A Generalized Robinson-Foulds Distance for Clonal Trees, Mutation Trees, and Phylogenetic Trees and Networks

Mercè Llabrés
University of the Balearic Islands
Palma de Mallorca, Spain
merce.llabres@uib.es

Francesc Rosselló
University of the Balearic Islands
Palma de Mallorca, Spain
cesc.rossello@uib.es

Gabriel Valiente
Technical University of Catalonia
Barcelona, Spain
gabriel.valiente@upc.edu

ABSTRACT

Cancer evolution is often modeled by clonal trees (whose nodes are labeled by multiple somatic mutations) or mutation trees (where nodes are labeled by single somatic mutations). Clonal trees are generated from sequence data with different computational methods that may produce different clone phylogenies, rendering their analysis and comparison necessary to infer mutation order and clone origin during tumor progression. In this paper, we present a distance metric for multi-labeled trees that generalizes the Robinson-Foulds distance for phylogenetic trees, allows for a similarity assessment at much higher resolution, and can be applied to trees and networks with different sets of node labels. The generalized Robinson-Foulds distance can be computed in time quadratic in the size of the input multisets of multisets of node labels, and is a metric for clonal trees, mutation trees, phylogenetic trees, and several classes of phylogenetic networks.

KEYWORDS

Cancer genomics, phylogenetics, Robinson-Foulds distance, metrics, multi-labeled tree, clonal tree, mutation tree, phylogenetic tree, phylogenetic network

1 BACKGROUND

The clonal theory of cancer evolution [40] establishes that tumor cell populations evolve from a single clone, that is, from somatic mutation in a single cell of origin. Thus, the evolution of a tumor cell population can be described by a multi-labeled tree, whose nodes are labeled by the set of somatic mutations that characterize the corresponding clone.

Recent studies provide evidence for the multi-clonal origin of tumor cell populations, that is, for tumors that arise from somatic mutation in two or more cells or clones of cells [41, 42]. However, the evolution of multi-clonal tumor cell populations can still be modeled by a set of multi-labeled trees, one for each founder clone, or by a single multi-labeled tree rooted at an additional node for healthy cells instead of the (multiple) founder clones.

The reconstruction of tumor evolution from bulk sequencing data (clonal deconvolution) and from single-cell sequencing data has received much attention over the last few years; see [3, 28, 48, 55] for recent reviews. Under perfect phylogeny [25], that is, under the infinite sites assumption [34, 37], by which a somatic mutation can be gained only once and never lost during tumor evolution, these methods infer multi-labeled trees that correspond to sets of clones.

Note that in multi-labeled trees inferred under the infinite sites assumption, for either binary characters [5] or multi-state phylogenies [21], a somatic mutation cannot appear more than once. However, tumor phylogenies inferred under more general models of evolution correspond to multi-labeled trees with multiple occurrences of some somatic mutations. For example, somatic mutations can be gained at most once but lost more than once in the Dollo model [22], and can be gained more than once but never lost in the Camin-Sokal model [10]. Thus, these methods infer multi-labeled trees that correspond to multisets of clones. See [6, 7] for recent approaches to cancer phylogeny reconstruction under these models of evolution.

We will make the infinite sites assumption in the rest of this paper. The multi-labeled trees that model tumor evolution under perfect phylogeny are called tumor trees in [24], labeled trees or mutation trees in [30], and clonal trees in [20, 32].

Definition 1.1. A clonal tree for a set X of somatic mutations is a rooted tree T such that

(1) each node $v \in T$ is labeled by a one or more mutations from X, denoted by $\ell(v)$,
(2) every mutation in X labels some node of T, and
(3) no mutation in X appears more than once in T.

That is, in a clonal tree T over a set of mutations X, we have $\emptyset \neq \ell(v) \subseteq X$ for all nodes $v \in T$, $\cup_{v \in T} \ell(v) = X$, and $\ell(v) \cap \ell(w) = \emptyset$ for all nodes $v \neq w \in T$.

A mutation tree is a particular case of a clonal tree whose nodes are labeled by only one mutation. They are called mutation trees in [1, 33] and 1-labeled trees in [30].

Definition 1.2. A mutation tree for a set X of somatic mutations is a clonal tree T for X in which each node is labeled by one mutation from X.

That is, in a mutation tree T over a set of mutations X, we have $|\ell(v)| = 1$ for all nodes $v \in T$. Mutation trees model tumor evolution at higher resolution than clonal trees. In fact, at the highest possible resolution [20].

On the other hand, phylogenetic trees are single-labeled trees that model the evolutionary history of a set of taxa, such as genes, species, populations, languages, etc.

Definition 1.3. A phylogenetic tree for a set X of taxa is a rooted tree T such that

(1) each leaf node $v \in T$ is labeled by an element from X, denoted by $\ell(v)$,
(2) every element in X labels some leaf node of T,
(3) the labels of different leaves are different, and
(4) every non-leaf node has at least two children.

The final publication is available at ACM via http://dx.doi.org/ 10.1145/3388440.3412479

Phylogenetic networks are a generalization of phylogenetic trees that also model reticulate evolutionary events, such as recombination, hybridization, and lateral gene transfer, in an evolutionary history [44, 45].

Definition 1.4. A phylogenetic network for a set X of taxa is a directed acyclic graph N such that

(1) each leaf node $v \in N$ is labeled by an element from X, denoted by $\ell(v)$,
(2) every element in X labels some leaf node of N,
(3) the labels of different leaves are different,
(4) every node with at least two parents has a single child, and
(5) every non-leaf node with at most one parent has at least two children.

Recall that the cluster associated with a node in a phylogenetic tree is the set of descendant leaf labels of the node in the tree, also called a clade or a monophyletic group, and the cluster representation of a phylogenetic tree [50, §2.2] is the set of clades for the nodes in the tree. In a similar vein, but moving up towards the root instead of moving down towards the leaves, let us define the clonal representation of a multi-labeled tree (under a given model of evolution, in particular under perfect phylogeny) as the multiset of sets of somatic mutations accumulated at the nodes of the tree during tumor evolution.

Figure 1 (left) shows the set of clones of the consensus clonal tree inferred by [24] for tumor sample CLL077 from [47]. Mutated genes are numbered as follows: (1) BCL2L13, (2) COL24A1, (3) DAZAP1, (4) EXOC6B, (5) GHDC, (6) GPR158, (7) HMCN1, (8) KLHDC2, (9) LRRC16A, (10) MAP2K1, (11) NAMPTL, (12) NOD1, (13) OCA2, (14) PLA2G16, (15) SAMHD1, (16) SLC12A1. Figure 1 (right) shows the multiset of clusters of a phylogenetic network constructed from 21 different isolates of the yeast *Cryptococcus gattii*, adapted from [54].

The Robinson-Foulds (RF) distance [46], a widely used metric for comparing phylogenetic trees, has been extended to phylogenetic networks [17] and can be computed very quickly on both phylogenetic trees [18, 43] and phylogenetic networks [2]. It was originally defined as the cardinality of the symmetric difference between the sets (or multisets) of clusters of the phylogenetic trees (or networks, respectively). When normalized to the unit interval, it is the Jaccard distance [29] on these sets or multisets: we recall its definition, for multisets, in Definition A.2 in the Appendix at the end of the paper.

Definition 1.5. Let $C(T)$ be the set of clusters in a phylogenetic tree T. The (normalized) Robinson-Foulds distance between two phylogenetic trees T_1 and T_2 is the Jaccard distance between their sets of clusters, that is,

$$d(T_1, T_2) = \frac{|C(T_1) \setminus C(T_2)| + |C(T_2) \setminus C(T_1)|}{|C(T_1) \cup C(T_2)|}$$

The Jaccard distance was shown to be a metric on sets in [36, 38]. See [26] for other distance metrics on sets.

Note that, in the RF distance between two phylogenetic trees, the distance between a cluster in one tree and a cluster in the other tree is either 0 (when the two trees share the cluster) or 1 (when the cluster belongs to only one of the two trees). In this paper, we generalize the RF distance by taking into account the overlap between each cluster in one tree and each cluster in the other tree. We extend the definition to multisets of multisets of node labels, thus the generalized RF distance can be applied to clonal trees and mutation trees (which we show below to be characterized by the set of clones of ascendant node labels), to phylogenetic trees (which are known to be characterized by the set of clusters of descendant node labels), and to some classes of phylogenetic networks (which are known to be characterized by the multiset of clusters of descendant node labels).

The following result was proved for subsets in [23], but a close analysis of the proof shows that it also holds for sub-multisets.

Theorem 1.6. *Let $d : X \times X \to \mathbb{R}$ be a distance metric on a non-empty set X, and let $\mathcal{M}(X)$ be the set of all non-empty finite multisets on X. Let $D_d : \mathcal{M}(X) \times \mathcal{M}(X) \to \mathbb{R}$ be the function defined, for each A and B in $\mathcal{M}(X)$, as*

$$D_d(A, B) = \frac{\sum_{a \in A} \sum_{b \in B \setminus A} d(a, b)}{|A \cup B| \cdot |A|} + \frac{\sum_{a \in A \setminus B} \sum_{b \in B} d(a, b)}{|A \cup B| \cdot |B|}$$

where in each sum, each element is counted with its multiplicity in the corresponding multiset. Then, D_d is a distance metric on $\mathcal{M}(X)$.

Notice that if $d : X \times X \to \mathbb{R}$ is the trivial metric

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

then

$$D_d(A, B) = \frac{|A| \cdot |B \setminus A|}{|A \cup B| \cdot |A|} + \frac{|A \setminus B| \cdot |B|}{|A \cup B| \cdot |B|} = \frac{|A \mathbin{\triangle} B|}{|A \cup B|}$$

is the Jaccard distance on $\mathcal{M}(X)$.

In this paper we shall be concerned with the following instantiation of this metric.

Definition 1.7. Let X be a non-empty set, let $\mathcal{M}(X)$ be the set of all non-empty finite multisets on X, and let $d_J : \mathcal{M}(X) \times \mathcal{M}(X) \to \mathbb{R}$ be the Jaccard distance on multisets. The *generalized Robinson-Foulds distance* D on $\mathcal{M}(X)$ is D_{d_J}.

In other words, for every pair of multisets A, B of multisets of elements of X,

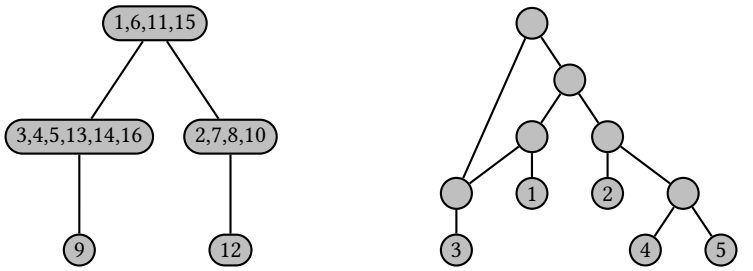
$$D(A, B) = \frac{\sum_{a \in A} \sum_{b \in B \setminus A} (|a \mathbin{\triangle} b| / |a \cup b|)}{|A \cup B| \cdot |A|} + \frac{\sum_{a \in A \setminus B} \sum_{b \in B} (|a \mathbin{\triangle} b| / |a \cup b|)}{|A \cup B| \cdot |B|}$$

where the union and difference operations, as well as the cardinalities, are understood to be of multisets.

Therefore, it can be used for comparing clonal trees and mutation trees (sets of clones), phylogenetic trees (sets of clusters), and some classes of phylogenetic networks (multisets of clusters). As a matter of fact, it can be used to compare any labeled structures that are characterized by sets or multisets of sets or multisets of labels.

The generalized RF distance can be computed in time quadratic in the size of the input, using simple algorithms and data structures that can be easily implemented in any modern programming language. In the following proof, we assume that the input consists of two sorted lists (multisets) of sorted lists (multisets). Sorting a list of lists A involves sorting each of the inner lists $a \in A$ and then sorting the outer list A, which takes time

$$\sum_{a \in A} O(|a| \log |a|) + O(|A|^2) = O(|A|^2)$$



$$\left\{ \begin{array}{l} \{1,2,6,7,8,10,11,12,15\},\{1,2,6,7,8,10,11,15\}, \\ \{1,3,4,5,6,11,13,14,15,16\}, \\ \{1,3,4,5,6,9,11,13,14,15,16\},\{1,6,11,15\} \end{array} \right\} \quad \left[\begin{array}{l} \{1\},\{1,2,3,4,5\},\{1,2,3,4,5\}, \\ \{1,3\},\{2\},\{2,4,5\},\{3\},\{3\}, \\ \{4\},\{4,5\},\{5\} \end{array} \right]$$

Figure 1: Set of clones of a clonal tree (left) and multiset of clusters of a phylogenetic network (right).

using comparison sort, but only $O(|A|)$ time when the elements are integers in the range $[1, |A|]$, using radix sort.

Lemma 1.8. *The generalized Robinson-Foulds distance $D(A, B)$ can be computed in $O(|Supp(A)| \cdot |B| + |Supp(B)| \cdot |A|)$ time.*

Proof. Assume the input consists of two sorted lists (multisets) of sorted lists (multisets). Computing the generalized RF distance between them requires the computation of the Jaccard distance $d(a, b)$ for each $a \in A$ and $b \in B$. Now, each such $d(a, b)$ can be computed in $O(|a| + |b|)$ time by simultaneous traversal of the two sorted lists to compute their multiset union and their multiset difference, with an overall asymptotic contribution of

$$\sum_{a \in A} \sum_{b \in B} (|a| + |b|) = |Supp(A)| \cdot |B| + |Supp(B)| \cdot |A|$$

to the running time. Computing the multiset of multisets union $A \cup B$ takes $O(|A| + |B|)$ time, again by simultaneous traversal of the two sorted lists of sorted lists. □

Notice that $\sum_{a \in A} \sum_{b \in B} (|a|+|b|) = O(m^3)$ on clonal or mutation trees for m mutations. The largest value is reached at trees with a clonal representation of the form $[[1], [1, 2], \ldots, [1, 2, \ldots, m]]$ and size $1 + 2 + \cdots + m = m(m + 1)/2 = O(m^2)$ and, in this case, the sum equals $m^2(m + 1) = O(m^3)$.

1.1 Related Work

The comparison of phylogenetic trees and networks, on the one hand, has been studied for several decades now; see [31] for a recent review. To the best of our knowledge, there is only one other generalization of the RF metric for the comparison of phylogenetic trees [4, 8]. Like the generalized RF distance we present in this paper, it does count not only identical but also overlapping clusters in the phylogenetic trees under comparison. However, computing it is an NP-hard problem, while the generalized RF distance can be computed in polynomial time.

On the other hand, several metrics have been proposed for the comparison of clonal trees and mutation trees in the last few years [1, 20, 32], including another generalization of the RF metric [30] into a set of distance metrics (the Bourque and k-Bourque distances) for comparing mutation trees. The Bourque distance coincides with the RF distance on clonal or mutation trees for the same set of mutations. We compare the generalized RF distance with other previous distances in Section 3.6. Studying the resolution of the k-Bourque distances is an interesting line of future work.

2 METHODS

We give below a series of simple and efficient algorithms to obtain the set or multiset of sets or multisets of node labels in a clonal or mutation tree, in a phylogenetic tree, and in a phylogenetic network. We also give simple and efficient algorithms to perform some set-theoretical operations on multisets of multisets of node labels, based on a simple representation of a multiset of node labels as a sorted list of pairs of node label and multiplicity, and a multiset of multisets of node labels as a sorted list of pairs of multiset of node labels and multiplicity. These basic algorithms allow us to compute the generalized RF distance in time quadratic in the size of the input sorted lists of sorted lists of node labels.

Notice that the set union operations in the pseudocode of the algorithms are actually list append operations that only take $O(1)$ time, as the clone or cluster and the node labels to append are always disjoint. Sorting the output of these algorithms using radix sort restores the simple multiset representation mentioned above.

2.1 Set of Clones of a Clonal or Mutation Tree

The set of clones (sets of ascendant node labels) of a clonal or mutation tree can be obtained by accumulating the sets of node labels in the ascendant nodes of each node, in time linear in the size of the tree, during a preorder traversal of the tree, see Algorithm 1.

2.2 Set of Clusters of a Phylogenetic Tree

The set of clusters (sets of descendant node labels) of a phylogenetic tree can be obtained by accumulating the sets of node labels in the descendant nodes of each node, in time linear in the size of the tree, during a postorder traversal of the tree, see Algorithm 2.

2.3 Multiset of Clusters of a Phylogenetic Network

The multiset of clusters (sets of descendant node labels) of a phylogenetic network can be obtained by accumulating the sets of node

Algorithm 1 Set C of clones of a clonal or mutation tree T

```
function SetOfClones(T)
    for all nodes v of T do
        C[v] = set of labels of v in T
    SetOfClones(T,root of T,C)
    return C
procedure SetOfClones(T,v,C)
    for all children w of v in T do
        C[w] = C[w] ∪ C[v]
        SetOfClones(T,w,C)
```

Algorithm 2 Set C of clusters of a phylogenetic tree T

```
function SetOfClusters(T)
    for all nodes v of T do
        if v is a leaf in T then
            C[v] = {label of v in T}
        else
            C[v] = ∅
    SetOfClusters(T,root of T,C)
    return C
procedure SetOfClusters(T,v,C)
    for all children w of v in T do
        SetOfClusters(T,w,C)
        C[v] = C[v] ∪ C[w]
```

labels in the descendant nodes of each node, in time linear in the size of the network, during a bottom-up traversal [52, §3.4] of the network. A simple and efficient algorithm is based on the idea of repeatedly deleting a vertex of in-degree zero to obtain a topological order of a directed acyclic graph [35, §2.2.3], but simulating the deletion of a vertex of out-degree zero instead, to perform a bottom-up traversal of a phylogenetic network. See Algorithm 3.

Algorithm 3 Multiset C of clusters of a phylogenetic network N

```
function MultisetOfClusters(N)
    let Q be an empty queue of nodes
    for all nodes v of N do
        outdeg[v] = number of children of v in N
        if outdeg[v] = 0 then
            C[v] = {label of v in N}
            enqueue(Q, v)
        else
            C[v] = ∅
    while Q is not empty do
        v = dequeue(Q)
        for all parent u of v in N do
            C[u] = C[u] ∪ C[v]
            outdeg[u] = outdeg[u] − 1
            if outdeg[u] = 0 then
                enqueue(Q, u)
    return C
```

2.4 Set-Theoretical Operations on Multisets

Computation of the generalized RF distance in time quadratic in the size of the input sorted list of sorted lists of node labels, requires the efficient computation of some set-theoretical operations on multisets of multisets.

The following algorithms are based on the idea behind the merge algorithm [39, §5.2] of the simultaneous traversal of two sorted lists or arrays. The union of two multisets of node labels, or two multisets of multisets of node labels, can both be computed in time linear in the size of the input, see Algorithm 4.

Algorithm 4 Union $C = A \cup B$ of two multisets of A and B

```
function MultisetUnion(A, B)
    C = ∅
    i = 1
    j = 1
    while i ⩽ |A| and j ⩽ |B| do
        if fst(A[i]) < fst(B[j]) then
            C = C ∪ A[i]
            i = i + 1
        else if fst(A[i]) = fst(B[j]) then
            C = C ∪ (fst(A[i]), max(snd(A[i]), snd(B[j])))
            i = i + 1
            j = j + 1
        else                                   fst(A[i]) > fst(B[j])
            C = C ∪ B[j]
            j = j + 1
    while i ⩽ |A| do
        C = C ∪ A[i]
        i = i + 1
    while j ⩽ |B| do
        C = C ∪ B[j]
        j = j + 1
    return C
```

The input sorted lists of node labels (for multisets) or sorted lists of sorted lists of node labels (for multisets of multisets) are assumed to be in the aforementioned simple format of sorted lists of pairs of node label and multiplicity (for multisets) or sorted lists of pairs of multiset of node labels and multiplicity (for multisets of multisets).

In a similar way, the difference of two multisets of node labels, or two multisets of multisets of node labels, can be computed in time linear in the size of the input, see Algorithm 5.

Using these algorithms, the generalized RF distance can be computed in time quadratic in the size of the input multisets of multisets of node labels, in a straightforward way.

3 RESULTS

We show below that the clonal representation of a clonal tree (the set of clones of ascendant node labels) and, in particular, of a mutation tree, characterizes, up to isomorphism, the tree, in much the same way as the cluster representation of a phylogenetic tree (the set of clusters of descendant node labels) or a phylogenetic network (the multiset of clusters of descendant node labels) characterizes, up to isomorphism, the tree or network.

Algorithm 5 Difference $C = A \setminus B$ of two multisets of A and B

```
function MultisetDifference(A, B)
    C = ∅
    i = 1
    j = 1
    while i ⩽ |A| and j ⩽ |B| do
        if fst(A[i]) < fst(B[j]) then
            C = C ∪ A[i]
            i = i + 1
        else if fst(A[i]) = fst(B[j]) then
            if snd(A[i]) > snd(B[j]) then
                C = C ∪ (fst(A[i]), snd(A[i]) − snd(B[j]))
            i = i + 1
            j = j + 1
        else                                   fst(A[i]) > fst(B[j])
            j = j + 1
    while i ⩽ |A| do
        C = C ∪ A[i]
        i = i + 1
    return C
```

We also show experimental results on all clonal trees for a given set of somatic mutations and a given number of nodes, all mutations trees for a given set of somatic mutations, all phylogenetic trees with a given set of leaf labels, and all phylogenetic networks with a given set of leaf labels. These results confirm the claim that the generalized RF distance is a generalization of the RF distance. As a matter of fact, for each value taken by the RF distance, the pairs of trees or networks at that distance value are split into many different values of the generalized RF distance, that is, the latter has a much higher resolution.

3.1 Clonal Representation

Let X be a set of somatic mutations, let $T = (V, E)$ be a clonal tree, and let $\ell : V \to \mathcal{P}(X)$ be an injective mapping such that different nodes have assigned disjoint subsets of mutations. We denote by $\ell(v) \in \mathcal{P}(X)$ the mutations of a node $v \in V$. In every rooted tree T, for every node $v \in V$ there is only one path from the root, r, to v. We denote this path by $p_v = (u_0, \ldots, u_k)$. where $u_0 = r$, $u_k = v$, and $(u_i, u_{i+1}) \in E$ for every $i = 1, \ldots, k-1$. The nodes $u_i \in p_v$ are the *ancestors* of v. We denote by $\mathcal{A}(v)$ the set of ancestors of v. That is, $\mathcal{A}(v) = \{u \in V \mid u \in p_v\}$. We define the *clone* of every node $v \in V$, and we denote it by $cl(v)$, as

$$cl(v) = \cup_{u_i \in \mathcal{A}(v)} \{\ell(u_i)\} \subseteq X\,.$$

We denote by $C(T) = \{cl(v) \mid v \in V\} \subseteq \mathcal{P}(X)$ the clonal representation of T.

Proposition 3.1. *The following properties hold:*

(1) *If $u \neq v \in V$, then $cl(u) \neq cl(v)$.*
(2) *For every $u, v \in V$, $\emptyset \neq cl(u) \cap cl(v) = cl(w)$, where w is the lowest common ancestor of u and v in T.*
(3) *$cl(u) \subseteq cl(v)$ if, and only if, $u \in \mathcal{A}(v)$.*
(4) *If $(u_1, u_2) \in E$, then $cl(u_1) \subseteq cl(u_2)$ and there is no $v \in V$ such that $cl(u_1) \subseteq cl(v) \subseteq cl(u_2)$.*

Now, we have the following result:

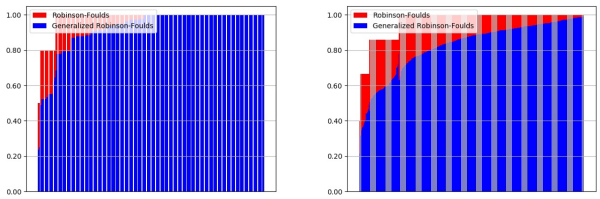


Figure 2: Pairwise RF and generalized RF distances on the 225 clonal trees for 5 mutations with 3 nodes (left) and the 640 clonal trees for 5 mutations with 4 nodes (right).

Proposition 3.2. *Let X be a set of labels, and let $T = (V, E)$ be a rooted tree with all its nodes injectively labeled on $\mathcal{P}(X)$. Then, the Hasse diagram of $(C(T), \subseteq)$ is a directed acyclic graph $H = (V_H, E_H)$ such that $\varphi : T \to H$, defined by $\varphi(v) = cl(v)$, is an isomorphism.*

Proof. $\varphi : T \to H$ is an isomorphism if it is a bijective function such that $(u_1, u_2) \in E(T)$ if and only if $(\varphi(u_1), \varphi(u_2)) \in E_H$. Since T is a rooted tree with all its nodes injectively labeled on $\mathcal{P}(X)$, the mapping $\varphi : T \to H$ defined by $\varphi(v) = cl(v)$ is clearly a bijective function. Now, by the properties in the above proposition we have that $(u_1, u_2) \in E$ if and only if $cl(u_1) \subseteq cl(u_2)$ and there is no any $v \in T$ such that $cl(u_1) \subseteq cl(v) \subseteq cl(u_2)$ if and only if $cl(u_1) \subseteq cl(u_2)$ and there is no any $Y \in C(T)$ such that $cl(u_1) \subseteq Y \subseteq cl(u_2)$ if and only if $(cl(u_1), cl(u_2)) \in E_H$. □

The previous results entail that the clonal representation characterizes, up to isomorphism, a clonal tree.

Proposition 3.3. *If T_1 and T_2 are clonal trees on a set of somatic mutations X such that $C(T_1) = C(T_2)$, then T_1 and T_2 are isomorphic clonal trees.*

Proof. Let T_1 and T_2 be clonal trees on a set of somatic mutations X and let H_1 and H_2 be the Hasse diagram of $(C(T_1), \subseteq)$ and $(C(T_2), \subseteq)$, respectively. If $C(T_1) = C(T_2)$ then, $H_1 = H_2$ and the previous proposition entails that T_1 and T_2 are isomorphic trees. As far as the nodes labels goes, they are uniquely determined by the fact that the label of the root is its clone, which is the minimum set, and if $(u, v) \in E$ then $\ell(v) = cl(v) \setminus cl(u)$. □

3.2 Clonal Trees

There are $\left\{{m \atop k}\right\}$ partitions of a set $X = \{1, \ldots, m\}$ into k non-empty subsets, and there are n^{n-1} rooted labeled trees with n nodes [51]. Thus, there are $\sum_{n=1}^{m} \left\{{m \atop n}\right\} n^{n-1}$ clonal trees with n nodes for a set X of m mutations. We generated all the clonal trees for a set $X = \{1, \ldots, m\}$ of somatic mutations and $n \leqslant m$ nodes, for $m = 3, 4, 5, 6$ and $n = 2, 3, \ldots, m$.

As can be seen in the pairwise distance plots in Figure 2, the generalized RF distance has a much higher resolution than the RF distance. As a matter of fact, there are only 2, 3, 4, 5, and 6 different values for the RF distance, but 2, 39, 752, 11,225, and 14,002 different values for the generalized RF distance on clonal trees for 6 mutations with 2, 3, 4, 5, and 6 nodes, respectively.

Table 1: Number of clonal trees for m somatic mutations with n nodes

m	n								
	2	3	4	5	6	7	8	9	10
2	2								
3	6	9							
4	14	54	64						
5	30	225	640	625					
6	62	810	4,160	9,375	7,776				
7	126	2,709	22,400	87,500	163,296	117,649			
8	254	8,694	108,864	656,250	2,068,416	3,294,172	2,097,152		
9	510	27,225	497,280	4,344,375	20,575,296	54,353,838	75,497,472	43,046,721	
10	1,022	83,970	2,182,720	26,578,125	177,502,752	691,776,120	1,572,864,000	1,937,102,445	1,000,000,000

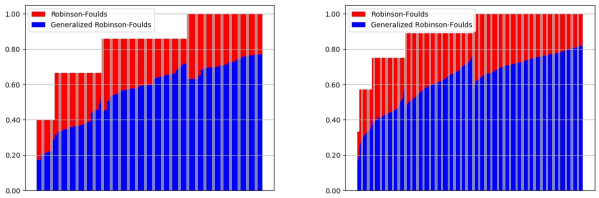


Figure 3: Pairwise RF and generalized RF distances on the 64 mutation trees for 4 mutations (left) and the 625 mutation trees for 5 mutations (right).

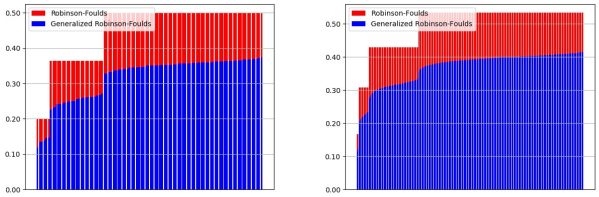


Figure 4: Pairwise RF and generalized RF distances on the 105 phylogenetic trees with 5 labeled leaves (left) and the 945 phylogenetic trees with 6 labeled leaves (right).

3.3 Mutation Trees

Mutation trees are clonal trees with m nodes for a set X of m somatic mutations. Thus, there are m^{m-1} mutation trees for a set X of m somatic mutations. We generated all the mutation trees for a set $X = \{1, \ldots, m\}$ of somatic mutations, for $m = 3, 4, 5, 6$.

As can be seen in the pairwise distance plots in Figure 3, the generalized RF distance has a much higher resolution than the RF distance. As a matter of fact, there are only 3, 4, 5, and 6 different values for the RF distance, but 12, 142, 2,363, and 14,002 different values for the generalized RF distance on mutation trees for 3, 4, 5, and 6 mutations, respectively.

3.4 Phylogenetic Trees

We generated all the fully-resolved phylogenetic trees with n labeled leaves, for $n = 3, 4, 5, 6$, using the algorithm described in [53, §5.3.3], as implemented in Bio::Phylo [15, 56], and computed the RF distance and the generalized RF distance between each pair of phylogenetic trees with the same number of labeled leaves.

As can be seen in the pairwise distance plots in Figure 4, the generalized RF distance has a much higher resolution than the RF distance. As a matter of fact, there are only 1, 2, 3, and 4 different values for the RF distance, but 1, 12, 149, and 1,406 different values for the generalized RF distance on phylogenetic trees with 3, 4, 5, and 6 labeled leaves, respectively.

3.5 Phylogenetic Networks

Several classes of phylogenetic networks have a unique, up to isomorphism, representation as a multiset of clusters, including binary galled trees [12], tree-child time-consistent phylogenetic networks [11, 16, 17], and semi-binary tree-sibling time-consistent phylogenetic networks [14]. Notice that more general classes of phylogenetic networks do not have such a unique representation [13].

We generated all the fully-resolved tree-child time-consistent phylogenetic networks with n labeled leaves, for $n = 3, 4$, using the algorithm described in [53, §8.3.3], as implemented in Bio::Phylo [15, 56], and computed the RF distance and the generalized RF distance between each pair of phylogenetic networks with the same number of labeled leaves. We also downloaded from [14, Suppl.] all the semi-binary tree-sibling time-consistent phylogenetic networks with n labeled leaves, for $n = 3, 4$, and computed the RF distance and the generalized RF distance between each pair of phylogenetic networks with the same number of labeled leaves.

As can be seen in the pairwise distance plots in Figure 5, the generalized RF distance has a much higher resolution than the RF distance. As a matter of fact, there are only 2 and 20 different values for the RF distance, but 6 and 546 different values for the generalized RF distance on fully-resolved tree-child time-consistent phylogenetic networks with 3 and 4 labeled leaves, respectively. Also, there are only 9 and 47 different values for the RF distance, but 17 and 9,059 different values for the generalized RF distance on semi-binary tree-sibling time-consistent phylogenetic networks with 3 and 4 labeled leaves, respectively.

Table 2: RF and generalized RF distances on clonal trees for m somatic mutations with n nodes

T_1		T_2		Robinson-Foulds			Generalized RF			T_1		T_2		Robinson-Foulds			Generalized RF		
m	n	m	n	min	max	count	min	max	count	m	n	m	n	min	max	count	min	max	count
3	3	3	2	0.7500	1.0000	2	0.4444	0.9583	12	6	3	3	2	0.3333	1.0000	3	0.1667	1.0000	20
4	2	3	2	0.6667	1.0000	2	0.3889	1.0000	4	6	3	3	3	0.5000	1.0000	3	0.2500	1.0000	59
4	2	3	3	0.7500	1.0000	2	0.4444	1.0000	12	6	3	4	2	0.3333	1.0000	3	0.1667	1.0000	20
4	3	3	2	0.3333	1.0000	3	0.1667	1.0000	20	6	3	4	3	0.5000	1.0000	3	0.2500	1.0000	59
4	3	3	3	0.5000	1.0000	3	0.2500	0.9778	56	6	3	4	4	0.6000	1.0000	3	0.3000	1.0000	234
4	3	4	2	0.7500	1.0000	2	0.4444	1.0000	13	6	3	5	2	0.3333	1.0000	3	0.1667	1.0000	20
4	4	3	2	0.8000	1.0000	2	0.4750	0.9750	25	6	3	5	3	0.5000	1.0000	3	0.2500	1.0000	59
4	4	3	3	0.2500	1.0000	4	0.1250	0.8472	97	6	3	5	4	0.6000	1.0000	3	0.3000	1.0000	234
4	4	4	2	0.8000	1.0000	2	0.4750	1.0000	26	6	3	5	5	0.6667	1.0000	3	0.3333	1.0000	684
4	4	4	3	0.6000	1.0000	3	0.3000	0.9583	190	6	3	6	2	0.7500	1.0000	2	0.4444	1.0000	13
5	2	3	2	0.6667	1.0000	2	0.3889	1.0000	4	6	4	3	2	0.5000	1.0000	3	0.2500	1.0000	55
5	2	3	3	0.7500	1.0000	2	0.4444	1.0000	12	6	4	3	3	0.2500	1.0000	4	0.1250	1.0000	289
5	2	4	2	0.6667	1.0000	2	0.3889	1.0000	4	6	4	4	2	0.5000	1.0000	3	0.2500	1.0000	55
5	2	4	3	0.7500	1.0000	2	0.4444	1.0000	12	6	4	4	3	0.2500	1.0000	4	0.1250	1.0000	307
5	2	4	4	0.8000	1.0000	2	0.4750	1.0000	27	6	4	4	4	0.4000	1.0000	4	0.1900	1.0000	1,050
5	3	3	2	0.3333	1.0000	3	0.1667	1.0000	20	6	4	5	2	0.8000	1.0000	2	0.4750	1.0000	26
5	3	3	3	0.5000	1.0000	3	0.2500	1.0000	59	6	4	5	3	0.2500	1.0000	4	0.1250	1.0000	307
5	3	4	2	0.3333	1.0000	3	0.1667	1.0000	20	6	4	5	4	0.4000	1.0000	4	0.1900	1.0000	1,054
5	3	4	3	0.5000	1.0000	3	0.2500	1.0000	59	6	4	5	5	0.5000	1.0000	4	0.2350	0.9938	3,851
5	3	4	4	0.6000	1.0000	3	0.3000	1.0000	234	6	4	6	2	0.8000	1.0000	2	0.4750	1.0000	26
5	3	5	2	0.7500	1.0000	2	0.4444	1.0000	13	6	4	6	3	0.6000	1.0000	3	0.3000	1.0000	233
5	4	3	2	0.5000	1.0000	3	0.2500	1.0000	55	6	5	3	2	0.6000	1.0000	3	0.3000	1.0000	167
5	4	3	3	0.2500	1.0000	4	0.1250	0.9861	286	6	5	3	3	0.4000	1.0000	4	0.2000	0.9905	984
5	4	4	2	0.8000	1.0000	2	0.4750	1.0000	26	6	5	4	2	0.8333	1.0000	2	0.4944	1.0000	74
5	4	4	3	0.2500	1.0000	4	0.1250	1.0000	307	6	5	4	3	0.4000	1.0000	4	0.2000	1.0000	1,037
5	4	4	4	0.4000	1.0000	4	0.1900	0.9750	952	6	5	4	4	0.2000	1.0000	5	0.0950	0.9833	4,396
5	4	5	2	0.8000	1.0000	2	0.4750	1.0000	26	6	5	5	2	0.8333	1.0000	2	0.4944	1.0000	74
5	4	5	3	0.6000	1.0000	3	0.3000	1.0000	233	6	5	5	3	0.6667	1.0000	3	0.3333	1.0000	732
5	5	3	2	0.8333	1.0000	2	0.4944	0.9833	72	6	5	5	4	0.2000	1.0000	5	0.0950	0.9938	4,821
5	5	3	3	0.4000	1.0000	4	0.2000	0.9111	526	6	5	5	5	0.3333	1.0000	5	0.1533	0.9760	9,960
5	5	4	2	0.8333	1.0000	2	0.4944	1.0000	74	6	5	6	2	0.8333	1.0000	2	0.4944	1.0000	74
5	5	4	3	0.6667	1.0000	3	0.3333	0.9733	611	6	5	6	3	0.6667	1.0000	3	0.3333	1.0000	732
5	5	4	4	0.2000	1.0000	5	0.0950	0.8750	1,745	6	5	6	4	0.5000	1.0000	4	0.2350	1.0000	3,874
5	5	5	2	0.8333	1.0000	2	0.4944	1.0000	74	6	6	3	2	0.8571	1.0000	2	0.5079	0.9881	174
5	5	5	3	0.6667	1.0000	3	0.3333	0.9905	728	6	6	3	3	0.5000	1.0000	4	0.2500	0.9417	1,514
5	5	5	4	0.5000	1.0000	4	0.2350	0.9625	2,647	6	6	4	2	0.8571	1.0000	2	0.5079	1.0000	184
6	2	3	2	0.6667	1.0000	2	0.3889	1.0000	4	6	6	4	3	0.7143	1.0000	3	0.3571	0.9815	1,588
6	2	3	3	0.7500	1.0000	2	0.4444	1.0000	12	6	6	4	4	0.3333	1.0000	5	0.1583	0.9236	5,346
6	2	4	2	0.6667	1.0000	2	0.3889	1.0000	4	6	6	5	2	0.8571	1.0000	2	0.5079	1.0000	185
6	2	4	3	0.7500	1.0000	2	0.4444	1.0000	12	6	6	5	3	0.7143	1.0000	3	0.3571	0.9931	1,812
6	2	4	4	0.8000	1.0000	2	0.4750	1.0000	27	6	6	5	4	0.5714	1.0000	4	0.2679	0.9750	9,261
6	2	5	2	0.6667	1.0000	2	0.3889	1.0000	4	6	6	5	5	0.1667	1.0000	6	0.0767	0.8972	21,278
6	2	5	3	0.7500	1.0000	2	0.4444	1.0000	12	6	6	6	2	0.8571	1.0000	2	0.5079	1.0000	186
6	2	5	4	0.8000	1.0000	2	0.4750	1.0000	27	6	6	6	3	0.7143	1.0000	3	0.3571	1.0000	1,836
6	2	5	5	0.8333	1.0000	2	0.4944	1.0000	73	6	6	6	4	0.5714	1.0000	4	0.2679	0.9889	11,915
										6	6	6	5	0.4286	1.0000	5	0.1957	0.9667	50,668

3.6 Comparison with Previous Distances

We have compared the generalized RF distance with the RF distance [46], the parent-child distance [24], the ancestor-descendant (AD) distance [24], and the clonal distance [24] on mutation trees. For each of them, we computed the 30,229,200 pairwise distances between the 7,776 mutation trees for 6 mutations.

As can be seen in the frequency distribution plots in Figure 6, there are 10 and 7 different values for the AD and the generalized

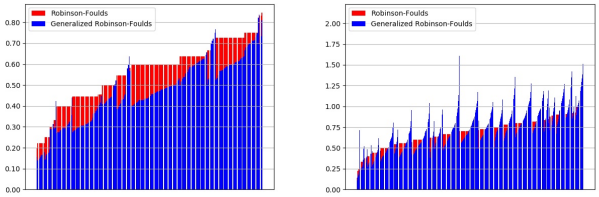


Figure 5: Pairwise RF and generalized RF distances on the 105 fully-resolved tree-child time-consistent phylogenetic networks with 4 labeled leaves (left) and the 444 semi-binary tree-sibling time-consistent phylogenetic networks with 4 labeled leaves (right).

Table 3: RF and generalized RF distances on phylogenetic trees with n labeled leaves

T_1	T_2	Robinson-Foulds			Generalized RF		
n	n	min	max	count	min	max	count
3	3	0.0000	0.3333	2	0.0000	0.2000	2
4	3	0.2857	0.6667	3	0.2286	0.5066	12
4	4	0.0000	0.4444	3	0.0000	0.3095	13
5	3	0.4444	0.7273	3	0.3644	0.6301	47
5	4	0.2222	0.6667	4	0.1810	0.5261	137
5	5	0.0000	0.5000	4	0.0000	0.3741	157
6	3	0.5455	0.7692	3	0.4558	0.7040	161
6	4	0.3636	0.7143	4	0.3017	0.6205	592
6	5	0.1818	0.6667	5	0.1498	0.5398	1,859
6	6	0.0000	0.5333	5	0.0000	0.4174	1,440

Table 4: Pearson correlation of the generalized RF distance and the RF, parent-child (PC), ancestor-descendant (AD), and clonal distances, on mutation trees for m mutations

m	RF	PC	AD	Clonal
3	0.95390660	0.29923048	0.97841199	0.81210677
4	0.93688530	0.40083128	0.97104421	0.79317898
5	0.90482204	0.37799780	0.96466932	0.75599857
6	0.86370536	0.32511976	0.96041726	0.71490774

RF distances, respectively, and only 5 for the other distances. The generalized RF distance stands out with a different distribution.

Also, the generalized RF distance has a strong Pearson correlation with the RF and ancestor-descendant distances, and a weaker Pearson correlation with the parent-child and clonal distances on the mutation trees for $3 \leqslant m \leqslant 6$ mutations, as shown in Table 4.

4 DISCUSSION

In this paper, we consider the generalized RF distance as a metric for different models of evolution. Namely, for clonal trees, mutation trees, phylogenetic trees, binary galled trees, tree-child time-consistent phylogenetic networks, and semi-binary tree-sibling

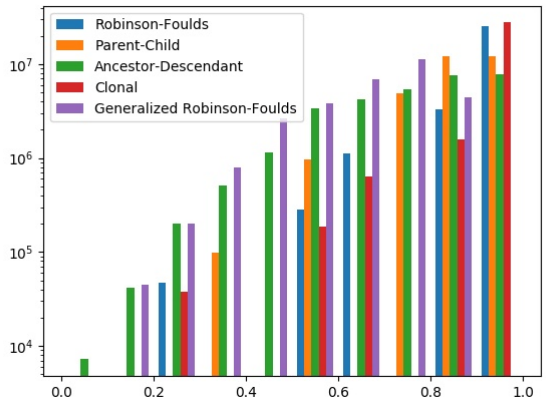


Figure 6: Frequency distribution of pairwise RF, parent-child, ancestor-descendant, clonal, and generalized RF distances on the 7,776 mutation trees for 6 mutations.

time-consistent phylogenetic networks. The fact that the generalized RF distance is a metric on multisets of multisets, provides a distance metric for every evolution model where the elements are uniquely determined by multisets of multisets (which includes sets of sets, sets of multisets, and multisets of sets), as in the case of the aforementioned models. In addition, whenever a function is defined on any evolution model that assigns a multiset of multisets to every element of the model, the generalized RF distance can be equally defined and used to compare elements because, despite the separation condition, the symmetry and triangle inequality still hold. Hence, it is a pseudo-metric that can be used under more general models of evolution of clonal trees than perfect phylogeny, as well as for the comparison of multi-labeled trees in phylogenetics [27].

We performed an all-against-all computation of the generalized RF distance on a number of clonal trees, mutation trees, phylogenetic trees, and phylogenetic networks. Beside allowing for a similarity assessment at much higher resolution than the RF distance, the generalized RF distance has the additional advantage that it makes it possible to compare labeled structures of different size. Tables 2 and 3 show the minimum value, maximum value, and number of different values taken by the RF and the generalized RF distances on pairs of clonal trees and phylogenetic trees of different size. As in the case of clonal trees, mutation trees, phylogenetic trees, and phylogenetic networks of the same size, the generalized RF distance also has much higher resolution than the RF distance when comparing clonal trees and phylogenetic trees of different size. Since the generalized RF distance seems to be a suitable metric for many evolution models, its analytical analysis (including the study of the minimum and maximum values it takes) as well as its statistical analysis, as it has been done for the RF distance [9], remain an interesting line of future work.

For those labeled structures that do not have a unique representation as a multiset of multisets of labels, such as clonal trees inferred under the Dollo model [22] or the Camin-Sokal model [10], it might

be interesting to consider a combined representation as a multiset of clones plus a multiset of clusters or, alternatively, as a multiset of pairs of clone and cluster. Uniqueness of such a combined representation would automatically turn the generalized RF distance into a metric for these models of evolution as well. Therefore, the study of this combined representation is another interesting line of future research.

ACKNOWLEDGMENTS

We thank David Posada for discussions on several aspects of this paper. This research was partially supported by the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund through project PGC2018-096956-B-C43 (FEDER/MICINN/AEI).

REFERENCES

[1] Nuraini Aguse, Yuanyuan Qi, and Mohammed El-Kebir. 2019. Summarizing the Solution Space in Tumor Phylogeny Inference by Multiple Consensus Trees. *Bioinformatics* 35, 14 (2019), i408–i416. https://doi.org/10.1093/bioinformatics/btz312

[2] Tetsuo Asano, Jesper Jansson, Kunihiko Sadakane, Ryuhei Uehara, and Gabriel Valiente. 2012. Faster Computation of the Robinson-Foulds Distance between Phylogenetic Networks. *Inf. Sci.* 197 (2012), 77–90. https://doi.org/10.1016/j.ins.2012.01.038

[3] Niko Beerenwinkel, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. 2015. Cancer Evolution: Mathematical Models and Computational Inference. *Syst. Biol.* 64, 1 (2015), e1–e25. https://doi.org/10.1093/sysbio/syu081

[4] Sebastian Böcker, Stefan Canzar, and Gunnar W. Klau. 2013. The Generalized Robinson-Foulds Metric. In *Proc. 13th Int. Workshop Algorithms in Bioinformatics (Lecture Notes in Computer Science)*, Aaron Darling and Jens Stoye (Eds.), Vol. 8126. Springer, Berlin, Heidelberg, 156–169. https://doi.org/10.1007/978-3-642-40453-5_13

[5] Paola Bonizzoni, Chiara Braghin, Riccardo Dondi, and Gabriella Trucco. 2012. The Binary Perfect Phylogeny with Persistent Characters. *Theor. Comput. Sci.* 454, 5 (2012), 51–63. https://doi.org/10.1016/j.tcs.2012.05.035

[6] Paola Bonizzoni, Simone Ciccolella, Gianluca Della Vedova, and Mauricio Soto. 2017. Beyond Perfect Phylogeny: Multisample Phylogeny Reconstruction via ILP. In *Proc. 2017 ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics*. Association for Computing Machinery, New York, NY, 1–10. https://doi.org/10.1145/3107411.3107441

[7] Paola Bonizzoni, Simone Ciccolella, Gianluca Della Vedova, and Mauricio Soto. 2019. Does Relaxing the Infinite Sites Assumption Give Better Tumor Phylogenies? An ILP-Based Comparative Approach. *IEEE ACM T. Comput. Bi.* 16, 5 (2019), 1410–1423. https://doi.org/10.1109/TCBB.2018.2865729

[8] Luka Borozan, Domagoj Matijević, and Stefan Canzar. 2019. Properties of the Generalized Robinson-Foulds Metric. In *Proc. 42nd Int. Convention on Information and Communication Technology, Electronics and Microelectronics*. Institute of Electrical and Electronic Engineers, New York, NY, 330–335. https://doi.org/10.23919/MIPRO.2019.8756638

[9] David Bryant and Mike Steel. 2009. Computing the Distribution of a Tree Metric. *IEEE ACM T. Comput. Bi.* 6, 3 (2009), 420–426. https://doi.org/10.1109/TCBB.2009.32

[10] Joseph H. Camin and Robert R. Sokal. 1965. A Method for deducing Branching Sequences in Phylogeny. *Evolution* 19, 3 (1965), 311–326. https://doi.org/10.1111/j.1558-5646.1965.tb01722.x

[11] Gabriel Cardona, Mercè Llabrés, Francesc Rosselló, and Gabriel Valiente. 2009. Metrics for Phylogenetic Networks I: Generalizations of the Robinson-Foulds Metric. *IEEE ACM T. Comput. Bi.* 6, 1 (2009), 46–61. https://doi.org/10.1109/TCBB.2008.70

[12] Gabriel Cardona, Mercè Llabrés, Francesc Rosselló, and Gabriel Valiente. 2011. Comparison of Galled Trees. *IEEE ACM T. Comput. Bi.* 8, 2 (2011), 410–427. https://doi.org/10.1109/TCBB.2010.60

[13] Gabriel Cardona, Mercè Llabrés, Francesc Rosselló, and Gabriel Valiente. 2014. The Comparison of Tree-Sibling Time Consistent Phylogenetic Networks is Graph Isomorphism-Complete. *Sci. World J.* 2014, 254279 (2014). https://doi.org/10.1155/2014/254279

[14] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. 2008. A Distance Metric for a Class of Tree-Sibling Phylogenetic Networks. *Bioinformatics* 24, 13 (2008), 1481–1488. https://doi.org/10.1093/bioinformatics/btn231

[15] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. 2008. A Perl Package and an Alignment Tool for Phylogenetic Networks. *BMC Bioinformatics* 9, 175 (2008). https://doi.org/10.1186/1471-2105-9-175

[16] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. 2008. Tripartitions do not always discriminate Phylogenetic Networks. *Math. Biosci.* 211, 2 (2008), 356–370. https://doi.org/10.1016/j.mbs.2007.11.003

[17] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. 2009. Comparison of Tree-Child Phylogenetic Networks. *IEEE ACM T. Comput. Bi.* 6, 4 (2009), 552–569. https://doi.org/10.1109/TCBB.2007.70270

[18] William H. E. Day. 1985. Optimal Algorithms for comparing Trees with Labeled Leaves. *J. Classif.* 2, 1 (1985), 7–28. https://doi.org/10.1007/BF01908061

[19] Michel Marie Deza and Elena Deza. 2009. *Encyclopedia of Distances*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30958-8

[20] Zach DiNardo, Kiran Tomlinson, Anna Ritz, and Layla Oesper. 2020. Distance Measures for Tumor Evolutionary Trees. *Bioinformatics* 36, 7 (2020), 2090–2097. https://doi.org/10.1093/bioinformatics/btz869

[21] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael. 2016. Inferring the Mutational History of a Tumor using Multi-State Perfect Phylogeny Mixtures. *Cell Syst.* 3 (2016), 43–53. https://doi.org/10.1016/j.cels.2016.07.004

[22] James S. Farris. 1977. Phylogenetic Analysis under Dollo's Law. *Syst. Zool.* 26, 1 (1977), 77–88. https://doi.org/10.1093/sysbio/26.1.77

[23] Osamu Fujita. 2013. Metrics based on Average Distance between Sets. *Japan J. Indust. Appl. Math.* 30, 1 (2013), 1–19. https://doi.org/10.1007/s13160-012-0089-6

[24] Kiya Govek, Camden Sikes, and Layla Oesper. 2018. A Consensus Approach to infer Tumor Evolutionary Histories. In *Proc. 2018 ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics*. Association for Computing Machinery, New York, NY, 63–72. https://doi.org/10.1145/3233547.3233584

[25] Dan Gusfield. 1991. Efficient Algorithms for inferring Evolutionary Trees. *Networks* 21, 1 (1991), 19–28. https://doi.org/10.1002/net.3230210104

[26] Kathy J. Horadam and Michael A. Nyblom. 2014. Distances between Sets based on Set Commonality. *Discr. Appl. Math.* 167 (2014), 310–314. https://doi.org/10.1016/j.dam.2013.10.037

[27] Katharina T. Huber, Andreas Spillner, Radosław Suchecki, and Vincent Moulton. 2011. Metrics on Multilabeled Trees: Interrelationships and Diameter Bounds. *IEEE ACM T. Comput. Bi.* 8, 4 (2011), 1029–1040. https://doi.org/10.1109/TCBB.2010.122

[28] Wazim Mohammed Ismail, Etienne Nzabarushimana, and Haixu Tang. 2019. Algorithmic Approaches to Clonal Reconstruction in Heterogeneous Cell Populations. *Quant. Biol.* 7, 4 (2019), 255–265. https://doi.org/10.1007/s40484-019-0188-3

[29] Paul Jaccard. 1912. The Distribution of Flora in the Alpine Zone. *New Phytol.* 11, 2 (1912), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

[30] Katharina Jahn, Niko Beerenwinkel, and Louxin Zhang. 2020. The Bourque Distances for Mutation Trees of Cancer. In *Proc. 20th Int. Workshop Algorithms in Bioinformatics (Leibniz International Proceedings in Informatics)*, Nadia Pisanti and Carl Kingsford (Eds.), Vol. 172. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 14:1–14:23. https://doi.org/10.4230/LIPIcs.WABI.2020.14

[31] Katherine St. John. 2017. The Shape of Phylogenetic Treespace. *Syst. Biol.* 66, 1 (2017), e83–e94. https://doi.org/10.1093/sysbio/syw025

[32] Nikolai Karpov, Salem Malikic, Md. Khaledur Rahman, and S. Cenk Sahinalp. 2019. A Multi-Labeled Tree Dissimilarity Measure for Comparing "Clonal Trees" of Tumor Progression. *Algorithms Mol. Biol.* 14, 17 (2019). https://doi.org/10.1186/s13015-019-0152-9

[33] Kyung In Kim and Richard Simon. 2014. Using Single Cell Sequencing Data to Model the Evolutionary History of a Tumor. *BMC Bioinformatics* 15, 27 (2014). https://doi.org/10.1186/1471-2105-15-27

[34] Motoo Kimura. 1969. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population due to Steady Flux of Mutations. *Genetics* 61, 4 (1969), 893–903.

[35] Donald E. Knuth. 1997. *The Art of Computer Programming* (3rd ed.). Vol. 1: Fundamental Algorithms. Addison-Wesley, Boston, MA.

[36] Michael Levandowsky and David Winter. 1971. Distance between Sets. *Nature* 234 (1971), 34–35. https://doi.org/10.1038/234034a0

[37] Jian Ma, Aakrosh Ratan, Brian J. Raney, Bernard B. Suh, Webb Miller, and David Haussler. 2008. The Infinite Sites Model of Genome Evolution. *PNAS* 105, 38 (2008), 14254–14261. https://doi.org/10.1073/pnas.0805217105

[38] John H. Mason. 1972. Distance between Sets. *Nat. Phys. Sci.* 235 (1972), 80. https://doi.org/10.1038/physci235080a0

[39] Kurt Mehlhorn and Peter Sanders. 2016. *Algorithms and Data Structures: The Basic Toolbox*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77978-0

[40] Peter C. Nowell. 1976. The Clonal Evolution of Tumor Cell Populations. *Science* 194, 4260 (1976), 23–28. https://doi.org/10.1126/science.959840

[41] Barbara L. Parsons. 2008. Many Different Tumor Types have Polyclonal Tumor Origin: Evidence and Implications. *Mutat. Res.* 659, 1 (2008), 232–247. https://doi.org/10.1016/j.mrrev.2008.05.004

[42] Barbara L. Parsons. 2018. Multiclonal Tumor Origin: Evidence and Implications. *Mutat. Res.* 777, 1 (2018), 1–18. https://doi.org/10.1016/j.mrrev.2018.05.001

[43] Nicholas D. Pattengale, Eric J. Gottlieb, and Bernard M. E. Moret. 2007. Efficiently Computing the Robinson-Foulds Metric. *J. Comput. Biol.* 14, 6 (2007), 724–735. https://doi.org/10.1089/cmb.2007.R012

[44] David Posada and Keith A. Crandall. 2001. Intraspecific Gene Genealogies: Trees Grafting into Networks. *Trends Ecol. Evol.* 16, 1 (2001), 37–45. https://doi.org/10.1016/S0169-5347(00)02026-7
[45] Mark A. Ragan. 2009. Trees and Networks before and after Darwin. *Biol. Direct* 4, 43 (2009). https://doi.org/10.1186/1745-6150-4-43
[46] David F. Robinson and L. R. Foulds. 1981. Comparison of Phylogenetic Trees. *Math. Biosci.* 53, 1–2 (1981), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2
[47] Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M. Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, Zoya Kingsbury, Shujun Luo, David McBride, Lisa Murray, Toshi Menju, Adele Timbs, Mark Ross, Jenny Taylor, and David Bentley. 2012. Monitoring Chronic Lymphocytic Leukemia Progression by Whole Genome Sequencing reveals Heterogeneous Clonal Evolution Patterns. *Blood* 120, 20 (2012), 4191–4196. https://doi.org/10.1182/blood-2012-05-433540
[48] Russell Schwartz and Alejandro A. Schäffer. 2017. The Evolution of Tumour Phylogenetics: Principles and Practice. *Nat. Rev. Genet.* 18, 4 (2017), 213–229. https://doi.org/10.1038/nrg.2016.170
[49] Mícheál O. Searcóid. 2007. *Metric Spaces.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-1-84628-627-8
[50] Mike Steel. 2016. *Phylogeny: Discrete and Random Processes in Evolution.* Society for Industrial and Applied Mathematics, Philadelphia, PA. https://doi.org/10.1137/1.9781611974485
[51] Lajos Takács. 1993. Enumeration of Rooted Trees and Forests. *Math. Scientist* 18, 1 (1993), 1–10.
[52] Gabriel Valiente. 2002. *Algorithms on Trees and Graphs.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-04921-1
[53] Gabriel Valiente. 2009. *Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R.* Chapman & Hall/CRC, Boca Raton, FL. https://doi.org/10.1201/9781420069747
[54] Leo van Iersel, Judith Keijsper, Steven Kelk, Leen Stougie, Ferry Hagen, and Teun Boekhout. 2009. Constructing Level-2 Phylogenetic Networks from Triplets. *IEEE ACM T. Comput. Bi.* 6, 4 (2009), 667–681. https://doi.org/10.1007/978-3-540-78839-3_40
[55] Fabio Vandin. 2017. Computational Methods for Characterizing Cancer Mutational Heterogeneity. *Front. Genet.* 8, 83 (2017). https://doi.org/10.3389/fgene.2017.00083
[56] Rutger A. Vos, Jason Caravas, Klaas Hartmann, Mark A. Jensen, and Chase Miller. 2011. Bio::Phylo: Phyloinformatic Analysis using Perl. *BMC Bioinformatics* 12, 63 (2011). https://doi.org/10.1186/1471-2105-12-63

A APPENDIX

Definition A.1. A metric space [19, 49] is an ordered pair (M, d), where M is a set and $d : M \times M \to \mathbb{R}$ is a metric, that is, a function such that, for any $x, y, z \in M$, the following holds:

Separation $d(x, y) = 0$ if and only if $x = y$,
Symmetry $d(x, y) = d(y, x)$, and
Triangular inequality $d(x, z) \leqslant d(x, y) + d(y, z)$.

When d satisfies the symmetry and the triangular inequality, d is a pseudo-metric and (M, d) is a pseudo-metric space.

Let us fix a set X, the meaning of whose elements will depend on the application (somatic mutations, node labels, etc.). A *sub-multiset* A of X, or simply a *multiset* if the universe X is clear from the context, is a mapping $X \to \mathbb{N}$ that assigns to each element $x \in X$ its *multiplicity* in A, which we shall denote by $m_A(x)$. The intuition behind this definition is that A consists of $m_A(x)$ copies of each $x \in X$, with $m_A(x) = 0$ meaning that x does not appear in A. A *set* is a multiset all whose multiplicities are $\leqslant 1$. The *support* $Supp(A)$ of a multiset is the set of elements of X that have positive multiplicity in A: $Supp(A) = \{x \in X \mid m_A(x) \neq 0\}$.

For instance, if $X = \mathbb{N}$, the multiset A defined by $m_A(1) = 3$, $m_A(2) = 2$, $m_A(4) = 1$, $m_A(5) = 2$, and $m_A(n) = 0$ for any other $n \in \mathbb{N}$ corresponds to $A = \{1, 1, 1, 2, 2, 4, 5, 5\}$, and its support is $Supp(A) = \{1, 2, 4, 5\}$. The standard definitions in sets are easily translated to multisets as follows:

Equality and sub-multiset $A = B \leftrightarrow \forall x \in X[m_A(x) = m_B(x)]$ and $A \subseteq B \leftrightarrow \forall x \in X[m_A(x) \leqslant m_B(x)]$.
Union and intersection $m_{A \cup B}(x) = \max(m_A(x), m_B(x))$ and $m_{A \cap B}(x) = \min(m_A(x), m_B(x))$.
Difference $m_{A \setminus B}(x) = \max(0, m_A(x) - m_B(x))$.
Symmetric difference $m_{A \triangle B}(x) = m_{A \setminus B}(x) + m_{B \setminus A}(x)$.
Cardinality $|A| = \sum_{x \in X} m_A(x)$.

A multiset A is *finite* when its cardinality is finite.

Moreover, multisets allow for another, specific, operation that generalizes the union of sets: the *sum* $m_{A+B}(x) = m_A(x) + m_B(x)$.

For instance, for the multisets $A = \{1, 1, 1, 2, 2, 4, 5, 5\}$ and $B = \{1, 2, 2, 2, 2, 3, 3, 5, 5, 7\}$, $A \cup B = \{1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 5, 7\}$, $A + B = \{1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 5, 5, 7\}$, $A \cap B = \{1, 2, 2, 5, 5\}$, $A \setminus B = \{1, 1, 4\}$, $B \setminus A = \{2, 2, 3, 3, 7\}$, $A \triangle B = \{1, 1, 2, 2, 3, 3, 4, 7\}$, $|A| = 8$, $|B| = 10$, $|A \cup B| = 13$, $|A + B| = 18$, $|A \cap B| = 5$, $|A \triangle B| = 8$.

The following properties hold on multisets:

(1) $|A \cup B| = |A| + |B| - |A \cap B|$
(2) $|A + B| = |A| + |B|$
(3) $|A \setminus B| = |A| - |A \cap B|$
(4) $|\emptyset| = 0$
(5) If $A \cap B = \emptyset$, then $A \setminus B = A$ and $A \cup B = A + B$
(6) $A \triangle B = (A \setminus B) + (B \setminus A) = A \cup B - (A \cap B)$

Definition A.2. For every pair of multisets A, B, their *Jaccard distance* is

$$d_J(A, B) = \frac{|A \triangle B|}{|A \cup B|} = \frac{|A \setminus B| + |B \setminus A|}{|A \cup B|}$$

For instance, for the pair of multisets A, B given above, $d(A, B) = 8/13$.