identical. Hence, the RF distance between the left and right trees is 4.

Like the phylogenetic tree case, it is easy to see that the RF satisfies the non-negativity, symmetry and triangle inequality conditions.

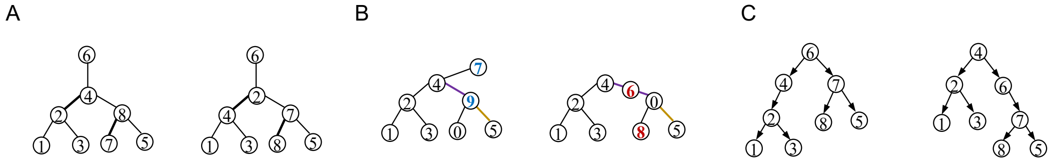

Figure 3 An illustration of the RF distance and the Bourque distance. **A.** Two unrooted 1-labeled trees. The RF distance between them is 4, as in the left tree, the edges $(2, 4)$ and $(7, 8)$ induces two partitions that are not found in the right tree and vice versa. **B.** The labels 0–5 are the labels appearing in the two trees. The Bourque distance between them is 9 (see the main text for details). **C.** The two labeled trees are rooted at different nodes. The RF distance between the left tree and the right tree is 2, as the partitions induced by $(6, 4)$ of Tree A and $(4, 6)$ of Tree B are different.

4 Generalizations of the RF distance for labeled trees

Let us consider labeled trees of different sizes or whose label sets are not the same (see the mutation trees studied in Section 7). The RF distance between any pair of such trees is simply equal to the total number of edges in the trees and thus fails to capture their dis-similarity. Here, we propose generalizations of the RF distance for measuring the dis-similarity of such trees better.

4.1 Bourque distances

For a labeled tree S, we use $\mathcal{L}(S)$ to denote the label set of S. Since each node of $V(S)$ is labeled with a non-empty subset of $\mathcal{L}(S)$, each edge $e = (u, v)$ induces the two-part partition $P(e) = \{L(u), L(v)\}$, where $L(u) = \cup_{x \in V(S): d(x,u) < d(x,v)} \ell(x)$ and $L(v) = \cup_{y \in V(S): d(y,v) < d(y,u)} \ell(y)$.

Let T be another labeled tree such that $C \triangleq \mathcal{L}(S) \cap \mathcal{L}(T) \neq \emptyset$. For $e' \in E(S)$ and $e'' \in E(T)$, we assume that the two-part partitions induced by e' and e'' are $P(e') = \{X, \mathcal{L}(S) \setminus X\}$ and $P(e'') = \{Y, \mathcal{L}(T) \setminus Y\}$, respectively, where $X \subset \mathcal{L}(S)$ and $Y \subset \mathcal{L}(T)$. $P(e')$ and $P(e'')$ are said to be *similar* if the following conditions are satisfied:

- $P(e') \neq P(e'')$;
- $X \cap C \neq \emptyset$ and $(\mathcal{L}(S) \setminus X) \cap C \neq \emptyset$;
- $\{X \cap C, (\mathcal{L}(S) \setminus X) \cap C\} = \{Y \cap C, (\mathcal{L}(T) \setminus Y) \cap C\}$.

We use $\sim$ to denote the similarity relationship of the edge-induced partitions of two trees. Note that the similarity relation is a many-to-many relation in the product space of edge-induced partitions $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(T)$.

▶ **Definition 5.** *Let S and T be two labeled trees and let $\mathcal{P}$ be the set of two-part partitions of $\mathcal{L}(S) \cap \mathcal{L}(T)$. The Bourque metric $B(S, T)$ between S and T is defined as:*

$$B(S,T) \triangleq |\mathcal{P}(T) \Delta \mathcal{P}(S)| - \sum_{P \in \mathcal{P}} \min \left(|\{Q' \in \mathcal{P}(S) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T) : Q'' \sim P\}| \right). \tag{2}$$

The intuition behind this definition is that we "correct" the RF distance by those partitions, that would be shared between both trees when labels unique to either of the two trees were ignored. For example, in Fig. 3B, the labels $\{7, 9\}$ that appear in the left tree are not found in the right tree, whereas the labels $\{6, 8\}$ that appear in the right tree are not found in the left tree. Therefore, none of the seven edge-induced partitions in either tree is found in the other. This implies that the RF distance between the two trees is 14. Since the labels appearing in both trees are $\{1, 2, 3, 4, 5\}$, the edge $(4, 9)$ (purple) of the left tree induces the same partition, $\{\{1,2,3,4\},\{0,5\}\}$ of $\{0,1,2,3,4,5\}$ as the edges $(4,6)$ and $(6,0)$ (purple) of the right tree. Furthermore, the edge $(1,2)$ (resp. $(2,3)$ and $(2,4)$) induces the same partition of $\{0,1,2,3,4,5\}$ in both trees; and the edge $(9,5)$ of the left tree induces the same partition of $\{1,2,3,4,5\}$ as the edge $(0,5)$ of the right tree. Therefore, the Bourque distance between both trees is $14-5=9$.

▶ **Proposition 6.** *Let S and T be two labeled trees with s and t nodes, respectively.*

(i) *If $\mathcal{L}(S)=\mathcal{L}(T)$, $2\times|s-t|\leq B(S,T)=\mathrm{RF}(S,T)$.*

(ii) *If $\mathcal{L}(S)\neq\mathcal{L}(T)$, $\max(s,t)-1\leq B(S,T)\leq\mathrm{RF}(S,T)=s+t-2$.*

(iii) *If $\mathcal{L}(S)\cap\mathcal{L}(T)=\emptyset$, $B(S,T)=\mathrm{RF}(S,T)=s+t-2$.*

(iv) *The Bourque metric is a distance metric; in other words, it satisfies the non-negativity, symmetry and triangle inequality conditions.*

Proof. The full proof appears in the Appendix. ◀

▶ **Proposition 7.** *The Bourque distance between two labeled trees S and T can be computed in linear time $O(|\mathcal{L}(S)|+\mathcal{L}(T)|$.*

Proof. We assume node labels are integers (otherwise, we apply hashing to convert the labels into integers). By indexing labels with integers and fill a hash table, we can determine the set C of node labels that are in both trees. We then remove all labels that are not in C from the two trees, as well as nodes that lost all labels in the process, so that the resulting trees have only nodes with labels from C. Lastly, we apply the algorithm developed by Day [8] for the RF distance to compute the negative term of Eqn. (2) in linear time.

The first term of Eqn. (2) is the RF distance and can thus be found in linear time. ◀

4.2 High-order Bourque distances

Like the RF distance, the Bourque distance has the tendency to overpenalize certain labeling differences and can saturate quickly. Thus it is not refined enough for some applications. In this subsection, we will use the Bourque distances between local subtrees and a matching algorithm ([2, 26, 33]) to define new distance metrics. The new metrics will take more values than the basic version.

Let T be a labeled tree and $u\in V(T)$. For an integer $k>0$, the k-star subtree $N_k(u)$ centered at u is defined as the subtree induced by the vertex set $\{v\in V(T)\mid d(u,v)\leq k\}$ in T. For any pair of labeled trees S and T of n and n' nodes, respectively, such that $n\geq n'$, define $\mathrm{BG}_k(S,T)$ as the weighted complete bipartite graph with two node parts $\{N_k(x)\ :\ x\in V(S)\}$ and $\{\emptyset_1,\cdots\emptyset_{n-n'}\}\cup\{N_k(y)\ :\ y\in V(T)\}$, where each $\emptyset_i$ is just the empty graph; the Bourque distance $B(N_k(x),N_k(y))$ is assigned to the edge $(N_k(x),N_k(y))$ as a weight for every $x\in V(S)$ and $y\in V(T)$ and $|N_k(x)|-1$ is assigned to the edge $(N_k(x),\emptyset_i)$ as a weight for any $\emptyset_i$. Although $N_k(x)$ can be identical for different nodes x, $\mathrm{BG}_k(S,T)$ always has $2n$ nodes.

▶ **Definition 8.** *Let S and T be two labeled trees. The k-Bourque distance $B_k(S,T)$ is defined as the minimum weight of a perfect matching in $\mathrm{BG}_k(S,T)$, $k\geq 1$.*

▶ **Proposition 9.** *The k-Bourque distances have the following properties:*

(1) *For any uniquely labeled trees S and T such that $|V(S)|=|V(T)|=n$, $B_k(S,T)=n\cdot B(S,T)$ for any $k\geq\max(\mathrm{diam}(S),\mathrm{diam}(T))$, where $\mathrm{diam}(X)$ is the diameter of X for $X=S,T$.*

(2) *$B_k(S,T)$ satisfies the non-negativity, symmetry and triangle inequality conditions for each $k\geq 1$.*

Proof. The full proof appears in the Appendix. ◀

▶ Remark. The run time of computing the k-Bourque distance for two labeled trees S and T with m and n nodes, respectively, is $O(\max(m,n)^3)$, as computing the Bourque distances between the k-star trees centered at tree nodes takes $O(\max(m,n)^2)$ in the worst case and computing the minimum weight perfect matching in $\mathrm{BG}_k(S,T)$ takes $O(\max(m,n)^3)$ time.

5 The Bourque distances for mutation trees

In this section, we will describe how to generalize the gNNI and Bourque distances to rooted labeled trees.

5.1 The gNNI

To transform a binary rooted phylogenetic tree into another in which the root is labeled differently, we add the so-called rotation operation that allows two nodes u and v that are connected by an edge to interchange not only one of their children but also their positions (right, Fig. 1B)[25]. A gNNI on a directed edge (a,b) of a rooted tree rewires some outgoing edges from a to b and vice versa and/or rewires the incoming edges to both a and b simultaneously (right, Figure 2B). More precisely, the gNNI is defined on rooted labeled trees as follows:

▶ **Definition 10.** *Let T be a rooted labeled tree and $e=(a,b)\in E(T)$ (where b is a child of a). An NNI operation on e transforms T by selecting a subset of edges $C_a=\{(a,x)\}$ that leave a, where $(a,b)\notin C_a$, and a subset of edges $C_b=\{(b,y)\}$ that leave b and then either (i) replacing each edge (a,x) of C_a with (b,x) and each edge (b,y) of C_b with (a,y) (left, Figure 2B) or (ii) rewiring the edges in C_a and C_b as in (i) as well as replacing the unique edge (z,a) that enters a and (a,b) with (z,b) and (b,a), respectively (right, Figure 2B).*

It is easy to see that for any pair of arbitrary labeled trees S and T, S can be transformed into T through a series of gNNIs as long as the labels appearing in the two trees are the same.

5.2 The RF and Bourque distances

In a rooted labeled tree, each directed edge also induces a 2-part partition on the label set. Therefore, the RF distance is well defined even for rooted trees that may not be uniquely labeled.

Let T be a rooted labeled tree. Recall that $\mathcal{L}(T)$ denotes the set of labels appearing in T. For a non-root node $u\in V(T)$, we use $L_T(u)$ to denote the set of the labels of u and its descendants. The unique edge entering u induces then an "ordered" two-part partition $(L_T(u),\mathcal{L}(T)\setminus L_T(u))$, which is an ordered pair of the two complementary subsets of $\mathcal{L}(T)$. Since the root of a rooted tree is a distinct node of the tree, we assume that the root is contained in the second part of an edge-induced partition. Hence, two edge-induced ordered partitions P' and P'' are *equal* if and only if the first part of P' is equal to the first component of P'' and the second part of P' is equal to the second component of P''. This is particularly useful when comparing two rooted trees with different roots. Let us define $\mathcal{OP}(T)$ to be the set of all edge-induced ordered partitions of T.

▶ **Definition 11.** *For two rooted labeled trees S and T, the RF distance $\mathrm{RF}(S,T)$ between S and T is defined as $|\mathcal{OP}(T)\ \Delta\ \mathcal{OP}(T)|$.*

For example, the two trees given in Figure 3C are obtained from rooting a unrooted labeled tree in different nodes. Only the partition induced by the edge $(6, 4)$ of the left tree is not found in the right tree. Conversely, the partition induced by the edge $(4, 6)$ in the right tree is not found in the left tree. Hence, the distance between these two trees is 2.

▶ **Proposition 12.** *Let S and T be two rooted labeled trees of equal size that have the same labels.*

(1) *Let $t \in V(T)$ such that it has the same label as the root r_S of S and let r_T be the root of T. We have that $RF(S, T) \geq 2d$, where d is the distance between r_T and t.*

(2) $\frac{1}{2}RF(S, T) \leq d_{\text{gnni}}(S, T) \leq RF(S, T)$.

Proof.

(1) Let the path between r_T and t be $r_T = t_0, t_1, t_2, \cdots, t_d = t$. All label sets $L_T(t_i)$ contain the label $\ell(r_S)$. However, only $L_T(t_0)$ is an element of $\{L_S(u) \mid u \in V(S)\}$. Furthermore, since both trees have the same number of nodes and edges, at least d subsets of $\{L_S(u) \mid u \in V(S)\}$ are not found in $\{L_T(v) \mid v \in V(T)\}$. Hence, $RF(S, T) \geq 2d$.

(2) The proof is similar to that of Proposition 4. ◀

Similarly, we can generate the similarity relationship of edge-induced partitions. For two non-root nodes $u \in V(S)$ and $v \in V(T)$, the ordered partitioned induced by the edges entering u and v are *similar* if and only if $(L_S(u), \mathcal{L}(S) \setminus L_S(u)) \neq (L_T(v), \mathcal{L}(T) \setminus L_T(v))$ but they are equal when restricted on $\mathcal{L}(S) \cap \mathcal{L}(T)$, denoted $(L_S(u), \mathcal{L}(S) \setminus L_S(u)) \sim (L_T(v), \mathcal{L}(T) \setminus L_T(v))$.

▶ **Definition 13.** *The Bourque distance $B(S, T)$ between two rooted labeled trees S and T is defined to be:*

$$|\mathcal{OP}(S)\Delta\mathcal{OP}(T)| - \sum_{P\in\mathcal{P}} \min(|\{(P' \in \mathcal{OP}(S)\ :\ P' \sim P\}|, |\{(P'' \in \mathcal{OP}(T)\ :\ P'' \sim P\}|). \qquad (3)$$

▶ **Proposition 14.** *The Bourque distance between two mutation trees S and T can be computed in linear time $O(|\mathcal{L}(S)| + \mathcal{L}(T)|$.*

The proof is analogous to Proposition 7, but we only count partitions that have no root label(s) in their first part.

5.3 High-order Bourque distances

Let S and T be two rooted labeled trees and $k \geq 1$. Set $L_T^{(k)}(u) = \{\ell(v) \in L_T(u) \mid d_T(u, v) \leq k\}$. By Proposition 12.1, we naturally define the k-Bourque distance $B_k(S, T)$ to be the minimum weight of a perfect matching in the complete bipartite graph $G_k(S, T)$. Here, if we assume $|V(S)| \leq |V(T)|$, $G_k(S, T)$ has the vertex set $\{\emptyset_i, L^{(k)}(s) \mid 1 \leq i \leq |V(T)| - |V(S)|;\ s \in S\} \cup \{L^{(k)}(t) \mid t \in T\}$ and the edge set $\{\emptyset_i, L^{(k)}(s) \mid s \in S\} \times \{L^{(k)}(t) \mid t \in T\}$, together with the weight function $B(x, y)$, where each $\emptyset_i$ is a copy of the empty graph.

6 Comparison of eight distance measures on rooted labeled trees

In this section, we compare the *Bourque distance* (BD) against the *1-Bourque distance* (1-BD), the *2-Bourque distance* (2-BD) and five previously published distance measures: *Common Ancestor Set* (CASet) [18], *Distinctly Inherited Set Comparison* (DISC) [18], an *Ancestor Difference measure* (AD) [18], a Triplet-based Distance (TD) [5] and the *Multi-Labeled Tree Dissimilarity* (MLTD) measure [20]. A detailed description of these measures is given in the Appendix. The gNNI distance is not included in this comparison, as there is no known method for its efficient computation.

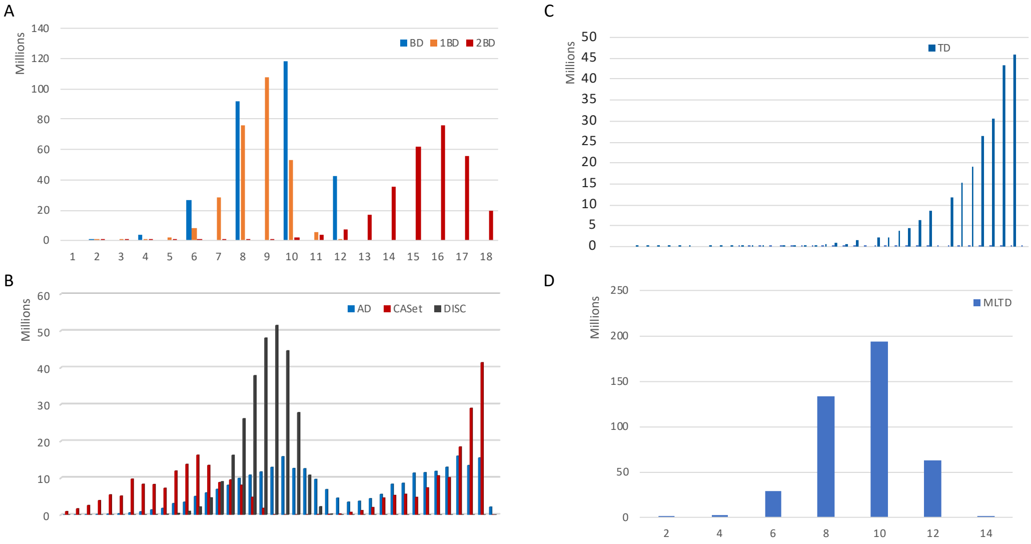

Figure 4 The frequency distribution of pairwise distances for the BD, 1-BD and 2-BD metrics (A), the AD, CASet and DISC measures (B), the TD measure (C) and the MLTD measure (D) in the space of rooted 1-labeled trees with 7 nodes. BD: Bourque distance; AD: Ancestor distance; CASet: Common Ancestor Set distance; DISC: Distinctly Inherited Set; TD: Triplet-based distance. MLTD: Multi-label tree distance.

6.1 Frequency distribution of the pair-wise distances in different metrics

There are $16,807$ unrooted and $7 \times 16,807$ rooted 1-labeled trees with seven nodes. Let R denote the set of such trees and let R_i denote the set of those rooted at Node i, where $1 \leq i \leq 7$. Let d be a distance function of rooted labeled trees. Clearly, for any i, $\{d(x,y) : x \in R_i, y \in R_i\} = \{d(x,y) : x \in R_1, y \in R_1\}$; for different nodes i and j, $\{d(x,y) : x \in R_i, y \in R_j\} = \{d(x,y) : x \in R_1, y \in R_2\}$. Therefore, we computed the pairwise Bourque distance (DB), 1-BD and 2-BD metrics between any $x \in R_1$ and any $y \in R_1 \cup R_2$ such that $x \neq y$. The frequency distributions of the three metrics are given in Fig. 4A, showing a Poisson distribution as the RF in the unrooted case [38].

The pairwise distances of AD, CASet, DISC and TD range from 0 to 1. We computed all the pair-wise distances for all possible pairs of distinct $x \in R_1$ and $y \in R_1 \cup R_2$. Because of the huge number of pair-wise distances, we binned them into 40 intervals $\left(\frac{i}{40}, \frac{i+1}{40}\right)$, $0 \leq i \leq 39$. The histograms for the frequency distributions of the pairwise distance values for the three measures are given in Fig. 4B. The AD and CASet measures have a similar distribution (blue and red in Fig. 4B), each having two peaks. The pairwise distances between trees rooted at the same node form the first peak, whereas the pairwise distances between trees rooted at different nodes form the second peak. These facts show that AD and CASet are sensitive to the root node. The frequency distribution (black) of the DISC measure appears to be again a kind of Poisson distribution. Whether the pairwise distances of the DISC, 1-BD and 2-BD between all 1-labeled trees with a given number of nodes follow a Poisson distribution or not needs further mathematical investigation. The bottom line is that the DISC measure and the Bourque metrics have different distributions of pairwise distances from the AD and DISC measures.

The frequency distribution of the TD is clearly different from the AD, CASet and DISC (Fig. 4C). More than 60% of the pairwise distances are greater than 0.9. For the discrete MLTD measure, we observe a Poisson-like distribution similar to the BD metric.

Lastly, for each of the AD, CASet, DISC, TD and MLTD measures, there are many pairs of trees with the same distance value, that have distinct distances in the BD metric. Figure S1 give an example for each.

6.2 Pairwise distances between random trees

We compared the BD, 1-BD, 2-BD, AD, CASet, DISC, TD and MLTD measures on rooted 1-labeled, 30-node trees that were randomly generated as follows. The tree generator first generated a random unrooted 1-labeled 30-node tree T_0 and then generated 20,000 random unrooted 1-labeled, 30-node trees in 400 iterations. In the i-th iteration, a tree generated in the $(i-1)$-th iteration was randomly selected. Next, five random trees were generated from the selected tree by applying a random NNI on an edge $e = (u, v)$ that was randomly generated, where u was an internal node. Here, a NNI just switched one subtree from the u side to v and one subtree from the v side to u if v was not a leaf and just moved a subtree from u to v if v was a leaf.

We computed the eight different distance values between T_0 rooted at Node 1 and the 20,000 trees rooted at Node 1, which are summarized in Fig. 5. This produced two interesting findings. First, the BD distances from T_0 to the random trees range from 0 to 58; the BD, 1-BD and 2-BD correlate well with each other, particularly when the Bourque distances ranged from 0 to 35. However, the distances between a pair of trees can be very different in the three metrics. For example, there are 3367 random trees that are 46 BD away from T_0. The 1-BDs between T_0 and the trees are from 32 to 45 (top left panel, Figure 5).

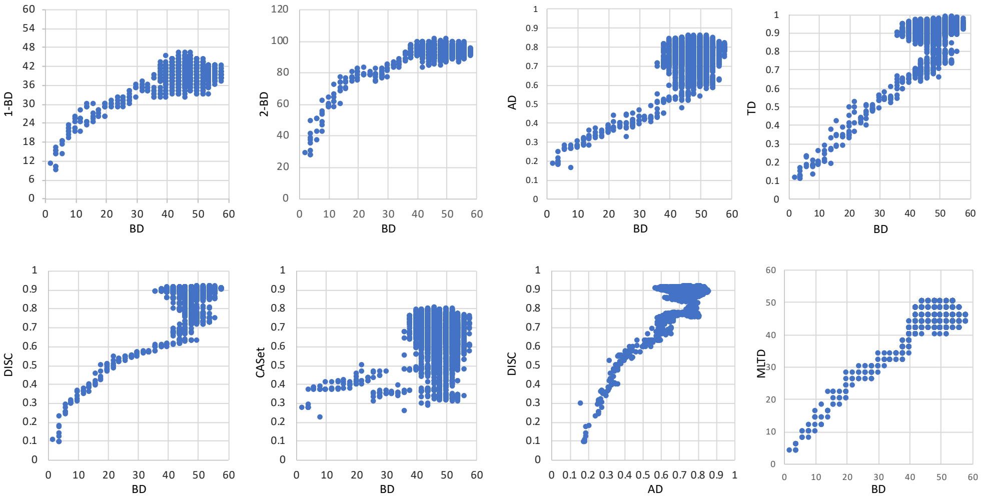

Figure 5 The scatter plots of the Bourque vs the other distance measures between a rooted 1-labeled tree and 20,000 random trees of 30 nodes. BD: Bourque distance; AD: Ancestor distance; CASet: Common Ancestor Set distance; DISC: Distinctly Inherited Set; MLTD: Multi-label tree distance; TD: Triplet-based distance.

Second, AD, DISC, MLTD and TD correlated with BD (and hence 1-BD and 2-BD) surprisingly well with Pearson correlation coefficients (PCC) from 0.38 to 0.543 even though they are defined differently. However, CASet and BD poorly correlated (middle panel, second row) with PCC 0.112.

7 Applications to mutation trees

7.1 The distances between three leukemia mutation trees

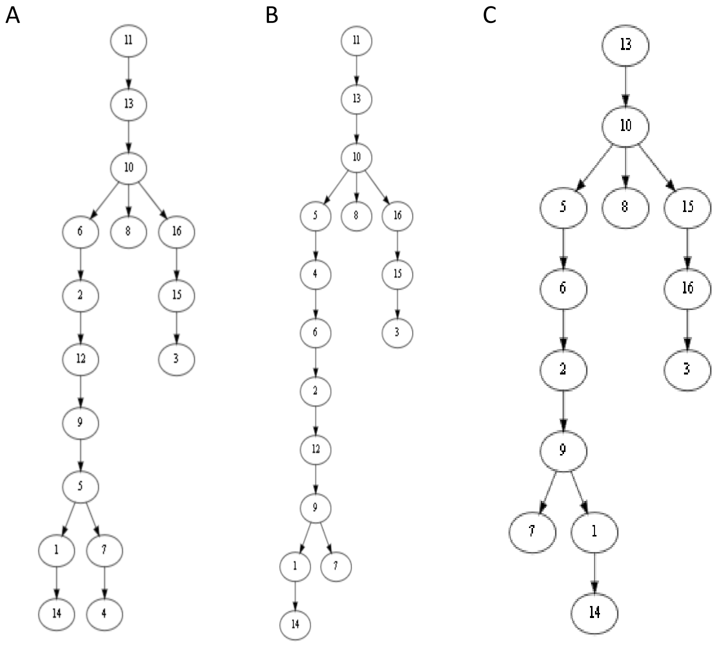

Figure 6 The mutation trees inferred by SCITE [19] (A), B-SCITE [28] (B) and PhISCS [30] (C) from single-cell sequencing data or with the bulk sequencing data for Patient 2 with childhood acute lymphoblastic leukemia that was reported in [16]. The mutation trees contain 16 mutated genes: *ATRNL1* (1), *BDNF_AS* (2), *BRD7P3* (3), *CMTM8* (4), *FAM105A* (5), *FGD4* (6), *INHA* (7), *LINXC00052* (8), *PCDH7* (9), *PLEC* (10), *RIMS2* (11), *RRP8* (12), *SIGLEC10* (13), *TRRAP* (14), *XPO7* (15), *ZC3H3* (16).

Single-cell sequencing data are prone to errors. Mutation trees inferred by different methods from the single-cell sequencing data of a patient are often different in both topology and labels of mutated genes. Fig. 6 shows mutation trees inferred by SCITE [19], B-SCITE [28] and PhISCS [30] for Patient 2, who had childhood acute lymphoblastic leukemia from [16]. Both the SCITE and B-SCITE trees (i.e. Tree A and Tree B) contain 16 mutations, whereas the PhISCS tree (i.e. Tree C) contains just 13 of the 16 mutations.

The pairwise distances between the trees were calculated using the eight distance measures (Table 1). Tree A and Tree B contain the same mutations. The difference between them is mainly the positions of Mutation 4 and Mutation 5 in the long chain on the left. The pairwise distance between them has the smallest value among the three trees for each of the eight measures. Tree B and Tree C have the same topology and are different only in that Mutations 4, 11 and 12 are missing in the latter. For each measure, the distance between Tree B and Tree C is smaller than or nearly equal to the distance between Tree A and Tree

Table 1 Pairwise distances between three mutation trees A, B, and C in Fig. 6 according to different metrics. The union extension of CASet and DISC were used to measure the difference between Tree A (or Tree B) and Tree C [10].

	A & B	A & C	B & C
BD	12	14	14
1-BD	9	26	23
2-BD	28	40	33
MLTD	4	7	5
CASet	0.1079	0.5495	0.5302
DISC	0.2394	0.4135	0.3468
AD	0.1699	0.5499	0.5276
TD	0.3607	0.6393	0.5821

C, consistent with intuition.

7.2 Distances between four simulated mutation trees

Figure 7 presents four simulated mutation trees downloaded from the OncoLib database for which the CASet and DISC disagreed significantly [10]. The pairwise distances between the four trees are given in Table 2. Note that the CASet and DISC distances between T_5 and T_{20} and between T_{14} and T_{26} are different from those reported in [10]. This is because a mutation appearing in a tree node is not an ancestor of another mutation in the same node in our distance calculation. Regardless of the differences between the definitions, our distance computing also shows the disagreement between the CASet and DISC distances. For example, the CASet distance between T_5 and T_{20} is four times as large as the CASet distance between T_{14} and T_{26}, whereas the DISC distance between the former is smaller than the DISC distance between the latter. This disagreement is also observed on the tree pairs $\{T_5, T_{14}\}$ and $\{T_{20}, T_{26}\}$.

Since these four different trees have only one internal edge, the Bourque distance between any two of them is 2. The pairwise 1-BD distances are not much different. However, their differences are reflected in the pairwise 2-BD distances.

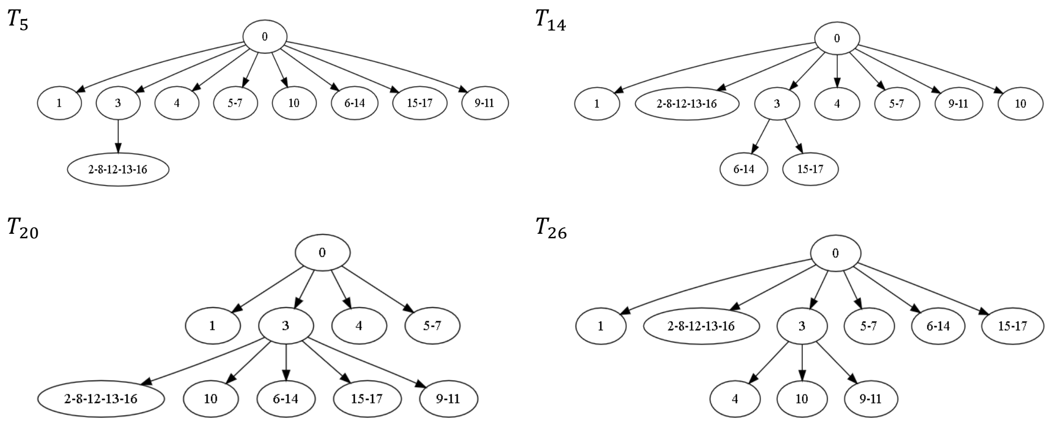

Figure 7 Four simulated mutation trees T_5, T_{14}, T_{20} and T_{26} from the OncoLib database [12].

Table 2 Pairwise distances between trees in Fig. 7 according to the eight distance measures.

	T_5 & T_{14}	T_5 & T_{20}	T_5 & T_{26}	T_{14} & T_{20}	T_{14} & T_{26}	T_{20} & T_{26}
BD	2	2	2	2	2	2
1-BD	11	12	12	12	13	12
2-BD	4	6	5	7	7	8
MLTD	6	10	6	14	10	12
CASet	0.0523	0.1830	0.0523	0.1961	0.0392	0.2157
DISC	0.3807	0.2402	0.3807	0.2483	0.3529	0.3039
AD	0.2500	0.1944	0.2500	0.2222	0.2222	0.2778
TD	0.1961	0.4363	0.2120	0.4669	0.2659	0.4951

8 Conclusions

We have introduced the Bourque and k-Bourque metrics for both unrooted labeled trees and mutation trees. These distances are the generalizations of the RF distance. We demonstrate, through a simulation, that they correlate with the CASet, DISC and AD distance measures for similar trees, but have different distributions of pairwise distances on between all 1-labeled trees with a fixed number of nodes. The advantages of the Bourque metric over CASet and DISC include that it satisfies the triangle inequality and it is computable in linear time. The k-Bourque metrics refine the Bourque metric.

Another contribution is a novel connection between the RF and gNNI metrics on labeled trees. A few theoretical questions arise from this connection between the RF and gNNI and related contributions. Is finding the gNNI distance for labeled trees NP-complete? What is the maximum value of the NNI distance between two binary 1-labeled trees? Can the RF distance be used to define a polynomial time algorithm with approximation ratio < 2 for the gNNi distance?

General mathematical questions also arise from the development of new metrics for comparisons of mutation trees. One is investigating mathematical relationships between the proposed metrics. Another is determining the distributions of pairwise distances between all the 1-labeled trees of the same size. For example, is the distribution Poisson for the Bourque metrics?

Finally, further generalisations of the Bourque distance will be interesting to study in the future, in particular for mutation trees where labels may occur multiple times in different nodes [5]. The motivation for this generalisation comes from the observation that in tumours the same mutations can happen independently in multiple subclones and can also be lost again over time [22].

References

1 Giulia Bernardini, Paola Bonizzoni, and Paweł Gawrychowski. On two measures of distance between fully-labelled trees. *arXiv preprint arXiv:2002.05600*, 2020.

2 Damian Bogdanowicz and Krzysztof Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160, 2011.

3 Michel Bourque. *Arbes de Steiner et réseaux dont certains sommets sont à localisation variable.* PhD thesis, Thèse (Ph. D.: Informatique)–Université de Montréal, 1978.

4 Samuel Briand, Christophe Dessimoz, Nadia El-Mabrouk, Manuel Lafond, and Gabriela Lobinska. A generalized robinson-foulds distance for labeled trees. In *Proceedings of APBC*, 2020.

5 Simone Ciccolella, Giulia Bernardini, Luca Denti, Paol Bonizzoni, Marco Previtali, and Gianluca Della Vedova. Triplet-based similarity score for fully multi-labeled trees with poly-occurring labels. *bioRxiv*, 2020.

6 Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. Inferring cancer progression from single cell sequencing while allowing loss of mutations. *bioRxiv*, page 268243, 2018.

7 Douglas E Critchlow, Dennis K Pearl, and Chunlin Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334, 1996.

8 William HE Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2(1):7–28, 1985.

9 Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):1–20, 2015.

10 Zach DiNardo, Kiran Tomlinson, Anna Ritz, and Layla Oesper. Distance measures for tumor evolutionary trees. *Bioinformatics*, 36(7):2090–2097, 2020.

11 Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.

12 Mohammed El-Kebir. Oncolib: Library for tumor heterogeneity. *GitHub repository*, 2018.

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

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

15 Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

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

17 Morris Goodman, John Czelusniak, G William Moore, Alejo E Romero-Herrera, and Genji Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 1979.

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

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

20 Nikolai Karpov, Salem Malikic, Md Khaledur Rahman, and S Cenk Sahinalp. A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression. *Algorithms for Molecular Biology*, 14(1):17, 2019.

21 Michelle Kendall and Caroline Colijn. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*, 33(10):2735–2743, 2016.

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

23 Shu-Yun Le, Ruth Nussinov, and Jacob V Maizel. Tree graphs of rna secondary structures and their comparisons. *Computers and Biomedical Research*, 22(5):461–473, 1989.

24 Ming Li, John Tromp, and Louxin Zhang. On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, 182(4):463–467, 1996.

25 Ming Li and Louxin Zhang. Twist–rotation transformations of binary trees and arithmetic expressions. *Journal of Algorithms*, 32(2):155–166, 1999.

26 Yu Lin, Vaibhav Rajan, and Bernard ME Moret. A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1014–1022, 2011.

27 Wayne P Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

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

29 Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.

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

31 G William Moore, M Goodman, and J Barnabas. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology*, 38(3):423–457, 1973.

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

33 Tom MW Nye, Pietro Lio, and Walter R Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22(1):117–119, 2006.

34 Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, 16(1):91, 2015.

35 David F Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119, 1971.

36 David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

37 Bruce A Shapiro and Kaizhong Zhang. Comparing multiple rna secondary structures using tree comparisons. *Bioinformatics*, 6(4):309–318, 1990.

38 Mike Steel and David Penny. Distributions of tree comparison metrics—some new results. *Systematic Biology*, 42:126–141, 1993.

39 Y Tateno, M Nei, and Tajima F. Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, 18:387–404, 1982.

40 Gabriel Valiente. *Algorithms on Trees and Graphs*, volume 2. Springer, New York, USA, 2013.

41 WT Williams and HT Clifford. On the comparison of two classifications of the same set of elements. *Taxon*, 20:519–522, 1971.

42 H Zafar, N Navin, K Chen, and L Nakhleh. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Research*, 29:1847–1859, 2019.

43 K Zhang and D Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18:1245–1262, 1989.

A Appendix: Proofs of Propositions 4 and 6

A.1 Proposition 4

Proposition 4. Let S and T be two labeled trees with s and t nodes, respectively.

(i) If $\mathcal{L}(S) = \mathcal{L}(T)$, $2 \times |s - t| \leq B(S,T) = \mathrm{RF}(S,T)$.

(ii) If $\mathcal{L}(S) \neq \mathcal{L}(T)$, $\max(s,t) - 1 \leq B(S,T) \leq \mathrm{RF}(S,T) = s + t - 2$.

(iii) If $\mathcal{L}(S) \cap \mathcal{L}(T) = \emptyset$, $B(S,T) = \mathrm{RF}(S,T) = s + t - 2$.

(iv) The Bourque metric is a distance metric; in other words, it satisfies the non-negativity, symmetry and the triangle inequality conditions.

Proof.

(i) Since the second term of (2) is non-positive, $B(S,T) \leq |\mathcal{P}(S) \Delta \mathcal{P}(T)| = \mathrm{RF}(S,T)$. Without loss of generality, we may assume $s \geq t$. By the definition of the similarity relation, $\mathcal{L}(S) = \mathcal{L}(T)$ implies that $\{(P,Q) \in \mathcal{P}(S) \times \mathcal{P}(T) \ : \ P \sim Q\} = \emptyset$ and thus $B(S,T) = RF(S,T) = 2|\mathcal{P}(S) \setminus \mathcal{P}(T)| \geq 2(s-t)$, proving the inequality.

(ii) If $\mathcal{L}(S) \neq \mathcal{L}(T)$, $|\mathcal{P}(T) \Delta \mathcal{P}(S)| = |\mathcal{P}(T)| + |\mathcal{P}(S)| = s + t - 2$. Moreover, we have:

$$\begin{aligned} & \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(S) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T) : Q'' \sim P\}|\right) \\ \leq \quad & \min\left(\sum_{P \in \mathcal{P}} |\{Q' \in \mathcal{P}(S) : Q' \sim P\}|, \sum_{P \in \mathcal{P}} |\{Q'' \in \mathcal{P}(T) : Q'' \sim P\}|\right) \\ \leq \quad & \min(|\mathcal{P}(T)|, |\mathcal{P}(T)|) = \min(s,t) - 1 \end{aligned}$$

and:

$$B(S,T) = s + t - 2 - (\min(s,t) - 1) = \max(s,t) - 1.$$

(iii) If $\mathcal{L}(S) = \mathcal{L}(T)$, the first term becomes $|\mathcal{P}(S)| + |\mathcal{P}(T)|$, which is $s + t - 2$; and the second term is zero. Therefore, the fact is true.

(iv) The non-negativity follows from (i) and (ii). The symmetric property of the Bourque metric follows from the definition of the Bourque distance. The triangle inequality is proved as follows.

Let T_1, T_2 and T_3 be three labeled trees. We consider the following three cases to prove $B(T_1, T_2) \leq B(T_1, T_3) + B(T_3, T_2)$.

Case 1. $\mathcal{L}(T_1) = \mathcal{L}(T_3) = \mathcal{L}(T_2)$. In this case, $B(T_i, T_j) = \mathrm{RF}(T_i, T_j)$. The triangle inequality for these three trees follows from the fact that the RF distance satisfies the triangle inequality.

Case 2. $\mathcal{L}(T_1) \neq \mathcal{L}(T_3)$ and $\mathcal{L}(T_3) \neq \mathcal{L}(T_2)$.

We have $B(T_1, T_2) \leq |\mathcal{P}(T_1) \Delta \mathcal{P}(T_2)| = |\mathcal{P}(T_1) \setminus \mathcal{P}(T_2)| + |\mathcal{P}(T_2) \setminus \mathcal{P}(T_1)|$.

On the other hand, by (ii), $\mathcal{L}(T_i) \neq \mathcal{L}(T_3)$ implies that $B(T_i, T_3) \geq \max\left(|\mathcal{P}(T_i)|, |\mathcal{P}(T_3)|\right)$ for $i = 1, 2$. Therefore,

$$\begin{aligned} B(T_1, T_3) + B(T_1, T_3) \quad & \geq \quad \max\left(|\mathcal{P}(T_1)|, |\mathcal{P}(T_3)|\right) + \max\left(|\mathcal{P}(T_2)|, |\mathcal{P}(T_3)|\right) \\ & \geq \quad |\mathcal{P}(T_1)| + |\mathcal{P}(T_2)| \geq B(T_1, T_2). \end{aligned}$$

Case 3. $\mathcal{L}(T_1) = \mathcal{L}(T_3) \neq \mathcal{L}(T_2)$ or $\mathcal{L}(T_1) \neq \mathcal{L}(T_3) = \mathcal{L}(T_2)$.

Note that the two conditions are symmetric. Hence, we just need to prove that the triangle inequality holds if the first condition is satisfied.

Let $\mathcal{P}$ be the set of 2-part partitions of $\mathcal{L}(T_1) \cap \mathcal{L}(T_2)$. Since $\mathcal{L}(T_1) = \mathcal{L}(T_3)$, $B(T_1, T_3) = |\mathcal{P}(T_1) \Delta \mathcal{P}(T_3)| = |\mathcal{P}(T_1)| + |\mathcal{P}(T_3)| - 2|\mathcal{P}(T_3) \cap \mathcal{P}(T_1)|$. Since $\mathcal{L}(T_1) \neq \mathcal{L}(T_2)$,

$$\begin{aligned} & B(T_1, T_2) \\ = \quad & |\mathcal{P}(T_1)| + |\mathcal{P}(T_2)| - \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right). \qquad (A.1) \end{aligned}$$

Similarly, since $\mathcal{L}(T_2) \neq \mathcal{L}(T_3)$,

$$\begin{aligned} & B(T_1, T_3) + B(T_3, T_2) \\ = \;& |\mathcal{P}(T_1)| + 2|\mathcal{P}(T_3) \setminus \mathcal{P}(T_1)| + |\mathcal{P}(T_2)| \\ & - \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right). \end{aligned} \tag{A.2}$$

Since

$$\begin{aligned} & \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ \leq \;& \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) \setminus \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ & + \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) \cap \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ \leq \;& \sum_{P \in \mathcal{P}} |\{Q' \in \mathcal{P}(T_3) \setminus \mathcal{P}(T_1) : Q' \sim P\}| \\ & + \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) \cap \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ \leq \;& |\mathcal{P}(T_3) \setminus \mathcal{P}(T_1)| + \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right), \end{aligned}$$

by Eqn. (A.1) and (A.2),

$$\begin{aligned} & B(T_1, T_3) + B(T_3, T_2) \\ = \;& |\mathcal{P}(T_1)| + 2|\mathcal{P}(T_3) \setminus \mathcal{P}(T_1)| + |\mathcal{P}(T_2)| \\ & - \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_3) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ \geq \;& |\mathcal{P}(T_1)| + |\mathcal{P}(T_2)| - \sum_{P \in \mathcal{P}} \min\left(|\{Q' \in \mathcal{P}(T_1) : Q' \sim P\}|, |\{Q'' \in \mathcal{P}(T_2) : Q'' \sim P\}|\right) \\ = \;& B(T_1, T_2). \end{aligned}$$

The triangle inequality is proved. ◀

A2. Proposition 6

Proposition 6. The k-Bourque distances have the following properties:

(1) For any uniquely labeled trees S and T such that $|V(S)| = |V(T)| = n$, $B_k(S, T) = n \cdot B(S, T)$ for any $k \geq \max(\text{diam}(S), \text{diam}(T))$, where $\text{diam}(X)$ is the diameter of X for $X = S, T$.

(2) $B_k(S, T)$ satisfies the non-negativity, symmetry and triangle inequality conditions for each $k \geq 1$.

Proof.

(1) If $k \geq \max(\text{diam}(S), \text{diam}(T))$, $N_k(u) = S$ for any $u \in V(S)$ and $N_k(v) = T$ for any $v \in V(T)$. This implies that every edge has the same weight $B(S, T)$ and every perfect matching has a weight of $n \times B(S, T)$ in the graph $\text{BG}_k(S, T)$.

(2) Obviously, $B_k(S, T)$ has the non-negativity and symmetry properties for each k. Let S, T and W be three labeled trees. We assume that $s = |V(S)| \geq |V(T)| = t$ and consider three cases to prove that $B_k(S, T) \leq B_k(S, W) + B_k(W, T)$ for $k \geq 1$.

Case 1. $w = |V(W)| \geq s$. Let us assume that:

$$f : \{N_k(v),\ \emptyset_i : v \in V(S), 1 \leq i \leq w - s\} \to \{N_k(v) : v \in V(W)\}$$

is a 1-to-1 function such that $\{(N_k(v), f(N_k(v))), (\emptyset_j, f(\emptyset_i) \ : \ v \in V(S), 1 \le i \le w - s\}$ is the minimum weight perfect matching in $\mathrm{BG}_k(S, W)$. Let us also assume that:

$$g : \{N_k(v) : v \in V(W)\} \to \{N_k(v),\ \emptyset_i \ : \ v \in V(T), 1 \le i \le w - t\}$$

is a 1-to-1 function such that $\{(N_k(v), g(N_k(v))) \ : \ v \in V(W)\}$ is the minimum weight perfect matching in $\mathrm{BG}_k(W, T)$.
We now define the following:

$$W_{00} = \{v \in V(W) \ : \ N_k(v) = f(\emptyset) \ \&\ g(N_k(v)) = \emptyset\},$$

$$W_{01} = \{v \in V(W) \ : \ N_k(v) = f(\emptyset) \ \&\ g(N_k(v)) \neq \emptyset\},$$

$$W_{10} = \{v \in V(W) \ : \ N_k(v) \neq f(\emptyset) \ \&\ g(N_k(v)) = \emptyset\},$$

$$W_{11} = \{v \in V(W) \ : \ N_k(v) \neq f(\emptyset) \ \&\ g(N_k(v)) \neq \emptyset\}.$$

Clearly, $|W_{10}| + W_{11}| = s$, $|W_{01}| + W_{11}| = t$ and thus $|W_{10}| - |W_{01}| = s - t$.
Let $W_{10} = \{a_1, a_2, \cdots, a_{k'}\}$ and $W_{01} = \{b_1, \cdots, b_k\}$, where $k' = k + s - t$. We then have:

$$\begin{aligned} \{(f^{-1}(N_k(v)), g(N_k(v))) \ : \ v \in W_{11}\} \quad &\cup \quad \{(f^{-1}(N_k(a_i)), g(N_k(b_i))) \ : \ 1 \le i \le k\} \\ &\cup \quad \{(f^{-1}(N_k(a_j)), \emptyset) \ : \ k < j \le k'\} \end{aligned}$$

is a perfect matching in $\mathrm{BG}_k(S, T)$ and its weight is:

$$\begin{aligned} C \quad = \quad & \sum_{v \in W_{11}} B\left(f^{-1}(N_k(v)), g(N_k(v))\right) + \sum_{1 \le i \le k} B\left(f^{-1}(N_k(a_i)), g(N_k(b_i))\right) \\ & + \sum_{k+1 \le i \le k'} B\left(f^{-1}(N_k(a_j)), \emptyset\right) \end{aligned}$$

Since the BD satisfies the triangle inequality (Proposition 6),

$$\begin{aligned} & B\left(f^{-1}(N_k(a_i)), g(N_k(b_i))\right) \\ \le \quad & B\left(f^{-1}(N_k(a_i)), N_k(a_i)\right) + B(N_k(a_i), \emptyset) + B(\emptyset, N_k(b_i)) + B\left(N_k(b_i), g(N_k(b_i))\right), 1 \le i \le k. \end{aligned}$$

$$\begin{aligned} C \quad \le \quad & \sum_{v \in W_{11}} \left[B\left(f^{-1}(N_k(v)), N_k(v)\right) + B(N_k(v), g(N_k(v)))\right] \\ & + \sum_{1 \le i \le k} \left[B\left(f^{-1}(N_k(a_i)), N_k(a_i)\right) + B(N_k(a_i), \emptyset) + B(\emptyset, N_k(b_i)) + B(N_k(b_i), g(N_k(b_i)))\right] \\ & + \sum_{k+1 \le i \le k'} \left[B\left(f^{-1}(N_k(a_j)), N_k(v)\right) + B(N_k(v), \emptyset)\right] \\ \le \quad & \sum_{v \in V(W)} B\left(f^{-1}(N_k(v)), N_k(v)\right) + \sum_{v \in V(W)} B(N_k(v), g(N_k(v))) \\ = \quad & B_k(S, W) + B_k(W, T). \end{aligned}$$

By definition, $B_k(S, T) \le C$, implying the triangle inequality.

Case 2. $t \ge w$.
Let us assume that

$$f : \{N_k(v) : v \in V(S)\} \to \{N_k(v),\ \emptyset_i : v \in V(W), 1 \le i \le s - w\}$$

is a 1-to-1 function such that $\{(N_k(v), f(N_k(v))) \ : \ v \in V(S)\}$ is a minimum weight perfect matching in $\mathrm{BG}_k(S, W)$, and assume that

$$g : \{N_k(v),\ \emptyset_i : v \in V(W), 1 \le i \le t - w\} \to \{N_k(v) : v \in V(T)\}$$

is a 1-to-1 function such that $\{(N_k(v), g(N_k(v))), (\emptyset_i, g(\emptyset_i)) \ : \ v \in V(W), 1 \leq i \leq t-w\}$ is the minimum weight perfect matching in $\mathrm{BG}_k(W,T)$. Then,

$$\begin{aligned} &\big\{\big(f^{-1}(N_k(v)), g(N_k(v))\big) \ : \ v \in V(W)\big\} \cup \big\{\big(f^{-1}(\emptyset_i), g(\emptyset_i)\big) \ : \ 1 \leq i \leq t-w\big\} \\ &\cup \big\{\big(f^{-1}(\emptyset_j), \emptyset_{j-t+w}\big) \ : \ t-w < j \leq s-w\big\} \end{aligned}$$

defines a perfect matching in $BG_k(S,T)$ and its weight C can be bounded by:

$$\begin{aligned} C &\leq \sum_{v \in V(W)} \left[B_k\left(f^{-1}(N_k(v)), N_k(v)\right) + B_k(N_k(v), g(N_k(v)))\right] \\ &\quad \sum_{1 \leq i \leq t-w} \left[B_k\left(f^{-1}(\emptyset_i), \emptyset_i\right) + B_k(\emptyset_i, g(\emptyset_i))\right] + \sum_{t-w<j\leq t-w} B_k(f^{-1}(\emptyset_i), \emptyset_{j-t+w}) \\ &= B_k(S,W) + B_k(W,T). \end{aligned}$$

Case 3. $s > w > t$. Let us assume that

$$f : \{N_k(v) : v \in V(S)\} \to \{N_k(v), \emptyset_i : v \in V(W), 1 \leq i \leq s-w\}$$

is a 1-to-1 function such that $\{(N_k(v), f(N_k(v))) \ : \ v \in V(S)\}$ is the minimum weight perfect matching in $\mathrm{BG}_k(S,W)$, and assume that

$$g : \{N_k(v) : v \in V(W)\} \to \{N_k(v), \ \emptyset_i \ : \ v \in V(T), 1 \leq i \leq w-t\}$$

is a 1-to-1 function such that $\{(N_k(v), g(N_k(v))) \ : \ v \in W\}$ is the minimum weight perfect matching in $\mathrm{BG}_k(W,T)$. Then,

$$\{(f^{-1}(N_k(v)), g(N_k(v))), (f^{-1}(\emptyset_j), \emptyset_j) \ : \ v \in V(W), 1 < j \leq s-w\}$$

is a perfect matching in $\mathrm{BG}_k(S,T)$ and its weight is:

$$\begin{aligned} C &= \sum_{v \in V(W)} B(f^{-1}(N_k(v)), g(N_k(v))) + \sum_{1 \leq i \leq s-w} B(f^{-1}(\emptyset_j), \emptyset_j)) \\ &\leq \sum_{v \in V(W)} \left[B(f^{-1}(N_k(v)), N_k(v)) + B(N_k(v), g(N_k(v)))\right] + \sum_{1 \leq i \leq s-w} B(f^{-1}(\emptyset_j), \emptyset_j)) \\ &\leq B_k(S,W) + B_k(W,T), \end{aligned}$$

where the inequality is derived from the triangle inequality. ◀

A3. Measures for comparing mutation trees

The CASet and DISC metrics

Recently, two metrics were introduced for mutation trees [18]. Let M be a label set and T be a rooted tree in which the nodes are uniquely labeled with the parts of a partitions of M. For a node $u \in V(T)$, we use $\ell(u)$ to denote the label of u. For each $m \in M$, we use $\ell^{-}(m)$ to denote the unique node whose label contains m.

Recall that $A_T(u)$ denotes the set of ancestors of u and $u \not\in A_T(u)$. For any $m \in M$, define $A_T(m) = \sum_{u \in A_T(\ell^{-}(m))} \ell(u)$. Note that $A_T(m') \cap A_T(m'')$ is equal to the set of their common ancestors for any m' and m'' of M.

Let S and T be two rooted labeled trees S and T whose nodes are uniquely labeled with the elements of M. The *Common Ancestor Set* (CASet) metric between S and T is defined as the average the Jaccard distance between the sets of common ancestors of two labels in S and T [18], i.e.,

$$\mathrm{CASet}(S,T) \triangleq \frac{1}{\binom{m}{2}} \sum_{i,j \in M : i<j} \frac{|(A_S(i) \cap A_S(j)) \triangle (A_T(i) \cap A_T(j))|}{|(A_S(i) \cap A_S(j)) \cup (A_T(i) \cap A_T(j))|},$$

where $A_S(i)$ is the empty set if i is not in the label set of S or it is an element of the label of the root of S. Here, the Jaccard distance between the empty set and itself is 0.

We use $D_S(i,j)$ to denote $A_S(i) \setminus A_S(j)$ for any two labels. The *Distinctly Inherited Set Comparison* (DISC) metric between S and T is defined to be [18]:

$$\text{DISC}(S,T) \triangleq \frac{1}{m(m-1)} \sum_{i,j\in M: i\neq j} \frac{|D_S(i,j) \triangle D_T(i,j)|}{|D_S(i,j) \cup D_T(i,j)|}.$$

In a mutation tree, the nodes are labeled with disjoint subsets of the label set; a label appearing in a tree may not appear in another tree inferred for the same patient. It is not hard to generalize the CASet and DISC in the context of mutation trees [18].

An ancestor difference metric

One reason to introduce the Bourque distance is that every uniquely labeled tree can be uniquely reconstructed from all its node-induced star subtrees. It is not hard to see that every rooted uniquely labeled tree can also be reconstructed from the paths from the root to all other nodes. Hence, the difference between two mutation trees on M can be measured by the Ancestor Difference (AD) metric defined by:

$$\text{AD}(S,T) \triangleq \sum_{m\in M} |A_S(m) \triangle A_T(m)|.$$

The AD metric has been used for comparing mutation trees in [18, 19]. Note that CASet, DISC and AD metrics do not satisfy the triangle inequality in general.

The triplet-based distance

The triplet distance has also been generalized to mutation trees [5]. In a mutation tree, any three labeled nodes induce a labeled tree that has three labeled nodes at most. The triplet-based distance (TD) between two mutation trees S and T with the same label set is defined by:

$$\text{TD}(S,T) = 1 - \frac{|\text{Triplets}(S) \cap \text{Triplets}(T)|}{\max\left(|\text{Triplets}(S)|, |\text{Triplets}(T)|\right)},$$

where Triplets(S) denotes the set of possible subtrees induced by three different labels.

A4. Supplementary Figure S1

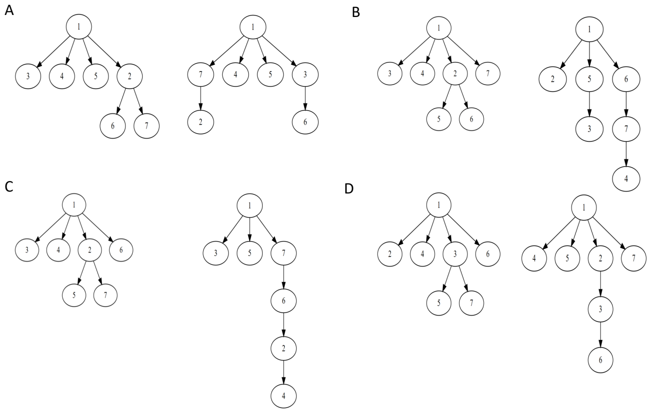

Figure S1 The two trees that have the same AD (A), CASet (B), DISC (C) and TD (D) distance but different DB distances from the 1-labeled star tree centered at Node 1.