Article

Imputing Phylogenetic Trees Using Tropical Polytopes over the Space of Phylogenetic Trees

Ruriko Yoshida

Naval Postgraduate School, Monterey, CA 93943-5219, USA; ryoshida@nps.edu; Tel.: +1-831-656-2973

Abstract: When we apply comparative phylogenetic analyses to genome data, it poses a significant problem and challenge that some of the given species (or taxa) often have missing genes (i.e., data). In such a case, we have to impute a missing part of a gene tree from a sample of gene trees. In this short paper, we propose a novel method to infer the missing part of a phylogenetic tree using an analogue of a classical linear regression in the setting of tropical geometry. In our approach, we consider a tropical polytope, a convex hull with respect to the tropical metric closest to the data points. We show a condition that we can guarantee that an estimated tree from the method has at most a Robinson–Foulds (RF) distance of four from the ground truth, and computational experiments with simulated data and empirical data from Clavicipitaceae, which contains more than 4000 genes, show the method works well.

Keywords: missing information; phylogenomics; phylogenetics; tropical geometry

MSC: 92B15

1. Introduction

Due to new technology, today we are able to generate sequences from the genome with lower cost. However, at the same time, we have a great challenge to analyze large scale datasets from genome sequences. In phylogenomics, a new field that applies tools from phylogenetics to genome datasets, we often conduct comparative phylogenetic analyses, that is, to compare evolutionary histories among a set of taxa between different genes from the genome (for example, see [1]). However, we often face the problem in this process that some taxa in the dataset have missing gene(s) [2,3]. When this happens, systematists infer missing part of a gene tree from other gene trees using a supervised learning method, such as the linear regression model.

A phylogenetic tree is a weighted tree which represents evolutionary history of a given set of taxa (or species). In phylogenetic trees, leaves represent species or taxa in the present time which we can observe and internal nodes in the tree, which represent ancestors of the species, do not have any labels. A gene tree is a phylogenetic tree reconstructed from an alignment of a gene in a genome. Gene trees with the same set of species or taxa do not have to have the same tree topology since each gene might have different mutation rates due to the selection pressures, etc. [4]. In a comparative phylogenetic analysis, we often compare gene trees (for example, we compare how they are different, how their mutation rates are different, and often we are interested in inferring the species tree).

To infer a missing part of a gene tree, we often apply a supervised method to regress the missing part. In this process, first, we compute an unique vector representation of each gene tree. Then we infer the missing components of the vector of the tree from the vectors computed from other gene trees in a dataset using a regression model, such as a multiple linear regression [3].

However, a set of all such vectors realizing all possible phylogenetic trees, which is called a *space of phylogenetic trees*, is not Euclidean. In fact, a space of phylogenetic trees

is an union of polyhedral cones with a large co-dimension, so this is not even convex in terms of Euclidean metrics. Therefore, it is not appropriate to apply classical regression models, such as linear regression or neural networks, since they assume convexity in terms of Euclidean geometry. Thus, in this short paper, we propose an analogue of a classical multiple linear regression in the setting of tropical geometry with the max-plus algebra—an application of tropical polytopes to infer the missing part of a phylogenetic tree.

Equidistant trees are used to model gene trees under the multi-species coalescent model [4]. Therefore, in this paper, we focus on an equidistant tree, which is a rooted phylogenetic tree such that the total weight on an unique path from its root to each leaf is the same, and we focus on the space of all possible equidistant trees. It is well-known that the space of all possible equidistant trees is a *tropical Grassmannian*, which means that it is a tropically linear space with respect to the *tropical metric* [5–7]. Therefore, with the tropical metric with the max-plus algebra, we can conduct statistical analyses using tropical linear algebra, which is an analogue of a classical linear algebra. In fact, there has been much development in statistical learning over the space of phylogenetic trees using tools from tropical geometry [7–12].

Since a tropical polytope is tropically convex and since the space of equidistant trees is tropically convex, if all vertices are equidistant trees, then a tropical polytope is contained in the space of equidistant trees. Thus, in this paper, we propose to use a tropical polytope over the space of equidistant trees to infer missing part of a phylogenetic trees. This proposed method has basically four main step: (1) compute induced trees on the set of leaves which we observe in T, a tree with missing leaf (leaves) from a training set; (2) compare T with these induced trees; (3) compute a tropical polytope with trees with full set of leaves whose induced trees have closest tree topologies with T; and (4) project T onto the tropical polytope computed in step (3).

In Section 2, we discuss basics from tropical geometry and in Section 3, we discuss basics from phylogenetics. In Section 4, we show our novel method to impute a missing part of a phylogenetic tree. Then, in Section 5, we show a theoretical condition of T that the worst case scenario for the estimated tree via our method has the *Robinson–Foulds distance* at most 4. Then, Section 6 shows computational experiments of our method against other methods including a multiple linear regression and our method performs well.

2. Basics in Tropical Geometry

In this section, we discuss basics from tropical geometry. We consider the *tropical projective torus*, $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ where $\mathbf{1} := (1, 1, \ldots, 1)$ is the vector with all ones in $\mathbb{R}^e$. Basically this means that any vectors in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ is invariant with $\mathbf{1}$, i.e., $(v_1 + c, \ldots, v_e + c) = (v_1, \ldots, v_e) = v$ for any element $v := (v_1, \ldots, v_e) \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. For more details, see [13,14].

Under the tropical semiring $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$, the tropical arithmetic operations of addition and multiplication are defined as:

$$a \oplus b := \max\{a, b\}, \quad a \odot b := a + b \quad \text{where } a, b \in \mathbb{R} \cup \{-\infty\}.$$

For any scalars $a, b \in \mathbb{R} \cup \{-\infty\}$ and for any vectors $x = (x_1, \ldots, x_e)$, $y = (y_1, \ldots, y_e) \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, we have tropical scalar multiplication and tropical vector addition defined as:

$$a \odot x \oplus b \odot y := (\max\{a + x_1, b + y_1\}, \ldots, \max\{a + x_e, b + y_e\}).$$

Definition 1. *Suppose we have a set $S \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. If*

$$a \odot x \oplus b \odot y \in S$$

for any $a, b \in \mathbb{R}$ and for any $x, y \in S$, then S is called tropically convex. *Suppose we have a finite subset $V = \{v^1, \ldots, v^s\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Then, the smallest tropically-convex subset containing V is called the* tropical convex hull *or* tropical polytope *of V.* $\mathrm{tconv}(V)$ *can also be written as:*

$$\mathrm{tconv}(V) = \{a_1 \odot v^1 \oplus a_2 \odot v^2 \oplus \cdots \oplus a_s \odot v^s \mid a_1, \dots, a_s \in \mathbb{R}\}.$$

A tropical line segment, Γ_{v^1, v^2}, between two points v^1, v^2 is a tropical convex hull of $\{v^1, v^2\}$.

Example 1. *Suppose we have three points* $(0,0,0), (0,1,1), (0,2,5) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. *Then the tropical convex hull of these three points is shown in Figure 1.*
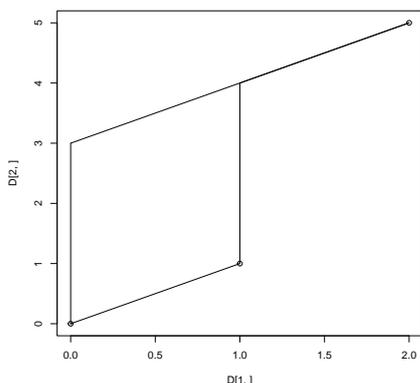
Figure 1. Tropical polytope of three points $(0,0,0), (0,1,1), (0,2,5) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$.

Remark 1. *By the definition, if a set $S \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ is tropically convex, then a tropical line segment between any two points in S must be contained in S.*

Definition 2. *For any points* $v := (v_1, \dots, v_e)$, $w := (w_1, \dots, w_e) \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, *the* tropical metric), d_{tr}, *between v and w is defined as:*

$$d_{\mathrm{tr}}(v, w) := \max_{i \in \{1, \dots, e\}} \{v_i - w_i\} - \min_{i \in \{1, \dots, e\}} \{v_i - w_i\}.$$

Example 2. *Suppose we have points* $(0,2,5), (0,5,5) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. *Then we have the distance between these two points in terms of the tropical metric is*

$$\begin{aligned} d_{\mathrm{tr}}((0,2,5), (0,5,5)) &= \max((0,2,5) - (0,5,5)) - \min((0,2,5) - (0,5,5)) \\ &= \max\{0, -3, 0\} - \min\max\{0, -3, 0\} = 3. \end{aligned}$$

Definition 3. *Let $V := \{v^1, \dots, v^s\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ and let $P = tconv(v^1, \dots, v^s) \subseteq \mathbb{R}^e/\mathbb{R}\mathbf{1}$ be a tropical polytope with its vertex set V. For $x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, let*

$$\pi_P(x) := \bigoplus_{l=1}^{s} \lambda_l \odot v^l, \quad \text{where} \quad \lambda_l = \min\{x - v^l\}. \tag{1}$$

Then $\pi_P(x)$ is a projection onto P with the property such that

$$d_{\mathrm{tr}}(x, \pi_P(x)) \le d_{\mathrm{tr}}(x, y)$$

for all $y \in P$.

Example 3. *Suppose we have three points* $v^1 = (0,0,0), v^2 = (0,1,1), v^3 = (0,2,5) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. *Also suppose the tropical convex hull P of these three points. Suppose we have a point $x = (0,2,2) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. Then,*

$$\begin{aligned} \lambda_1 &= \min\{0, 2-0, 2-0\} = 0, \\ \lambda_2 &= \min\{0, 2-1, 2-1\} = 0, \\ \lambda_3 &= \min\{0, 2-2, 2-5\} = -3. \end{aligned}$$

Thus, we have

$$
\begin{aligned}
\pi_P(x) &= \max\{(0+0, 0+0, 0+0), (0+0, 0+1, 0+1), (-3+0, -3+2, -3+5)\}, \\
&= \max\{(0,0,0), (0,1,1), (-3,-1,2)\}, \\
&= (0,1,2).
\end{aligned}
$$

Thus, $\pi_P(x) = (0,1,2)$.

3. Basics in Phylogenetic Trees

Let $[m] := \{1, \ldots, m\}$. A phylogenetic tree T on $[m]$ is a weighted tree of m leaves with the labels $[m]$ and internal nodes in the tree not labeled. A subtree T' in T on $a \subset [m]$ is a subtree of T with leaves a. An *equidistant tree* on $[m]$ is a rooted phylogenetic tree on $[m]$ such that the total weight on the path from its root to each leaf i in $[m]$ has the same distance for each $i \in [m]$. In this paper, we assume on equidistant trees.

In order to conduct a statistical analysis, we have to convert a phylogenetic tree into a vector. Here, we discuss one way to convert a phylogenetic tree into a vector.

Definition 4. *Suppose we have a* dissimilarity map $D : [m] \times [m] \to \mathbb{R}$ *such that*

$$
D(i,j) = \begin{cases} D(i,j) \geq 0 & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases}
$$

If there exists a phylogenetic tree on $[m]$ such that $D(i,j)$ is the total weight on the unique path from a leaf $i \in [m]$ to a leaf $j \in [m]$, then we call D as a tree metric.

Remark 2. *Since a tree metric of a phylogenetic tree on $[m]$ is symmetric and its diagonal is 0, we consider an upper triangular matrix of the tree metric and we consider the upper triangular matrix of the tree metric as a vector in $e = \binom{m}{2}$.*

Definition 5. *Let $D : [m] \times [m] \to \mathbb{R}$ be a metric over $[m]$, namely, D is a map from $[m] \times [m]$ to $\mathbb{R}$ such that*

$$
\begin{aligned}
D(i,j) &= D(j,i) & \text{for all } i,j \in [m] \\
D(i,j) &= 0 & \text{if and only if } i = j \\
D(i,j) &\leq D(i,k) + D(j,k) & \text{for all } i,j,k \in [m].
\end{aligned}
$$

Suppose D is a metric on $[m]$. Then if D satisfies

$$
\max\{D(i,j), D(i,k), D(j,k)\} \tag{2}
$$

is attained at least twice for any $i,j,k \in [m]$, then D is called an ultrametric.

Example 4. *Suppose we have $m = 3$ and*

$$
D(1,2) = 0, D(1,3) = 2, D(2,3) = 2.
$$

Then, D is an ultrametric.

It is well-known that if we have an ultrametric on $[m]$, then there is an unique equidistant tree on $[m]$ by the following theorem:

Theorem 1 ([15]). *Suppose we have an equidistant tree T with a leaf label set $[m]$ and suppose $D(i,j)$ for all $i,j \in [m]$ is the distance from a leaf i to a leaf j. Then, D is an ultrametric if and only if T is an equidistant tree on $[m]$.*

Therefore by Theorem 1, in this paper, we consider the set of ultrametrics, $\mathcal{U}_m \subset \mathbb{R}^e / \mathbb{R}\mathbf{1}$, on m as the space of equidistant trees on $[m]$.

Definition 6. *Let $x, y \subset [m]$ such that $x \cup y = [m]$ and $x \cap y = \emptyset$. Suppose we have an equidistant phylogenetic tree T with the leave set $[m]$. A* clade *of T with leaves $x \subset [m]$ is an equidistant tree on x constructed from T by adding all common ancestral interior nodes of any combinations of only leaves x and excluding common ancestors including any leaf from $[m] - x$ in T, and all edges in T connecting to these ancestral interior nodes and leaves x.*

Definition 7. *For a rooted phylogenetic tree, a* nearest neighbor interchange (NNI) *is an operation of a phylogenetic tree to change its tree topology by picking three mutually exclusive leaf sets $X_1, X_2, X_3 \subset X$ and changing a tree topology of the clade, possibly the whole tree, consisting with three distinct clades with leaf sets X_1, X_2, and X_3.*

Remark 3. *Since there are three possible ways of connecting three distinct clades, NNI possibly creates two new tree topologies on $[m]$.*

Definition 8. *Suppose we have rooted phylogenetic trees T_1, T_2 on $[m]$. The* Robinson–Foulds (RF) distance *is the number of operations that the subtree of T_1 has the same tree topology as the subtree of T_2 by removing a leaf of T_1 and the subtree of T_2 has the same tree topology as the subtree of T_1 by removing a leaf of T_2.*

Remark 4. *The Robinson–Foulds distance is always divisible by 2.*

Remark 5. *One can see that the Robinson–Foulds distance between two trees which is one NNI move apart is 2.*

4. Method

In this section, we introduce a method to infer a missing part of an equidistant tree using tools from tropical geometry. Let $RF(T_1, T_2)$ be the Robinson–Foulds distance between T_1 and T_2. The algorithm on this method is shown in Algorithm 1.

Algorithm 1: Imputation with a tropical polytope

    Data: Equidistant tree T' on $a \subset [m]$, and a sample of equidistant trees on $[m]$,
        $\{T_1, \ldots, T_n\}$
    Result: Estimated imputed tree $\hat{T}$ on $[m]$

1 Let T_i' be a tree dropped leaves $[m] - a$ from each T_1 for $i \in \{1, \ldots, n\}$;
2 Drop leaves $[m] - a$ from each T_1 for $i \in \{1, \ldots, n\}$;
3 Set $S = \emptyset$;
4 Set $d \leftarrow m^2$;
5 **for** $i = 1, \ldots, n - 1$ **do**
6   **if** $d > RF(T_i', T_j')$ **then**
7        $d \leftarrow RF(T_i', T')$;

8 **for** $i = 1, \ldots, n - 1$ **do**
9   **if** $RF(T_i', T') == d$ **then**
10       $S \leftarrow T_i' \cup S$;

11 Compute the tropical polytope P of ultrametrics computed from trees T_i for all
      trees $T_i' \in S$;
12 Let T_{tmp} be a tree attached leaves $[m] - a$ to the root of T';
13 Convert an ultrametric $u \in \mathcal{U}_m$ computed from T_{tmp};
14 Let v be a projection of u onto P;
15 Realize an equidistant tree $\hat{T}$ from v and return $\hat{T}$

---

5. Theoretical Results

Let $x, y \subset [m]$ such that $x \cup y = [m]$ and $x \cap y = \varnothing$. Let $\{T_1, \ldots, T_n\}$ be a sample of equidistant trees with m leaves. Let T'_i for $i = 1, \ldots n$ be an equidistant tree with x by dropping tips y from T_i, i.e., T'_i is an induced tree on a.

Theorem 2. *Suppose $\{T_1, \ldots, T_n\}$ is a sample of equidistant trees with $[m]$ and let $T_i = T'_i \cup T''_i$ such that T'_i is a subtree on x which is an equidistant tree with x by dropping tips y from T_i and T''_i is an subgraph graph with y by adding all common ancestral interior nodes of any combinations of only leaves x and excluding common ancestors including any leaf from $[m] - x$ in T_i for $i = 1, \ldots, n$. Suppose T'_i and T' have the same tree topology for $i = 1, \ldots, n$. If T''_i are clade in T_i for $i = 1, \ldots, n$ and T'' is also a clade in T, then an estimated tree $\hat{T}$ via our method with the tropical polytope $P := tconv\, (T_1, \ldots, T_n)$ and T differ at most the Robinson–Foulds distance = 4.*

Proof. Since T''_i are connected trees for $i = 1, \ldots, n$, T''_i forms a clade in T_i for $i = 1, \ldots, n$. Also T'' is a connected tree, so that T'' is also a clade in T. This means that T_i and T_j for any $i, j \in \{1, \ldots, n\}$ have only one NNI move distance. Since T'_i and T' have the same tree topology and since T'' is also a clade in T, T_i, and T differ only by one NNI move distance. Note that T_i and T_j have at most the Robinson–Foulds distance = 2 since T_i and T_j have just one NNI move difference, and so as with T. Let U_i be an ultrametric form a tree T_i for $i = 1, \ldots, n$. Then, take any tropical line segment Γ_{u_i, u_j}. Since T_i and T_j have just one NNI move difference, by Theorem 8 in [16], any tree topology of the tree realized from an ultrametric in Γ_{u_i, u_j} has the same tree topology of T_i or T_j. Since P is tropically convex, any point in P is a tropically convex combination of T_i for $i = 1, \ldots, n$. Thus, the tree topology of the tree realized by an ultrametric in P has at most one NNI move different. Since the estimate $\hat{T} \in P$, and any tree realized from an ultrametric in P to T has at most 2 NNI move difference. Thus, we have the result. $\square$

6. Computational Experiments

In this section, we apply this method to simulated data sets and compare its performance with the baseline model, which uses means of each missing element in an ultrametric computed from a tree, and the multiple linear regression model proposed by Yasui et al. in [3].

6.1. Simulated Data

To assess a performance of this method, we use simulated datasets generated from the multi-species coalescent model using the software `Mesquite` [17].

Under the multi-species coalescent model, there are two parameters: species depth SD and effective population size N_e. In this paper, we fix the effective population size as $N_e = 10,000$ and we vary SD as we vary the ratio

$$R = \frac{SD}{N_e}.$$

6.2. Experimental Design

Here we vary $R = 0.25, 0.5, 1, 2, 5, 10$. For this experiment, we fix the number of leaves as 10. Therefore $e = 45$. For each value of R, we generate a random species tree via the Yule model first. Then we generate the set of 1000 gene trees from the multi-species coalescent model given the species tree. Therefore, for each R, we have a simulated dataset with size 1000.

Note that when R is larger we have tighter constraints to gene tree topologies by its species tree. Therefore, we do not have large variance for generating gene trees so that it is easier to estimate the missing part of a gene tree. On the other hand, if we have small R, then we have a large variance for gene tree topologies, the coalescent model is becoming more like a random process [4].

For estimating the performance of our method when we vary the number of leaves missing, we set three different cases: one leaf out of 10 leaves is removed, two leaves out of 10 leaves are removed, and three leaves out of 10 leaves are removed. For each scenario in terms of R and in terms of the number of leaves removed, we pick 200 random observations from the data set of 1000 trees as a test set.

To compare the performance of our method, we use the baseline model, i.e., we fill missing values of an ultrametric by taking the mean of observations with the full set of leaves and the multiple linear regression model proposed by Yasui et al. [3]. For the multiple linear regression model, we set a missing element as a response variable and observed elements in an ultrametric as predictors [3].

6.3. Results

To assess a performance of our method against the baseline and the method proposed by Yasui et al., we use the Robinson–Foulds distance between an estimated tree $\hat{T}$ and T. The results are shown in Table 1 and Figure 2. Note that the smaller the Robinson–Foulds distance between two trees, the closer their tree topologies are. When the Robinson–Foulds distance is 0, their tree topologies are the same. Figure 3 shows that the performance of the baseline and the performance of the linear regression model have very similar results for one, two, and three, the number of leaves removed and all ratio of R.

Table 1. These numbers shown in this table are average Robinson–Foulds distances between estimated trees and true trees. For each category, we infer 200 trees from 800 trees. The smaller the Robinson–Foulds distance is, the better the performance.

# of Leaves Removed	Method	R					
		10	**5**	**2**	**1**	**0.5**	**0.25**
1 leaf removed	Tropical Metric	0.76	2.30	4.11	6.11	7.91	9.13
	Baseline	2.63	6.34	8.71	11.36	12.43	12.87
	Linear Regression	2.70	6.34	8.66	11.26	12.44	12.89
2 leaves removed	Tropical Metric	0.84	2.65	5.36	6.95	8.26	9.20
	Baseline	2.55	6.28	8.32	11.02	12.55	13.00
	Linear Regression	2.66	6.29	8.58	10.96	12.51	13.06
3 leaves removed	Tropical Metric	1.32	2.95	6.06	7.93	9.85	9.84
	Baseline	2.46	6.12	7.95	10.80	12.73	13.10
	Linear Regression	2.56	6.16	8.05	10.94	12.70	13.23

According to computational experiments with simulated datasets shown in Table 1 and Figure 2, our method has smaller Robinson–Foulds distances in any cases compared to other methods. It is interesting that the number of leaves removed seems to greatly affect the results in general while R clearly affects performances of all three methods compared.

If we have only one missing leaf and larger R, often the average Robinson–Foulds distances between inferred trees and true trees is less than 1 because we have often the condition satisfied in Theorem 2 due to very high constraints on tree topologies of gene trees.
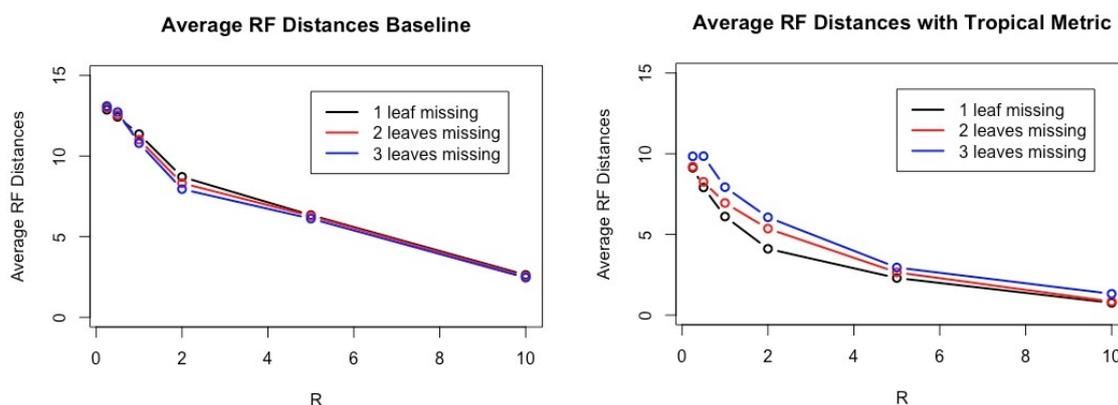
Figure 2. This figure shows performance on the baseline (**Left**) and our method using a tropical poltyope (**Right**). For each category, we infer 200 trees from 800 trees. The x-axis represents the ratio R and the y-axis shows the average Robinson–Foulds distances between estimated trees and true trees for 200 trees. The smaller the Robinson–Foulds distance is, the better the performance.
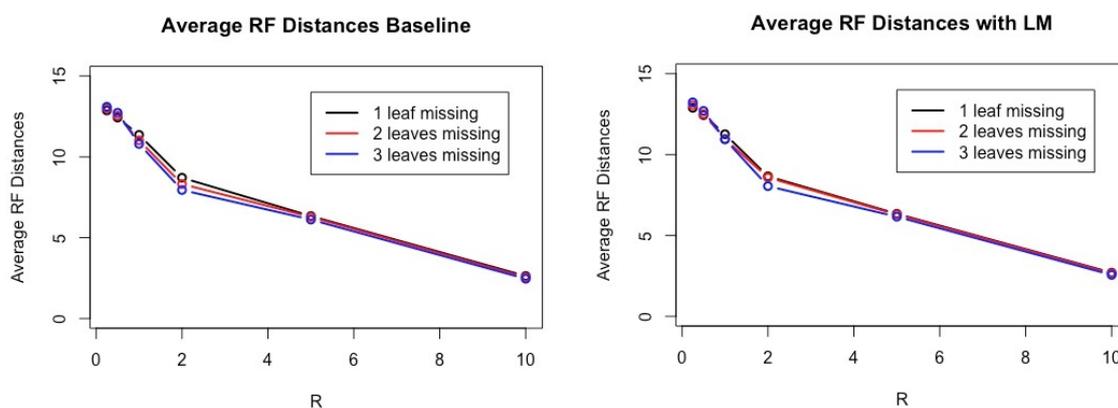
Figure 3. This figure shows performance on the baseline model (**left**) and linear regression models (**Right**). For each category, we infer 200 trees from 800 trees. The x-axis represents the ratio R and the y-axis shows the average Robinson–Foulds distances between estimated trees and true trees for 200 trees. The smaller the Robinson–Foulds distance is, the better the performance. As one can see, these results are very close to each other for all R.

7. Application to Empirical Data

7.1. Empirical Dataset

We apply our imputation method to a set of annotated genomes of the fungal family, Clavicipitaceae from [18]. Kang et al. annotated genome sequences for 12 species from the fungal family Clavicipitaceae with MAKER. The 12 species in Clavicipitaceae include *Epichloë festucae* Fl1 (designated E0894), *Aciculosporium take* (A7993), *Atkinsonella texensis* B6155, *Balansia obtecta* B249 (B0249), *Claviceps purpurea* 20.1 (C0201), *Epichloë amarillans* E4668, *Epichloë inebrians* E818 (E0818), *Epichloë glyceriae* E277 (E0277), *Epichloë mollis* AL9923 (E3601), *Epichloë typhina* subsp. *poae* E5819, *Metarhizium robertsii* ARSEF 23 (M0023), and *Periglandula ipomoeae* P4806. For details on phylogenetic reconstructions on this dataset, see [18]. In this dataset there are 3408 house keeping genes out of 4268 genes. Among them 860 gene trees, including gene trees for ergot alkaloid biosynthesis (EAS) genes have missing leaf (leaves).

7.2. Results

After imputation, we apply the tropical kernel density estimator (KDE) [19] to find possible outlier(s) from the dataset. With the significance level $\alpha = 0.01$ and bandwidth estimated by the nearest neighbors, we found 43 possible outliers including 12 core house-

keeping genes and a mating type gene (*mtAC*) with those of the five *EAS* genes, *easG*, *easC*, *easD*, *cloA*, and *easA*.

Since the tropical polytope defined by the set of all house keeping genes with full leaves is in $\binom{12}{2}$ dimension. Thus, in order to visualize the tropical polytope we apply a tropical principal component analysis (PCA) developed by Yoshida et al. [7,8]. We use the tropical PCA to visualize in the two dimension shown in Figure 4.
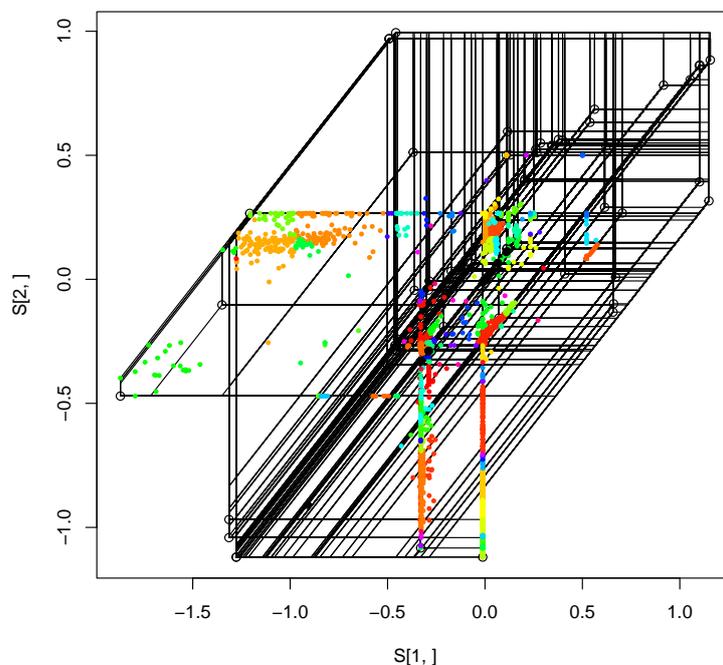
Figure 4. The tropical polytope of gene trees for the housekeeping genes. Black points in the figures are outliers found by the tropical KDE. In order to visualize this, we apply the tropical PCA. Each color represents an unique tree topology.

8. Discussion

In this short paper, we show a novel method to impute a missing part of an equidistant tree on [*m*] using a tropical polytope, which is an analogue of a linear regression in the setting of tropical geometry. From simulated data generated from the multi-species coalescent model, we show that this method works very well. In addition, we show a condition that the estimate tree and the true tree have at most a Robinson–Foulds distance of 4 (Theorem 2).

In order to demonstrate our method, we apply it to simulated data as well as an empirical data from Clavicipitaceae. This dataset contains more than 4000 genes and 860 out of 4268 gene trees miss some of leaves. Some trees miss 8 out of 12 tips. Then, we apply the tropical KDE developed by [19] to find outlying gene trees and the results are consistent with the conclusion in [18].

In the future, we can investigate applying the "tropical principal component analysis (PCA)" proposed by Yoshida, et al. in [8] to an imputation of trees since the classical PCA can be viewed as a multivariate linear regression model with orthogonal projections.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declare no conflict of interest.

References

1. Koonin, E.; Puigbó, P.; Wolf, Y. Comparison of phylogenetic trees and search for a central trend in the "forest of life". *J. Comput. Biol.* **2011**, *18*, 917–924. [CrossRef] [PubMed]
2. Peters, C.; Geary, M.; Nelson, H.; Rusk, B.; von Hardenberg, A.; Muir, A.P. Phylogenetic placement and life history trait imputation for Grenada Dove Leptotila wellsi, Generalized fuzzy trees. *Int. J. Comput. Intell. Syst.* **2023**, *10*, 711–720.
3. Yasui, N.; Vogiatzis, C.; Yoshida, R.; Fukumizu, K. imPhy: Imputing Phylogenetic Trees with Missing Information Using Mathematical Programming. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 1222–1230. [CrossRef] [PubMed]
4. Maddison, W. Gene Trees in Species Trees. *Syst. Biol.* **1997**, *46*, 523–536. [CrossRef]
5. Ardila, F.; Klivans, C.J. The Bergman Complex of a Matroid and Phylogenetic Trees. *J. Comb. Theory Ser. B* **2006**, *96*, 38–49. [CrossRef]
6. Speyer, D.; Sturmfels, B. Tropical mathematics. *Math. Mag.* **2009**, *82*, 163–173. [CrossRef]
7. Page, R.; Yoshida, R.; Zhang, L. Tropical principal component analysis on the space of phylogenetic trees. *Bioinformatics* **2020**, *36*, 4590–4598.
8. Yoshida, R.; Zhang, L.; Zhang, X. Tropical Principal Component Analysis and its Application to Phylogenetics. *arXiv* **2017**, arXiv:1710.02682.
9. Monod, A.; Lin, B.; Yoshida, R.; Kang, Q. Tropical Geometry of Phylogenetic Tree Space: A Statistical Perspective. 2019. Available online: https://arxiv.org/pdf/1805.12400.pdf (accessed on 27 June 2023).
10. Lin, B.; Sturmfels, B.; Tang, X.; Yoshida, R. Convexity in Tree Spaces. *SIAM J. Discret. Math.* **2017**, *3*, 2015–2038. [CrossRef]
11. Yoshida, R. Tropical Balls and its Applications to K Nearest Neighbor over the Space of Phylogenetic Trees. *Mathematics* **2021**, *9*, 779. [CrossRef]
12. Yoshida, R. Tropical Data Science over the Space of Phylogenetic Trees. In *IntelliSys 2021: The Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2021; Volume 295, Chapter 26, pp. 340–361.
13. Joswig, M. *Essentials of Tropical Combinatorics*; Graduate Studies in Mathematics; American Mathematical Society: Providence, RI, USA, 2021.
14. Maclagan, D.; Sturmfels, B. *Introduction to Tropical Geometry*; American Mathematical Society: Providence, RI, USA, 2015; Volume 161.
15. Buneman, P. A note on the metric properties of trees. *J. Comb. Theory Ser. B* **1974**, *17*, 48–50. [CrossRef]
16. Yoshida, R.; Cox, S. Tree Topologies along a Tropical Line Segment. *Vietnam J. Math.* **2022**, *50*, 395–419. [CrossRef]
17. Maddison, W.P.; Maddison, D. Mesquite: A Modular System for Evolutionary Analysis. Version 2.72. 2009. Available online: http://mesquiteproject.org (accessed on 28 June 2023).
18. Kang, Q.; Schardl, C.; Moore, N.; Yoshida, R. CURatio: Genome-wide phylogenomic analysis method using ratios of total branch lengths. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *17*, 981–989. [CrossRef] [PubMed]
19. Yoshida, R.; Miura, K.; Barnhill, D.; Howe, D. Tropical Density Estimation of Phylogenetic Trees. 2022. Available online: https://arxiv.org/abs/2206.04206 (accessed on 30 June 2023).