

Catalog of Small Trees

Marta Casanellas Luis David Garcia Seth Sullivan

This chapter is concerned with the description of the Small Trees website which can be found at the following web address:

<http://www.math.tamu.edu/~lgp/small-trees/small-trees.html>

The goal of the website is to make available in a unified format various algebraic features of different phylogenetic models. In the first section, we describe a detailed set of notational conventions for describing the phylogenetic models on trees which are listed on this website. This includes conventions for writing down the parameterizations given a tree as well as describing the Fourier transform and writing down phylogenetic invariants in Fourier coordinates. The second section gives a brief description of each of the types of algebraic information which are associated to a model and a tree on the Small Trees website. The third section contains an example of a page on the website. The final section is concerned with simulation studies of using algebraic invariants to recover phylogenies using the invariants for the Kimura 3-parameter model.

1 Notational Conventions

1.1 Labeling trees

We assume that each phylogenetic model is presented with a particular tree T together with a figure representing that tree. The figures of trees with up to five leaves will be the ones that can be found on the Small Trees website.

1.1.1 Rooted trees

If T is a rooted tree, there is a distinguished vertex of T called the root and labeled by the letter r . The tree T should be drawn with the root r at the

top of the figure and the edges of the tree below the root. Each edge in the tree is labeled with a lowercase letter a, b, c, \dots . The edges are labeled in alphabetical order starting at the upper left hand corner, proceeding left to right and top to bottom. The leaves are labeled with the numbers $1, 2, 3, \dots$ starting with the left-most leaf and proceeding left to right. Figure 1 shows the “giraffe” tree with four leaves and its labeling.

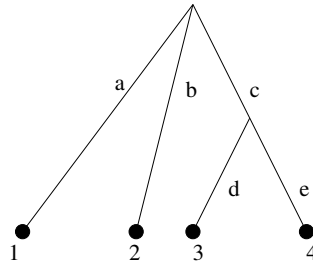


Figure 1: The giraffe tree on four leaves

1.1.2 Unrooted trees

If T is an unrooted tree, it should be drawn with the leaves in a circle. The edges of T are labeled with lower-case letters a, b, c, \dots in alphabetical order starting at the upper left-hand corner of the figure and proceeding left to right and top to bottom. The leaves are labeled with the numbers $1, 2, 3, \dots$ starting at the first leaf “left of 12 o’clock” and proceeding counterclockwise around the perimeter of the tree. Figure 2 illustrates this on the “quartet” tree.

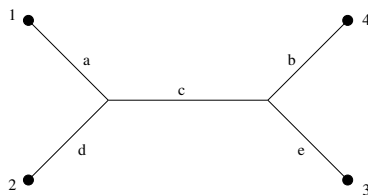


Figure 2: The quartet tree on four leaves

1.2 Parameterizations

Associated to each node in a model is a random variable with two or four states depending on whether we are looking at binary data or DNA data. In the case of binary data these states are $\{0, 1\}$ and for DNA data they are $\{A, C, G, T\}$ in this order.

1.2.1 Root Distribution

The root distribution is a vector of length two or four depending on whether the model is for binary or DNA sequences. The name of this vector is r . Its entries are parameters r_0, r_1, r_2, \dots and are filled in from left to right and are recycled as the model requires.

Example 1. In the general strand symmetric model r always denotes the vector

$$r = (r_0, r_1, r_1, r_0).$$

We tacitly assume that the entries in r sum to 1, though we do not eliminate a parameter to take this into account. If the model assumes a uniform root distribution, then r has the form $r = (1/2, 1/2)$ or $r = (1/4, 1/4, 1/4, 1/4)$ according to whether the model is for binary or DNA data.

1.2.2 Transition Matrices

In each type of model, the letters a, b, c, \dots which label the edges are also the transition matrices in the model. These are either 2×2 or 4×4 matrices depending on whether the model is a model for binary data or DNA data. In each case, the matrix is filled from left to right and top to bottom with unknown parameters, recycling a parameter whenever the model requires it. For the transition matrix of the edge labeled with x these entries are called x_0, x_1, x_2, \dots

Example 2. For example, in the Kimura 3-parameter model the letter a represents the matrix

$$a = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_0 & a_3 & a_2 \\ a_2 & a_3 & a_0 & a_1 \\ a_3 & a_2 & a_1 & a_0 \end{pmatrix}.$$

The Kimura 2-parameter and Jukes-Cantor models give rise to specializations of the parameters in the Kimura 3-parameter model, and hence the letters denoting the parameters are recycled. For instance, the letter c in the Jukes-Cantor DNA model and the letter d in the Kimura 2-parameter model represent the following matrices

$$c = \begin{pmatrix} c_0 & c_1 & c_1 & c_1 \\ c_1 & c_0 & c_1 & c_1 \\ c_1 & c_1 & c_0 & c_1 \\ c_1 & c_1 & c_1 & c_0 \end{pmatrix}, \quad d = \begin{pmatrix} d_0 & d_1 & d_2 & d_1 \\ d_1 & d_0 & d_1 & d_2 \\ d_2 & d_1 & d_0 & d_1 \\ d_1 & d_2 & d_1 & d_0 \end{pmatrix}.$$

In the general strand symmetric model the letter e always represents the matrix

$$e = \begin{pmatrix} e_0 & e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 & e_7 \\ e_7 & e_6 & e_5 & e_4 \\ e_3 & e_2 & e_1 & e_0 \end{pmatrix}.$$

We assume that the entries of these matrices satisfy additional linear constraints which make them into transition matrices. For instance, in the Jukes-Cantor DNA model, this constraint is $c_0 + 3c_1 = 1$ and in the general strand symmetric model the two linear relations are $e_0 + e_1 + e_2 + e_3 = 1$ and $e_4 + e_5 + e_6 + e_7 = 1$. We do not, however, use these linear relations to eliminate parameters.

1.2.3 Molecular Clock Assumption

The molecular clock (MC) assumption for a rooted tree T is defined as the assumption that, for each subtree, along each path from the root of that subtree to any leaf i the product of the transition matrices corresponding to the edges are identical. As the edges in the path are read down the tree, the matrices are multiplied left to right.

Example 3. For the giraffe tree in Figure 1 the MC assumption translates into the following identities:

$$a = b = cd = ce \text{ and } d = e.$$

These equalities of products of parameter matrices suggest that some parameter matrices should be replaced with products of other parameter

matrices and their inverses. This makes the parameterization involve rational functions (instead of just polynomials).

Here is a systematic rule for making these replacements. Starting from the bottom of the tree, make replacements for transition matrices. Each vertex in the tree induces equalities among products of transition matrices along all paths emanating downward from this vertex. Among the edges emanating downward from a given vertex, all but one of the transition matrices for these edges will be replaced by a product of other transition matrices and their inverses. When choosing replacements, always replace the transition matrix which belongs to the shorter path to a leaf. If all such paths have the same length, replace the matrices which belong to the left most edges emanating from a vertex.

Example 4. In the 4-leaf giraffe tree from the previous example, we replace the matrix d with e and we replace the matrices a and b with ce . Thus, when we write the parameterization in probability coordinates only the letters c and e will appear in the parameterizing polynomials.

1.2.4 Specifying the Joint Distribution

The probabilities of the leaf colorations of a tree with n leaves are denoted by p_W where W is a word of length n in the alphabet $\{0, 1\}$ or $\{A, C, G, T\}$. Every probability indeterminate p_W is a polynomial in the parameters of the model. Two of these probabilities p_W and p_U are equivalent if their defining polynomials are identical. This divides the 2^n or 4^n probabilities into equivalence classes. The elements of each class are ordered lexicographically, and the classes are ordered lexicographically by their lexicographically first elements.

Example 5. For the Jukes–Cantor DNA model with uniform root distribution on a three taxa claw tree there are five equivalence classes:

- **Class 1:** $p_{AAA} p_{CCC} p_{GGG} p_{TTT}$
- **Class 2:** $p_{AAC} p_{AAG} p_{AAT} p_{CCA} p_{CCG} p_{CCT} p_{GGA} p_{GGC} p_{GGT} p_{TTA} p_{TTC} p_{TTG}$
- **Class 3:** $p_{ACA} p_{AGA} p_{ATA} p_{CAC} p_{CGC} p_{CTC} p_{GAG} p_{GCG} p_{GTG} p_{TAT} p_{TCT} p_{TGT}$

- **Class 4:** $p_{ACC} p_{AGG} p_{ATT} p_{CAA} p_{CGG} p_{CTT} p_{GAA} p_{GCC} p_{GTT} p_{TAA}$
 $p_{TCC} p_{TGG}$
- **Class 5:** $p_{ACG} p_{ACT} p_{AGC} p_{AGT} p_{ATC} p_{ATG} p_{CAG} p_{CAT} p_{CGA} p_{CGT}$
 $p_{CTA} p_{CTG} p_{GAC} p_{GAT} p_{GCA} p_{GCT} p_{GTA} p_{GTC} p_{TAC} p_{TAG} p_{TCA} p_{TCG}$
 $p_{TGA} p_{TGC}$

For each class i there will be an indeterminate p_i which denotes the sum of the probabilities in the class i . For these N probabilities the expression for the probability p_i as a polynomial or rational function in the parameters appears on the webpage (if these expressions are small enough) or in a separate linked page for longer expressions.

Example 6. In the 3-taxa claw tree with Jukes–Cantor model and uniform root distribution these indeterminates are:

$$\begin{aligned}
p_1 &= a_0 b_0 c_0 + 3a_1 b_1 c_1 \\
p_2 &= 3a_0 b_0 c_1 + 3a_1 b_1 c_0 + 6a_1 b_1 c_1 \\
p_3 &= 3a_0 b_1 c_0 + 3a_1 b_0 c_1 + 6a_1 b_1 c_1 \\
p_4 &= 3a_1 b_0 c_0 + 3a_0 b_1 c_1 + 6a_1 b_1 c_1 \\
p_5 &= 6a_0 b_1 c_1 + 6a_1 b_0 c_1 + 6a_1 b_1 c_0 + 6a_1 b_1 c_1
\end{aligned}$$

Note that $p_1 + p_2 + p_3 + p_4 + p_5 = 1$ after substituting $a_0 = 1 - 3a_1$, $b_0 = 1 - 3b_1$ and $c_0 = 1 - 3c_1$.

1.3 Fourier Coordinates

Often we will describe these phylogenetic models in an alternate coordinate system called the Fourier coordinates. This change of coordinates happens simultaneously on the parameters and on the probability coordinates themselves.

1.3.1 Full Fourier Transform

Each of the 2^n or 4^n Fourier coordinate are denoted by q_W where W is a word in either $\{0, 1\}$ or $\{A, C, G, T\}$.

The Fourier transform from p_U to q_W is given by the following rule:

$$p_{i_1 \dots i_n} = \sum_{j_1, \dots, j_n} \chi^{j_1}(i_1) \cdots \chi^{j_n}(i_n) q_{j_1 \dots j_n},$$

$$q_{i_1 \dots i_n} = \frac{1}{k^n} \sum_{j_1, \dots, j_n} \chi^{i_1}(j_1) \cdots \chi^{i_n}(j_n) p_{i_1 \dots i_n}.$$

Here χ^i is the character of the group associated to the i th group element. The character tables of the groups we use, namely \mathbb{Z}_2 and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are:

	0	1			A	1	1	1	1
0	1	1	and		C	1	-1	1	-1
1	1	-1			G	1	1	-1	-1
					T	1	-1	-1	1

In other words, $\chi^i(j)$ is the (i, j) entry in the appropriate character table. One special feature of this transformation is that the Fourier transform of the joint distribution has a parameterization that can be written in product form; we refer to [2, 5, 6] for a detailed treatment of the subject. Equivalently, the Fourier transform simultaneously diagonalizes all transition matrices. Therefore, we replace the transition matrices a, b, c, \dots for diagonal matrices denoted A, B, C, \dots , where A has diagonal elements A_1, A_2, A_3, A_4 ; B has diagonal elements B_1, B_2, B_3, B_4 ; etc. Since we will only use the entries of the previous diagonal matrices, there will be no confusion, for example, between the matrix A and the base A . Furthermore, these parameters must satisfy the relations imposed by the corresponding model and the Molecular Clock assumption. For instance, in the Jukes–Cantor model we have the relations $A_2 = A_3 = A_4, B_2 = B_3 = B_4$.

The q_W are polynomials or rational functions in the transformed parameters. They are given parametrically as

$$q_{i_1 \dots i_n} := \begin{cases} \prod_{e \in E} M_e(k_e) & \text{if } i_n = i_1 + i_2 + \dots + i_{n-1} \text{ in the group} \\ 0 & \text{otherwise} \end{cases}$$

where M_e is the corresponding diagonal matrix associated to edge e , and k_e is the sum (in the corresponding group) of the labels at the leaves that are “beneath” the edge e .

We say that q_W and q_U are equivalent if they represent the same polynomial in terms of these parameters. These Fourier coordinates are grouped

into equivalence classes. The elements in the equivalence classes are ordered lexicographically. Most of the Fourier coordinates q_W are zero and these are grouped in class 0. The others are ordered Class 1, Class 2, lexicographically by their lexicographically first element.

Example 7. Here we display the classes of Fourier coordinates for the Jukes–Cantor DNA model on the 3 leaf claw tree.

- **Class 0:** $q_{AAC} q_{AAT} \dots$
- **Class 1:** q_{AAA}
- **Class 2:** $q_{ACC} q_{AGG} q_{ATT}$
- **Class 3:** $q_{CAC} q_{GAG} q_{TAT}$
- **Class 4:** $q_{CCA} q_{GGA} q_{TTA}$
- **Class 5:** $q_{CGT} q_{CTG} q_{GCT} q_{GTC} q_{TCG} q_{TGC}$

We replace each of the Fourier coordinates in class i by the new Fourier coordinate q_i . We take q_i to be the *average* of the q_W in class i since this operation is better behaved with respect to writing down invariants.

1.3.2 Specialized Fourier Transform

We also record explicitly the linear transformation between the p_i and the q_i by recording a certain rational matrix which describes this transformation. This is the *specialized Fourier transform*. In general, this matrix will not be a square matrix. This is because there may be additional linear relations among the p_i which are encoded in the different q_i classes. Because of this ambiguity, we also explicitly list the inverse map.

It is possible to obtain the matrix that represents the specialized Fourier transform from the matrix that represents the full Fourier transform. If M represents the matrix of the full Fourier transform and N the matrix of the specialized Fourier transform, then N_{ij} , (the entry indexed by the i th Fourier class and the j th probability class) is given by the formula:

$$N_{ij} = \frac{1}{|C_i||D_j|} \sum_{U \in C_i} \sum_{W \in D_j} M_{UW}$$

where C_i is the i th equivalence class of Fourier coordinates and D_j is the j th equivalence class of probability coordinates. We do not include the 0th equivalence class of Fourier coordinates in the previous formula.

Example 8. In the Jukes–Cantor DNA model on the 3 leaf claw tree the specialized Fourier transform matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{3} & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3} \\ 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1 \end{pmatrix}.$$

2 Description of website features

We give a brief description of the various items described on the website.

Dimension (D): The dimension of the model.

Degree (d): The degree of the model. Algebraically, this is defined as the number of points in the intersection of the model and a generic (i.e. “random”) subspace of dimension 4^n minus the dimension of the model.

Maximum Likelihood Degree (mld): The maximum likelihood degree of the model. See section ??.

Number of Probability Coordinates (np): Number of equivalence classes of the probability coordinates. See the preceding section.

Number of Fourier Coordinates (nq): Number of equivalence classes of Fourier coordinates without counting Class 0 (see the preceding section). This is also the dimension of the smallest linear space that contains the model.

Specialized Fourier Transform: See the preceding section for a description.

Phylogenetic Invariants: A list of generators of the prime ideal of phylogenetic invariants. These are given in the Fourier coordinates.

Singularity Dimension (sD): The dimension of the set of singular points on the model.

Singularity Degree (sd): The algebraic degree of the set of singular points on the model.

3 Example

Here we describe the Jukes–Cantor model on the quartet tree (see Figure 2) in more detail.

Dimension: $D = 5$ (note that there are only 5 independent parameters, one for each transition matrix.)

Degree: $d = 34$.

Number of Probability Coordinates: $np = 15$ and the classes are represented by:

$$\begin{aligned}
p_1 &= a_0b_0c_0d_0e_0 + 3a_1b_0c_1d_1e_0 + 3a_0b_1c_1d_0e_1 + 3a_1b_1c_0d_1e_1 + 6a_1b_1c_1d_1e_1, \\
p_2 &= 3(a_0b_1c_0d_0e_0 + 3a_1b_1c_1d_1e_0 + a_0b_0c_1d_0e_1 + 2a_0b_1c_1d_0e_1 \\
&\quad + a_1b_0c_0d_1e_1 + 2a_1b_1c_0d_1e_1 + 2a_1b_0c_1d_1e_1 + 4a_1b_1c_1d_1e_1), \\
p_3 &= 3(a_0b_1c_1d_0e_0 + a_1b_1c_0d_1e_0 + 2a_1b_1c_1d_1e_0 + a_0b_0c_0d_0e_1 \\
&\quad + 2a_0b_1c_1d_0e_1 + 2a_1b_1c_0d_1e_1 + 3a_1b_0c_1d_1e_1 + 4a_1b_1c_1d_1e_1), \\
p_4 &= 3(a_0b_0c_1d_0e_0 + a_1b_0c_0d_1e_0 + 2a_1b_0c_1d_1e_0 + a_0b_1c_0d_0e_1 \\
&\quad + 2a_0b_1c_1d_0e_1 + 2a_1b_1c_0d_1e_1 + 7a_1b_1c_1d_1e_1), \\
p_5 &= 6(a_0b_1c_1d_0e_0 + a_1b_1c_0d_1e_0 + 2a_1b_1c_1d_1e_0 + a_0b_1c_0d_0e_1 + a_0b_0c_1d_0e_1 \\
&\quad + a_0b_1c_1d_0e_1 + a_1b_0c_0d_1e_1 + a_1b_1c_0d_1e_1 + 2a_1b_0c_1d_1e_1 + 5a_1b_1c_1d_1e_1), \\
p_6 &= 3(a_1b_0c_1d_0e_0 + a_0b_0c_0d_1e_0 + 2a_1b_0c_1d_1e_0 + a_1b_1c_0d_0e_1 \\
&\quad + 2a_1b_1c_1d_0e_1 + 2a_1b_1c_0d_1e_1 + 3a_0b_1c_1d_1e_1 + 4a_1b_1c_1d_1e_1), \\
p_7 &= 3(a_1b_1c_1d_0e_0 + a_0b_1c_0d_1e_0 + 2a_1b_1c_1d_1e_0 + a_1b_0c_0d_0e_1 + 2a_1b_1c_1d_0e_1 \\
&\quad + 2a_1b_1c_0d_1e_1 + a_0b_0c_1d_1e_1 + 2a_1b_0c_1d_1e_1 + 2a_0b_1c_1d_1e_1 + 2a_1b_1c_1d_1e_1), \\
p_8 &= 6(a_1b_1c_1d_0e_0 + a_0b_1c_0d_1e_0 + 2a_1b_1c_1d_1e_0 + a_1b_1c_0d_0e_1 \\
&\quad + a_1b_0c_1d_0e_1 + a_1b_1c_1d_0e_1 + a_1b_0c_0d_1e_1 + a_1b_1c_0d_1e_1 \\
&\quad + a_0b_0c_1d_1e_1 + a_1b_0c_1d_1e_1 + 2a_0b_1c_1d_1e_1 + 3a_1b_1c_1d_1e_1), \\
p_9 &= 3(a_1b_1c_0d_0e_0 + a_0b_1c_1d_1e_0 + 2a_1b_1c_1d_1e_0 + a_1b_0c_1d_0e_1 + 2a_1b_1c_1d_0e_1 \\
&\quad + a_0b_0c_0d_1e_1 + 2a_1b_1c_0d_1e_1 + 2a_1b_0c_1d_1e_1 + 2a_0b_1c_1d_1e_1 + 2a_1b_1c_1d_1e_1), \\
p_{10} &= 3(a_1b_0c_0d_0e_0 + a_0b_0c_1d_1e_0 + 2a_1b_0c_1d_1e_0 + 3a_1b_1c_1d_0e_1 \\
&\quad + a_0b_1c_0d_1e_1 + 2a_1b_1c_0d_1e_1 + 2a_0b_1c_1d_1e_1 + 4a_1b_1c_1d_1e_1),
\end{aligned}$$

$$\begin{aligned}
p_{11} &= 6(a_1b_1c_0d_0e_0 + a_0b_1c_1d_1e_0 + 2a_1b_1c_1d_1e_0 + a_1b_0c_1d_0e_1 \\
&\quad + 2a_1b_1c_1d_0e_1 + a_1b_0c_0d_1e_1 + a_0b_1c_0d_1e_1 + a_1b_1c_0d_1e_1 \\
&\quad + a_0b_0c_1d_1e_1 + a_1b_0c_1d_1e_1 + a_0b_1c_1d_1e_1 + 3a_1b_1c_1d_1e_1), \\
p_{12} &= 6(a_1b_1c_1d_0e_0 + a_1b_1c_0d_1e_0 + a_0b_1c_1d_1e_0 + a_1b_1c_1d_1e_0 \\
&\quad + a_1b_1c_0d_0e_1 + a_1b_0c_1d_0e_1 + a_1b_1c_1d_0e_1 + a_0b_0c_0d_1e_1 \\
&\quad + a_1b_1c_0d_1e_1 + 2a_1b_0c_1d_1e_1 + 2a_0b_1c_1d_1e_1 + 3a_1b_1c_1d_1e_1), \\
p_{13} &= 6(a_1b_1c_1d_0e_0 + a_1b_1c_0d_1e_0 + a_0b_1c_1d_1e_0 + a_1b_1c_1d_1e_0 \\
&\quad + a_1b_0c_0d_0e_1 + 2a_1b_1c_1d_0e_1 + a_0b_1c_0d_1e_1 + a_1b_1c_0d_1e_1 \\
&\quad + a_0b_0c_1d_1e_1 + 2a_1b_0c_1d_1e_1 + a_0b_1c_1d_1e_1 + 3a_1b_1c_1d_1e_1), \\
p_{14} &= 6(a_1b_0c_1d_0e_0 + a_1b_0c_0d_1e_0 + a_0b_0c_1d_1e_0 + a_1b_0c_1d_1e_0 + a_1b_1c_0d_0e_1 \\
&\quad + 2a_1b_1c_1d_0e_1 + a_0b_1c_0d_1e_1 + a_1b_1c_0d_1e_1 + 2a_0b_1c_1d_1e_1 + 5a_1b_1c_1d_1e_1), \\
p_{15} &= 6(a_1b_1c_1d_0e_0 + a_1b_1c_0d_1e_0 + a_0b_1c_1d_1e_0 + a_1b_1c_1d_1e_0 + a_1b_1c_0d_0e_1 \\
&\quad + a_1b_0c_1d_0e_1 + a_1b_1c_1d_0e_1 + a_1b_0c_0d_1e_1 + a_0b_1c_0d_1e_1 + a_0b_0c_1d_1e_1 \\
&\quad + a_1b_0c_1d_1e_1 + a_0b_1c_1d_1e_1 + 4a_1b_1c_1d_1e_1).
\end{aligned}$$

Number of Fourier Coordinates: $nq = 13$. The classes are:

$$\begin{aligned}
q_1 &= q_{AAAA}, \\
q_2 &= q_{AACC}, q_{AAGG}, q_{AATT}, \\
q_3 &= q_{ACAC}, q_{AGAG}, q_{ATAT}, \\
q_4 &= q_{ACCA}, q_{AGGA}, q_{ATTA}, \\
q_5 &= q_{ACGT}, q_{ACTG}, q_{AGCT}, q_{AGTC}, q_{ATCG}, q_{ATGC}, \\
q_6 &= q_{CAAC}, q_{GAAG}, q_{TAAT}, \\
q_7 &= q_{CACA}, q_{GAGA}, q_{TATA}, \\
q_8 &= q_{CAGT}, q_{CATG}, q_{GACT}, q_{GATC}, q_{TACG}, q_{TAGC}, \\
q_9 &= q_{CCAA}, q_{GGAA}, q_{TTAA}, \\
q_{10} &= q_{CCCC}, q_{CCGG}, q_{CCCT}, q_{GGCC}, q_{GGGG}, q_{GGTT}, \\
&\quad q_{TTCC}, q_{TTGG}, q_{TTTT}, \\
q_{11} &= q_{CGAT}, q_{CTAG}, q_{GCAT}, q_{GTAC}, q_{TCAG}, q_{TGAC}, \\
q_{12} &= q_{CGCG}, q_{CGGC}, q_{CTCT}, q_{CTTC}, q_{GCCG}, q_{GCGC}, \\
&\quad q_{GTGT}, q_{GTTG}, q_{TCCT}, q_{TCTC}, q_{TGGT}, q_{TGTG}, \\
q_{13} &= q_{CGTA}, q_{CTGA}, q_{GCTA}, q_{GTCA}, q_{TCGA}, q_{TGCA}.
\end{aligned}$$

The invariants of degree 3 associated to the right interior vertex are:

$$\begin{aligned}
& q_1 q_5 q_5 - q_3 q_4 q_2, q_1 q_5 q_8 - q_3 q_7 q_2, q_1 q_5 q_{12} - q_3 q_{13} q_2, \\
& q_1 q_8 q_5 - q_6 q_4 q_2, q_1 q_8 q_8 - q_6 q_7 q_2, q_1 q_8 q_{12} - q_6 q_{13} q_2, \\
& q_1 q_{12} q_5 - q_{11} q_4 q_2, q_1 q_{12} q_8 - q_{11} q_7 q_2, q_1 q_{12} q_{12} - q_{11} q_{13} q_2, \\
& q_9 q_5 q_5 - q_3 q_4 q_{10}, q_9 q_5 q_8 - q_3 q_7 q_{10}, q_9 q_5 q_{12} - q_3 q_{13} q_{10}, \\
& q_9 q_8 q_5 - q_6 q_4 q_{10}, q_9 q_8 q_8 - q_6 q_7 q_{10}, q_9 q_8 q_{12} - q_6 q_{13} q_{10}, \\
& q_9 q_{12} q_5 - q_{11} q_4 q_{10}, q_9 q_{12} q_8 - q_{11} q_7 q_{10}, q_9 q_{12} q_{12} - q_{11} q_{13} q_{10}.
\end{aligned}$$

The maximum likelihood degree, the singularity dimension and the singularity degree are computationally difficult to achieve and we have not been able to compute them using computer algebra programs.

4 Using the invariants

In this section we report some of the experiments we have made for inferring small trees using phylogenetic invariants. These experiments were made using the invariants for trees with 4 taxa on the Kimura 3-parameter model that can be found in our website [1] which were computed using the Sturmfels-Sullivant theorem [5]. The results obtained show that phylogenetic invariants are an efficient method for tree reconstruction.

We implemented an algorithm that performs the following tasks. Given 4 DNA sequences s_1, s_2, s_3, s_4 , it first counts the number of occurrences of each pattern for the topology $((s_1, s_2), s_3, s_4)$. Then it changes these absolute frequencies to Fourier coordinates. From this, we have the Fourier transforms in the other two possible topologies for trees with 4 species. We then evaluate all the phylogenetic invariants for the Kimura 3-parameter model in the Fourier coordinates of each tree topology. We call s_f^T the absolute value of this evaluation for the polynomial f and tree topology T . From these values $\{s_f^T\}_f$, we produce a score for each tree topology T , namely $s(T) = \sum_f |s_f^T|$. The algorithm then chooses the topology that has minimum score.

There was an attempt to define the score as the Euclidean norm of the values s_f^T , but from our experiments, we deduced that the 1-norm chosen above performs better.

We then tested this algorithm for different sets of sequences. We used the program *evolver* from the package PAML [7] to generate sequences according to the Kimura 2-parameter model with transition/transversion ratio equal

to 2 (typical value of mammalian DNA). In what follows we describe the different tests we made and the results we obtained.

We generated 4-taxa trees with random branch lengths uniformly distributed between 0 and 1. We performed 600 tests for sequences of lengths between 1000 and 10,000. The percentage of trees correctly reconstructed can be seen in Figure 3.

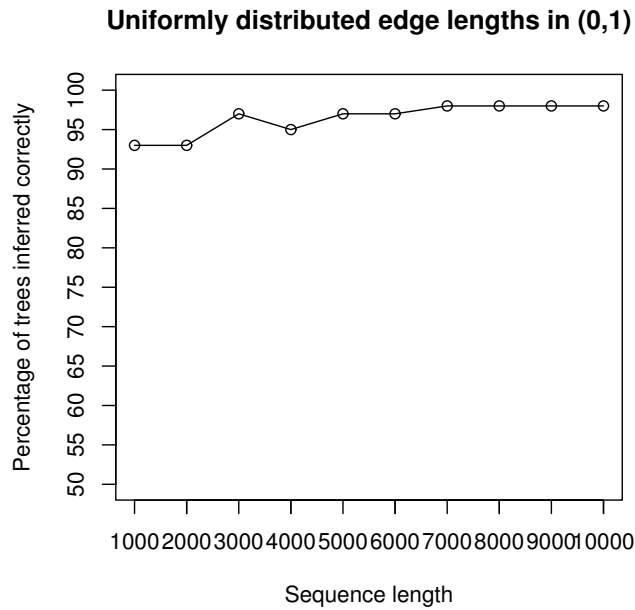


Figure 3: Percentage of trees correctly reconstructed with random branch lengths uniformly distributed between 0 and 1.

We observed that our method fails to reconstruct the right tree mainly when the length of the interior edge of the tree is small compared to the other branch lengths. More precisely, in the trees that cannot be correctly inferred, the length of the interior edge is about 10% the average length of the other edges.

Our method was also tested by letting the edge lengths be normally distributed with a given mean μ . We chose the values 0.25, 0.05, 0.005 for the mean μ , following [3]. We also let the standard deviation be 0.1μ . In this case, we tested DNA sequences of lengths ranging from 50 to 10,000. Here, we only

display the results for sequences of length up to 1000, because we checked that for larger sequences, we always infer the correct tree. For each sequence length in $\{50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$, we generated edge lengths normally distributed with mean μ using the data analysis program *R* [4]. and we performed 10 tests with our algorithm. We repeated this process 10 times, so that we generated 100 sequences for each mean and sequence length. The results are presented in Figure 4.

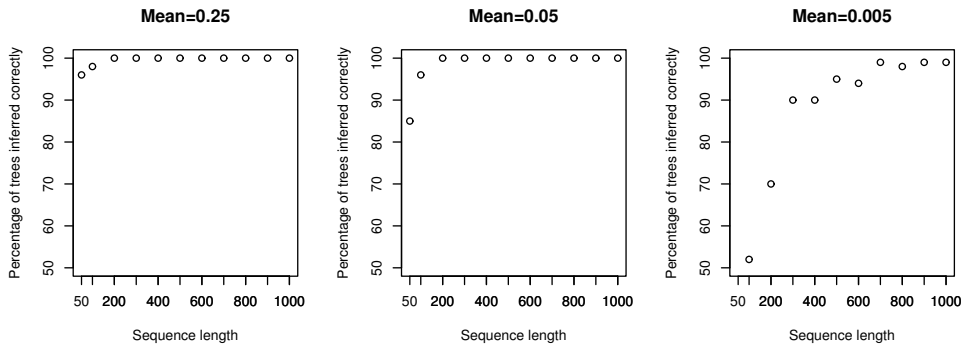


Figure 4: Percentage of trees correctly reconstructed with edge lengths normally distributed with mean equal to 0.25, 0.05, 0.005.

From Figure 4, we see that for $\mu = 0.25$ or $\mu = 0.05$, it is enough to consider sequences of length 200 to obtain a 100% efficiency. A much smaller mean such as $\mu = 0.0005$ was also tested. In this case, an efficiency over 90% was only obtained for sequences of length ≥ 3000 .

The method presented here is by no means the unique form of using these invariants, so different ways of using them can even improve the tests.

Acknowledgments

Marta Casanellas was partially supported by RyC program of “Ministerio de Ciencia y Tecnologia”, BFM2003-06001 and BIO2000-1352-C02-02 of “Plan Nacional I+D” of Spain. Luis David Garcia was a postdoctoral fellow at the Mathematical Science Research Institute. Seth Sullivant was supported by a NSF graduate research fellowship. Much of the research in this chapter occurred during a visit to University of California, Berkeley, and we are

grateful for their hospitality. We like to thank Serkan Hoşten for all his help and guidance throughout this project.

References

- [1] M. Casanellas, L.D. Garcia, and S. Sullivant, *Small phylogenetic trees*, <http://www.math.tamu.edu/~lgp/small-trees/small-trees.html>, 2004.
- [2] S Evans and T Speed, *Invariants of some probability models used in phylogenetic inference*, *The Annals of Statistics* **21** (1993), 355–377.
- [3] K. St. John, T. Warnow, B. Moret, and L. Vawter, *Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor joining*, *Journal of Algorithms* **48** (2003), 174–193.
- [4] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004, 3-900051-07-0.
- [5] B Sturmfels and S Sullivant, *Toric ideals of phylogenetic invariants*, 2004.
- [6] L. A. Székely, M. A. Steel, and P. L. Erdős, *Fourier calculus on evolutionary trees*, *Adv. in Appl. Math.* **14** (1993), no. 2, 200–210. MR MR1218244 (94i:92009)
- [7] Z. Yang, *Paml: A program package for phylogenetic analysis by maximum likelihood*, *CABIOS* **15** (1997), 555–556.