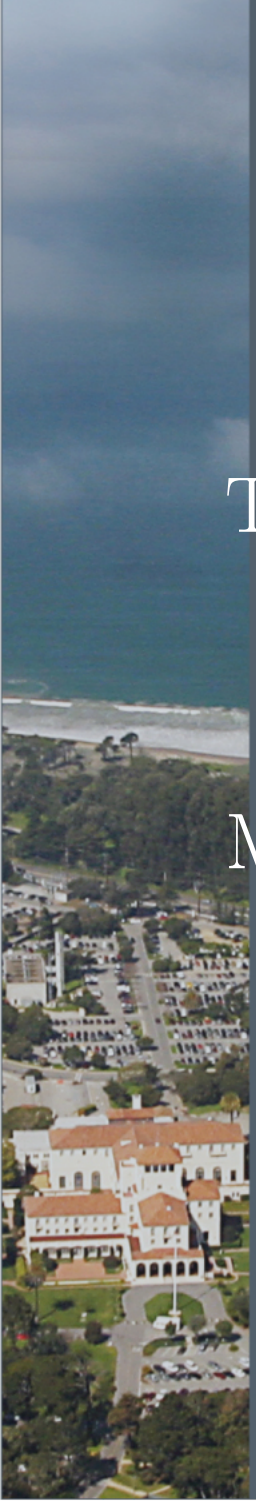


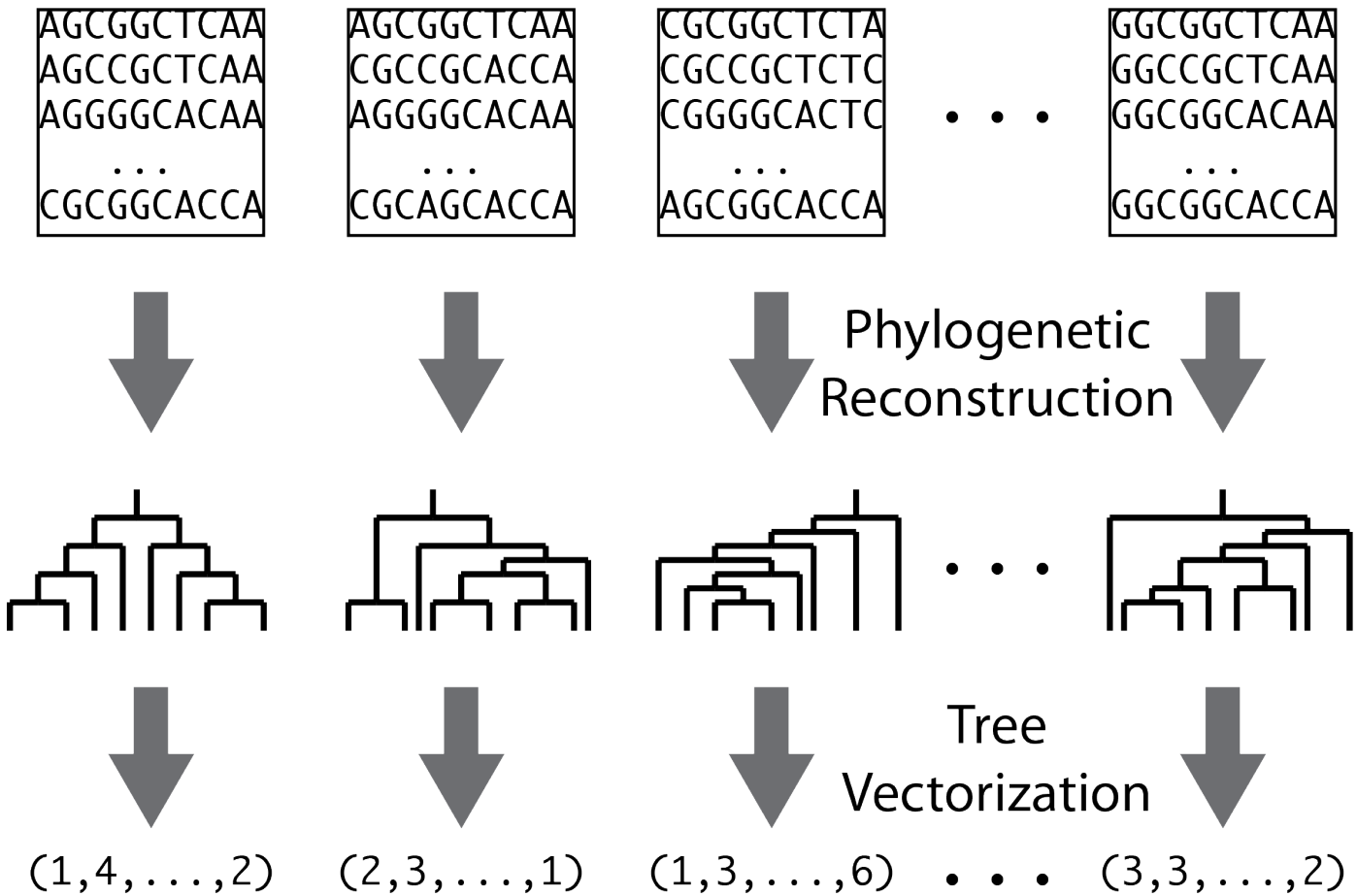
Tropical Geometry Tools for Machine Learning and Phylogenomics

From Phylogenetics to Phylogenomics: Mathematical and Statistical Challenges in the Era of Big Data

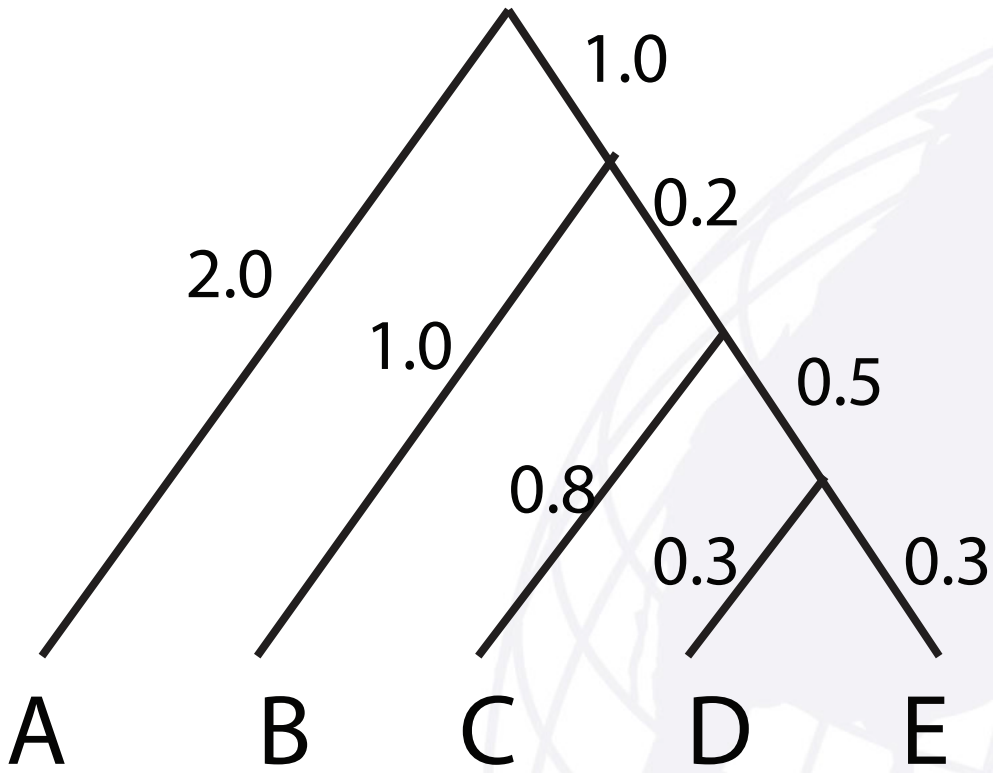
Ruriko Yoshida

Naval Postgraduate School

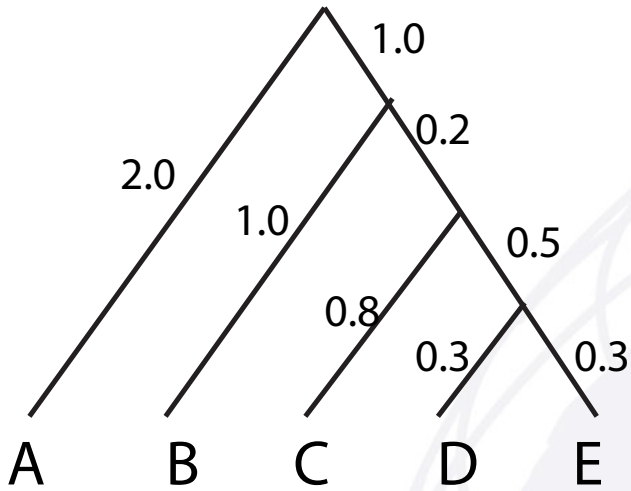
Phylogenetics to Phylogenomics



Equidistance tree

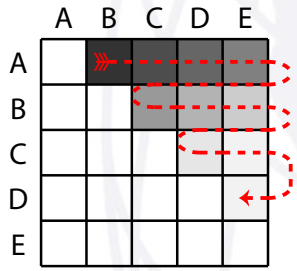


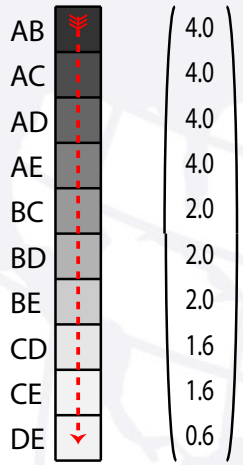
Vectorize a tree



Dissimilarity Map

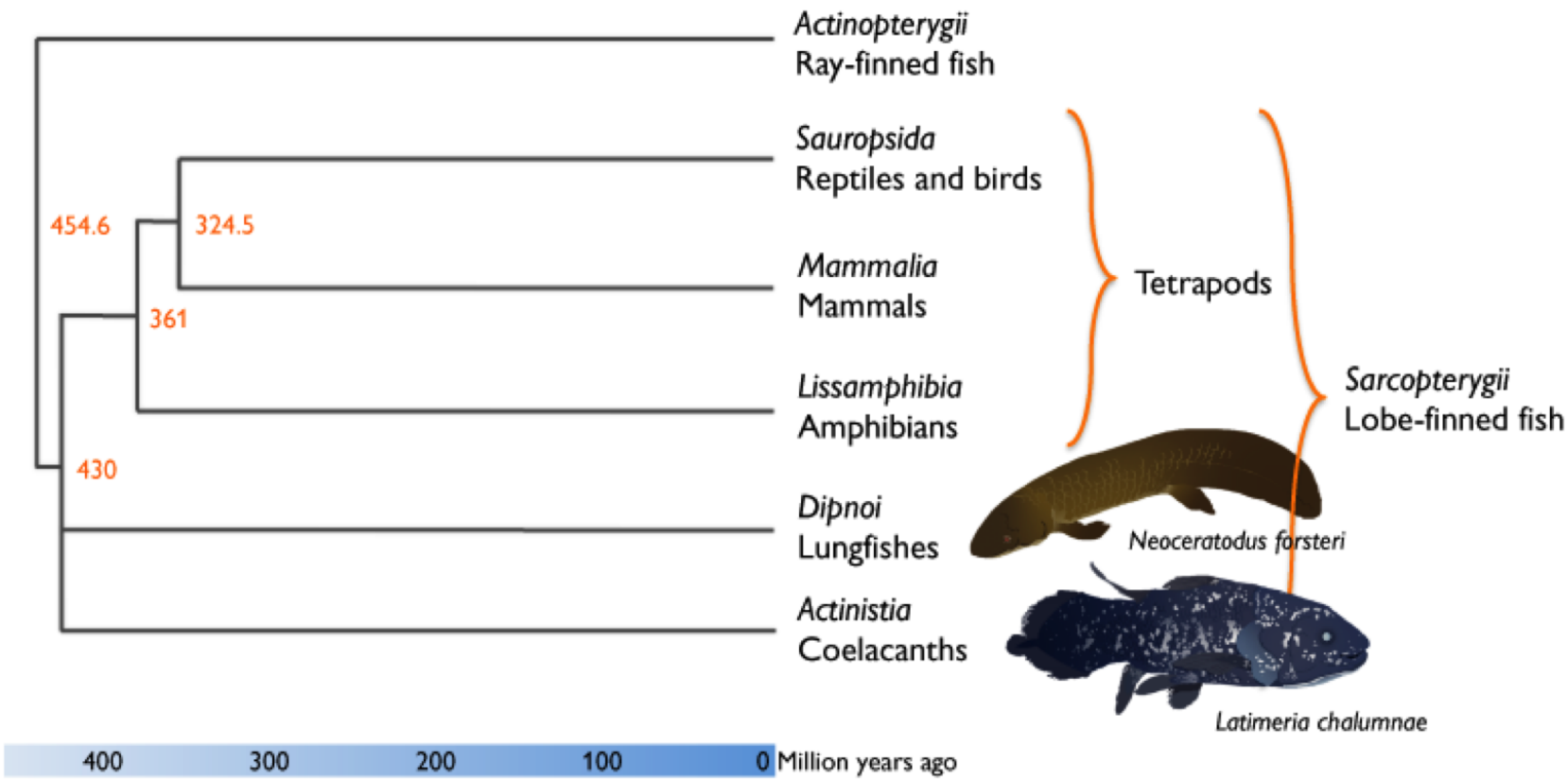
	A	B	C	D	E
A	0	4.0	4.0	4.0	4.0
B	4.0	0	2.0	2.0	2.0
C	4.0	2.0	0	1.6	1.6
D	4.0	2.0	1.6	0	0.6
E	4.0	2.0	1.6	0.6	0





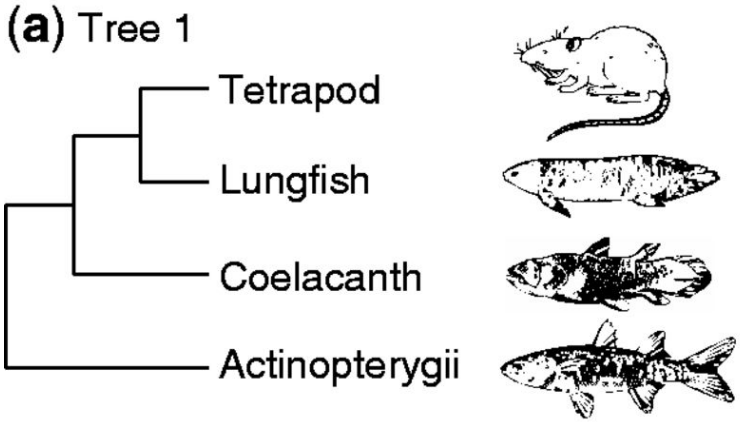
4.0
4.0
4.0
4.0
2.0
2.0
2.0
1.6
1.6
0.6

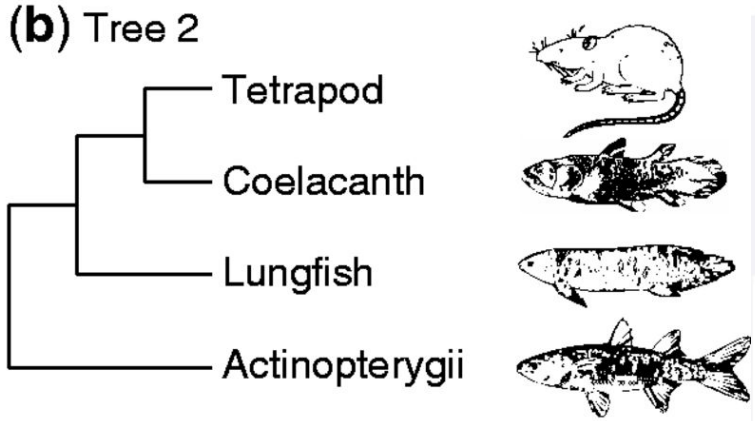
Coelacant data

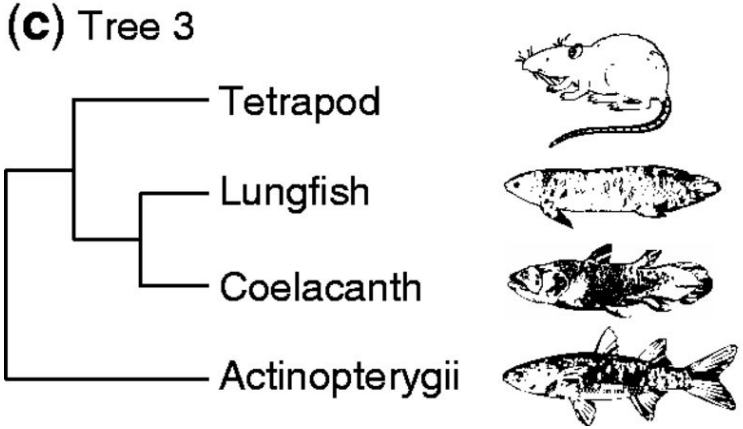


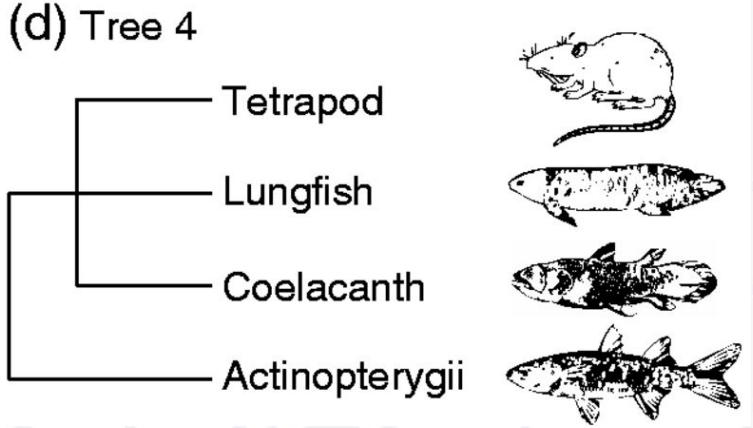
Coelacant data

They have used the data set of 1290 gene trees with 10 species to reconstruct a phylogenetic trees from the concatenated data set. There are about 25% of them are falling toint each category.

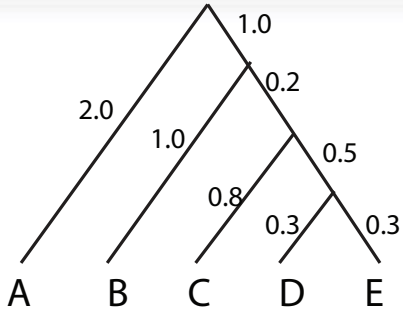




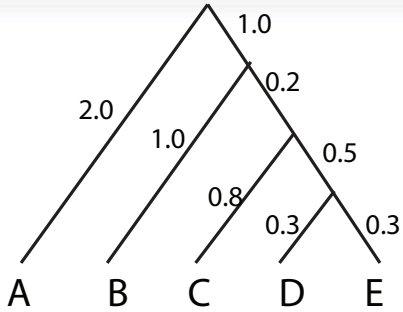




Vectorize a tree



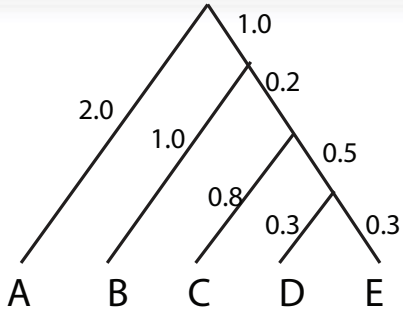
Vectorize a tree



Remark

A distance matrix D realizes an equidistant tree iff D is an ultrametric, i.e. $\max\{D_{ij}, D_{jk}, D_{ik}\}$ for distinct leaves i, j, k achieves at least twice.

Vectorize a tree

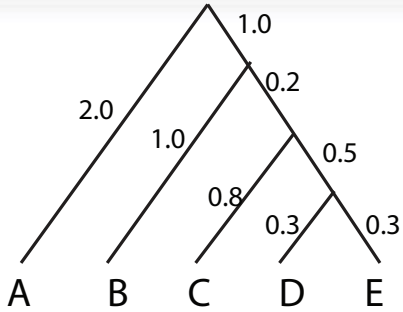


Remark

A distance matrix D realizes an equidistant tree iff D is an ultrametric, i.e. $\max\{D_{ij}, D_{jk}, D_{ik}\}$ for distinct leaves i, j, k achieves at least twice.

Thus, we consider the space of all ultrametrics as a tree space. This space is an union of lower dimensional polyhedral cones.

Vectorize a tree



Remark

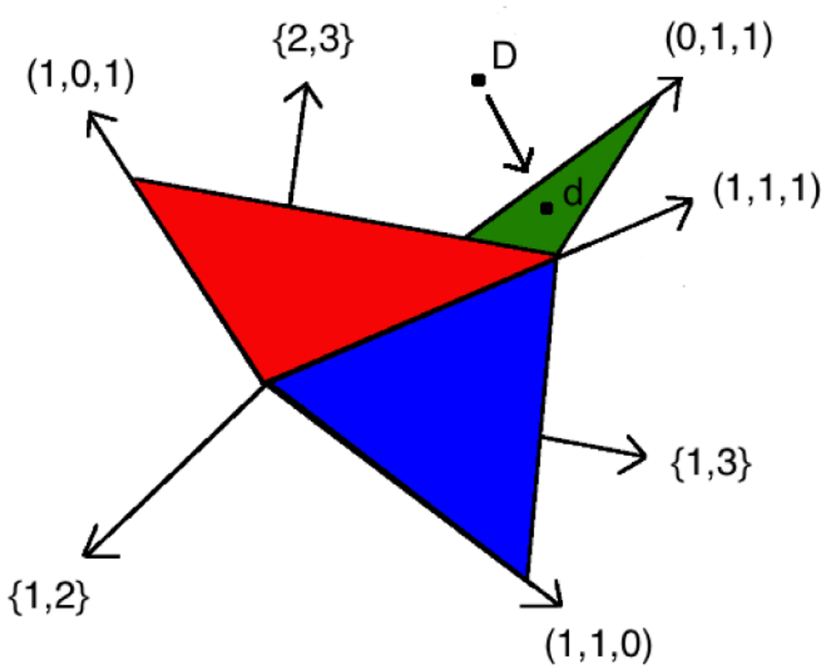
A distance matrix D realizes an equidistant tree iff D is an ultrametric, i.e. $\max\{D_{ij}, D_{jk}, D_{ik}\}$ for distinct leaves i, j, k achieves at least twice.

Thus, we consider the space of all ultrametrics as a tree space. This space is an union of lower dimensional polyhedral cones.

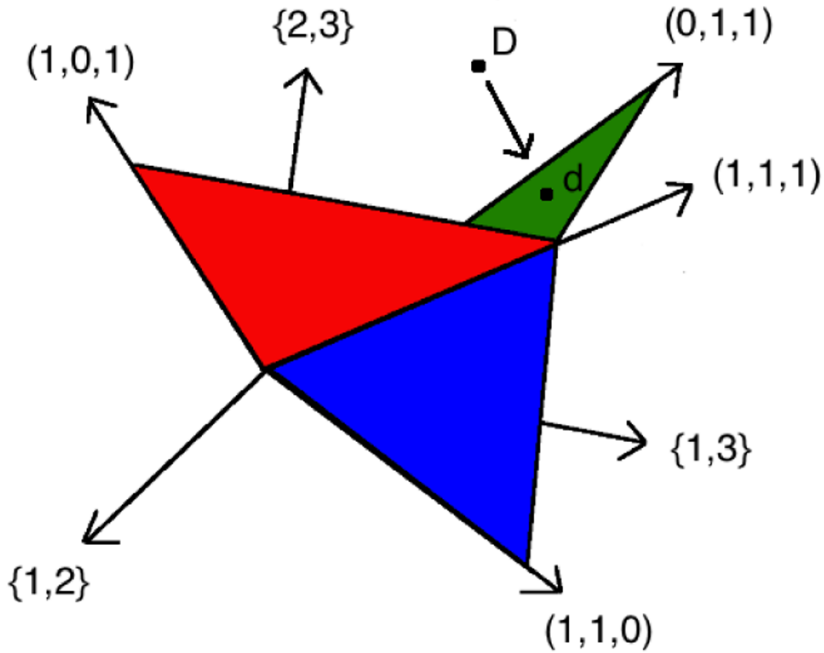
Theorem (Ardila and Klivans (2006) and Page et al. (2020))

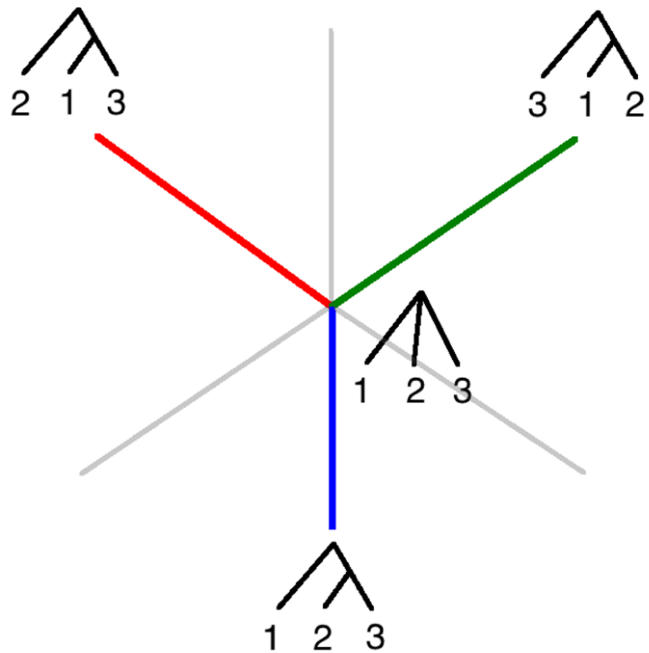
A space of ultrametrics $\mathcal{U}_n$ is the solution set of the tropicalization of linear equations.

Tree Space

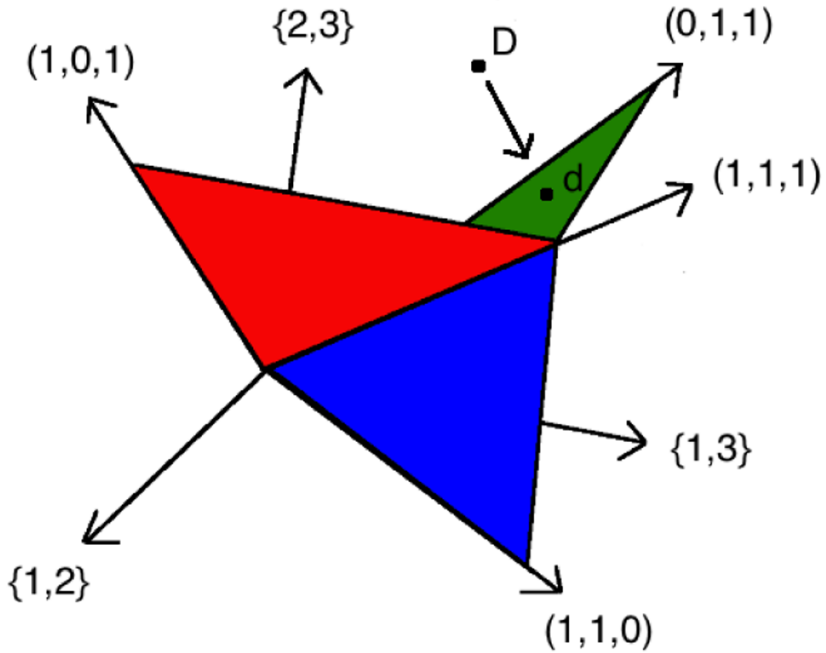


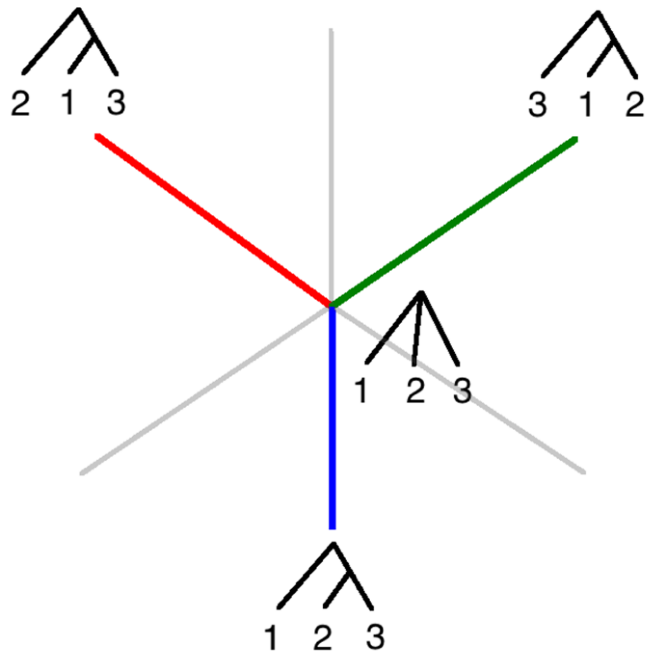
Tree Space





Tree Space

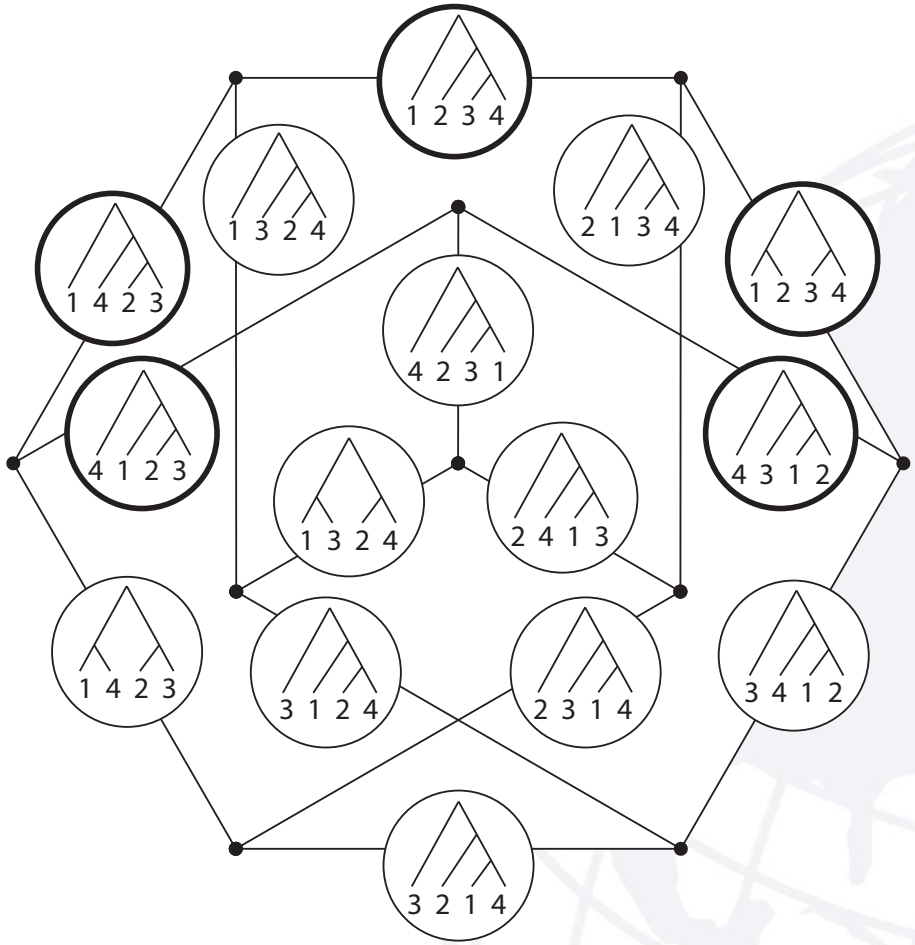




Estimating a phylogenetic tree

Polyhedral cone of equidistant tree metrics on trees with three leaves. A least squared method projects a distance matrix down to the tree space.

Treespace


1 2 3 4
1 3 2 4
2 1 3 4
1 4 2 3
1 2 3 4
4 2 3 1
4 1 2 3
4 3 1 2
1 3 2 4
2 4 1 3
1 4 2 3
3 1 2 4
2 3 1 4
3 4 1 2
3 2 1 4

Tropical distance

Definition (Max-plus algebra)

Here we use max-plus algebra, i.e.,

$$a \oplus b := \max\{a, b\} \text{ and } a \odot b := a + b.$$

Tropical distance

Definition (Max-plus algebra)

Here we use max-plus algebra, i.e.,

$$a \oplus b := \max\{a, b\} \text{ and } a \odot b := a + b.$$

Definition (Tropical distance)

The tropical distance between two points is computed as follows:

$$d_{\text{tr}}(v, w) \;=\; \max(v - w) - \min(v - w).$$

This metric is also known as the **generalized Hilbert projective metric**.

Tropical distance

Definition (Max-plus algebra)

Here we use max-plus algebra, i.e.,

$$a \oplus b := \max\{a, b\} \text{ and } a \odot b := a + b.$$

Definition (Tropical distance)

The tropical distance between two points is computed as follows:

$$d_{\text{tr}}(v, w) \;=\; \max(v - w) - \min(v - w).$$

This metric is also known as the **generalized Hilbert projective metric**.

Proposition

The function d_{tr} is a well-defined metric on $\mathbb{R}^e/\mathbb{R}\mathbf{1}$.

Tropical polytope

Definition

A subset S of $\mathbb{R}^e$ is called *tropically convex* if it contains the point $a \odot x \oplus b \odot y$ for all $x, y \in S$ and all $a, b \in \mathbb{R}$. The *tropical convex hull* or *tropical polytope* of a given subset $V \subset \mathbb{R}^e$ is the smallest tropically convex subset containing V of $\mathbb{R}^e$.

$$\text{tconv}(V) = \{a_1 \odot v_1 \oplus \cdots \oplus a_r \odot v_r : v_1, \ldots, v_r \in V \text{ and } a_1, \ldots, a_r \in \mathbb{R}\}.$$

Tropical polytope

Definition

A subset S of $\mathbb{R}^e$ is called *tropically convex* if it contains the point $a \odot x \oplus b \odot y$ for all $x, y \in S$ and all $a, b \in \mathbb{R}$. The *tropical convex hull* or *tropical polytope* of a given subset $V \subset \mathbb{R}^e$ is the smallest tropically convex subset containing V of $\mathbb{R}^e$.

$$\text{tconv}(V) = \{a_1 \odot v_1 \oplus \cdots \oplus a_r \odot v_r : v_1, \ldots, v_r \in V \text{ and } a_1, \ldots, a_r \in \mathbb{R}\}.$$

Remark

A tropical line segment between two points is a tropical polytope of two points. TML Demo

Tropical polytope

Definition

A subset S of $\mathbb{R}^e$ is called *tropically convex* if it contains the point $a \odot x \oplus b \odot y$ for all $x, y \in S$ and all $a, b \in \mathbb{R}$. The *tropical convex hull* or *tropical polytope* of a given subset $V \subset \mathbb{R}^e$ is the smallest tropically convex subset containing V of $\mathbb{R}^e$.

$$\text{tconv}(V) = \{a_1 \odot v_1 \oplus \cdots \oplus a_r \odot v_r : v_1, \ldots, v_r \in V \text{ and } a_1, \ldots, a_r \in \mathbb{R}\}.$$

Remark

A tropical line segment between two points is a tropical polytope of two points. TML Demo

Theorem (Monod et al. (2020))

A tropical line segment between two points $u, v \in \mathcal{U}_m$ is an intrinsic geodesic in $\mathcal{U}_m$ and unique.

Example over $\mathbb{R}^3/\mathbb{R}\mathbf{1}$

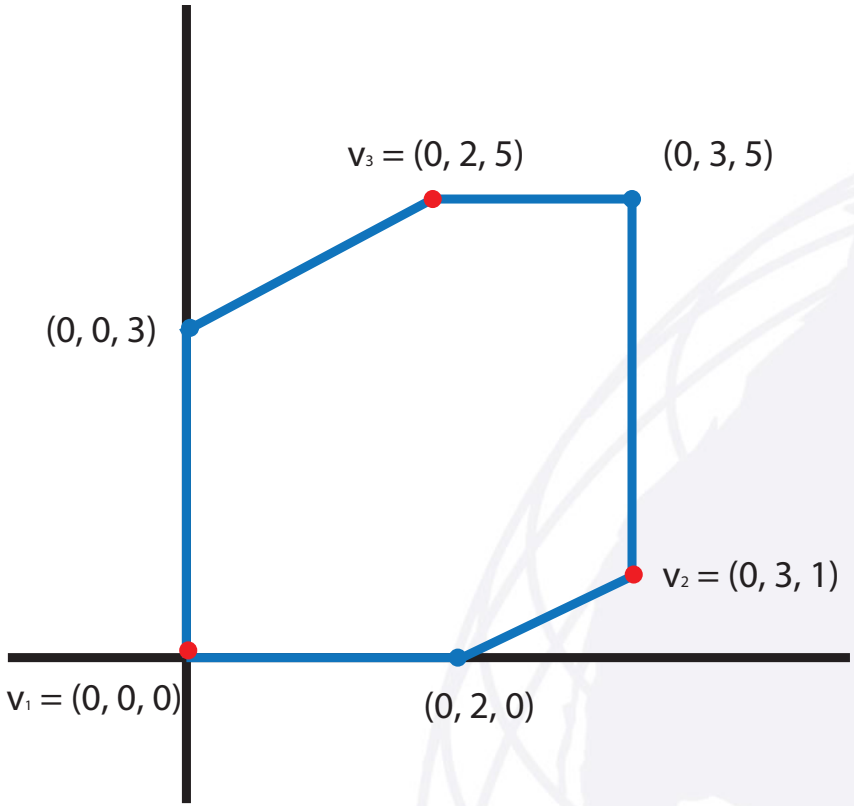
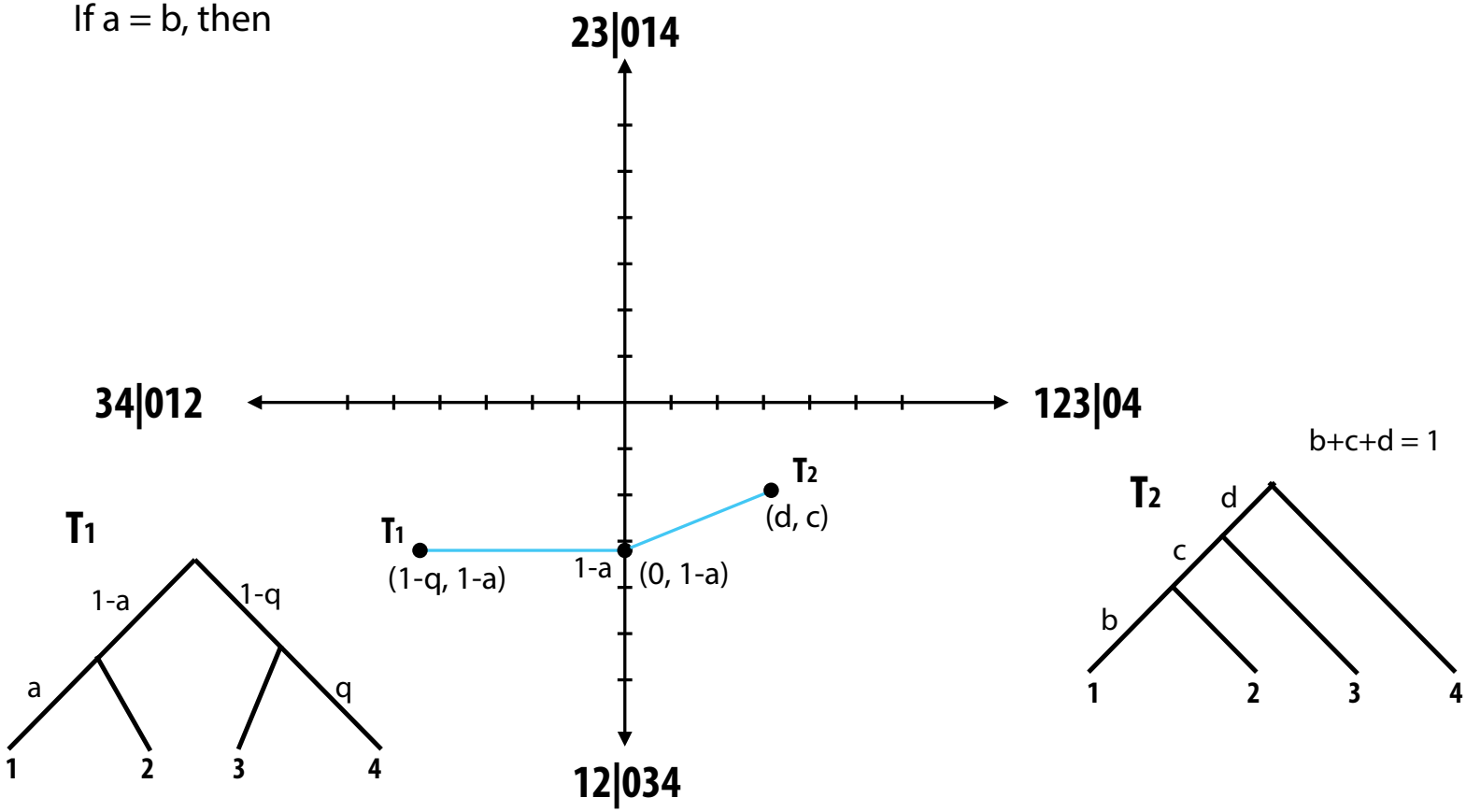


Figure: Tropical polytope of three points $(0,0,0)$, $(0,3,1)$, $(0,2,5)$ in $\mathbb{R}^3/\mathbb{R}\mathbf{1}$.

Example over $\mathcal{U}_4$

If a = b, then



Tropical projection

Proposition

Consider a tropical polytope $\mathcal{P} = \operatorname{tconv}(D^{(1)}, D^{(2)}, \ldots, D^{(s)})$.

$$\pi_{\mathcal{P}}(D) = \lambda_1 \odot D^{(1)} \oplus \lambda_2 \odot D^{(2)} \oplus \cdots \oplus \lambda_s \odot D^{(s)},$$

where $\lambda_k = \min(D - D^{(k)})$.

Tropical projection

Proposition

Consider a tropical polytope $\mathcal{P} = \text{tconv}(D^{(1)}, D^{(2)}, \ldots, D^{(s)})$.

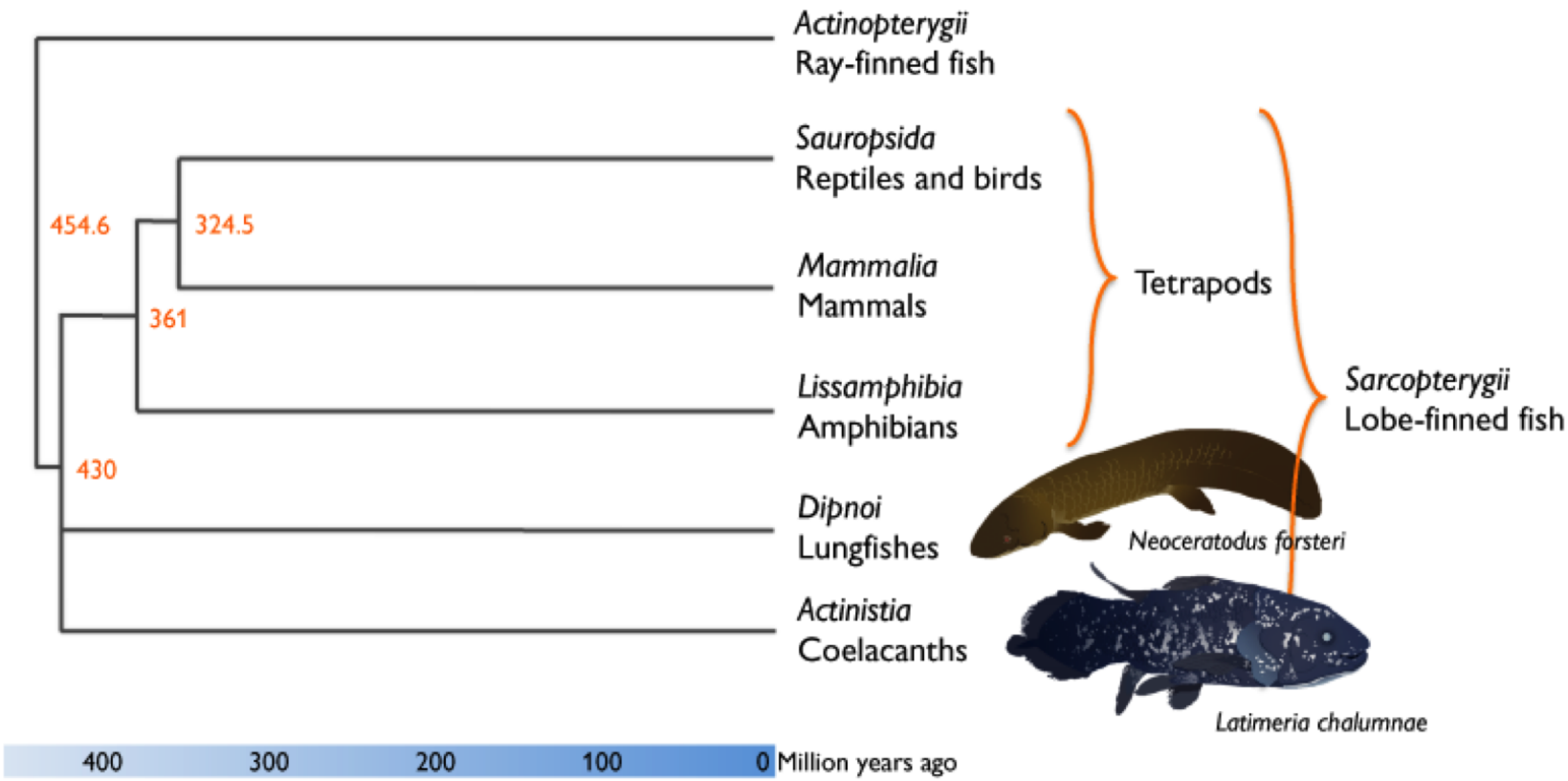
$$\pi_{\mathcal{P}}(D) = \lambda_1 \odot D^{(1)} \oplus \lambda_2 \odot D^{(2)} \oplus \cdots \oplus \lambda_s \odot D^{(s)},$$

where $\lambda_k = \min(D - D^{(k)})$.

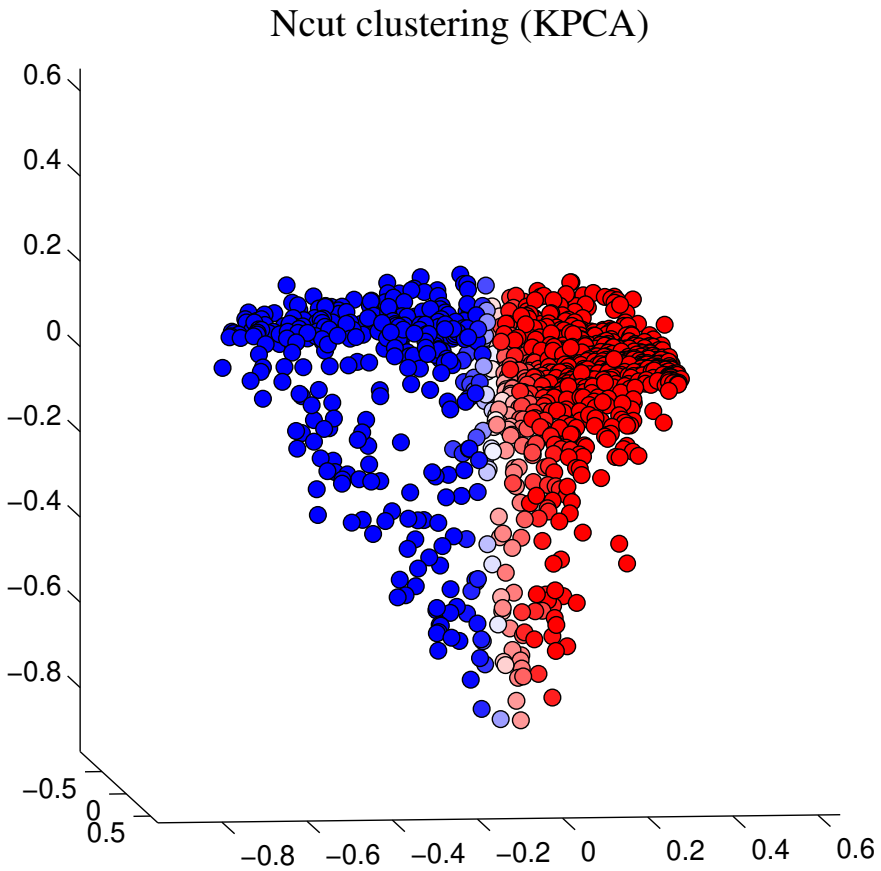
Remark

A tropical triangle in ultrametric tree space $\mathcal{U}_m$ *has dimension has at most dimension* 2. *A tropical convex hull of* s *many trees in ultrametric tree space* $\mathcal{U}_m$ *has dimension has at most dimension* $s - 1$.

Application 1: Tropical PCA

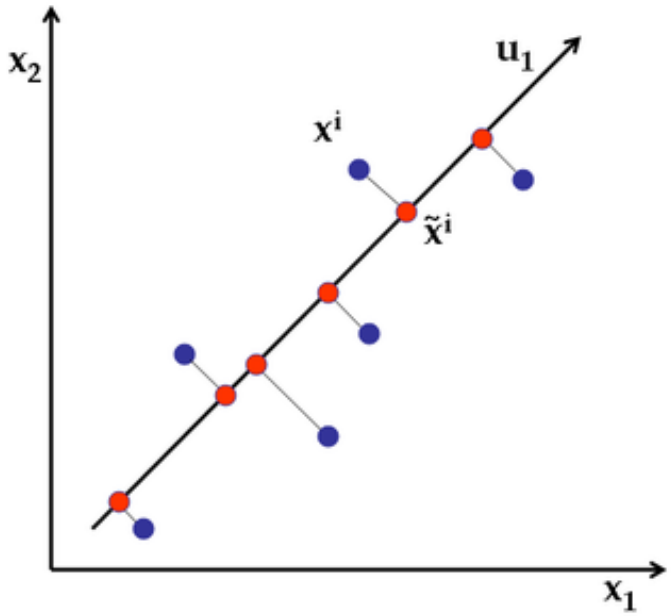


PCA over Euclidean space



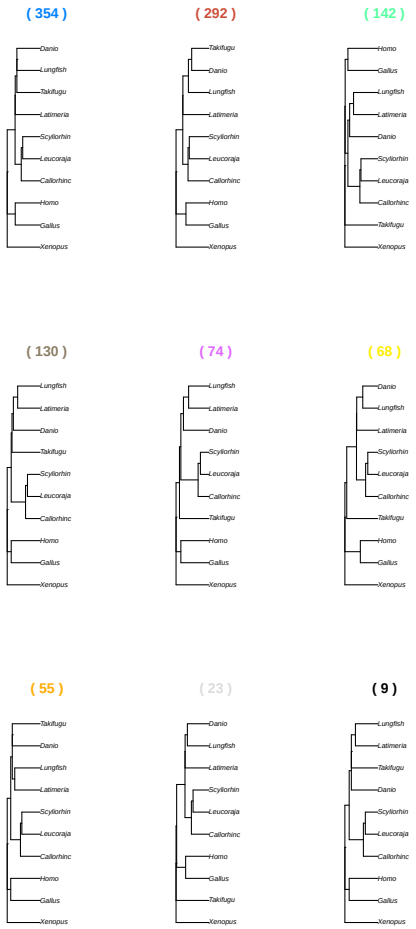
First, we applied PCA dimensionality reduction on the data sets and then we have applied the normalized cut clustering method (R.Y., K. Fukumizu, and C. Vogiatzis, 2017).

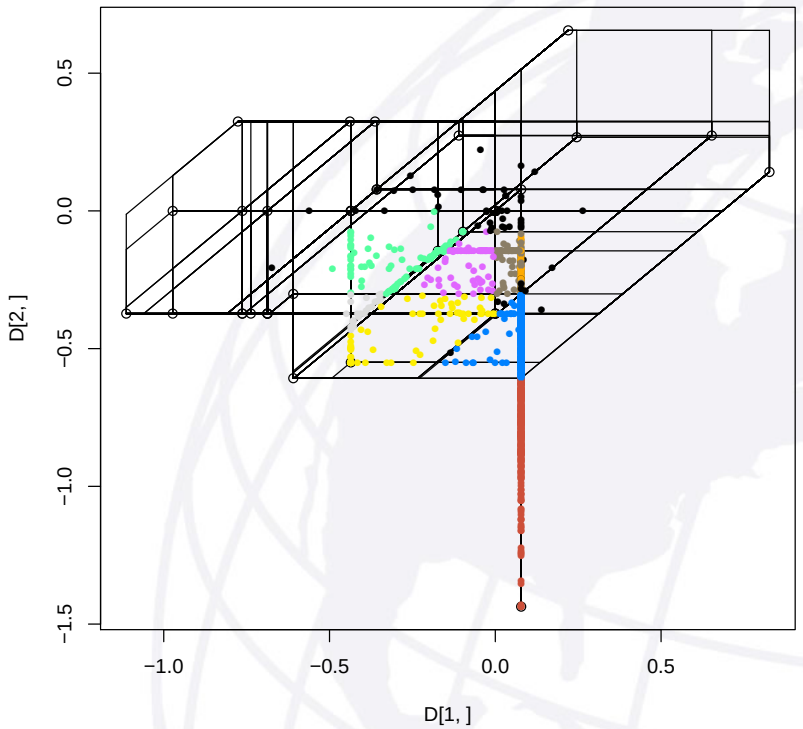
Dimensionality reduction



Given the set of estimated gene trees $S = \{T_1, T_2, \ldots, T_s\}$ for s genes across the genome. We want to estimate a plane/convex hull on a "treespace" such that the sum of distances between each T_i and its projection on the convex hull/plane is minimized.

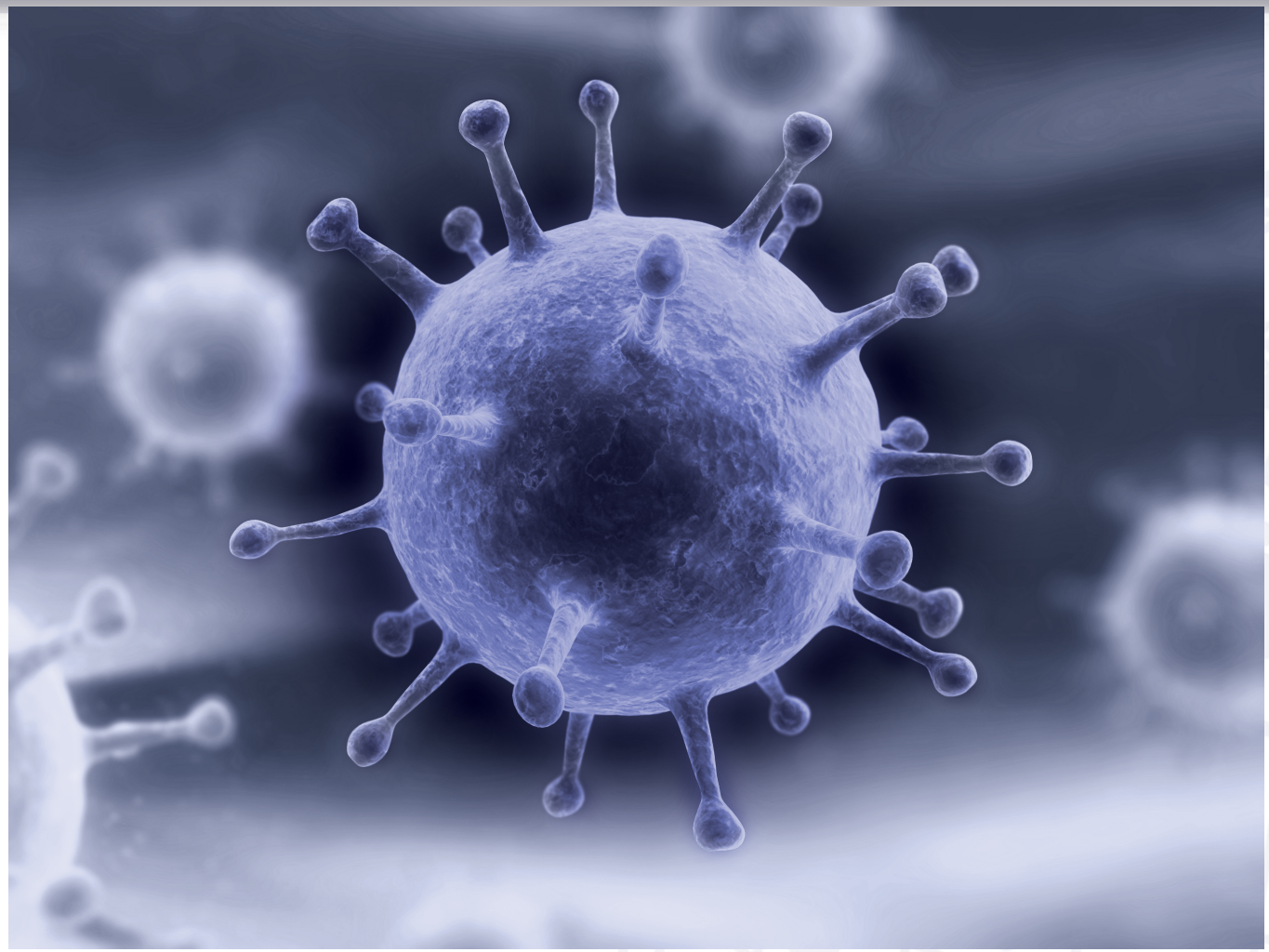
Tropical PCA on Coelacant data





TML Demo

Hemagglutinin (HA) sequences for influenza

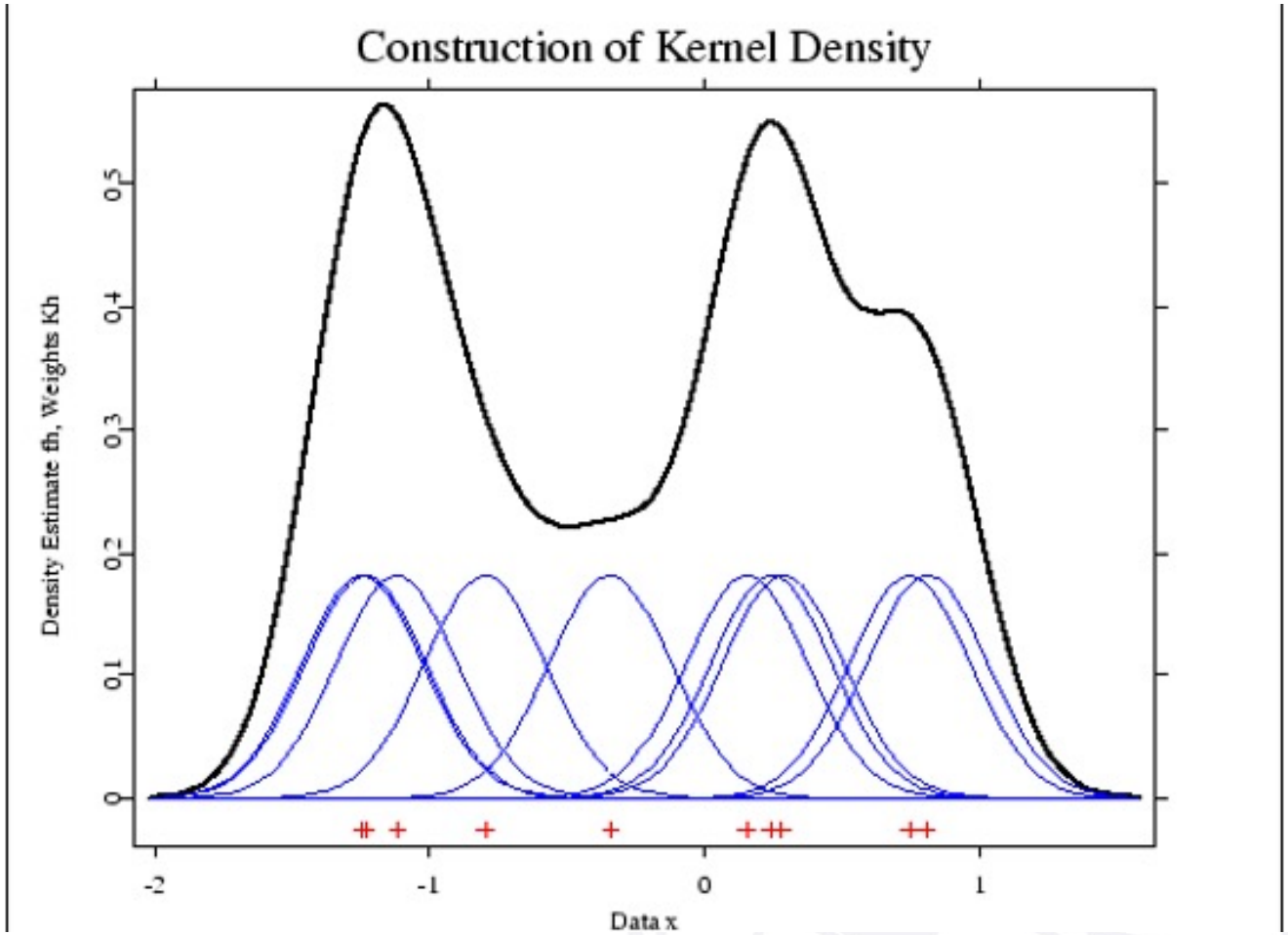


1089 full length sequences of hemagglutinin (HA) for influenza A H3N2 from 1993 to 2017 in the state of New York were obtained from the GI-SAID EpiFlu

Hemagglutinin (HA) sequences for influenza

Year	Tree with 4 leaves		Tree with 5 leaves	
	Tropical PCA	BHV	Tropical PCA	BHV
1993	0.9981(0.9559)	0.7099	0.8962(0.7269)	0.3019
1994	0.9997(0.9426)	0.4611	0.9559(0.8505)	0.4347
1995	0.9999(0.8665)	0.19	0.9787(0.9577)	0.3151
1996	0.9997(0.9821)	0.215	0.9851(0.7482)	0.5025
1997	0.9930(0.9532)	0.0069	0.9430(0.8437)	0.0505
1998	>0.9999(0.9395)	0.0452	0.9264(0.879)	0.6408
1999	>0.9999(0.9069)	0.0038	0.9798(0.8564)	0.9524
2000	0.9892(0.9132)	0.9555	0.9302(0.794)	0.0014
2001	>0.9999(0.9088)	0.9402	0.9526(0.8302)	0.9488
2002	0.9995(0.9863)	0.0107	0.9956(0.9525)	0.8962
2003	0.9995(0.9848)	0.0972	0.9685(0.8622)	0.4927
2004	0.9982(0.9505)	0.4272	0.9502(0.7931)	0.3651
2005	0.9998(0.9949)	0.4628	0.9770(0.8304)	0.3634
2006	0.9972(0.9643)	0.0951	0.8350(0.73)	0.2383
2007	0.9926(0.9381)	0.5562	0.8912(0.6995)	0.2727
2008	0.9920(0.8813)	0.4887	0.7753(0.4637)	0.0460
2009	0.9860(0.8926)	0.0763	0.9034(0.6289)	0.1563
2010	0.9995(0.8886)	0.0329	0.8603(0.6665)	0.1935
2011	0.9999(0.9016)	0.3592	0.6888(0.5920)	0.2771
2012	0.9930	0.2756	0.7177(0.5568)	0.1998
2013	0.9499(0.7935)	0.3612	0.7433(0.5624)	0.1279
2014	0.9727	0.1383	N/A	N/A

Application 2: Tropical KDE



Kernel Density Estimator

Suppose we have an i.i.d. sample of trees
$\mathcal{S} := \{T_1, \ldots, T_N\} \subset \mathcal{U}_m$.
Our non-parametric density estimator with the tropical metric over the space of ultrametrics formulated as

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

where k is a non-negative function defined on trees, such that

$$k(T, T_i) = \frac{1}{C} \exp\left(-\left(\frac{d_{\mathrm{tr}}(T, T_i)}{\sigma}\right)\right),$$

where $C > 0$ **is the normalizing constant, so that** $\sum_T k(T, T_i) = 1$ and $\sigma > 0$ is a "bandwidth" parameter that controls, in the summation formula for $\hat{f}(T)$, how tightly each contribution $k(T, T_i)$ will be centered on T_i.

Simulation Experiment

For this computational experiments, we generate gene trees from the multispecies coalescent models with a given species tree via the software `Mesquite`. We fixed the effective population size $N_e = 100,000$ and varied

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

where SD is the species depth.

Simulation Experiment

For this computational experiments, we generate gene trees from the multispecies coalescent models with a given species tree via the software `Mesquite`. We fixed the effective population size $N_e = 100,000$ and varied

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

where SD is the species depth.
We generate 1000 gene trees for each species tree. In these simulated experiments, we vary the ratio
$R = 0.25, 0.5, 1, 2, 5, 10$.

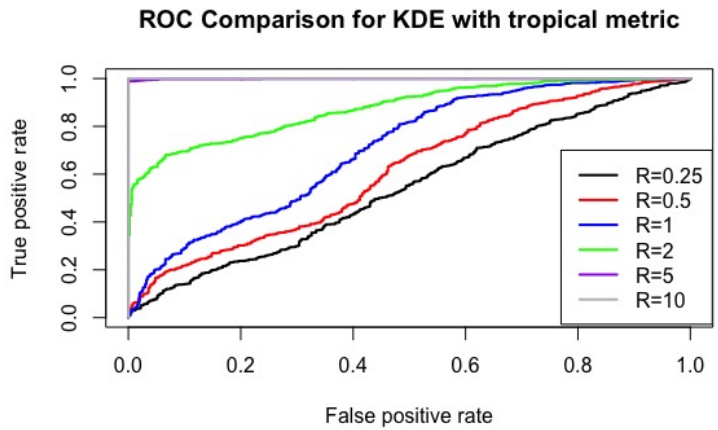
Simulation Experiment

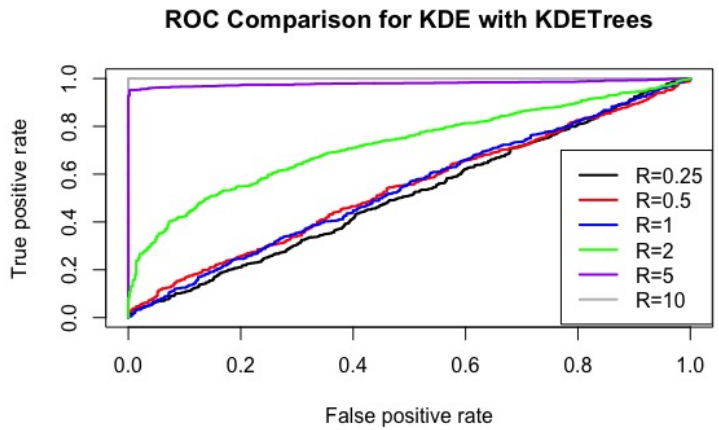
For this computational experiments, we generate gene trees from the multispecies coalescent models with a given species tree via the software `Mesquite`. We fixed the effective population size $N_e = 100,000$ and varied

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

where SD is the species depth.
We generate 1000 gene trees for each species tree. In these simulated experiments, we vary the ratio
$R = 0.25, 0.5, 1, 2, 5, 10$.
Then we estimated the KDE using the tropical KDE and with `KDETrees`. We iterate this process for 500 times. Therefore, we have estimated probabilities for 1000 trees in $\mathbb{T}_1$ and for 500 trees in $\mathbb{T}_2$.

Results

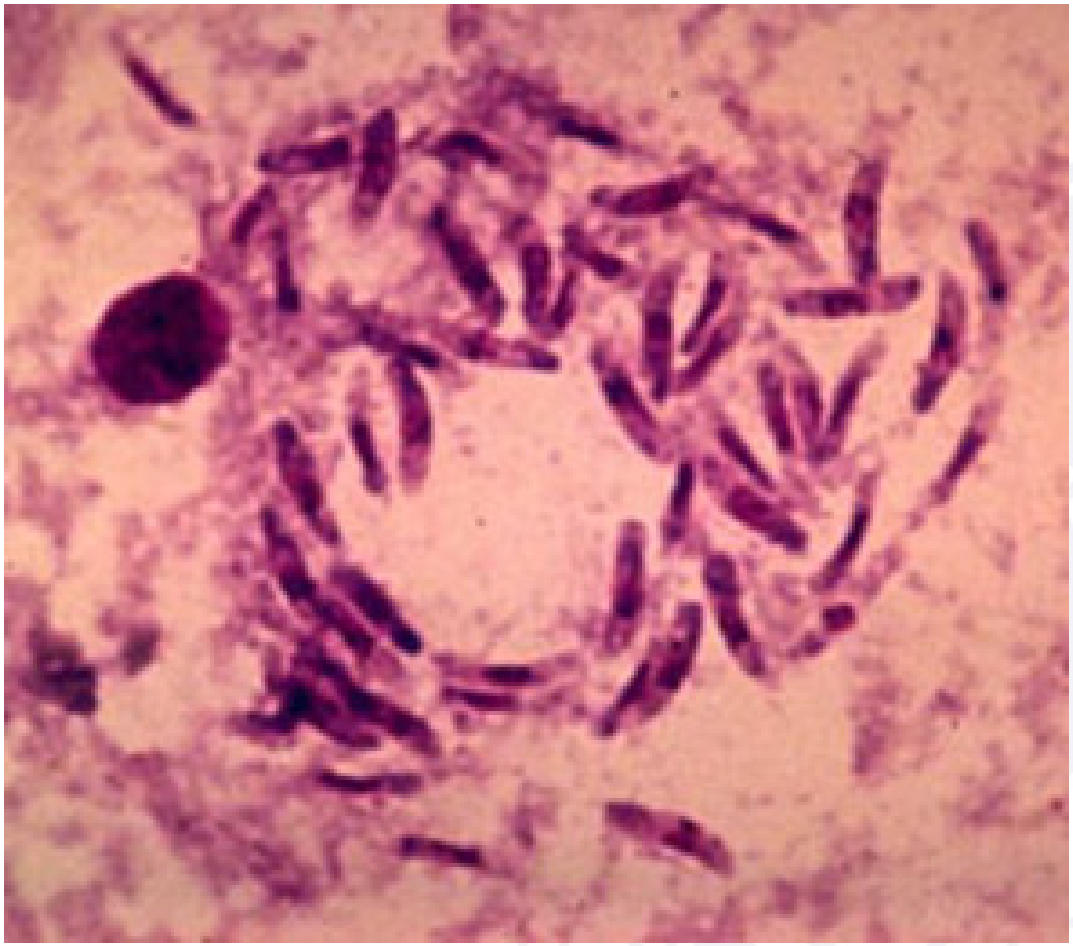




R	0.25	0.5	1	2	5	10
Tropical	0.542	0.614	0.706	0.876	0.999	1
BHV	0.511	0.539	0.537	0.721	0.980	1

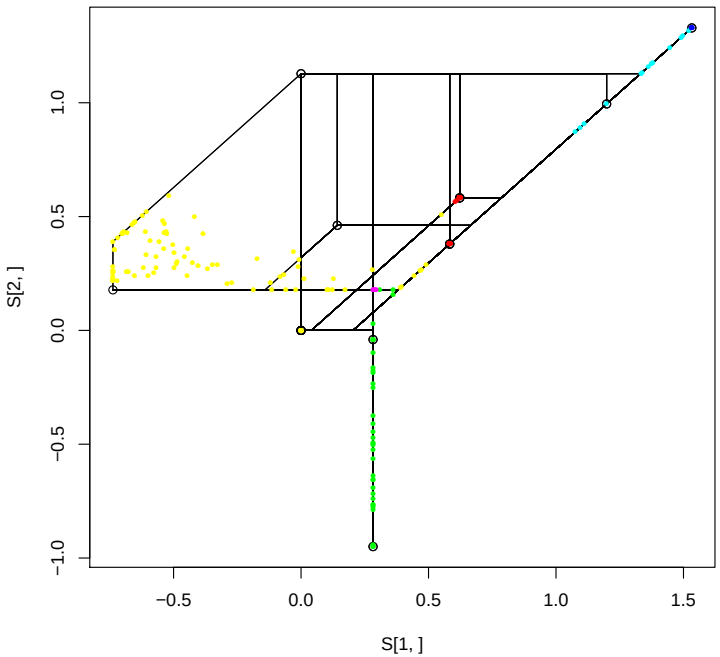
Tropical: 9.54 seconds and `KDETrees`: 1.27 minutes.

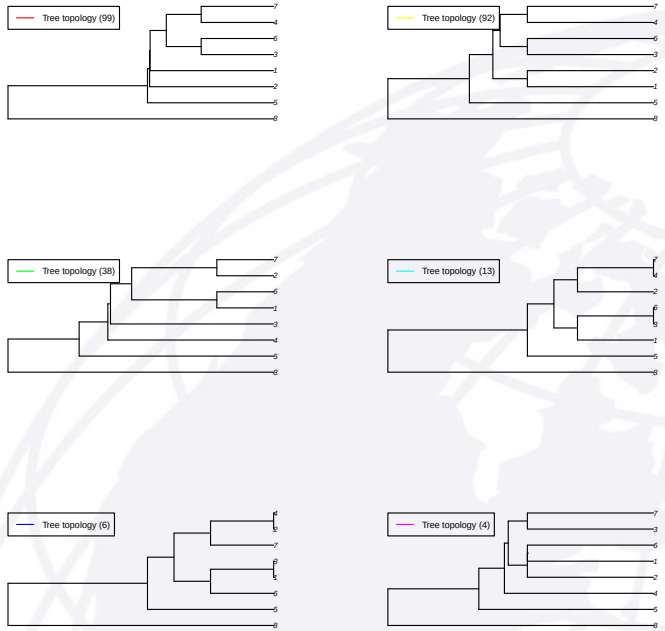
Apicomplexa



The Apicomplexa (also called Apicomplexia) are a large phylum of parasitic alveolates including malaria and Toxoplasma.

Apicomplexa





Outliers via tropical KDE: IDs 691, 566, 650, 730, 615, 712, 630, 625, 755, 708, 497, 690, 503 (ordered by the smallest probabilities to the largest).

#	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

Statistical Learning Models with Tropical Metric

- Tropical Support Vector Machines [Yoshida et al. (2023)]: Phylogenomics and Neuron Data;
- Logistic Regression with Tropical Metric [Aliatimis, et al (2024)]: Phylogenomics;
- Tropical PCA [Page et al. (2020), Miura and Yoshida (2023)]: Phylogenomics and Neuron Data;
- Tropical Density Estimation [Yoshida et al. (2024)]: Phylogenomics;
- KNN with Tropical Metric [Yoshida (2022)]: Phylogenomics;
- Neural Network with ReLU is a tropical "rational function" [Zhang et al. (2018)]; and
- etc.

Advertising

We are hiring open ranked position at the Operations Research Department at NPS!!

If you are interested in please let me know!
https://main.hercjobs.org/jobs/20428959/assistant-associate-full-professor-ad-3-5-7

Please submit via email your application package to Faculty Hiring Committee, Operations Research Department, Naval Postgraduate School, at or_jobs@nps.edu.

Thank you for your attention!

Questions?

"Tropical Geometric Tools for Machine Learning: the TML package" (with D. Barnhill, G. Aliatimis and K. Miura). Journal of Software for Algebra and Geometry 14-1 (2024), 133–174. DOI 10.2140/jsag.2024.14.133.

R Package `TML`:
https://cran.r-project.org/web/packages/TML/index.html

Supported by NSF.