Classifying and counting linear phylogenetic invariants for the Jukes Cantor model

by

M.A. Steel
Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand.

Y.X. Fu
Human Genetics Center
University of Texas at Houston, Texas, 77030, USA

No. 117 November, 1994

Keywords: Phylogenetic invariants, trees, forests, Hadamard matrix, Jukes-Cantor model.

ABSTRACT

Linear invariants are useful tools for testing phylogenetic hypotheses from aligned DNA/RNA sequences, particularly when the sites evolve at different rates. Here we give a simple, graph theoretic classification, for each phylogenetic tree T, of its associated vector space I(T) of linear invariants under the Jukes-Cantor one parameter model of nucleotide substitution. We also provide an easily-described basis for I(T), and show that if T is a binary (fully resolved) phylogenetic tree with n sequences at its leaves then :

$$\dim[\mathbf{I(T)}] = 4^{\mathbf{n}} - \mathbf{F_{2n-2}}$$

where $\mathbf{F_n}$ is the n-th Fibonacci number. Our method applies a recently-developed Hadamard-matrix based technique to describe elements of I(T) in terms of edge-disjoint packings of subtrees in T, and thereby complements earlier more algebraic treatments.

INTRODUCTION

The Jukes-Cantor model

Tree-based Markov models provide a simple mechanism for describing nucleotide substitution, and thereby estimating the underlying tree from aligned sequence data. In these models there is an underlying rooted tree T that represents the evolutionary history of the species being considered. Generally this tree is not known, and is the object to be estimated. The species are the leaves (degree 1 vertices) of this tree, labelled $1, 2, ..., n$, and the interior vertices correspond to (unknown) ancestral species, including a global ancestor, called the root of the tree. Such a tree is called a (rooted) *phylogenetic tree* (on $\{1, 2, ..., n\}$) though we will simply call it a *tree*. If all vertices other than the leaves and the root are of degree 3, the tree is said to be *binary*.

Each site in the aligned sequences is assumed to have evolved down the tree from an unknown random state $\alpha \in$ { A,C,G,T} at the root, to the extant states at the leaves, according to a Markov process. The simplest such process is the (equilibrium) Jukes-Cantor (JC) model. In this model, the states evolve according to a continuous-time, stationary Markov process, in which each state occurs with equal probability at the root, and the transition

rates on any edge of the tree are the same for each possible substitution (thus, on each edge e of the tree, there is just one mutation rate $\mu_e > 0$, for all substitutions $\alpha \to \beta$, $\alpha \neq \beta$). Consequently, for each edge, e, of T, the probability of any particular net substitution $\alpha \to \beta$ between the endpoints of the edge takes the same value for all $\alpha \neq \beta$, and we denote this value as p_e. Note that p_e may vary across edges, and between sites. Note also that $p_e = \frac{1}{4}[1 - \exp(-4\mu_e t_e)]$ where t_e is the temporal length of the edge, and so p_e lies in the half-open interval $[0, 0.25)$ (a further constraint, which we do not impose here, is given by the *molecular clock hypothesis* which requires μ_e to be the same for each edge e). The expected number of substitutions on edge e, which we denote as γ_e, is given by:

$$\gamma_e = -\frac{3}{4}\log_e(1 - 4p_e)$$

Since the probability P that the endpoints of edge e are in different states is given by $P = 3p_e$, this gives the well-known "Jukes-Cantor correction" relationship: $\gamma_e = -\frac{3}{4}\log_e(1 - \frac{4}{3}P)$. For more details on tree-based Markov type models see Rodriguez *et al.* (1990). The assignment of states to the leaves $1, ..., n$ of T will be called a *pattern*, and denoted χ. We let $P_\chi = P_\chi(T, \mathbf{p})$ denote the probability of generating χ under the JC model, where $\mathbf{p} = [p_e]$, and we let $\mathbf{P}(T, \mathbf{p})$ denote the (χ-indexed) column vector $[P_\chi(T, \mathbf{p})]$. Since each state at the root has probability $\frac{1}{4}$, we have:

$$P_\chi = \frac{1}{4}\sum_{\chi^*:\chi^*|\{1...n\}=\chi}\left[\prod_{e=(u,v):\chi^*(u)\neq\chi^*(v)} p_e \prod_{e=(u,v):\chi^*(u)=\chi^*(v)} (1 - 3p_e)\right]$$

where χ^* ranges over all assignments of states to the vertices of T that extend χ.

In Lemma 1 (below), we give an alternative formula for P_χ which, surprisingly, involves a tree-independent summation.

Linear Phylogenetic Invariants

A linear (phylogenetic) invariant for T under the JC model is a linear function $L(\mathbf{x}) = \sum_\chi \lambda_\chi x_\chi$ in indeterminants x_χ, indexed over all patterns, and with real coefficients λ_χ which satisfies the property:

$$\text{If } \mathbf{x} = \mathbf{P}(T, \mathbf{p}) \text{ then } L(\mathbf{x}) = 0.$$

Thus L vanishes whenever it is evaluated on the probability distribution arising from T (for any choice of the edge parameters p_e) under the JC model. The set of phylogenetic invariants for T forms a real vector space, which we denote as $I(T)$. Thus, $I(T) = \{\boldsymbol{\lambda}^t\mathbf{x} : \boldsymbol{\lambda}^t\mathbf{P}(T,\mathbf{p}) = 0, \text{ for all } \mathbf{p}\}$, where $\boldsymbol{\lambda}$ is the column vector $[\lambda_\chi]$, and where superscript t denotes transpose.

Let $M = \cap_T I(T)$, the vector space of *model invariants* of the JC model. Any L in M is an invariant for every tree, and so provides no information to discriminate between trees (in applications such invariants test the adequacy of the JC model). If $L \in I(T) - M$, for some T, then L is said to be *(phylogenetically) informative.*

Linear invariants for several models including some more general than Jukes-Cantor have been considered by other authors (see Lake (1987), Felsenstein (1991), Fu and Li (1992), and Nguyen and Speed (1992)); our aim here is to obtain, for the JC model, more information (including an exact enumeration) for linear invariants on any number of sequences, and a more tree-based representation of them; in addition, many of the linear invariants in the JC model do not exist in more general models.

The motivation for studying linear invariants comes from their application to sequences, which we now outline briefly: suppose each site in a collection of aligned DNA sequence data evolves according to the JC model, with underlying tree T – where the continuous parameter $\mathbf{p}$ may vary from site to site. Suppose that L is a linear phylogenetic invariant for T. Then, letting x_χ be the observed number of sites at which pattern χ occurs in the sequences, the expected value $\mathcal{E}(L)$ of $L = L(\mathbf{x})$ is identically zero (for any sequence length). Furthermore, provided the sites evolve independently, then L is approximately normally distributed for reasonable length sequences, and its variance can be estimated by conventional methods, thereby allowing statistical tests (see, for example, Navidi *et al.* (1991)) whereby trees whose associated invariants take sufficiently non-zero values for the data are rejected as potential candidates for the underlying evolutionary tree. The principal advantages of linear (over nonlinear) invariants is that (i) $\mathcal{E}(L)$ is known exactly for finite sequences, and (ii) the sites are not required to evolve identically.

For the Jukes-Cantor model, the space M has been characterized by Fu (1994), who showed that:

$$dim[M] = \frac{23 \times 4^{n-1} - 3 \times 2^{n-1} - 2}{6} \tag{1}$$

In addition, Fu (1994) constructed bases for $I(T)$ for binary trees with n

leaves, $1 \leq n \leq 7$, and showed that $dim[I(T)]$ is $3, 14, 59, 243, 990, 4007, 16151$, respectively.

If linear invariants are being used, in applications, to distinguish between trees (rather than to test the JC model) then the quotient space $I(T)/M$ is a more natural vector space to work in (selecting invariants $L_1, ..., L_k$ where $M+L_1, ...M+L_k$ is a basis for $I(T)/M$), and it is helpful that the dimension of this space for larger values of n is much smaller than $I(T)$ - for instance, when $n = 7$ we have $dim[I(T)/M] = 482$. In the Theorem we give an exact formula for $dim[I(T)]$ (in terms of the Fibonacci numbers) and we see that the ratio of $dim[I(T)/M]$ to $dim[I(T)]$ approaches $1/24$ as $n \to \infty$. For testing one tree T against another T' a natural space to work in is the quotient space $I(T)/[I(T) \cap I(T')]$, whose dimension grows more slowly than $dim[I(T)/M]$ (the ratio of dimensions tending to 0 as $n \to \infty$ by part (4) of our Theorem).

Results

Notation

We write $[n] = \{1, ..., n\}$ and $2^{[n]}$ for the power set of $[n]$. It is convenient to code the nucleotides A,C,G,T as elements of the *Klein four-group*, $\mathbf{Z}_2 \times \mathbf{Z}_2$, as in Evans and Speed (1993). Thus,

$$\text{A} = (0,0), \text{C} = (1,0), \text{G} = (0,1), \text{ and } \text{T} = (1,1),$$

with addition $\oplus$ carried out in this group (i.e. componentwise, modulo 2) so that, for example, $\text{C} \oplus \text{C} = \text{A}$, and $\text{C} \oplus \text{T} = \text{G}$.

In this way each pattern χ is associated in a one-to-one fashion with a pair (θ, α), where $\alpha \in \{\text{A,C,G,T}\}$ is the state assigned to leaf n, and $\theta = (\sigma_1, \sigma_2)$ is a pair of subsets of $[n-1]$ determined by:

$$\sigma_1 := \{i : \chi_i \oplus \chi_n = \text{C or T}\};\ \sigma_2 := \{i : \chi_i \oplus \chi_n = \text{G or T}\}.$$

where χ_i and χ_n are respectively the states of leaf i and n.

We denote this association by writing $\chi = \chi(\theta, \alpha)$. In case $\alpha = A$, we write this more briefly as $\chi(\theta)$. For example, the pattern $\chi = \text{CACTGA} = \chi(\theta)$, where $\theta = (\{1,3,4\}, \{4,5\})$, while $\chi = \text{TTGCCC}= \chi(\theta, \alpha)= \chi((\{3\}, \{1,2,3\}), \text{C})$.

Now, $\chi(\theta, \alpha)$ is obtained from $\chi(\theta)$ by applying a permutation (dependent on α) on {A,C,G,T}. Thus, by the equilibrium assumption of states at the

root of T, and the form of the transition matrices in the JC model, $P_\chi(\theta, \alpha)$ is the same for each value of α. Thus, $P_\chi(\theta, \alpha) = P_\chi(\theta)$ for all α, and so, if we let

$$L_{\theta,\alpha}(\mathbf{x}) := x_{\chi(\theta)} - x_{\chi(\theta,\alpha)}$$

then $L_{\theta,\alpha}$ is a nonzero linear model invariant for each θ, and $\alpha \neq$ A. Moreover, the collection:

$$B(n) := \{L_{\theta,\alpha} : \alpha = \text{ C,G,T } ; \theta \in 2^{[n-1]} \times 2^{[n-1]}\} \tag{2}$$

consists of $3 \times 4^{n-1}$ linearly independent elements (but not a basis!) of M.

To proceed further we need to describe how to represent $\mathbf{P}(T, \mathbf{p})$ by a formula (Lemma 1) that involves two sets of paths in T, where each set of paths is edge-disjoint. It is convenient to think of these two sets of paths as forming a packing of edge-disjoint subtrees of T (Lemma 2), as this simplifies their enumeration for binary trees (Lemma 3) and for establishing our other main results.

We first describe how the two sets of edge-disjoint paths arise. We will let ω denote throughout any pair (X_1, X_2), where each $X_i (i = 1, 2)$ is a subset of $[n]$ of even cardinality. Now, for any phylogenetic tree on $[n]$, each X_i induces a unique set of edges, denoted $\mathcal{P}(T, X_i)$, as follows: pair off, in any way, the leaves of T that are labelled by elements of X_i. Then $\mathcal{P}(T, X_i)$ is the set of edges that lie in an odd number of paths (this is independent of the pairing on X_i). We can always choose the paths to be edge disjoint. In addition, for binary trees we can insist that the paths be vertex disjoint as well, in which case the collection of paths is unique (this is not necessarily true for nonbinary trees). For any such $\omega = (X_1, X_2)$, let

$$\mathcal{P}(T, \omega) := \mathcal{P}(T, X_1) \cup \mathcal{P}(T, X_2) \ .$$

For $\theta = (\sigma_1, \sigma_2)$ let $\theta\omega = |\sigma_1 \cap X_1| + |\sigma_2 \cap X_2|$ and let H denote the $4^{n-1} \times 4^{n-1}$ matrix $[(-1)^{\theta\omega}]$, with rows indexed by the $\theta's$ and columns indexed by the $\omega's$. Let $z_e = (1 - 4p_e) = \exp(-\frac{4}{3}\gamma_e)$, and let $\mathbf{z}$ be the ω-indexed column vector $\mathbf{z} = [z_\omega]$ with $z_\omega = \prod_{e \in \mathcal{P}(T,\omega)} z_e$.

Lemma 1 *[Székely et al. (1993)]. H is a Hadamard matrix, and*

$$\mathbf{P}(T, \mathbf{p}) = \frac{1}{4^n} H\mathbf{z}.$$

Thus, Lemma 1 asserts that $HH^t = H^tH = 4^{n-1}I$, and if $\chi = \chi(\theta, \alpha)$, for any α, then $P_\chi(T, \mathbf{p}) = \frac{1}{4^n}\sum_\omega (-1)^{\theta\omega} \prod_{e \in P(T,\omega)} z_e$. This result is the restriction of a more general result (for the Kimura 3ST model) in Székely *et al.* (1993) to the JC model. The more general result has been useful for classifying the nonlinear phylogenetic invariants of the Kimura 3ST model in Steel *et al.* (1993) (see also Evans and Speed, (1993)) and is a generalization of the pioneering work of Hendy and Penny (1989) on a similar representation for the two-state (Cavender-Farris) model.

Definition Given a phylogenetic tree T, a *subforest* $\mathfrak{J}$ of T is a set of edge-disjoint subtrees of T which have all their degree-one vertices in the leaves of T. We allow $\mathfrak{J} = \emptyset$, and let $s(T)$ be the set of subforests of T. If $\mathcal{P}(T, \omega)$ is the set of edges of a subforest $\mathfrak{J}$ of T we write $\omega \rightarrow_T \mathfrak{J}$.

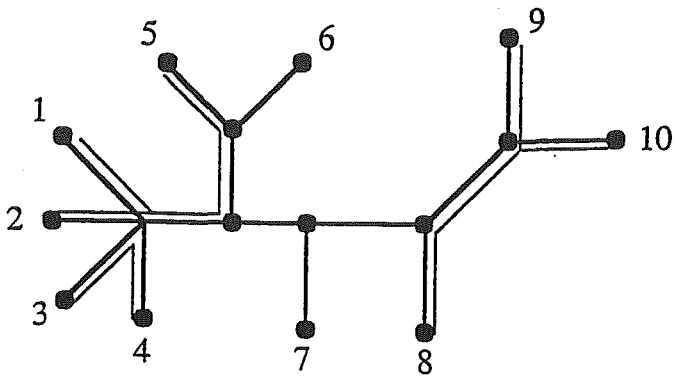

FIG. 1. A subforest $\mathfrak{J}$ of T consisting of three edge-disjoint trees. An example of a pair ω of even cardinality subsets of $[n]$ ($n = 10$) for which $\omega \rightarrow_T \mathfrak{J}$ is $\omega = (\{1,2,3,4,8,9\}, \{1,5,8,10\})$.

An example of a subforest is given in Fig. 1. Note that, for a binary tree, the component trees in any subforest $\mathfrak{J}$ are vertex disjoint and so $\mathfrak{J}$ is determined uniquely by its set of edges, but this is not necessarily true for a

nonbinary tree.

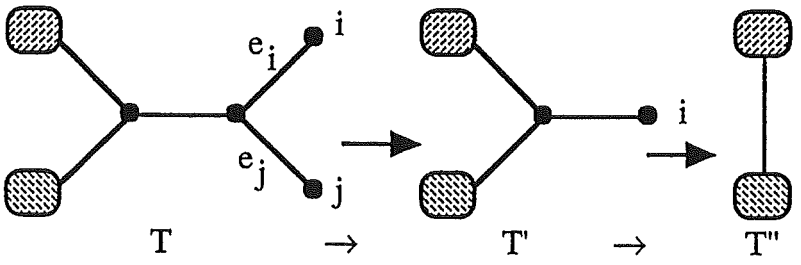

FIG. 2. Two succesive pruning operations to obtain binary trees T' and T'' from a binary tree T.

Lemma 2. *$\mathfrak{J}$ is a subforest of T precisely if $\omega \rightarrow_T \mathfrak{J}$ for some ω.*

Proof. $\mathcal{P}(T,\omega)$ is clearly the edge set of a subforest of T. Conversely, suppose $\mathfrak{J}$ is a subforest of T. We show that induction on n, the number of leaves of T that $\omega \rightarrow_T \mathfrak{J}$ for some ω. The result holds for $n < 4$, so suppose T has $n \geq 4$ leaves. Then we can select a vertex v of T that is adjacent to leaves $x_1, ..., x_r$, for some $r > 1$, and is adjacent to one other vertex w. Suppose, firstly, that one such leaf x_i is not in any component of $\mathfrak{J}$. In that case delete x_i and its incident edge to obtain a tree T' with one less leaf. By induction $\omega \rightarrow_{T'} \mathfrak{J}$ for some w, hence $\omega \rightarrow_T \mathfrak{J}$, also. On the other hand, if each x_i is contained in a component of $\mathfrak{J}$, then, since $r > 1$, and since the edge $e = [v, w]$ of T can appear in at most one component of $\mathfrak{J}$, it follows that two of the above leaves - say x_i, x_j - are in the same component tree $t \in \mathfrak{J}$. Now, if these are the only two leaves of t, let T'' be the tree obtained from T by deleting these two leaves, and their incident edges (and if $r = 2$ delete also the vertex v and its incident edge, as in Fig.2). Letting $\mathfrak{J}' = \mathfrak{J} - \{t\}$, by induction we have $\omega \rightarrow_{T''} \mathfrak{J}'$ and so $\omega^* \rightarrow_T \mathfrak{J}$, where ω^* is obtained by adding x_i and x_j to one of the two even cardinality sets in ω. For the remaining

case, when t has at least three leaves, delete x_j and its incident edge from T to obtain a tree T'. Let $\mathfrak{J}'$ be the forest obtained from $\mathfrak{J}$ by deleting from T leaf x_j and its incident edge. Again invoking the induction hypothesis we have $\omega \rightarrow_{T'} \mathfrak{J}'$, for some $\omega = (X_1, X_2)$. If $x_i \in X_1 \cap X_2$, let $X_1' = X_1$, $X_2' = (X_2 - \{x_i\}) \cup \{x_i, x_j\}$; if x_j is just in one of the components of ω, say X_1, let $X_1' = X_1$, $X_2' = X_2 \cup \{x_i, x_j\}$ (similarly if $x_i \in X_2 - X_1$). In either case we obtain a pair ω of even cardinality subsets of $[n]$, for which $\omega \rightarrow_T \mathfrak{J}$, as required.

Lemma 3. *A binary tree T with n leaves has precisely F_{2n-2} subforests, where F_n is the n-th Fibonacci number.*

Proof. Suppose firstly that $n > 3$. Select a pair of leaves (i, j) of T which are separated by just two edges, e_i and e_j (see Fig. 2). Let T' be the binary trees obtained from T by deleting j and its incident edge e_j. Let T'' be the binary tree obtained from T' by deleting leaf i and its incident edge, and making the resulting tree homeomorphically irreducible (suppressing the vertex of degree 2) as shown in Fig. 2. We claim that, for $n > 3$:

$$|s(T)| = 3|s(T')| - |s(T'')| \tag{3}$$

Let $s_1(T') = \{\mathfrak{J} \in s(T') : j \text{ is a leaf of } \mathfrak{J}\}$; $s_2(T') = s(T') - s_1(T')$, and let $E(\mathfrak{J})$ denote the set of edges of any subforest $\mathfrak{J}$. Each $\mathfrak{J} \in s_1(T')$ produces three subforests of T, namely (i) $\mathfrak{J}$, (ii) the subforest obtained from $E(\mathfrak{J})$ by replacing e_i and e_j, and (iii) the subforest obtained by adding e_j to $E(\mathfrak{J})$. Each $\mathfrak{J} \in s_2(T')$ produces two subforests of T, namely (i) $\mathfrak{J}$, and (ii) the subforest which adds e_i and e_j to $E(\mathfrak{J})$. Since each subforest of T arises in precisely one such way from either $s_1(T')$ or $s_2(T')$, we get:

$$|s(T)| = 3|s_1(T')| + 2|s_2(T')| = 3|s(T')| - |s_2(T')|.$$

Equation (3) now follows by identifying $s_2(T')$ with $s(T'')$. From Equation (3) an inductive argument shows that $|s(T)|$ depends only on n. Thus, letting $s(n) = |s(T)|$ for any binary tree on n leaves, we have $s(n) = 3s(n-1) - s(n-2)$, with starting values $s(2) = 2$, $s(3) = 5$, and this recursion is satisfied by $s(n) = F_{2n-2}$.

Before presenting our main results, it is necessary to recall (for parts (4) and (5)) the notions of compatibility and (strict) consensus. Given trees T and T' on $[n]$ write $T' \preceq T$ if T' can be obtained from T by collapsing

a subset of the edges of T. Then $\preceq$ is a partial order, and any collection of trees has a unique, and easily-computed greatest lower bound under this ordering, called the *strict consensus* of the collection. If T and T' have an upper bound (i.e. a tree T'' with $T, T' \preceq T''$)) then T and T' are said to be *compatible*, otherwise they are *incompatible* (for example, two different binary trees are necessarily incompatible).

Theorem *(1) $L(\mathbf{x}) = \sum_\chi \lambda_\chi x_\chi$ is a phylogenetic invariant for T under the JC model, if and only if, for all subforests $\mathfrak{J}$ of T,*

$$\sum_\theta \left[\sum_{\omega:\ \omega \to_T \mathfrak{J}} (-1)^{\theta\omega} \right] \mu_\theta = 0$$

where $\mu_\theta = \sum_\alpha \lambda_{\chi(\theta,\alpha)}$

(2) For each $\mathfrak{J} \in s(T)$, select any ω for which $\omega \to_T \mathfrak{J}$, and call it $\omega(\mathfrak{J})$ (this is possible by Lemma 2). A basis for the space $I(T)$ of phylogenetic invariants for T is the (disjoint) union of $B(n)$ and $B(T)$, where $B(n)$ is given by Equation (2), and where $B(T)$ is the collection of invariants $L_{\mathfrak{J},\omega}$ of the form:

$$L_{\mathfrak{J},\omega}(\mathbf{x}) = \sum_\theta \left[(-1)^{\theta\omega(\mathfrak{J})} - (-1)^{\theta\omega} \right] x_{\chi(\theta)}$$

over all pairs $(\mathfrak{J}, \omega)$, where $\mathfrak{J}$ is a subforest of T and $\omega \to_T \mathfrak{J}$, $\omega \neq \omega(\mathfrak{J})$.

(3) $dim[I(T)] = 4^n - |s(T)|$. In particular, if T is a binary tree,

$$dim[I(T)] = 4^n - F_{2n-2},$$
$$dim[I(T)/M] = \frac{(2^{n-1}+1)(2^{n-1}+2)}{6} - F_{2n-2};$$

(4) If T and T' are two binary trees, and T^* is their strict consensus, then

$$dim[I(T)/I(T) \cap I(T')] \leq F_{2n-2} - |s(T^*)| \leq F_{2n-2} - 2^n + n$$

(5) For any tree T there is an associated linear invariant L_T which has the property that for any tree T' incompatible with T we have:

$$\text{If } \mathbf{x} = \mathbf{P}(T', \mathbf{p}), \mathbf{p} > \mathbf{0}, \text{ then } L_T(\mathbf{x}) > 0.$$

Proof. It is helpful to introduce two new matrices K and A. Let $K = [K_{\mathfrak{J},\omega}]$ be the $(0,1)$-matrix with $K_{\mathfrak{J},\omega} = 1$, if $\omega \rightarrow_T \mathfrak{J}$, and $K_{\mathfrak{J},\omega} = 0$ otherwise. Let $A = KH^t$. Thus $A = [A_{\mathfrak{J},\theta}]$ is the $|s(T)| \times 4^{n-1}$ matrix with

$$A_{\mathfrak{J},\theta} := K_{\mathfrak{J}} H_\theta^t = \sum_{\omega:\omega \rightarrow_T \mathfrak{J}} (-1)^{\theta\omega}$$

(1) Write $L(\mathbf{x}) = \sum_\chi \lambda_\chi x_\chi = \sum_\theta \sum_\alpha \lambda_{\chi(\theta,\alpha)} x_{\chi(\theta,\alpha)}$. Now, $P_{\chi(\theta)} = P_{\chi(\theta,\alpha)}$, so that if L is an invariant for T we have,

$$\sum_\theta \mu_\theta P_{\chi(\theta)} = 0 \tag{4}$$

Equation (4) is equivalent, by Lemma 1, to requiring that the column vector $\boldsymbol{\mu} = [\mu_\theta]$ satisfies $\boldsymbol{\mu}^t(H\mathbf{z}) = 0$. Now, $\mathbf{z} = K^t\mathbf{w}$, where $\mathbf{w}$ is the $\mathfrak{J}$-indexed column vector $[w_{\mathfrak{J}}]$ with $w_{\mathfrak{J}} = \prod_{e \in E(\mathfrak{J})} z_e$ (where $E(\mathfrak{J})$ is the set of edges of $\mathfrak{J}$), and so,

$$\boldsymbol{\mu}^t(H\mathbf{z}) = \boldsymbol{\mu}^t(HK^t\mathbf{w}) = \boldsymbol{\mu}^t(A^t\mathbf{w}) = (A\boldsymbol{\mu})^t\mathbf{w}. \tag{5}$$

Thus, (from Equation (5)) we see that Equation (4) implies $(A\boldsymbol{\mu})^t\mathbf{w} = \mathbf{0}$. This must hold for all choices of $\mathbf{p} \in [0, 0.25)^\epsilon$ (where ϵ = number of edges of T), and hence for all choices of $\mathbf{z} \in (0,1]^\epsilon$. Thus, if we regard the z_e's as indeterminants in the real polynomial ring $\mathcal{R}[z_{e^1}, ..., z_{e^\epsilon}]$ (where $e^1, ..., e^\epsilon$ are the edges of T), then $(A\boldsymbol{\mu})^t\mathbf{w}$ must be the zero element of this ring and hence the coefficient of every monomial $w_{\mathfrak{J}} = \prod_{e \in E(\mathfrak{J})} z_e$ in $(A\boldsymbol{\mu})^t\mathbf{w}$ must be zero. But the coefficient of $w_{\mathfrak{J}}$ in $(A\boldsymbol{\mu})^t\mathbf{w}$ is just $(A\boldsymbol{\mu})_{\mathfrak{J}}$, and so $(A\boldsymbol{\mu})_{\mathfrak{J}} = 0$ for all $\mathfrak{J}$, that is $A\boldsymbol{\mu} = \mathbf{0}$, which translates into the condition described in part (1).

(2). From part (1) of this Theorem, the space of (real) vectors $\boldsymbol{\mu} = [\mu_\theta]$ for which $\sum_\theta \mu_\theta x_{\chi(\theta)} \in I(T)$ is precisely the (right) null-space of A,

$$NS(A) := \{\mathbf{y} : A\mathbf{y} = \mathbf{0}\}.$$

Since H is a Hadamard matrix, and so is of full rank, any basis $\{\mathbf{w}^1, ..., \mathbf{w}^r\}$ of $NS(AH)$, provides a basis $\{H\mathbf{w}^1, ..., H\mathbf{w}^r\}$ for $NS(A)$. Now, since $A = KH^t$ and since H is Hadamard, $AH = 4^{n-1}K$, and hence:

$$NS(AH) = NS(K). \tag{6}$$

Furthermore, since each column of K contains exactly one nonzero entry, to find a basis for $NS(K)$ it suffices to take the union (over $\mathfrak{J}$) of the bases for $NS(K_{\mathfrak{J}})$, where $K_{\mathfrak{J}}$ is the $\mathfrak{J}$-indexed row of K. Since the entries in $K_{\mathfrak{J}}$ are 0's and 1's a basis for $NS(K_{\mathfrak{J}})$ is the set of the vectors $\{\mathbf{e}^{\omega(\mathfrak{J})} - \mathbf{e}^{\omega} : \omega \rightarrow_T \mathfrak{J}, \omega \neq \omega(\mathfrak{J})\}$, where $\mathbf{e}^{\omega}$ is the unit vector with a 1 in the ω-position, and 0 elsewhere. Taking the union of these bases, we obtain, from (4), a basis for AH, and thereby (applying H) the given basis $B(T)$ for the subspace $I_0(T)$ of $I(T)$ of invariants of the forms $\sum_{\theta} \mu_{\theta} x_{\chi(\theta)}$. Thus, since $B(n)$ is a basis for the null space of the linear transformation from $I(T)$ to $I_0(T)$ given by $\boldsymbol{\lambda}^t \mathbf{x} \rightarrow \boldsymbol{\mu}^t \mathbf{x}$, it follows that $B(T) \cup B(n)$ is a basis for $I(T)$.

(3) From part (2),

$$\begin{aligned} dim[I(T)] &= |B(T)| + |B(n)| = (4^{n-1} - |s(T)|) + 3 \times 4^{n-1} \\ &= 4^n - |s(T)|. \end{aligned}$$

For a binary tree, T, $|s(T)| = F_{2n-2}$ from Lemma 3 and $dim[I(T)/M] = dim[I(T)] - dim[M]$ where $dim[M]$ is given by Equation (1).

(4) The tree T^* is obtained from T and T' by collapsing edges, and so, for any edge parameters $\mathbf{p}$ for T^*, we have $\mathbf{P}(T^*, \mathbf{p}) = \mathbf{P}(T, \mathbf{p}^1) = \mathbf{P}(T', \mathbf{p}^2)$ for suitable $\mathbf{p}^1, \mathbf{p}^2$ (which assign probability 0 to any edge of T [resp. T'] that is collapsed). It follows that $I(T)$ and $I(T')$ are both subspaces of $I(T^*)$. Thus, the direct sum $I(T) + I(T')$ is also a subspace of $I(T^*)$, and so:

$$\begin{aligned} dim[I(T^*)] &\geq dim[I(T) + I(T')] \\ &= dim[I(T)] + dim[I(T')] - dim[I(T) \cap I(T')] \\ &= 2(4^n - F_{2n-2}) - dim[I(T) \cap I(T')]. \end{aligned}$$

Thus, since $dim[I(T^*)] = 4^n - |s(T^*)|$, $dim[I(T) \cap I(T')] \geq 4^n - 2F_{2n-2} + |s(T^*)|$. Again invoking part (3),

$$\begin{aligned} dim[I(T)/I(T) \cap I(T')] &= dim[I(T)] - dim[I(T) \cap I(T')] \\ &= (4^n - F_{2n-2}) - dim[I(T) \cap I(T')] \end{aligned}$$

which combined with the previous inequality, establishes the first inequality in part (4). The second inequality in part (4) follows from the observation that any subset of $[n]$, other than a singleton subset, determines a subforest of T^* (the minimal subtree of T^* connecting the leaves in this subset) and thus $|s(T^*)| \geq 2^n - n$.

(5) Consider the binary tree $T = ij|kl$ on four leaves (in which the path connecting leaves i and j is disjoint from the path connecting leaves k and l). Note that:

$$\mathfrak{J} = \mathcal{P}(T, (\{i,j\},\{k,l\})) = \mathcal{P}(T, (\{i,j,k,l\}, \emptyset)).$$

Thus we obtain a linear invariant $L_{\mathfrak{J},\omega}$ for T, by taking $\omega(\mathfrak{J}) = (\{i,j,k,l\}, \emptyset)$, $\omega = (\{i,j\},\{k,l\})$. Furthermore this invariant has the property that it is always strictly positive when evaluated at $\mathbf{x} = \mathbf{P}(T', \mathbf{p})$ for either of the other two binary trees T' on four leaves, provided $p_e > 0$ on the internal edge of T'. To see this note that $L_{\mathfrak{J},\omega} = (H^t\mathbf{x})_{\omega(\mathfrak{J})} - (H^t\mathbf{x})_\omega$ where $\mathbf{x} = [x_\theta]$; setting $\mathbf{x} = \mathbf{P}(T', \mathbf{p})$, we have, from Lemma 1, that $\mathbf{x} = 4^{-n}H\mathbf{z}'$, and so, $L_{\mathfrak{J},\omega}(\mathbf{x}) = z'_{\omega(\mathfrak{J})} - z'_\omega > 0$ (provided $p_e > 0$ on the internal edge e of T'). Thus, the claim holds for four leaves. Of course if T has more than four leaves and has the above tree $ij|kl$ as a subtree (when attention is restricted to i,j,k,l and degree two vertices are ignored) then we obtain a linear invariant for T by simply summing out the states of all the other leaves. Let us denote this linear invariant as $L_T(ij|kl)$. Now, two trees T and T' are incompatible precisely if T has a quartet of leaves $\{i,j,k,l\}$ which is resolved into two different binary trees by T and by T'. Thus, if we let L_T be the sum of the invariants $L_T(ij|kl)$ for all induced subtrees $ij|kl$ of T we obtain the claimed result.

REMARKS

(i) $\mathfrak{J} = \emptyset$ in (1) gives $\sum_\theta \mu_\theta = 0$ as a necessary condition for a linear invariant.

(ii) In the proof of (5) we gave an example of a phylogenetically informative invariant in $B(T)$ for the tree $T = ij|kl$. Another informative invariant is given by the identity:

$$\mathfrak{J} := \{T\} = \mathcal{P}(T, (\{i,k\},\{j,l\})) = \mathcal{P}(T, (\{i,l\},\{j,k\}))$$

however this does not share the strong property enjoyed by the invariant described in (5). Note that for the three binary trees T, T', T'' on four leaves we have (from the Theorem) that $dim[I(T)/I(T) \cap I(T') \cap I(T'')] = 2$; $dim[I(T)/I(T) \cap I(T')] = 1$.

(iii) One advantage of the type of invariants described in part (5) is that they allow one-sided statistical tests, rather than two-sided tests. Note that

the condition $\mathbf{p} > \mathbf{0}$ is stronger than really necessary (p_e can be zero on any edge e of T that is incident with a leaf).

(iv) The space of model invariants, M, comprises "most" of $I(T)$ in the sense that the ratio of their dimensions tends to $23/24$ as n tends to infinity.

ACKNOWLEDGEMENT:

For support and encouragement to visit Houston, the first author kindly thanks Professor W. H. Li.

REFERENCES

Evans, S. N. and Speed, T. P. 1993. Invariants of some probability models used in phylogenetic inference. *Ann. Stat.* **21**, 355-377.

Felsenstein, J. 1991. Counting phylogenetic invariants in some simple cases. *J. Theor. Biol.* **7**, 357-76.

Fu, Y.X. 1994. Linear invariants under Jukes and Cantor's one parameter model, submitted to *J. Theor. Biol.*

Fu, Y. X. and Li, W. H. 1992. Construction of linear invariants in phylogenetic inference. *Math. Biosci.* **109**, 209-228.

Hendy M.D. and Penny, D. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38** (4), 297-309.

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

Navidi, W.C., Churchill, G.A. and von Haeseler, A. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.* **8** (1), 128-143.

Nguyen, T. and Speed, T. P. 1992. A derivation of all linear invariants for a nonbalanced transversion model. *J. Mol. Evol.* **35**, 60-76.

Rodriguez, F., Oliver, J.L., Marin, A. and Medina, J.R. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485-501.

Steel, M.A., Székely, L.A., Erdös, P.L. and Waddell, P. 1993. A complete family of phylogenetic invariants for any number of taxa under the Kimura 3ST model. *NZ. J. Botany* **31**, 289-296.

Székely, L.A., Erdös, P.L., Steel, M.A. and Penny, D. 1993. A Fourier inversion formula for evolutionary trees, *Appl. Math. Lett.* **6** (2), 13-16.