TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/11 pp490-499
Volume 18, Number 5, October 2013

Distances Between Phylogenetic Trees: A Survey

Feng Shi, Qilong Feng*, Jianer Chen, Lusheng Wang, and Jianxin Wang

Abstract: Phylogenetic trees have been widely used in the study of evolutionary biology for representing the tree-like evolution of a collection of species. However, different data sets and different methods often lead to the construction of different phylogenetic trees for the same set of species. Therefore, comparing these trees to determine similarities or, equivalently, dissimilarities, becomes the fundamental issue. Typically, Tree Bisection and Reconnection (TBR) and Subtree Prune and Regraft (SPR) distances have been proposed to facilitate the comparison between different phylogenetic trees. In this paper, we give a survey on the aspects of computational complexity, fixed-parameter algorithms, and approximation algorithms for computing the TBR and SPR distances of phylogenetic trees.

Key words: phylogenetic tree; tree bisection and reconnection; subtree prune and regraft; fixed-parameter algorithm; approximation algorithm

1 Introduction

In biology, phylogenetic trees are used to describe the evolutionary relationships among groups of species (e.g., organisms and populations); these relationships are derived from the molecular sequencing data and morphological data matrices. The leaves of phylogenetic trees are labeled as the species, and the internal nodes correspond to speciation events. If the evolutionary origin is given, then the phylogenetic tree is termed as *rooted*; otherwise, it is termed as *unrooted*.

Constructing a phylogenetic tree is the fundamental computational problem in phylogenetics. Several methods have been proposed based on various criteria, including (not exhaustively) parsimony[1-3], compatibility[4], distance[5,6], and maximum likelihood[1,7,8]. Therefore, given the same set of species, different data sets and different methods result in the construction of different trees. Therefore, it is worthwhile to compare such different phylogenetic trees. In order to facilitate the comparison of different phylogenetic trees, several metrics for measuring the distance between phylogenetic trees have been proposed, such as *Robinson-Foulds* distance[9], *Nearest Neighbor Interchange* (NNI) distance[10-19], *Tree Bisection and Reconnection* (TBR) distance and *Subtree Prune and Regraft* (SPR) distance[20-23]. Among them, the TBR and SPR distances have been extensively studied in the literature.

In this paper, we provide an analysis of the computational complexity, fixed-parameter algorithms, and approximation algorithms for the TBR and SPR distance problems as well as the *Maximum Agreement Forest* (MAF) problem on phylogenetic trees.

2 Related Terminologies

The following definitions follow the ones in Refs. [24, 25].

Given a fixed label-set X, each label in X corresponds to a specific extant species. An *unrooted*

• Feng Shi, Qilong Feng, Jianer Chen, and Jianxin Wang are with School of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: {fengshi, csufeng, jianer, jxwang}@csu.edu.cn.

• Jianer Chen is also with Department of Computer Science and Engineering, Texas A&M University, College Station, Texas 77843-3112, USA.

• Lusheng Wang is with Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China. E-mail: lwang@cs.cityu.edu.hk.

* To whom correspondence should be addressed.

Manuscript received: 2013-08-06; revised: 2013-08-15; accepted: 2013-08-20

binary phylogenetic tree—simply, an unrooted binary *X-tree*—is a tree whose leaves are labeled bijectively by the label-set X; further, each internal node is unlabeled and has a degree 3. If a particular leaf is designated as the *root* (such that it is both a root and a leaf labeled with the special symbol ρ) that specifies a unique ancestor-descendant relationship in the tree, then it is called a *rooted binary phylogenetic tree*—simply, a rooted binary *X-tree*. Recently, the *unrooted multifurcating phylogenetic tree*—simply, an unrooted multifurcating *X-tree*—has received considerable attention; the leaves of such an *X-tree* are also labeled bijectively by the label-set X and each internal node is unlabeled; however, the degree of each internal node is not less than 3.

A *forced contraction* is an operation performed on a phylogenetic tree that replaces each degree-2 vertex v and its incident edges with a single edge connecting the two neighbors of v and removes each unlabeled vertex that has degree smaller than 2.

An SPR operation on a binary X-tree T is defined as the removal of any edge in T, and therefore, pruning a subtree T' and then regrafting the subtree T' with the same removed edge to a new vertex obtained by subdividing a pre-existing edge in T. A forced contraction is applied to the resulting tree in order to delete the degree-2 vertex. To distinguish between the operations on rooted and unrooted trees, we will refer to the corresponding operations as rSPR and uSPR. Figure 1 shows a schematic representation of the SPR operation.

A TBR operation on a binary X-tree T is defined as the removal of any edge in T, resulting in two subtrees T_1 and T_2, which are then reconnected by creating a new edge between the midpoints of any edge in T_1 and any edge in T_2. A forced contraction is applied to the resulting tree. The TBR operation is always defined on an unrooted phylogenetic tree. Figure 1 also shows the schematic representation of a TBR operation.

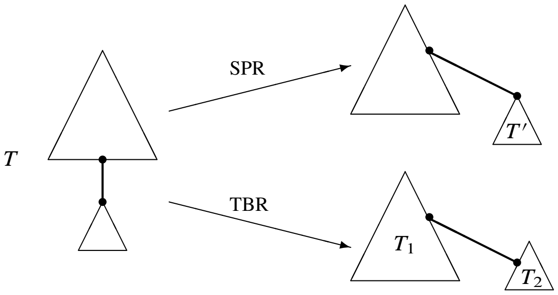

Fig. 1 Schematic representations of SPR and TBR operations.

The SPR and TBR distances between two X-trees T_1 and T_2 with identical label-sets X are defined as the minimum number of SPR and TBR operations required to transform T_1 into T_2, which are denoted by $d_{\mathrm{SPR}}(T_1, T_2)$ and $d_{\mathrm{TBR}}(T_1, T_2)$, respectively. Evidently, $d_{\mathrm{SPR}}(T_1, T_2) = d_{\mathrm{SPR}}(T_2, T_1)$ and $d_{\mathrm{TBR}}(T_1, T_2) = d_{\mathrm{TBR}}(T_2, T_1)$. Since the TBR operation is a generalization of the SPR operation, the TBR operation can be simulated using two SPR operations; therefore, we have $d_{\mathrm{SPR}}(T_1, T_2) \leqslant 2d_{\mathrm{TBR}}(T_1, T_2)$.

The definitions of the TBR and SPR distance problems are given below.

• TBR (SPR) Distance Problem: Given two phylogenetic trees T_1 and T_2 with identical label-sets, use the minimum number of TBR (SPR) operations to transform T_1 into T_2.

• Parameterized TBR (SPR) Distance Problem: Given two phylogenetic trees T_1 and T_2 with identical label-sets and a parameter k, can T_1 be transformed into T_2 by performing no more than k TBR (SPR) operations?

Given two phylogenetic trees T_1 and T_2 with identical label-sets, the MAF models—graphical theoretical models—are formulated for T_1 and T_2 involving the TBR and SPR distances[26]. Before the MAF is defined, we describe a few related terminologies.

A *subtree* T' of an unrooted X-tree T is a connected subgraph of T that contains at least one leaf in T (if T' consists of only one leaf, then it is a *single-vertex tree*). A *subforest* of an unrooted X-tree T is a subgraph of T. An unrooted X*-forest* F is a *subforest* of an unrooted X-tree T that contains all the leaves of T such that each connected component of F contains at least one leaf in T. Therefore, an unrooted X-forest F is a collection of leaf-labeled trees whose label-sets are disjoint such that the union of the label-sets is equal to X. Assume that the forced contraction operation is applied immediately whenever applicable. Two labels a and b in an unrooted X-forest F are called *siblings* if any one of them is adjacent to the same non-leaf vertex in F, which is called the "parent" of a and b.

A *subtree* T' of a rooted X-tree T is a connected subgraph of T that contains at least one leaf in T. In order to preserve the ancestor-descendant relationship in T, the root of the subtree of T should be defined. If

T' contains the leaf ρ, then it is the root of the subtree; otherwise, the node in T' that is the least common ancestor of the leaves in T' is defined as the root of T'. A (rooted) X*-forest* F is a *subforest* of a rooted X-tree T that contains a collection of subtrees whose label-sets are disjoint such that the union of the label-sets is equal to X. Therefore, one of the subtrees in a rooted X-forest F should have the vertex labeled ρ as its root. Assume that the forced contraction operation is applied immediately whenever applicable. However, if the root r of a subtree T' is of degree 2, then the forced contraction operation is not applied on r in order to preserve the ancestor-descendant relationship in T.

If a leaf-labeled forest F' is isomorphic to a subforest of an X-forest F (up to the forced contraction), then we simply say that F' is a *subforest* of F. An X-forest is an *agreement forest* for two X-trees if it is a *subforest* of any trees. We define the *order* of a forest F as the number of connected components in F, and it is denoted by $\mathrm{Ord}(F)$. The MAF for two trees is an agreement forest of the minimum order.

The MAF and parameterized MAF problems are formally defined as follows:

- MAF Problem: Given two phylogenetic trees T_1 and T_2 with identical label-sets, construct an MAF for T_1 and T_2.
- Parameterized MAF Problem: Given two phylogenetic trees T_1 and T_2 with identical label-sets and a parameter k, is there an agreement forest having a size of at the most k for T_1 and T_2?

3 TBR Distance

3.1 TBR distance and order of MAF

Allen and Steel[24] proved that the TBR distance between two unrooted binary phylogenetic trees is equal to the order of the MAF of the two trees minus 1 and they used the proof derived by Hein et al.[26] to show that the TBR distance problem on two unrooted phylogenetic trees is NP-hard.

Yang[27] studied the MAF problem on two unrooted multifurcating trees and proved that the TBR distance between these trees is also equal to the order of the MAF minus 1. In the following, we provide a detailed proof for two unrooted multifurcating trees, which is also applicable to unrooted binary trees.

First, we extend the definition of the TBR operation with respect to multifurcating trees.

Definition 1[27] A TBR operation on an unrooted multifurcating phylogenetic tree T is defined as the removal of any edge, yielding two subtrees T_1 and T_2, which are then reconnected by a new edge e. Each end of the edge e can be on either a non-leaf vertex or the midpoint of the edges in T_1 and T_2.

Note that in the definition of the TBR operation on binary trees, it is required that the ends of the new edge e be on the midpoints of the edges in T_1 and T_2 to ensure that the resulting tree is a binary tree. For multifurcating trees, we relax this condition and allow the ends of the new edge e to join the non-leaf vertices in T_1 and T_2.

Theorem 1[27] Let T_1 and T_2 be any two unrooted multifurcating phylogenetic trees with identical label-sets; then, $d_{\mathrm{TBR}}(T_1, T_2) = \mathrm{MAF}(T_1, T_2) - 1$, where $\mathrm{MAF}(T_1, T_2)$ is the order of the MAF for T_1 and T_2.

Proof We prove the theorem by applying mathematical induction on $d_{\mathrm{TBR}}(T_1, T_2)$ and $\mathrm{MAF}(T_1, T_2)$.

For $d_{\mathrm{TBR}}(T_1, T_2) = 0$, T_1 and T_2 are isomorphic; T_1 itself is an MAF for T_1 and T_2, so $\mathrm{MAF}(T_1, T_2) = 1$; hence, the theorem holds true. For $d_{\mathrm{TBR}}(T_1, T_2) = 1$, we remove an edge e in T_1, resulting in two subtrees T' and T''; then, they are connected by a new edge e' to obtain T_2. Therefore, $\{T', T''\}$ is an agreement forest for T_1 and T_2. Because $T_1 \neq T_2$, $\mathrm{MAF}(T_1, T_2) > 1$, so $\{T', T''\}$ is an MAF for T_1 and T_2, i.e., $\mathrm{MAF}(T_1, T_2) = 2$. The theorem again holds true.

Now, suppose that the hypothesis holds true for pairs of unrooted multifurcating phylogenetic trees with a TBR distance of $d \geqslant 1$ and suppose $d_{\mathrm{TBR}}(T_1, T_2) = d + 1$. For T_1 and T_2, there must exist a tree T_3 such that $d_{\mathrm{TBR}}(T_1, T_3) = d$ and $d_{\mathrm{TBR}}(T_3, T_2) = 1$. Therefore, by the inductive hypothesis, there exists an MAF $F = \{T_1', T_2', \cdots, T_d', T_{d+1}'\}$ of size $d + 1$ for T_1 and T_3 and an MAF $F' = T_3 \setminus \{e\}$ of size 2 for T_3 and T_2. Since F is a subforest of T_1, $F \setminus \{e\}$ is also a subforest of T_1. Since F is a subforest of T_3, $F \setminus \{e\}$ is also a subforest of T_2. Therefore, $F \setminus \{e\}$ is an agreement forest for T_1 and T_2. The order of forest $F \setminus \{e\}$ is at the most $d + 2$, i.e., $\mathrm{MAF}(T_1, T_2) \leqslant d + 2$. This shows that $d_{\mathrm{TBR}}(T_1, T_2) \geqslant \mathrm{MAF}(T_1, T_2) - 1$.

Now, we again use mathematical induction on $\mathrm{MAF}(T_1, T_2)$ to show that $d_{\mathrm{TBR}}(T_1, T_2) \leqslant \mathrm{MAF}(T_1, T_2) - 1$. For $\mathrm{MAF}(T_1, T_2) = 1$, T_1 itself is an MAF for T_1 and T_2, i.e., $d_{\mathrm{TBR}}(T_1, T_2) = 0$; therefore, the theorem holds true. For $\mathrm{MAF}(T_1, T_2) = 2$, we can obtain an MAF by removing a single edge from each T_1 and T_2; hence, $d_{\mathrm{TBR}}(T_1, T_2) = 1$. Suppose that the hypothesis holds true for pairs of unrooted multifurcating phylogenetic trees with an MAF $F = \{T'_1, T'_2, \cdots, T'_m\}$ of size $m \geqslant 1$ and suppose $\text{MAF}(T_1, T_2) = m + 1$. Since $T'_1, T'_2, \cdots, T'_m, T'_{m+1}$ are disjoint in T_1, there exists a simple path P in T_1 that connects two trees in F such that no internal vertex in P exists in F. Without loss of generality, suppose that the path P has its two end-vertices v_1 and v_2 in T'_1 and T'_2, respectively. Now, we construct a new tree T_3 as follows. First, add a new edge e_1 between v_1 and v_2 in T_2, which causes a unique cycle in $T_2 \cup \{e_1\}$; therefore, there exists an edge e_2 in the cycle that does not exist in F. Now, T_3 is constructed by removing the edge e_2 in $T_2 \cup \{e_1\}$. Note that the subtree $T' = T'_1 \cup T'_2 \cup \{e_1\}$ is a subtree in both T_1 and T_3. Therefore, $\{T', T'_3, \cdots, T'_m, T'_{m+1}\}$ is an agreement forest for T_1 and T_3, i.e., $\text{MAF}(T_1, T_3) \leqslant m$; therefore, by the inductive hypothesis, $d_{\text{TBR}}(T_1, T_3) \leqslant m - 1$. Note that T_2 differs from T_3 by exactly one TBR operation, i.e., $d_{\text{TBR}}(T_3, T_2) = 1$. Therefore, $d_{\text{TBR}}(T_1, T_2) \leqslant d_{\text{TBR}}(T_1, T_3) + d_{\text{TBR}}(T_3, T_2) \leqslant m = \text{MAF}(T_1, T_2) - 1$. This completes the proof of the theorem. ■

3.2 Fixed-parameter algorithms for TBR distance

The TBR distance problem is NP-hard; therefore, we need to study approximation algorithms or fixed-parameter algorithms for determining its solution. A parameterized problem is *Fixed-Parameter Tractable* (FPT)[28] if it is solvable in a time of $f(k)n^{O(1)}$.

Allen and Steel[24] proved that the parameterized TBR distance problem on two unrooted binary phylogenetic trees is FPT by proving the problem is kernelizable[29] with the following reduction rules.

- **Subtree Reduction Rule**. Replace any pendant subtree that occurs identically in both the trees by a single leaf with a new label.
- **Chain Reduction Rule**. Replace any chain of pendant subtrees that occur identically in both the trees by three new leaves with new labels correctly oriented to preserve the direction of the chain.

Given two phylogenetic trees T_1 and T_2, the kernelization algorithm applies the above rules on T_1 and T_2 whenever applicable. After recursively applying these rules, the resulting trees have size $n' \leqslant 4c(k-1)$, where c is a constant and k is the given parameter such that the TBR distance between the resulting trees remains the same as the TBR distance between the two original trees. Note that n' is independent of the leaf-set size n of the original trees T_1 and T_2. There are $O(k^3)$ possible TBR operations that can be performed on the resulting trees and an exhaustive search can be used to determine if there exist at the most k operations that can transform T_1 into T_2. Therefore, Allen and Steel[24] proposed a fixed-parameter algorithm with a running time of $O(k^{3k} + p(n))$, where $p(n)$ is the running time required to apply the reduction rules. Therefore, the parameterized TBR distance problem is FPT.

Hallett and McCartin[30] developed a parameterized algorithm with a running time of $O(4^k k^5 + n^{O(1)})$ for the MAF problem on two unrooted binary phylogenetic trees. Their algorithm proceeds in two phases. In the first phase, their algorithm determines all the possible minimal incompatible quartets, each of which need to be eliminated to construct the MAF for the two given trees T_1 and T_2. They have shown that such a quartet can be removed in exactly four ways, leading to four branches in the search tree, with a single edge removed in each case. In the second phase, their algorithm determines all the possible obstructions, each of which needs to be eliminated to construct the MAF. They have shown that each obstruction can be removed in two ways, leading to two branches in the search tree, with a single edge removed in each case. They have also defined the minimal incompatible quartet and obstruction. Note that at the most $k - 1$ edges are removed to yield any solution. Therefore, the depth of the search tree is bounded by k. Hence, the size of the search tree is bounded by 4^k. Each iteration takes a time of $O(n^5)$. Hence, the algorithm takes a time of $O(4^k n^5)$.

Whidden and Zeh[31] further improved the time complexity to $O(4^k n)$. Given two unrooted binary phylogenetic forests F_1 and F_2 with identical label-sets, their algorithm fully utilizes the relationship among the sibling leaves in trees. For an arbitrary sibling pair (a, b) in F_2, their algorithm undertakes corresponding operations according to the three cases for a and b in F_1, until no sibling pair exists in F_2. Suppose that F^* is a fixed MAF for F_1 and F_2. Case 1: a and b are siblings in F_1. Then, a and b must be siblings in F^*; a and b can be merged, and the parents of a and b in both the forests are labeled with a new label. Case 2: a and b are in different components of F_1. Then, at least one of a or b is a single-vertex tree in F^*. Therefore, either the edge incident to a or that to b in forest F_1 can be removed, leading to two branches in the search tree. Case 3: a and b are in the same component of F_1. There are three possibilities for a and b in F^*: a

is a single-vertex tree, b is a single-vertex tree, or a and b are siblings. Let $P = \{a, c_1, c_2, \cdots, c_r, b\}$ be the unique path in F_1 that connects a and b, where $r \geqslant 2$. The cases in which either a or b is a single-vertex tree in F^* result in the removal of the edge incident to a or b in F_1. In order to ensure a and b become siblings in F_1, at the most one of the edges that is not on the path P but is incident to a vertex in P can be retained. Since the subtree in an unrooted forest does not need to preserve any ancestor-descendant relationship, anyone of these edges can be retained. On the other hand, since $r \geqslant 2$, at least one of the two edges, which is not on the path P but is incident to c_1 and c_r, needs to be removed. Therefore, there are two branches that require removing—either the edge incident to c_1 or the edge incident to c_r. Consequently, there are four branches in the search tree for Case 3, with a single edge removed from the forest F_1 in each branch. Since at the most $k - 1$ edges can be removed to yield any solution, the depth of the search tree is bounded by k and its size is bounded by 4^k.

3.3 Approximation algorithms for TBR distance

The parameterized algorithm proposed by Whidden and Zeh[31] for the MAF problem on two unrooted binary phylogenetic trees can lead to an approximation algorithm with a ratio of 4, but the ratio was improved to 3 in Ref. [31]. Let $e(F_1, F_2, F)$ denote the size of the smallest edge set E such that $F \setminus E$ is an agreement forest of F_1 and F_2, where F is a subforest of F_1. Suppose a and b are siblings in F_2 and are not siblings in F_1, and c is the node having a common parent with a in F_1 and e_c is the edge between c and its parent. Let e_a and e_b be the edges incident to a and b in F_1, respectively. They have shown that $e(F_1, F_2, F \setminus \{e_a, e_b, e_c\}) \leqslant e(F_1, F_2, F) - 1$. Therefore, the number of edges removed in F is at the most three times of $e(F_1, F_2, F_1)$.

3.4 TBR distance between two unrooted multifurcating phylogenetic trees

Most of the earlier studies on MAF have been restricted to binary trees. The TBR distance problem on multifurcating trees has been recently investigated. Yang[27] proved that the TBR distance between two unrooted multifurcating phylogenetic trees is equal to the order of the MAF minus 1 and presented an FPT algorithm for the MAF problem on multifurcating phylogenetic trees with a running time of $O(4^k n^5)$. Their algorithm closely follows the idea of the algorithm on binary phylogenetic trees in Ref. [30] in which the minimal incompatible quartets and obstructions are eliminated. However, the algorithm in Ref. [27] needs to handle the additional star quartet structures and non-binary structures in its analysis, which are much more complicated than those of binary trees.

Chen et al.[32] proposed a $O(3^k n)$-time parameterized algorithm for the MAF problem on two unrooted multifurcating trees, which is also currently the best available algorithm for the MAF problem on two unrooted binary phylogenetic trees. A *Bottommost Sibling Set* (BSS) is a maximal sibling set X such that either the degree of the parent of X is at the most $|X| + 1$ or X is the leaf set of a single-edge tree. Given two unrooted phylogenetic forests F_1 and F_2 with identical label-sets, their algorithm arbitrarily selects a BSS from F_2 and analyzes the possible cases for the BSS in F_1.

Based on the analysis of BBS, Chen et al.[32] also developed an approximation algorithm with a ratio of 3 for the MAF problem on unrooted multifurcating trees, which is the first constant-ratio approximation algorithm for the MAF problem on unrooted multifurcating trees.

4 SPR Distance

A uSPR operation on an unrooted binary phylogenetic tree T is defined as the removal of any edge (u, v) in T, and therefore, pruning two subtrees T_u and T_v and then either regrafting the subtree T_u by connecting the node u to a new vertex obtained by subdividing a pre-existing edge in T_v or regrafting the subtree T_v by connecting the node v to a new vertex obtained by subdividing a pre-existing edge in T_u. An rSPR operation on a rooted binary phylogenetic tree T is defined as the removal of any edge (u, v) in T, where u is on the path from the root of the tree to v, and therefore, pruning a subtree T_v, and then regrafting the subtree T_v by connecting the node v to a new vertex obtained by subdividing a pre-existing edge in the component C_u that contains u.

Hein et al.[26] proved that the SPR distance problem is NP-hard. First, their reductions transform an instance of a known NP-complete problem, namely, the exact cover by 3-sets (X3C), into an instance of MAF of two rooted phylogenetic trees with identical label-sets; then, the order of the MAF of two phylogenetic trees is transformed into the SPR distance. They have specified the reduction from the order of the MAF to the SPR distance; that is, the order of the MAF for T_1 and T_2 is one more than the SPR distance for any pair of rooted (or unrooted) phylogenetic trees T_1 and T_2 with identical label-sets. Unfortunately, Allen and Steel[24] found subtle mistakes in the proofs of the reduction in Ref. [26] while transforming the order of the MAF of two phylogenetic trees into the SPR distance and provided counterexamples to show that the reduction in Ref. [26] is neither true for unrooted trees, nor for SPR transformations on rooted trees.

Bordewich and Semple[33] used a revised definition of the MAF to prove that the computation of the rSPR distance between two rooted trees is NP-hard. Hickey et al.[34] used the polynomial-time reduction from X3C to MAF to prove that the computation of the uSPR distance between two unrooted trees is NP-hard.

4.1 uSPR distance problem is FPT

The two reductions in Ref. [24] are essential to the proof of FPT for TBR[24], providing a method to reduce the initial trees to smaller trees (with equivalent distances) whose sizes are bounded by the distance between the trees. The first reduction rule (subtree reduction) preserves the uSPR distances between the trees. It is unknown whether the second reduction rule (chain reduction) preserves the uSPR distance between the trees. In order to prove that the parameterized uSPR distance problem on two unrooted binary trees is FPT, Bonet and John[35] introduced a new reduction rule that is a variant of the chain reduction rule.

- ***c*-chain reduction rule**. Replace a chain of pendant leaves that occurs identically in both the trees by c new leaves with new labels correctly oriented to preserve the direction of the chain.

Given two unrooted binary phylogenetic trees T_1 and T_2 with identical label-sets, the kernelization algorithm applies the subtree reduction rule and $9k$-chain reduction rule on T_1 and T_2 if possible. Let T_1' and T_2' be the resulting trees when no reduction rules can be applied on T_1 and T_2. It is shown in Ref. [35] that $d_{\text{uSPR}}(T_1, T_2) = d_{\text{uSPR}}(T_1', T_2')$ and if $d_{\text{uSPR}}(T_1, T_2) \leqslant k$, then $|T_1'| \leqslant 76k^2$. The size of T_1' is bounded by $76k^2$, which is independent of the leaf set size n of T_1. There are at the most $O(k^4)$ possible uSPR operations that can be performed on the resulting trees. Then, an exhaustive search can be performed to determine if there exists at the most k operations that can transform T_1 into T_2. Therefore, there exists a parameterized algorithm for solving the parameterized uSPR distance problem with a running time of $O(k^{4k} + p(n))$, where $p(n)$ is the running time of the kernelization algorithm. Therefore, the parameterized uSPR distance problem is FPT.

4.2 Fixed-parameter algorithms for rSPR distance

The subtree reduction rule proposed by Allen and Steel[24] preserves the rSPR distance, but is not preserved by the chain reduction rule[24]. In order to prove that the parameterized rSPR distance problem is FPT, Bordewich and Semple[33] proposed a modified version of the chain reduction rule that preserves the rSPR distance.

- **Rooted chain reduction rule**. Replace any chain of the pendant subtrees that occur identically and with the same orientation relative to the root in both the trees by three new leaves with new labels correctly oriented to preserve the direction of the chain.

Let T_1 and T_2 be two rooted binary phylogenetic trees with identical label-sets. The kernelization algorithm in Ref. [33] applies the subtree reduction rule and rooted chain reduction rule on T_1 and T_2 if applicable. Let T_1' and T_2' be the resulting trees when no reduction rules can be applied on T_1 and T_2. It is shown in Ref. [33] that $d_{\text{rSPR}}(T_1, T_2) = d_{\text{rSPR}}(T_1', T_2')$ and if $d_{\text{rSPR}}(T_1, T_2) \leqslant k$, then $|T_1'| \leqslant 28k$. There are at the most $(56k)^2$ possible rSPR operations that can be performed on the resulting trees. Then, an exhaustive search can be used to determine if there exists at the most k operations that can transform T_1 into T_2. Therefore, Bordewich and Semple[33] proposed a parameterized algorithm with a running time of $O((56k)^{2k} + p(n))$, where $p(n)$ is the running time of the kernelization algorithm. Therefore, the parameterized rSPR distance problem is FPT.

Bordewich and Semple[33] proved that the rSPR distance between two rooted binary phylogenetic trees is equal to the order of the rooted version of the MAF minus 1.

For the MAF problem on two rooted binary phylogenetic trees, Bordewich et al.[36] developed a parameterized algorithm with a running time of $O(4^k k^4 + n^3)$. The idea of their algorithm closely follows that of the parameterized algorithm proposed by Hallett and McCartin[30] for the MAF problem on two unrooted binary phylogenetic trees. First, the algorithm in Ref. [36] finds a minimal incompatible triple that can be deleted by removing each of the associated four edges in four ways, leading to four branches in the search tree. The minimal incompatible triple is defined in Ref. [36]. When there are no more incompatible triples between F and T, the algorithm iteratively finds the components of F overlapping in T that can be deleted by removing each of the two associated edges in exactly two ways, leading to two branches in the search tree. Since the algorithm separates into at the most four branches in each iteration and each iteration takes a time of $O(n^4)$, it follows that the algorithm takes a time of $O(4^k n^4)$.

Whidden et al.[37] developed a parameterized algorithm with a running time of $O(2.42^k k + n^3)$, which is currently the best available algorithm for the MAF problem on two rooted binary phylogenetic trees. The idea of the algorithm in Ref. [37] follows the algorithm in Ref. [31] for the MAF problem on unrooted trees. For an arbitrary sibling pair (a, b) in the rooted binary phylogenetic forest F_2, the algorithm in Ref. [37] performs corresponding operations according to the three cases for a and b in the rooted binary phylogenetic forest F_1 until no sibling pair exists in F_2. Suppose that F^* is a fixed MAF for F_1 and F_2. Case 1: a and b are siblings in F_1. Then, a and b must be siblings in F^*; a and b can be merged and the parents of a and b in both the forests are labeled with a new label. Case 2: a and b are in different components of F_1. Then, at least one of a and b is a single-vertex tree in F^*. Therefore, either the edge incident to a or the edge incident to b in forest F_1 can be removed, resulting in two branches in the search tree. Case 3: a and b are in the same component of F_1. Let $P = \{a, c_1, c_2, \cdots, c_h, \cdots, c_r, b\}$ be the unique path in F_1 that connects a and b, where c_h is the least common ancestor of a and b, $1 \leqslant h \leqslant r$. Since a and b are not siblings in F_1, $r \geqslant 2$. There are three possibilities for a and b in F^*: a is a single-vertex tree, b is a single-vertex tree, or a and b are siblings. For the cases in which either a or b is a single-vertex tree in F^*, the edge incident to a or b in F_1 is removed. If a and b are siblings in F^*, all the edges that are not on the path P but are incident to a vertex c_j in P, where $j \neq h$, should be removed. Based on the above analysis, the size of the search tree is bounded by $O(3^k)$. In order to get an improved result, Whidden et al.[37] performed further analysis of Case 3. Case 3-1: $r = 2$. Their algorithm only removes the edge incident to c_i, $i \neq h$. Case 3-2: $r \geqslant 3$. Their algorithm separates into three branches: remove e_a or e_b or all the edges that are not on the path P but are incident to a vertex c_j in P, where $j \neq h$. Since there is a branch that removes at least two edges in Case 3-2, the size of the search tree is reduced to $O(2.42^k)$.

Chen and Wang[38] implemented the $O(3^k n)$-time algorithm proposed by Whidden et al.[37] for computing a maximum (acyclic) agreement forest, which can output all the maximum (acyclic) agreement forests. The program can be augmented to construct an optimal hybridization network for each given maximum (acyclic) agreement forest. To the best of our knowledge, this is the first time that optimal hybridization networks could be rapidly constructed.

4.3 Approximation algorithms for rSPR distance

Hein et al.[26] developed a 3-approximation algorithm for the MAF problem on two rooted binary trees. For a sibling pair of a and b in the tree T_2, if a and b are siblings in the tree T_1, their algorithm replaces this pair of a and b with a new leaf labeled as (a, b) in both the trees. Otherwise, the algorithm removes the corresponding edge set in T_1 according to the relationship between a and b in T_1 until a and b become siblings or become separated. They have investigated five cases based on the relationship between a and b in T_1. Eventually, both the trees are cut into the same forest. Rodrigues et al.[39] described a family of instances that shows that the approximation ratio of the algorithm in Ref. [26] is not 3, but 4. Rodrigues et al.[39] also proposed two 3-approximation algorithms for the MAF problem on two rooted binary trees which are similar to the algorithm in Ref. [26].

However, Bonet et al.[40] proved that both the ratios of the algorithms in Refs. [26] and [39] are 5 for the rSPR distance on two rooted binary phylogenetic trees. A 5-approximation algorithm with a linear running time was proposed by Bonet et al.[40].

Based on the parameterized algorithm for the MAF problem on two rooted binary trees, Bordewich et al.[36] proposed a 3-approximation algorithm with a running time of $O(n^5)$. Given two rooted binary phylogenetic trees T_1 and T_2 with identical label-sets, their approximation algorithm proceeds by removing the edges from T_1 to obtain a forest F of T_1; until F yields an agreement forest for T_1 and T_2. To obtain such a forest, their algorithm iteratively finds a minimal incompatible triple $ab|c$ of F with respect to T_2, and removes the associated edges e_a, e_c, and e_r from F. When there is no incompatible triple of F

with respect to T_2, the algorithm iteratively finds the components T_s and T_t of F that overlap in T_2, and removes the associated edges e_s and e_t. When there are no overlapping components in T_2, F becomes an agreement forest for T_1 and T_2. Let $e(T_1, T_2, F)$ denote the size of the smallest edge set E such that $F \setminus E$ is an agreement forest for T_1 and T_2, where F is a subforest of T_1. They have shown that whenever a set of edges is removed from F corresponding to either an incompatible triple of F with respect to T_2 or a pair of components in F that overlap in T_2, the value $e(T_1, T_2, F)$ decreases by at least one. Since the number of edges removed from F at each iteration is at the most 3, the number of edges that is removed from T_1 is at the most three times of $e(T_1, T_2, T_1)$.

The parameterized algorithm proposed by Whidden et al.[37] for the MAF problem on two rooted binary phylogenetic trees leads to a 3-approximation algorithm with a linear running time.

Rodrigues et al.[39] extended the approximation algorithm for the MAF problem on two rooted binary trees to two rooted multifurcating trees with bounded degree $d \geqslant 2$ and has a ratio of $d + 1$, where d is the maximum number of children that are possible for a node in the trees.

5 MAF Problem for Multiple Binary Phylogenetic Trees

There are several methods for constructing phylogenetic trees. Therefore, for the same set of species, more than two phylogenetic trees can be constructed. Hence, it is worthwhile to investigate the MAF problem on more than two phylogenetic trees. An MAF of order k for a set of phylogenetic trees implies that for any two phylogenetic trees in a given set, one of them can be obtained from the other by performing no more than $k - 1$ TBR operations (for unrooted trees) or no more than $k - 1$ rSPR operations (for rooted trees).

Shi et al.[25] focused on the MAF problem on multiple (i.e., two or more) binary phylogenetic trees for both the rooted and unrooted cases. An $O(3^k n)$-time parameterized algorithm for the MAF problem on multiple rooted binary phylogenetic trees and an $O(4^k n)$-time parameterized algorithm for the MAF problem on multiple unrooted binary phylogenetic trees are presented in Ref. [25]. Let $\mathcal{C} = \{T_1, T_2, \cdots, T_m\}$ be a collection of rooted or unrooted binary phylogenetic trees. Note that for any MAF F of order k for the trees in $\mathcal{C}$, F must be an agreement forest for the trees T_1 and T_2, which is not necessarily the maximum.

Shi et al.[25] analyzed all the agreement forests for T_1 and T_2 and proposed the terminology of *maximal agreement forest*. The *maximal agreement forest* F for T_1 and T_2 is an agreement forest that there is no agreement forest F' for T_1 and T_2 such that F is a subforest of F' and $\text{Ord}(F') < \text{Ord}(F)$. For any agreement forest F for T_1 and T_2, F must be a subforest of a maximal agreement forest of T_1 and T_2. An MAF for T_1 and T_2 is also a maximal agreement forest for T_1 and T_2. They have shown that for every MAF F for $\{T_1, T_2, \cdots, T_m\}$, there is a maximal agreement forest F^* for T_1 and T_2 such that F is also an MAF for $\{F^*, T_3, \cdots, T_m\}$.

The general idea of the algorithms in Ref. [25] works as follows: (1) construct a collection $\mathcal{C}$ of agreement forests for T_1 and T_2 that contains all the maximal agreement forests F^* for T_1 and T_2 with $\text{Ord}(F^*) \leqslant k$; (2) for each agreement forest F for T_1 and T_2 constructed in (1), perform the operation recursively on the instance $(F, T_3, \cdots, T_m; k)$.

6 Conclusions

In this paper, we survey the results involving the computational complexity, fixed-parameter algorithms, and approximation algorithms for computing the TBR and SPR distances for phylogenetic trees.

Several interesting and challenging questions were raised as follows:

(1) Approximation algorithms and improved fixed-parameter algorithms for uSPR distance

No approximation algorithms have been investigated with regard to the uSPR distance problem in the literature thus far. Although Bonet and John[35] proposed a parameterized algorithm with a running time of $O(k^{4k} + p(n))$ for the parameterized uSPR disatance problem, there is still considerable room for further improvement.

(2) Extend existing algorithms for binary phylogenetic trees to multifurcating phylogenetic trees

For several biological data sets[41], the constructed phylogenetic trees are multifurcating. Yang[27] and Chen et al.[32] studied the TBR distance problem on two unrooted multifurcating trees. Therefore, it is worthwhile to study the problems related to multifurcating phylogenetic trees, such as uSPR and rSPR distance problems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61103033, 61173051, 61232001, and 70921001).

References

[1] A. W. F. Edwards and L. L. Cavalli-Sforza, The reconstruction of evolution, *Annals of Human Genetics*, vol. 27, pp. 105-106, 1963.

[2] W. Fitch, Toward defining the course of evolution: Minimum change for a specified tree topology, *Systematic Biology*, vol. 20, no. 4, pp. 406-416, 1971.

[3] D. Sankoff, Minimal mutation trees of sequences, *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 35-42, 1975.

[4] W. J. Le Quesne, The uniquely evolved character concept and its cladistic application, *Systematic Biology*, vol. 23, no. 4, pp. 513-517, 1974.

[5] W. M. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science*, vol. 155, no. 3760, pp. 279-284, 1967.

[6] N. Saitou and M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406-425, 1987.

[7] J. Felsenstein, Evolutionary trees for DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368-376, 1981.

[8] D. Barry and J. A. Hartigan, Statistical analysis of hominoid molecular evolution, *Statistical Science*, vol. 2, no. 2, pp. 191-210, 1987.

[9] D. Robinson and L. Foulds, Comparison of phylogenetic trees, *Mathematical Biosciences*, vol. 53, nos. 1-2, pp. 131-147, 1981.

[10] D. Robinson, Comparison of labeled trees with valency three, *Journal of Combinatorial Theory, Series B*, vol. 11, no. 2, pp. 105-119, 1971.

[11] K. Culik and D. Wood, A note on some tree similarity measures, *Information Processing Letters*, vol. 15, no. 1, pp. 39-42, 1982.

[12] W. Day, Properties of the nearest neighbor interchange metric for trees of small size, *Journal of Theoretical Biology*, vol. 101, no. 2, pp. 275-288, 1983.

[13] R. Boland, E. Brown, and E. Day, Approximating minimum-length-sequence metrics: A cautionary note, *Mathematical Social Sciences*, vol. 4, no. 3, pp. 261-270, 1983.

[14] J. Jarvis, J. Luedeman, and D. Shier, Counterexamples in measuring the distance between binary trees, *Mathematical Social Sciences*, vol. 4, no. 3, pp. 271-274, 1983.

[15] J. Jarvis, J. Luedeman, and D. Shier, Comments on computing the similarity of binary trees, *Journal of Theoretical Biology*, vol. 100, no. 3, pp. 427-433, 1983.

[16] M. Krivanke, Computing the nearest neighbor interchange metric for unlabeled binary trees is NP-complete, *Journal of Classification*, vol. 3, no. 1, pp. 55-60, 1986.

[17] V. King and T. Warnow, On measuring the nni distance between two evolutionary trees, presented at the DIMACS Mini Workshop on Combinatorial Structures in Molecular Biology, South Plainfield, USA, 1994.

[18] M. Li, J. Tromp, and L. Zhang, Some notes on the nearest neighbor interchange distance, in *Proc. 2nd Annual International Computing and Combinatorics Conference (COCOON)*, Hong Kong, China, 1996, pp. 343-351.

[19] M. Li, J. Tromp, and L. Zhang, On the nearest neighbor interchange distance between evolutionary trees, *Journal on Theoretical Biology*, vol. 182, no. 4, pp. 463-467, 1996.

[20] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, *Mathematical Biosciences*, vol. 98, no. 2, pp. 185-200, 1990.

[21] J. Hein, A heuristic method to reconstruct the history of sequences subject to recombination, *Journal of Molecular Evolution*, vol. 36, no. 4, pp. 396-405, 1993.

[22] P. Buneman, The recovery of trees from measures of dissimilarity, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall, and P. T. Tautu, Ed. Edinburgh, UK: Edinburgh University Press, 1971, pp. 387-395.

[23] D. Swofford, G. Olsen, P. Waddell, and D. Hillis, Phylogenetic inference, in *Molecular Systematics*, 1996, pp. 407-513.

[24] B. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, *Annals of Combinatorics*, vol. 5, no. 1, pp. 1-15, 2001.

[25] F. Shi, J. Chen, Q. Feng, and J. Wang, Parameterized algorithms for maximum agreement forest on multiple trees, in *Proc. 19th Annual International Computing and Combinatorics Conference (COCOON)*, Hangzhou, China, 2013, pp. 567-578.

[26] J. Hein, T. Jiang, L. Wang, and K. Zhang, On the complexity of comparing evolutionary trees, *Discrete Applied Mathematics*, vol. 71, no. 1-3, pp. 153-169, 1996.

[27] Y. Yang, A fixed-parameterized algorithm for the maximum agreement forest problem on multifurcating trees, Master dissertation, Central South University, Changsha, China, 2012. .

[28] R. Downey and M. Fellows, *Parameterized Complexity*. New York, USA: Springer, 1999.

[29] J. Chen, Parameterized computation and complexity: A new approach dealing with NP-hardness, *Journal of Computer Science and Technology*, vol. 20, no. 1, pp. 18-37, 2005.

[30] M. Hallett and C. McCartin, A faster FPT algorithm for the maximum agreement forest problem, *Theory of Computing Systems*, vol. 41, no. 3, pp. 539-550, 2007.

[31] C. Whidden and N. Zeh, A unifying view on approximation and FPT of agreement forests, in *Proc. Algorithms in Bioinformatics*, Philadelphia, PA, USA, 2009, pp. 390-402.

[32] J. Chen, J. Fan, and S. Sze, Parameterized and approximation algorithms for the MAF problem in multifurcating trees, presented at the 39th International Workshop on Graph-Theoretic Concepts in Computer Science, Lübeck, Germany, 2013.

[33] M. Bordewich and C. Semple, On the computational complexity of the rooted subtree prune and regraft distance, *Annals of Combinatorics*, vol. 8, no. 4, pp. 409-423, 2005.

[34] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin, SPR distance computation for unrooted trees, *Evolutionary Bioinformatics*, vol. 4, pp. 17-27, 2008.

[35] M. Bonet and K. St. John, On the complexity of uSPR distance, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 572-576, 2010.

[36] M. Bordewich, C. McCartin, and C. Semple, A 3-approximation algorithm for the subtree distance between phylogenies, *Journal of Discrete Algorithms*, vol. 6, no. 3, pp. 458-471, 2008.

[37] C. Whidden, R. Beiko, and N. Zeh, Fixed-parameter and approximation algorithms for maximum agreement forests, CoRR. abs/1108.2664, 2011.

[38] Z. Chen and L. Wang, HybridNET: A tool for constructing hybridization networks, *Bioinformatics*, vol. 26, no. 22, pp. 2912-2913, 2010.

[39] E. Rodrigues, M. Sagot, and Y. Wakabayashi, The maximum agreement forest problem: Approximation algorithms and computational experiments, *Theoretical Computer Science*, vol. 374, nos. 1-3, pp. 91-110, 2007.

[40] M. Bonet, K. St. John, R. Mahindru, and N. Amenta, Approximating subtree distances between phylogenies, *Journal of Computational Biology*, vol. 13, no. 8, pp. 1419-1434, 2006.

[41] O. Paun, C. Lehnebach, J. Johansson, P. Lockhart, and E. Hrandl, Phylogenetic relationships and biogeography of Ranunculus and allied genera (Ranunculaceae) in the Mediter-ranean region and in the European Alpine System, *Taxon*, vol. 54, no. 4, pp. 911-932, 2005.

Feng Shi is a PhD candidate of Department of Computer Science and Engineering in Central South University. He received his BS degree in computer science from Central South University in 2011. His main research interests include computer algorithms and parameterized algorithms.

Qilong Feng received his PhD degree in computer science from Central South University, China, in 2010. His current research interests include computer algorithms and parameterized algorithms.

Jianer Chen received his PhD degree in computer science from Courant Institute of New York University in 1987 and PhD degree in mathematics from Columbia University in 1990. He is currently a professor of computer science at Texas A&M University, and Central South University in Changsha, China. His main research is centered on computer algorithms and their applications. His current research projects include exact and parameterized algorithms, computer graphics, computer networks, and computational biology.

Lusheng Wang received his PhD degree from McMaster University, Hamilton, Ontario, Canada, in 1995. Currently, he is a professor in the Department of Computer Science, City University of Hong Kong. His research interests include algorithms, bioinformatics, and computational biology. He is a member of the IEEE.

Jianxin Wang received his PhD degree in computer science from Central South University, China, in 2001. Currently, he is a professor at School of Information Science and Engineering, Central South University, China. His current research interests include algorithm analysis and optimization, computer network, and bioinformatics. He has published more than 100 papers in various international journals and refereed conferences. He is serving as the program committee chair or member of several international conferences. He is a senior member of IEEE.