New Zealand Journal of Botany, 1993, Vol. 31: 289–296
0028–825X/93/3103–0289 $2.50/0

A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model

MIKE STEEL
Mathematics Department
Massey University
Palmerston North, New Zealand*

LASZLO SZÉKELY
Department of Mathematics
University of New Mexico
Albuquerque, NM 87131, U.S.A.

PETER L. ERDÖS
Hungarian Academy of Science
and CWI, Amsterdam,
The Netherlands

PETER WADDELL
Plant Biology Department
Massey University
Palmerston North, New Zealand

*Present address: Department of Mathematics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

Abstract We describe a new family of phylogenetic invariants that arise from the recently developed spectral analysis approach to tree reconstruction. These invariants, which are valid for Kimura's 3ST model, possess four important properties—they are defined equally easily for any number of taxa, their description is tree-independent, they apply even when the distribution of the four nucleotides in the ancestral taxon is unknown, and they can be modified to deal with sequence sites that do not mutate independently with identical distribution.

B93011
Received 19 January 1993; accepted 16 April 1993

Keywords genetic sequences; phylogenetic trees; phylogenetic invariants; Kimura's three-parameter model; convergence in probability

INTRODUCTION

Recently there has been considerable interest in phylogenetic invariants (Cavender 1989, 1991; Fu & Li 1991, 1992; Nguyen & Speed 1992; Evans & Speed 1993; Ferretti & Sankoff 1993). Broadly speaking, phylogenetic invariants are functions which, when evaluated on "ideal" sequence data, take a value that depends only on the (dimensionless) underlying evolutionary tree linking the taxa. The motivation for their development is that they allow, in principle, the consistent reconstruction of an evolutionary tree from sequence data without having to deal with a large number of unknown parameters.

All phylogenetic invariants are based on some stochastic model describing the observed differences in the aligned segments of genetic sequences for a collection of extant species. The validity of a particular invariant therefore depends on the correctness of its relevant model. This model, which we shall denote throughout as M, encapsulates the underlying probabilistic mechanism by which the unknown sequence of the ancestral taxon mutated randomly over time to result in the observed forms in extant taxa.

It is important to distinguish two aspects of M—the *stochastic features* (e.g., whether or not mutations at different sequence sites and/or places in the tree are independent) and the *parameters*. The parameters of M generally comprise one discrete variable—the true tree, denoted T, that is, the historically correct evolutionary (gene) tree connecting the extant species—as well as continuous variables, denoted P, which mostly pertain to edges of the true tree (and may be related to time) or perhaps, additionally, describe the distribution of rates and correlations of mutation at different sites on the sequence.

A desirable aspect of M is that it imparts to the

data sufficient traces of the true tree connecting the species so as to enable this tree to be reconstructed, at least from "ideal" sequences. In practice this generally means that M should incorporate some independence assumptions as stochastic features, and not possess too many freely adjustable parameters. Increasing the number of parameters and weakening the assumptions in M allows M to better approximate nature; however, this brings a risk of overfitting the data; indeed, with sufficient flexibility, any sequence data can be described perfectly by any tree!

Attempts to reconstruct T without having to worry about the continuous parameters motivated the study of invariants, which began around 1987 from two quite different directions. Cavender & Felsenstein (1987) published a seminal paper which derived quadratic invariants for Cavender/Farris's symmetric two-character state nucleotide model (with four taxa). Lake (1987) independently developed his "evolutionary parsimony" method, which used linear polynomials to identify a tree under a different stochastic model of four-character state nucleotide sequences, still with four taxa. Subsequent work has attempted, with mixed success, to (1) generalise these methods systematically to sets of more than four taxa, (2) extend to more general models, and (3) develop statistical tests by which invariants can discriminate between two trees, or construct confidence intervals of trees (as in Navidi et al. 1991).

Fu & Li (1991) and Evans & Speed (1993) have analysed a quite different model due to Kimura (1981), which has three parameters describing the rates of transition and the two types of transversion substitutions—these rates being allowed to vary for each edge of the tree. Fu & Li (1991) show the transition matrices for this model form a semigroup and allow certain events to be pairwise independent, and furthermore, that no larger semigroup can allow this independence. Evans & Speed (1993) provide a complete characterisation of when any given polynomial is an invariant for a tree under this model. Their characterisation is general (although it assumed equal frequencies of the four states at the root, something we dispense with) and is valid for any number of taxa. They did not explicitly construct the simple and complete family of invariants for any number of taxa that we describe below. Evans & Speed exploited the fact that Kimura's model is essentially described by a finite group—the Klein four-group. It was this observation that allowed for the construction of our invariants, which are essentially a consequence of the recent extension of Hendy & Penny's (1993) "spectral analysis" method from two-state to four-state character sequences (see Steel et al. 1992).

In parallel, much attention has been devoted to constructing, counting, and classifying linear invariants (e.g., Nguyen & Speed 1992; Fu & Li 1992). There are good statistical reasons to prefer linear invariants over higher order polynomial or more general invariants (see Navidi et al. 1991); however, for several models, such as Kimura's three-parameter model, there exist no linear invariants (apart from the trivial one). Furthermore, restricting attention to linear polynomial invariants generally neglects a large amount of information contained in non-linear invariants. Among non-linear invariants, it appears that only polynomial functions have been considered, though these seem to have few statistical (or other) advantages over invariants based on other analytical functions.

The organisation of this paper is as follows. We first provide a definition for phylogenetic invariants. We then describe the three-parameter model for nucleotide substitution proposed by Kimura (1981). Under the further assumption that sequence sites evolve "independently and identically" (the i.i.d. assumption) we construct a collection of invariants. An extension is then described that provides invariants in the case where this i.i.d. assumption is relaxed, to allow for different rates across sites, and a limited degree of dependence between sites. In both cases, these invariants asymptotically, uniquely identify the tree that generated the data. An obvious weakness in this last statement is the word "asymptotically"—statistical tests based on the nonasymptotic properties of invariants (Waddell et al. in press) are required for them to be useful in practise, and so our results are just a first step in this direction.

PHYLOGENETIC INVARIANTS – DEFINITIONS

Throughout this paper we adopt the conventions of letting $[n]$ denote the set $\{1,...,n\}$, and $[X_i]$ denote the column vector whose i-th component is X_i. We use standard terminology for describing aspects of phylogenetic trees (e.g., Steel 1992), which are trees whose degree 1 vertices (leaves) are labelled and whose remaining vertices are unlabelled and of degree ≥ 3, though we refer throughout to phylogenetic trees simply as trees. Also, we adopt the following convention: if a rooted tree T, as in Fig.

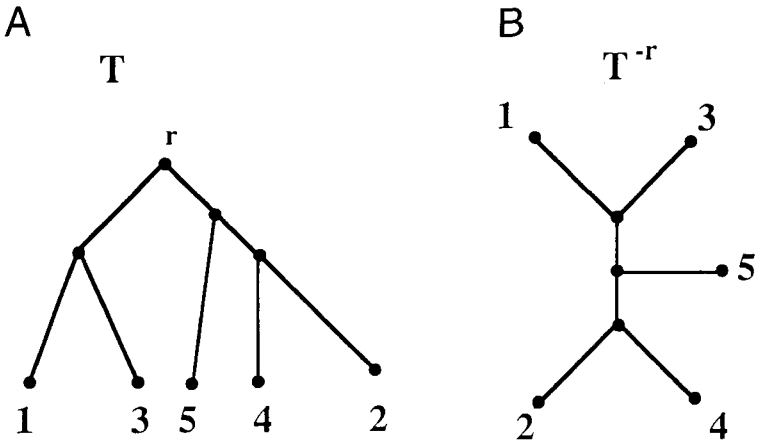

Fig. 1 **A**, A rooted phylogenetic tree T having leaves labelled $[5] = \{1,...,5\}$. **B**, The unrooted phylogenetic tree on $[5]$ obtained from T by suppressing the root r, denoted T^{-r}.

1A, has a root r, of degree 2, let T^{-r} denote the tree obtained from T by deleting r and replacing its two incident edges with a single edge, as in Fig. 1B. If the degree of r is more than 2, we can take T^{-r} to be simply T itself, regarding r as just one of the unlabelled internal vertices of T.

A number of essentially similar definitions for phylogenetic invariants have been proposed—here our definition emphasises the fact that phylogenetic invariants are calculated on observed data rather than "ideal" data. If the taxa are labelled $1,...,n$, then the collection of states for the different taxa at a given site in the aligned sequences is an ordered n-tuple $\pi = (S_1,...,S_n)$ where each S_i is either A, C, G or T (U if using RNA sequences), and this π is often called a *pattern.* For each pattern π, let D_π denote the proportion of sites in the sequences where this pattern occurs. A *phylogenetic invariant* (or more briefly, an *invariant*) for a model M with parameters (T,P) is a function $f = f([X_\pi])$ in variables X_π ("indeterminants") that are indexed over all the possible patterns, π, and so that the following property holds for all P:

> If D is stochastically generated according to M, and f is evaluated with $X_\pi \equiv D_\pi$, then as the sequences become long, f tends to 0 with certainty (i.e., with probability 1).

More succinctly, we write:

$$f([D_\pi]) \to_p 0 \qquad (1)$$

where $\to_p$ denotes convergence in probability (see Rényi 1970). When the sequence site substitutions are independent then condition (1) is equivalent to the more usual formulation

$$f([E_\pi]) = 0 \qquad (1^*)$$

where E_π is the expected value of D_π for model M.

The important words in these definitions are the words "for all P"—that is, the truth of (1) or (1*) does not depend on the actual values the unknown edge parameters take, although the rate of convergence in (1) may. Often f is a polynomial function, though this is not necessary. The linear function:

$$f([X_\pi]) = \sum_\pi X_\pi - 1$$

is a noninformative invariant for all M, the **trivial invariant.**

If f satisfies (1) for all trees T, we say it is a *model invariant*, otherwise it is a *phylogenetically informative invariant.* Model invariants do not distinguish between trees, but are useful as a check on the correctness of the model M. Phylogenetically informative invariants can, in principle, place restrictions upon which trees adequately describe the data. Such invariants, taken together, may single out the correct tree by a process of elimination. Thus, we say that a collection of invariants for model M *identifies all trees* if each possible value for T^{-r} (i.e. the tree T without specifying the placement of the root) can be distinguished by those entries in the collection which satisfy (1) (or (1*)). Thus, for each unrooted tree T there is a set of invariants $I(T)$ so that if $T \neq T'$ then $I(T) \neq I(T')$. One can ask for more than this however—we will say a collection of phylogenetic invariants is *complete* if, for each unrooted tree T^{-r}, the invariants in $I(T^{-r})$ all satisfy (1*) and the remaining invariants are all positive if *and only if* M has parameters (T,P) for some P.

The following proposition is proved in Appendix 1.

Proposition 1 If a model has a collection of invariants which identify all three fully resolved (i.e. binary) trees on four taxa then

(1) for each fully resolved, unrooted tree T on $n \geq 4$ taxa, there exists an invariant $f_T([X_\pi])$ such that

$$f_T([D_\pi]) \to_p 0$$

if and only if T is the (unrooted) tree parameter in M.

(2) for $n \geq 4$ taxa there is a collection of $\binom{n}{3}$ invariants that identifies all trees.

TREES AS SET SYSTEMS

Normally a phylogenetic tree is thought of as a graph, that is, as a collection of vertices joined by edges. However, there is a natural way to represent a phylogenetic tree as a collection of subsets, and this representation is an essential ingredient in our construction of a complete collection of invariants. If we take an unrooted tree, T, whose leaves are labelled by the set $[n]$ (=$\{1,...,n\}$) and we delete an edge of T, this breaks the tree into two connected components and thereby partitions $[n]$ into a pair of sets; this pair is frequently referred to as a *split*. One of these two sets will *not* contain the last label n; if we select this set, and do this for all the edges of the tree T we obtain a collection, $\sigma = \sigma(T)$, of subsets of $[n-1]$ which have the following two properties:

(i) $[n-1] \in \sigma$ and $\{i\} \in \sigma$, for all $i \in [n-1]$.

(ii) if $\rho, \rho' \in \sigma$, then $\rho \cap \rho' \in \{\rho, \rho', \varnothing\}$.

Condition (ii) is often expressed by saying that ρ and ρ' are *compatible*. It is easily shown that $\sigma(T)$ has at most $2n-3$ sets, and this upper bound is achieved precisely if T is a binary (i.e. fully resolved) tree. For example, for the tree T^{-r}, in Fig. 1B, we have $\sigma(T^{-r}) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,2,3,4\}, \{1,3\}, \{2,4\}\}$.

Conversely, any collection, σ, of subsets of $[n-1]$ which satisfy (i) and (ii) corresponds to $\sigma(T)$ for a unique phylogenetic tree T on $[n]$. This fundamental result is due to Buneman (1971). Furthermore, T can easily and quickly be recovered from σ (e.g., by Meacham's TREE POPPING method; see Bandelt & Dress 1986). More generally, Buneman (1971) described how to construct a graph from any collection σ of sets, and showed that this graph is a tree precisely if the sets in σ are all pairwise compatible. For further details the interested reader should consult Barthélemy & Guénoche (1991).

KIMURA'S THREE-PARAMETER MODEL

Kimura (1981) defined a three substitution-type (3ST) model for nucleotide substitution, for which there is a rate α for "transition" type substitutions, and rates β and γ for two types of "transversion" substitution, as illustrated in Fig. 2A.

Following Kimura, we denote these three types of substitution by the letters P,R,Q, respectively. Let 0 denote the "null substitution", which fixes the four nucleotides. Then the collection $\{0,P,Q,R\}$ is closed under composition $\oplus$ (the effect of doing one substitution followed by another). For example, P followed by Q is the same substitution as R (regardless of which nucleotide it is applied to). In fact, the composition table of this collection, shown in Fig. 2B, forms a *group*—the so-called Klein 4-group, as first pointed out by Evans & Speed (1993). Ultimately, it is this property of the 3ST model which leads to the results described below. Note that Kimura's two-parameter model and the Jukes-Cantor model are both special cases of the 3ST model (put $\beta = \gamma$; $\alpha = \beta = \gamma$, respectively).

Now, consider a collection of taxa, labelled $1,...,n$. These taxa form the leaves (endpoints) of their evolutionary tree, T, which is a rooted tree, as in Fig. 1A. The root, r, of T represents the most recent common ancestor of the taxa. Consider the evolution of a single site on a segment of aligned DNA or RNA sequence from the ancestor r to its present observed forms. For this site the state—A, C, G, or T (U for RNA)—at the root will always be unknown. As the root sequence, and its descendant varieties (in the intermediate and extant species) evolve, substitutions of type P,Q,R will occur randomly with rates α,γ,β. We do not assume these parameters are constant over the evolutionary tree—they may vary freely from edge to edge (i.e. over time, and across different lines of descent). We assume that all the rates are strictly positive on any edge of the evolutionary tree.

The key assumption in the 3ST model is that these three types of substitutions take place independently in the tree—so that, for example, a transition at a site, in some intermediate species at some time in its evolution does not affect the probability of (say) an R-type transversion at that same site, either some time later in its evolution, or in another intermediate species in a different line of descent on the tree.

This assumption concerns a single site. The usual way to extend this to a collection of sites is to assume that:

(A) there is (statistical) independence across the sites of the sequence (as well as across the tree). This

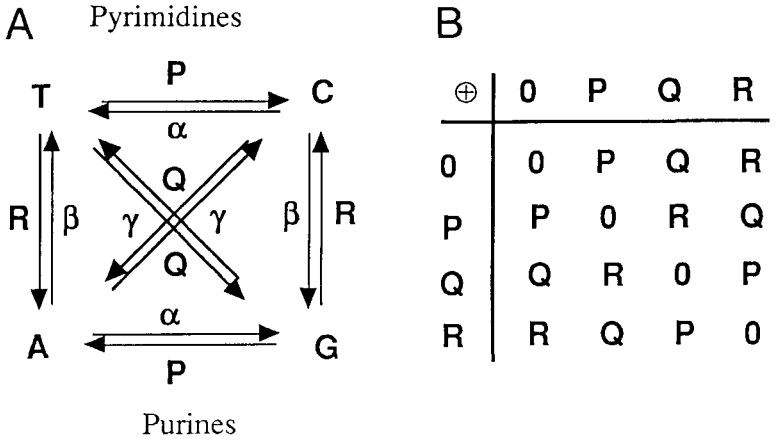

Fig. 2 **A**, The three types of substitutions (P,Q,R) and their rate parameters (α,γ,β) in Kimura's 3ST model. **B**, The substitutions, together with the "null substitution" 0, form a group (the Klein 4-group) under the operation of composition, denoted $\oplus$.

assumption implies, for example, that a transition at a site should not influence the likelihood of a transition occuring simultaneously, or subsequently, at a neighbouring site in some extant or ancestral taxon.

(B) the rates of the three types of substitution at any place in the tree is the same for all sites (these rates may, however, vary across the tree).

Statistically, regarding the sites as random variables, condition (A) states the sites are independently distributed, while (B) states that they are identically distributed; together this is commonly referred to as the "i.i.d." assumption.

Both the above assumptions seem particularly severe; for example, the second will not hold for sites subject to linkage or selection. However, it is not possible to say anything useful by completely abandoning assumptions (A) and (B); a tractable extension replaces conditions (A) and (B) by:

(A′) Each site influences only a fixed number of other sites—a precise statement is given later.

(B′) Each site can evolve at a different rate, but at any point in the tree, the ratios of the rate parameters α,β,γ at that point are the same across the sites (these ratios may vary across the tree).

A further aspect of the model concerns the distribution of the states at the root. In what follows we do not make any assumption whatsoever about this distribution (in particular, unlike the invariants described by Evans & Speed (1993), we do not need to assume that the distribution of the four states at the root is uniform) in other words, the model has *arbitrary root distribution.*

We show the i.i.d. assumptions ((A), (B)) lead to polynomial invariants, while the more general assumptions ((A′), (B′)) lead to analytical invariants.

INVARIANTS UNDER THE i.i.d. MODEL

We now construct a complete collection of polynomial invariants for the 3ST model with arbitrary root distribution, under the i.i.d. assumptions. Essentially we take certain linear combinations of the variables (the values $L_\theta([X_\pi])$ below), and our invariants are simply the difference of two terms, each of which is a product of these linear combinations. Furthermore, the linear combinations involved are such that each variable X_π occurs in each combination, with coefficient either +1 or −1, depending on the particular combination.

We begin with some definitions. A *quadripartition* (of $[n]$) is a pair $\theta = (\sigma_1,\sigma_2)$, where each σ_i is a subset of $[n-1]$. For quadripartitions $\theta = (\sigma_1,\sigma_2)$, and $\theta^* = (\mu_1,\mu_2)$, define

$$h(\theta^*,\theta) = \begin{cases} +1, \text{ if } |\sigma_1\cap\mu_1| + |\sigma_2\cap\mu_2| \text{ is even} \\ -1, \text{ if } |\sigma_1\cap\mu_1| + |\sigma_2\cap\mu_2| \text{ is odd.} \end{cases} \quad (2)$$

(where $|S|$ denotes the size of set S).

For a pattern $\pi = (S_1,\ldots,S_n)$ let $\theta(\pi)$ denote the quadripartition (σ_1,σ_2) where:

$\sigma_1 = \{i\colon S_n \to S_i \text{ is a R-type or Q-type substitution}\}$

$\sigma_2 = \{i\colon S_n \to S_i \text{ is a R-type or P-type substitution}\}$.

Thus, for example, σ_1 is precisely the set consisting of those taxa i whose state differs from taxon n by a transversion, as described in Fig. 2A. Note that precisely four patterns induce the same quadripartition.

For a quadripartition $\theta = (\sigma_1, \sigma_2)$, define the linear combination:

$$L_\theta([X_\pi]) = \sum_\pi h(\theta, \theta(\pi)) X_\pi \quad (3)$$

where the summation is over all patterns. Note that the coefficient of each X_π in $L_\theta([X_\pi])$ is either +1 or −1. For a quadripartition θ define the polynomial $f_\theta = f_\theta[X_\pi]$ by:

$$f_\theta = \prod_{\theta^*: h(\theta,\theta^*)=1} L_{\theta^*}([X_\pi]) - \prod_{\theta^*: h(\theta,\theta^*)=-1} L_{\theta^*}([X_\pi]) \quad (4)$$

These are our invariants. To state which ones apply for a particular tree, we need to relate quadripartitions to trees by the representation of trees as set systems. Firstly, recall that T^{-r} is the tree obtained from T by suppressing its root. The tree T^{-r} defines a special set of quadripartitions, which we now describe. Let

$\Delta = \{\rho^{(i)} : [n-1] \supseteq \rho \neq \varnothing\}$

where

$\rho^{(1)} = (\rho, \varnothing)$, $\rho^{(2)} = (\varnothing, \rho)$, $\rho^{(3)} = (\rho, \rho)$

and let

$C(T^{-r}) = \{(\varnothing,\varnothing)\} \cup \{\rho^{(i)}: \rho \in \sigma(T^{-r}), i=1,2,3\}$.

Note that $C(T^{-r})$ determines T^{-r} (since $\sigma(T^{-r})$ does), and has size three times the number of edges of T, so is therefore no larger than $3(2n-3)$, this bound being taken precisely if T is fully resolved. The following proposition, which provides the promised class of invariants for the 3ST model, follows from Theorem 10 of Székely et al. (1993).

Proposition 2 For the 3ST model with underlying tree T, and arbitrary root distribution, under the i.i.d. assumption, let E_π denote the expected value of D_π.

(1) $f_\theta([E_\pi]) = 0$ if and only if $\theta \notin C(T^{-r})$.

(2) $\{f_\theta : \theta \neq (\varnothing,\varnothing)\}$ is a complete collection of invariants, consisting of $3\times(2^{n-1}-1)$ phylogenetically informative invariants, namely $\{f_\theta: \theta \in \Delta\}$, and

$4^{n-1} - 3\times(2^{n-1}-1) - 1$ model invariants. Each unrooted tree t with e edges has

$4^{n-1} - 3e - 1$ associated invariants, namely $\{f_\theta : \theta \notin C(t)\}$.

EXTENSIONS TO NON-i.i.d. MODELS

In this section we adopt the following notation. If ψ is a function defined on real numbers, and $\mathbf{x}$ a real $n\times 1$ vector, then $\psi\mathbf{x}$ is the vector whose i-th component is $\psi(\mathbf{x}_i)$, for $i = 1,...,n$.

Define the linear form:

$$\chi_\theta([X_\pi]) = \sum_{\pi:\theta(\pi)=\theta} X_\pi \quad (5)$$

Indexing the quadripartitions, the $h(\theta,\theta')$ values of Equation (2) form a 4^{n-1} by 4^{n-1} symmetric Hadamard matrix, denoted $\mathbf{H}$, so that Equation (3) can be rewritten:

$L_\theta([X_\pi]) = (\mathbf{Hx})_\theta$

where $\mathbf{x} = [\mathbf{x}_\theta]$, and $\mathbf{x}_\theta = \chi_\theta([X_\pi])$ is given by Equation (5).

Let $e = [e_\theta]$, where $e_\theta = \chi_\theta([E_\pi])$, where E_π is, as usual, the expected value of D_π. Then it can be shown that all the components of $\mathbf{H}e$ are positive, so that, since $\mathbf{H}^{-1} = 4^{1-n}\mathbf{H}$, Proposition 2(1) can be restated as follows: under the 3ST model with i.i.d.

$(\mathbf{H}^{-1}\log\mathbf{H}e)_\theta = 0$ if and only if $\theta \notin C(T^{-r})$

which is the first half of Theorem 1 of Steel et al. (1992). The second half of that theorem states:

$(\mathbf{H}^{-1}\log\mathbf{H}e)_\theta = E_\rho^i$, if $\theta = \rho^{(i)} \in C(T^{-r})$

where E_ρ^i is the expected number of type-i substitutions on the edge of T^{-r} corresponding to ρ, where a type −1,2,3 substitution is a Q,P,R substitution, respectively. (There is a minor but important technical qualification required in case T has exactly two edges incident with its root, and ρ corresponds to the edge of T^{-r} which replaces those two edges. In this case, ρ corresponds to two edges of T, and then E_ρ^i is the total expected number of type-i substitutions on both these two edges. Note also that we are assuming throughout that $E_\rho^i > 0$ for all i and $\rho \in \sigma(T^{-r})$).

We wish to generalise these two results, thereby providing invariants under more general conditions than the i.i.d. assumption. We first present a more general version of the independence assumption, which is a limiting statement that, informally, says that non-independence arises only between a fixed number of sites (condition (A′)). Precisely, suppose the patterns in the sequences are ordered $\pi_1,...,\pi_c$ (this may be chosen to differ from the sequence ordering if the new ordering better reflects the interactions arising in three dimensions by having most interaction occurring between neighbours). Let

$$\delta(j,\theta) = \begin{cases} 1 \text{ if } \theta(\pi_j) = \theta \\ 0 \text{ otherwise} \end{cases}$$

Let ρ_{ij} denote the maximum (over all θ) correlation of $\delta(i,\theta)$ with $\delta(j,\theta)$. Then we can take, for condition (A′), the following statement:

$$(A')\rho_{ij} < \frac{1}{|i-j|} \qquad \text{for } i \neq j.$$

The numerator, 1, can be replaced by any constant, and there are other slightly different alternatives to (A′), however we do not pursue these here.

As for condition (B′), this states, informally, that the substitution process at different sites is essentially the same, except that it is proceeding at different speeds. More precisely, for the edge e of T corresponding to $\rho \in [n-1]$, and a site j on the aligned sequences, let $E_\rho{}^i(j)$ be the expected number of type-i substitutions on edge e. Then condition (B′) states:

(B′) $E_\rho{}^i(j)$ can be written in the form $E_\rho{}^i \times \lambda_j$.

Here λ_j can be thought of as the rate at which substitutions occur at site j, and $E_\rho{}^i$ is the average (over all the sites) of the expected number of type-i substitutions on the edge of T^{-r} corresponding to ρ, divided by the average value of the λ_j's.

Let $M(x)$ denote the limiting average value of the numbers $e^{x\lambda_j}$ (averaged over all sites j) as the sequences length c becomes large. That is, let

$$M(x) = \lim_{c\to\infty} c^{-1} \sum_j e^{x\lambda_j}$$

For example, if the rate parameters λ_j are drawn independently according to some distribution then $M(x)$ is the moment generating function of this distribution. Now, λ_j is positive for all j, and so $M(x)$ is monotone increasing, and therefore has a unique left functional inverse, which we denote as $\phi(x)$. That is, ϕ is the function for which $\phi(M(x)) = x$ for all real x. Then we have the following result, which is proved in Appendix 2.

Proposition 3 For the 3ST model with underlying tree T and arbitrary root distribution, under conditions (A′) and (B′) let $d = [d_\theta]$, $d_\theta = \chi_\theta([D_\pi])$. Then,

$$(\mathbf{H}^{-1}\phi\mathbf{H}d)_\theta \to_p \begin{cases} 0, & \text{if and only if } \theta \notin C(T^{-r}). \\ E_\rho{}^i, & \text{if } \theta = \rho^{(i)} \in C(T^{-r}) \end{cases}$$

EXAMPLES: (1) In case all the sites evolve at the same rate ($=\lambda$), we have $M(x) = e^{\lambda x}$, giving $\phi(x) = (1/\lambda)\log(x)$, and so Proposition 3 is just a special case of Proposition 2.

(2) Jin & Nei (1990) suggest that the gamma distribution

$$f(x) = \frac{e^{-\upsilon x} x^{k-1} \upsilon^k}{\Gamma(k)}, \quad x > 0,$$

may be appropriate for the λ_j's. In this case, $M(x) = (\nu/(\nu-x))^k$ so that $\phi(x) = \nu(1-x^{-1/k})$. Letting $\phi_k(x) = x^{-1/k}$, the invariants can then be written, independent of ν, as

$$(e_1 - \mathbf{H}^{-1}\phi_k\mathbf{H}d)_\theta \to_p 0, \text{ if and only if } \theta \notin C(T^{-r}),$$

where $e_1 = [1,0,0,0...]^t$.

ACKNOWLEDGMENT

The work of the first author (MAS) was supported by a grant from the New Zealand Lotteries Commission.

REFERENCES

Bandelt, H-J.; Dress, A. 1986: Reconstructing the shape of a tree from observed dissimilarity data. *Advances in applied mathematics 7*: 309–343.

Barthélemy, J.-P; Guénoche, A. 1991: Trees and proximity representations. England, John Wiley and Sons Ltd. Pp. 117–205.

Buneman, P. 1971: The recovery of trees from measures of dissimilarity. *In*: Hodson, F. R.; Kendall, D. G.; Tautu, P. *ed.* Mathematics in the archaeological and historical sciences. Edinburgh, Edinburgh University Press. Pp. 387–395.

Cavender, J. A. 1989: Mechanized derivation of linear invariants. *Molecular biology and evolution 6*: 301–316.

Cavender, J. A. 1991: Necessary conditions for the method of inferring phylogeny by linear invariants. *Mathematical biosciences 103*: 69–75.

Cavender, J. A.; Felsenstein, J. 1987: Invariants of phylogenies: simple case with discrete states. *Journal of classification 4*: 57–71.

Evans, S. N.; Speed, T. P. 1993: Invariants of some probability models used in phylogenetic inference. *Annals of statistics 21*: 355–377.

Ferretti, V.; Sankoff, D. 1993: The empirical discovery of phylogenetic invariants. *Advances in applied probability 25*: 290–302.

Fu, Y-X.; Li, W. H. 1991: Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree. *Mathematical biosciences 105*: 229–238.

Fu, Y-X.; Li, W. H. 1992: Construction of linear invariants in phylogenetic inference. *Mathematical biosciences 109*: 201–228.

Hendy, M. D.; Penny, D. 1993: Spectral analysis of phylogenetic data. *Journal of classification 10*: 1–20.

Jin, L.; Nei, M. 1990: Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular biology and evolution 7 (1)*: 82–102.

Kimura, M. 1981: Estimation of evolutionary sequences between homologous nucleotide sequences. *Procedings of the National Academy of Sciences U.S.A. 78*: 454–458.

Lake, J. A. 1987: A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular biology and evolution 4*: 167–191.

Navidi, W. C.; Churchill, G. A.; von Haeseler, A. 1991: Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Molecular biology and evolution 8 (1)*: 128–143.

Nguyen, T.; Speed, T. P. 1992: A derivation of all linear invariants for a non-balanced transversion model. *Journal of molecular evolution 35*: 60–76.

Rényi, A. 1970: Probability theory. Amsterdam, North Holland Publishing.

Steel, M. A. 1992: The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of classification 9*: 91–116.

Steel, M. A.; Hendy, M. D.; Székely, L. A.; Erdös, P. L. 1992: Spectral analysis and a closest tree method for genetic sequences. *Applied mathematics letters 5 (6)*: 63–67.

Székely, L. A.; Steel M. A.; Erdös, P. L. 1993: Fourier calculus on evolutionary trees. *Advances in applied mathematics 14*: 200–216.

Waddell, P.; Penny, D.; Hendy, M.; Arnold, G. in press: The sampling distributions and covariance matrix of phylogenetic spectra. *Molecular biology and evolution.*

APPENDIX 1: Proof of Proposition 1

(1) If T has four leaves, set

$$f_T([X_\pi]) = \sum_{f \in C} f([X_\pi])^2 \qquad \text{(A1)}$$

Then, since C identifies all trees on four leaves, $f_T([D_\pi]) \to_p 0$ if and only if T is the unrooted tree parameter in M. Now suppose T has $n > 4$ leaves. Applying Proposition 2(3) of Steel (1992), since T is fully resolved there exists a set Q of $n-3$ phylogenetic subtrees of T, each having four leaves of T, such that the trees in Q collectively define T. Thus, let

$$f_T([X_\pi]) = \sum_{t \in Q} f_t([X(t)_\rho]) \qquad \text{(A2)}$$

where, if t has leaves {i,j,k,l}, $X(t)_\rho$ is the sum of X_π over all patterns π which extend the pattern ρ on {i,j,k,l} and f_t is given by (A1) with $T=t$. In this way, $f_T([D_\pi]) \to_p 0$ if and only if T is the unrooted tree parameter in M.

(2) Each phylogenetic tree is defined by how it resolves all its subtrees on four leaves, containing a fixed leaf, say n (Bandelt & Dress 1986). Let this collection of subtrees be R. Then the collection of invariants $\{f_t([X(t)_\rho]): t \in R\}$, where $f_t([X(t)_\rho])$ is as defined in part (1), identifies all trees on n taxa.

APPENDIX 2: Proof of Proposition 3

Let $\gamma = [\gamma_\theta]$ where $\gamma_\theta = E_\rho{}^i$, if $\theta = \rho^{(i)} \in C(T^{-r})$, while $\gamma_\theta = 0$ otherwise except for $\gamma_{(\emptyset,\emptyset)}$ which is chosen so that the components of γ sum to 0. Then the expected value of $\delta(j,\theta)$, denoted $<\delta(j,\theta)>$, is given by Theorem 1 of Steel et al. (1992) as

$$<\delta(j,\theta)> = (\mathbf{H}^{-1}\exp\mathbf{H}(\lambda_j\gamma))_\theta$$

Thus, $$c^{-1}\sum_j <\delta(j,\theta)> = c^{-1}\sum_j (\mathbf{H}^{-1}\exp\mathbf{H}(\lambda_j\gamma))_\theta$$

$$= \left(\mathbf{H}^{-1}c^{-1}\sum_j \exp\mathbf{H}(\lambda_j\gamma)\right)_\theta$$

Now, $$\lim_{c\to\infty} c^{-1}\sum_j (\exp\mathbf{H}(\lambda_j\gamma))_{\theta'} = \lim_{c\to\infty} c^{-1}\sum_j (\exp\lambda_j(\mathbf{H}\gamma))_{\theta'} = M((\mathbf{H}\gamma)_{\theta'})$$

Thus, $$\lim_{c\to\infty} c^{-1}\sum_j <\delta(j,\theta)> = (\mathbf{H}^{-1}M\mathbf{H}\gamma)_\theta,$$

Applying Bernstein's Theorem (see Rényi 1970, p. 379), we deduce that, for each θ,

$$d_\theta \to_p \lim_{c\to\infty} c^{-1}\sum_j <\delta(j,\theta)> = (\mathbf{H}^{-1}M\mathbf{H}\gamma)_\theta,$$

thus, $$d \to_p \mathbf{H}^{-1}M\mathbf{H}\gamma \qquad \text{(A3)}$$

Now, since ϕ is continuous, so too is $\mathbf{H}^{-1}\phi\mathbf{H}$, and if we have (for general random vectors) $Z \to_p z$, then it is easily checked that $\psi(Z) \to_p \psi(z)$ for any continuous function ψ. Applying this to (A3) we deduce that $\mathbf{H}^{-1}\phi\mathbf{H}d \to_p \mathbf{H}^{-1}\phi\mathbf{H}\mathbf{H}^{-1}M\mathbf{H}\gamma = \gamma$, as required.