JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 31, Number 1, 2024

Pp. 1–13
DOI: 10.1089/cmb.2023.0262

Research Article

Consensus Tree Under the Ancestor–Descendant Distance is NP-Hard

YUANYUAN QI[1] and MOHAMMED EL-KEBIR[1,2]

ABSTRACT

Due to uncertainty in tumor phylogeny inference from sequencing data, many methods infer multiple, equally plausible phylogenies for the same cancer. To summarize the solution space $\mathcal{T}$ of tumor phylogenies, consensus tree methods seek a single best representative tree S under a specified pairwise tree distance function. One such distance function is the ancestor–descendant (AD) distance $d(T, T')$, which equals the size of the symmetric difference of the transitive closures of the edge sets $E(T)$ and $E(T')$. Here, we show that finding a consensus tree S for tumor phylogenies $\mathcal{T}$ that minimizes the total AD distance $\sum_{T\in\mathcal{T}} d(S, T)$ is NP-hard.

Keywords: cancer, consensus tree, infinite sites assumption, intra-tumor heterogeneity.

1. INTRODUCTION

Cancer results from an evolutionary process during which somatic mutations accumulate in a population of cells (Nowell, 1976). To study tumor evolution, researchers apply phylogeny inference algorithms to sequencing data of tumors (Schwartz and Schäffer, 2017). Due to uncertainty in tumor phylogeny inference from sequencing data, these methods typically yield multiple candidate trees $\mathcal{T}$ for the same tumor (El-Kebir et al., 2016; Qi et al., 2019).

To summarize this solution space, several works have been proposed to infer a consensus tree S that best represents the set $\mathcal{T}$ of candidate trees. More formally, these consensus tree methods typically employ a distance function $d(T_1, T_2)$ that compares two trees T_1 and T_2, and seek a consensus tree S that minimizes the total distance $\sum_{T\in\mathcal{T}} d(S, T)$ (Aguse et al., 2019; DiNardo et al., 2020; Fu and Schwartz, 2021; Govek et al., 2018, 2020; Guang et al., 2023; Karpov et al., 2019).

As many tumor phylogeny inference methods make the infinite sites assumption, which states that each mutation is gained exactly once on the tree and never subsequently lost (Kimura, 1969), tumor phylogenies that adhere to this assumption are typically represented as mutation trees. These are rooted trees where each

[1]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.
[2]Cancer Center at Illinois, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.

node represents a clone composed of the mutations labeling the nodes of the unique path to the root (Fig. 1a). For example, GraPhyC (Govek et al., 2018, 2020) finds an optimal consensus mutation tree based on the parent–child (PC) distance in polynomial time.

Briefly, the PC distance of trees T_1 and T_2 is defined as the size of the symmetric difference between the edge sets $E(T_1)$ and $E(T_2)$. Aguse et al. (2019) generalized the problem to identify multiple consensus trees under the PC distance, and Christensen et al. (2020) considered a multiple-choice version of the problem to identify repeated evolutionary trajectories in cancer phylogeny cohort data.

Although computationally tractable, DiNardo et al. (2020) suggest the PC distance may not provide enough resolution for tumor phylogeny comparison. Recently, Guang et al. (2023) developed TuELiP, which uses integer linear programming to identify the optimal consensus tree with the ancestor–descendant (AD) distance, originally proposed by Govek et al. (2018). The AD distance equals the size of the symmetric difference of the transitive closures of the directed edge sets $E(T_1)$ and $E(T_2)$.

In other words, the AD distance equals the number of ordered pairs of vertices (u, v) where u is an ancestor of v that are unique to either T_1 or T_2 (Fig. 1b). Compared with the PC distance, the AD distance provides greater resolution to detect subtle differences between the two trees. Importantly, the hardness of the consensus tree problem under the AD distance is unknown (Fig. 1c), leaving the existence of a polynomial-time algorithm as an open problem.

In this study, we show that finding the optimal consensus tree under the AD distance is NP-hard. Therefore, unlike the PC distance consensus tree problem, for which a polynomial-time algorithm exists (Govek et al., 2020; Govek et al., 2018), there is no polynomial-time algorithm for the consensus tree problem under the AD distance unless $P = NP$.

2. PROBLEM STATEMENT

We consider mutation trees T, which are rooted, vertex-labeled trees with vertex set $V(T)$ and edge set $E(T)$. Intuitively, each vertex i of a mutation tree T corresponds to a tumor clone comprising the mutations that label the vertices of the unique path from i to the root of T (Fig. 1a). We write $i \prec_T j$ if (i) vertex i is an ancestor of vertex j and (ii) $i \neq j$. We write $i \perp_T j$ if vertices i and j occur on distinct root-to-leaf paths of T, that is, $i \not\prec_T j$ and $j \not\prec_T i$. While $\perp_T$ is symmetrical, that is, $i \perp_T j$ if and only if $j \perp_T i$, the relation $\prec_T$ is not symmetrical.

Neither $\prec_T$ nor $\perp_T$ are reflexive, that is, it does not hold that $i \prec_T i$, nor does it hold that $i \perp_T i$ for any vertex i. To compute the distance between two rooted trees T_1 and T_2 on the same vertex set, we compare the AD sets $A(T_1)$ and $A(T_2)$ of T_1 and T_2, respectively. More formally, $A(T_1)$ equals the transitive closure of $E(T_1)$.

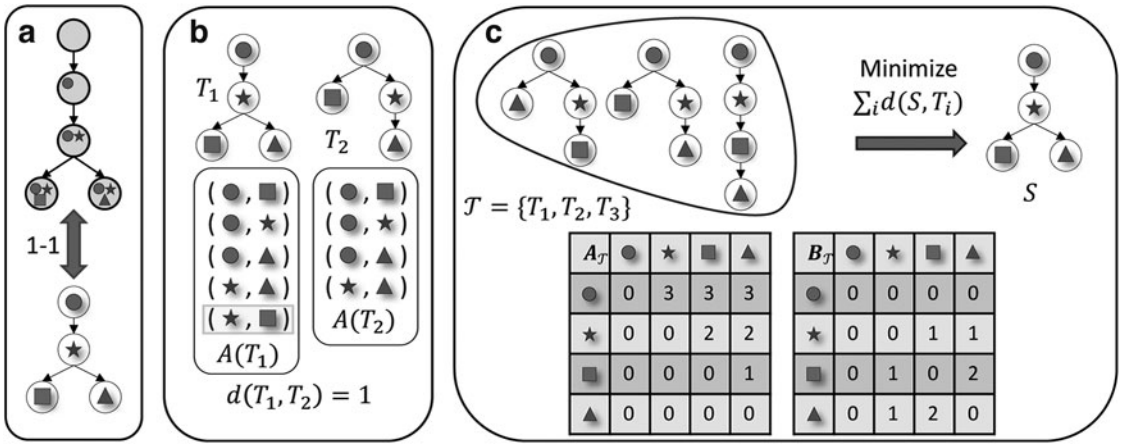

FIG. 1. Overview of the ADCT problem. **(a)** There is a bijection between phylogenies under the infinite sites assumption and mutation trees. **(b)** The AD distance $d(T_1, T_2)$ of mutation trees T_1 and T_2 equals the size of the symmetric difference of the AD sets $A(T_1)$ and $A(T_2)$. Here, $d(T_1, T_2) = 1$ due to the unmatched pair of $A(T_1)$ indicated in a box. **(c)** In the ADCT problem, we seek a consensus mutation tree S that minimizes the sum of the distances to the STEP trees $\mathcal{T}$. AD, ancestor–descendant; ADCT, Ancestor–Descendant Consensus Tree.

Definition 1. *The* AD set $A(T)$ *of a rooted tree T consists of all ordered pairs* (i, j) *of vertices such that i is ancestor of j,* that is, $A(T) = \{(i, j) \in V(T) \times V(T) | i \prec_T j\}$.

The AD distance $d(T_1, T_2)$ equals the size of the symmetric difference of $A(T_1)$ and $A(T_2)$, more formally defined as follows. See Figure 1b for an example.

Definition 2. *Given two rooted trees* T_1, T_2 *on the same vertex set, the* AD distance $d(T_1, T_2)$ *equals the size of the symmetric difference of* $A(T_1)$ *and* $A(T_2)$, that is, $d(T_1, T_2) = |A(T_1) \backslash A(T_2)| + |A(T_2) \backslash A(T_1)|$.

This leads to the following problem.

Problem 2.3 (Ancestor–Descendant Consensus Tree [ADCT]). *Given a multi-set* $\mathcal{T} = \{T_1, \ldots, T_m\}$ *of rooted trees on the same vertex set* $V(\mathcal{T})$*, find a rooted tree S on vertex set* $V(\mathcal{T})$ *such that the sum* $\sum_{i=1}^{m} d(S, T_i)$ *of the distances from S to each input tree* $T \in \mathcal{T}$ *is minimum.*

3. COMBINATORIAL CHARACTERIZATION

For any unordered pair $\{i, j\}$ of distinct vertices in a mutation tree T, it must hold that $i \prec_T j$, $j \prec_T i$ or $i \perp_T j$. We indicate the first two cases using $\mathbf{1}\{i \prec_T j\}$ such that $\mathbf{1}\{i \prec_T j\} = 1$ if $i \prec_T j$ and 0 otherwise, and the third case using $\mathbf{1}\{i \perp_T j\}$ such that $\mathbf{1}\{i \perp_T j\} = 1$ if $i \perp_T j$ and 0 otherwise. As such, the distance $d(T_1, T_2)$ can be decomposed as follows.

Lemma 1. *The AD distance* $d(T_1, T_2)$ *for trees* T_1 *and* T_2 *on the same vertex set* $[n]$ *equals*

$$d(T_1, T_2) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(T_1, T_2) \tag{1}$$

where $d^{i,j}(T_1, T_2)$ *is the distance contributed by the unordered pair* $\{i, j\}$ *of distinct vertices defined as*

$$\begin{aligned} d^{i,j}(T_1, T_2) \quad &= \mathbf{1}\{i \prec_{T_1} j\}(2 \cdot \mathbf{1}\{j \prec_{T_2} i\} + \mathbf{1}\{i \perp_{T_2} j\}) \\ &\quad + \mathbf{1}\{j \prec_{T_1} i\}(2 \cdot \mathbf{1}\{i \prec_{T_2} j\} + \mathbf{1}\{i \perp_{T_2} j\}) \\ &\quad + \mathbf{1}\{i \perp_{T_1} j\}(\mathbf{1}\{i \prec_{T_2} j\} + \mathbf{1}\{j \prec_{T_2} i\}). \end{aligned} \tag{2}$$

Proof. See Section 6. □

We can similarly decompose the total AD distance $d(S, \mathcal{T})$ between a tree S and trees $\mathcal{T}$ by first defining the AD matrix and the branching matrix as follows.

Definition 3. *An* $n \times n$ *matrix* $A_{\mathcal{T}} = [a_{i,j}]$ *is an* AD matrix for trees $\mathcal{T}$ *provided each entry* $a_{i,j}$ *equals* $\sum_{T \in \mathcal{T}} \mathbf{1}\{i \prec_T j\}$.

Definition 4. *An* $n \times n$ *matrix* $B_{\mathcal{T}} = [b_{i,j}]$ *is a* branching matrix for trees $\mathcal{T}$ *provided each entry* $b_{i,j}$ *equals* $\sum_{T \in \mathcal{T}} \mathbf{1}\{i \perp_T j\}$.

While $A_{\mathcal{T}}$ may not be symmetric, matrix $B_{\mathcal{T}}$ is symmetric due to symmetry of the relation $\perp_T$. Moreover, the diagonal of both matrices consist of 0s. Finally, note that $a_{i,j} + a_{j,i} + b_{i,j} = |\mathcal{T}|$ if $i \neq j$. See Figure 1c for an example.

Lemma 2. *The AD distance* $d(S, \mathcal{T})$ *between a tree S and trees* $\mathcal{T}$ *on the same vertex set* $[n]$ *equals*

$$d(S, \mathcal{T}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(S, \mathcal{T}) \tag{3}$$

where $d^{i,j}(S, \mathcal{T})$ *is the distance contributed by the unordered pair* $\{i, j\}$ *of distinct vertices defined as*

$$\begin{aligned} d^{i,j}(S, \mathcal{T}) = \mathbf{1}\{i \prec_S j\}\left(2a_{j,i} + b_{i,j}\right) + \mathbf{1}\{j \prec_S i\}\left(2a_{i,j} + b_{i,j}\right) \\ + \mathbf{1}\{i \perp_S j\}\left(a_{i,j} + a_{j,i}\right). \end{aligned} \tag{4}$$

Proof. See Section 6. □

4. COMPLEXITY

Our main result is as follows.

Theorem 1. *The* ADCT *problem is NP-hard.*

We show NP-hardness by giving a polynomial-time reduction from the MAX-CLIQUE problem, defined as follows.

Problem 4.1 (MaxClique). *Given an undirected graph G with $n = |V(G)|$ vertices and $m = |E(G)|$ edges, find a clique $C \subseteq V(G)$ such that $|C|$ is maximum.*

The MAXCLIQUE problem is NP-hard (Cook, 1971; Karp, 1972). In the following reduction, we assume the undirected graph G contains at least three vertices, that is, $n > 2$. This assumption does not affect the hardness of the MAXCLIQUE problem. We impose an arbitrary ordering on the vertices $V(G)$ such that $V(G) = [n] = \{1, \ldots, n\}$. For each vertex $i \in V(G)$, let $\delta(i)$ be the set of vertices adjacent to i in G, that is, $\delta(i) = \{j \in [n] | (i, j) \in E(T)\}$.

Using the ordering, we further split the set $\delta(i)$ of neighboring vertices $\delta(i)$ into vertices $\delta^{>}(i) = \{j \in [n] | j \in \delta(i), i < j\}$ that are adjacent to i and occur after i in the ordering. The vertex set $V(\mathcal{T})$ of the corresponding ADCT problem instance includes $2n+1$ vertices, labeled $\{0, 1, \ldots, n, n+1, \ldots, 2n\}$. Vertex 0 denotes a special vertex that is the shared root of all trees $\mathcal{T}$, and a set $\{n+1, \ldots, 2n\}$ of n vertices that forms a chain in all trees. We construct the following multi-set $\mathcal{T}$ of rooted trees on the new vertex set $V(\mathcal{T}) = \{0, \ldots, 2n\}$, with three types of trees (Fig. 2b).

First, let T_0 be a chain tree whose vertices are in ascending order, that is, $E(T_0) = \{(i, i+1) | 0 \leq i < 2n\}$. The multi-set $\mathcal{T}_0$ comprises $n^3 - 2n^2 + 2n - 3$ copies of T_0. Second, for each vertex i in the undirected graph $V(G)$, let T_i be a tree rooted at 0. The edge set $E(T_i)$ consists of (i) edges from 0 to every vertex in G that is either at most i or is not adjacent to i, that is, $\{(0, j) | j \in V(G) \backslash \delta^{>}(i)\}$; (ii) edges from 0 to every vertex in G that is greater than i and adjacent to i, that is, $\{(i, j) | j \in \delta^{>}(i)\}$; (iv) the edge $\{(i, n+1)\}$; and (v) a chain from $n+1$ to $2n$ in ascending order, that is, $\{(j, j+1) | n < i < 2n\}$.

The multi-set $\mathcal{T}_i$ comprises $n^2 + 1$ copies of T_i. Third, for each vertex $i \in V(G)$, let T'_i be a tree rooted at 0. The edge set $E(T'_i)$ consists of (i) edges from 0 to every vertex in G, that is, $\{(0, j) | j \in V(G)\}$; (ii) the edge $\{(i, n+1)\}$; and (iii) a chain from $n+1$ to $2n$ in ascending order, that is, $\{(j, j+1) | n+1 \leq j < 2n\}$. The multi-set $\mathcal{T}'_i$ comprises only one copy of T'_i.

The multi-set $\mathcal{T}$ of trees corresponding to MAXCLIQUE, for instance G comprises the sum of multi-sets $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_n, \mathcal{T}'_i, \ldots, \mathcal{T}'_n$. Note that the sum of two multi-sets X and X' results in a multi-set Y

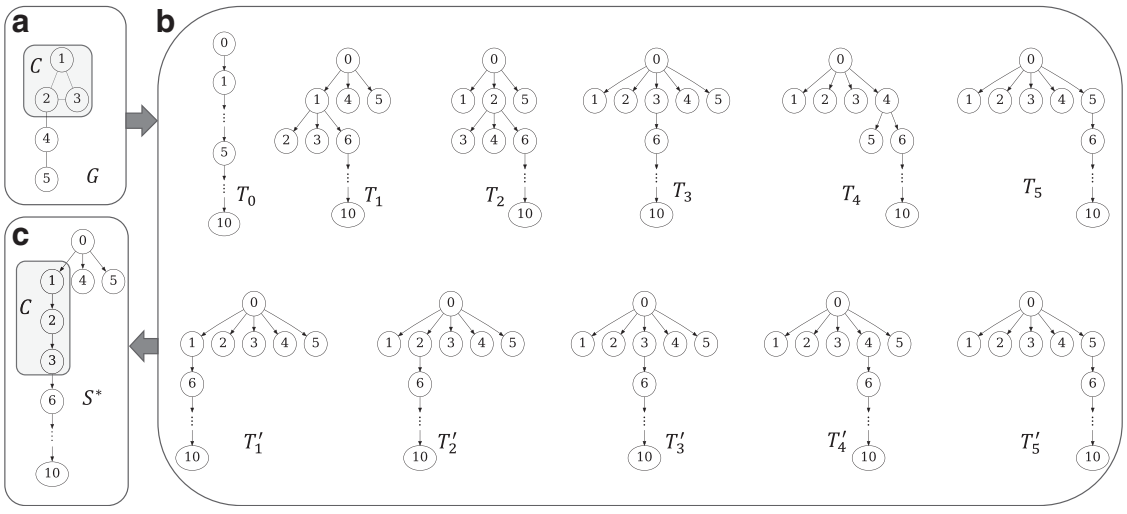

FIG. 2. An example reduction from MaxClique to ADCT. **(a)** An undirected graph G with $n = 5$ vertices and $m = 5$ edges with a maximum clique C of size 3. Here, $\delta(2) = \{1, 3, 4\}$ and $\delta^{>}(2) = \{3, 4\}$. **(b)** The corresponding trees in $\mathcal{T}$, with $n^3 - 2n^2 + 2n - 3 = 82$ copies of T_0, $n^2 + 1 = 26$ copies of T_i and one copy of T'_i for each vertex $i \in V(G)$. **(c)** The optimal consensus tree S^*. The vertices on the directed path between 0 and $n+1 = 6$ indicate the maximum clique C.

whose unique elements have a multiplicity equal to the sum of the multiplicities of that element in X and X'. As such, the total number $\mathcal{T}$ of trees equals $2n^3-2n^2+4n-3$. Clearly, the reduction can be completed in polynomial time. We have the following two lemmas characterizing the AD and branching matrix of $\mathcal{T}$, respectively.

Lemma 3. *For any $i,j \in V(\mathcal{T})$, the entry $a_{i,j}$ of the AD matrix $A_{\mathcal{T}}$ equals:*

$$a_{i,j}=\begin{cases} 2n^3-2n^2+4n-3, & \text{if } 0 = i<j \le 2n,\\ n^3-n^2+2n-2, & \text{if } 0<i<j\le n,\ (i,j)\in E(G),\\ n^3-2n^2+2n-3, & \text{if } 0<i<j\le n,\ (i,j)\notin E(G),\\ n^3-n^2+2n-1, & \text{if } 0<i\le n<j\le 2n,\\ 2n^3-2n^2+4n-3, & \text{if } n\le i<j\le 2n,\\ 0, & \text{if } 0\le j<i\le 2n. \end{cases} \tag{5}$$

Proof. See Section 6. □

Lemma 4. *For any $i,j \in V(\mathcal{T})$ such that $i<j$, entries $b_{i,j}=b_{j,i}$ of the branching matrix $B_{\mathcal{T}}$ equal:*

$$b_{i,j}=b_{j,i}=\begin{cases} 0, & \text{if } 0 = i<j \le 2n,\\ n^3-n^2+2n-1, & \text{if } 0<i<j\le n,(i,j)\in E(G),\\ n^3+2n, & \text{if } 0<i<j\le n,(i,j)\notin E(G),\\ n^3-n^2+2n-2, & \text{if } 0<i\le n<j\le 2n,\\ 0, & \text{if } n\le i<j\le 2n. \end{cases}$$

Proof. Recall that $a_{i,j}+a_{j,i}+b_{i,j}=|\mathcal{T}|$. Let $i,j\in V(\mathcal{T})$ such that $i<j$. Since $a_{j,i}=0$ by Lemma 3, we have $a_{i,j}+b_{i,j}=|\mathcal{T}|$. As such, $b_{i,j}=b_{j,i}=|\mathcal{T}|-a_{i,j}$. This lemma follows using the values of $a_{i,j}$ from Lemma 3. □

We prove the following lower bound on the distance $d(S,\mathcal{T})$ of any tree S on vertex set $V(\mathcal{T})$.

Lemma 5. *If S is a tree on $V(\mathcal{T})$, then $d(S,\mathcal{T})$ is at least $L=\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{i,j}+\sum_{i=1}^{n}\sum_{j=n+1}^{2n} b_{i,j}$.*

Proof. Recall that for any pair (i,j) of vertices in any mutation tree S, exactly one of $i\prec_S j$, $j\prec_S i$, $i\perp_S j$ is true. Therefore, a trivial lower bound on $d^{i,j}(S,\mathcal{T})$ can be obtained from Eq. (4): $d^{i,j}(S,\mathcal{T})\ge d^{i,j}_{\min}(\mathcal{T})=\min\{2a_{j,i}+b_{i,j}, 2a_{i,j}+b_{i,j}, a_{i,j}+a_{j,i}\}$. Note that $b_{i,j}=b_{j,i}$ by Definition 4. As such, $d^{i,j}_{\min}(\mathcal{T})=d^{j,i}_{\min}(\mathcal{T})$. Using Lemma 3, if $i<j$, we have $d^{i,j}_{\min}(\mathcal{T})=\min\{2a_{j,i}+b_{i,j}, 2a_{i,j}+b_{i,j}, a_{i,j}+a_{j,i}\}=\min\{b_{i,j}, 2a_{i,j}+b_{i,j}, a_{i,j}\}=\min\{b_{i,j}, a_{i,j}\}$. Further, for $i<j$, by Lemma 4, we obtain

$$d^{i,j}_{\min}(\mathcal{T})=\begin{cases} 0, & \text{if } 0 = i<j\le 2n,\\ a_{i,j}, & \text{if } 0<i<j\le n,\\ b_{i,j}, & \text{if } 0<i\le n<\text{j}\le 2n,\\ 0, & \text{if } n\le i<j\le 2n. \end{cases} \tag{6}$$

Plugging this into Eq. (3) of Lemma 2, we finally obtain

$$\begin{aligned} d(S,\mathcal{T}) &= \sum_{i=0}^{2n}\sum_{j=i+1}^{2n} d^{i,j}(S,\mathcal{T}) \ge \sum_{i=0}^{2n}\sum_{j=i+1}^{2n} d^{i,j}_{\min}(\mathcal{T})\\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{i,j}+\sum_{i=1}^{n}\sum_{j=n+1}^{2n} b_{i,j}. \end{aligned} \tag{7}$$

□

We define a C-constrained tree as follows—shown in Figure 3a.

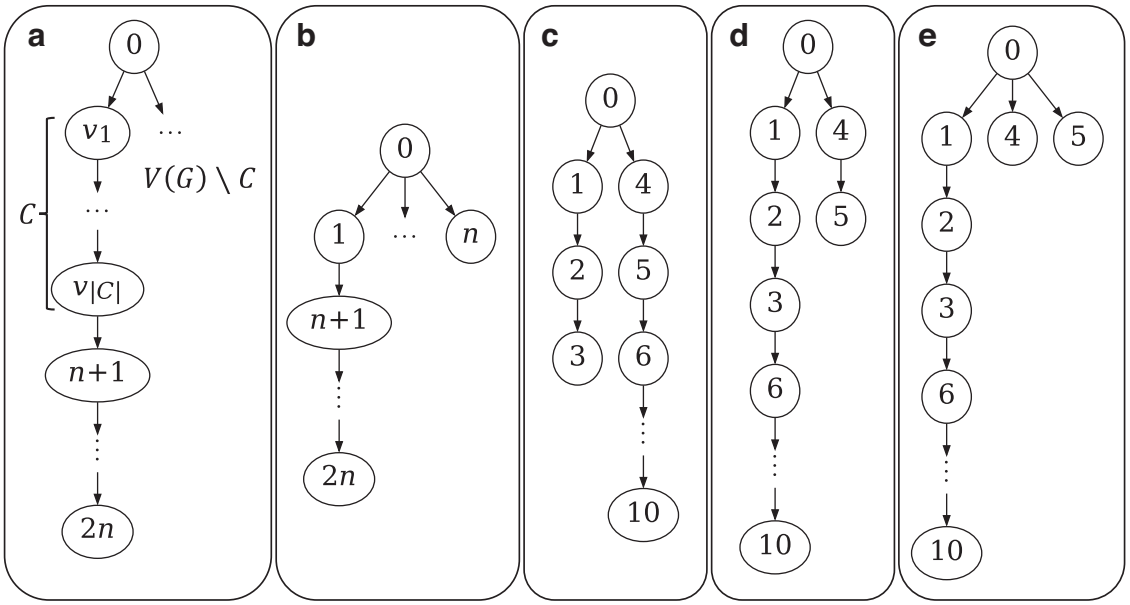

FIG. 3. **(a)** The structure of a C-constrained tree S_c as well as an optimal consensus tree S^*. **(b)** The tree used to prove an upper bound on $d(S^*, \mathcal{T})$ in Lemma 7. **(c)** An example tree based on the instance shown in Figure 2 used in Lemma 12, where the chain $\{n+1, \ldots, 2n\}$ is attached to vertex 5 which has smaller depth than vertex 3. **(d)** An example tree based on the instance shown in Figure 2 used in Lemmas 12 and 14. In the former lemma, the chain $\{n+1, \ldots, 2n\}$ is attached to vertex 5, which has higher depth than vertex 3. In the latter lemma, vertex 5 is a leaf whose parent is not 0. **(e)** An example tree based on the instance shown in Figure 2 used in Lemma 14 where the parent of 5 is 0 instead of 4.

Definition 5. *For vertices $C = \{v_1, \ldots, v_k\} \subseteq V(G)$ of G such that $v_1 < \ldots < v_k$, the* C-constrained tree S_C *has vertex set $\{0, \ldots, 2n\}$ such that (i) vertex 0 is the root, (ii) there is an edge $(0, i)$ for each vertex $i \in \{1, \ldots, n\} \setminus C$, and (iii) there is a chain $0 \to v_1 \to \ldots \to v_k \to n+1 \to \ldots \to 2n$.*

If C is a clique in G then the corresponding tree S_C induces the following distance $d(S_C, \mathcal{T})$.

Lemma 6. *For any clique C of size k of G, we have $d(S_C, \mathcal{T}) = L + n^2 - nk + k(k-1)/2$.*

Proof. Using Eq. (3) in Lemmas 2, 3, and 4 and Eq. (6) in Lemma 5, we discuss the following six cases for the difference between $d^{i,j}(S_C, \mathcal{T})$ and $d^{i,j}_{\min}(\mathcal{T})$ of vertices $0 \le i < j \le 2n$.

First, we consider $0 = i < j \le 2n$. Since $i = 0$ is the root vertex of S_C. Therefore, it holds that $i \prec_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = 2a_{j,i} + b_{i,j} = 0 = d^{i,j}_{\min}(\mathcal{T})$. Second, we consider $0 < i < j \le n$ and $i, j \in C$. In this case, i, j are on the same branch in S_C. Therefore, it holds that $i \prec_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = 2a_{j,i} + b_{i,j} = b_{i,j}$. Since C is a clique, we have $(i, j) \in E(G)$.

Therefore, $d^{i,j}(S_C, \mathcal{T}) - d^{i,j}_{\min}(\mathcal{T}) = b_{i,j} - a_{i,j} = 1$. Third, we consider $0 < i < j \le n$, and $i \notin C$ or $j \notin C$. In this case, it holds that $i \perp_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = a_{i,j} + a_{j,i} = a_{i,j} = d^{i,j}_{\min}(\mathcal{T})$. Fourth, we consider $0 < i \le n < j \le 2n$ and $i \in C$. In this case, it holds that $i \prec_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = 2a_{j,i} + b_{i,j} = b_{i,j} = d^{i,j}_{\min}(\mathcal{T})$. Fifth, we consider $0 < i \le n < j \le 2n$ and $i \notin C$.

In this case, it holds that $i \perp_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = a_{i,j} + a_{j,i} = a_{i,j}$. Therefore, $d^{i,j}(S_C, \mathcal{T}) - d^{i,j}_{\min}(\mathcal{T}) = a_{i,j} - b_{i,j} = 1$. Sixth, we consider $n \le i < j \le 2n$. It holds that $i \prec_{S_C} j$. As such, $d^{i,j}(S_C, \mathcal{T}) = 2a_{j,i} + b_{i,j} = 0 = d^{i,j}_{\min}(\mathcal{T})$.

Thus, only the second and fifth case have a non-zero value for $(d^{i,j}(S_C, \mathcal{T}) - d^{i,j}_{\min}(\mathcal{T}))$. Putting everything together, we have that $d(S_C, \mathcal{T}) - L$ equals

$$\sum_{i=0}^{2n} \sum_{j=0}^{2n} (d^{i,j}(S_C, \mathcal{T}) - d^{i,j}_{\min}(\mathcal{T})) = \sum_{i<j,\, i,j \in C} 1 + \sum_{0<i\le n,\, i \notin C} \sum_{j=n+1}^{2n} 1$$
$$= n^2 - nk + k(k-1)/2.$$

This proves the lemma. □

Our goal now is to show that an optimal consensus tree S^* of the multi-set $\mathcal{T}$ obtained from the undirected graph G is a C-constrained tree such that C is a clique of G. To this end, we establish the following useful upper bound on $d(S^*, \mathcal{T})$.

Lemma 7. *It holds that $d(S^*, \mathcal{T})$ is at most $L+n^2-n$.*

Proof. To prove the lemma, we consider a tree $S=(V, E(S))$, $E(S)=\{(0,i)|0<i\leq n\}\cup\{1,(n+1)\}\cup\{(i,i+1)|n<i<2n\}$ as shown in Figure 3b. By Lemmas 2, 3, and 4, we have $d(S,\mathcal{T})=\sum_{i=0}^{2n}\sum_{j=i+1}^{2n} d^{i,j}(S,\mathcal{T})$ where

$$d^{i,j}(S,\mathcal{T})=\begin{cases}0, & \text{if } 0=i<j\leq 2n,\\ a_{i,j}, & \text{if } 0<i<j\leq n,\\ b_{i,j}, & \text{if } 1=i\leq n<j\leq 2n,\\ a_{i,j}, & \text{if } 1<i\leq n<j\leq 2n,\\ 0, & \text{if } n\leq i<j\leq 2n.\end{cases}$$

Observe that $a_{i,j}=b_{i,j}+1$ for $0<i\leq n<j\leq 2n$ in Lemmas 3 and 4. Using the lower bound L established in Lemma 5, we have

$$\begin{aligned}d(S,\mathcal{T})&=\sum_{i=0}^{2n}\sum_{j=i+1}^{2n} d^{i,j}(S,\mathcal{T})=\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{i,j}+\sum_{j=n+1}^{2n} b_{1,j}+\sum_{i=2}^{n}\sum_{j=n+1}^{2n}(b_{i,j}+1)\\&=L+n^2-n.\end{aligned}$$

Hence, $d(S^*,\mathcal{T})\leq d(S,\mathcal{T})=L+n^2-n$. □

We now reason about the topology of S^*. The following lemma shows that j cannot be an ancestor of i in S^* if $i<j$.

Lemma 8. *For any pair (i,j) of vertices such that $0\leq i<j\leq 2n$, either $i\prec_{S^*} j$ or $i\perp_{S^*} j$.*

Proof. See Section 6. □

Our reduction enforces that 0 is the root of S^* and that $2n$ is a leaf.

Lemma 9. *Vertex 0 is the root of S^*.*

Proof. Suppose for a contradiction that $0<j\leq 2n$ is the root of S^*. Consider vertex 0. Since j is the root, it holds that $j\prec_{S^*}0$. However, since $0<j$, by Lemma 8, it must hold that either $0\prec_{S^*} j$ or $0\perp_{S^*} j$, yielding a contradiction. Hence, vertex 0 must be the root of S^*. □

Lemma 10. *The subgraph of S^* induced by vertices $\{0,\ldots,n\}$ forms a tree.*

Proof. It suffices to prove that no vertex $\{n+1,\ldots,2n\}$ is an ancestor of a vertex $\{1,\ldots,n\}$ in S^*. Suppose for a contradiction there exist vertices $0<i\leq n<j\leq 2n$ such that $j\prec_{S^*} i$. Since $i<j$, by Lemma 8, it must hold that either $i\prec_{S^*} j$ or $i\perp_{S^*} j$, yielding a contradiction. □

Moreover, vertices $\{n+1,\ldots,2n\}$ form a chain from $n+1$ to $2n$ in ascending order as shown by the following lemma.

Lemma 11. *For any pair (i,j) of vertices such that $n<i<j\leq 2n$, it holds that $i\prec_{S^*} j$.*

Proof. Suppose for a contradiction that $u\not\prec_{S^*} v$ for some $n<u<v\leq 2n$. By Lemma 8, we have $v\not\prec_{S^*} u$. Therefore, $u\perp_{S^*} v$, that is, u and v are branched in S^*. By Eq. (3) in Lemmas 2 and 3, we have $d^{u,v}(S^*,\mathcal{T})=a_{u,v}$. As such,

$$\begin{aligned}d(S,\mathcal{T})=\sum_{i=0}^{2n}\sum_{j=i+1}^{2n} d^{i,j}(S,\mathcal{T})&=d^{u,v}(S,\mathcal{T})+\sum_{0<i<j\leq 2n,\,(i,j)\neq(u,v)} d^{i,j}(S,\mathcal{T})\\&\geq a_{u,v}+\sum_{0<i<j\leq n,\,(i,j)\neq(u,v)} d^{i,j}_{\min}(S,\mathcal{T})\\&=L+2n^3-2n^2+4n-3\end{aligned}$$

Since $2n^3 - 2n^2 + 4n - 3 > n^2 - n$, Lemma 7, which states that $d(S^*, \mathcal{T}) \leq L + n^2 - n$, implies that S^* is not an optimal consensus tree, a contradiction. □

The root vertex r of a mutation tree S has $\text{depth}_S(r) = 0$, and every vertex $v \neq r$ with parent u has $\text{depth}_S(v) = \text{depth}_S(u) + 1$. In Lemma 11, we have shown that the chain $\{n+1, \ldots, 2n\}$ remains intact in an optimal consensus tree S^*. In the following lemma, we will show that this chain is attached to a maximum-depth vertex among $\{0, \ldots, n\}$ in S^*.

Lemma 12. *The parent of vertex $n+1$ in S^* is a vertex in the set $\{0, \ldots, n\}$ with maximum depth.*

Proof. Let i be the parent of vertex $n+1$. Lemma 11 states that the chain $n+1 \to \ldots \to 2n$ is kept intact in S^*. This means that i must be in $\{0, \ldots, n\}$. Suppose for a contradiction that $\text{depth}_{S^*}(i)$ does not have the maximum depth among vertices $\{0, \ldots, n\}$. Therefore, there is a vertex $0 \leq j \leq n$ such that $\text{depth}_{S^*}(j) > \text{depth}_{S^*}(i)$—as illustrated in Figure 3c. Let P_i be the unique path from 0 to i. Let P_j be the unique path from 0 to j.

Since $\text{depth}_{S^*}(j) > \text{depth}_{S^*}(i)$, we have $|V(P_i)| < |V(P_j)|$. We remove the chain and re-attach it to the higher-depth vertex j, yielding $S = (V, E(S))$, where $E(S) = (E(S^*) \setminus \{(i, n+1)\}) \cup \{(j, n+1)\}$ as shown in Figure 3d. By Lemma 10, S is a tree. We will show that $d(S, \mathcal{T}) < d(S^*, \mathcal{T})$ by distinguishing four cases regarding the placement of vertices $0 \leq u < v \leq 2n$.

First, we consider $0 \leq u < v \leq n$ or $n < u < v \leq 2n$. In the former case, u, v are located outside the chain and in the latter case, inside the chain. In both cases, the relation between u and v is the same in both S^* and S. As such, $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = 0$ by Eq. (3) in Lemma 2. Second, we consider $0 \leq u \leq n < v \leq 2n$ and $u \in V(P_i) \cap V(P_j)$. The relation between u and v also stays the same, and u is an ancestor of v in both S^* and S.

Similar to the previous case, we have $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = 0$. Third, we consider $0 \leq u \leq n < v \leq 2n$ and $u \in V(P_i) \setminus V(P_j)$. Thus, u is an ancestor of v in S^*; however, they are branched in S. By Eq. (3) in Lemmas 2, 3, and 4, $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = b_{u,v} - a_{u,v} = -1$. Fourth, we consider $0 \leq u \leq n < v \leq 2n$ and $u \in V(P_j) \setminus V(P_i)$. Thus, u is an ancestor of v in S; however, they are branched in S^*. Similar to the previous case, we have $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = a_{u,v} - b_{u,v} = 1$. Therefore,

$$\begin{aligned} d(S^*, \mathcal{T}) - d(S, \mathcal{T}) &= \sum_{u \in V(P_i) \Delta V(P_j)} \sum_{v=n+1}^{2n} (d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T})) \\ &= (|V(P_j)| - |V(P_i)|)n > 0. \end{aligned}$$

Note that Δ indicates the symmetric difference. This contradicts that S^* is optimal. □

We have that non-adjacent vertices i, j in G must be branched in an optimal consensus tree S^*.

Lemma 13. *For any pair $i, j \in V(G)$ of distinct vertices where $(i, j) \notin E(G)$, it holds that $i \perp_{S^*} j$.*

Proof. Suppose for a contradiction there exist vertices $u, v \in V(G)$ such that $(u, v) \notin E(G)$ and $u \not\perp_{S^*} v$. WLOG, we assume $u < v$. By Lemma 8, either $u \prec_{S^*} v$ or $u \perp_{S^*} v$. Therefore, it holds that $u \prec_{S^*} v$. By Eq. (3) in Lemma 2 and Lemma 3, $d^{u,v}(S, \mathcal{T}) = b_{u,v}$. As such, $d(S, \mathcal{T})$ equals

$$\begin{aligned} \sum_{i=0}^{2n} \sum_{j=i+1}^{2n} d^{i,j}(S, \mathcal{T}) &= d^{u,v}(S, \mathcal{T}) + \sum_{0 < i < j \leq 2n,\, (i,j) \neq (u,v)} d^{i,j}(S, \mathcal{T}) \\ &\geq b_{u,v} + \sum_{(i<j),\, (i,j) \in V,\, (i,j) \neq (u,v)} d^{i,j}_{\min}(S, \mathcal{T}) \\ &= L + b_{u,v} - a_{u,v} = L + 2n^2 + 3 \end{aligned}$$

Since $2n^2 + 3 > n^2 - n$, Lemma 7, which states that $d(S^*, \mathcal{T}) \leq L + n^2 - n$, implies S^* is not an optimal consensus tree, a contradiction. □

In the following lemma, we show that S^* is a star tree except for one linear branch containing a subset $C \subseteq \{1, \ldots, n\}$ of vertices and terminating with the chain $\{n+1, \ldots, 2n\}$ (Fig. 3a, e).

Lemma 14. *Let C be the vertices on the unique path from vertex 0 to vertex $n+1$, excluding 0 and $n+1$. Then, vertex 0 is the parent of all vertices $\{1, \ldots, n\} \setminus C$ in S^*.*

Proof. Suppose for a contradiction that there are vertices in $\{1, \ldots, n\}\backslash C$ whose parents are not 0 in S^*. Among these vertices, consider a leaf vertex i. Let vertex $j \neq 0$ be the parent of i (see Fig. 3d where $i=5$ and $j=4$). Recall that by Lemma 9, vertex 0 is the root of S^*. Let V_i be the vertices on the unique path from vertex 0 to i, excluding 0 and i. By Lemma 8, V_i consists of vertices u such that $u < i$. Consider the tree S where we attach vertex i to the root 0, that is, $E(S)=E(S^*)\backslash\{(j, i)\} \cup \{(0, i)\}$. See Figure 3e for an example.

To compute $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T})$, we distinguish two cases for vertices $0 \leq u < v \leq 2n$. First, we consider $v=i$ and $u \in V_i$. Thus, $u \prec_{S^*} v$. By the contrapositive of Lemma 13, we have $(u, i) \in E(G)$ for any $u \in V_i$. Since $u \perp_S v$, by Eq. (3) in Lemmas 2, 3, and 4, we have $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = b_{u,v} - a_{u,v} = 1$. Second, we consider the case where $v \neq i$ or $u \notin V_i$. The relationship between u and v is the same in S and S^*. As such, $d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = 0$ by Eq. (3) in Lemma 2. Therefore,

$$d(S^*, \mathcal{T}) - d(S, \mathcal{T}) = \sum_{u \in V_i} d^{u,v}(S^*, \mathcal{T}) - d^{u,v}(S, \mathcal{T}) = |V_i| > 0$$

This contradicts that S^* is optimal and thus proves the lemma. □

Finally, we show that vertices C of S^* are, indeed, a clique of G.

Lemma 15. *The vertices C of S^* on the unique path from vertex 0 to vertex $n+1$, excluding 0 and $n+1$, form a clique in G.*

Proof. By Lemma 9, vertex 0 is the root of S^*. Therefore, for any $i, j \in C$, we have $i \not\perp_{S^*} j$. By the contrapostive of Lemma 13, $(i, j) \in E(G)$ for all $i, j \in C$. Hence, C is a clique of G. □

Corollary 1. *Any optimal consensus tree S^* is a C-constrained tree such that C is a clique of G.*

Lemma 16. *For any subset $C \subseteq V(G)$ of vertices, the C-constrained tree S_C is an optimal consensus tree of $\mathcal{T}$ if and only if C is a maximum clique in G.*

Proof. ($\Rightarrow$) Let S_C be an optimal C-constrained consensus tree. By Lemma 15, we know that C is a clique. Let $|C|=k$. By Lemma 6, we have $d(S_C, \mathcal{T}) = L + n^2 - nk + k(k-1)/2$. Suppose for a contradiction that C is not a maximum clique of G. By our premise, there must exist another clique C' such that $|C'|=\ell > k = |C|$. Let $S_{C'}$ be the corresponding C'-constrained tree following Definition 5. By Lemma 6, we have $d(S_{C'}, \mathcal{T}) = L + n^2 - n\ell + \ell(\ell-1)/2$. Since $n \geq \ell \geq k+1$,

$$\begin{aligned} d(S_{C'}, \mathcal{T}) - d(S_C, \mathcal{T}) &= (n + \frac{1}{2})(k-\ell) + \frac{(k+\ell)(\ell-k)}{2} \\ &< (n + \frac{1}{2})(k-\ell) + n(\ell-k) \leq -\frac{1}{2} < 0, \end{aligned}$$

which contradicts that S_C is optimal.

($\Leftarrow$) Let C be a maximum clique of G such that $|C|=k$. Suppose for a contradiction that the corresponding C-constrained tree S_C is not an optimal consensus tree of $\mathcal{T}$. By Lemma 6, we have $d(S_C, \mathcal{T}) = L + n^2 - nk + k(k-1)/2$. Therefore, by Corollary 1, there exists an optimal C'-constrained consensus tree $S_{C'}$, where $|C'|=\ell$, such that the distance $d(S_{C'}, \mathcal{T})$ is strictly less than $d(S_C, \mathcal{T})$. By Lemma 6, $d(S_{C'}, \mathcal{T}) = L + n^2 - n\ell + \ell(\ell-1)/2$. We have

$$\begin{aligned} d(S_{C'}, \mathcal{T}) - d(S_C, \mathcal{T}) &= (n + \frac{1}{2})(k-\ell) + \frac{(k+\ell)(\ell-k)}{2} \\ &= (n + \frac{1}{2} - \frac{k+\ell}{2})(k-\ell) < 0. \end{aligned}$$

Since $k, \ell \leq n$, we have $(k+\ell)/2 \leq n$. This implies that $k - \ell < 0$, which contradicts that C is a maximum clique of size k. □

5. DISCUSSION

In this work, we demonstrated the NP-hardness of the consensus tree problem under the AD distance. While the problem of finding a maximum clique for a graph with n vertices is hard to approximate within a

factor of $O(n^{1-\varepsilon})$ for any real number $\varepsilon > 0$ unless $P = \text{NP}$ (Zuckerman, 2006), our reduction is not approximation-factor preserving. As such, one might be able to achieve better approximation factors for the consensus tree problem under the AD distance, including possibly constant factors. We will investigate this in future work.

ACKNOWLEDGMENTS

An earlier draft of this manuscript was posted as a preprint at bioRxiv (doi: 10.1101/2023.07.17.549375).

AUTHORS' CONTRIBUTIONS

Y.Q. Conceptualization, formal analysis, and writing-original draft. M.E.-K.: Conceptualization, validation, writing–review, and editing.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

M.E.-K. was supported by the National Science Foundation (CCF-2046488) as well as funding from the Cancer Center at Illinois.

REFERENCES

Aguse N, Qi Y, El-Kebir M. Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. Bioinformatics 2019;35(14):i408–i416.

Christensen S, Kim J, Chia N, et al. Detecting evolutionary patterns of cancers using consensus trees. Bioinformatics 2020;36(Suppl. 2):i684–i691.

Cook SA. The complexity of theorem-proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing. 1971; pp. 151–158.

DiNardo Z, Tomlinson K, Ritz A, et al. Distance measures for tumor evolutionary trees. Bioinformatics 2020;36(7): 2090–2097.

El-Kebir M, Satas G, Oesper L, et al. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Systems 2016;3(1):43–53.

Fu X, Schwartz R. Contreedp: A consensus method of tumor trees based on maximum directed partition support problem. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021; pp. 125–130.

Govek K, Sikes C, Oesper L. A consensus approach to infer tumor evolutionary histories. BCB 2018;63–72.

Govek K, Sikes C, Zhou Y, et al. Graphyc: Using consensus to infer tumor evolution. IEEE/ACM Trans Comp Biol Bioinform 2020;19(1):465–478.

Guang Z, Smith-Erb M, Oesper L. A weighted distance-based approach for deriving consensus tumor evolutionary trees. Bioinformatics 2023;39(Suppl. 1):i204–i212; doi: 10.1093/bioinformatics/btad230.

Karp RM. Reducibility among combinatorial problems. In: Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department. (Miller RE, Thatcher JW, Bohlinger JD. eds.) Springer US: Boston, MA, USA; 1972; pp. 85–103.

Karpov N, Malikic S, Rahman M, et al. A multi-labeled tree dissimilarity measure for comparing ''clonal trees'' of tumor progression. Algorithms Mol Biol 2019;14(1):1–18.

Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 1969;61(4):893.

Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194(4260):23–28.
Qi Y, Pradhan D, El-Kebir M. Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors. Algorithms Mol Biol 2019;14:1–14.
Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: Principles and practice. Nat Rev Genet 2017;18(4):213–229.
Zuckerman D. Linear degree extractors and the inapproximability of max clique and chromatic number. In: Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing. 2006; pp. 681–690.

Address correspondence to:
Dr. Mohammed El-Kebir
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801
USA

E-mail: melkebir@illinois.edu

6. APPENDIX

6.1. Supplementary proofs

Lemma 1. *The AD distance $d(T_1, T_2)$ for trees T_1 and T_2 on the same vertex set $[n]$ equals*

$$d(T_1, T_2) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(T_1, T_2) \tag{8}$$

where $d^{i,j}(T_1, T_2)$ is the distance contributed by the unordered pair $\{i, j\}$ of distinct vertices defined as

$$\begin{aligned} d^{i,j}(T_1, T_2) \quad &= \mathbf{1}\{i \prec_{T_1} j\}(2 \cdot \mathbf{1}\{j \prec_{T_2} i\} + \mathbf{1}\{i \perp_{T_2} j\}) + \mathbf{1}\{j \prec_{T_1} i\}(2 \cdot \mathbf{1}\{i \prec_{T_2} j\} + \mathbf{1}\{i \perp_{T_2} j\}) \\ &\quad + \mathbf{1}\{i \perp_{T_1} j\}(\mathbf{1}\{i \prec_{T_2} j\} + \mathbf{1}\{j \prec_{T_2} i\}). \end{aligned} \tag{9}$$

Proof. By Definition 2, we have

$$\begin{aligned} d(T_1, T_2) &= |A(T_1) \backslash A(T_2)| + |A(T_2) \backslash A(T_1)| \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{i \prec_{T_1} j\} \mathbf{1}\{i \not\prec_{T_2} j\} + \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{i \prec_{T_2} j\} \mathbf{1}\{i \not\prec_{T_1} j\} \\ &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\mathbf{1}\{i \prec_{T_1} j\} \mathbf{1}\{i \not\prec_{T_2} j\} + \mathbf{1}\{j \prec_{T_1} i\} \mathbf{1}\{j \not\prec_{T_2} i\}) \\ &\quad + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\mathbf{1}\{i \prec_{T_2} j\} \mathbf{1}\{i \not\prec_{T_1} j\} + \mathbf{1}\{j \prec_{T_2} i\} \mathbf{1}\{j \not\prec_{T_1} i\}\right) \\ &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\mathbf{1}\{i \prec_{T_1} j\}(\mathbf{1}\{j \prec_{T_2} i\} + \mathbf{1}\{i \perp_{T_2} j\}) + \mathbf{1}\{i \prec_{T_2} j\}(\mathbf{1}\{j \prec_{T_1} i\} + \mathbf{1}\{i \perp_{T_1} j\}) \\ &\quad + \mathbf{1}\{j \prec_{T_1} i\}(\mathbf{1}\{i \prec_{T_2} j\} + \mathbf{1}\{i \perp_{T_2} j\}) + \mathbf{1}\{j \prec_{T_2} i\}(\mathbf{1}\{i \prec_{T_1} j\} + \mathbf{1}\{i \perp_{T_1} j\})] \\ &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(T_1, T_2). \end{aligned}$$

□

Lemma 2. *The AD distance $d(S, \mathcal{T})$ between a tree S and trees $\mathcal{T}$ on the same vertex set $[n]$ equals*

$$d(S, \mathcal{T}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(S, \mathcal{T}) \tag{10}$$

where $d^{i,j}(S, \mathcal{T})$ is the distance contributed by the unordered pair $\{i, j\}$ of distinct vertices defined as

$$d^{i,j}(S, \mathcal{T}) = \mathbf{1}\{i \prec_S j\}\left(2a_{j,i} + b_{i,j}\right) + \mathbf{1}\{j \prec_S i\}\left(2a_{i,j} + b_{i,j}\right) + \mathbf{1}\{i \perp_S j\}\left(a_{i,j} + a_{j,i}\right). \tag{11}$$

Proof. We apply Lemma 1 and obtain

$$\begin{aligned}
d(S, \mathcal{T}) &= \sum_{T \in \mathcal{T}} d(S, T) = \sum_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(S, T) \\
&= \sum_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\mathbf{1}\{i \prec_S j\}(2 \cdot \mathbf{1}\{j \prec_T i\} + \mathbf{1}\{i \perp_T j\}) \\
&\quad + \mathbf{1}\{j \prec_S i\}(2 \cdot \mathbf{1}\{i \prec_T j\} + \mathbf{1}\{i \perp_T j\}) \\
&\quad + \mathbf{1}\{i \perp_S j\}(\mathbf{1}\{i \prec_T j\} + \mathbf{1}\{j \prec_T i\})] \\
&= \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\mathbf{1}\{i \prec_S j\}\left(2 \sum_{T \in \mathcal{T}} \mathbf{1}\{j \prec_T i\} + \sum_{T \in \mathcal{T}} \mathbf{1}\{i \perp_T j\}\right) \\
&\quad + \mathbf{1}\{j \prec_S i\}\left(2 \sum_{T \in \mathcal{T}} \mathbf{1}\{i \prec_T j\} + \sum_{T \in \mathcal{T}} \mathbf{1}\{i \perp_T j\}\right) \\
&\quad + \mathbf{1}\{i \perp_S j\}\left(\sum_{T \in \mathcal{T}} \mathbf{1}\{i \prec_T j\} + \sum_{T \in \mathcal{T}} \mathbf{1}\{j \prec_T i\}\right)] = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^{i,j}(S, \mathcal{T}).
\end{aligned}$$

□

Lemma 3. *For any $i, j \in V(\mathcal{T})$, the entry $a_{i,j}$ of the AD matrix $A_{\mathcal{T}}$ equals:*

$$a_{i,j} = \begin{cases} 2n^3 - 2n^2 + 4n - 3, & \text{if } 0 = i < j \leq 2n, \\ n^3 - n^2 + 2n - 2, & \text{if } 0 < i < j \leq n,\ (i,\ j)\ \in\ E(G), \\ n^3 - 2n^2 + 2n - 3, & \text{if } 0 < i < j \leq n,\ (i,\ j)\ \notin\ E(G), \\ n^3 - n^2 + 2n - 1, & \text{if } 0 < i \leq n < j \leq 2n, \\ 2n^3 - 2n^2 + 4n - 3, & \text{if } n \leq i < j\ \leq 2n, \\ 0, & \text{if } 0 \leq j < i\ \leq 2n. \end{cases} \tag{12}$$

Proof. We prove the lemma by examining each of the six cases separately. For the first case, consider a pair (i, j) such that $0 = i < j \leq 2n$. Recall that $i = 0$ is the root vertex of all trees in $\mathcal{T}$. Thus, it holds that $i \prec_T j$ for any $T \in \mathcal{T}$. Therefore, $a_{i,j} = |\mathcal{T}| = 2n^3 - 2n^2 + 4n - 3$. For the second case, consider a pair (i, j) such that $0 < i < j \leq n, (i, j) \in E(G)$.

Then, $i \prec_T j$ for all trees T in $\mathcal{T}_0$ and $\mathcal{T}_i$. However, $i \not\prec_T j$ for any tree T in the remaining multi-sets different from $\mathcal{T}_0$ and $\mathcal{T}_i$. Therefore, $a_{i,j} = |\mathcal{T}_0| + |\mathcal{T}_i| = (n^3 - 2n^2 + 2n - 3) + (n^2 + 1) = n^3 - n^2 + 2n - 2$. For the third case, consider a pair (i, j) such that $0 < i < j \leq n (i, j) \notin E(G)$. Then, $i \prec_T j$ for all trees T in $\mathcal{T}_0$. However, $i \not\prec_T j$ for any tree T in the remaining multi-sets different from $\mathcal{T}_0$. Therefore, $a_{i,j} = |\mathcal{T}_0| = n^3 - 2n^2 + 2n - 3$.

For the fourth case, consider a pair (i, j) such that $0 < i \leq n < j \leq 2n$. Then, $i \prec_T j$ for all trees T in the multi-sets $\mathcal{T}_0, \mathcal{T}_i$ and $\mathcal{T}_i'$. However, $i \not\prec_T j$ for any tree T in the remaining multi-sets. Therefore,

$a_{i,j} = |\mathcal{T}_0| + |\mathcal{T}_i| + |\mathcal{T}'_i| = (n^3 - 2n^2 + 2n - 3) + (n^2 + 1) + 1 = n^3 - n^2 + 2n - 1$. For the fifth case, consider a pair (i, j) such that $n < i < j \leq 2n$.

By construction, the chain $n+1, \ldots, 2n$ is kept intact in every tree. Thus, $i \prec_T j$ for any tree $T \in \mathcal{T}$. Therefore, $a_{i,j} = |\mathcal{T}| = 2n^3 - 2n^2 + 4n - 3$. Finally, consider (i, j) such that $0 \leq j < i \leq 2n$. For any $T \in \mathcal{T}$ and each edge $(i, j) \in E(T)$, it holds that $i < j$. Therefore, it holds that $i \not\prec_T j$ and thus $a_{i,j} = 0$ if $i > j$. □

Lemma 8. *For any pair (i, j) of vertices such that $0 \leq i < j \leq 2n$, either $i \prec_{S^*} j$ or $i \perp_{S^*} j$.*

Proof. To prove this lemma, consider a tree S such that $v \prec_S u$ for some $0 \leq u < v < 2n$. By Eq. (3) in Lemma 2, $d^{u,v}(S, \mathcal{T}) = 2a_{u,v} + b_{u,v}$. We distinguish three cases regarding the occurrence of u and v, and show for each case that the resulting distance $d(S, \mathcal{T})$ will exceed the upper bound established in Lemma 7. First, consider $0 < u < v \leq n$. Then, $d^{u,v}_{\min}(\mathcal{T}) = a_{u,v}$ by Eq. (6), yielding

$$\begin{aligned} d(S, \mathcal{T}) = \sum_{i=0}^{2n} \sum_{j=i+1}^{2n} d^{i,j}(S, \mathcal{T}) &= d^{u,v}(S, \mathcal{T}) + \sum_{0<i<j\leq 2n,\,(i,j)\neq(u,v)} d^{i,j}(S, \mathcal{T}) \\ &\geq 2a_{u,v} + b_{u,v} + \sum_{0<i<j\leq 2n,\,(i,j)\neq(u,v)} d^{i,j}_{\min}(\mathcal{T}) \\ &= L + a_{u,v} + b_{u,v} = L + 2n^3 - 2n^2 + 4n - 3. \end{aligned}$$

Since $2n^3 - 2n^2 + 4n - 3 > n^2 - n$, Lemma 7, which states that $d(S^*, \mathcal{T}) \leq L + n^2 - n$, implies S is not an optimal consensus tree.

Second, consider $0 < u \leq n < v < 2n$. Then, $d^{u,v}_{\min}(\mathcal{T}) = b_{u,v}$ by Eq. (6), yielding

$$\begin{aligned} d(S, \mathcal{T}) = \sum_{i=0}^{2n} \sum_{j=i+1}^{2n} d^{i,j}(S, \mathcal{T}) &= d^{u,v}(S, \mathcal{T}) + \sum_{0<i<j\leq 2n,\,(i,j)\neq(u,v)} d^{i,j}(S, \mathcal{T}) \\ &\geq 2a_{u,v} + b_{u,v} + \sum_{0<i<j\leq n,\,(i,j)\neq(u,v)} d^{i,j}_{\min}(\mathcal{T}) \\ &= L + 2a_{u,v} = L + 2n^3 - 2n^2 + 4n - 4 \end{aligned}$$

Since $2n^3 - 2n^2 + 4n - 4 > n^2 - n$, Lemma 7, which states that $d(S^*, \mathcal{T}) \leq L + n^2 - n$, implies S is not an optimal consensus tree.

Third, consider $u = 0$ or $n < u < v \leq 2n$. Then, $d^{u,v}_{\min}(\mathcal{T}) = b_{u,v} = 0$ by Eq. (6), yielding

$$\begin{aligned} d(S, \mathcal{T}) = \sum_{i=0}^{2n} \sum_{j=i+1}^{2n} d^{i,j}(S, \mathcal{T}) &= d^{u,v}(S, \mathcal{T}) + \sum_{0<i<j\leq 2n,\,(i,j)\neq(u,v)} d^{i,j}(S, \mathcal{T}) \\ &\geq 2a_{u,v} + b_{u,v} + \sum_{0<i<j\leq n,\,(i,j)\neq(u,v)} d^{i,j}_{\min}(\mathcal{T}) \\ &= L + 4n^3 - 4n^2 + 8n - 6 \end{aligned}$$

Since $4n^3 - 4n^2 + 8n - 6 > n^2 - n$, Lemma 7, which states that $d(S^*, \mathcal{T}) \leq L + n^2 - n$, implies S is not an optimal consensus tree. □