Analyzing Phylogenetic Trees with a Tree Lattice Coordinate System and a Graph Polynomial

Pengyu Liu*, Priscila Biller, Matthew Gould, and Caroline Colijn
Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

**E-mail: pengyu_liu@sfu.ca.*

Supplementary Material

Tree Representations and Metrics

Polynomial representation and metric In this section, we provide more details about the polynomial representation for tree shapes, the lattice representation for trees with branch lengths, and their induced metrics. A tree is a connected acyclic graph. A clade in a rooted tree is the subtree consisting of an internal node and all its descendant nodes in the tree. A tree shape or a tree topology is an unlabelled tree. A tree shape T can be uniquely represented by a bivariate graph polynomial $P(T, x, y)$. Let T be a tree shape with n tips, the coefficients of its polynomial $P(T, x, y)$ can be displayed as the coefficient matrix $C(T)$.

$$C(T) = \begin{array}{c} \\ 1 \\ x \\ x^2 \\ \vdots \\ x^n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & y & y^2 & \ldots & y^n \end{array} \\ \begin{bmatrix} c^{(0,0)} & c^{(0,1)} & c^{(0,2)} & \ldots & c^{(0,n)} \\ c^{(1,0)} & c^{(1,1)} & c^{(1,2)} & \ldots & c^{(1,n)} \\ c^{(2,0)} & c^{(2,1)} & c^{(2,2)} & \ldots & c^{(2,n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^{(n,0)} & c^{(n,1)} & c^{(n,2)} & \ldots & c^{(n,n)} \end{bmatrix} \end{array} = (c^{(i,j)})$$

For a rooted tree shape T with n tips, the coefficient $c^{(i,j)}$ in its polynomial $P(T, x, y)$ can be interpreted as the number of options to choose j clades in T such that these clades include $n-i$ tips of T in total. In particular, the second column of $C(T)$ shows the numbers of clades of each size in the rooted tree T.

Let T_1, T_2 be two tree shapes and $C(T_1) = (c_1^{(i,j)})$, $C(T_2) = (c_2^{(i,j)})$ be the coefficient matrices of the polynomials $P(T_1, x, y)$, $P(T_2, x, y)$. We define a function

$$\gamma(c_1, c_2) = \begin{cases} |c_1 - c_2| / (c_1 + c_2) & \text{if } c_1 \neq 0 \text{ or } c_2 \neq 0 \\ 0 & \text{if } c_1 = 0 \text{ and } c_2 = 0 \end{cases}$$

and the polynomial metric by

$$d_P(T_1, T_2) = \sum_{0 \leqslant i,j \leqslant n} \gamma(c_1^{(i,j)}, c_2^{(i,j)})$$

This distance is also known as the Canberra distance. When comparing two tree shapes of different sizes, we assign a coefficient zero to each term that is absent in a polynomial.

Lattice representation and metric To represent trees with branch lengths, we introduce the lattice representation for rooted phylogenetic trees. A binary tree lattice or simply a tree lattice is the infinite full binary tree with successive integral node labels from the root to its descendants and from left to right, that is, the root has label 1 and the left child of the node with label k has label $2k$ and the right child has label $2k+1$; see Figure 1a. We call these nodes and their labels the lattice positions of the tree lattice.

In order to compute the lattice representation of a tree T with branch lengths, we need to determine the lattice positions for every node in T. For a node N and its child nodes N_1 and N_2, we compute the polynomials $P(S_1, x, y)$ and $P(S_2, x, y)$ of the clades S_1 and S_2 rooted at N_1 and N_2 respectively. Let $C(S_1) = (c_1^{(i,j)})$ and $C(S_2) = (c_2^{(i,j)})$ be the coefficient matrices of the polynomials $P(S_1, x, y)$, $P(S_2, x, y)$. We compare the corresponding coefficients $c_1^{(i,j)}$ and $c_2^{(i,j)}$ in the alphabetic order of (i, j), that is, in the order of $(0,0), (0,1), (0,2), ..., (1,0), (1,1), (1,2)...$ traversing each row of the coefficient matrices. If the first distinct pair of coefficients has $c_1^{(i,j)} > c_2^{(i,j)}$, then N_1 is the left child of N, otherwise N_2 is the left child. If the polynomials $P(S_1, x, y)$, $P(S_2, x, y)$ are identical, then the node with larger branch length is the left child node of N. We call this presented form of the tree T the polynomial ladderized form and the list of which lattice position is occupied by a node of the tree T together with the list of corresponding branch length the lattice representation of the tree T. As an example, Table 1 shows the lattice representation of the tree T displayed in Figure 1b. Note that lattice representations can also represent tree shapes by setting the branch lengths to one.

Let T be a tree with branch lengths and l^i be the branch length of T at lattice position i. If the lattice position is not occupied, then $l^i = 0$. Similarly, for two trees T_1, T_2 with branch lengths, we define the lattice metric by

$$d_L(T_1, T_2) = \sum_{i=1}^{\infty} \gamma(l_1^i, l_2^i)$$

The Canberra distance The Canberra distance, hence the polynomial and the lattice distance, are genuine mathematical metrics. Here is an elementary proof. It is easy to check that $d(T_1, T_2) = 0$ if and only if $T_1 \simeq T_2$, and $d(T_1, T_2) = d(T_2, T_1)$. We show that the triangular inequality is true for the metric, that is, $d(T_1, T_3) \leqslant d(T_1, T_2) + d(T_2, T_3)$. We only need to prove the following inequality holds for $a, b, c \geqslant 0$.

$$\frac{|a-c|}{a+c} \leqslant \frac{|a-b|}{a+b} + \frac{|b-c|}{b+c}$$

Note that if $a \geqslant c \geqslant b$ or $c \geqslant a \geqslant b$, we have

$$\frac{|a-c|}{a+c} \leqslant \frac{|a-b|}{a+c} + \frac{|b-c|}{a+c} \leqslant \frac{|a-b|}{a+b} + \frac{|b-c|}{b+c}$$

If $a \geqslant b \geqslant c$, we have

$$\frac{a-c}{a+c} \leqslant \frac{2b(a-c)}{(a+b)(b+c)} = \frac{a-b}{a+b} + \frac{b-c}{b+c}$$

This is equivalent to $b^2 + ac \leqslant ab + bc$, which is true because $ac - bc \leqslant ab - b^2$. Similarly, the equality also holds when $c \geqslant b \geqslant a$.

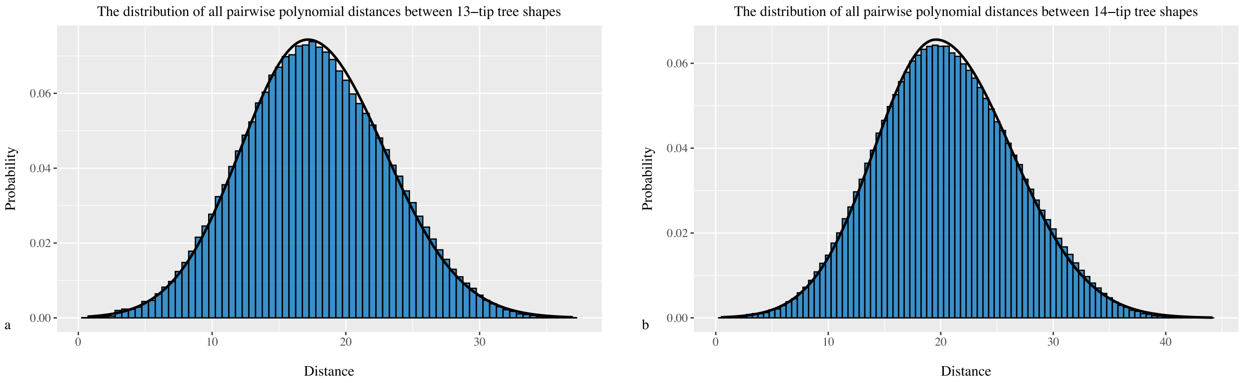
If $b \geqslant a \geqslant c$, we have

$$\frac{a-c}{a+c} \leqslant \frac{2(b^2-ac)}{(a+b)(b+c)} = \frac{b-a}{a+b} + \frac{b-c}{b+c}$$

This is equivalent to $ab(b-a) + 3c(b^2 - a^2) + c^2(b-a) \geqslant 0$, which is true as $b \geqslant a$.

Similarly, the equality also holds when $b \geqslant c \geqslant a$. Therefore the polynomial metric is a genuine metric.

The distribution of polynomial distances The distribution of polynomial distances between all pairs of tree shapes with n tips resembles a normal distribution. Supplementary Figure 1 displays the distribution for tree shapes with 13 and 14 tips, where the black solid curves are normal fits. For the distribution for 13-tip trees, the estimated mean value is 17.70, the estimated standard deviation is 5.37, and Shapiro-Wilk normality test has W of 0.99 and p-value of 6.21×10^{-15}. For the distribution for 14-tip trees, the estimated mean value is 20.54, the estimated standard deviation is 6.10, and Shapiro-Wilk normality test has W of 0.99 and p-value of 4.43×10^{-15}.



Supplementary Figure 1. a: the distributions of all pairwise polynomial distances between all rooted binary tree shapes with 13 tips. b: the distributions of all pairwise polynomial distances between all rooted binary tree shapes with 14 tips. The black solid curves are normal fits.

Tree Clustering

We display analogous visualizations of different pairwise distances between tree shapes and trees with branch lengths for the same datasets as displayed in Figure 1. Visualizations are displayed in Supplememntary Figure 2.

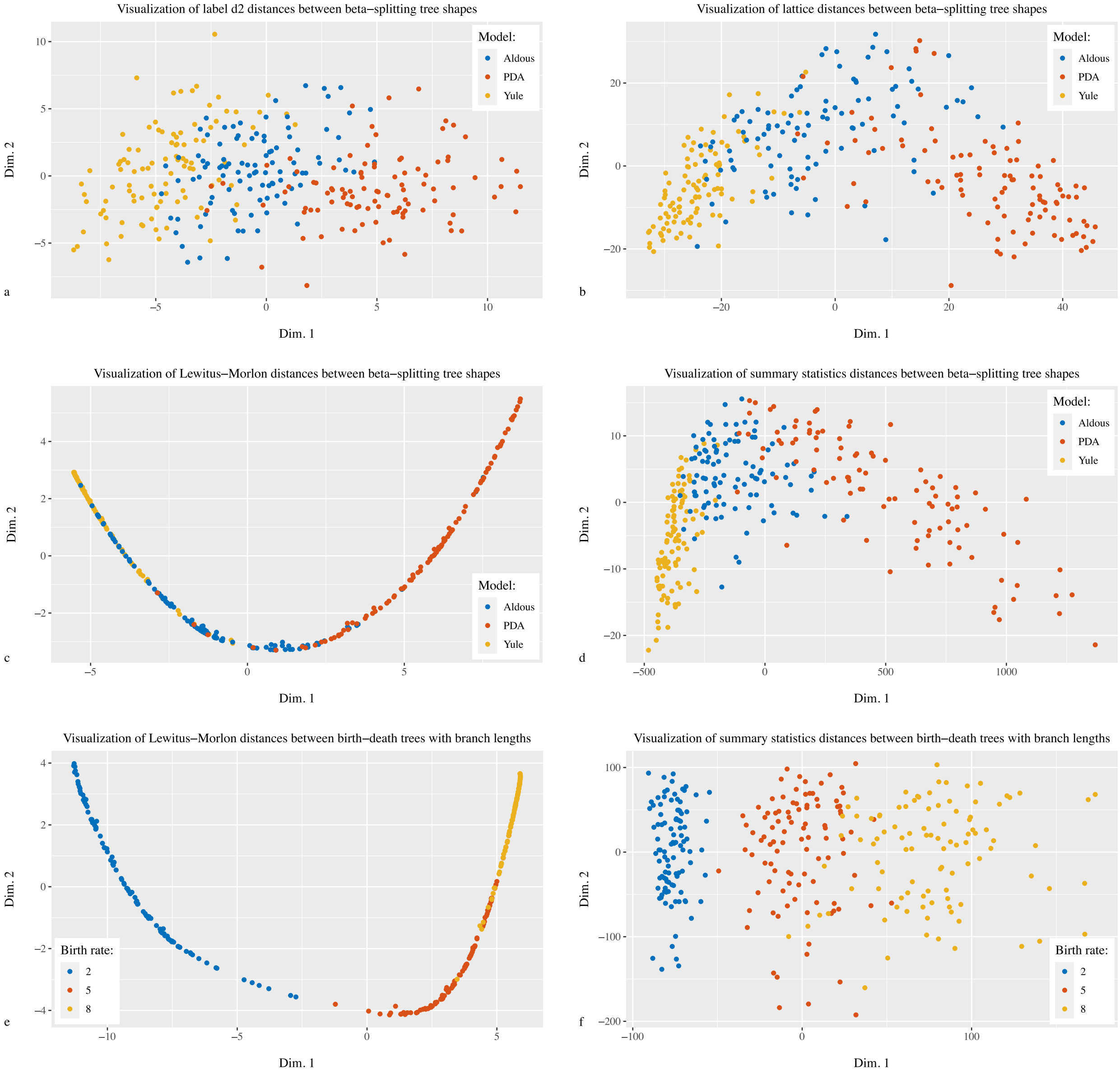
Parameter Estimation and Model Selection

We conduct the 20 more parameter estimation experiments for the 100 sets of birth-death trees of various sizes by approximate Bayesian computation (ABC) method on the summary statistics and on flattened lattice distance, constructed by attaching the list of branch lengths to the end of the list of occupied lattice positions. Supplementary Figure 3a shows the average relative error, and Supplementary Figure 3c shows the computational time for each of the experiments. The results are compared with the results by k-nearest neighbors regression on the lattice metric.

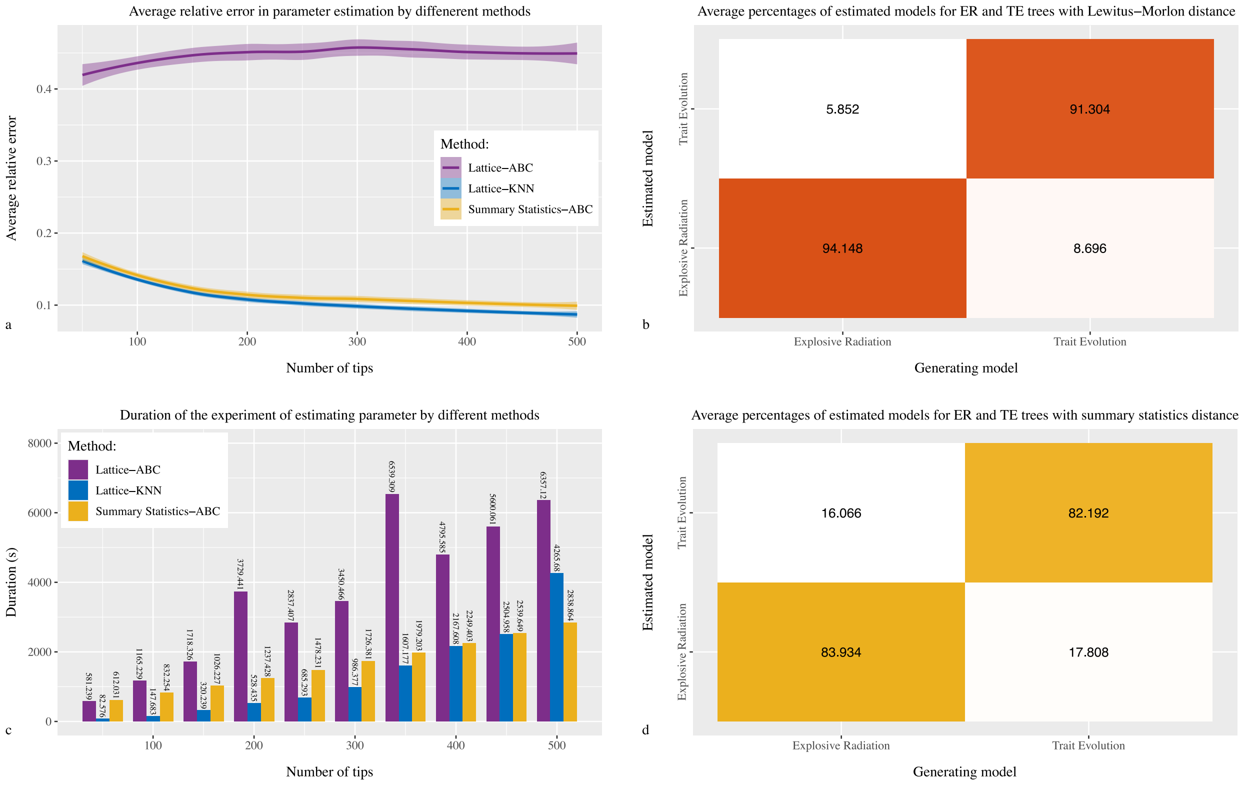
Supplementary Figure 3b and Supplementary Figure 3d show the average percentages of estimated models over 100 sets for explosive radiation trees and trait evolution trees by k-nearest neighbors classification and the Lewitus-Morlon distance and the summary statistics distance respectively.

Applications to Real Data

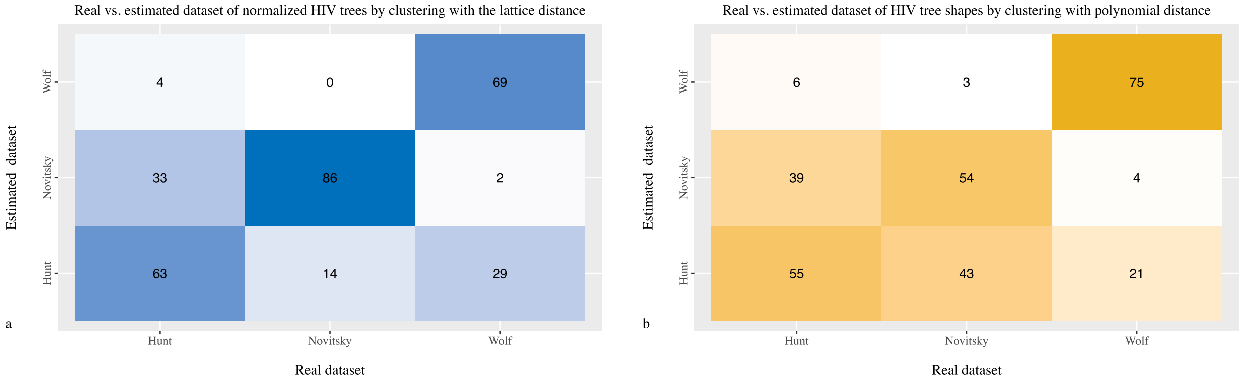
In Supplementary Figure 4, we show the average percentages of estimated dataset for normalized HIV trees by k-medoids clustering and the lattice distance and for HIV tree shapes by k-medoids clustering and the polynomial distance, in correspondence with the visualizations in Figure 4a and Figure 4c.



Supplementary Figure 2. a: visualizations of pairwise label d_2 distances between beta-splitting tree shapes. b: visualizations of pairwise lattice distances between beta-splitting tree shapes. c: visualizations of pairwise Lewitus-Morlon distances between beta-splitting tree shapes. d: visualizations of pairwise summary statistics distances between beta-splitting tree shapes. e: visualizations of pairwise Lewitus-Morlon distances between birth-death trees with branch lengths. f: visualizations of pairwise summary statistics distances between birth-death trees with branch lengths. All visualizations are by multi-dimensional scaling and each dot represents a tree shape or a tree with branch lengths in these plots.



Supplementary Figure 3. a: average relative error in estimating parameters for trees of different sizes by different methods; curves represent mean values over 100 sets and bands represent 95% confidence intervals of the mean values. b: average percentages of estimated models over 100 sets for explosive radiation trees and trait evolution trees by k-nearest neighbors classification and the Lewitus-Morlon distance. c: duration of the experiment of parameter estimation with different methods and trees of various sizes. d: average percentages of estimated models over 100 sets for explosive radiation trees and trait evolution trees by k-nearest neighbors classification and the summary statistics distance.



Supplementary Figure 4. a: the estimated dataset of normalized HIV trees by k-medoids clustering and the lattice distance, with misclassification rate 0.273. b: the estimated dataset of HIV tree shapes by k-medoids clustering and the polynomial distance, with misclassification rate 0.423.