Optimal Algorithm to Reconstruct a Tree from a Subtree Distance

Takanori Maehara
RIKEN Center for Advanced Intelligence Project
takanori.maehara@riken.jp

Kazutoshi Ando
Department of Mathematical and Systems Engineering, Shizuoka University
ando.kazutoshi@shizuoka.ac.jp

Abstract
This paper addresses the problem of finding a representation of a subtree distance, which is an extension of the tree metric. We show that a minimal representation is uniquely determined by a given subtree distance, and give a linear time algorithm that finds such a representation. This algorithm achieves the optimal time complexity.

Keywords: graph algorithm, phylogenetic tree, tree metric, subtree distance

1. Introduction and the results

A phylogenetic tree represents an evolutionary relationship among the species that are investigated. Estimating a phylogenetic tree from experimental data is a fundamental problem in phylogenetics. One of the commonly used approaches to achieve this task is the use of a distance-based method. In this approach, we first compute the dissimilarity (i.e., a nonnegative and symmetric function) between the species by, e.g., the edit distance between the genome sequences. Then, we find a weighted tree having the shortest path distance that best fits the given dissimilarity. The most popular method for this approach is the neighbor-joining method.

A weighted tree $T$ is specified by the set of vertices $V(T)$, the set of edges $E(T)$, and the nonnegative edge weight $l : E(T) \rightarrow \mathbb{R}_+$. Let us consider the case in which a given dissimilarity $d : X \times X \rightarrow \mathbb{R}$ exactly fits some weighted tree; i.e., there exists
a weighted tree $T$ and a mapping $\psi : X \to V(T)$ such that

$$d(x, y) = d_T(\psi(x), \psi(y)) \quad (x, y \in X),$$

(1.1)

where $d_T(u, v)$ is the distance between $u$ and $v$ in $T$ for $u, v \in V(T)$. In this case, the dissimilarity $d : X \times X \to \mathbb{R}$ is called a tree metric, and the pair $(T, \psi)$ is called a representation of $d$. It is known that a dissimilarity $d$ is a tree metric if and only if it satisfies an inequality called the four-point condition \[9, 2\], which is given by

$$d(x, y) + d(z, w) \leq \max\{d(x, z) + d(y, w), d(x, w) + d(y, z)\}$$

(1.2)

for any $x, y, z, w \in X$. For any tree metric $d$, there exists a unique minimal representation $(T, \psi)$ of $d$ \[2\]. Here, a representation is minimal if there is no representation $(T', \psi')$ of $d$, such that $T'$ is obtained by removing some vertices and edges and/or by contracting some edges of $T$ (i.e., $T'$ is a proper minor of $T$). Furthermore, such a representation is constructed in $O(n^2)$ time \[3\], where $n = |X|$. In some applications, we are interested in the distance between groups of species (e.g., genus, tribe, or family). In such a case, we aim to identify a group as a connected subgraph in a phylogenetic tree. The subtree distance, which was introduced by Hirai \[5\], is an extension of the tree metric that can be adopted for use in such situations. A function $d : X \times X \to \mathbb{R}$ is called a subtree distance if there exists a weighted tree $T$ and a mapping $\phi : X \to 2^{V(T)}$ such that $\phi(x)$ induces a subtree of $T$ (i.e., a connected subgraph of $T$) for $x \in X$ and equations

$$d(x, y) = d_T(\phi(x), \phi(y)) \quad (x, y \in X),$$

(1.3)

hold, where $d_T(U, W) = \min\{d_T(u, w) \mid u \in U, w \in W\}$ for $U, W \subseteq V(T)$. We say that a pair $(T, \phi)$ is a representation of $d$. Note that a subtree distance is not necessarily a metric because it may not satisfy the non-degeneracy ($d(x, y) > 0$ for $x \neq y$) and the triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$). Hirai proposed a characterization of subtree distances in which a dissimilarity $d : X \times X \to \mathbb{R}$ is a subtree distance if and only if it satisfies an inequality called the extended four-point condition, which is given by

$$d(x, y) + d(z, w) \leq \max \left\{ \frac{d(x, z) + d(y, w) + d(x, w) + d(y, z)}{2}, \frac{d(x, y) + d(z, w) + d(x, w) + d(y, z)}{2}, \frac{d(x, y) + d(z, w) + d(x, w) + d(y, z)}{2}, \frac{d(x, y) + d(z, w) + d(x, w) + d(y, z)}{2} \right\}$$

(1.4)

for any $x, y, z, w \in X$. The extended four-point condition yields an $O(n^4)$ time algorithm to recognize a subtree distance.
This paper addresses the following problem, which we call the *subtree distance reconstruction problem*.

**Problem 1.** Given a subtree distance \( d : X \times X \to \mathbb{R} \) on a finite set \( X \), find a representation \((T, \phi)\) of \( d \).

Ando and Sato [1] proposed an \( O(n^3) \) time algorithm for this problem. Their algorithm consists of three steps: (1) identify a subset \( V_0 = \{ x \in X \mid d(y, z) \leq d(x, y) + d(x, z) \} \), (2) find a representation \((T, \phi)\) for the restriction of \( d \) onto \( V_0 \), and (3) for \( x \in X \setminus V_0 \), locate \( \phi(x) \) in \( T \) by examining a connected components of \( T \setminus \phi(x) \).

In this study, we propose the following theorems. We define a minimal representation for a subtree distance in the same manner as a minimal representation for a tree metric.

**Theorem 2.** For a subtree distance \( d : X \times X \to \mathbb{R} \), a minimal representation \((T, \phi)\) is uniquely determined by \( d \).

**Theorem 3.** There exists an \( O(n^2) \) time algorithm that finds, for any subtree distance \( d : X \times X \to \mathbb{R} \), its unique minimal representation, where \( n = |X| \).

The proof of Theorem 3 is constructive. Similar to [1], our algorithm consists of three parts: (1) identify the set of objects \( L \) that are *mapped to the leaves*, (2) find the minimal representation \((T, \phi)\) for the restriction of \( d \) onto \( L \), and (3) for \( x \in X \setminus L \), locate \( \phi(x) \) in \( T \) by measuring the distances from the leaves. Since Steps 1 and 3 can be implemented with a time complexity of \( O(n^2) \), and there is an \( O(n^2) \) time algorithm for Step 2 [3], the total time complexity of the algorithm is \( O(n^2) \). Note that even if we know \( d : X \times X \to \mathbb{R} \) is a tree metric, \( \Omega(n^2) \) time is required to reconstruct a tree [4]. Therefore, our algorithm achieves the optimal time complexity.

This algorithm can also be used to recognize a subtree distance by checking the failure or inconsistency during the process and by verifying equations (1.3) after the reconstruction.

**Corollary 4.** There exists an \( O(n^2) \) time algorithm that determines whether a given input \( d : X \times X \to \mathbb{R} \) is a subtree distance or not, where \( n = |X| \).

2. **Proofs**

We assume that there are no objects \( x, y \in X \) such that \( d(x, z) = d(y, z) \) for all \( z \in X \). This assumption is satisfied by removing such elements after lexicographic sorting, which requires \( O(n^2) \) time [7]. Clearly, this preprocessing does not change
the minimal representation. We also assume that |X| ≥ 3. Otherwise, the theorems trivially hold.

First, we prove Theorem 2. We identify the properties of a minimal representation.

**Lemma 5.** Let \((T, \phi)\) be a minimal representation of a subtree distance \(d : X \times X \to \mathbb{R}\). Then, the following properties hold.

1. For each edge \(e \in E(T)\), the length of \(e\) is positive.
2. For each leaf vertex \(u \in V(T)\), there exists \(x \in X\) such that \(\phi(x) = \{u\}\).

**Proof.**

1. If there is an edge \(e\) of zero length, we can contract \(e\) from the representation.
2. Let \(u\) be a leaf vertex of \(T\). If there is no \(x \in X\) with \(u \in \phi(x)\), we can remove \(u\) from the representation to obtain a smaller representation. Suppose that, for all \(x \in X\) with \(u \in \phi(x)\), \(\phi(x)\) contains at least two elements. Then, these \(\phi(x)\)s contain the unique adjacent vertex \(v\) of \(u\). Since \(d(v, w) \leq d(u, w)\) for all \(w \in V(T) \setminus \{u\}\), \(u\) does not contribute any shortest paths in the tree. Therefore, we can remove \(u\) from the representation.

Motivated by Property 2 in Lemma 5, we introduce the following definition. For a minimal representation \((T, \phi)\), an object \(x \in X\) is a **leaf object** if \(\phi(x) = \{u\}\) for some leaf \(u \in V(T)\).

We defined a leaf object by specifying a minimal representation. However, as shown below, the set of leaf objects is uniquely determined by \(d\). First, we show that there exists an object that is a leaf object for any minimal representation.

**Lemma 6.** Let \((r, r') \in \text{argmax}_{x,y \in X} d(x, y)\). Then, for any minimal representation, \(r\) and \(r'\) are leaf objects.

**Proof.** For any tree, the farthest pair is attained by a pair of leaves.

Next, we show that the leaf objects are characterized by \(d\) and a leaf object \(r\).

**Lemma 7.** Let \((T, \phi)\) be a minimal representation of subtree distance \(d : X \times X \to \mathbb{R}\), and let \(r \in X\) be a leaf object. An object \(x \in X \setminus \{r\}\) is a leaf object if and only if \(d(y, r) < d(x, y) + d(x, r)\) for all \(y \in X \setminus \{x, r\}\).

**Proof.** (The “if” part). Suppose \(x\) is not a leaf object. Then, there is another leaf object \(y \in X \setminus \{x, r\}\) such that the path from \(\phi(r)\) to \(\phi(y)\) intersects \(\phi(x)\). By considering distances among \(\phi(r), \phi(y)\) and \(\phi(x)\) on this path, we have \(d(r, y) \geq d(x, r) + d(x, y)\).
(The “only if” part). Suppose $x$ is a leaf object. Then, for any $y \in X \setminus \{x, r\}$ the shortest path from $\phi(y)$ to $\phi(r)$ never intersects $\phi(x)$. Thus, we have $d(y, r) < d(x, y) + d(x, r)$. Here, we used the assumption that there is no $x, y \in X$ such that $d(x, z) = d(y, z)$ for all $z \in X$.

Since we can take a leaf object $r$ universally by Lemma 6, and the condition in Lemma 7 is described without specifying the underlying representation, we can conclude that the set of leaf objects is uniquely determined by $d$. Thus, we obtain the following corollary.

**Corollary 8.** Any minimal representation of a subtree distance has the same leaf objects.

Now, we observe that the dissimilarity $d|_L : L \times L \to \mathbb{R}$ obtained by restricting $d : X \times X \to \mathbb{R}$ on the set of leaf objects $L \subseteq X$ forms a tree metric because the leaf objects are mapped to singletons. Since the minimal representation of a tree metric is unique [2], any minimal representation of $d$ has the same topology as the minimal representation of $d|_L$.

The remaining issue is to show that each non-leaf object $x \in X \setminus L$ is uniquely located in the minimal representation $(T, \phi)$ of $d|_L$. This is clear because any connected subgraph in a tree is uniquely identified by the distances from the leaves. More precisely, we obtain the following explicit representation.

We first consider $T$ as a continuous object. We fix a leaf object $r \in L$. For each leaf object $x \in L \setminus \{r\}$, there exists a unique path $\phi(r), \phi(x)$ from $\phi(r)$ to $\phi(x)$ in $T$. Let $I(r, a; x, b)$ be the interval on the path having a distance of at least $a$ from $\phi(r)$ and at least $b$ from $\phi(x)$, i.e.,

$$I(r, a; x, b) = \{u \in \text{path}(\phi(r), \phi(x)) : d_T(\phi(r), u) \geq a, d_T(\phi(x), u) \geq b\}.$$

For $U \subseteq V(T)$, we denote by $\overline{U}$ the subgraph of $T$ induced by $U$. Note that both $I(r, a; x, b)$ and $\overline{U}$ are continuous objects. By using these notations, we obtain the following.

**Lemma 9.** Let $d : X \times X \to \mathbb{R}$ be a subtree distance and $L$ be the set of leaf objects. Let $(T, \phi)$ be the minimal representation of $d|_L$. Fix a leaf object $r \in L$. Then, we have for each non-leaf object $z \in X \setminus L$

$$\overline{\phi(z)} = \bigcup_{x \in L \setminus \{r\}} I(r, d(r, z); x, d(x, z)).$$

5
Proof. Since $\overline{\phi(z)}$ is a connected subgraph, the intersection of $\overline{\phi(z)}$ and the path from $\phi(r)$ to $\phi(x)$ is the interval $I(r, d(r, z); x, d(x, z))$. Since any tree is covered by the paths from a fixed leaf $\phi(r)$ and the other leaves $\phi(x)$ for $x \in L \setminus \{r\}$, we have

$$\overline{\phi(z)} = \left( \bigcup_{x \in L \setminus \{r\}} \text{path}(\phi(r), \phi(x)) \right) \cap \overline{\phi(z)}$$

$$= \bigcup_{x \in L \setminus \{r\}} \left( \text{path}(\phi(r), \phi(x)) \cap \overline{\phi(z)} \right)$$

$$= \bigcup_{x \in L \setminus \{r\}} I(r, d(r, z); x, d(x, z)). \quad (2.3)$$

By placing the vertices on the boundaries of $\overline{\phi(z)}$ for $z \in X \setminus L$ and then letting $\phi(z)$ be the vertices intersecting $\overline{\phi(z)}$ for $z \in X \setminus L$, we obtain the minimal representation of $d$. This completes the proof of Theorem 2. Note that the number of the vertices of the minimal representation is $O(n^2)$ since each $\overline{\phi(z)}$ has at most $|L| = O(n)$ boundaries.

Now, we prove Theorem 3. The above proof of Theorem 2 is constructive, and it provides the following algorithm:

1. Identify the set $L$ of leaf objects by Lemmas 6 and 7.
2. Find the minimal representation of $d|_L$ by the existing algorithm.
3. Locate the non-leaf objects by Lemma 9.

We evaluate the time complexity of this algorithm. Step 1 is conducted in $O(n^2)$ time for finding a leaf object $r \in X$ and $O(n^2)$ time for finding other leaf objects. Step 2 is performed in $O(n^2)$ time by using Culberson and Rudnicki’s algorithm [3]. Also, Step 3 is performed in $O(n^2)$ time by equation (2.2) since, for each $z$, it processes each edge at most once. Hence, Theorem 3 is proved.

Acknowledgments

The authors thank Hiroshi Hirai and anonymous referees for helpful comments. This work was supported by the Japan Society for the Promotion of Science, KAKENHI Grant Number 15K00033.
References

[1] Kazutoshi Ando and Koki Sato. An algorithm for finding a representation of a subtree distance. *Journal of Combinatorial Optimization*, 36(3):742–762, 2018.

[2] Peter Buneman. The recovery of trees from measures of dissimilarity. In D.G. Kendall and P. Tautu, editors, *Mathematics in the Archeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.

[3] Joseph C. Culberson and Piotr Rudnicki. A fast algorithm for constructing trees from distance matrices. *Information Processing Letters*, 30(4):215–220, 1989.

[4] Jotun J. Hein. An optimal algorithm to reconstruct trees from additive distance data. *Bulletin of Mathematical Biology*, 51(5):597–603, 1989.

[5] Hiroshi Hirai. Characterization of the distance between subtrees of a tree by the associated tight span. *Annals of Combinatorics*, 10(1):111–128, 2006.

[6] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[7] Juraj Wiedermann. The complexity of lexicographic sorting and searching. In *Proceedings of the 8th International Symposium on Mathematical Foundations of Computer Science*, pages 517–522, 1979.

[8] Edward O. Wiley and Bruce S. Lieberman. *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons, 2011.

[9] K. A. Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6):90–92, 1965.