# Toric Ideals of Phylogenetic Invariants

Bernd Sturmfels and Seth Sullivant
Department of Mathematics, University of California, Berkeley

**Abstract**

Statistical models of evolution are algebraic varieties in the space of joint probability distributions on the leaf colorations of a phylogenetic tree. The phylogenetic invariants of a model are the polynomials which vanish on the variety. Several widely used models for biological sequences have transition matrices that can be diagonalized by means of the Fourier transform of an abelian group. Their phylogenetic invariants form a toric ideal in the Fourier coordinates. We determine minimal generators and Gröbner bases for these toric ideals. For the Jukes-Cantor and Kimura models on a binary tree, our Gröbner basis consists of quadrics, cubics and quartics.

## 1 Introduction

Cavender and Felsenstein [3] and Lake [8] introduced phylogenetic invariants as an algebraic tool for reconstructing evolutionary trees from biological sequence data. Such invariants exist for any tree-based Markov model, and they uniquely characterize that model. While partial lists of invariants have been described for various models [refs here], the literature still conveys a sense that phylogenetic invariants and algebraic algorithms for computing them are not useful for any problem whose size is of biological interest. In his book *Inferring Phylogenies*, Felsenstein sums this up from the perspective of molecular biology as follows: *... (algebraic) invariants are worth attention, not for what they do for us now, but what they might lead to in the future...* [6, page 390]. A similar tone is expressed in the final section of the book *Phylogenetics* by Semple and Steel [10, page 212].

But "the future" could be closer than readers of these two excellent books might think. For the general Markov model, considerable progress has been made in the recent work of Allman and Rhodes [1, 2]. See [9, Conjecture 5] for a determinantal formula for the Allman-Rhodes ideal in the binary case. The present paper is not concerned with the general Markov model but with a class of special models, namely, the *group-based models* [10, §8.10]. The problem of finding invariants for these models was studied by many authors including Evans-Speed [4], Székely-Steel-Erdös [13], Steel-Fu [11] and Evans-Zhou [5]. The class of group-based models includes the *Jukes-Cantor model*, for either binary or DNA sequences, and the *Kimura models*, with two or three parameters.

The main result of this paper is an explicit description of a Gröbner basis and a generating set for the ideal of phylogenetic invariants of such a model. Here is a rough statement:

**Theorem 1.** *For any group based model on a phylogenetic tree $T$, the prime ideal of phylogenetic invariants is generated by the invariants of the local submodels around each interior node of $T$, together with the quadrics which express conditional independence statements along the splits of $T$.*

The precise form of this theorem and its proof will be given in Section 5. We continue here by reviewing Markov models on trees and by stating our results for some well-known group-based models. Let $T$ be a rooted tree with $m$ leaves and $\mathcal{V}(T)$ denote the set of nodes of $T$. To each node $v \in \mathcal{V}(T)$ we associate a $k$-ary random variable $X_v$. For biological reasons the most common values of $k$ are $2, 4$, and 20: these correspond to random variables which encode base pairs, individual nucleotides, or amino acids in a protein. The probability $P(X_v = i)$ is the probability that $X_v$ is in state $i$: in applications to DNA sequences this probability represents the proportion of characters in the sequence at $v$ which is a particular nucleotide, namely, $A$, $C$, $G$ or $T$.

The relationship between the random variables $X_v$ is encoded by the structure of the tree. Let $\pi$ be a distribution of the random variable $X_r$ at the root node $r$. For each node $v \in \mathcal{V}\backslash\{r\}$, let $a(v)$ be the unique parent of $v$. The transition from $a(v)$ to $v$ is given by a $k \times k$-matrix $A^{(v)}$ of probabilities. Then the probability distribution at each node is computed recursively by the rule

$$P(X_v = i) \quad = \quad \sum_{j=1}^{k} A_{ij}^{(v)} \cdot P(X_{a(v)} = j). \tag{1}$$

This rule induces a joint distribution on all the random variables $X_v$. We label the leaves of $T$ by $1, 2, \ldots, m$, and we abbreviate the joint distribution on the variables at the leaves as follows:

$$p_{i_1 i_2 \ldots i_m} \quad = \quad P(X_1 = i_1, X_2 = i_2, \ldots, X_m = i_m). \tag{2}$$

In biological applications, one estimates (some of) these $k^m$ probabilities from $m$ aligned sequences on $k$ letters, and the aim is to reconstruct the tree. The root distribution $\pi$ and the transition matrices $A^{(v)}$ are typically unknown. In the general Markov model of [1], each matrix entry $A_{ij}^{(v)}$ is an independent model parameter. For the group-based models, to be studied in this paper, the number of model parameters is smaller because some of the entries of $A^{(v)}$ are assumed to coincide.

A *phylogenetic invariant* of the model is a polynomial in the leaf probabilities $p_{i_1 i_2 \cdots i_m}$ which vanishes for every choice of model parameters. The set of these polynomials forms a prime ideal in the polynomial ring over the unknowns $p_{i_1 i_2 \cdots i_m}$, Our objective is to compute this ideal as explicitly as possible. In the language of algebraic geometry, we seek to determine the variety parametrized by the rational map induced by joint distribution on the leaves. The study of such varieties for various statistical models is a central theme in the emerging field of *algebraic statistics* [refs here].

In this paper, we determine the ideal of invariants for models whose structure is governed by an abelian group. Four models used in computational biology have this structure: the Jukes-Cantor models and the Kimura models. Theorem 2 below summarizes our results for these models.

The *Jukes-Cantor model* on two bases ($k = 2$) is the model with transition matrices

$$A^{(v)} \quad = \quad \begin{pmatrix} 1 - a_v & a_v \\ a_v & 1 - a_v \end{pmatrix},$$

where $a_v$ is the probability of making a transition between the states along the edge from $a(v)$ to

$v$. The *Kimura 3 parameter model* on $k = 4$ bases (for DNA sequences) has the transition matrices

$$A^{(v)} \quad = \quad \begin{pmatrix} 1-a_v-b_v-c_v & a_v & b_v & c_v \\ a_v & 1-a_v-b_v-c_v & c_v & b_v \\ b_v & c_v & 1-a_v-b_v-c_v & a_v \\ c_v & b_v & a_v & 1-a_v-b_v-c_v \end{pmatrix}$$

where $a_v$ is the probability of a transition and the $b_v$ and $c_v$ are transversion probabilities. The Kimura 2-parameter model arises as the subvariety defined by taking $b_v = c_v$ for all $v$ and the Jukes-Cantor 4 base model is the subvariety defined by setting $a_v = b_v = c_v$ for all $v$.

Evans and Speed [4] introduced a linear change of coordinates, based on the discrete Fourier transform, which diagonalizes the parametrization of these models. In Section 2 we will review this construction at the level of generality proposed by Székely-Steel-Erdös [13]. The crucial idea is to label the states of the random variables $X_v$ by a finite abelian group ($\mathbb{Z}_2$ for the Jukes-Cantor 2-base model and $\mathbb{Z}_2 \times \mathbb{Z}_2$ for the other three models) in such a way that the probability of transitioning from $g_i$ to $g_j$ is seen to depend only on the difference $g_i - g_j$. Replacing the original coordinates $p_{i_1 \cdots i_m}$ by Fourier coordinates $\widehat{p}_{i_1 \cdots i_m}$, the ideal of phylogenetic invariants becomes a toric ideal. Recall (e.g. from [12]) that a *toric ideal* is a prime ideal generated by differences of monomials.

As an example consider the Jukes-Cantor 2-base model for $m = 4$. The Fourier coordinates are

$$\widehat{p}_{ijkl} \quad = \quad \sum_{r=0}^{1} \sum_{s=0}^{1} \sum_{t=0}^{1} \sum_{u=0}^{1} (-1)^{ir+js+kt+lu} \cdot p_{rstu}, \qquad \text{where } i,j,k,l \in \mathbb{Z}_2. \tag{3}$$

If $T$ is the balanced binary tree of height two, then this model has the parametric representation

$$\widehat{p}_{ijkl} \quad \mapsto \quad a_i \cdot b_{i+j} \cdot c_j \cdot d_{i+j+k+l} \cdot e_k \cdot f_{k+l} \cdot g_l. \tag{4}$$

Disregarding the trivial invariant $\widehat{p}_{0000} - 1$, the toric ideal of phylogenetic invariants is generated by 20 linearly independent quadrics. These arise as the $2 \times 2$-minors of the four $2 \times 4$-matrices

$$\begin{pmatrix} \widehat{p}_{0i00} & \widehat{p}_{0i01} & \widehat{p}_{0i10} & \widehat{p}_{0i11} \\ \widehat{p}_{1i00} & \widehat{p}_{1i01} & \widehat{p}_{1i10} & \widehat{p}_{1i11} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widehat{p}_{00i0} & \widehat{p}_{01i0} & \widehat{p}_{10i0} & \widehat{p}_{11i0} \\ \widehat{p}_{00i1} & \widehat{p}_{01i1} & \widehat{p}_{10i1} & \widehat{p}_{11i1} \end{pmatrix} \quad \text{for } i = 0, 1. \tag{5}$$

Moreover, these quadrics form a Gröbner basis for a suitable term order. This generalizes as follows:

**Theorem 2.** *Let $T$ be an arbitrary binary rooted tree. Modulo the trivial invariant $\widehat{p}_{00\cdots0} - 1$,*

*(a) the ideal of the Jukes Cantor 2-base model is generated by polynomials of degree 2,*

*(b) the ideal of the Jukes Cantor 4-base model is generated by polynomials of degree 1, 2 and 3,*

*(c) the ideal of the Kimura 2-parameter model is generated by polynomials of degree 1, 2, 3 and 4,*

*(d) the ideal of the Kimura 3-parameter model is generated by polynomials of degree 2, 3 and 4.*

*Each of these generating sets has an explicit combinatorial description and it is a Gröbner basis.*

The outline for the paper is as follows. In the next section we discuss the question whether one really needs a full set of generators for the ideal of phylogenetic invariants. We argue that the answer is affirmative, by showing that, in contrast to what was suggested in [4], [7], and [14], for most models it does not suffice to take algebraically independent invariants. This theme will be picked up again in Section 7, where we describe how our Gröbner bases might be used for possible biological applications. In Section 3 we review the Fourier transform technique introduced by Evans and Speed [4] for diagonalizing group based models. This is done for arbitrary finite abelian groups, as in [13], and it reduces to our problem to computing the kernel of a monomial map as in (4).

Section 4 turns rooted trees on $m$ leaves into unrooted trees on $m+1$ leaves, and it introduces "friendly labelings" on abelian groups. These labelings are used to classify the linear model invariants, and to set up a coordinate system modulo the linear invariants. For the Jukes-Cantor 4-base model, this is precisely the construction involving Fibonacci numbers and sub-forests due to Steel and Fu [11].

In Section 5 we state and prove the precise form Theorem 1, our main result, both for binary and non-binary trees. Theorem 2 is derived as a corollary in Section 6. The generators and Gröbner bases of the of the Jukes-Cantor ideals and Kimura ideals are described in explicit combinatorial terms. Conclusions, algorithmic implications and open problems are presented in Section 7.

## 2 How Many Invariants are Needed ?

Each algebraic variety $X$ to be discussed in this paper lives in an ambient space of $k^m$ dimensions. The coordinates of the ambient space are the probabilities $p_{i_1 i_2 \cdots i_m}$, or their Fourier transforms $\widehat{p}_{i_1 i_2 \cdots i_m}$, which will be defined in general in the next section. The dimension of the model $X$ is the number of algebraically independent model parameters, and the *codimension* of the model $X$ is

$$\operatorname{codim}(X) \quad = \quad k^m \,-\, \dim(X).$$

This is the number of local equations needed to describe the variety $X$ at a smooth point **??**. However, in general, the number of equations needed to describe $X$ at a singular point, or the number of equations needed to define a variety $X$ globally, can be much larger than the codimension of $X$.

Several papers on phylogenetic invariants give the impression that to characterize a model $X$, it suffices to take only $\operatorname{codim}(X)$ polynomial invariants, and some authors raised the question whether there is a complete list of algebraically independent invariants. We wish to argue that, from the perspective of algebraic geometry, it is quite misleading to ask for only $\operatorname{codim}(X)$ polynomial invariants. Most models in algebraic statistics, including the group-based evolutionary models treated in this paper, require many more polynomial equations than their codimension, even if one is only interested in strictly positive probability distributions. In our view, a given system of polynomial invariants cannot be considered "complete" unless it generates the prime ideal of $X$.

We illustrate this issue for the case when $X$ is the Jukes-Cantor 2-base model for $m = 4$, with parametric representation given by (4). The variety $X$ has codimension 8. The homogeneous prime ideal of the model is given by the $2 \times 2$-minors of the four $2 \times 4$-matrices in (5). This ideal requires 20 minimal generators. Can we replace these 20 quadrics by a smaller subset? Don't eight suffice?

The answer is clearly "no" when $X$ is the complex variety defined by requiring that the matrices (5) have rank one. However, more than eight equations are needed even if we consider a small neighborhood of the centroid of the probability simplex. This centroid is the uniform distribution on the leaf colorations. In Fourier coordinates, this neighborhood is given by setting $\widehat{p}_{0000} = 1$ and by assuming that the other 15 coordinates $\widehat{p}_{ijkl}$ are real numbers of small absolute value.

If we add the trivial invariant $\widehat{p}_{0000} - 1$ to our 20 quadrics, then the resulting ideal in the polynomial ring in 15 unknowns still has codimension 8 but it is now minimally generated by ten equations. The first five of these ten equations express five of the unknowns in terms of the others:

$$\begin{aligned} &\widehat{p}_{1010} - \widehat{p}_{1000}\widehat{p}_{0010}\,, \quad \widehat{p}_{1001} - \widehat{p}_{1000}\widehat{p}_{0001}\,, \quad \widehat{p}_{1011} - \widehat{p}_{1000}\widehat{p}_{0011}\,, \\ &\widehat{p}_{0101} - \widehat{p}_{0001}\widehat{p}_{0100}\,, \quad \widehat{p}_{1001} - \widehat{p}_{0001}\widehat{p}_{1000}\,, \quad \widehat{p}_{1101} - \widehat{p}_{0001}\widehat{p}_{1100}. \end{aligned}$$

Here either the second or the fifth equation is redundant. What remains is an ideal of codimension three which is minimally generated by five homogeneous quadrics. The five remaining quadrics are the five $2 \times 2$-minors which do not involve the upper left corner in the following $3 \times 3$-matrix:

$$\begin{pmatrix} \bullet & \widehat{p}_{0010} & \widehat{p}_{0011} \\ \widehat{p}_{0100} & \widehat{p}_{0110} & \widehat{p}_{0111} \\ \widehat{p}_{1100} & \widehat{p}_{1110} & \widehat{p}_{1111} \end{pmatrix}$$

If we remove any of these five quadrics then the zero set of the remaining four equations contains points which are not in the model, even in a neighborhood the uniform distribution. For example, we get extraneous solutions by placing random small reals $\epsilon_{ijkl}$ in the matrices

$$\begin{pmatrix} \bullet & 0 & 0 \\ 0 & \epsilon_{0110} & \epsilon_{0111} \\ 0 & \epsilon_{1110} & \epsilon_{1111} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \bullet & \epsilon_{0010} & 0 \\ \epsilon_{0100} & \epsilon_{0110} & 0 \\ \epsilon_{1100} & \epsilon_{1110} & 0 \end{pmatrix}$$

Notice that matrices with these entries are near the centroid of the probability simplex and satisfy all but one of the five $2 \times 2$-minors of the matrix. Thus we need all five quadrics to define our variety, even set-theoretically, and even locally around the uniform distribution. We regard the determinantal formula (5) as the best representation of the ideal of phylogenetic invariants, even though it involves more than $\operatorname{codim}(X) = 8$ polynomials.

The failure to describe a phylogenetic model $X$ set-theoretically becomes much more dramatic if we replace the ideal generators derived in this paper with the *canonical invariants* introduced by Székely, Steel and Erdös [13]. The number of canonical invariants is always equal to the codimension of $X$, but, as we have argued, this means that they are far from having the correct zero set. For the specific Jukes-Cantor 2-base model with $m = 4$ discussed above, there are eight canonical

invariants. From [13, Theorem 10], we see that they are the following binomials of degree eight:

$$\widehat{p}_{0000}\widehat{p}_{0010}\widehat{p}_{0100}\widehat{p}_{0110}\widehat{p}_{1001}\widehat{p}_{1011}\widehat{p}_{1101}\widehat{p}_{1111} - \widehat{p}_{0001}\widehat{p}_{0011}\widehat{p}_{0101}\widehat{p}_{0111}\widehat{p}_{1000}\widehat{p}_{1010}\widehat{p}_{1100}\widehat{p}_{1110},$$
$$\widehat{p}_{0000}\widehat{p}_{0010}\widehat{p}_{0101}\widehat{p}_{0111}\widehat{p}_{1000}\widehat{p}_{1010}\widehat{p}_{1101}\widehat{p}_{1111} - \widehat{p}_{0001}\widehat{p}_{0011}\widehat{p}_{0100}\widehat{p}_{0110}\widehat{p}_{1001}\widehat{p}_{1011}\widehat{p}_{1100}\widehat{p}_{1110},$$
$$\widehat{p}_{0000}\widehat{p}_{0010}\widehat{p}_{0101}\widehat{p}_{0111}\widehat{p}_{1001}\widehat{p}_{1011}\widehat{p}_{1100}\widehat{p}_{1110} - \widehat{p}_{0001}\widehat{p}_{0011}\widehat{p}_{0100}\widehat{p}_{0110}\widehat{p}_{1000}\widehat{p}_{1010}\widehat{p}_{1101}\widehat{p}_{1111},$$
$$\widehat{p}_{0000}\widehat{p}_{0011}\widehat{p}_{0100}\widehat{p}_{0111}\widehat{p}_{1001}\widehat{p}_{1010}\widehat{p}_{1101}\widehat{p}_{1110} - \widehat{p}_{0001}\widehat{p}_{0010}\widehat{p}_{0101}\widehat{p}_{0110}\widehat{p}_{1000}\widehat{p}_{1011}\widehat{p}_{1100}\widehat{p}_{1111},$$
$$\widehat{p}_{0000}\widehat{p}_{0011}\widehat{p}_{0101}\widehat{p}_{0110}\widehat{p}_{1001}\widehat{p}_{1010}\widehat{p}_{1100}\widehat{p}_{1111} - \widehat{p}_{0001}\widehat{p}_{0010}\widehat{p}_{0100}\widehat{p}_{0111}\widehat{p}_{1000}\widehat{p}_{1011}\widehat{p}_{1101}\widehat{p}_{1110},$$
$$\widehat{p}_{0000}\widehat{p}_{0011}\widehat{p}_{0101}\widehat{p}_{0110}\widehat{p}_{1000}\widehat{p}_{1011}\widehat{p}_{1101}\widehat{p}_{1110} - \widehat{p}_{0001}\widehat{p}_{0010}\widehat{p}_{0100}\widehat{p}_{0111}\widehat{p}_{1001}\widehat{p}_{1010}\widehat{p}_{1100}\widehat{p}_{1111},$$
$$\widehat{p}_{0000}\widehat{p}_{0001}\widehat{p}_{0110}\widehat{p}_{0111}\widehat{p}_{1010}\widehat{p}_{1011}\widehat{p}_{1100}\widehat{p}_{1101} - \widehat{p}_{0010}\widehat{p}_{0011}\widehat{p}_{0100}\widehat{p}_{0101}\widehat{p}_{1000}\widehat{p}_{1001}\widehat{p}_{1110}\widehat{p}_{1111},$$
$$\widehat{p}_{0000}\widehat{p}_{0001}\widehat{p}_{0100}\widehat{p}_{0101}\widehat{p}_{1010}\widehat{p}_{1011}\widehat{p}_{1110}\widehat{p}_{1111} - \widehat{p}_{0010}\widehat{p}_{0011}\widehat{p}_{0110}\widehat{p}_{0111}\widehat{p}_{1000}\widehat{p}_{1001}\widehat{p}_{1100}\widehat{p}_{1101}.$$

The zero set of these equations has codimension three (!), and has many irreducible components. The structure of the primary decomposition of the ideal of canonical equations is very complicated. For instance, among the irreducible components, there are 48 linear spaces of codimension three, e.g.

$$\widehat{p}_{1001} \;=\; \widehat{p}_{1000} \;=\; \widehat{p}_{1010} \;=\; 0.$$

Among all the probability distributions which satisfy these invariants, the distributions which come from the models are a very low dimensional portion. The canonical equations correspond to a lattice basis for the toric ideal of phylogenetic invariants. It follows from general theory in commutative algebra the toric ideal can computed from the canonical equations by the process of saturation (as described in [12, Algorithm 12.3]), but this is non-trivial and time-consuming computation. What we offer in this paper is an explicit description of a list of phylogenetic invariants which minimally generates the toric ideals of interest. But in all cases (with the exception of a few trivial ones), the number of our polynomial invariants will be considerably larger than the codimension of the model, a feature which is unavoidable in algebraic geometry.

## 3 A Linear Change of Coordinates

In this section we describe the Fourier transform, which is an orthogonal linear change of coordinates that turns the irreducible variety of distributions of a group-based model into a toric variety. We refer the reader to [13] for detailed version of the proofs we describe below.

Let $G$ be a finite Abelian group, of order $|G| = k$. The *dual group* $\widehat{G}$ (or *character group*) of $G$ is defined as

$$\widehat{G} = \mathrm{Hom}(G, \mathbb{C}^\times),$$

the group of all group homomorphisms from $G$ into the multiplicative group of complex numbers. The elements of $\widehat{G}$ are called characters of $G$ and a typical element of $\widehat{G}$ is denoted by the letter $\chi$.

Given any function $f : G \to \mathbb{C}$, the Fourier transform $\hat{f}$ of $f$ (over $G$) is the function $\hat{f} : \widehat{G} \to \mathbb{C}$ defined by

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g) f(g).$$

Given two functions $f_1$ and $f_2$ on $G$, their *convolution* $f_1 * f_2$ is the new function

$$f_1 * f_2(g) = \sum_{h \in G} f_1(h) f_2(g-h).$$

The main facts we will need about the dual group, the Fourier transform, and convolution products is summarized in the following lemma.

**Lemma 3.** *Let $G$ be an abelian group and $f_1, f_2$ two function from $G$ to $\mathbb{C}$ and $\mathbf{1}$ the constant function. Then*

1. $G \cong \widehat{G}$,
2. $\widehat{f_1 * f_2} = \widehat{f_1} \cdot \widehat{f_2}$, *and*
3. $$\widehat{\mathbf{1}}(\chi) = \begin{cases} |G| & \text{if } \chi \equiv 1 \\ 0 & \text{otherwise} \end{cases}$$

Our interest in the Fourier transform comes from the fact that it can be used to simplify the parameterizations which arise from the Jukes-Cantor and Kimura models. This simplification was originally described in (*EVANS - SPEED*) but we will provide an elementary proof of this fact.

Recall that the joint distribution of every model for phylogenetic evolution has the form

$$P(X_1 = g_1, \ldots, X_m = g_m) = p_{g_1,\ldots,g_m} = \sum_{v \text{ not a leaf}} \pi_{g_r} \prod_{v \in \mathcal{V}(T) \setminus r} A^{(v)}_{g_{a(v)}, g_v}.$$

In the case of any group based model, we may write this more conveniently (for the ease of proof) as

$$p(g_1, \ldots, g_m) = \sum_{v \text{ not a leaf}} \pi(g_r) \prod_{v \in \mathcal{V}(T) \setminus r} f^{(v)}(g_{a(v)} - g_v).$$

The main theorem about discrete Fourier analysis and group based models is that that the Fourier tranform of the joint distribution has a parametrization that can be written in product form. In particular, the variety of distributions for a phylogenetic model with group structure is a toric variety. Note that if the abelian group for the group model is any group besides $\mathbb{Z}_2^l$, this implies that even though the ideal of phylogenetic invariants is guaranteed to have generators with rational coordinates, there is a natural and generally simpler generating set which contains polynomials with complex coefficients. Before we state and prove the main result about the Fourier transform for group models, we will illustrate the ideal with a small example.

**Example 4.** Let $T$ be the tree $K_{1,m}$ whose only nodes are the $m$ leaves and the root. The joint probability of a group model at the leaves is given by

$$p(g_1, g_2, \ldots, g_m) = \sum_{h \in G} \pi(h) \prod_{i=1}^{m} p_i(g_i - h).$$

We will take the Fourier transform of this probability density with respect to the group $G^m$. To do this, we introduce the new function $\widetilde{\pi}$ defined as

$$\widetilde{\pi}(h_1, \ldots, h_m) = \begin{cases} \pi(h_1) & \text{if } h_1 = h_2 = \cdots = h_m \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$p(g_1, g_2, \ldots, g_m) = \sum_{(h_1, \ldots, h_m) \in G^m} \tilde{\pi}(h_1, \ldots, h_m) \prod_{i=1}^{m} p_i(g_i - h_i),$$

so that $p$ is a convolution over $G^m$. Taking the Fourier transform yields

$$\widehat{p}(\chi_1, \ldots, \chi_m) = \widehat{\widetilde{\pi}}(\chi_1, \ldots, \chi_m) \prod_{i=1}^{m} \widehat{p}_i(\chi_i)$$

by the convolution formula and using the independence of the $p_i$ in the Fourier transform. Furthermore

$$\begin{aligned} \widehat{\widetilde{\pi}}(\chi_1, \ldots, \chi_m) &= \sum_{(g_1, \ldots, g_m) \in G^m} \langle (\chi_1, \ldots, \chi_m), (g_1, \ldots, g_m) \rangle \cdot \widetilde{\pi}(g_1, \ldots, g_m) \\ &= \sum_{g \in G} \chi_1 \chi_2 \cdots \chi_m(g) \pi(g) = \widehat{\pi}(\chi_1 \chi_2 \cdots \chi_m), \end{aligned}$$

and hence

$$\widehat{p}(\chi_1, \ldots, \chi_m) = \widehat{\pi}(\chi_1 \cdots \chi_m) \prod_{i=1}^{m} \widehat{p}_i(\chi_i).$$

Example 4 is the first step in the induction needed to prove the following crucial result.

**Lemma 5.** *Let $p(g_1, \ldots, g_m)$ be the joint distribution parameterized by a group based model for the tree $T$. Then the Fourier transform of $p$ has the form*

$$\widehat{p}(\chi_1, \ldots, \chi_m) = \widehat{\pi} \chi_1 \cdots, \chi_m) \prod_{v \in \mathcal{V}(T) \setminus r} \widehat{p}_v( \prod_{l \in \mathcal{L}(v)} \chi_l)$$

*where $\mathcal{L}(v)$ is the set of leaves which have $v$ as a common ancestor.*

## 4 Edge Labellings and Linear Invariants

At the end of the previous section, we presented a monomial parameterization for group based models where each set of parameters is focused at the vertices. In this section, we will first change the problem slightly by adding a new edge to our tree (at the root) which has the effect of associating parameters to each edge of the tree. Then to each leaf coloration by group elements $(g_1, \cdots, g_m)$ we associate a labelled tree. These labelled trees have the advantage of providing a description of the linear invariants of these models as well a setting up a convenient coordinate system modulo the linear invariants.

Let $G$ be a finite abelian group, and $T$ a rooted tree with $m$ leaves. We are interested in understanding the kernel of the monomial parameterization given by

$$\widehat{p}(\chi_1, \ldots, \chi_m) = \widehat{\pi}(\chi_1 \cdots, \chi_m) \prod_{v \in \mathcal{V}(T) \setminus r} \widehat{p}_v( \prod_{l \in \mathcal{L}(v)} \chi_l).$$

This is a toric ideal in the Fourier coordinates $\widehat{p}$. We will see later that this corresponds to an ideal of phylogenetic invariants for the group based models we are interested in. Since $G \cong \widehat{G}$ we replace the characters $\chi_i$ by $g_i$ and the products of characters by sums of elements in the abelian group. That is, we are interested in understanding the kernel of the monomial parameterization

$$\widehat{p}(g_1, \ldots, g_m) = \widehat{\pi}(g_1, \cdots, g_m) \prod_{v \in \mathcal{V}(T) \setminus r} \widehat{f}_v( \sum_{l \in \mathcal{L}(v)} g_l).$$

This parameterization has the structure of requiring a set of parameters for each node of the tree. We add an extra edge at the root of $T$ to achieve a new tree $T'$ with $m+1$ leaves, and now we associate a set of parameters to each edge of the tree, by "moving" the parameters from a given vertex to the edge directly above it. The root node

Given an assignment of group elements $(g_1, \ldots, g_m)$ to the $m$ leaves of $T$, we get, for each edge $e$ of $T'$ an assignment of a group element $A(e)$. The assignment function $A(e)$ is defined by

$$A(e) = \sum_{g_l \in \mathcal{L}(e)} g_l$$

where $\mathcal{L}(e)$ is the set of leaves below $e$. So we are interested in studying the kernel of the monomial parameterization

$$\widehat{p}(g_1, \ldots, g_m) = \prod_{e \in E(T')} f_e(A(e))$$

where we have eliminated the special distinction of the root distribution. Now the function $f_e$ might not be one-to-one and this is where we introduce labellings. Let $\mathcal{L}$ be a set of labels and $L$ a labelling function

$L : G \to \mathcal{L}$.

The labelling function $L$ encodes the instances where $f_e(g_i) = f_e(g_j)$. *For the time being, we will assume that the labelling function associated to each edge of the tree is the same for every edge.*

However, we will show later that this assumption can be dropped in some special instances. Given this labelling function, we wish to understand the kernel of the following monomial parameterizations

$$\phi_{G,T,L} : \mathbb{C}[\widehat{p}_{g_1,\ldots,g_m}] \to \mathbb{C}[a_{l_i}^{(e)} | e \in E(T), l_i \in \mathcal{L}$$
$$\widehat{p}_{g_1,\ldots,g_m} \longmapsto \prod_{e \in E(T)} a_{L(A(e))}^{(e)}.$$

where the $a_{l_i}^{(e)}$ are the parameters coming from each edge $e$. The kernel of this ring homomorphism is a toric ideal of phylogenetic invariants in the Fourier transform of the probabilities. We denote this ideal by $I_{T,L}$ suppressing dependence on the group $G$.

From this description, we immediately can deduce the structure of the linear invariants of these ideals.

**Theorem 6 (Linear Invariants).** *Every linear invariant of the ideal $I_{T,L}$ is of the form*

$$\widehat{p}_{g_1,ldots,g_m} = \widehat{p}_{h_1,\ldots,h_m}$$

*where $(g_1, \ldots, g_m)$ and $(h_1, \ldots, h_m)$ induce the same edge labelling on $T'$.*

The labelling function $L$ induces a map

$$\widetilde{L} : G^m \to \mathcal{L}^{|E(T')|},$$

and we denote by $C(T, L)$ the image of this map which we call the set of consistent labellings. We are then left with the problem of studying the kernel of the ring map

$$\widetilde{\phi}_{G,T,L} : \mathbb{C}[\widehat{p}_l | l \in C(T, L)] \to \mathbb{C}[a_{\ell_i}^{(e)} | \ell_i \in \mathcal{L}]$$
$$\widehat{p}_L \mapsto \prod_{e \in E(T)} a_{\ell_e}^{(e)}.$$

The kernel of this ring map is the ideal $I_{T,L}$ modulo linear invariants.

While we would like to understand $I_{T,L}$ for every labelling function, we will restrict attention to a class of labelling functions which arises naturally in the course of studying phylogenetic invariants: the friendly labellings.

**Definition 7.** Let $L$ be a labelling function $L : G \to \mathcal{L}$. For $m \geq 3$ let $Z \subset G^m$ be the set

$$Z = \{(g_1, \ldots, g_m) \in G^m | \sum_{i=1}^{m-1} g_i = g_m\}.$$

Consider the induced map $\widetilde{L} : Z \subset G^m \to \mathcal{L}^m$ and denote by $\pi_i$ the projection $\pi_i : G^n \to G$ onto the $i$-th coordinate. The function $L$ is called *$m$-friendly* if for every $l = (l_1, \ldots, l_m) \in \widetilde{L}(Z) \subset \mathcal{L}^m$ and for all $i$,

$$\pi_i(\widetilde{L}^{-1}(l)) = L^{-1}(l_i).$$

A labelling function is *friendly* if it is $m$-friendly for all $m \geq 3$.

**Lemma 8.** *Labelling functions that are* 3*-friendly are friendly.*

Lemma 8 says that checking whether a labelling is friendly can be done simply with a finite computation. The point of studying friendly labellings is that consistent labellings "glue" together. We will now make this statement explicit. Let $e$ be an internal edge of the tree $T'$. Denote by $T_{e^-}$ the tree obtained from $T'$ by taking the edge $e$ and all the edges below $e$. Denote by $T_{e^+}$ the tree obtained from $T'$ be taking the edge $e$ and all edges not in $T_{e^-}$. Then we have the following

**Lemma 9.** *Let $l^1 \in C(T_{e^-}, L)$ and $l^2 \in C(T_{e^+}, L)$ and suppose that the label assigned to edge $e$ in both $l^1$ and $l^2$ is the same: $l^1_e = l^2_e$. Then the labelling $l$ for $T'$ obtained from $l^1$ and $l^2$ by labelling edges of $T'$ appropriately is consistent: $l \in C(T, L)$.*

Lemma 9 is surprisingly simple but it is the main technical result upon which all our combinatorial constructions of generators and Gröbner bases rest. Indeed, as we will see, it implies that phylogenetic invariants of group based models with friendly labellings are only determined by local features of the tree.

We now conclude this section with some examples of friendly labellings.

**Example 10.** Let $G$ be any group. Any function $L : G \to \mathcal{L}$ that is injective is friendly for trivial reasons (all the sets described in the definition of friendly are one element sets and trivially nonempty). For similar reasons, if $\mathcal{L}$ consists of elements of a group and $L$ is a group homomorphism then $L$ is friendly.

Let $\mathcal{L} = \{0, 1\}$ and $L$ be the function

$$L(g) = \begin{cases} 0 & \text{if } g = id \\ 1 & \text{otherwise} \end{cases}$$

which we call the Jukes-Cantor labelling function. This $L$ is friendly for any group. Note that this labelling only corresponds to the usual Jukes-Cantor models when $G = \mathbb{Z}_2^r$.

Finally, suppose that $G = \mathbb{Z}_2^2$, $\mathcal{L} = \{0, 1, 2\}$ then the labelling function $f$ with

$$L((0,0)) = 0, L((0,1)) = 1, L((1,0)) = L((1,1)) = 2$$

is friendly. We call this labelling function the Kimura 2-parameter labelling function.

Note that most labelling function are, in fact, not friendly and non-friendly labellings are easy to construct.

**Example 11.** Let $G = \mathbb{Z}_4$ and $\mathcal{L} = \{0, 1, 2\}$. Then the labelling function $L$ defined by

$$L(0) = 0, f(1) = 1, L(2) = L(3) = 2$$

is not friendly.

## 5 The Main Result

We will now state and prove our main result which shows how to build Gröbner bases and generating sets for group based models with friendly labelling functions out of purely local information in the tree. We express all the necessary nonlinear polynomials in terms of the labelled variables $\widehat{p}_l$. *Throughout this section we always assume the $L$ is a friendly labelling.*

For the ease of notation and proof, we will write binomials the the $\widehat{p}_l$ variables using *tableau notation*. That is, for any monomial $M = \widehat{p}_{l^1}\widehat{p}_{l^2}\cdots\widehat{p}_{l^d}$ we encode this monomial as a $d \times |E(T')|$ matrix of labels

$$M = \begin{bmatrix} l^1 \\ l^2 \\ \vdots \\ l^d \end{bmatrix}$$

where each column of the matrix is indexed by a particular edge of the tree $T'$. Binomials $M - M'$ are represented as formal differences of tableau. Notice that given a formal difference of tableau representing a binomial $b = M - M'$, it is easy to check whether or not $b \in I_{T,L}$. First, each row of $M$ and $M'$ should be a consistent labelling. Second, for each edge of $T'$ the multiset of labels appearing should be the same in the respective columns of $M$ and $M'$. We are now ready to construct the binomials that will constitute the Gröbner bases and generating sets that we have hinted at throughout.

If $e$ is an interior edge of $T'$ then we can consider the trees $T_{e^-}$ and $T_{e^+}$ which we described in the previous section. Without loss of generality, we may write our tableau in three grouped columns as

$$M = \begin{bmatrix} l^1 & m^1 & n^1 \\ l^2 & m^2 & n^2 \\ \vdots & \vdots & \vdots \\ l^d & m^d & n^d \end{bmatrix}$$

where the left-most columns correspond to the edges in $T_{e^-} \setminus \{e\}$, the middle index corresponds to the edge $e$ and the right-most columns correspond to the edges in $T_{e^-} \setminus \{e\}$.

**Lemma 12.** *Let $(l_1, m, n_1)$ and $(l_2, m, n_2)$ be labellings of $T$ in $C(f, T)$. Then*

$$g = \begin{bmatrix} l^1 & m & n^1 \\ l^2 & m & n^2 \end{bmatrix} - \begin{bmatrix} l^1 & m & n^2 \\ l^2 & m & n^1 \end{bmatrix}$$

*is a binomial in $I_{T,f}$.*

**Definition 13.** Denote by $Ind(T_{e^+}, T_{e^-})$ the union of all the quadratic binomials from Lemma 17.

Now suppose that $v$ is any interior vertex of $T'$. Without loss of generality, we may write a tableau as

$$M = \begin{bmatrix} (l^1, x^1) & (m^1, y^1) & \cdots & (n^1, z^1) \\ (l^2, x^2) & (m^2, y^2) & \cdots & (n^2, z^2) \\ \vdots & \vdots & \vdots & \vdots \\ (l^d, x^d) & (m^d, y^d) & \cdots & (n^d, z^d) \end{bmatrix}$$

where the total number of column groups is equal to the number of edges which are incident to $v$ and the groups are labelled by these edges. In each grouping the first group of labels corresponds to all the edges of $T'$ which are on the side of the corresponding edge away from $v$, and the second set of labels is the single label corresponding to the edge incident to $v$. For each interior vertex, we can form the subtree $T_v$ which consists of only the edges incident to $v$.

**Lemma 14.** *Let $M - M'$ written in tableau notation as*

$$M - M' = \begin{bmatrix} (l^1, x^1) & (m^1, y^1) & \cdots & (n^1, z^1) \\ (l^2, x^2) & (m^2, y^2) & \cdots & (n^2, z^2) \\ \vdots & \vdots & \vdots & \vdots \\ (l^d, x^d) & (m^d, y^d) & \cdots & (n^d, z^d) \end{bmatrix}$$

INSERT DESCRIPTION OF LOCAL EXT
INSERT MAIN THEOREM
******

We will describe a gluing theorem which reduces the computation of the generators of these ideals to finding generators of the corresponding ideals for smaller trees. The main idea of the construction is to break the problem into smaller pieces along the splits in the graph.

Let $T$ be a tree with at least one interior edge (a non-leaf edge) $e$. Define the split trees associated to $e$ to be the trees $T_{e^+}$ and $T_{e^-}$ the two trees which are obtained from $T$ by having $e$ by a leaf on the left or the right.

Binomials in $I_{T,f}$ can be represented by tableau. That is, to each monomial in the labelled variables, we associate a matrix where the columns are indexed by the edges of the tree and the rows consists of the labelled variables. Binomials in $I_{T,f}$ are represented by formal difference of the tableau corresponding to the monomials for the terms.

If we fix a specific interior edge $e$ we can always write our tableaux so that the edges from the tree $T_{e^+}$ appear on the left half of the tableau and the edges for the tree $T_{e^-}$ appear on the right half of the tableau. That is, the tableau for a binomial $g$ can be arranged to look like

$$g = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m'_1 & n'_1 \\ \vdots & \vdots & \vdots \\ l'_d & m'_d & n'_d \end{bmatrix}$$

where the $m_i$ and $m'_i$ are single labels corresponding to the edge $e$, the $l_i$ and $l'_i$ are strings of labels corresponding to the edges in $T_{e^+} \setminus e$, and the $n_i$ and $n'_i$ are strings of labels corresponding to edges in $T_{e^-}$. Note that since $g$ belongs to the ideal $I_{T,f}$, the multiset of labels which appears on the edge

$e$ must be the same for both the leading and trailing terms of $g$. Hence, after rearranging the rows of the tableau we may write

$$g = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 & n'_1 \\ \vdots & \vdots & \vdots \\ l'_d & m_d & n'_d \end{bmatrix}.$$

Every binomial in $I_{T,f}$ restricts to a binomial in $I_{T_{e^+},f}$ by deleting columns in the tableau. This is, if $g$ is a binomial in $I_{T,f}$ as written above then the binomial

$$g|_{e^+} = \begin{bmatrix} l_1 & m_1 \\ \vdots & \vdots \\ l_d & m_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 \\ \vdots & \vdots \\ l'_d & m_d \end{bmatrix}$$

belongs to $I_{T_{e^+},f}$. Similarly, deleting the $l_i$ and $l'_i$ columns yields a binomial $g|_{e^-}$ in $I_{T_{e^-},f}$. There is also a constructive converse, from which binomials in $I_{T_{e^+},f}$ and $I_{T_{e^-},f}$ can be extended to binomials in $I_{T,f}$.

**Lemma 15.** *Let $g$ be a binomial in $I_{T_{e^+},f}$ written tableau notation as*

$$g = \begin{bmatrix} l_1 & m_1 \\ \vdots & \vdots \\ l_d & m_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 \\ \vdots & \vdots \\ l'_d & m_d \end{bmatrix}.$$

*Let $n_1, \ldots, n_d$ be sequences of labels such that each of $(m_i, n_i) \in C(f, T_{e^-})$. Then the binomial $g^*$ defined by*

$$g^* = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l'_d & m_d & n_d \end{bmatrix},$$

*belongs to $I_{T,f}$.*

*Proof.* Restricting the tableau to the tree $T_{e^+}$ and $T_{e^-}$ shows that the multiset of labels which appears on each edge are the same: $g^*|_{T_{e^+}} = g$ and $g^*|_{T_{e^-}} = 0$. We must also show that each of the labellings $(l_i, m_i, n_i)$ and $(l'_i, m_i, n_i)$ are in $C(f, T)$ so that $g^*$ actually represents a binomial in $\mathbb{C}[\widehat{p}_L]$. This follows from the fact that $(l_i, m_i) \in C(f, T_{e^+})$ and $(m_i, n_i) \in C(f, T_{e^i})$ and since $f$ was assumed to be friendly $(l_i, m_i, n_i) \in C(f, T)$. □

Of course, a similar extension procedure works for binomials in $I_{T_{e^-},f}$.

**Definition 16.** Let $G \subset I_{T_{e^+},f}$ be a collection of binomials. We define $Ext(G \to T)$ to be the union of the binomials $g^*$ from Lemma 15 as $g$ ranges over $G$. Similarly define $Ext(T \leftarrow G)$ for $G \subset I_{T_{e^-},f}$ a collection of binomials.

The second part of our construction is a family of quadratic binomials associated to the interior edge $e$.

**Lemma 17.** *Let $(l_1, m, n_1)$ and $(l_2, m, n_2)$ be labellings of $T$ in $C(f, T)$. Then*

$$g = \begin{bmatrix} l_1 & m & n_1 \\ l_2 & m & n_2 \end{bmatrix} - \begin{bmatrix} l_1 & m & n_2 \\ l_2 & m & n_1 \end{bmatrix}$$

*is a binomial in $I_{T,f}$.*

*Proof.* Clearly the construction guarantees that the same set of indices appears on each edge. We must show that the index labels in each tableau actually correspond variables in the ring $\mathbb{C}[\widehat{p}_L]$. However, since $(l_1, m, n_1)$ and $(l_2, m, n_2)$ are labellings of $T$ in $C(f, T)$ then $(l_1, m)$ and $(l_2, m)$ are in $C(f, T_{e^+})$ and $(m, n_1)$ and $(m, n_2)$ are in $C(f, T_{e^-})$. But then the fact that $f$ is friendly implies that $(l_1, m, n_2)$ and $(l_2, m, n_1)$ are in $C(f, T)$. □

**Definition 18.** Denote by $Ind(T_{e^+}, T_{e^-})$ the union of all the quadratic binomials from Lemma 17.

The main result of this paper is the following.

**Theorem 19.** *Let $T$ be a tree that has an interior edge $e$ and $f$ a friendly labelling. Suppose that $G_1$ is a binomial generating set for $I_{T_{e^+},f}$ and $G_2$ is a binomial generating set for $I_{T_{e^-}}$. Then*

$$Ext(G_1 \rightarrow T) \cup Ext(T \leftarrow G_2) \cup Ind(T_{e^+}, T_{e^-})$$

*is a generating set for $I_{T,f}$.*

We will prove this result by first proving the following strengthened statement.

**Theorem 20.** *Let $T$ be a tree that has an interior edge $e$ and $f$ a friendly labelling. Suppose that $G_1$ is a binomial Gröbner basis for $I_{T_{e^+},f}$ and $G_2$ is a binomial Gröbner basis for $I_{T_{e^-}}$. Then there exists a term order such that*

$$Ext(G_1 \rightarrow T) \cup Ext(T \leftarrow G_2) \cup Ind(T_{e^+}, T_{e^-})$$

*is a Gröbner basis for $I_{T,f}$.*

# 6 Evolutionary Models for DNA Sequences

In this section, we discuss consequences of the main result from Section 5 for models for DNA sequences. In particular, we will give explicit descriptions of the quadrics, cubics, and quartics needed to generate the phylogenetic ideals for the Jukes-Cantor and Kimura models on binary trees.

### 6.1 Jukes-Cantor 2-base model

Let $T$ be a binary tree with $m$ leaves. The Jukes-Cantor 2-base model has transition matrices which look like, in homogeneous coordinates,

$$\begin{pmatrix} b_v & a_v \\ a_v & b_v \end{pmatrix}.$$

We label the states by elements of the group $\mathbb{Z}_2$ and observe that this a a group based model. After applying the Fourier transform, we arrive at the monomial parameterization

$$\widehat{p}_{g_1,\ldots,g_m} \mapsto \prod_{v \in \mathcal{V}(T)} a_{g_v}^{(v)}$$

where $g_v = \sum g_l$ where $l$ ranges over the leaves of $T$ below $v$. Adding an extra edge and passing to the labelled coordinates, we observe that each labelled coordinate $\widehat{p}_l$ gives a labelling $l$ as a set of nonintersecting paths through $T'$. We will now go through the description of invariants from sections 4 and 5 to describe all the invariants for this model for any binary tree.

Since the labelling function for this model is injective (the number of parameters in each transition matrix is equal to the order of the group) there are no linear invariants. Furthermore, all the sets $Ext(T_v \to T)$ are empty since one easily checks that ideal $I_{T_v,L} =< 0 >$ for any $T_v$ where $T_v$ is the three leaf tree at a vertex. Hence, the ideal $I_{T,L}$ for the Jukes-Cantor 2-base model is generated by the quadratic polynomials coming from the splits, which we will associate with rank conditions of matrices.

INSERT DESCRIPTION OF QUADRICS

**Theorem 21.**

### 6.2 Jukes-Cantor 4-base model

### 6.3 Kimura 2-parameter model

### 6.4 Kimura 3-parameter model

## 7 Conclusion

## References


[1] E. Allman and J. Rhodes.

[2] E. Allman and J. Rhodes.

[3] J. A. Cavender and J. Felsenstein. Invariants of phylogenies: a simple case with discrete states. *J. Classif.* 4:57-71 (1987)

[4] S. Evans and T. Speed.

[5] S. Evans and Zhou.

[6] J. Felsenstein. *Inferring Phylogenies*

[7] Hagedorn.

[8] J. A. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.* 4:167-191 (1987)

[9] L. Pachter and B. Sturmfels.

[10] Semple and Steel. *Phylogenetics*

[11] Steel and Fu

[12] B. Sturmfels. *Gröbner Bases and Convex Polytopes*, American Mathematical Society, Providence, 1995.

[13] Székely, Steel and Erdös

[14] Székely, Steel, Erdös and W.