Tropical Density Estimation of Phylogenetic Trees

Ruriko Yoshida, David Barnhill, Keiji Miura, and Daniel Howe

Abstract—Much evidence from biological theory and empirical data indicates that, gene trees, phylogenetic trees reconstructed from different genes (loci), do not have to have exactly the same tree topologies. Such incongruence between gene trees might be caused by some "unusual" evolutionary events, such as meiotic sexual recombination in eukaryotes or horizontal transfers of genetic material in prokaryotes. However, most of the gene trees are constrained by the tree topology of the underlying species tree, that is, the phylogenetic tree depicting the evolutionary history of the set of species under consideration. In order to discover "outlying" gene trees which do not follow the "main distribution(s)" of trees, we propose to apply the "tropical metric" with the max-plus algebra from tropical geometry to a non-parametric estimation of gene trees over the space of phylogenetic trees. In this research we apply the "tropical metric," a well-defined metric over the space of phylogenetic trees under the max-plus algebra, to non-parametric estimation of gene trees distribution over the tree space. Kernel density estimator (KDE) is one of the most popular non-parametric estimation of a distribution from a given sample, and we propose an analogue of the classical KDE in the setting of tropical geometry with the tropical metric which measures the length of an intrinsic geodesic between trees over the tree space. We estimate the probability of an observed tree by empirical frequencies of nearby trees, with the level of influence determined by the tropical metric. Then, with simulated data generated from the multispecies coalescent model, we show that the non-parametric estimation of the gene tree distribution using the tropical metric performs better than one using the Billera-Holmes-Vogtmann (BHV) metric developed by Weyenberg et al. in terms of computational times and accuracy. We then apply it to Apicomplexa data.

Index Terms—Apicomplexa, Kernel Density Estimator, Phylogenomics, Space of Phylogenetic Trees, Tropical Geometry

1 Introduction

Due to efficient genome sequencing technologies in terms of time and cost, it is essential to develop efficient bioinformatic methods to analyze genome structure and evolution. In this paper, we focus on correlations between *gene trees*, that is, phylogenetic trees reconstructed from alignments of genes in a genome. Ané et al., for example, applied a Bayesian method to estimate concordance among gene trees from multiple loci in [1]. However, most of such estimators assume that concordance is present among the given set of gene trees. In practice, however, numerous evolutionary processes can reduce correlations between gene trees. For example, such evolutionary processes include negative or balancing selection on a locus, which might increase the chance for ancestral gene copies to maintain through speciation events [2], and horizontal gene transfer, which shuffles divergent genes among different species [3].

In this paper, we propose a method analogous to estimate a distribution of gene trees over the *space of phylogenetic trees* as a whole. Especially using this estimated distribution of trees, our interest is to identify gene trees which exhibit significant discordance among gene trees. These "uncommon" genes can be generated from evolutionary processes, such as, paralogy, neofunctionalization, horizontal gene transfer or periods of rapid molecular evolution, and they might come from processes of data analyses, such as incorrect sequencing, alignment, tree reconstruction or annotation [4].

While parametric statistical methods are available or under development, we propose in this paper a *nonparametric* approach which offers particular advantages in phylogenomic analyses. In particular, problems of estimation and potentially incorrect selection of model parameters, which could obscure the search for outlier trees, are obviated. Here we propose a method analogue to kernel density estimator to estimate a distribution of trees over the space of trees using tools from tropical geometry.

One of the most challenging problems in phylogenomics is to study correlations among gene trees over the space of phylogenetic trees. Ideally, we could apply conventional statistical methods directly to a sample of gene trees, however, the *space of phylogenetic trees*, the space of all possible phylogenetic trees with m leaves, is not Euclidean. In fact it is an union of lower dimensional cones over $\mathbb{R}^e$, where $e = \binom{m}{2}$ and it is not convex [5]. Therefore, we cannot just apply conventional statistical models in data science to a set of phylogenetic trees because these methods assume Euclidean spaces [6].

The notion of the space of phylogenetic trees with m leaves comes from the work by Billera-Holmes-Vogtmann (BHV) in [7]. Billera-Holmes-Vogtmann defined the space by gluing $m-2$ dimensional positive orthants, where each orthant represents all possible rooted phylogenetic trees with a fixed tree topology with m leaves. Over the tree space with the *BHV metric*, two orthants with coordinates defined by edge lengths of interior edges are glued to each other if the tree topology for one orthant differs by one nearest

- *R. Yoshida and D. Barnhill are with the department of Operations Research, Naval Postgraduate School, Monterey, CA, 93943, USA. E-mail: see http://polytopes.net*
- *K. Miura is with School of Biological and Environmental Sciences, Kwansei Gakuin University, Sanda, Hyogo, 669-1330, Japan.*
- *D. Howe is with Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA.*

Manuscript received XXX, 2022; revised XXX, 2022.

neighbor interchange (NNI) move from the tree topology for the other orthant. In their work, Billera-Holmes-Vogtmann also showed that this tree space is $\mathrm{CAT}(0)$ space. This means that for any two trees in the tree space there is a unique shortest connecting path, called a geodesic, defined by the $\mathrm{CAT}(0)$-metric.

Shortly after that, in 2004, Speyer and Sturmfels showed that the space of phylogenetic trees with m leaves is a tropical Grassmanian [8], which is a *tropicalization* of the set of all solutions for a system of certain linear equations [6] under max-plus arithmetic. Several researchers then showed that the *tropical metric* with max-plus algebra on the space of *equidistant trees* with m leaves behaves very well [9], [10], [11]. For example, the dimension of the convex hull, the smallest tropical convex set, of s points with the tropical metric over the *tropical projective space* $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$ is at most $s-1$ while this is not the case with the BHV metric [11]. Therefore developing a machine learning algorithm that predicts based on the tropical metric as a data proximity measure is crucial for novel discovery.

Applications of the tropical metric to phylogenomics have been often done over the space of equidistant trees. An equidistant tree is a rooted phylogenetic tree whose distance between its root to each leaf is the same for all leaves in the tree. In terms of biology, this can be seen as a phylogenetic tree with a molecular clock. Also, the multispecies coalescent model assumes that all gene trees are equidistant. Therefore this is a natural assumption in evolutionary biology [12]. For example, Yoshida et al. [6] and Page et al. [13] developed *tropical principal component analysis*. In their work they use the fact that the space of equidistant trees with m leaves is a tropically convex set over the tropical projective space in terms of the tropical metric and the *tropical line segment* between any trees over the space is intrinsically geodesic and is unique [14].

A kernel density estimator (KDE) is a non-parametric density estimator using kernel functions, which is useful for, say, discovering outliers. Weyenberg et al. developed a non-parametric density estimator over the space of phylogenetic trees in terms of the BHV metric by mimicking a classical KDE [15]. The biggest problem Weyenberg et al. encountered was that the kernel function normalizing constant varies depending on the location of the center of the function. In addition, even though Weyenberg et al. developed a method to approximate the normalizing constant for a kernel function with the BHV metric over the space of phylogenetic trees, there is still no explicit method to compute the normalizing constant.

In this paper, since the space of equidistant trees is a tropical convex set [6], we apply a Hit and Run (HAR) sampler from tropically convex sets with the tropical metric developed by Yoshida et al. [16] to estimate the normalizing constant of a kernel function with the tropical metric over the space of equidistant trees. Computationally, we show that the normalizing constant of a kernel function is independent from a central location of the function over the space of equidistant trees. Then we develop an analogue of a classical KDE with the tropical metric over the space of equidistant trees and, with simulated data generated from the multispecies coalescent model, we show that the KDE with the tropical metric performs better than the one with the BHV developed by Weyenberg et al. [17] in terms of computational time and accuracy. We also apply it to Apicomplexa data from [18].

This paper is organized as follows. In Section 2, we first outline basics on tropical geometry using the max-plus algebra over the tropical semiring. Then, we outline our non-parametric estimation of the gene tree distribution over the space of phylogenetic trees with a given set of leaves defined by the tropical metric. In Section 3, we show how we set up simulation studies with our method. Then we show the results from computational experiments with simulated data generated from the multispecies coalescent model and with the empirical data of Apicomplexa from [18]. In Section 4, we discuss the results from computational experiments and we end with future work and an open problem in Section 5.

2 METHODS

2.1 Basics of Tropical Geometry

Throughout this paper, like [8], we consider the *tropical projective torus* $\mathbb{R}^e/\mathbb{R}\mathbf{1}$, which is isomorphic to $\mathbb{R}^{e-1}$. For more details, see [19], [20].

Definition 2.1 (Tropical Arithmetic Operations). Under the tropical semiring $(\mathbb{R} \cup \{-\infty\}, \boxplus, \odot)$, the tropical arithmetic operations of addition and multiplication are defined as:

$$c_1 \boxplus c_2 := \max\{c_1, c_2\}, \quad c_1 \odot c_2 := c_1 + c_2,$$

where $c_1,\ c_2 \in \mathbb{R} \cup \{-\infty\}$. Over the tropical semiring, the identity element under addition is $-\infty$ and the identity element under multiplication is 0.

Definition 2.2 (Tropical Scalar Multiplication and Vector Addition). For any scalars $c_1,\ c_2 \in \mathbb{R} \cup \{-\infty\}$ and for any vectors $v = (v_1, \ldots, v_e), w = (w_1, \ldots, w_e)$ over the *tropical projective space* $(\mathbb{R} \cup -\{\infty\})^e/\mathbb{R}\mathbf{1}$, we have tropical scalar multiplication and tropical vector addition as:

$$c_1 \odot v \boxplus c_2 \odot w := (\max\{c_1+v_1, c_2+w_1\}, \ldots, \max\{c_1+v_e, c_2+w_e\}).$$

Definition 2.3. Suppose we have $S \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. S is *tropically convex* if

$$c_1 \odot v \boxplus c_2 \odot w \in S$$

for any $c_1, c_2 \in \mathbb{R}$ and for any points $v, w \in S$. Suppose $V = \{v^1, \ldots, v^s\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. The smallest tropically-convex subset containing V is called the *tropical convex hull* or *tropical polytope* of V which can be written as the set of all tropical linear combinations of V as:

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

A *tropical line segment* between two points $v^1,\ v^2$ is a tropical polytope of a set of two points $\{v^1,\ v^2\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$.

Definition 2.4 (Generalized Hilbert Projective Metric). For any vectors $v := (v_1, \ldots, v_e),\ w := (w_1, \ldots, w_e) \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, the *tropical distance* d_{tr} between v and w is defined as:

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

This distance measure is a well-defined metric over the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ [11].

2.2 Basics of Ultrametrics

Suppose we have $[m] := \{1, \ldots, m\}$ and let $d : [m] \times [m] \to \mathbb{R}$ be a metric over $[m]$, that is, d is a map from $[m] \times [m]$ to $\mathbb{R}$ such that

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

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

$$\max\{d(i,j), d(i,k), d(j,k)\} \tag{1}$$

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

Example 2.5. Suppose $m = 3$. Let d be a metric on $[m] := \{1, 2, 3\}$ such that

$$d(1,2) = 2,\ d(1,3) = 2,\ d(2,3) = 1.$$

Since the maximum is achieved twice, d is an ultrametric.

A phylogenetic tree is a weighted tree whose internal nodes do not have labels and whose external nodes, i.e., leaves, have labels $[m]$. Throughout this paper, we consider a rooted phylogenetic tree with a leaf label set $[m]$.

Definition 2.6. Suppose we have a rooted phylogenetic tree T with a leaf label set $[m]$. If the distance from its root to each leaf $i \in [m]$ is the same distance for all $i \in [m]$, then we call T an *equidistant tree*.

In order to conduct a statistical analysis, we need to map a phylogenetic tree on $[m]$ to a vector representation. There are many ways to map a phylogenetic tree to a vector, including the BHV coordinates [7]. In this paper, we vectorize phylogenetic tree as dissimilarity maps. Dissimilarity maps are maps $d : [m] \times [m] \to \mathbb{R}$ such that $d(i,i) = 0$ and $d(i,j) = d(j,i)$. In phylogenetics, we consider dissimilarity maps over the product of a leaf set $[m]$ such that $d(i,j)$ is the pairwise distance between a leaf $i \in [m]$ and a leaf $j \in [m]$. Throughout this paper we consider a vector of all possible pairwise distances in T between any two leaves in $[m]$ as a vector representation of a phylogenetic tree T with $[m]$. Then we have the following theorem.

Theorem 2.7 ([21]). *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.*

Example 2.8. Suppose we have $m = 5$. Then, the phylogenetic tree shown in Fig. 1 is an equidistant tree with a leaf label set $[5] := \{A, B, C, D, E\}$ and its pairwise distances are

$$u = (4, 4, 4, 4, 2, 2, 2, 1.6, 1.6, 0.6)$$

which is an ultrametric.

Using Theorem 2.7, if we wish to consider all possible equidistant trees, then it is equivalent to consider the space of ultrametrics as the space of phylogenetic trees on $[m]$. Here we define $\mathcal{U}_m$ as the space of ultrametrics with a set of leaf labels $[m]$.

Throughout this paper, we assume we have a sample of gene trees which are equidistant. This assumption is not unusual in phylogenomics since the multispecies coalescent model assumes that all gene trees are equidistant trees in order to conduct the inference on the species tree from a sample of gene trees [12].

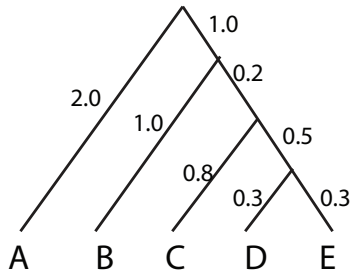

Fig. 1. Example of an equidistant tree with a leaf label set $[5]$.

Theorem 2.9 ([5], [13]). *Suppose we have a* classical *linear subspace $L_m \subset \mathbb{R}^e$ defined by the linear equations $x_{ij} - x_{ik} + x_{jk} = 0$ for $1 \leq i < j < k \leq m$. Let Trop$(L_m) \subseteq \mathbb{R}^e/\mathbb{R}\mathbf{1}$ be the* tropicalization *of the linear space $L_m \subset \mathbb{R}^e$, that is, we replace the classical addition by the tropical addition $\boxplus$ and we replace the classical multiplication by the tropical multiplication $\odot$ in the equations defining the linear subspace L_m, so that all points $(x_{12}, x_{13}, \ldots, x_{m-1,m})$ in Trop(L_m) satisfy the condition:*

$$\max_{i,j,k \in [m]} \{v_{ij}, v_{ik}, v_{jk}\}$$

is achieved at least twice. Then the image of $\mathcal{U}_m$ inside of the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ is equal to Trop(L_m).

Remark 1. Since $\mathcal{U}_m \subseteq \mathbb{R}^e/\mathbb{R}\mathbf{1}$ is the tropicalization of the linear subspace, $\mathcal{U}_m$ is tropically convex. Therefore, if we take a tropical line segment $\Gamma_{u,v}$ between any two ultrametrics $u, v \in \mathcal{U}_m$, then since $\Gamma_{u,v}$ is also tropically convex, $\Gamma_{u,v}$ is contained in $\mathcal{U}_m$, i.e., $\Gamma_{u,v} \subset \mathcal{U}_m$. Further, Monod et al. in [14] showed that $\Gamma_{u,v}$ is a unique geodesic between $u, v \in \mathcal{U}_m$. Therefore, $d_{\mathrm{tr}}(u,v)$ measures the length of $\Gamma_{u,v}$ which is an *intrinsic metric* between $u, v \in \mathcal{U}_m$.

2.3 Non-parametric Estimation of Gene Tree Distribution

Suppose we have an i.i.d. sample of trees $\mathcal{S} := \{T_1, \ldots, T_N\} \subset \mathcal{U}_m$. Our goal is to estimate the gene tree distribution from $\mathcal{S}$ over the space of ultrametrics $\mathcal{U}_m$. Here we assume that the 'non-outlying trees' are independently sampled from some unknown distribution which we are interested in estimating and 'outlying trees' are sampled from a different distribution. Our non-parametric density estimator with the tropical metric over the space of ultrametrics $\mathcal{U}_m$ mimics a classical kernel density estimator (KDE) formulated as:

$$\hat{f}(T) \propto \frac{1}{N} \sum_{i=1}^{N} k(T, T_i) \tag{2}$$

where k is a non-negative function defined over $\mathcal{U}_m$ such that

$$k(T, T_i) = \exp\left(-\left(\frac{d_{\mathrm{tr}}(T, T_i)}{\sigma}\right)\right), \tag{3}$$

where $\sigma > 0$ is a user specified parameter to define "bandwidth" which controls, how tightly each contribution of a function $k(T, T_i)$ will be centered around $T_i \in \mathcal{S}$ in terms

of d_{tr} (See Formula (2)). In [15], [17], the default set up of this user-defined parameter is determined by the nearest neighbor of each $T_i \in \mathcal{S}$. Ideally the normalizing constant

$$C(T_i) = \int_{\mathcal{U}_m} k(T, T_i) dT$$

does not depend on $T_i \in \mathcal{U}_m$ so that we do not have to compute the normalizing constant for each $T_i \in \mathcal{S}$ as was required in [17]. In achieving this, our proposed method in this section will be more analogous to a kernel density estimation. Since our experiments in the following subsection show that the normalizing constant $C(T_i)$ does not vary for any $T_i \in \mathcal{U}_m$, we assume that the normalizing constant $C(T_i)$ is a constant for any $T_i \in \mathcal{U}_m$.

In this paper we are interested in detecting outliers $T_j \in \mathcal{S}$ similar to [15], [17]. Therefore, we consider the estimation

$$\hat{g}(T_j) \propto \frac{1}{N-1} \sum_{i \neq j} k(T_j, T_i)$$

for $T_j \in \mathcal{S}$. As is the case in [15], [17], after we estimated $\hat{g}(T_j)$ for each $T_j \in \mathcal{S}$, we classify T_j as an outlying tree if $\hat{g}(T_j)$ is less than $Q_1 - \kappa IQR$, where Q_1 is the first quartile and IQR is the interquartile range of the set of all scores for all trees in $\mathcal{S}$. κ is a tuning parameter and it is set to 1.5 as a default [22].

2.4 Approximating Normalizing Constants

In [15], [17], the authors considered the function

$$k_{\mathrm{BHV}}(T, T_i) \propto \exp\left(-\left(\frac{d_{\mathrm{BHV}}(T, T_i)^2}{\sigma}\right)\right), \qquad (4)$$

where d_{BHV} is the BHV metric defined by Billera, Holmes and Vogtmann over $\mathcal{T}_m$, the space of phylogenetic trees with m leaves using the BHV metric [7]. In [17], Weyenberg et al. showed that $C_{\mathrm{BHV}}(T_i)$ varies on $T_i \in \mathcal{T}_m$ where

$$C_{\mathrm{BHV}}(T_i) = \int_{\mathcal{T}_m} k_{\mathrm{BHV}}(T, T_i) dT.$$

Therefore, Weyenberg et al. in [15] developed an algorithm to approximate $C_{\mathrm{BHV}}(T_i)$ for any $T_i \in \mathcal{T}_m$. When T_i is the star tree, i.e., the tree with no internal branch, $C_{\mathrm{BHV}}(T_i)$ achieves its largest values. Therefore, in this section, we apply a *Hit and Run* sampler developed by Yoshida et al. [16] to approximate the normalizing constant of $k(T, T_i)$ for $T_i \in \mathcal{U}_m$. Especially, we compare the normalizing constant of $k(T, T_i)$ where T_i is the star tree and T_i is a binary random tree for $m = 10$.

Example 2.10. In this example, we use three different trees as the centroid of each distribution. One is shown in the left picture of Fig. 2. The second is shown in the right picture of Fig. 2 and the last is the star tree of ten leaves with the length of each branch equal to 1. Using 1,000 samples and letting $\sigma = 1.5, 2, 5$ we achieve the results shown in Table 1. From this result, it seems that the normalizing constant $D(T_i)$ in terms of d_{tr} for any $T_i \in \mathcal{U}_m$ is invariant under the different tree topologies.

Estimating the normalizing constants

$\sigma \backslash$ Tree Type	Tree in left Fig. 2	Tree in right Fig. 2	Star Tree
1.5	117.21	120.59	117.78
2	199.24	199.24	201.23
5	521.23	521.12	524.45

TABLE 1
Results for estimating the normalizing constants for different centroids with varying σ.

2.5 Computational time

The computational time complexity of the tropical distance $d_{\mathrm{tr}}(T_1, T_2)$ between two trees $T_1, T_2 \in \mathcal{U}_m$ is $O(m^2)$. Therefore, when computing the normalizing constant, for each $T \in \mathcal{U}_m$, the time complexity of computing $f(T)$ is $O(Nm^2)$ while with the BHV metric it is $O(Nm^6)$ for each $T \in \mathcal{T}_m$.

3 Results

3.1 Simulated Experiments

For this computational experiment, we generate gene trees from the multispecies coalescent models with a given species tree via the software `Mesquite` [12]. We fixed the effective population size $N_e = 100,000$ and varied $R = \frac{SD}{N_e}$ where SD is the species depth which is the number of generations from the common ancestor (the root) to the taxa (leaves).

Algorithm 1 Generating a set of gene trees from the multispecies coalescent model

Input: The number of leaves m; R, the ratio of the species depth and effective population size; and the number of gene trees N.
Output: A sample of gene trees $\mathbb{T}$.
Set the labels for leaves to the species tree and gene trees using m.
Use the Yule model to generate a random species tree T.
Using the species tree T with the ratio R, generate N gene trees $\mathbb{T}$. **return** $\mathbb{T}$.

To sample trees randomly from two different distributions, we fix the number of leaves as $m = 10$ and generate two different species trees T_1, T_2 using the Yule process. Then using the coalescent model for gene trees within the species tree, we generate 1000 gene trees for each species tree via Algorithm 1. In these simulated experiments, we vary the ratio $R = 0.25, 0.5, 1, 2, 5, 10$. Let $\mathbb{T}_1$ be the set of gene trees with the species tree T_1 and let $\mathbb{T}_2$ be the set of gene trees with the species tree T_2. It is worth noting that when we have small R, gene trees generated from a coalescent model within a given species tree are similar to random trees. Thus, it becomes harder to distinguish between two distributions of gene trees with two different species trees as R becomes smaller [24].

To get the ROCs for the two samples, we conduct experiments described in Algorithm 2 with $r = 500$ and $g = 1000$. More specifically, for each R, we take all 1000 trees from $\mathbb{T}_1$ and we take one tree from $\mathbb{T}_2$. Then we estimate probability distribution of gene trees using the tropical density estimator described in Equation (2) (Fig. 3, Left) and with `KDETrees` (Fig. 3, Right). We iterate this process 500 times.

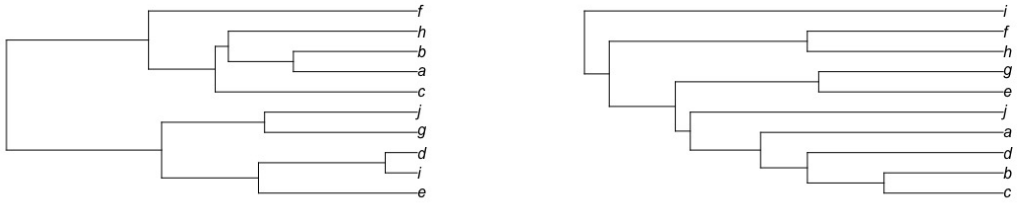

Fig. 2. (Left) The centroid tree used for computational experiments. Its ultrametric is $u = (0, 0.446, 2, 2, 0.942, 2, 0.348, 2, 2, 0.446, 2, 2, 0.942, 2, 0.348, 2, 2, 2, 2, 0.942, 2, 0.446, 2, 2, 0.124, 2, 0.845, 2, 0, 0.845, 2, 0.845, 2, 0.124, 0.845, 2, 0.942, 2, 2, 2, 0.845, 0.079, 2, 2, 0.845)$. (Right) The second centroid tree for Example 2.10. The picture is produced by the `R` package `ape` [23].

Algorithm 2 Experiments on a Sample Generating from Coalescent

Input: $g > 1$ many non-outlier gene trees $T_1, \ldots, T_g$; and $r \geq 1$ many outlier gene trees $T'_1, \ldots, T'_r$. Density Estimator M.

Output: Estimated probabilities for g many non-outlier gene trees and r many outlier gene trees.

for $j = 1, \ldots, r$, **do**

  for $i = 1, \ldots, g$, **do**

    Compute estimated probability $\hat{f}(T_i)$ of T_i via M with a sample of gene trees $\{T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_g, T'_j\}$.

    Compute estimated probability $\hat{f}(T'_j)$ of T'_j via M with a sample of gene trees $\{T_1, \ldots, T_g\}$.

  end for

end for

return $\hat{f}(T_1), \ldots, \hat{f}(T_g)$ and $\hat{f}(T'_1), \ldots, \hat{f}(T'_r)$.

Area Under the Curves (AUCs)

R	0.25	0.5	1	2	5	10
Tropical	0.54	0.61	0.71	0.88	1.00	1
BHV	0.51	0.54	0.54	0.72	0.98	1

TABLE 2

Area Under the Curves (AUCs) for the KDE with the tropical metric and the BHV metric via `KDETrees`.

Therefore, we have estimated probabilities for 1000 trees in $\mathbb{T}_1$ and for 500 trees in $\mathbb{T}_2$.

In this next experiment (Fig. 3, Right), we compare the results against `KDETrees` from [15], [17] with the Billera-Holmes-Vogtmann (BHV) metric [7]. We run computational experiments in MACPRO with 2.4 GHz 8-Core Intel Core i9 processor and 64 GB 2667 MHz DDR4 memory. The computational time for one iteration with our tropical KDE is 9.54 seconds and with `KDETrees` is 1.27 minutes.

3.2 Applications to Apicomplexa Data

In this section we apply a tropical KDE with the HAR algorithm over the space of ultrametrics to the Apicomplexa dataset which consists of 268 orthologous sequences with eight species of protozoa from [18]. There are eight species in each alignment in the set: *Babesia bovis* (Bb), *Cryptosporidium parvum* (Cp), *Eimeria tenella* (Et) [15], *Plasmodium falciparum* (Pf) [11], *Plasmodium vivax* (Pv), *Theileria annulata* (Ta), and *Toxoplasma gondii* (Tg). An outgroup is a free-living ciliate, *Tetrahymena thermophila* (Tt).

Apicomplexa gene sets identified as outliers by KDE with the tropical metric

#	Gene ID	Function
691	PFA0310c	calcium-transporting ATPase
566	PF13_0257	glutamate–tRNA ligase
650	PF11_0358	DNA-directed RNA polymerase, beta subunit, putative
730	PFL0930w	clathrin heavy chain, putative
615	PF13_0063	26S proteasome regulatory subunit 7, putative
712	MAL13P1.274	serine/threonine protein phosphatase pfPp5
630	PFL2120w	hypothetical protein, conserved
625	PFD1090c	clathrin assembly protein, putative
755	PF10_0148	hypothetical protein
708	PFC0140c	N-ethylmaleimide-sensitive fusion protein, putative
497	PF13_0228	40S ribosomal subunit protein S6, putative
690	MAL8P1.134	hypothetical protein, conserved
503	PF13_0178	translation initiation factor 6, putative

TABLE 3

Apicomplexa gene sets identified as outliers by KDE with the tropical metric. All annotations except 728 are putative. Based on the gene set designations in [18]. Gene set represented by GeneID for *P.falciparum*.

The gene trees in the 0.05 lower tail of the estimated distribution of gene trees using the tropical KDE are trees with their IDs 691, 566, 650, 730, 615, 712, 630, 625, 755, 708, 497, 690, 503 (ordered by the smallest probabilities to the largest). Details of these outlying gene trees can be found in Table 3.

In addition, we visualize the distribution of gene trees of Apicomplexa using tropical principal component analysis (PCA) developed by Yoshida et al. [6] via the best-fitted tropical triangle, principal tropical polytope, shown in Fig. 4. Suppose we have an $s \times e$ matrix $\mathcal{D}$ whose rows represent vertices of the best-fitted tropical polytope over $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ via the tropical PCA. Recall that the tropical convex hull of the rows of $\mathcal{D}$ is isometric (linear translation) to the tropical convex hull of the columns of $\mathcal{D}$ [20, Theorem 5.2.21]. In our case we have $s = 3$ and $e = \binom{8}{2} = 28$. Therefore, these unfilled circles in Fig. 4 represent the columns of $\mathcal{D}$ whose rows are the vertices of the best-fitted tropical triangle for

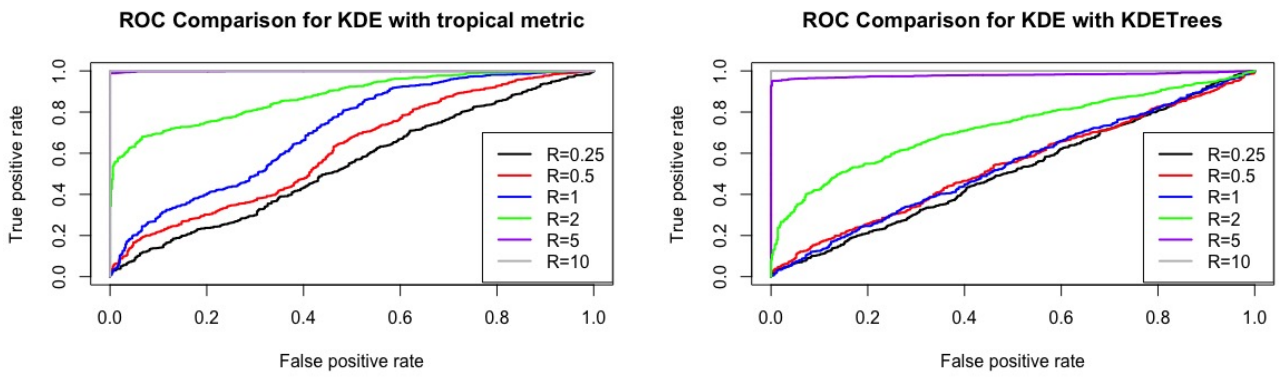

Fig. 3. (Left) ROC curves for the KDE with the tropical metric. (Right) ROC curves for `KDETrees` [15], [17].

Apicomplexa data set from [18]. Filled black points in Fig. 4 represent "outlying" gene trees. The bottom of Fig. 4 shows tree topologies which appear most frequently in the best-fit tropical triangle estimated. The number next to each tree topology in the plot is the number of observations appear in the best-fit tropical triangle.

4 Discussion

4.1 Simulation Study

In [15], [17], Weyenberg et al. showed that `KDETrees` outperformed software `Phylo-MCOA`. In general it works well. However, the biggest problem when using `KDETrees` is that the normalizing constant, $C(T_i)$, of the function $k_{\text{BHV}}(T, T_i)$ for all $T_i \in \mathcal{T}_m$ varies. However, using the tropical metric it seems that we do not have the same issue. From the computation using the HAR sampler from [16], we estimate that C, the normalizing constant of $kK(K.T_i)$, is constant for all $T_i \in \mathcal{U}_m$ while with the BHV metric, the normalizing constant $C(T_i)$ for $kK_{rmBHV}(T, T_i)$ varies for $T_i \in \mathcal{T}_m$. Therefore, we do not have to compute the normalizing constant for each observation in $\mathcal{S}$ when using the tropical metric in order to estimate the gene tree distribution from a sample $\mathcal{S} \subset \mathcal{U}_m$. This is not the case when using the BHV metric, since the normalizing constant varies with each T_i, requiring computation for each sample. This makes the computational time for estimating the gene tree distribution much faster with the tropical metric as compared with using the BHV metric.

Weyenberg et al. in [17] estimate the normalizing constant $C(T_i)$ for each observation T_i in a sample by using the *cone* distance between two trees on the BHV coordinates where one tree in the BHV coordinates goes through on the straight line to the origin (the star tree) and then goes through on the straight line to the other tree. This can lead to large errors in the estimation and may affect the performance of `KDETrees`. As we can see from Table 2 and Figure 3, our proposed non-parametric estimation of the gene tree distribution outperforms `KDETrees` proposed by Weyenberg et al. [15], [17] for all $R = 0.25, 0.5, 1, 2, 5, 10$.

4.2 Apicomplexa

In this section we summarize the analysis on outliers identified from the non-parametric estimation of the gene tree distribution we propose in this research. Sequence alignments used to derive gene trees were judged to be poor if gene annotation errors were evident and likely reduced the accuracy of the alignment. Here we have Pf = *Plasmodium falciparum*, Pv = *Plasmodium vivax*, Bb = *Babesia bovis*, Ta = *Theileria annulata*, Et = *Eimeria tenella*, Tg = *Toxoplasma gondii*, Cp = *Cryptosporidium parvum*, and Tt = *Tetrahymena thermophila* (outgroup).

- **PFA0310c**: Generally good alignment of sequences. The tree topology is mostly consistent with species phylogeny, except Tg and Et are clustered with the outgroup Tt rather than the expected Cp.
- **PF13_0257**: Poor alignment in the N-terminal portion of the sequences. Long C-terminal extension in the outgroup Tt. There are several anomalies in the tree topology. The outgroup Tt clustered with the piroplasms Ta and Bb. The intestinal parasite Cp clustered with malaria parasites Pv and Pf.
- **PF11_0358**: Good sequence alignment in blocks. Longer sequences for the malaria parasites Pf and Pv, including aN-terminal extension and several internal insertions. These potentially reflect incorrect gene annotation. Pf and Pv branch deeper than the Tt outgroup branch.
- **PFL0930w**: Good sequence alignment in blocks, but with multiple assorted insertions in the gene for different taxa. The tree topology is inconsistent with phylogeny. The outgroup Tt branched internally and clustered with the coccidian parasites Tg and Et.
- **PF13_0063**: Overall good sequence alignment. The protein horter sequence for Et is shorter. There is a 50 amino acid repetitive insertion in Et, possibly reflecting a gene annotation error. The tree is generally consistent with phylogeny. The intestinal parasite Cp is on the basal branch with outgroup Tt.
- **MAL13P1.274**: Good alignment in the C-terminal half of the protein sequences. Inconsistent alignment in the N-terminal half with an approximately100 amino acid. extension in Pf and Pv. The outgroup Tt clustered with the malaria parasites Pf and Pv; otherwise, the tree topology is largely consistent with phylogeny.

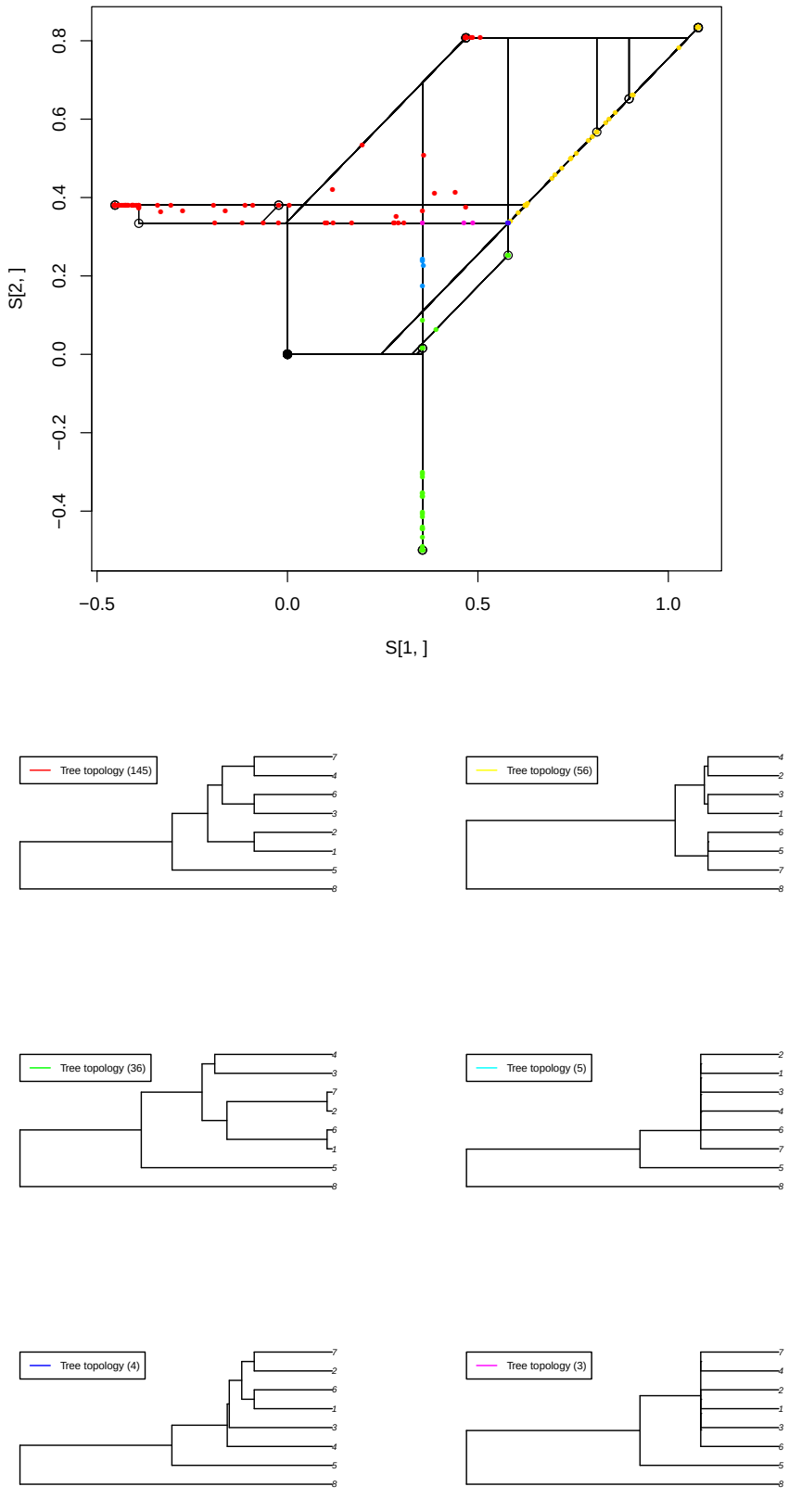

Fig. 4. Estimated tropical PCA developed by Yoshida et al. in [6] using the Hit and Run algorithm developed by Yoshida et al. [16] with the Apicomplexa dataset from [18]. In the figures above, a leaf label 1 is for the label "Pv", a leaf label 2 is for the label "Pf", a leaf label 3 is for the label "Tg", a leaf label 4 is for the label "Et", a leaf label 5 is for the label "Cp", a leaf label 6 is for the label "Ta", a leaf label 7 is for the label "Bb", a leaf label 8 is for the label "Tt", the outgroup. (Top) 1000 iterations were conducted for each vertex of the principal tropical triangle. The black points represent "outlying" gene trees. (Bottom) the top 6th tree topologies of the projected observations onto the best-fit tropical triangle after 1000 iterations. The number inside of the parentheses for each tree topology is the number of projected trees which have the particular tree topology. The sum of residuals is 307.682.

- **PFL2120w**: Poor sequence alignment, with multiple sequence insertions in different species. The intestinal parasite Cp clustered with the piroplasms Bb and Ta. The outgroup Tt clustered with the malaria parasites Pf and Pv.
- **PFD1090c**: Good sequence alignment. There is a long N-terminal extension in Et with homopolymeric stretches, likely reflecting incorrect gene annotation. The tree is very inconsistent with phylogeny. The piroplasm Ta clustered with the intestinal parasite Cp. The piroplasm Bb clustered with the malaria parasites Pv and Pf. The outgroup Tt is located on an internal branch with the coccidian parasites Tg and Et.
- **PF10_0148**: Generally good alignment in the N-terminal half of the proteins. Insertion present in the gene sequences for Pf and Pv. There is an approximate 100 residue C-terminal extension in Tg. The tree topology is inconsistent with phylogeny. The intestinal parasite Cp clustered with the malaria parasites Pf and Pv. The outgroup Tt clustered with the coccidian parasites Tg and Et.
- **PFC0140c**: Good alignment in the central portion of the gene sequences. The gene sequence for Et is much shorter. The coccidian parasite Tg clustered with the malaria parasites Pf and Pv.
- **PF13_0228**: The sequence for the outgroup Tt is much longer than all others with long N- terminal and C-terminal extensions. There is very good sequence alignment in blocks, but with lengthy insertions for the outgroup Tt, possibly reflecting incorrect annotation of the gene. The piroplasms Bb and Ta do not form a monophyletic taxon.
- **MAL8P1.134**: Good alignment in blocks. There is a much longer sequence for the outgroup Tt. in the malaria parasites Pf and Pv share a sequence insertion. The tree has the malaria parasites Pf and Pv clustered with the coccidian parasites Tg and Et.
- **PF13_0178**: Good sequence alignment. The tree has Tg and Et branched as the basal taxa.

5 Conclusion

From computational experiments, the tropical metric outperforms the BHV metric proposed by Weyenberg et al. [15], [17] when using this approach for a non-parametric estimation of the gene tree distribution in terms of accuracy and computational time. Therefore, we intend to extend our methods to large-scale codivergence studies that will describe the tree space encompassing such ancestral gene pools. After doing so, outliers in that tree space will represent such events in genome evolution as gene duplications, lateral gene transfer between species, retention of ancestral polymorphisms by balancing selection, or accelerated evolution by neofunctionalization. Even phylogenetic codivergence of regions within enzyme sequences are of interest due to the possibility of module or domain shuffling in gene evolution. Outlier trees may represent erroneous gene models, correction of which can enhance genome annotations; or they may represent genes with unusual evolutionary histories caused by horizontal gene transfer, trans-species (ancient) polymorphisms, or accelerated evolution due to positive selection and neofunctionalization.

With the combination of visualization via tropical principal component analysis (PCA) developed by Yoshida et al. [6], we can see how gene trees in a given sample are distributed over the space of phylogenetic trees. Fig. 4 shows the visualization via the tropical PCA with annotations of outlying gene trees written in black. From Fig. 4, it seems that all outlying gene trees are projected onto the same point in the two dimensional tropical triangle. It is not clear whether this is a unique case or it happens often.

It is well-known that if we reconstruct a phylogenetic tree from a concatenated alignment from gene alignments (for example, [25]), an estimated phylogenetic tree is not statistically consistent. This means that no matter how large an input alignment is, the reconstructed phylogenetic tree from the concatenated alignment might not be converging to the true tree. However, without the methodology proposed in this research, we might be able to obtain an *interval estimation* of a phylogenetic tree from a set of gene trees instead of a point estimation of a tree based on a concatenated alignment from gene alignments.

There are still some open problems. For example, with the Billera-Holmes-Vogtmann metric [17], the normalizing constant $C(T_i)$ for the function $k_{\mathrm{BHV}}(T, T_i)$ varies for $T_i \in \mathcal{T}_m$. While the geodesic between random two trees under the BHV metric over the tree space goes through the origin, i..e., the star tree, with positive probability [26], Yoshida and Cox showed that under the tropical metric, the tropical line segment (geodesic under the tropical metric) between two random trees on the tree space does not go through the origin, the star tree, with probability one [27] if $m \geq 5$. Therefore, for small trees with $m < 5$, the normalizing constant for $k(T, T_i)$ with the tropical metric for $T_i \in \mathcal{U}_m$ might vary depending on their central location $T_i \in \mathcal{U}_m$. However, for $m \geq 5$, the normalizing constant for $k(T, T_i)$ with the tropical metric for $T_i \in \mathcal{U}_m$ seems to be constant for any point in $\mathcal{U}_m$ as we see from the example 2.10 for $m = 10$. However, it is not proven mathematically. Thus, we have the following conjecture:

Conjecture 1. The integration

$$C(T_i) = \int_{\mathcal{U}_m} k(T, T_i) dT$$

is constant for any fixed $T_i \in \mathcal{U}_m$ for $m \geq 5$.

Funding

RY and DB are partially funded by NSF DMS 1916037. KM is partially funded by JSPS KAKENHI 18K11485 and 22H02364.

Acknowledgments

The authors would like to thank all editor and referees for useful comments in order to improve this manuscript.

REFERENCES

[1] C. Ane, B. Larget, D. A. Baum, S. D. Smith, and A. Rokas, "Bayesian estimation of concordance among gene trees," *Mol. Biol. Evol.*, vol. 24, pp. 412–426, 2007.

[2] N. Takahata and M. Nei, "Allelic genealogy under overdominant and frequency-dependent selectionand polymorphism of major histocompatibility complex loci," *Genetics*, vol. 124, pp. 967–978, 1990.

[3] L. Liu and D. K. Pearl, "Species trees from gene trees," *Syst. Biol.*, 2007, in press.

[4] D. Horner and G. Pesole, "Phylogenetic analyses: a brief introduction to methods and their application," *Expert Rev. Mol. Diagn.*, pp. 339–350, 2004.

[5] F. Ardila and C. J. Klivans, "The bergman complex of a matroid and phylogenetic trees. journal of combinatorial theory," *Series B*, vol. 96, no. 1, pp. 38–49, 2006.

[6] R. Yoshida, L. Zhang, and X. Zhang, "Tropical principal component analysis and its application to phylogenetics," *Bulletin of Mathematical Biology*, vol. 81, pp. 568–597, 2019.

[7] L. Billera, S. Holmes, and K. Vogtmann, "Geometry of the space of phylogenetic trees," *Adv Appl Math*, vol. 27, no. 4, pp. 733–767, 2001.

[8] D. Speyer and B. Sturmfels, "Tropical mathematics," *Mathematics Magazine*, vol. 82, pp. 163–173, 2009.

[9] M. Akian, S. Gaubert, N. Viorel, and I. Singer, "Best approximation in max-plus semimodules," *Linear Algebra Appl.*, vol. 435, pp. 3261–3296, 2011.

[10] G. Cohen, S. Gaubert, and J. Quadrat, "Duality and separation theorems in idempotent semimodules," *Linear Algebra Appl.*, vol. 379, pp. 395–422, 2004.

[11] B. Lin, B. Sturmfels, X. Tang, and R. Yoshida, "Convexity in tree spaces," *SIAM Discrete Math*, vol. 3, pp. 2015–2038, 2017.

[12] W. P. Maddison and D. Maddison, "Mesquite: a modular system for evolutionary analysis. version 2.72," 2009, available at http://mesquiteproject.org. [Online]. Available: http://mesquiteproject.org

[13] R. Page, R. Yoshida, and L. Zhang, "Tropical principal component analysis on the space of phylogenetic trees," *Bioinformatics*, vol. 36, no. 17, pp. 4590–4598, 06 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa564

[14] A. Monod, B. Lin, Q. Kang, and R. Yoshida, "Tropical foundations for probability & statistics on phylogenetic tree space," 2019.

[15] G. Weyenberg, P. M. Huggins, C. L. Schardl, D. K. Howe, and R. Yoshida, "kdetrees: non-parametric estimation of phylogenetic tree distributions," *Bioinformatics*, vol. 30, no. 16, pp. 2280–2287, 04 2014. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu258

[16] R. Yoshida, K. Miura, and D. Barnhill, "Hit and run sampler from tropically convex sets," *Algebraic Statistics*, 2023, to appear.

[17] G. Weyenberg, R. Yoshida, and D. Howe, "Normalizing kernels in the Billera-Holmes-Vogtmann treespace," *IEEE ACM T. Comput. Bi.*, p. doi:10.1109/TCBB.2016.2565475, 2016.

[18] C. Kuo, J. P. Wares, and J. C. Kissinger, "The apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees," *Mol Biol Evol*, vol. 25, no. 12, pp. 2689–2698, 2008.

[19] M. Joswig, *Essentials of tropical combinatorics*, ser. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2021.

[20] D. Maclagan and B. Sturmfels, *Introduction to Tropical Geometry*, ser. Graduate Studies in Mathematics. Providence, RI: Graduate Studies in Mathematics, 161, American Mathematical Society, 2015, vol. 161.

[21] P. Buneman, "A note on the metric properties of trees," *J. Combinatorial Theory Ser. B.*, vol. 17, pp. 48–50, 1974.

[22] J. Tukey, *Exploratory Data Analysis*. Boston, MA: Addison-Wesley, 1977.

[23] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, pp. 289–290, 2004.

[24] B. Rannala, S. V. S. V. Edwards, A. Leaché, and Z. Yang, "The Multi-species Coalescent Model and Species Tree Inference," in *Phylogenetics in the Genomic Era*, C. Scornavacca, F. Delsuc, and N. Galtier, Eds. No commercial publisher — Authors open access book, 2020, pp. 3.3:1–3.3:21. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02535622

[25] L. S. Kubatko and J. H. Degnan, "Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence," *Systematic Biology*, vol. 56, no. 1, pp. 17–24, 02 2007. [Online]. Available: https://doi.org/10.1080/10635150601146041

[26] M. Owen and S. Provan, "A fast algorithm for computing geodesic distances in tree space," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 8, pp. 2–13, 2011.

[27] R. Yoshida and S. Cox, "Tree topologies along a tropical line segment," *Vietnam Journal of Mathematics*, vol. 50, pp. 395–419, 2022.

Ruriko Yoshida Ruriko Yoshida is a Professor of Operations Research at the Naval Postgraduate School. Her research topics cover a wide variety of areas: applications of algebraic combinatorics to statistical problems, such as goodness of fit tests, optimized camera placement in sensor networks, applications of online optimization to UxVs, statistical learning using networks. She received her Ph.D. (2004) in Mathematics from the University of California, Davis. She then went to the University of California, Berkeley as a postdoctoral researcher, and then Duke University for her postdoctoral research from 2004 to 2006. She was at the University of Kentucky from 2006 to 2016 as an assistant and then as associate professor. In 2016, she joined the operations research department at the Naval Postgraduate School.

David Barnhill Commander David Barnhill is an officer in the U.S. Navy and an assistant professor at the department of mathematics in the United States Naval Academy. He earned his bachelor of science degree in history from the United States Naval Academy in 2005 and a master of science degree in operations research from NPS in 2021. He earned his Ph.D. degree in operations research from the Naval Postgraduate School (NPS) in 2024. His Ph.D. research touches on various aspects of algebraic geometry, focusing on supervised and unsupervised statistical and machine learning techniques applied to tropically convex sets.

Keiji Miura Keiji Miura received the Ph.D. degree in science from Kyoto University, Kyoto, Japan, in 2006. From April 2006 to September 2008, he was a JSPS Research fellow with the University of Tokyo. From October 2008 to March 2011, he was a JST PRESTO researcher with Harvard University. From April 2011 to March 2015, he was an assistant professor with Tohoku University. From April 2015 to March 2021, he was an associate professor with Kwansei Gakuin University. Since 2021 he has been working as a professor with Kwansei Gakuin University. His research area is mathematical neuroscience.

Daniel Howe Daniel Howe received the PhD degree in molecular parasitology from Purdue University in 1992. He was at the Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, as a post-doctoral research associate from 1993 to 1999. Since 1999, has worked in the Department of Veterinary Science, MH Gluck Equine Research Center, University of Kentucky. He is currently a professor in the Department of Veterinary Science.