Cophenetic Median Trees Under the Manhattan Distance

Alexey Markin
Department of Computer Science
Iowa State University
Ames, IA 50011
amarkin@iastate.edu

Oliver Eulenstein
Department of Computer Science
Iowa State University
Ames, IA 50011
oeulenst@iastate.edu

ABSTRACT

Computing median trees from gene trees using path-difference metrics has provided several credible species tree estimates. Similar to these metrics is the cophenetic family of metrics that originates from a dendrogram comparison metric introduced more than 50 years ago. Despite the tradition and appeal of the cophenetic metrics, the problem of computing median trees under this family of metrics has not been analyzed. Like other standard median tree problems relevant in practice, as we show here, this problem is also NP-hard. NP-hard median tree problems have been successfully addressed by local search heuristics that are solving thousands of instances of a corresponding local search problem. For the local search problem under a cophenetic metric the best known (naïve) algorithm has a time complexity that is typically prohibitive for effective heuristic searches. Focusing on the Manhattan norm (Manhattan cophenetic metric), we describe an efficient algorithm for this problem that improves on the naïve solution by a factor of n, where n is the size of the input trees. We demonstrate the performance of our local search algorithm in a comparative study using published empirical data sets.

CCS CONCEPTS

• **Applied computing** → **Molecular evolution**; **Bioinformatics**;

KEYWORDS

Cophenetic distance, median trees, phylogenetics, local search, SPR

ACM Reference format:
Alexey Markin and Oliver Eulenstein. 2017. Cophenetic Median Trees Under the Manhattan Distance. In *Proceedings of ACM-BCB'17, August 20-23, 2017, Boston, MA, USA., ,* 9 pages.
DOI: http://dx.doi.org/10.1145/3107411.3107443

Reconstruction of the evolutionary history, commonly modeled as a phylogenetic tree, is one of the central problems in biology. Solutions to this problem have direct applications in almost every discipline of natural sciences including medicine, epidemiology, biochemistry, agronomy, environmental sciences, and linguistics [15, 17, 18, 28, 32]. While the paramount goal of evolutionary biology is the representation of the true phylogenetic tree for millions of existing species, today the computation of phylogenies even for hundreds of species is a greatly challenging problem.

A well-established framework for large-scale phylogenetic tree inference comprises the construction of a species tree, called a *supertree*, from a collection of smaller trees each covering a subset of the species in question. Such collections of trees appear naturally either as so-called *gene trees* – the trees describing the evolution of a gene shared by species [16, 27]; or as trees collected from various sources in an attempt to unify the previously inferred knowledge and present a single phylogenetic tree with high confidence in it [5].

Supertree construction is a challenging problem on its own, since in practice the input trees often appear to be in disagreement, meaning that they suggest discordant evolutionary histories [5]. A classic approach for addressing this problem entails the search for a supertree that *fits* all the input trees as best as possible. The goodness of fit is mathematically assessed by a *cost function* that computes dissimilarity between any two phylogenetic trees. The resulting supertree search problem is called the *median tree problem* under the respective cost function. The problem has been extensively studied both from the theoretical and applications perspectives [5]. While all studied median tree problems of interest are inherently hard, various heuristics estimating median trees, such as the classic maximum representation with parsimony (MRP), hold enormous value for the biological community and have become an essential tool in practice [8, 19, 30, 38, 40].

The recent surge of interest in median trees under one of the oldest and extensively studied metrics in comparative phylogenetics, the path-difference distance [25, 26], opened a promising path for studies of median trees under other vector-based objectives. Perhaps, the most well-established one of them is the *cophenetic distance* [9]. The cophenetic distance was established based on one of the most popular dendrogram (which can be seen to be equivalent to a bijectively labeled weighted phylogenetic tree) comparison methods introduced by Sokal and Rohlf more than 50 years ago [35].

Similarly to the path-difference distance that uses the encoding of trees as path-length vectors, the cophenetic distance uses a vector encoding called a *cophenetic vector*. A cophenetic vector contains information about a distance from the least common ancestor (LCA) of two taxa to the root of the tree for *each* pair of taxa in a given rooted tree. A cophenetic vector *equivalently* encodes a phylogenetic tree [9]; hence one can measure a distance between two trees in the cophenetic vector space. Such distance can be formulated in terms of common vector norms, such as the Manhattan norm (also known as the Taxicab norm) and the Euclidean norm; the corresponding tree metrics are called the *Manhattan cophenetic distance* and the *Euclidean cophenetic distance* respectively. In

ACM-BCB'17, August 20-23, 2017, Boston, MA, USA.

DOI: http://dx.doi.org/10.1145/3107411.3107443

contrast to other popular comparative metrics, such as Robinson-Foulds, the cophenetic metrics can be similarly defined for weighted phylogenetic trees as well, which gives it an additional perspective advantage.

While both the cophenetic metrics and the path-difference metrics use vector distances, apart from that the two distance families do not bare much similarity. Primarily, the cophenetic distance relies on the LCA-mappings which brings it semantically closer to the classic model-based *deep coalescence* cost function [23] and the related *duplications with losses* cost function [6]. Indeed, in this work we experimentally establish that the cophenetic metrics are more closely correlated with the classic model-based deep coalescence cost function than the path-difference metrics. Further, in contrast to the path-difference metrics, which are only insignificantly perturbed by the rootings of the trees under consideration, the cophenetic metrics rely heavily on the rootings, and therefore, are better suited for the inference of rooted phylogenies.

As most of the other median tree problems are inherently complex, unsurprisingly, here we show that the cophenetic median tree problem is NP-hard. In fact, even the commonly practiced local search estimation approach implemented naïvely turns out to be computationally heavy and already infeasible for an instance of the cophenetic median tree problem involving less than a hundred of species. In our work we move the feasibility bound for the local search heuristic much farther by designing an efficient local neighborhood search algorithm. The algorithm results from (i) an extensive analysis of properties of the cophenetic vectors and (ii) the adapted preprocessing part of the recently developed algorithmic framework targeting the path-difference median tree heuristics. With the efficient heuristic at hand we were able to compute the first large-scale cophenetic median tree estimates and evaluate them against supertrees constructed by other standard supertree methods on benchmark empirical datasets.

Related work. The family of cophenetic metrics has been first extensively studied by Cardona et.al. [9]. Their work, on one hand, addresses the minimum and maximum values as well as distributions of the cophenetic metrics; on the other hand, compares the cophenetic metrics to other popular comparative metrics, such as path-difference distances and the Robinson-Foulds metric. The best known algorithm for computing the cophenetic distance between any two phylogenetic trees of size n requires $O(n^2)$ time.

There has been a large body of work focusing on the biological, mathematical, and algorithmic properties of median trees adopting various definitions of distance measures that have been effectively used in comparative phylogenetics [5]. As most of the studied median tree problems are NP-hard, the classic median tree estimation algorithms, including MRP, effectively employ the local search (hill-climbing) heuristics [2, 10, 22, 24, 39], which have provided credible estimates of large-scale species trees [24, 39]. Local search heuristics typically operate in a search space of all existing supertrees (candidate median trees) for the given collection of input trees. The search starts with a supertree called a *seed* and it maintains the candidate median tree updating it on each iteration. An iteration encompasses a search of a tree with the minimum distance to the input trees in the neighborhood of the current candidate tree. The neighborhood is typically defined in terms of a tree edit operation of choice. One of the most popular such operations is called *subtree prune and regraft (SPR)*, whose respective neighborhood contains $\Theta(n^2)$ trees, where n is the size of a median tree. Consequently, the SPR-neighborhood search problem under the Manhattan cophenetic metric can be solved naïvely in $\Theta(kn^4)$ time, where k is the number of input trees.

The effectiveness of standard local search heuristics is typically highly dependent on the choice of the starting tree. Traditionally, greedy heuristics are employed to efficiently construct well-fitting starting trees; that is, trees that consistently have a significantly smaller distance to the input trees than a randomly chosen tree. An extension of this approach, a *hybrid heuristic*, was introduced as a method to improve the power of the traditional approach by applying the local search heuristic in multiple stages of constructing the starting tree itself [25].

Our contribution. The focus of this work is the development and the applicability study of effective heuristics estimating cophenetic median trees. The design of a heuristic is a necessary component, as we show that the *cophenetic median tree problem* under any vector norm is NP-hard.

Focusing on the Manhattan norm, we reveal that Manhattan cophenetic median trees can be successfully estimated via the local search approach. To make this possible, we propose an efficient algorithm that exploits the properties of the cophenetic vectors and enables the heuristical computation of Manhattan cophenetic median trees for sufficiently large phylogenetic datasets relating hundreds of species. The latter is demonstrated in our first experimental study that explores the runtime of the local search heuristic. The proposed algorithm solves the local SPR-neighborhood search problem in $O(kn^3)$ improving on the naïve runtime by a factor of n.

Further, in our experiments, we use the developed algorithm as a core of the state-of-the-art hybrid heuristic to evaluate cophenetic median trees on published empirical datasets. We evaluate the obtained cophenetic median trees against supertrees constructed by several other classic supertree methods. Finally, motivated by the previous work and our results, we study the correlation of the Manhattan cophenetic metric to the classic model-based cost functions including deep coalescence, duplications, and duplications with losses.

1 BASICS AND PRELIMINARIES

Basic definitions. Throughout this paper we adhere to the definitions and notation introduced in [25]. A *(phylogenetic) tree* T is a rooted binary tree, where each leaf is uniquely labeled with a taxon, each internal node v has exactly two children nodes, denoted by $\mathrm{Ch}_T(v)$, and each node u except for the root has a single parent node denoted by $\mathrm{Pa}_T(u)$. In addition, we denote the node set, edge set and leaf set of T by $V(T)$, $E(T)$ and $\mathsf{L}(T)$ respectively. We denote the root by $\mathrm{Rt}(T)$ and a sibling of each non-root node u by $\mathrm{Sb}(u)$. We also set $T(v)$ to be a subtree of T rooted at $v \in V(T)$, and $\overline{T(v)}$ to be a tree obtained by pruning $T(v)$ from T. Occasionally we use the standard nested parenthesis notation to represent small trees.

We define a partial order $\preceq_T$ on the vertex set $V(T)$, such that $u \preceq v$, if v is a node on the path from u to $\mathrm{Rt}(T)$. Additionally, we say $u \prec v$, if $u \preceq v$ and $u \neq v$. The *least common ancestor (LCA)* of

two nodes $u, v \in V(T)$, $\mathsf{LCA}_T(u, v)$, is the furthest from the root node, w, such that $v \preceq w$ and $u \preceq w$.

A set of leaves $\mathsf{L}(T(v))$ is called a *cluster* of the node v, and is denoted by C_v. Note that for convenience we identify the leaves in a phylogenetic tree with the respective labels (taxa).

Let $L \subseteq \mathsf{L}(T)$ and T' be the minimal subtree of T with leaf set L. We define the *leaf-induced subtree* $T[L]$ of T to be the tree obtained from T' by successively removing each node of degree two (except for the root) and adjoining its two neighbors (a parent and a child).

Let $\mathcal{P}$ be a set of phylogenetic trees $\{G_1, \ldots, G_k\}$. We extend the definition of a leaf set to a set of trees as follows: $\mathsf{L}(\mathcal{P}) := \cup_{i=1}^{k} \mathsf{L}(G_i)$. A tree S is called a *supertree* of $\mathcal{P}$, if $\mathsf{L}(S) = \mathsf{L}(\mathcal{P})$. A set of trees $\mathcal{P}$ is called *compatible* if there exist a supertree T consistent with every tree in $\mathcal{P}$, and a tree T is *consistent* with a tree G if $T[\mathsf{L}(G)] \equiv G$ up to isomorphism of rooted semi-labeled trees [33].

Cophenetic distance. In this section we follow the definitions presented in [9]. Given a phylogenetic tree T, let the *cophenetic value* of $u, v \in V(T)$, denoted by $\phi_{u,v}(T)$, be the length (measured in the number of edges) of the path from $\mathsf{LCA}_T(u, v)$ to $\mathsf{Rt}(T)$. Additionally, $\delta_u := \phi_{u,u}(T)$ is the *depth* of the node u. Given that, the *cophenetic vector* of T is

$$\phi(T) = (\phi_{i,j}(T))_{1 \le i \le j \le |\mathsf{L}(T)|},$$

for some fixed ordering of leaves in T. The *cophenetic distance* between two trees G and S over the same leaf set is defined as

$$\mathsf{d}_{\phi,p}(G, S) := ||\phi(G) - \phi(S)||_p,$$

where $|| \cdot ||_p$ denotes an L_p norm of a vector for some fixed $p \in [1, \infty)$.

Next, let $\Phi(G, S)$ be the *cophenetic difference matrix* of size $|\mathsf{L}(G)| \times |\mathsf{L}(G)|$, such that for all $i, j \in \mathsf{L}(G), \Phi_{i,j}(G, S) = \phi_{i,j}(G) - \phi_{i,j}(S)$. Note that $\mathsf{L}(G)$ should be equivalent to $\mathsf{L}(S)$.

We further extend the definition of cophenetic distance to a set of trees. Given a set of trees $\mathcal{P}$ and a supertree of $\mathcal{P}$, S, the cophenetic distance between S and $\mathcal{P}$ is

$$\mathsf{d}_{\phi,p}(\mathcal{P}, S) := \sum_{G \in \mathcal{P}} \mathsf{d}_{\phi,p}(G, S[\mathsf{L}(G)]).$$

On practice the input trees are of different sizes and might be involved this different sets of taxa. Consequently, the supertrees are typically larger than the input trees. Note however that we defined the cophenetic distance only for trees over the same taxon set. Hence, in the above equation we use the *minus method* [13] to account for the cophenetic distance between the supertree S and an input tree G (that is, we prune the extra information from S when comparing to G).

Path-induced subtrees. Let $P = (u_0, u_1, \ldots, u_k)$ be a simple path in a tree T, then *i-th exit node*, denoted by $e_i(P)$, for $0 < i < k$ is defined as follows:

(a) **If** $u_{i-1} \prec u_i \prec u_{i+1}$, then $e_i := \mathsf{Sb}_T(u_{i-1})$; Figure 1 depicts an example for this case;
(b) **If** $u_{i+1} \prec u_i \prec u_{i-1}$, then $e_i := \mathsf{Sb}_T(u_{i+1})$.

Additionally, if $u_{k-1} \prec u_k$, then $e_k := \mathsf{Sb}_T(u_{k-1})$; otherwise, $e_k := u_k$. For brevity, $E_i(P) := C_{e_i(P)}$ (i-th *exit cluster*), or simply E_i when the path P can be inferred from the context. Figure 2 illustrates the exit nodes induced by two different types of paths.

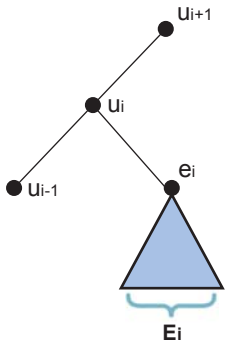

Figure 1: An example of an exit-node and the corresponding subtree

2 COPHENETIC MEDIAN TREE PROBLEMS

In this section we formulate and explore the basic properties of the median tree problems defined for cophenetic distances under different vector norms. Based on the definitions presented in the previous section we introduce a class of *cophenetic median tree problems* as follows:

Problem 2.1 (Cophenetic median tree (for an L_p norm) – decision version).
Instance: A set of input trees $\mathcal{P}$ and a real number q;
Question: Determine whether there exists a supertree S, such that $\mathsf{d}_{\phi,p}(\mathcal{P}, S) \le q$.

2.1 Cophenetic median tree problems are NP-hard

We prove that Problem 2.1 is NP-hard under any L_p vector norm. The proof is based on the NP-hardness proof of the related path-difference median tree problem [26]. To show NP-hardness we provide a reduction from the NP-complete MaxRTC [7].

Problem 2.2 (Maximum Compatible Subset of Rooted Triplets – MaxRTC).
Instance: A set of rooted triplets $\mathcal{R}$ and an integer $0 \le c \le |\mathcal{R}|$;
Question: Is there a subset $\mathcal{R}' \subseteq \mathcal{R}$, such that $\mathcal{R}'$ is compatible and $|\mathcal{R}'| \ge c$.

Theorem 2.1. *The cophenetic median tree problem under an L_p norm is NP-hard for any $p \ge 1$.*

Proof. Consider a rooted triplet R and a tree S, such that $\mathsf{L}(R) \subseteq \mathsf{L}(S)$. Note that if S is *consistent* with R, then $\mathsf{d}_{\phi,p}(R, S[\mathsf{L}(R)]) = 0$; otherwise, $\mathsf{d}_{\phi,p}(R, S[\mathsf{L}(R)]) = 4^{\frac{1}{p}}$. The latter relationship can be easily verified by, for example, computing a cophenetic distance between two incompatible triplets $((a, b), c)$ and $((a, c), b)$.

Given that $\mathsf{d}_{\phi,p}(R, S[\mathsf{L}(R)])$ is a constant that depends only on whether S is consistent with R or not, we can map an instance $\langle \mathcal{R}, c \rangle$ of the MaxRTC problem to an instance $\langle \mathcal{R}, (|\mathcal{R}| - c)4^{\frac{1}{p}} \rangle$ of the cophenetic median tree problem. It is not difficult to observe that $\langle \mathcal{R}, (|\mathcal{R}| - c)4^{\frac{1}{p}} \rangle$ is a *yes*-instance if and only if $\langle \mathcal{R}, c \rangle$ is a *yes*-instance of MaxRTC. □

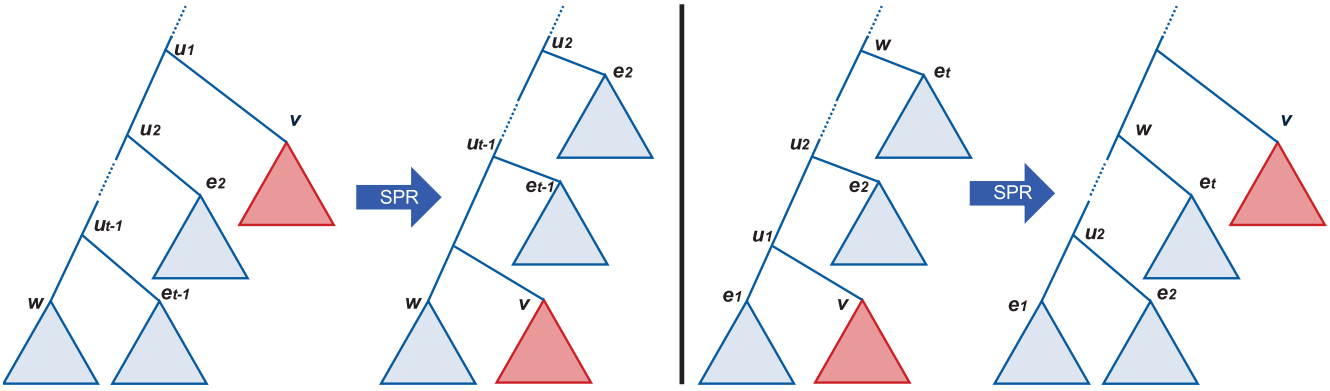

Figure 2: (left) shows how the $SPR(v, w)$ operation changes the tree structure, when w is an descendant of $\mathrm{Sb}(v)$; (right) shows how the $SPR(v, w)$ operation changes the tree, when w is an ancestor of v and $\mathrm{Sb}(v)$.

3 LOCAL SEARCH FOR THE COPHENETIC MEDIAN TREE PROBLEM

In this section we describe an efficient algorithm for approximating cophenetic median trees using a standard local search approach under the classic SPR tree edit operation.

3.1 *SPR*-Local search framework

Given a node $v \in V(S)\backslash\{\mathrm{Rt}(S)\}$, and a node $u \in V(\overline{S(v)})$, $SPR_S(v, u)$ is a tree obtained by the following modifications of the tree $S' = \overline{S(v)}$:

(1) If u is a root of S', then a new root w' is introduced, so that u is a child of w'. Otherwise, an edge $(\mathrm{Pa}(u), u)$ is subdivided by a new node w'.
(2) Connect the pruned subtree $S(v)$ to the node w'.

Further, we define the following sets of trees that can be obtained from S by performing SPR:

$$SPR_S(v) := \bigcup_u SPR_S(v, u); \qquad SPR_S := \bigcup_{v,u} SPR_S(v, u).$$

SPR_S is called an *SPR-neighborhood* of a tree S. It is easy to see from the definition that $|SPR_S| = O(n^2)$, where $n = |\mathsf{L}(S)|$.

Given a set of input trees $\mathcal{P} = \{G_1, ..., G_k\}$, the search space in an SPR local search problem could be viewed as a graph $\mathcal{T}$, where nodes represent all existing supertrees (candidate median trees) of $\mathcal{P}$. $\{S_1, S_2\}$ is an edge in $\mathcal{T}$, if S_1 could be transformed to S_2 with a single SPR operation.

At each iteration local search heuristic finds a candidate tree S' in the neighborhood of a current tree S, such that S' minimizes the cost function that we are interested in. In case $S \equiv S'$, the local search stops (reaches a local minimum). Otherwise, it proceeds to the next iteration with a tree S'. An instance (single iteration) of the SPR-based local neighborhood search problem could be formalized as follows:

Problem 3.1 (Cophenetic metric local neighborhood search).
Instance: An input set $\mathcal{P}$ and a candidate tree (supertree) S;
Question: Find a tree $S' = \underset{S' \in SPR_S}{\arg\min}\, \mathrm{d}_{\phi,p}(\mathcal{P}, S')$.

Naïve algorithm for the local search problems. Given two trees S and G, one can compute $\mathrm{d}_{\phi,p}(S, G)$ in $O(n^2)$ time for any p. Therefore, direct computation of the $\phi_p(\mathcal{P}, S')$ score for each $S' \in SPR_S$ would take $O(n^4 k)$ time, where $n = |\mathsf{L}(\mathcal{P})|$ and $k = |\mathcal{P}|$. Next, we show how to improve on this complexity under $p = 1$ (the *Manhattan* distance).

To fix the set up, let $\mathrm{G} \in \mathcal{P}$ be a fixed input tree, and let S_i be a supertree in the i-th iteration of the local search. Throughout the next section we refer to the restricted tree $S_i[\mathsf{L}(G)]$ as simply S.

3.2 Local search environment for cophenetic median trees

To design a faster algorithm for the cophenetic local search problem we examine the structure of the SPR-environment of a candidate median tree. We are interested in the structure of the cophenetic difference matrix $\Phi(T, S)$ for some $T = SPR_S(v, w)$. Let $U_T = (v = u_0, \ldots, u_t = w)$ be the path between v and w in T, and let u_h be the node closest to the root of S on that path (i.e., $u_h \succ u_i$ for all $0 \le i \le t, i \ne h$). Below we show the structure of the matrix $\Phi(S, T)$ by considering a few major cases that provably cover the whole matrix.

(i) $\mathbf{Sb}(\boldsymbol{v}) \prec \boldsymbol{w}$, i.e., the path U_T is a part of the path from v to $\mathrm{Rt}(S)$. This case is depicted in Figure 2 (right-hand side).
  (a) $\forall \boldsymbol{i} \in \boldsymbol{C_v}, \boldsymbol{j} \in \boldsymbol{E_p} : \phi_{i,j}(T) = \phi_{i,j}(S) - (t - p)$ for $1 \le p \le t - 1$. This case characterizes the change in depths of LCAs between leaves in C_v and leaves in exit clusters of the path U_T. Note that while $\mathrm{LCA}_S(i, j) = u_p$, after regrafting we have $\mathrm{LCA}_T(i, j) = \mathrm{Pa}_T(w) = \mathrm{Pa}_T(v)$.
  (b) $\forall \boldsymbol{i} \in \boldsymbol{C_w}, \boldsymbol{j} \in \boldsymbol{C_w}\backslash\boldsymbol{E_1} : \phi_{i,j}(T) = \phi_{i,j}(S) + 1$. This change is due to the fact that we add a new node, $\mathrm{Pa}_T(w)$, on the path from the nodes u_p (for $2 \le p \le t$) to the root. Note, however, that the depth of the node e_1 remains unchanged.
(ii) $\mathbf{Sb}_S(\boldsymbol{v}) \succeq \boldsymbol{w}$. This case is similar to the one considered above and it is depicted on the left-hand side of Figure 2.

(a) $\forall i \in C_v, j \in E_p : \phi_{i,j}(T) = \phi_{i,j}(S) + (p-2)$ for $2 \leq p \leq t$. Due to the observation that $\mathsf{LCA}_S(i,j) = u_1$, while after regrafting $\mathsf{LCA}_T(i,j) = \mathsf{Pa}_T(e_p)$.

(b) $\forall i \in C_{u_2}, j \in C_{u_2} \backslash E_t : \phi_{i,j}(T) = \phi_{i,j}(S) - 1$. We removed the node, $\mathsf{Pa}_S(v)$, from the paths from the nodes u_p (for $2 \leq p \leq t-1$) to the root. However, the depth of the node $e_t = w$ remains unchanged.

(iii) $\mathbf{Sb}(v) \not\prec w, \mathbf{Sb}(v) \not\succeq w$. That is, u_h is some node on the path, which is neither w nor v.

(a) $\forall i \in E_1, j \in E_1 : \phi_{i,j}(T) = \phi_{i,j}(S) - 1$. Node e_1 becomes one edge closer to the root.

(b) $\forall i \in C_w, j \in C_w : \phi_{i,j}(T) = \phi_{i,j}(S) + 1$. Node w becomes one edge further from the root.

(c) $\forall i \in C_v, j \in E_p : \phi_{i,j}(T) = \phi_{i,j}(S) + (p-h)$ for $1 \leq p \leq t, p \neq h$. For $p < h$, we have $\mathsf{LCA}_S(i,j) = u_p$ and $\mathsf{LCA}_T(i,j) = u_h$; as for $p > h$, we have $\mathsf{LCA}_S(i,j) = u_h$ and $\mathsf{LCA}_T(i,j) = \mathsf{Pa}_T(e_p)$.

(iv) For all three cases outlined above: $\forall i, j \in C_v : \phi_{i,j}(T) = \phi_{i,j}(S) - \delta_v(S) + \delta_w(S) + 1$. Since we regraft the subtree $S(v)$ above w, depths of all nodes inside this subtree increase by $(-\delta_v(S) + \delta_w(S) + 1)$; hence, the change in the cophenetic vector.

(v) Observe that for any other choice of i and j, the corresponding cophenetic value, $\phi_{i,j}$, is not affected.

Overall, the following clusters are involved in the changes in the cophenetic vector of S: i is one of the clusters in $\mathscr{C}_i = \{C_v, C_w, C_{u_2}, E_1\}$, and j appears in $\mathscr{C}_j = \{C_v, C_w, C_{u_2}, E_1, \ldots, E_t\}$. The key observation is that regardless of the form of U_T, there are only $O(t)$ pairs of clusters $(C_i, C_j) \in \mathscr{C}_i \times \mathscr{C}_j$ such that the respective cophenetic values are altered.

Let $d(C_i, C_j)$ be the value, such that $\forall (i,j) \in C_i \times C_j,\ \phi_{i,j}(T) = \phi_{i,j}(S) + d(C_i, C_j)$ according to the cases outlined above. For example, $d(C_v, C_v) = -\delta_v(S) + \delta_w(S) + 1$. Given that, Equation 1 shows how the cophenetic distance $\phi(S,G)$ is effected by an SPR operation. The equation has been adopted from the work on Manhattan path-difference median trees [26]. For technical reasons we define

$$\Delta_{i,j}(C_1, C_2) := \begin{cases} 0, & \text{if } i > j \text{ and } (j,i) \in C_1 \times C_2 \\ \Phi_{i,j}(S,G), & \text{otherwise.} \end{cases}$$

In most cases the fixed clusters C_1 and C_2 are clear from the context, and hence we use the shorthand notation, d and $\Delta_{i,j}$ for $d(C_1, C_2)$ and $\Delta_{i,j}(C_1, C_2)$ respectively. Below we provide the final equation.

$$\mathsf{d}_{\phi,1}(T,G) - \mathsf{d}_{\phi,1}(S,G) = \sum_{\substack{C_1 \in \mathscr{C}_i \\ C_2 \in \mathscr{C}_j}} \sum_{\substack{i \in C_1 \\ j \in C_2}} \left(|\Delta_{i,j} + d| - |\Delta_{i,j}| \right)$$

$$= \sum_{\substack{C_1 \in \mathscr{C}_i \\ C_2 \in \mathscr{C}_j}} \begin{pmatrix} d \cdot \#\{(i \in C_1, j \in C_2) | \Delta_{i,j} \geq -d)\} \\ -d \cdot \#\{(i \in C_1, j \in C_2) | \Delta_{i,j} < -d)\} \\ +2 \sum\limits_{\substack{i \in C_1, j \in C_2: \\ -d \leq \Delta_{i,j} < 0}} \Delta_{i,j} \ - \ 2 \sum\limits_{\substack{i \in C_1, j \in C_2: \\ 0 \leq \Delta_{i,j} < -d}} \Delta_{i,j} \end{pmatrix}. \quad (1)$$

3.3 Efficient algorithm

To make use of the above analysis we adopt the data structures developed in the work on Manhattan path-difference median trees [26]. In this section we briefly describe the resulting preprocessing idea.

Let $\mathscr{C}(S)$ be a set of all clusters in a tree S. For $\mathsf{L}_1, \mathsf{L}_2 \in \mathscr{C}(S)$, we define $\Sigma_{\geq}(\mathsf{L}_1, \mathsf{L}_2)$ to be a vector indexed from $-n$ to n, such that

$$\Sigma_{\geq}(\mathsf{L}_1, \mathsf{L}_2)[x] = \Sigma_{\geq}(\mathsf{L}_2, \mathsf{L}_1)[x] = \sum_{\substack{i \in \mathsf{L}_1, j \in \mathsf{L}_2: \\ i \leq j,\ \Delta_{i,j} \geq x}} \Delta_{i,j}(L_i, L_j)$$

Similarly, we define a vector $\#_{\geq}(\mathsf{L}_1, \mathsf{L}_2)$ indexed from $-n$ to n, such that

$$\#_{\geq}(\mathsf{L}_1, \mathsf{L}_2)[x] = \#\{(i \in \mathsf{L}_1, j \in \mathsf{L}_2 - \mathsf{L}_1) | i \leq j, \Delta_{i,j}(L_1, L_2) \geq x\}$$

It is not difficult to check that given such vectors it is possible to compute $\mathsf{d}_{\phi,1}(T,G) - \mathsf{d}_{\phi,1}(S,G)$ for an arbitrary $T \in SPR_S$ in $O(n)$ time using Equation 1. For example, let's consider some $T = SPR_S(v,w)$, such that the corresponding path U_T is a part of a path from v to the root (i.e., $v \prec w$). That means that if we choose $C_1 = C_v$ and $C_2 = E_1$, then $d(C_1, C_2) = -t + 1$ according to the analysis presented in Section 3.2, where $t \geq 2$ is the length of U_T in edges. We can now use the vectors defined above to find a part of the sum in Equation 1 that corresponds to the chosen clusters C_1 and C_2. That is, we observe the following relations:

- $\#\{(i \in C_1, j \in C_2) | \Delta_{i,j} \geq t-1\}$ is simply $\#_{\geq}(C_1, C_2)[t-1]$.
- $\#\{(i \in C_1, j \in C_2) | \Delta_{i,j} < t-1\} = \big(\#_{\geq}(C_1, C_2)[-n] - \#_{\geq}(C_1, C_2)]t-1]\big)$.
- $\sum\limits_{\substack{i \in C_1, j \in C_2: \\ 0 \leq \Delta_{i,j} < t-1}} \Delta_{i,j} = \Sigma_{\geq}(C_1, C_2)[0] - \Sigma_{\geq}(C_1, C_2)[t-1]$.

Complexity analysis. An algorithm for computing vectors $\Sigma_{\geq}$ and $\#_{\geq}$ efficiently was presented in [26]. Although there the vectors were defined in a slightly different way, the algorithm can be adopted for our needs (we omit the technical details for brevity).

The time complexity for computing these vectors is $O(n^3)$ for a fixed input tree G. Having these vectors computed we can calculate the value $\mathsf{d}_{\phi,1}(T,G)$ in $O(n)$ for all $T \in SPR_S$. Given that $|SPR_S| = O(n^2)$, and the number of input trees is k, the overall time complexity is $O(kn^3)$.

4 EXPERIMENTAL EVALUATION

Here we evaluate the local search approach enabled by the algorithm described in the previous section as applied to the Manhattan cophenetic median tree problem. In our first study we evaluate the local search approach in terms of time that it takes to converge to a local minimum on artificially constructed datasets of different sizes. In the second study we apply the state-of-the-art local search heuristic to sufficiently large empirical datasets and compare obtained cophenetic median trees to supertrees obtained by using other popular supertree and median tree methods.

4.1 Scalability analysis

We compare the runtime of pure local search strategies (when a starting tree is chosen randomly) for the estimation of Manhattan median trees using the naïve and improved algorithms described in Sections 3.1 and 3.3 respectively.

Data sets. We estimate the runtime for randomly generated sets of input trees. We generated 12 random input sets with 10 trees in each over the number of taxa varying from 10 in the smallest dataset to 120 in the largest one, with a step of 10.

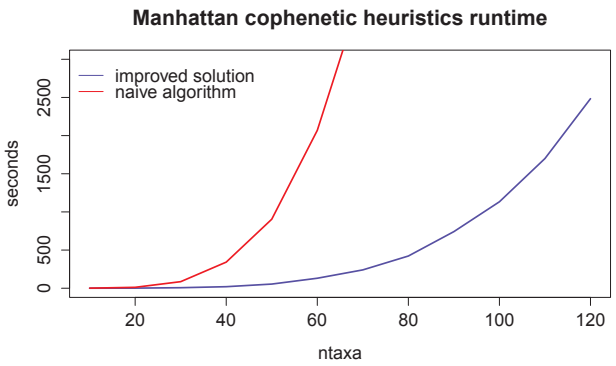

Figure 3: The growth of runtime until convergence with a gradual increase of the number of taxa in input datasets. Mean runtimes among five trials are presented.

Experimental setting. For each of the generated datasets we ran both the naïve and improved heuristics five times each. In Figure 3 we report the mean runtimes over the five trials.

Results. Figure 3 depicts that the runtime for the improved local search algorithm grows *substantially slower* than the runtime under the naïve algorithm. Without the improved algorithm the taxa limit for local search heuristic is reached already for 70 taxa. However, the presented here analysis and the resulting algorithm allows us to estimate cophenetic median trees for much larger instances, which we demonstrate in our second study.

4.2 Empirical study

In this section we compare heuristically estimated Manhattan cophenetic median trees to path-difference median trees and other supertrees constructed by using several highly recognized supertree methods. The study is conducted over two standard empirical datasets. The goal is to evaluate the applicability of the local search approach to the cophenetic median tree problem and to find relations between various supertree and median tree methods.

Data sets. Following the original work on path-difference median trees, we evaluate our cophenetic median tree heuristic on published baseline datasets [25, 26]
(i) *Cetartiodactyla*: contains 201 trees over 299 taxa overall [30];
(ii) *Marsupials*: contains 158 trees over 272 taxa overall [8].
These datasets are considerably large and serve as benchmarks for phylogenetic studies (see [3, 11, 21, 34]).

Methods. In order to obtain credible estimates for Manhattan cophenetic median trees (MCMT) we used the *hybrid heuristic* framework that was successfully applied to path-difference median trees outperforming other standard local search paradigms [26]. For comparison we use two path-difference median tree heuristics – one for Manhattan median trees (MMT) and another for Euclidean median trees (EMT).

Additionally, following the preceding studies [21, 25, 26], we include the following methods in our study: *modified min-cut* (MMC) algorithm that computes supertrees in polynomial time and was suggested for use on large-scale datasets [29]; two *triplet median tree heuristics* (TH) that approach the corresponding triplet median tree problem using SPR and TBR tree edit operations respectively [21]. Note that TBR (stands for tree bisection and reconnection) is an extension of the SPR operation, where the pruned subtree is allowed to be re-rooted before regrafting it. Finally, we include the classic *maximum parsimony with representation heuristic* (MRP). MRP was recognized as the most applied supertree method among practitioners [5]. Here we use the implementation of MRP heuristic in the popular software package, PAUP* [38], under the TBR branch swapping [21].

Experimental setting. To compare the methods under consideration we used the results of their execution over both datasets (each method was executed 10 times, except for MMC, which is a deterministic method). We further evaluated each of the generated supertrees with the respective input dataset using 6 relevant objectives: the Manhattan cophenetic distance, the Manhattan and Euclidean path-difference distances, triplet similarity (the objective function for triplet heuristics), the average maximum agreement subtree (MAST) similarity, and the parsimony score.

The best scores among the ten trials under each objective are presented in Table 1

Results. From Table 1 we observe that our objective function, the Manhattan cophenetic distance, significantly differs from others in terms of the distribution of scores across the methods. That is, as expected, the here introduced method, MCMT, performs best in terms of this objective. We also observe that the MRP and TH heuristics produce trees that are better in terms of cophenetic distance than the trees produced by path-difference median tree methods. On the other hand, MCMT produces trees that score better in terms of path-difference objectives than trees generated by MRP and TH heuristics. That is, we observe a rather asymmetric behavior for the three vector-based objectives and their respective median tree estimates (MMT, EMT, and MCMT).

4.3 Correlation with other cost functions

Cardona et.al. studied correlations between the cophenetic metrics and other popular tree comparison metrics including the path-difference distance and the classic Robinson-Foulds metric (RF) [9]. They have demonstrated that the Manhattan cophenetic distance (i) does not bare strong correlation with the Manhattan and Euclidean path-difference distances (Spearman correlation coefficient of ≈ 0.45), and (ii) has almost no significant correlation with the Robinson-Foulds distance (Spearman correlation coefficient of approximately ≈ -0.0008). As was mentioned in the introduction, the cophenetic distance is dependent on the LCA mappings, and therefore, can be expected to be more closely related to the cost functions originating from the gene tree parsimony (GTP) problem than the path-difference distance or RF. We consider the following GTP related cost functions: gene duplications (GD), deep coalescence (DC), and duplications with losses (DL). In fact, based on the formal definitions, the cophenetic distance is most closely related to the deep coalescence cost function, as both take into account the path-lengths between LCA mappings. In this section we test our hypothesis that the two cost functions are indeed correlated.

Experimental setting. In order to assess correlations between different cost function we follow the Cardona et.al. setting. That is,

Data set	Method	L_1 cophen.	L_1 PDD	L_2 PDD	Triplet-sim	MAST	Pars. score
Marsup 158 input trees 272 taxa	MMC	1,564,728	1,681,015	16,670.45	51.73 %	53.4 %	3901
	MRP	122,459	515,257	5,694.59	98.29 %	**71.6** %	**2274**
	TH(SPR)	143,398	515,906	5,866.27	**98.99** %	70.3 %	2312
	TH(TBR)	143,501	517,274	5,888.22	**98.99** %	70.4 %	2317
	EMT	260,787	327,379	**4,380.77**	85.24 %	67.0 %	2869
	MMT	286,357	**323,909**	5,063.34	54.68 %	57.6 %	3817
	MCMT	**60,737**	372,719	4,974.58	90.90 %	64.8 %	3036
Cetartio 201 input trees 299 taxa	MMC	1,004,359	918,639	16,206.17	70.03 %	51.5 %	4929
	MRP	186,582	365,870	6,991.36	96.49 %	**65.2** %	**2603**
	TH(SPR)	168,620	403,233	7,630.03	**97.28** %	63.1 %	2754
	TH(TBR)	168,497	401,327	7,591.13	**97.28** %	63.0 %	2754
	EMT	209,680	258,836	**5,639.24**	85.98 %	61.0 %	3394
	MMT	225,705	**258,424**	6,142.98	66.28 %	54.2 %	4218
	MCMT	**74,135**	288,411	6,620.88	87.80 %	58.1 %	3895

Table 1: Empricial evaluation of supertree methods over two published phylogenetic datasets. The best scores under each objective function are shown in bold.

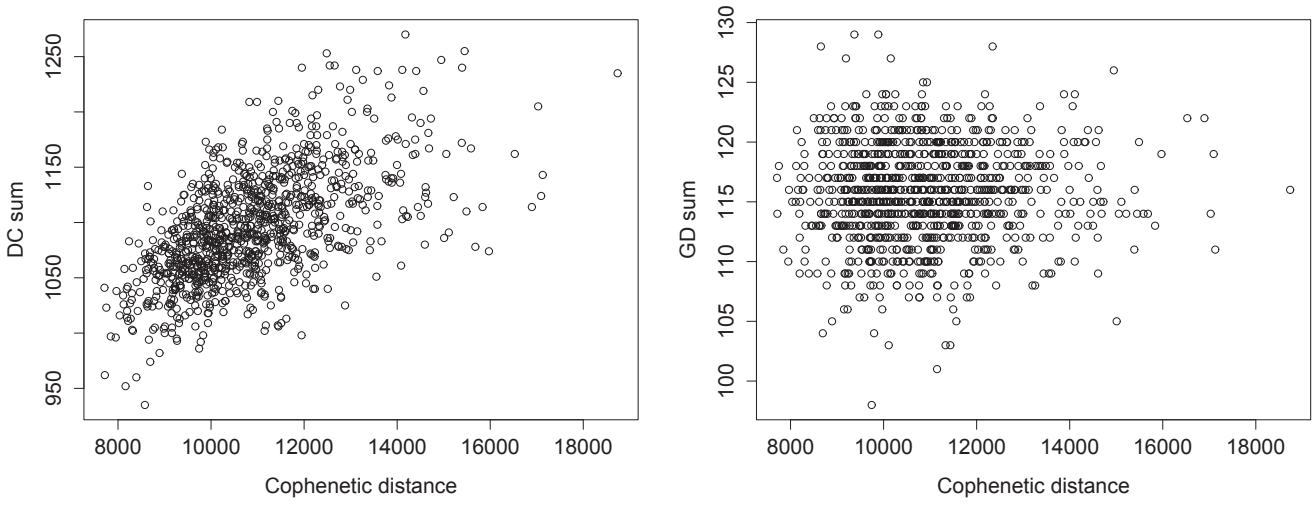

Figure 4: The figures depict the cophenetic distances between pairs of trees in comparison to the respective DC cost (left-hand side) and the GD cost (right-hand side).

we generated 1000 random pairs of bifurcating phylogenetic trees with 100 of labeled leaves each. The trees were drawn from the uniform distribution. Next, for each pair of trees, (T_1, T_2), we computed the Manhattan cophenetic distance as well as duplications, deep coalescence, and duplications with losses costs. Observe that the GD, DC, and DL cost functions are not symmetric. Thus, in order to compare it to the symmetric cophenetic distance, we computed the cost sums $C(T_1, T_2) + C(T_2, T_1)$, where $C \in \{GD, DC, DL\}$.

Based on the obtained scores for a thousand of tree pairs we computed Spearman correlation coefficients for each pair of cost functions (see Table 2). The data was also plotted on Figure 4 to emphasize the correlation patterns.

	DC sum	GD sum	DL sum
L_1 cophenetic	0.602	0.023	0.544
DC sum	-	0.331	0.975
GD sum	-	-	0.517

Table 2: Spearman correlation coefficients for the four LCA-based cost functions.

Results. Observe from Table 2 that the two most correlated cost functions are deep coalescence and duplications with losses. This was highly expected, since the DL cost function can be represented as a linear combination of DC and GD, where the DC cost is much

more significant than the GD cost [41]. Further, in justification to our hypothesis, we observe a significant correlation between the Manhattan cophenetic distance and DC. Observe that the Spearman correlation for these two cost functions is higher than the respective correlation coefficient between the L_1 cophenetic distance and the path-difference distances. The observed correlation is further illustrated in Figure 4 (left-hand side) as opposed to the low correlation plot for the GD cost function (on the right-hand side).

5 CONCLUSION AND OUTLOOK

The problem of discordance in phylogenetic trees has been addressed by the means of the median tree approach for over 20 years [4, 5, 31]. Median tree methods employ various objective cost functions, which can be classified into mathematically informed costs, and biologically informed costs. Biological costs are based on evolutionary processes causing discordance between two trees (e.g., deep coalescences, gene duplications, and gene duplications with losses [14]), which are typically used when gene trees are compared with species trees [14]. In contrast, independent of any evolutionary causes, mathematical costs between two trees are measuring the amount of elementary evolutionary information that is common (or different) in these trees, and are thought to be applicable as a tool of error-correction and formal maximization of common information among the input trees [5]. Another distinction between mathematical and biological costs is that the former typically satisfy the properties of a metric [36], while the latter once are not symmetric and do not satisfy the triangle inequality [14]. Despite the differences between mathematical and biological costs, in the median tree setting the Manhattan cophenetic metric, a mathematical cost, and the deep coalescence cost function, a classic biological cost, showed to be strongly correlated in our experiments. This suggests that the cophenetic median tree methods may be used as a universal solution to the generalized supertree problem. Note also that as a metric, the cophenetic model provides valuable mathematical properties, which are not met by most of the biologically informed cost functions. The fact that the cophenetic metrics are not tied to a biological model also gives an advantage, as they can be naturally generalized for the comparison of weighted phylogenetic trees. Given the increasing interest in the time-annotated median trees [20], this property provides great applicability perspectives for the weighted cophenetic median trees.

In this work we presented a new method for median tree estimation based on the Manhattan cophenetic metric, and studied the correlations of this metric with other classic cost functions that have been applied and recognized in the context of median trees.

We devised an effective heuristic that enabled the computation of the first Manhattan cophenetic median tree estimates on sufficiently large phylogenetic datasets, which became only possible due to the efficient algorithm for the local neighborhood search that we put forward in this work. This algorithmic advancement allowed us to perform a first applicability study that evaluates the Manhattan cophenetic median tree heuristic against other supertree methods on benchmark datasets. The results of this study motivates much broader future investigations into the significance of cophenetic median trees from the evolutionary perspective.

Furthermore, the special properties of the cophenetic metrics, such as dependence on the LCAs, also indicate that it is possible to devise an efficient local search heuristic under the TBR edit operation [1, 12, 37], which will be feasible on large-scale datasets and complement our current work.

6 ACKNOWLEDGMENTS

We thank the four anonymous reviewers for their constructive comments. This material is based upon work supported by the National Science Foundation under Grant No. 1617626.

REFERENCES

[1] Benjamin L. Allen and Mike Steel. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics* 5 (2001), 1–15.
[2] Mukul S. Bansal, J. Gordon Burleigh, and Oliver Eulenstein. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC Bioinformatics* 11 Suppl 1 (2010), S42. https://doi.org/10.1186/1471-2105-11-S1-S42
[3] Mukul S. Bansal, J. Gordon Burleigh, Oliver Eulenstein, and David Fernández-Baca. 2010. Robinson-Foulds Supertrees. *Algorithms for Molecular Biology* 5, 1 (2010), 1–12. https://doi.org/10.1186/1748-7188-5-18
[4] Bernard R. Baum. 1992. Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. *Taxon* 41, 1 (1992), 3–10.
[5] Olaf R.P. Bininda-Emonds (Ed.). 2004. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life.* Computational Biology, Vol. 4. Springer Verlag.
[6] Paola Bonizzoni, Gianluca Della Vedova, and Riccardo Dondi. 2005. Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical computer science* 347, 1-2 (2005), 36–53.
[7] David Bryant. 1997. *Hunting for trees, building trees andcomparing trees: theory and method in phylogeneticanalysis.* Ph.D. Dissertation. Department of Mathematics, University of Canterbury, New Zeeland.
[8] Marcel Cardillo, Olaf R. P. Bininda-Emonds, Elizabeth Boakes, and Andy Purvis. 2004. A species-level phylogenetic supertree of marsupials. *Journal of Zoology* 264 (2004), 11–31. Issue 01. https://doi.org/10.1017/S0952836904005539
[9] Gabriel Cardona, Arnau Mir, Francesc Rosselló, Lucía Rotger, and David Sánchez. 2013. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinformatics* 14, 1 (2013), 3. https://doi.org/10.1186/1471-2105-14-3
[10] Ruchi Chaudhary, Mukul S. Bansal, André Wehe, David Fernández-Baca, and Oliver Eulenstein. 2010. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11 (2010), 574. https://doi.org/10.1186/1471-2105-11-574
[11] Duhong Chen, Oliver Eulenstein, David Fernández-Baca, and JGordon Burleigh. 2006. Improved heuristics for minimum-flip supertree construction. *Evolutionary Bioinformatics* 2 (2006).
[12] Duhong Chen, Oliver Eulenstein, David Fernández-Baca, and J Gordon Burleigh. 2006. Improved heuristics for minimum-flip supertree construction. *Evol Bioinform Online* 2 (2006), 347–56.
[13] James A Cotton and Mark Wilkinson. 2007. Majority-rule supertrees. *Syst Biol* 56, 3 (2007), 445–452. https://doi.org/10.1080/10635150701416682
[14] O. Eulenstein, S. Huzurbazar, and D.A. Liberles. 2010. *Evolution after Gene Duplication.* John Wiley, Chapter Reconciling Phylogenetic Trees.
[15] Peter Forster and Colin Renfrew. 2006. *Phylogenetic methods and the prehistory of languages.* McDonald Inst of Archeological.
[16] Henry Gee. 2003. Evolution: ending incongruence. *Nature* 425, 6960 (Oct 2003), 782. https://doi.org/10.1038/425782a
[17] Simon R. Harris, Edward J.P. Cartwright, M Estéée Török, Matthew T.G. Holden, Nicholas M. Brown, Amanda L. Ogilvy-Stuart, Matthew J. Ellington, Michael A. Quail, Stephen D. Bentley, Julian Parkhill, and Sharon J. Peacock. 2013. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *Lancet Infect Dis* 13, 2 (2013), 130–6. https://doi.org/10.1016/S1473-3099(12)70268-2
[18] Ruth. A. Hufbauer, Robin A. Marrs, Aaron K. Jackson, René Sforza, Harsh Pal Bais, Jorge M. Vivanco, and Shannan E. Carney. 2003. Population structure, ploidy levels and allelopathy of Centaurea maculosa (spotted knapweed) and C. diffusa (diffuse knapweed) in North America and Eurasia. In *Proceedings of the XI International Symposium on Biological Control of Weeds, Canberra Australia.* USDA Forest Service. Forest Health Technology Enterprise Team, Morgantown, WV., 121–126.
[19] Martyn Kennedy, Roderic DM Page, and R Prum. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119, 1 (2002), 88–108.

[20] Adam D. Leaché. 2010. The Timetree of Life. S. Blair Hedges and Sudhir Kumar, editors. *Integrative and Comparative Biology* 50, 1 (2010), 141–142. https://doi.org/10.1093/icb/icp110 arXiv:http://icb.oxfordjournals.org/content/50/1/141.full.pdf+html

[21] Harris T Lin, J Gordon Burleigh, and Oliver Eulenstein. 2009. Triplet supertree heuristics for the tree of life. *BMC Bioinformatics* 10, Suppl 1, Article S8 (2009). https://doi.org/10.1186/1471-2105-10-S1-S8

[22] Harris T. Lin, J. Gordon Burleigh, and Oliver Eulenstein. 2012. Consensus properties for the deep coalescence problem and their application for scalable tree search. *BMC Bioinformatics* 13 Suppl 10 (2012), S12. https://doi.org/10.1186/1471-2105-13-S10-S12

[23] Wayne P Maddison. 1997. Gene trees in species trees. *Systematic biology* 46, 3 (1997), 523–536.

[24] Wayne P. Maddison and L. Lacey Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol* 55, 1 (2006), 21–30. https://doi.org/10.1080/10635150500354928

[25] Alexey Markin and Oliver Eulenstein. 2016. Path-Difference Median Trees. In *Bioinformatics Research and Applications: 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings*, Anu Bourgeois, Pavel Skums, Xiang Wan, and Alex Zelikovsky (Eds.). Springer International Publishing, Cham, 211–223. https://doi.org/10.1007/978-3-319-38782-6_18

[26] Alexey Markin and Oliver Eulenstein. 2017. Computing Manhattan Path-Difference Median Trees: a Practical Local Search Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* PP (2017). https://dx.doi.org/10.1109/TCBB.2017.2718507

[27] James O. McInerney, James A. Cotton, and Davide Pisani. 2008. The prokaryotic tree of life: past, present... and future? *Trends Ecol Evol* 23, 5 (May 2008), 276–81. https://doi.org/10.1016/j.tree.2008.01.008

[28] Serena Nik-Zainal and et al. 2012. The life history of 21 breast cancers. *Cell* 149, 5 (2012), 994–1007.

[29] Roderic D. M. Page. 2002. Modified Mincut Supertrees. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI '02)*. Springer-Verlag, London, UK, 537–552.

[30] Samantha A. Price, Olaf R. P. Bininda-Emonds, and John L. Gittleman. 2005. A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla). *Biological Reviews* 80, 3 (2005), 445–473. https://doi.org/10.1017/S1464793105006743

[31] Mark A Ragan. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular phylogenetics and evolution* 1, 1 (1992), 53–58.

[32] Johannes J. Le Roux, Ania M. Wieczorek, Mohsen M. Ramadan, and Carol T. Tran. 2006. Resolving the native provenance of invasive fireweed (Senecio madagascariensis Poir.) in the Hawaiian Islands as inferred Poir.) in the Hawaiian Islands as inferred from phylogenetic analysis. *Diversity and Distributions* 12 (2006), 694–702.

[33] Charles Semple and Mike A. Steel. 2003. *Phylogenetics*. University Press, Oxford.

[34] S. Snir and S. Rao. 2010. Quartets MaxCut: A Divide and Conquer Quartets Algorithm. *IEEE/ACM TCBB* 7, 4 (2010), 704–718. https://doi.org/10.1109/TCBB.2008.133

[35] Robert R. Sokal and F. James Rohlf. 1962. The Comparison of Dendrograms by Objective Methods. *Taxon* 11, 2 (1962), 33–40. http://www.jstor.org/stable/1217208

[36] M. A. Steel. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9 (1992), 91–116.

[37] David L. Swofford. 1990. Phylogeny reconstruction. *Molecular Systematics* (1990), 411–501.

[38] David L. Swofford. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts. (2002).

[39] Cuong Than and Luay Nakhleh. 2009. Species tree inference by minimizing deep coalescences. *PLoS Comput Biol* 5, 9 (2009), e1000501. https://doi.org/10.1371/journal.pcbi.1000501

[40] Martin F Wojciechowski, Michael J Sanderson, Kelly P Steele, and Aaron Liston. 2000. Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach. *Advances in legume systematics* 9 (2000), 277–298.

[41] Louxin Zhang. 2011. From gene trees to species trees II: species tree inference by minimizing deep coalescence events. *IEEE/ACM Trans Comput Biol Bioinform* 8, 6 (2011), 1685–91. https://doi.org/10.1109/TCBB.2011.83