CONSTRUCTING AND COUNTING PHYLOGENETIC INVARIANTS

STEVEN N. EVANS AND XIAOWEN ZHOU

ABSTRACT. The method of invariants is an approach to the problem of reconstructing the phylogenetic tree of a collection of m taxa using nucleotide sequence data. Models for the respective probabilities of the 4^m possible vectors of bases at a given site will have unknown parameters that describe the random mechanism by which substitution occurs along the branches of a putative phylogenetic tree. An invariant is a polynomial in these probabilities that, for a given phylogeny, is zero for all choices of the substitution mechanism parameters. If the invariant is typically non–zero for another phylogenetic tree, then estimates of the invariant can be used as evidence to support one phylogeny over another.

Previous work of Evans and Speed showed that, for certain commonly used substitution models, the problem of finding a minimal generating set for the ideal of invariants can be reduced to the linear algebra problem of finding a basis for a certain lattice (that is, a free $\mathbb{Z}$-module). They also conjectured that the cardinality of such a generating set can be computed using a simple "degrees of freedom" formula. We verify this conjecture. Along the way, we explain in detail how the observations of Evans and Speed lead to a simple, computationally feasible algorithm for constructing a minimal generating set.

1. INTRODUCTION

The *method of invariants* is a probability–based technique for inferring phylogenetic relations among a group of taxa using nucleotide sequence data. The essential idea behind the method is the following. Suppose that we have aligned DNA sequence data for a m taxa. For a given position in the sequence we have a stochastic model for the base each taxon exhibits at that position. That is, we have a model giving the 4^m joint probabilities

$$p_{B_1 \dots B_m} := \mathbb{P}\{Y_1 = B_1, \dots, Y_m = B_m\},$$

where Y_i is the base observed for the i^{th} taxon and B_i is one the four possible bases A, G, C, T. The model typically involves a putative phylogenetic tree and other unknown parameters that describe the random mechanism by which substitution of bases has occurred through time along the branches of the tree. An *invariant* is a polynomial function in the 4^m variables $p_{B_1 \dots B_m}$, $(B_1, \dots, B_m) \in \{A, G, C, T\}^m$. For a particular phylogeny, it is zero for all choices of the substitution mechanism

Date: May 18, 1998.

1991 *Mathematics Subject Classification.* Primary: 62P10, 13P10. Secondary: 68Q40, 20K01, 60B15.

Key words and phrases. invariant, phylogeny, tree, discrete Fourier analysis, elimination ideal, lattice.

Research supported in part by NSF grant DMS-9703845.

parameters. If the invariant is typically non-zero for other phylogenies, then estimates of the value of the invariant can be used as evidence for or against the putative phylogeny.

Invariants were first introduced by Cavender and Felsenstein 1987 and Lake 1987. Substantial work has been done on the construction of linear invariants (see, for example, Fu 1995, Fu and Li 1992, Hendy and Penny 1996, Nguyen and Speed 1992 and Steel and Fu 1995). As to results on non-linear polynomial invariants, Székely et al. 1993 extend the Fourier analytic approach of Evans and Speed to groups other than the group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ that arises with 4 bases. Ferretti and Sankoff 1993, 1995 and Ferretti et al. 1994 present an "empirical" approach to finding invariants by enlightened trial–and–error. Steel et al. 1993 apply spectral analysis techniques to tree reconstruction and construct all the invariants for the Kimura's 3ST model. Counting formulae for invariants are obtained for certain models in Felsenstein 1991, Steel et al. 1993 and Steel and Fu 1995.

In algebraic parlance, the collection of invariants form an *ideal*: the sum of two invariants is an invariant, and the product of an invariant and any polynomial is also an invariant. More specifically, the ideal of invariants is nothing other than the *elimination ideal* for the set of model probabilities $\{p_{B_1 \dots B_m}\}$ viewed as a set of functions of the parameters describing the substitution mechanism; that is, the ideal of invariants is the totality of algebraic relations between these functions. When the model can be parametrised so that the model probabilities $p_{B_1 \dots B_m}$ are polynomials in the substitution mechanism parameters, then there are standard algorithms using Gröbner bases that, in principle, produce a basis (that is, a minimal generating set) for this ideal (see, for example, Chapter 3 of Cox et al. 1992). In practice, however, such procedures appear to be computationally infeasible for a "generic" elimination ideal problem involving the number of polynomials and variables encountered with just 4 taxa. In order to proceed, it is therefore necessary to uncover structure that is specific to this particular instance of the elimination ideal problem.

Evans and Speed 1993 used some discrete Fourier analysis to develop a procedure for building a basis of the ideal of invariants when the substitution mechanism is given by the *Kimura three–parameter model* and two special cases of it, the *Kimura two–parameter model* and *Jukes–Cantor model* (see Section 2 below for definitions). They showed that the problem could be reduced to one of finding a basis for a certain lattice (that is, a free $\mathbb{Z}$-module). Unfortunately, they did not make it sufficiently clear that the latter problem is just one of linear algebra that can be efficiently solved using Gaussian elimination. A subsidiary aim of this paper is to give explicit algorithms for constructing invariants. These algorithms have been implemented in *Mathematica* and can be obtained from the authors upon request.

Evans and Speed also noted that, in the particular examples they computed, there is a basis of the ideal of invariants with cardinality the same as the number of "degrees of freedom" in the model obtained by an informal parameter counting argument. The main aim of this paper is to establish that this observation is true in complete generality.

The plan of the rest of the paper is as follows. We first give a brief review of the models and related terminology in Section 2, then introduce the algorithms of constructing all independent invariants in Kimura and Jukes-Cantor models with both arbitrary and uniform distributions in Section 3. Proofs of Evans and Speed's conjectures are included in the subsequent Sections.

2. Models

In this section we describe the models which are amenable to the Fourier approach of Evans and Speed and for which we can obtain the number of algebraically independent invariants.

Let $\mathbf{T}$ be a finite rooted tree. Write ρ for the root of $\mathbf{T}$, $\mathbf{V}$ for the set of vertices of $\mathbf{T}$, and $\mathbf{L} \subset \mathbf{V}$ for the set of leaves. We regard $\mathbf{T}$ as a directed graph with edge directions leading away from the root. The elements of $\mathbf{L}$ correspond to the taxa, the tree $\mathbf{T}$ is the phylogenetic tree for the taxa, and the elements of $\mathbf{V}\backslash\mathbf{L}$ can be thought of as unobserved ancestors of the taxa. Enumerate $\mathbf{L}$ as $(l_1, \ldots, l_m)$ and $\mathbf{V}$ as $(v_1, \ldots, v_n)$, with the convention that $l_j = v_j$ for $j = 1, \ldots, m$ and $\rho = v_n$.

Each vertex $v \in \mathbf{V}$ other than the root ρ has a a *father* $\sigma(v)$ (that is, there is a unique $\sigma(v) \in \mathbf{V}$ such that the directed edge $(\sigma(v), v)$ is in the rooted tree $\mathbf{T}$). If v_α and v_ω are two vertices such that there exist vertices $v_\beta, v_\gamma \ldots, v_\xi$ with $\sigma(v_\beta) = v_\alpha$, $\sigma(v_\gamma) = v_\beta$, $\ldots$, $\sigma(v_\omega) = v_\xi$ (that is, there is a directed path in $\mathbf{T}$ from α to ω), then we say that v_ω is a descendent of v_α or that v_α is an ancestor of v_ω and we write $v_\alpha \leq v_\omega$ or $v_\omega \geq v_\alpha$. Note that a vertex is its own ancestor and its own descendent. The *outdegree* $\mathrm{outdeg}(u)$ of $u \in \mathbf{V}$ is the number of *children* of u, that is, the number of $v \in \mathbf{V}$ such that $u = \sigma(v)$. To avoid degeneracies we will always suppose that $\mathrm{outdeg}(v) \geq 2$ for all $v \in \mathbf{V}\backslash\mathbf{L}$.

As far as we are aware, all the probability models proposed in the literature for the bases exhibited by the taxa have the following general form. Let π be a probability distribution on $\{A, G, C, T\}$. We will refer to π as the *root distribution*, and the probability $\pi(B)$ is the probability that the common ancestor species at the root exhibits base B. For each vertex $v \in \mathbf{V}\backslash\{\rho\}$, let $P^{(v)}$ be a stochastic matrix on $\{A, G, C, T\}$. We will refer to $P^{(v)}$ as the *substitution matrix* associated with the edge $(\sigma(v), v)$. The entry $P^{(v)}(B, B')$ is the conditional probability that the species at vertex v exhibits base B' given that the species at vertex $\sigma(v)$ exhibits base B.

Define a probability distribution μ on $\{A, G, C, T\}^{\mathbf{V}}$ by setting

$$\mu((B_v)_{v \in \mathbf{V}}) := \pi(B_\rho) \prod_{v \in \mathbf{V}\backslash\{\rho\}} P^{(v)}(B_{\sigma(v)}, B_v).$$

The distribution μ is the joint distribution of the bases exhibited by all of the species in the tree, both the taxa and the unobserved ancestors. The induced marginal distribution on $\{A, G, C, T\}^{\mathbf{L}}$ is

$$p_{(B_l)_{l \in \mathbf{L}}} := \sum_{v \in \mathbf{V}\backslash\mathbf{L}} \sum_{B_v} \mu(((B_v)_{v \in \mathbf{V}\backslash\mathbf{L}}, (B_l)_{l \in \mathbf{L}})),$$

where each of the dummy variables B_v, $v \in \mathbf{V}\backslash\mathbf{L}$, is summed over the set $\{A, G, C, T\}$. The distribution p is the joint distribution of the bases exhibited by the taxa. Notice that μ is the joint distribution of a $\{A, G, C, T\}^{\mathbf{V}}$–valued, tree–indexed Markov random field with transition probability $P^{(v)}(i_{\sigma(v)}, i_v)$ at each $v \in \mathbf{V}$. The Markov property may be stated as follows: for any two vertices v' and v'', the base at v' and the base at v'' are conditionally μ–independent given the base at any vertex v on the unique (undirected) path connecting v' and v''.

For the tree shown in Figure 1, $\mathbf{V} = \{1,2,3,4,5\}$, $\rho = 5$, $\mathbf{L} = \{1,2,3\}$, and

$$p_{B_1B_2B_3} = \sum_{B_4,B_5 \in \{A,G,C,T\}} \pi(B_5)P^{(1)}(B_5,B_1)P^{(4)}(B_5,B_4)P^{(2)}(B_4,B_2)P^{(3)}(B_4,B_3).$$

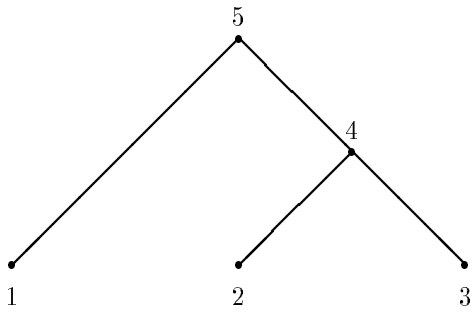

Figure 1

The models of this form which appear in the literature usually take each substitution matrix to be the transition matrix at some point in time of a continuous time Markov chain on the state space $\{A,G,C,T\}$ (which particular point in time is possibly different for each edge, and these variables constitute unknown parameters in the model). We will be particularly interested in a sub-family of Markov chains described in terms of the infinitesimal generator matrix of the chain. Kimura 1981 presents such a model in which the infinitesimal generator matrix is of the form

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix} \end{array},$$

where $\alpha, \beta, \gamma \geq 0$. The value of the triple (α, β, γ) is possibly different for each edge, and these variables also constitute unknown parameters in the model. We will refer to this model as the *Kimura three–parameter model*. If we further restrict the class of allowable infinitesimal generator matrices by imposing the extra condition that $\beta = \gamma$ then we obtain the model considered by Kimura 1980. We will refer to this model as the *Kimura two–parameter model*. Finally, if we require that $\alpha = \beta = \gamma$ we obtain the model considered in Jukes and Cantor 1969 and more explicitly in Neyman 1971 , which we will refer to as the *Jukes-Cantor model*.

One key observation in Evans and Speed 1993 is that there is a group structure inherent in these models. More precisely, the set of bases $\{A,G,C,T\}$ can be identified as an Abelian group, $\mathbb{G}$, with the group operation defined by the following addition table:

$$\begin{array}{c} + \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} A & G & C & T \\ G & A & T & C \\ C & T & A & G \\ T & C & G & A \end{pmatrix} \end{array}.$$

This group is isomorphic to the *Klein 4-group* $\mathbb{Z}_2 \bigoplus \mathbb{Z}_2$ (that is, the group consisting of the elements $\{(0,0),(0,1),(1,0),(1,1)\}$ with the group operation being coordinate wise addition modulo 2). One possible isomorphism is given by $A \leftrightarrow (0,0)$, $G \leftrightarrow (0,1)$, $C \leftrightarrow (1,0)$ and $T \leftrightarrow (1,1)$. Then it is straightforward to check that the infinitesimal generator matrices is nothing other than the infinitesimal generator matrix for a random walk on the group.

In particular, the resulting substitution matrices are of the form $P^{(v)}(B,B') = \pi^{(v)}(B'-B)$ for some probability vector $\pi^{(v)}$ on $\mathbb{G}$. Consequently, if $(Z_v)_{v\in\mathbf{V}}$ is a vector of independent $\mathbb{G}$-valued random variables, with Z_ρ having distribution π, and Z_v, $v \in \mathbf{V}\backslash\{\rho\}$, having distribution $\pi^{(v)}$, then p is the joint distribution of $(Y_l)_{l\in\mathbf{L}}$, where

$$Y_l := \sum_{v \le l} Z_v.$$

The tool used in Evans and Speed 1993 to exploit this last remark is Fourier analysis on $\mathbb{G}$. Let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ denote the unit circle in the complex plane, and regard $\mathbb{T}$ as an Abelian group with the group operation being ordinary complex multiplication. The *characters* of $\mathbb{G}$ are the group homomorphisms mapping $\mathbb{G}$ into $\mathbb{T}$. That is, $\chi : \mathbb{G} \to \mathbb{T}$ is a character if $\chi(g_1+g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in \mathbb{G}$. The characters form an Abelian group under the operation of pointwise multiplication of functions. This group is called the *dual group* of $\mathbb{G}$ and is denoted by $\hat{\mathbb{G}}$. The groups $\mathbb{G}$ and $\hat{\mathbb{G}}$ are isomorphic. Given $g \in \mathbb{G}$ and $\chi \in \hat{\mathbb{G}}$, write $\langle g, \chi\rangle$ for $\chi(g)$. One may write $\hat{\mathbb{G}} = \{1, \phi, \psi, \phi\psi\}$, where the following table gives the values of $\langle g, \chi\rangle$ for $g \in \mathbb{G}$ and $\chi \in \hat{\mathbb{G}}$:

$$\begin{array}{c} \\ 1 \\ \phi \\ \psi \\ \phi\psi \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{array}\right) \end{array}.$$

3. Algorithms

In this section we use the observations in Evans and Speed 1993 to give explicit algorithms for constructing a basis of the ideal of invariants for the models introduced in Section 2. We note that for any choice of substitution mechanisms and any tree we always have the trivial invariant

$$\sum_{(B_l)_{l\in\mathbf{L}}} p_{(B_l)_{l\in\mathbf{L}}} - 1 = 0.$$

We call this invariant the *sum constraint*.

3.1. Three–parameter Kimura model, arbitrary root distribution.

We begin with an explicit algorithm for constructing a basis for the ideal of invariants for the three–parameter Kimura model with arbitrary root distribution. This algorithm and algorithms given later in this section for other models are justified by the results in Evans and Speed 1993 .

We first need some notation. We call a vector $(\chi_{l_1}, \ldots, \chi_{l_m}) \in \hat{\mathbb{G}}^m$ an *allocation of characters to leaves*. Such an allocation of characters to leaves induces an *allocation of characters to vertices* $(\chi_{v_1}, \ldots, \chi_{v_n}) \in \hat{\mathbb{G}}^n$ as follows. The character χ_{v_i}

is the product of the χ_{l_j} for all leaves ℓ_j that are descendents of v_i, that is,

$$\chi_{v_i} := \prod_{l_j \geq v_i} \chi_{l_j}.$$

In particular, if v_i is a leaf (and hence the leaf l_i by our numbering convention), then $\chi_{v_i} = \chi_{l_i}$.

For example, in the 3 taxa case of Figure 1, write $(l_1, l_2, l_3) = (1, 2, 3)$ and $(v_1, v_2, v_3, v_4, v_5) = (1, 2, 3, 4, 5)$ the allocation of characters to leaves $(\phi, \psi, \phi\psi)$ induces an allocation of characters to vertices $(\phi, \psi, \phi\psi, \phi, 1)$.

Let

$$\{(\chi_{i,1}, \ldots, \chi_{i,n}),\ i = 1, \ldots, 4^m - 1\}$$

be an enumeration of the various allocations of characters to vertices induced by the $4^m - 1$ different allocations of characters to leaves other than the allocation $(1, \ldots, 1)$. Define $3n$ vectors $\{\mathbf{x}_{v,\theta} = (x^{(1)}_{v,\theta}, \ldots, x^{(4^m-1)}_{v,\theta}),\ v \in \mathbf{V},\ \theta \in \{\phi, \psi, \phi\psi\}\}$ of dimension $4^m - 1$ by setting

$$x^{(i)}_{v_j,\theta} := \begin{cases} 1, \text{ if } \chi_{i,j} = \theta, \\ 0, \text{ otherwise,} \end{cases} \tag{3.1}$$

for $i = 1, \ldots, 4^m - 1$, $j = 1, \ldots, n$ and $\theta \in \{\phi, \psi, \phi\psi\}$.

Let $\{(a_{1,r}, ..., a_{4^m-1,r}),\ r = 1, ..., k\}$ be a basis of the null space of the real vector space generated by $\{\mathbf{x}_{v,\theta},\ v \in \mathbf{V},\ \theta \in \{\phi, \psi, \phi\psi\}\}$. Such a basis is readily constructed using Gaussian elimination (see, for example, Section 2.1 of Cox et al. 1992). It is apparent from the Gaussian elimination algorithm that for $r = 1, \ldots, 4^m - 1$, each $a_{i,r}$ can be taken to be an integer and the greatest common divisor of $|a_{i,r}|$, $i = 1, \ldots, 4^m - 1$, can be taken to be 1. Under these conditions the collection $\{(a_{1,r}, ..., a_{4^m-1,r}),\ r = 1, ..., k\}$ is also a basis for the lattice (that is, the free $\mathbb{Z}$-module)

$$\{\alpha \in \mathbb{Z}^{4^m-1} : \sum_{i=1}^{4^m-1} \alpha_i x^{(i)}_{v,\theta} = 0,\ v \in \mathbf{V},\ \theta \in \{\phi, \psi, \phi\psi\}\}.$$

The collection of polynomials consisting of the sum constraint and the k polynomials

$$\begin{aligned} &\prod_{\{i:a_{i,r}>0\}} \left(\mathbb{E}\left[\prod_{j=1}^{m} \langle Y_j, \chi_{i,j} \rangle \right] \right)^{a_{i,r}} - \prod_{\{i:a_{i,r}<0\}} \left(\mathbb{E}\left[\prod_{j=1}^{m} \langle Y_j, \chi_{i,j} \rangle \right] \right)^{-a_{i,r}} \\ &= \prod_{\{i:a_{i,r}>0\}} \left(\sum_{(B_1,\ldots,B_m)\in\mathbb{G}^m} \prod_{j=1}^{m} \langle B_j, \chi_{i,j} \rangle p_{B_1 \ldots B_m} \right)^{a_{i,r}} \\ &\quad - \prod_{\{i:a_{i,r}<0\}} \left(\sum_{(B_1,\ldots,B_m)\in\mathbb{G}^m} \prod_{j=1}^{m} \langle B_j, \chi_{i,j} \rangle p_{B_1 \ldots B_m} \right)^{-a_{i,r}}, \quad r = 1, \ldots, k, \end{aligned}$$

is a basis for the ideal of invariants.

In the 3 taxa case of Figure 1, there are 15 vectors $\mathbf{x}_{v,\theta}$ of dimension 63. The null space of the real vector space generated by $\{\mathbf{x}_{v,\theta}\}$ has dimension 48. An example of an element of the null space is a 63 dimensional vector with one entry 1, two

entries -1 and other entries 0, where the entry 1 corresponds to the allocation of characters (ϕ,ψ,ψ), and the two entries -1 correspond to the allocations $(\phi,1,1)$ and $(1,\psi,\psi)$. The corresponding invariant is the quadratic polynomial

$$\begin{aligned}
&\mathbb{E}[\langle Y_1,\phi\rangle\langle Y_2,\psi\rangle\langle Y_3,\psi\rangle] - \mathbb{E}[\langle Y_1,\phi\rangle]\,\mathbb{E}[\langle Y_2,\psi\rangle\langle Y_3,\psi\rangle] \\
&= \sum_{(B_1,B_2,B_3)\in\mathbb{G}^3} \langle B_1,\phi\rangle\langle B_2,\psi\rangle\langle B_3,\psi\rangle p_{B_1B_2B_3} \\
&\quad - \left\{\sum_{(B_1,B_2,B_3)\in\mathbb{G}^3} \langle B_1,\phi\rangle p_{B_1B_2B_3}\right\}\left\{\sum_{(B_1,B_2,B_3)\in\mathbb{G}^3} \langle B_2,\psi\rangle\langle B_3,\psi\rangle p_{B_1B_2B_3}\right\}.
\end{aligned}$$

3.2. **Three–parameter Kimura model, uniform root distribution.** We now consider the three–parameter Kimura model with a uniform distribution at the root. Now there will be two classes of invariants, linear invariants that arise because of the uniform distribution at the root and non-linear ones similar to those that arise in the arbitrary root distribution case and reflect the dependence structure of the model (see Theorem 6.1 of Evans and Speed 1993).

Given an allocation of characters to leaves $(\chi_{l_1},\ldots,\chi_{l_m})$, the character allocated to the root in the induced allocation of characters to vertices is $\prod_{j=1}^m \chi_{l_j}$. In order that $\prod_{j=1}^m \chi_{l_j} = 1$, the characters χ_{l_j}, $j = 1,\ldots,m-1$, can be chosen arbitrarily and then there is a corresponding unique choice of χ_{l_m}. There are thus 4^{m-1} different allocations of characters to leaves such that $\prod_{j=1}^m \chi_{l_j} = 1$.

If $\prod_{j=1}^m \chi_{l_j} \neq 1$, then a simple calculation shows that $\mathbb{E}[\langle Y_\rho, \prod_{j=1}^m \chi_{l_j}\rangle] = 0$, which corresponds to a linear invariant

$$\sum_{(B_1,\ldots,B_m)\in\mathbb{G}^m} \prod_{j=1}^m \langle B_j,\chi_{l_j}\rangle p_{B_1\ldots B_m}.$$

The invariants corresponding to different such allocations are algebraically independent. There are a total of $4^m - 4^{m-1}$ such invariants. In the abovementioned three taxa example there are $4^3 - 4^2 = 48$ such invariants. A typical one is the one derived from $\mathbb{E}[\langle Y_1,\phi\rangle\langle Y_2,\psi\rangle\langle Y_3,\phi\psi\rangle]$.

Now consider the allocations of characters to leaves $(\chi_{l_1},\ldots,\chi_{l_m})$ such that $\prod_{j=1}^m \chi_{l_j} = 1$. There are total $4^{m-1}-1$ different such allocations other than $(1,\ldots,1)$. Reusing notation from Subsection 3.1, let

$$\{(\chi_{i,1},\ldots,\chi_{i,n}),\, i = 1,\ldots,4^{m-1}-1\}$$

be an enumeration of the induced allocations of characters to vertices. Define $3(n-1)$ vectors $\{\mathbf{x}_{v,\theta},\, v\in\mathbf{V}\backslash\{\rho\},\, \theta\in\{\phi,\psi,\phi\psi\}\}$ of dimension $4^{m-1}-1$ as in Subsection 3.1 (notice that $\mathbf{x}_{\rho,\theta} = 0$ for $\theta\in\{\phi,\psi,\phi\psi\}\}$). Following exactly the same algorithm described in Subsection 3.1 we can recover another collection of algebraically independent invariants. Theorem 6.1 in Evans and Speed 1993 gives that the union of these two collections and the sum constraint constitutes a basis for the ideal of invariants.

3.3. **Two–parameter Kimura model.** In a Kimura two–parameter model with arbitrary root distribution, the algorithm is similar to that in Subsection 3.1. Let $\{(\chi_{i,1},\ldots,\chi_{i,n}),\, i = 1,\ldots,4^m-1\}$ be an enumeration of the various allocations of characters to vertices induced by the 4^m-1 different allocations of

characters to leaves other than the allocation $(1,\dots,1)$. Define $2n+1$ vectors $\{\mathbf{x}_{\rho,\theta} = (x_{\rho,\theta}^{(1)},\dots,x_{\rho,\theta}^{(4^m-1)}),\ \theta \in \{\phi,\psi,\phi\psi\}\} \cup \{\mathbf{x}_{v,\theta} = (x_{v,\theta}^{(1)},\dots,x_{v,\theta}^{(4^m-1)}),\ v \in \mathbf{V}\backslash\{\rho\},\ \theta \in \{\phi,\psi\}\}$ of dimension 4^m-1 by setting

$$x_{\rho,\theta}^{(i)} = x_{v_n,\theta}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,n} = \theta, \\ 0, \text{ otherwise,} \end{cases} \tag{3.2}$$

for $i = 1,\dots,4^m-1$, $\theta \in \{\phi,\psi,\phi\psi\}$,

$$x_{v_j,\phi}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,j} \in \{\phi,\phi\psi\}, \\ 0, \text{ otherwise,} \end{cases} \tag{3.3}$$

and

$$x_{v_j,\psi}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,j} = \psi, \\ 0, \text{ otherwise,} \end{cases} \tag{3.4}$$

for $i = 1,\dots,4^m-1$, $j = 1,\dots,n-1$.

A suitable basis for the null space of the real vector space generated by $\{\mathbf{x}_{\rho,\theta},\ \theta \in \{\phi,\psi,\phi\psi\}\} \cup \{\mathbf{x}_{v,\theta},\ v \in \mathbf{V}\backslash\{\rho\},\ \theta \in \{\phi,\psi\}\}$ gives rise to a basis for the ideal of invariants in exactly the same way as in the algorithm of Subsection 3.1.

In a two–parameter Kimura model with uniform distribution, the algorithm is similar to that for the three–parameter model in Subsection 3.2. We first distinguish two cases, $\prod_{j=1}^m \chi_{l_j} \neq 1$ or $\prod_{j=1}^m \chi_{l_j} = 1$. In the later case, there are $2(n-1)$ vectors of dimension $4^{m-1}-1$ defined in the same way as (3.3) and (3.4).

3.4. **Jukes-Cantor model.** The algorithms for Jukes-Cantor model are similar to those in the abovementioned two models. Here, for example, the equivalent of (3.1) is

$$x_{\rho,\theta}^{(i)} = x_{v_n,\theta}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,n} = \theta, \\ 0, \text{ otherwise,} \end{cases} \tag{3.5}$$

for $i = 1,\dots,4^m-1$, $\theta \in \{\phi,\psi,\phi\psi\}$, and

$$x_{v_j}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,j} \in \{\phi,\psi,\phi\psi\}, \\ 0, \text{ otherwise,} \end{cases} \tag{3.6}$$

for $i = 1,\dots,4^m-1$, $j = 1,\dots,n-1$.

4. Statement of results

When we say that there are N algebraically independent invariants, we mean that the ideal of invariants has a basis with N elements. Recall that our tree $\mathbf{T}$ has n vertices and m leaves. Evans and Speed 1993 observed that for Kimura three–parameter models with arbitrary root distribution, the marginal distribution of leaves can take 4^m different values, the root distribution contributes 3 parameters, and the substitution matrix for each edge contributes 3 parameters; and this suggests that the number of "degrees of freedom" is 4^m-3n, and in all the examples they computed the number of algebraically independent invariants always coincides with the number of degrees of freedom obtained from this informal parameter counting procedure. (Note that in order to identify invariants with elements of an elimination ideal we are taking the sum constraint to be an invariant. This

differs from the convention in Evans and Speed 1993 and so our counting differs by one from their counting.) They conjecture that such a counting formula holds in general. The following theorem verifies this conjecture.

Theorem 4.1. *Consider a Kimura three–parameter model with arbitrary root distribution. There are* $4^m - 3n$ *algebraically independent invariants.*

If the root distribution is uniform, then the root distribution does not contribute any parameters. Moreover, if $\text{outdeg}(\rho) = 2$, then the contribution of the two edges connected to the root would be the same as that of a single edge connecting the two children of the root. For example, the tree in Figure 2 is equivalent to the tree in Figure 3 in the sense that for the three–parameter Kimura substitution mechanism, the class of possible probability vectors $(p_{B_1 \ldots B_m})$ that can be produced by the two trees coincides when the distribution at the root 7 in Figure 2 and the distribution at the root 6 in Figure 3 is taken to be uniform. In other words, if $\text{outdeg}(\rho) = 2$ in the uniform root distribution case, then the number of parameters contributed by the substitution matrices for the two edges connected to the root is just 3. The following counting formulae are therefore expected.

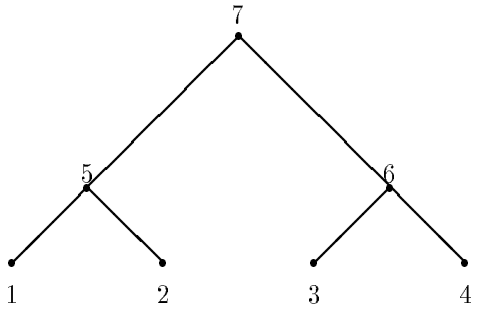

Figure 2

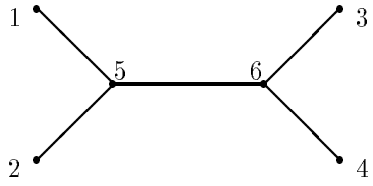

Figure 3

Theorem 4.2. *Consider a Kimura three–parameter model with uniform root distribution. If* $\text{outdeg}(\rho) = 2$*, then there are* $4^m - 3(n-2)$ *algebraically independent invariants. If* $\text{outdeg}(\rho) > 2$*, then there are* $4^m - 3(n-1)$ *algebraically independent invariants.*

Similar counting rules apply for Kimura two–parameter models and Jukes–Cantor models. They are formulated in Theorem 4.3 and Theorem 4.4.

Theorem 4.3. *Consider a Kimura two–parameter model.*

(i) With arbitrary root distribution, there are $4^m - 3 - 2(n-1)$ *algebraically independent invariants.*
(ii) With uniform root distribution and $\mathrm{outdeg}(\rho) = 2$*, there are* $4^m - 2(n-2)$ *algebraically independent invariants.*
(iii) With uniform root distribution and $\mathrm{outdeg}(\rho) > 2$*, there are* $4^m - 2(n-1)$ *algebraically independent invariants.*

Theorem 4.4. *Consider a Jukes–Cantor model.*

(i) With arbitrary root distribution, there are $4^m - 3 - (n-1)$ *algebraically independent invariants.*
(ii) With uniform root distribution and $\mathrm{outdeg}(\rho) = 2$*, there are* $4^m - (n-2)$ *algebraically independent invariants.*
(iii) With uniform root distribution and $\mathrm{outdeg}(\rho) > 2$*, there are* $4^m - (n-1)$ *algebraically independent invariants.*

5. Proof of Theorem 4.1

In this section the $4^m - 1$-dimensional vectors $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V},\, \theta \in \{\phi, \psi, \phi\psi\}\}$ are as defined in Subsection 3.1. In a Kimura three–parameter model with arbitrary distribution, the algorithm in Subsection 3.1 shows that the number of algebraically independent invariants is the 1 plus the dimension of the null space of the real vector space generated by $4^m - 1$-dimensional vectors $\mathbf{x}_{v,\theta}$. The latter dimension is $4^m - 1$ minus the dimension of the vector space generated by the collection $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V},\, \theta \in \{\phi, \psi, \phi\psi\}\}$. Hence, Theorem 4.1 will follow if we can show that this collection is linearly independent.

Lemma 5.1. *The vectors* $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V},\, \theta \in \{\phi, \psi, \phi\psi\}\}$ *are linearly independent.*

Proof. Suppose we have real numbers $b_{j,\theta}$, $j = 1, \ldots, n$, $\theta \in \{\phi, \psi, \phi\psi\}$, satisfying

$$\sum_{\theta \in \{\phi,\psi,\phi\psi\}} \sum_{j=1}^{n} b_{j,\theta} x^{(i)}_{v_j,\theta} = 0, \quad i = 1, \ldots, 4^m - 1. \tag{5.1}$$

We need to establish that all the $b_{j,\theta}$ are zero. The proof will proceed by induction on n, the number of vertices of $\mathbf{T}$.

The case $n = 1$ is straightforward.

Suppose for some integer $N \geq 1$ that the assertion is true for all $n \leq N$ and consider the assertion for $n = N + 1$.

Choose any two leaves with a common father. (Note that such leaves exist by our standing assumption that all vertices other than leaves have outdegree at least 2.) Without loss of generality and recalling our labeling convention, we may assume that the leaves have been numbered in such a way that these leaves are $l_1 = v_1$ and $l_2 = v_2$. Denote the common father by v^*. Let $v^* = v_{j_1} \geq v_{j_2} \geq \ldots \geq v_{j_k} = v_n = \rho$ be, in reverse order, the vertices along the directed path joining ρ to v^*. That is, $\sigma(v_{j_p}) = v_{j_{p+1}}$ for $p = 1, \ldots, k-1$.

Consider $\theta \in \{\phi, \psi, \phi\psi\}$. The instance of equation (5.1) that arises when the index i corresponds to the allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j = 1, \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{1,\theta} + b_{j_1,\theta} + \ldots + b_{j_k,\theta} = 0. \tag{5.2}$$

The instance of equation (5.1) that arises when the index i corresponds to the allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j = 2\ , \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{2,\theta} + b_{j_1,\theta} + \ldots + b_{j_k,\theta} = 0 \tag{5.3}$$

The instance of equation (5.1) that arises when the index i corresponds to the allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j \in \{1, 2\}, \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{1,\theta} + b_{2,\theta} = 0 \tag{5.4}$$

(recall that $\theta^2 = 1$).

Combining equations (5.2), (5.3) and (5.4) gives that $b_{1,\theta} = b_{2,\theta} = 0$.

Suppose that the leaves have been numbered so that the children of v^* are $l_1 = v_1, \ldots, l_s = v_s$. The argument we have just been through establishes that $b_{j,\theta} = 0$ for all $j = 1, \ldots, s$. (It also establishes that $b_{j,\theta} = 0$ for all $j = 1, \ldots, m$, but we will not use this fact.) Therefore, equation (5.1) reduces to

$$\sum_{\theta \in \{\phi,\psi,\phi\psi\}} \sum_{j=s+1}^{n} b_{j,\theta} x^{(i)}_{v_j,\theta} = 0, \quad i = 1, \ldots, 4^m - 1. \tag{5.5}$$

Consider $v \in \mathbf{V} \backslash \{l_1, \ldots, l_s\}$ and the character χ_v allocated to the vertex v when the characters $(\chi_{l_1}, \ldots, \chi_{l_m})$ are allocated to the leaves. If $l_1, \ldots, l_s$ are not descendents of v (that is, if v^* is not a descendent of v – we emphasise that vertices are their own descendents), then χ_v is independent of $\chi_{l_1}, \ldots, \chi_{l_s}$. On the other hand, if $l_1, \ldots, l_s$ are descendents of v (that is, if v^* is a descendent of v), then χ_v only depends on $\chi_{l_1}, \ldots, \chi_{l_s}$ through the value of $\prod_{j=1}^{s} \chi_{l_j} = \chi_{v^*}$. Moreover, as $\chi_{l_1}, \ldots, \chi_{l_s}$ vary, χ_{v^*} ranges over all of $\{1, \phi, \psi, \phi\psi\}$. Consequently, the system of equations (5.5) is just the system of equations (5.1) corresponding to a new tree $\mathbf{T}^*$ obtained by removing the leaves $l_1, \ldots, l_s$ and the edges connecting them to v^*. Since the number of vertices of $\mathbf{T}^*$ is $N + 1 - s < N$ we can apply the inductive assumption to conclude that $b_{j,\theta} = 0$ for all $j = 1, \ldots, n$ and $\theta \in \{\phi, \psi, \phi\psi\}$. (Note that $\mathbf{T}^*$ satisfies our standing assumption that all vertices other than leaves have outdegree at least 2.) □

6. Proof of Theorem 4.2

For this section, the $4^{m-1} - 1$–dimensional vectors $\mathbf{x}_{v,\theta}$ are as defined in Subsection 3.2. In a Kimura three–parameter model with uniform distribution, the algorithm in Subsection 3.2 shows that there are $4^m - 4^{m-1}$ algebraically independent linear invariants corresponding to the $4^m - 4^{m-1}$ allocations of characters to leaves satisfying $\prod_{j=1}^{m} \chi_{l_j} \neq 1$ and there are $4^{m-1} - 1 - \dim \text{span}\{\mathbf{x}_{v,\theta},\ v \in$

$\mathbf{V}\backslash\{\rho\}$, $\theta \in \{\phi, \psi, \phi\psi\}\}$ algebraically independent nonlinear invariants derived from $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V},\, \theta \in \{\phi, \psi, \phi\psi\}\}$. Therefore, the following two results establish Theorem 4.2.

Lemma 6.1. *Suppose that* $\mathrm{outdeg}(\rho) = 2$. *Let* v' *and* v'' *be the two children of* ρ. *Then* $\mathbf{x}_{v',\theta} = \mathbf{x}_{v'',\theta}$ *for* $\theta \in \{\phi, \psi, \phi\psi\}$ *and the vectors* $\mathbf{x}_{v,\theta}$, $v \in \mathbf{V}\backslash\{\rho, v''\}$, $\theta \in \{\phi, \psi, \phi\psi\}$, *are linearly independent.*

Proof. Since $1 = \prod_{j=1}^{m} \chi_{j_l} = \chi_{v'}\chi_{v''}$ it follows that $\chi_{v'} = \chi_{v''}$ and hence $\mathbf{x}_{v',\theta} = \mathbf{x}_{v'',\theta}$ for $\theta \in \{\phi, \psi, \phi\psi\}$.

The proof of the second claim will proceed via an induction similar to that used in the proof of Theorem 4.1. In order to verify the inductive step, assume that $n > 3$. By interchanging the designations of v' and v'', we can suppose that v' is not a leaf. We can also assume that we have numbered the vertices so that $v'' = v_{n-1}$ and $\rho = v_n$. Suppose that we have real numbers $b_{j,\theta}$, $j = 1, \ldots, n-2$, $\theta \in \{\phi, \psi, \phi\psi\}$, satisfying

$$\sum_{\theta \in \{\phi,\psi,\phi\psi\}} \sum_{j=1}^{n-2} b_{j,\theta} x^{(i)}_{v_j,\theta} = 0, \quad i = 1, \ldots, 4^{m-1} - 1. \tag{6.1}$$

Choose any two leaves that are descendents of v' and have a common father. Without loss of generality, we may assume that the leaves have been numbered in such a way that these leaves are $l_1 = v_1$ and $l_2 = v_2$. Denote the common father by v^*. Let $v^* = v_{j_1} \geq v_{j_2} \geq \ldots \geq v_{j_k} = v' \geq \rho$ be, in reverse order, the vertices along the directed path connecting ρ to v^*. Choose a leaf v^{**} that is not a descendent of v' (and hence is a descendent of v''). Let $v^{**} = v_{j'_1} \geq v_{j'_2} \geq \ldots \geq v_{j'_q} = v'' \geq \rho$ be, in reverse order, the vertices along the directed path connecting ρ to v^{**}. Of course, $\{v_1, v_2, v_{j_1}, \ldots, v_{j_k}\} \cap \{v_{j'_1}, \ldots, v_{j'_q}\} = \emptyset$.

Consider $\theta \in \{\phi, \psi, \phi\psi\}$. The instance of equation (6.1) that arises when the index i corresponds to the character allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j \in \{1, j'_1\}, \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{1,\theta} + \sum_{i=1}^{k} b_{j_i,\theta} + \sum_{i=1}^{q} b_{j'_i,\theta} = 0 \tag{6.2}$$

The instance of equation (6.1) that arises when the index i corresponds to the character allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j \in \{2, j'_1\}, \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{2,\theta} + \sum_{i=1}^{k} b_{j_i,\theta} + \sum_{i=1}^{q} b_{j'_i,\theta} = 0 \tag{6.3}$$

The instance of equation (6.1) that arises when the index i corresponds to the character allocation $(\chi_{l_1}, \ldots, \chi_{l_m})$ given by

$$\chi_{l_j} = \begin{cases} \theta, \text{ if } j \in \{1,2\}, \\ 1, \text{ otherwise,} \end{cases}$$

is

$$b_{1,\theta} + b_{2,\theta} = 0 \tag{6.4}$$

Combining equations (6.2), (6.3) and (6.4) gives $b_{1,\theta} = b_{2,\theta} = 0$. The inductive step can now be completed as in Lemma 5.1. □

The next lemma follows from similar considerations.

Lemma 6.2. *If* $\text{outdeg}(\rho) > 2$, *then the vectors* $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V}\backslash\{\rho\},\, \theta \in \{\phi, \psi, \phi\psi\}\}$ *are linearly independent.*

7. Proofs of Theorem 4.3 and Theorem 4.4

The following lemmas are key to the proofs of Theorem 4.3 and Theorem 4.4. We leave the details to the reader.

Lemma 7.1. *Consider a Kimura two–parameter model and use the notation of Subsection 3.3.*

(i) Suppose that the root distribution is arbitrary. Then the vectors

$$\{\mathbf{x}_{\rho,\theta},\, \theta \in \{\phi, \psi, \phi\psi\}\} \cup \{\mathbf{x}_{v,\theta},\, v \in \mathbf{V}\backslash\{\rho\},\, \theta \in \{\phi, \psi\}\}$$

are linearly independent.

(ii) Suppose that the root distribution is uniform and that $\text{outdeg}(\rho) = 2$. *Let* v' *and* v'' *be the two children of* ρ. *Then* $\mathbf{x}_{v',\theta} = \mathbf{x}_{v'',\theta}$ *for* $\theta \in \{\phi, \psi\}$ *and the vectors* $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V}\backslash\{\rho, v''\},\, \theta \in \{\phi, \psi\}\}$, *are linearly independent.*

(iii) Suppose that the root distribution is uniform and that $\text{outdeg}(\rho) > 2$. *Then the vectors* $\{\mathbf{x}_{v,\theta},\, v \in \mathbf{V}\backslash\{\rho\},\, \theta \in \{\phi, \psi\}\}$ *are linearly independent.*

Lemma 7.2. *Consider a Jukes-Cantor model and use the notation of Subsection 3.4.*

(i) Suppose that the root distribution is arbitrary. Then the vectors

$$\{\mathbf{x}_{\rho,\theta},\, \theta \in \{\phi, \psi, \phi\psi\}\} \cup \{\mathbf{x}_v,\, v \in \mathbf{V}\backslash\{\rho\}\}$$

are linearly independent.

(ii) Suppose that the root distribution is uniform and that $\text{outdeg}(\rho) = 2$. *Let* v' *and* v'' *be the two children of* ρ. *Then* $\mathbf{x}_{v'} = \mathbf{x}_{v''}$ *and the vectors* $\{\mathbf{x}_v,\, v \in \mathbf{V}\backslash\{\rho, v''\}\}$ *are linearly independent.*

(iii) Suppose that the root distribution is uniform and that $\text{outdeg}(\rho) > 2$. *Then the vectors* $\{\mathbf{x}_v, v \in \mathbf{V}\backslash\{\rho\}\}$ *are linearly independent.*

Acknowldgement: The authors thank Joe Felsenstein, David Sankoff, Terry Speed and Bernd Sturmfels for helpful conversations.

References

[1] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. Classification*, 4:57–71, 1987.

[2] D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms : an introduction to computational algebraic geometry and commutative algebra*. New York : Springer-Verlag, 1992.

[3] S.N. Evans and T.P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21:355–377, 1993.

[4] Joseph Felsenstein. Counting phylogenetic invariants in some simple cases. *J. Theor. Biol.*, 152:357–376, 1991.

[5] V. Ferretti, B.F. Lang, and D. Sankoff. Skewed base composition, asymmetric transition matrices, and phylogenetic invariants. *J. Comput. Biol.*, 1(1):77–92, 1994.

[6] V. Ferretti and D. Sankoff. The empirical discovery of phylogenetic invariants. *Adv. Appl. Prob.*, 25:290–302, 1993.

[7] V. Ferretti and D. Sankoff. Phylogenetic invariants for more general evolutionary models. *J. Theor. Biol.*, 173:147–162, 1995.

[8] Yun-Xin Fu. Linear invariants under Jukes' and Cantor's one-parameter model. *J. Theor. Biol.*, 173:339–352, 1995.

[9] Yun-Xin Fu and Wen-Hsiung Li. Construction of linear invariants in phylogenetic inference. *Math. Biosci.*, 109:201–228, 1992.

[10] Michael D. Hendy and David Penny. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *J. Comput. Biol.*, 3(1):19–31, 1996.

[11] T.H. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. New York: Academic Press, 1969.

[12] M. Kimura. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[13] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78:454–458, 1981.

[14] J.A. Lake. A rate-independent technique for analysis of nucleic acid sequences:evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191, 1987.

[15] J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S.S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. New York: Academic Press, 1971.

[16] Trang Nguyen and T. P. Speed. A derivation of all linear invariants for a nonbalanced transversion model. *J. Mol. Evol.*, 35:60–76, 1992.

[17] M.A. Steel and Y. X. Fu. Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model. *J. Comput. Biol.*, 2(1):39–47, 1995.

[18] Mike Steel, Laszlo Székely, Peter L. Erdös, and Peter Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3st model. *N.Z. J. Bot.*, 31:289–296, 1993.

[19] L.A. Székely, M.A. Steel, and P.L. Erdös. Fourier calculus on evolutionary trees. *Adv. Appl. Math.*, 14:200–216, 1993.

E-mail address: evans@stat.Berkeley.EDU

E-mail address: xzhou@stat.Berkeley.EDU

Department of Statistics #3860, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A. phone: (510)-642-2777, fax: (510)-642-7892