DIFFERENT TREES HAVE DISTINCT 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 collection of 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 possible 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. We show for a widely used, general class of substitution mechanisms that given two different trees there is always a polynomial that is an invariant for one tree but not an invariant for the other. Thus estimates of invariants can always be used to discriminate between competing phylogenies.

1. INTRODUCTION

The *method of invariants* is a statistical technique for inferring phylogenetic relations among a group of m taxa using aligned DNA sequence data. For a given position in the sequence we have a stochastic 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. This model involves a phylogenetic tree and parameters that describe the random mechanism by which substitution of bases has occurred through time along the branches of the tree. An *invariant* for a particular phylogeny is a polynomial function in the 4^m variables $p_{B_1 \dots B_m}$, $(B_1, \dots, B_m) \in \{A, G, C, T\}^m$, that is zero for all choices of the substitution mechanism parameters. If a polynomial is an invariant for one phylogeny but not for another, then an estimate of the value of the polynomial can be used to discriminate between the two trees.

Invariants were first introduced by Cavender and Felsenstein [CF87] and Lake [Lak87]. We refer the reader to [EZ98] for a recent, fairly extensive bibliography of the area.

Note that the sum of two invariants is an invariant, and the product of an invariant and any polynomial is also an invariant. In algebraic terminology, the collection of invariants is an *ideal* in the ring of polynomials. Evans and Speed [ES93] used discrete Fourier analysis to produce a minimal generating set for the ideal of invariants when the substitution mechanism is given by the *three–parameter Kimura*

Date: January 24, 2001.

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

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

Research supported in part by NSF grants DMS-9703845 and DMS-0071468.

model and two special cases of it, the *two–parameter Kimura 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 free $\mathbb{Z}$-module. The algorithm implicit in their method was explained more fully in [EZ98], where a conjecture of Evans and Speed on the number of algebraically independent invariants for various models was also established.

In order for the method of invariants to be useful for deciding between two possible phylogenies, it must be the case that there is an invariant for one tree which is not an invariant for the other. However, we are not aware of any work in the literature showing in any generality that this will be so. Our main aim in this paper is to establish that this is indeed the case for the models to which the methods of Evans and Speed apply – see Section 3. The observations made in the proof are also used to give an alternative proof of the counting conjectures of [ES93].

We further point out Section 4 that in order to find an invariant that discriminates between two trees, one can restrict attention to invariants that just take into account the bases observed at 3 taxa. In principle, therefore, we only need to do a once–off computation of the invariants for the 4 trees that one can build on 3 taxa (we require that the non-leaf vertices of our trees are "really there" in the sense that they have outdegree at least 2, so that there are 3 trees on 3 leaves with 2 non-leaf vertices and 1 tree on 3 leaves with 1 non-leaf vertex). This observation has obvious computational advantages, because the algorithm in [ES93, EZ98] involves Gaussian elimination on a $4^m \times 3n$ matrix when we are dealing with the three–parameter Kimura model and a tree that has m leaves and n vertices in total. On the other hand, it may be of limited statistical utility, because it seems reasonable that if one is trying to decide between two markedly different phylogenies, then just concentrating on 3 taxa could discard potentially useful information.

2. Models

In this section we describe the models to which the Fourier approach of Evans and Speed applies.

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 $(\ell_1, \dots, \ell_m)$ and $\mathbf{V}$ as $(v_1, \dots, v_n)$, with the convention that $\ell_j = v_j$ for $j = 1, \dots, m$ 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 $v_\beta, v_\gamma \dots, v_\xi$ with $\sigma(v_\beta) = v_\alpha$, $\sigma(v_\gamma) = v_\beta$, $\dots$, $\sigma(v_\omega) = v_\xi$ (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 $v_\alpha \le v_\omega$ or $v_\omega \ge v_\alpha$. Note that a vertex is its own ancestor and its own descendent. The *outdegree* 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 outdeg$(v) \ge 2$ for all $v \in \mathbf{V}\backslash\mathbf{L}$.

Let $\pi^{(\rho)}$ be a probability distribution on $\{A, G, C, T\}$. We will refer to $\pi^{(\rho)}$ as the *root distribution*, and the probability $\pi^{(\rho)}(B)$ 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

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

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

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

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''.

Kimura [Kim81] introduced such a model in which the substitution matrices are of the form

$$\begin{array}{c|cccc} & A & G & C & T \\ \hline A & w & x & y & z \\ G & x & w & z & y \\ C & y & z & w & x \\ T & z & y & x & w \end{array},$$

where $0 \leq w, x, y, z \leq 1$ and $w + x + y + z = 1$. The value of (w, x, y, z) is possibly different for each edge, and these variables also constitute unknown parameters in the model. We will refer to this model as the *three–parameter Kimura model.* If we further restrict the class of allowable substitution matrices by imposing the extra condition that $y = z$ then we obtain the model considered in [Kim80]. We will refer to this model as the *two–parameter Kimura model.* Finally, if we require that $x = y = z$ we obtain the model considered in [JC69] and more explicitly in [Ney71], which we will refer to as the *Jukes-Cantor model.* We are considering the case where the root distribution $\pi^{(\rho)}$ is arbitrary. The root distribution is sometimes fixed to be uniform, and [ES93, EZ98] discuss the modifications that are necessary to handle the construction of invariants in this case. For the sake of brevity and because it does not appear to have as much practical application, we will not discuss it here.

One key observation in [ES93] 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:

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

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 $T \leftrightarrow (1,1)$.

It follows that the 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 $\pi^{(\rho)}$, 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

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

The tool used in [ES93] 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 $\chi \in \hat{\mathbb{G}}$:

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

We now outline the observations of [ES93, EZ98] concerning the ideal of invariants for the three–parameter Kimura model. We refer the reader to these references for more details and the modifications that are required for the two–parameter Kimura model and the Jukes–Cantor model.

We first need some notation. We call a vector $(\chi_{\ell_1}, \ldots, \chi_{\ell_m}) \in \hat{\mathbb{G}}^m$ 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 is the product of the χ_ℓ for all leaves ℓ that are descendents of v, that is,

$$\chi_v := \prod_{\ell \geq v} \chi_\ell.$$

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

Let

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

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

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

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

Write $\mathcal{R}(\mathbf{T})$ for the free $\mathbb{Z}$–module generated by the set $\{\mathbf{x}_{v,\theta} : v \in \mathbf{V}, \theta = \phi, \psi, \phi\psi\}$; that is, $\mathcal{R}(\mathbf{T})$ is the collection of integer vectors of dimension 4^m consisting of $\mathbb{Z}$-linear combinations of the $\mathbf{x}_{v,\theta}$. Set

$$\mathcal{N}(\mathbf{T}) := \{a \in \mathbb{Z}^{4^m} : \sum_{i=1}^{4^m} a_i x_{v,\theta}^{(i)} = 0,\ v \in \mathbf{V},\ \theta = \phi, \psi, \phi\psi\},$$

so that $\mathbb{Z}^{4^m} = \mathcal{R}(\mathbf{T}) \oplus \mathcal{N}(\mathbf{T})$.

For $a \in \mathbb{Z}^{4^m}$, the polynomial

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

is an invariant if and only if $a \in \mathcal{N}(\mathbf{T})$. Moreover, if $\{(a_{1,r}, ..., a_{4^m,r}),\ r = 1, ..., k\}$ is a $\mathbb{Z}$-basis for the free $\mathbb{Z}$-module $\mathcal{N}(\mathbf{T})$ (it is shown in [EZ98] that the $3n$ vectors $\{\mathbf{x}_{v_j,\theta} : j = 1, \ldots, n,\ \theta = \phi, \psi, \phi\psi\}$ are linearly independent and so the rank r is $4^m - 3n$), then the set of polynomials of the form

$$\prod_{\{i:a_{i,r} \geq 0\}} \left(\mathbb{E}\left[\prod_{j=1}^m \langle Y_j, \chi_{i,j}\rangle\right]\right)^{a_{i,r}} - \prod_{\{i:a_{i,r} \leq 0\}} \left(\mathbb{E}\left[\prod_{j=1}^m \langle Y_j, \chi_{i,j}\rangle\right]\right)^{-a_{i,r}}$$

generates the ideal of invariants but no subset thereof does.

3. Invariants always discriminate

We begin with the natural notion of equivalence for trees with labelled leaves. We say that two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves are *identical* if there is a bijection τ from the set of vertices $\mathbf{V}'$ of $\mathbf{T}'$ to the set of vertices $\mathbf{V}''$ of $\mathbf{T}''$ such that $\tau(\ell) = \ell$ for each leaf $\ell \in \mathbf{L}$ and $u \in \mathbf{V}'$ is the father of $v \in \mathbf{V}'$ in $\mathbf{T}'$ if and only if $\tau(u) \in \mathbf{V}''$ is the father of $\tau(v) \in \mathbf{V}''$ in $\mathbf{T}''$. This is equivalent to requiring that $\tau(\ell) = \ell$ for each leaf $\ell \in \mathbf{L}$ and $u \in \mathbf{V}'$ is the ancestor of $v \in \mathbf{V}'$ in $\mathbf{T}'$ if and only if $\tau(u) \in \mathbf{V}''$ is the ancestor of $\tau(v) \in \mathbf{V}''$ in $\mathbf{T}''$. It is not hard to see that two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves are identical if and only if for each $v' \in \mathbf{V}'$ the set of leaves descended from v' is equal to the set of leaves descended from some $v'' \in \mathbf{V}''$ and vice-versa.

Given two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves, write $\nu(\mathbf{T}', \mathbf{T}'')$ for the number of vertices v'' of $\mathbf{T}''$ such that the collection of leaves descended from v'' (that is, $\{\ell \in \mathbf{L} : \ell \geq v''\}$) is not the collection of leaves descended from any vertex of $\mathbf{T}'$. If $\mathbf{T}'$ and $\mathbf{T}''$ are not identical, then either $\nu(\mathbf{T}', \mathbf{T}'') > 0$ or $\nu(\mathbf{T}'', \mathbf{T}') > 0$. The following result thus gives that if $\mathbf{T}'$ and $\mathbf{T}''$ are not identical, then (under the three–parameter Kimura model) there is an invariant for one tree that is not an invariant for the other tree. In fact, there are $3\nu(\mathbf{T}', \mathbf{T}'')$ algebraically independent invariants for the tree $\mathbf{T}'$ that are not invariants for the tree $\mathbf{T}''$, and similarly with the roles of $\mathbf{T}'$ and $\mathbf{T}''$ interchanged. Analogous results hold for the two–parameter Kimura model and the Jukes–Cantor model, with 3 being replaced by 2 and 1, respectively. We omit the proofs of these latter two results, since they follow the same pattern as the one given below.

Theorem 3.1. *Suppose that $\mathbf{T}'$ and $\mathbf{T}''$ are two trees with the same set $\mathbf{L}$ of leaves. The rank of the free $\mathbb{Z}$–module $\mathcal{N}(\mathbf{T}') \cap \mathcal{R}(\mathbf{T}'')$ is $3\nu(\mathbf{T}', \mathbf{T}'')$.*

Proof. Note that

$$\begin{aligned}\operatorname{rank}(\mathcal{N}(\mathbf{T}') \cap \mathcal{R}(\mathbf{T}'')) &= \operatorname{rank}(\mathcal{R}(\mathbf{T}'')) - \operatorname{rank}(\mathcal{R}(\mathbf{T}') \cap \mathcal{R}(\mathbf{T}'')) \\ &= \operatorname{rank}(\mathcal{R}(\mathbf{T}') + \mathcal{R}(\mathbf{T}'')) - \operatorname{rank}(\mathcal{R}(\mathbf{T}')).\end{aligned}$$

Write $\mathbf{V}'$ and $\mathbf{V}''$ for the vertices of $\mathbf{T}'$ and $\mathbf{T}''$, respectively, and let $\tilde{\mathbf{V}}''$ denote the set of vertices v'' of $\mathbf{T}''$ such that the collection of leaves descended from v'' (that is, $\{\ell \in \mathbf{L} : \ell \geq v''\}$) is not the collection of leaves descended from any vertex of $\mathbf{T}'$. Hence $|\tilde{V}''| = \nu(\mathbf{T}', \mathbf{T}'')$. Of course, if $v'' \in \mathbf{V}'' \backslash \tilde{\mathbf{V}}''$, then there is a vertex $v' \in \mathbf{V}'$ such that the assignment of characters to v' and v'' for each assignment of characters to leaves are the same, and hence the vector $\mathbf{x}_{v',\theta}$ (calculated for $\mathbf{T}'$) is the same as the vector $\mathbf{x}_{v'',\theta}$ (calculated for $\mathbf{T}''$). The result will thus follow if we can show that the vectors

$$\{\mathbf{x}_{v',\theta} : v' \in \mathbf{V}',\, \theta = \phi, \psi, \phi\psi\} \cup \{\mathbf{x}_{v'',\theta} : v'' \in \tilde{\mathbf{V}}'',\, \theta = \phi, \psi, \phi\psi\}$$

are linearly independent over the integers (equivalently, over the reals).

Let $\mathbf{X}$ denote the $4^m \times 3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)$ matrix with columns indexed by $(\mathbf{V}' \cup \tilde{\mathbf{V}}'') \times \{\phi, \psi, \phi\psi\}$ that has the column corresponding to (v', θ), $v' \in \mathbf{V}'$ (resp. (v'', θ), $v'' \in \tilde{\mathbf{V}}''$) given by $\mathbf{x}_{v',\theta}$ (resp. $\mathbf{x}_{v'',\theta}$). We need to show that $\mathbf{X}$ has (real) rank $3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)$, and this is equivalent to showing that the associated $3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|) \times 3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)$ Gram matrix $\mathbf{X}^t\mathbf{X}$ has full rank (see 0.4.6(d) of [HJ85]).

The entry of $\mathbf{X}^t\mathbf{X}$ with indices $((v^*, \theta^*), (v^{**}, \theta^{**}))$, $v^*, v^{**} \in \mathbf{V}' \cup \tilde{\mathbf{V}}''$, $\theta^*, \theta^{**} \in \{\phi, \psi, \phi\psi\}$, is the usual scalar product of $\mathbf{x}_{v^*,\theta^*}$ with $\mathbf{x}_{v^{**},\theta^{**}}$, which is just the number of assignments of characters to leaves that assign θ^* to v^* and θ^{**} to v^{**}. We can compute this number of assignments as follows.

If $v^* = v^{**}$ and $\theta^* = \theta^{**}$, then it is clear by symmetry that this entry is 4^{m-1}, whereas if $v^* = v^{**}$ and $\theta^* \neq \theta^{**}$, then this entry is obviously 0.

Consider now the case where $v^* \neq v^{**}$, so that the collection of leaves descended from v^* in its tree is not the same as the collection of leaves descended from v^{**} in its tree. We claim that the entry of $\mathbf{X}^t\mathbf{X}$ with indices $((v^*, \theta^*), (v^{**}, \theta^{**}))$ is 4^{m-2}. To see this, write $\mathbf{L}^*$ and $\mathbf{L}^{**}$ for the leaves descended from v^* and v^{**}, respectively. Suppose first that $\mathbf{L}^* \backslash \mathbf{L}^{**}$, and $\mathbf{L}^{**} \backslash \mathbf{L}^*$ are both non-empty. If we have an assignment of characters to leaves that assigns the characters η^* to v^* and

η^{**} to v^{**}, then replacing the character assigned to some $\ell^* \in \mathbf{L}^* \backslash \mathbf{L}^{**}$ from χ^* (say) to $\rho^* \eta^* \chi^*$ and replacing the character assigned to some $\ell^{**} \in \mathbf{L}^{**} \backslash \mathbf{L}^*$ from χ^{**} (say) to $\rho^{**} \eta^{**} \chi^{**}$ gives a new assignment of characters to leaves that assigns ρ^* to v^* and ρ^{**} to v^{**} (recall that characters for $\mathbb{G}$ are their own inverses). It follows that number of assignments of characters to leaves that assign θ^* to v^* and θ^{**} to v^{**} is indeed 4^{m-2} when $\mathbf{L}^* \backslash \mathbf{L}^{**}$, and $\mathbf{L}^{**} \backslash \mathbf{L}^*$ are both non-empty. Similar arguments handle the cases $\mathbf{L}^* \subsetneq \mathbf{L}^{**}$ and $\mathbf{L}^{**} \subsetneq \mathbf{L}^*$, and we leave these to the reader.

We conclude that $\mathbf{X}^t \mathbf{X}$ can be partitioned into 3×3 blocks so that the blocks down the diagonal are all of the form

$$\begin{pmatrix} 4^{m-1} & 0 & 0 \\ 0 & 4^{m-1} & 0 \\ 0 & 0 & 4^{m-1} \end{pmatrix},$$

while the off–diagonal blocks are all of the form

$$\begin{pmatrix} 4^{m-2} & 4^{m-2} & 4^{m-2} \\ 4^{m-2} & 4^{m-2} & 4^{m-2} \\ 4^{m-2} & 4^{m-2} & 4^{m-2} \end{pmatrix}.$$

Now

$$\mathbf{X}^t \mathbf{X} = 4^{m-2}(\mathbf{D} + \mathbf{1}\mathbf{1}^t)$$

where $\mathbf{1}$ is the (column) vector with all entries equal to 1 and $\mathbf{D}$ is a matrix partitioned into 3×3 blocks with the blocks down the diagonal all of the form

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix},$$

and the off–diagonal blocks all zero. Note that $\mathbf{D}$ is invertible with inverse a partitioned matrix that has blocks down the diagonal all of the form

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix},$$

and the off–diagonal blocks all zero. A standard result on inverses of small rank perturbations (see 0.7.4 of [HJ85]) gives that $\mathbf{X}^t \mathbf{X}$ is indeed invertible (and hence full rank), with inverse

$$\begin{aligned} & 4^{-(m-2)} \left(\mathbf{D}^{-1} - \frac{1}{1 + \mathbf{1}^t \mathbf{D}^{-1} \mathbf{1}} \mathbf{D}^{-1} \mathbf{1}\mathbf{1}^t \mathbf{D}^{-1} \right) \\ & = 4^{-(m-2)} \left(\mathbf{D}^{-1} - \frac{1}{1 + 3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)} \mathbf{1}\mathbf{1}^t \right). \end{aligned}$$

□

Remark 3.2. The calculations in the above proof show that if $\mathbf{T}$ has m leaves and n vertices in total, then $\operatorname{rank}\mathcal{R}(\mathbf{T}) = 3n$ and hence $\operatorname{rank}\mathcal{N}(\mathbf{T}) = 4^m - 3n$. This gives another proof of the result from [EZ98] that the three–parameter Kimura model (with arbitrary root distribution) has $4^m - 3n$ algebraically independent invariants. The analogous counting results for the two–parameter Kimura and Jukes–Cantor models can be obtained similarly.

4. Three–leaved subtrees suffice

We begin with some general comments about trees. Suppose that we have a tree $\mathbf{T}$ with vertices $\mathbf{V}$ and leaves $\mathbf{L}$. Each pair of (possibly equal) leaves $\ell^*, \ell^{**} \in \mathbf{L}$ has a *most recent common ancestor* $\ell^* \wedge \ell^{**} \in \mathbf{V}$. That is, $\ell^* \wedge \ell^{**} \le \ell^*$ and $\ell^* \wedge \ell^{**} \le \ell^{**}$, and if v is another vertex with $v \le \ell^*$ and $v \le \ell^{**}$, then $v \le \ell^* \wedge \ell^{**}$. Each vertex $v \in \mathbf{V}$ is of the form $\ell^* \wedge \ell^{**}$ for some pair of leaves $\ell^*, \ell^{**} \in \mathbf{L}$.

Given a subset of leaves $\tilde{\mathbf{L}} \subseteq \mathbf{L}$, we can define the *reduced subtree induced by* $\tilde{\mathbf{L}}$. This is as tree $\tilde{\mathbf{T}}$ with leaf set $\tilde{\mathbf{L}}$ and vertex set $\tilde{\mathbf{V}} \subseteq \mathbf{V}$ given by the set of $\ell^* \wedge \ell^{**}$ for $\ell^*, \ell^{**} \in \tilde{\mathbf{L}}$. The tree structure is that inherited from $\mathbf{T}$, that is, the father of $v \in \tilde{\mathbf{V}}$ is the greatest element of $\tilde{\mathbf{V}}$ strictly less than v in the order $\le$ on $\mathbf{T}$. In other words, $\tilde{\mathbf{T}}$ is just like the usual graph–theoretic subtree spanned by $\tilde{\mathbf{L}}$, except that we "erase" non–leaf vertices that have outdegree 1.

Proposition 4.1. *Two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same leaf set $\mathbf{L}$ are identical if and only if for every subset of $\mathbf{L}$ of size 3 the reduced subtrees induced by this set in $\mathbf{T}'$ and $\mathbf{T}''$ are identical.*

Proof. The "only if" direction is obvious, so we consider the "if" direction.

Suppose that $\mathbf{T}'$ and $\mathbf{T}''$ are two trees with the same leaf set $\mathbf{L}$ such that for every subset of $\mathbf{L}$ of size 3 the reduced subtrees induced by this set in $\mathbf{T}'$ and $\mathbf{T}''$ are identical. Consider distinct leaves $\ell^*, \ell^{**} \in \mathbf{L}$. Write v' (resp. v'') for the most recent common ancestor of ℓ^* and ℓ^{**} in $\mathbf{T}'$ (resp. $\mathbf{T}''$).

The result will follow if we can show that the set of leaves descended from v' equals the set of leaves descended from v''. By symmetry, it further suffices to show that the set of leaves descended from v' is contained in the set of leaves descended from v''.

If ℓ^* and ℓ^{**} are the only leaves descended from v' in $\mathbf{T}'$, then we are done. Suppose, therefore, the ℓ is another leaf descended from v' in $\mathbf{T}'$. Because the reduced subtree induced by ℓ^*, ℓ^{**}, ℓ in $\mathbf{T}'$ is identical to the reduced subtree induced by ℓ^*, ℓ^{**}, ℓ in $\mathbf{T}''$, it is immediate that ℓ is descended from v'' in $\mathbf{T}''$, as required. □

Observe now that if we have a three–parameter Kimura model on a tree $\mathbf{T}$ with $m \ge 3$ leaves and we just observe the bases at 3 leaves, say the bases Y_1, Y_2, Y_3 at the leaves ℓ_1, ℓ_2, ℓ_3, then these bases are governed by a three–parameter Kimura model on the reduced subtree induced by ℓ_1, ℓ_2, ℓ_3. Therefore, if we have an invariant for the induced subtree it "lifts" to an invariant for $\mathbf{T}$ in the obvious way: the variable $p_{B_1 B_2 B_3}$ in the polynomial for the subtree is replaced by the "marginal probability" $\sum_{B_4} \cdots \sum_{B_m} p_{B_1 B_2 B_3 B_4 \ldots B_m}$ to produce a polynomial for $\mathbf{T}$.

Moreover, by letting certain of the random variables Z_v in the model on $\mathbf{T}$ be identically 0 (equivalently, letting certain of the distributions $\pi^{(v)}$ be point masses at 0) we see that as we range over all choices of parameters for a three–parameter Kimura model on $\mathbf{T}$ we also range over all choices of parameters for a three–parameter Kimura model on the induced subtree. Therefore, if a polynomial is not an invariant for the subtree, then lifting this polynomial to $\mathbf{T}$ in the manner described above produces a polynomial which is not an invariant for $\mathbf{T}$. In fact, the lifted polynomial will be non–zero for a dense subset of the three–parameter Kimura model substitution parameters for $\mathbf{T}$.

It follows from these observations, Proposition 4.1, and Theorem 3.1 that given two different trees on the same set of $m \ge 3$ leaves, we can always find 3 leaves and a three-parameter Kimura model invariant for a 3–leaved tree such that the lifted polynomial is a three-parameter Kimura model invariant for one of the trees but not the other. This observation also holds in the two–parameter Kimura and Jukes–Cantor cases.

Acknowledgements: The authors thank Tom Hagedorn and Susan Holmes for useful conversations.

References

[CF87] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. Classification*, 4:57–71, 1987.

[ES93] S.N. Evans and T.P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21:355–377, 1993.

[EZ98] S.N. Evans and X. Zhou. Constructing and counting phylogenetic invariants. *J. Comput. Biol.*, 5:713–724, 1998.

[HJ85] R.A. Horn and C.R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, 1985.

[JC69] 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.

[Kim80] 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.

[Kim81] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78:454–458, 1981.

[Lak87] J.A. Lake. A rate-independent technique for analysis of nucleic acid sequences:evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191, 1987.

[Ney71] 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.

E-mail address: evans@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

Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC V6T 2G3, Canada phone: (604)-822-2251, fax: (604)-822-6074
E-mail address: zhou@math.ubc.ca