Reconstructing Trees When Sequence Sites Evolve at Variable Rates

M.A. STEEL,¹ L.A. SZÉKELY,² and M.D. HENDY³

ABSTRACT

For a sequence of colors independently evolving on a tree under a simple Markov model, we consider conditions under which the tree can be uniquely recovered from the "sequence spectrum"—the expected frequencies of the various leaf colorations. This is relevant for phylogenetic analysis (where colors represent nucleotides or amino acids; leaves represent extant taxa) as the sequence spectrum is estimated directly from a collection of aligned sequences. Allowing the rate of the evolutionary process to vary across sites is an important extension over most previous studies—we show that, given suitable restrictions on the rate distribution, the true tree (up to the placement of its root) is uniquely identified by its sequence spectrum. However, if the rate distribution is unknown and arbitrary, then, for simple models, it is possible for every tree to produce the same sequence spectrum. Hence there is a logical barrier to accurate, consistent phylogenetic inference for these models when assumptions about the rate distribution are not made. This result exploits a novel theorem on the action of polynomials with non-negative coefficients on sequences.

INTRODUCTION

The question of how best to reconstruct evolutionary trees is both controversial and challenging. On the one hand, different methods and/or different data often give rise to different trees, leading to arguments over whose method (or data!) was "correct." The challenge arises because, unlike other branches of theoretical biology (for instance, population genetics), one is primarily estimating quantities that cannot (even in principle) be observed or measured directly. Thus, one relies on an underlying theory of how the estimated quantity (the tree) is related to observable quantities (genetic sequences, fossils, morphological/behavioral/biochemical evidence). A major problem is that there is much uncertainty as to the exact nature of this link, and which assumptions are justified in any underlying stochastic model.

The simplest and earliest approaches to tree reconstruction were direct methods that were not based on any stochastic model. For sequences, the maximum parsimony method and related compatibility methods have been used; while, for pairwise distance measures, several methods were devised to recover a tree under the assumptions that the distances correspond to the lengths of paths in the (edge-weighted) tree. These methods are still widely used today on (uncorrected) sequences and distance measures (for a recent survey, see Felsenstein, 1988), despite the fact that since the late 1970s it has been known that these methods could lead to incorrect trees, under simple models of sequence evolution (Felsenstein, 1978).

¹Mathematics and Statistics Department, University of Canterbury, Christchurch, NZ.
²Department of Computer Science, Eötvös University, Budapest, Hungary.
³Mathematics Department, Massey University, Palmerston North, NZ.
This inconsistency motivated the development of three types of statistically-based methods: (i) maximum likelihood (Felsenstein, 1981), (ii) phylogenetic invariants (Cavender and Felsenstein 1987; Lake, 1987), and (iii) correctional transformations to the data.

This third category outputs either new sequences or distances that can then be fed into the simpler methods described earlier, without leading to inconsistencies of the type described by Felsenstein. These transformations include spectral analysis (Hendy, 1989; Steel et al. 1992), LogDet/paralinear transformations (Lake, 1994; Lockhart et al., 1994), and the related, but more restrictive, transformation described by Rodriguez et al. (1990).

One problem with all three classes of methods is that, at present, they are based on models that are too restrictive to describe the underlying biology adequately.

In particular, there are two types of assumptions that are problematic: first the imposition of unrealistic constraints on the stochastic model of site mutations, for instance, that it is governed by a reversible and/or stationary Markov-style model. Such restrictions cannot easily explain how marked variation in nucleotide frequencies (for instance “GC richness”) evolved between different sequences. A second type of assumption concerns the way in which site mutations translate into sequence evolution. Generally, it is assumed that sites evolve independently and identically (the i.i.d. assumption), but this is also usually unrealistic.

The first type of assumption is easier to relax, because it has recently been shown (Steel, 1993) that under very general conditions (but still retaining the i.i.d. assumption) the expected distribution of patterns appearing at sites in the sequences defines the tree. Thus, methods like maximum likelihood will identify the correct tree, given sufficiently long sequences. However under this general model, the location of the root (ancestral taxon) cannot be determined; we prove this here in Theorem 2, thereby extending an earlier result due to Felsenstein (1981).

In this paper, we are mostly concerned with relaxations of the i.i.d. assumption, in particular with allowing sites to evolve at different rates. This is the simplest relaxation over the rigid i.i.d. assumption, and appears to be biologically relevant. Yang (1993) has taken a useful step in this direction by showing how to modify the maximum likelihood method to allow a gamma distribution of rates across sites, under a particular underlying model on four taxa.

However, as in any area of modeling, making a model more flexible (by allowing more parameters) generally results in less predictive power. In modeling sequence evolution, it is easy to devise stochastic models that are so general that they give no hint of the underlying tree from the observed sequences. Thus, it is important to identify the boundary between having enough structure (to find the tree) and too much flexibility (and consequent loss of the tree) when sequence sites are allowed to evolve at different rates.

In Theorem (3)1(i) we show that in the case that sites evolve independently, but with varying rates, then the tree can still be uniquely recovered for simple symmetric models, provided the distribution of rates is known. This raises the question of whether uniqueness also holds when no assumption is made regarding the distribution of rates across sites.

In certain cases this is so; Lake (1987) showed that a certain model, which has linear phylogenetic invariants (tree-related linear equations between the expected frequencies of the patterns), allows the tree to be recovered even under (unknown) site-to-site variation of rates. However, the existence of linear invariants seems to require special properties for the model, and without them the reconstruction story changes dramatically.

We show here in Theorem 3(2) that starting with even the simplest two-state model (the Cavender–Farris model, which allows different but symmetric transition matrices for each edge), and modifying it to allow a variation of rates across sites can lead to a complete inability to recover the underlying tree (by any method), since every tree can induce the same probability distribution on the sequence patterns. Consequently, no tree reconstruction method can consistently recover trees under this model without regard to this rate-across-sites distribution. This result extends to four-state models, such as the Kimura’s 3ST model, as this contains the symmetric two-state model as a submodel. Thus, if the evolutionary tree is always to be uniquely recoverable under models that do not possess linear invariants, some restrictions on the rate-across-sites distribution need to be imposed, perhaps given further knowledge of the causes of site heterogeneity (be they selection, local interaction, or the covarion hypothesis).

For example, a simple restriction is that an unknown number of sites are invariant, that is, have a zero rate of evolution, while the remaining sites have an (unknown) identical rate of evolution. We show in Theorem 3(1)(ii) that under this restriction, applied to our model, the tree is uniquely defined by the expected distribution of the patterns in the sequences. Alternatively, if a molecular clock is imposed, then uniqueness also holds under this model (Theorem 3(1)(iii)). We suggest these as first steps toward further determining the conditions, under which this model, and more general models, can always allow the tree to be recovered.

The proof that uniqueness is lost without these constraints relies on a novel and apparently new result for
identifying the components of monotonic sequences by polynomials possessing nonnegative coefficients. We state this result here, and give a proof, due to one of us (L.A.S.), in the Appendix. Here and later, we adopt the following convention: given a vector $x$, and a function $f: \mathbb{R} \rightarrow \mathbb{R}$ we let $f(x)$ denote the vector whose $i$th component is $f(x_i)$.

**Theorem 1** For any $k$ vectors in $\mathbb{R}^n$, $x_1, \ldots, x_k$ with $0 < (x_1)_j < (x_k)_j < 1$ for $i = 1, 2, \ldots, k$, and $j = 1, 2, \ldots, n - 1$, there exist nonconstant polynomials $p_i: i = 1, 2, \ldots, k$, each having nonnegative coefficients, which sum to 1, and such that $p_1(x_1) = p_2(x_2) = \ldots = p_k(x_k)$, as vectors.

**EVOLOVUTION OF SITES**

Evolutionary relationships between extant taxa are generally represented by a leaf-labeled phylogenetic tree. Such a tree, denoted $T^\rho$, has leaves, labeled $1, \ldots, n$, which correspond to the extant taxa, and a labeled root vertex $\rho$ (of degree at least 2) representing the global ancestral taxon. The remaining vertices are unlabeled and correspond to intermediate ancestral taxa. We let $T$ denote the unique unrooted (phylogenetic) tree, obtained from $T^\rho$ by unlabeling $\rho$, and deleting any vertices of degree two (their incident edges being identified), as illustrated in Fig. 1.

To analyze the evolution of aligned sequences, it is useful to consider first the evolution of a single site in those sequences (we return to sequences again in the next section). Substitutions (point mutations) at a site are generally modeled by a probability distribution $\pi$ on a set of $r \geq 1$ colors (states) at the root $\rho$ of $T^\rho$, together with an $r \times r$ transition matrix $M_e$ for each edge $e$ of $T$ (see Fig. 1). The colors are the character states under consideration, so that $r = 4$ (or 2) for genetic sequences, $r = 20$ for amino acid sequences. The (random) color at the root "evolves" down the tree, thereby assigning colors randomly to the vertices, from the root down to the leaves. For each edge $e = (i,j)$, with $i$ between $j$ and the root, $(M_e)_{ij}$ is the probability that $j$ is coloured $\beta$ given that $i$ was colored $\alpha$. It is assumed that the random assignment of a color to a vertex $v$ is dependent only on the colors of its immediate ancestor.

Under this model, each coloration $\chi$ of the leaves of $T$ has a well-defined probability, which we denote by $f_\chi(T^\rho, P)$, where $P: E(T^\rho) \cup \{\rho\} \rightarrow \mathbb{R}^\alpha$, is defined by $P(e) = M_e$ for each edge $e \in E(T^\rho)$, the set of edges of $T^\rho$, and $P(\rho) = \text{diag}[\pi]$, the diagonal matrix whose $jj$ entry is $\pi_j$. Ordering the leaf colorations $\chi$, the $f_\chi(T^\rho, P)$ form a vector that we write as $f(T^\rho, P)$. In Steel (1993) it is shown that, under the general model, which assumes only:

$$\det(M_e) \in \{0, 1, -1\} \text{ for all } e \in E(T^\rho); \pi_\alpha \neq 0 \text{ for all colors } \alpha$$

(1)

the vector $f(T^\rho, P)$ is sufficient to uniquely recover $T$ (and in polynomial-time), since the matrix-based quantities:

$$\phi_{xy} := -\ln |\det F_{xy}|$$

where $(F_{xy})_{\alpha\beta} = \text{Prob}[\chi(x) = \alpha \& \chi(y) = \beta]$ defined for each pair $x, y \in \{1, \ldots, n\}$, $x \neq y$, satisfy the four-point condition on $T$ (see, for instance, Bandelt and Dress, 1986).

However, the location of $\rho$ (in forming $T^\rho$) can never be determined under the general model (without imposing additional assumptions) as we now demonstrate. This is the analogue (for the general model) of an earlier nonuniqueness result for a more restrictive (reversible) model, due to Felsenstein (1981).

**FIG. 1.** A phylogenetic tree $T^\rho$ on the taxa set 1,2,3,4,5, together with a distribution $\pi$ of colors (states) at the root vertex (ancestral taxon) $\rho$. Each edge $e$ of $T^\rho$ has an associated transition matrix $M_e$. The unrooted tree $T$ is obtained from $T^\rho$ by deleting $\rho$, and identifying its two incident edges.
Theorem 2. [Nonlocatability of p from \( f(T^{op}, P) \) under the general model (1)].
Let \( T \) be an unrooted phylogenetic tree with vertices \( p_1, p_2 \) of degree \( \geq 2 \), and let \( T_1, T_2 \) be the trees obtained by rooting \( T \) at \( p_1 \) and at \( p_2 \). Then for any \( P_1 \) satisfying (1) for \( T_1 \), there exists a \( P_2 \) satisfying (1) for \( T_2 \) such that \( f(T_1, P_1) = f(T_2, P_2) \).

Proof: (We show first that this result holds if \( p_1, p_2 \) are adjacent in \( T \), then by transitivity it can be extended to any pair of vertices in \( T \).)
Suppose \( p_1 \) and \( p_2 \) are adjacent in \( T \) with \( e = (p_1, p_2) \). Given \( P_1 \) on \( T_1 \) let \( \pi_1, \pi_2 \) be the distributions at \( p_1 \) and \( p_2 \), then with \( M_e = P_1(e) \), \( \pi_2 = \pi_1 M_e \). As \( M_e \) has nonnegative components and no column consisting entirely of zeros (since \( \det(M_e) \neq 0 \) and since \( \pi_1 > 0 \), then \( \pi_2 > 0 \). Now,

\[
f_2(T_1, P) = \sum_{x,y} \text{Prob}[\chi(x(p_1) = x & \chi(p_2) = y) \times \text{Prob}[\chi(x(p_1) = x & \chi(p_2) = y]}
\]

where the summation is over all pairs of colors, \( x, y \).

\[
\text{Prob}[\chi(x(p_1) = x & \chi(p_2) = y] = (M_e)_{x,y}(\pi_1) \]

and so, if we define a matrix \( M'_{e} \) as follows:

\[
(M'_{e})_{x,y} = (M_e)_{x,y}(\pi_1) \]

then \( M'_{e} \) is a transition matrix with \( \det(M'_{e}) \neq (0, 1, -1) \).

Let \( P_2 \) be the function which assigns \( \text{diag}[\pi_2] \) to \( p_2 \), \( M'_{e} \) to edge \( e \) and assigns the same transition matrices as \( P_1 \) to the other edges of \( T \). Note, if we consider the Markov chain proceeding from \( p_2 \) to \( p_1 \), the joint probability that \( \chi(p_1) = x \) & \( \chi(p_2) = y \) is: \( (M'_{e})_{x,y}(\pi_2) = (M_e)_{x,y}(\pi_1) \), as before, and so we can apply (2) to deduce that:

\[
f_2(T_2, P_2) = f_2(T_1, P_1) \quad \text{for all } \chi.
\]

The result can now be extended by induction using sequences of adjacent rerootings, for \( T \) being rooted at two nonadjacent roots.

For the remainder of this paper, we consider, for simplicity, that evolution at a single site is described by the simplest two-color model, the Cavender–Farris model (Farris, 1973; Cavender, 1978) which assumes each matrix \( M_e \) is symmetric. However, many of the results apply a fortiori to more general models, and to the case \( r = 4 \) (under, say, the Kimura 3ST model); thus, the restriction is in no way serious. We let \( p_e \) denote throughout the off-diagonal entry in \( M_e \).

Even for the Cavender–Farris model, the vector \( f(T^{op}, P) \) does not determine the position of the root in \( T^{op} \) if the distribution of colors at a root is uniform (although the root can always be located if the molecular clock hypothesis is assumed, or the distribution of colors at the root is not uniform (Steel et al., 1994). However, from the above discussion, we can always recover \( T \) from \( f(T^{op}, P) \) provided \( M_e \notin (0, 1, -1) \) that is, provided \( p_e \notin (0, 0.5) \).

The stochastic mechanism generating a color change between the ends of an edge is often taken to be a continuous-time Markov process, with rate \( \lambda_e > 0 \). If \( t_e > 0 \) denotes the time for which such a process operates for edge \( e \), and if \( q_e \) denotes the expected number of color changes associated with \( e \), we have that

\[
q_e = \lambda_e t_e
\]

and, under such a process, it is easily shown (Hendy 1989) that:

\[
p_e = 0.5(1 - \exp(-2q_e)).
\]

Thus, a further restriction in the Cavender–Farris model, is that

\[
0 < p_e < 0.5, \quad \text{for all } e \in E(T^{op}).
\]

The molecular clock hypothesis states that \( \lambda_e > 0 \) is a constant across edges (unless stated otherwise, we do not assume this here). Since the \( t_e \) values correspond to time, it is implicit that the sum of the \( t_e \) values from the root to any leaf is the same. Thus, the molecular clock hypothesis is equivalent to requiring that the expected number of color changes on the path in \( T \) from the root to a leaf is the same for each leaf.
EVOLUTION OF SEQUENCES

The central problem in phylogenetic analysis is how to reconstruct, from aligned r-state sequences, the underlying evolutionary tree, and perhaps also to provide information about the times between branchings (the \( t_e \) values).

The models described in the previous section concern the evolution of a color at a single site in a collection of aligned sequences. There are a number of ways to extend this to a model describing the evolution of the entire frame of aligned sequences. The simplest is to suppose that each site evolves identically and independently (the i.i.d. assumption), and this was Cavender’s original proposal for his model (Cavender, 1978). In this case, by the law of large numbers, the proportion of sites that correspond to a pattern \( \chi \) converges to \( f_\chi(T^{10},P) \) with probability 1, as the number of sites grows. Thus, under the i.i.d. assumption the inversion problem in phylogenetic analysis is asymptotically equivalent to the problem of reconstructing \( T \) from \( f(T^{10},P) \). Thus, sufficiently long sequences (dependent on \( P \)) determine \( T \) (assuming (3)).

However it is well known (Jin and Nei, 1990; Reeves, 1990) that the i.i.d. assumption is invalid for many sequences; in particular, the assumption of an identical process at each site is often unrealistic, with certain sites and regions apparently evolving faster (“hot spots”), while other regions are more conserved, and some sites may or may not be able to change color at all. Thus, a more realistic model would allow, for each site \( i \), the associated rate parameter \( \lambda_i \) to be multiplied by a factor \( \mu_i \geq 0 \). In this case, considering site \( i \) (for which the standard Cavender–Farris model applies), the expected number of mutations on edge \( e \), which we denote as \( q_e(\mu_i) \), equals \( \mu_i \lambda_i \) if we take the values \( \lambda_i \) to be constant across sites, let us call this value \( q_e \), as in the standard Cavender–Farris model. Thus, we have:

\[
q_e(\mu_i) = \mu_i q_e
\]

Note that we do not impose any additive constraint on the \( \mu_i \) values, for instance, they need not sum to 1.

There are three ways to describe \( \mu_i \): (i) as a well-defined, but unknown number; (ii) as randomly and independently selected from a distribution \( \varnothing \) which is constant over all \( i \); (iii) as randomly and independently selected from a distribution that varies with the site \( i \). Note that (i) and (ii) are both special cases of (iii). We will call (ii) the generalized Cavender–Farris model (the preference for (ii) over (i) has little consequence for the questions we consider—for example, it can be shown that an analogous version of Theorem 3(2), below, holds under description (i) of the \( \mu_i \) values, but the details are slightly more involved, and we omit them here—for more details see the remark at the end of the Appendix).

Under this model, let \( f_\chi = f_\chi(T^{10},P,\varnothing) \) denote the probability of generating at any site the pattern \( \chi \) (equivalently, this is the expected proportion of sites in the sequence for which pattern \( \chi \) occurs). We refer to the association \( \chi \rightarrow f_\chi \) as the sequence spectrum. We assume throughout that \( \varnothing \) does not assign \( \mu_i = 0 \) with probability 1.

**Theorem 3** Assume the generalized Cavender–Farris model.

1. The tree \( T \) is determined from its sequence spectrum if either:
   (i) \( \varnothing \) is known,
   (ii) \( \varnothing \) is unknown, but it has positive measure only on 0 and one other (unknown) value,
   (iii) \( \varnothing \) is neither known nor constrained, but we assume the molecular clock hypothesis (i.e., \( \lambda_e = \text{constant} \)).

2. If none of the conditions (i)–(iii) hold, then \( T \) may no longer be determined by its sequence spectrum—indeed each tree, with an associated \( \varnothing = 0 \), can induce an identical sequence spectrum.

**Remarks:** Part (1) remains true for four-state sequences under the generalized Kimura 3ST model where sites can evolve at varying rates. We leave the proof to the reader. In condition (i) of part (1) we need to know only the moment generating function of the distribution \( \varnothing \) defined on the negative real line (note that this function always exists over this domain for any distribution \( \varnothing \)). Part (1)(ii) models the situation where an unknown set of sites is unable to change, while the remaining sites evolve independently and identically. Note that part (2) would be trivial if we allowed \( \lambda_e = 0 \) on edges not incident with leaves, indeed, inserting or contracting such edges does not change the sequence spectra (Székely et al., 1993). Note also that in case (2) we are not allowing an unconstrained and arbitrary process (i.e., free choice of transition matrices \( M_e \)) at each site, since we are insisting that the ratio of edge lengths (i.e., the ratio of the \( q_e(\mu_i) \) values for pairs of edges) is the same for each site \( i \). In the more general model where the process can vary between sites, it is interesting that the maximum
likelihood tree(s) coincide exactly with the maximum parsimony tree(s) (see Penny et al., 1994). Finally, we note that, if, instead of the generalized Cavender–Farris (or generalized Kimura 3ST model), we were to consider a model possessing linear phylogenetic invariants (as in Lake, 1987), and these invariants were sufficient to distinguish between trees, then (2) would no longer hold—for these models variation of rates between sites is not a theoretical problem for tree reconstruction. Unfortunately, such models tend to be quite special.

Proof: Actually for part (1) we do not require complete knowledge of the sequence spectrum, just the expected frequencies of the "essentially different" patterns. Thus, suppose $T$ has leaf set $\{1, \ldots, n\}$ and, for a subset $\sigma$ of $\{1, \ldots, n-1\}$, let $s_\sigma$ be the probability of generating, under the generalized Cavender–Farris model, either of the two colorations for which $\sigma$ is the set of leaves which are colored differently to leaf $n$. Thus, $s_\sigma$ is a sum of two $f_\sigma$ values. We show that the collection $(s_\sigma)$ determines $T$ given conditions (i) or (ii). The proof relies on a useful description of $(s_\sigma)$, derived for the case $\mu = \text{constant}$, by Hendy (1989) (for the Cavender–Farris model) and by Steel et al., (1992) (for the Kimura 3ST model) and extended to the general case by Steel et al., (1993, 1994). Specifically, let us order the subsets of $\{1, \ldots, n-1\}$ so that the $s_\sigma$ form a vector, $s$. For an edge $e$ of $T$, let $s_e$ denote the subset of leaves that become disconnected from leaf $n$ when edge $e$ is deleted from $T$ (so $s_e \subseteq \{1, \ldots, n-1\}$), and let $\sigma(T) = \{s_e : e \in E(T)\}$. Note that $T$ can be uniquely reconstructed from $\sigma(T)$, and in time which is linear in $n$ (see, for instance, Gusfield, 1991). For $e \in E(T)$, let:

$$
\gamma_e = \begin{cases} 
q_e, & \text{if the root } \rho \text{ of } T^{op} \text{ has degree } 2, \\
q_e + q_e, & \text{if } e \text{ is the edge of } T \text{ subdivided to create } \rho, \\
0, & \text{if } \sigma = \sigma_e \text{ and both } e_1, e_2 \text{ are edges of } T^{op} \text{ incident with } \rho.
\end{cases}
$$

Let $Y_\sigma = \sum_{e \in E(T)} \gamma_e$ if $\sigma = \sigma_e$.

Then, from Steel et al., (1993, 1994), in the generalized Cavender–Farris model, $s = H^{-1}M(J_1)$, where $H$ is the $2^{n-1} \times 2^{n-1}$ Hadamard matrix $H = ((-1)^{(\sigma, \sigma')})$ (where $\sigma, \sigma' \subseteq \{1, \ldots, n-1\}$), while $M(x)$ is the moment generating function for $\sigma$ (applied componentwise to $H$) defined over the restricted domain $x \in (-\infty, 0]$ (Rényi, 1970). Note that, since $H$ is symmetric, $H^{-1} = 2^{1-n} H$. Thus, $\gamma = H^{-1}\phi(Hs)$, where $\phi$ is the functional (left) inverse of $M$, which exists since, over its restricted domain $(-\infty, 0)$, $M$ always exists, and is monotonically increasing. Now, $\sigma(T)$ and hence $T$ is determined by $\gamma$, since $\sigma(T) = \{\sigma : Y_\sigma > 0\}$, and so this establishes part (1), case (i).

For the remainder of the proof we need to introduce an alternative description of $H$, due, originally, to Hendy (1989). For a subset $X$ of $\{1, \ldots, n\}$ of even cardinality, let $P(T, X)$ denote the unique set of edges of $T$ that exists in any collection of edge-disjoint paths of $T$ that connect pairs of leaves from $X$. Note that $P(T, X)$ is well defined, even though, for nonbinary trees, there may be more than one matching of $X$ leading to edge-disjoint paths. Order the even cardinality subsets of $X$ as follows: we already have an ordering on the subsets $\sigma$ of $\{1, \ldots, n-1\}$ so that:

$$
X_\sigma = \begin{cases} 
\sigma, & \text{if } |\sigma| = 0 \text{ (mod 2)}, \\
\sigma \cup \{n\} & \text{if } |\sigma| = 1 \text{ (mod 2)}
\end{cases}
$$

Then, Lemma 5 of Steel et al. (1994), states for the Cavender–Farris model that:

$$
(H\gamma)_\sigma = -2 \sum_{e \in P(T, X_\sigma)} q_e.
$$

Regarding case (ii), we first note that it suffices to establish the claim in the case that $T$ is a tree on four leaves, since once this is established, the result extends to all $T$. This is because every leaf-labeled (unrooted) phylogenetic tree is characterized by the phylogenetic subtree it induces on each subset of four leaves (Bandelt and Dress, 1986), and by assumption each of these would be uniquely defined by considering the marginal sequence spectrum for that subset of leaves.

Thus, suppose two trees on four leaves (with their associated distributions $\phi$ satisfying the conditions of part(ii)) induce the same sequence spectra, and hence the same $s$ vector. Let $T_1$ and $T_2$ denote, respectively, the two unrooted trees obtained from the original trees by deleting the root (and any vertices of degree 2). As de-
scribed above, the common s vector derived from either of the two parent trees is a function of a vector \( \gamma_1, \gamma_2 \) defined on the edges on \( T_1, T_2 \), respectively. Thus, suppose \( T_1 \) and \( T_2 \) are different (we will derive a contradiction). We have, \( H^{-1} M_1(H_T) = H^{-1} M_2(H_T) \), where \( M_j(x) \) is the moment generating function for the distribution associated with \( T_j \). Thus, we have the vector equality:

\[
M_1(H_T) = M_2(H_T).
\]

For the conditions on the rate distribution prescribed by condition (ii) we have, for \( j = 1, 2 \), that \( M_j(x) = 1 - \alpha_j + \alpha_j \exp(\mu_j x) \), for unknown \( \alpha_j \in (0, 1), \mu_j > 0 \). Thus, letting \( r' \) and \( r \) denote, respectively, the vectors: \( \exp(\mu_1 H_T) \) and \( \exp(\mu_2 H_T) \), we have:

\[
r' = \beta r + 1 - \beta \text{, where } \beta = \alpha_2 / \alpha_1.
\]

Since \( T_1 \) and \( T_2 \) are different, at least one of them, say \( T_2 \), is fully resolved, that is, has an edge which separates a pair of leaves from leaf 4 Without loss of generality, we may suppose that these two leaves are 1 and 3, that is, \( \{1, 3\} \in \sigma(T_2) \).

Now, regarding \( T_1 \), since this tree differs from \( T_2 \) there is a leaf \( j \neq 3 \) such that the path connecting leaves 1 and \( j \) is edge disjoint from the path connecting the remaining two leaves. Without loss of generality we may assume that \( j = 2 \). Thus, from (4) we have

\[
r''_{[1, 2, 3]} = r'_{[1, 2]} r''_{[3]}
\]

From (5),

\[
\beta (r_{[1, 2]} r''_{[3]} - r_{[1, 2]} - r_{[3]} + 1) + (r_{[1, 2]} - r_{[1, 3]} + 1) = 0.
\]

Now, \( \beta \neq 0 \), so we have:

\[
\beta = \frac{(1 + r_{[1, 2]} - r_{[1, 2]} - r_{[3]})}{(1 - r_{[1, 2]})(1 - r_{[3]})}
\]

Now, \( \{1, 3\} \in \sigma(T_2) \), so, from (4), \( r_{[1, 2, 3]} > r_{[1, 2]} r''_{[3]} \), and thus the numerator of (6) exceeds the denominator, which is positive since \( r_{[1, 2]}, r_{[3]} < 1 \). Thus, \( \beta > 1 \). Now, from (5), \( r = \beta r' + 1 - \beta' \), where \( \beta' = \beta^{-1} \), so repeating an analogous argument, starting with the identity:

\[
r_{[1, 2, 3]} = r_{[1, 3]} r''_{[2]}
\]

and applying (from 4) the inequality

\[
r_{[1, 2, 3]} \geq r'_{[1, 2]} r''_{[2]}
\]

we would deduce that \( \beta' \geq 1 \). But since \( \beta' = \beta^{-1} \), this gives \( \beta \leq 1 \), the required contradiction.

Regarding part (iii), we show that not only \( T \) but \( T^{op} \) is determined from the sequence spectrum. It suffices to establish this stronger claim for all rooted trees with just 3 leaves (by an argument analogous to that given for the proof of (ii)).

Thus, suppose \( s(T_1) = s(T_2) \) for \( T_1 \neq T_2 \), being two distinct rooted trees on leaf set \( \{1, 2, 3\} \). As before this would imply:

\[
M_1(H_T) = M_2(H_T)
\]

for \( \gamma_1, \gamma_2 \) derived from \( T_1, T_2 \), respectively. Since \( M_1 \) and \( M_2 \) are strictly monotone increasing, it follows that \( H_T \) and \( H_T \) are ordered equivalently (we say two vectors \( x \) and \( y \) are ordered equivalently provided \( x_1 < x_2 \Leftrightarrow y_1 < y_2 \)). Now, suppose in \( T_1 \) leaf 1 is adjacent to the root, but leaves 2 and 3 are not. Then, from (4), we have:

\[
(H_T)_{[1]} = -2 \sum_{e \in P(T_1, 1, 3)} q_e
\]

\[
(H_T)_{[1, 2]} = -2 \sum_{e \in P(T_1, 1, 2)} q_e
\]

\[
(H_T)_{[2]} = -2 \sum_{e \in P(T_1, 2, 3)} q_e
\]
Thus, by the molecular clock hypothesis: $(H_{T_1})_{121} > (H_{T_2})_{111} = (H_{T_2})_{121}$. Similarly, since $T_2 \neq T_1$, we may suppose that leaf 2 is adjacent to the root of $T_2$. Then $(H_{T_2})_{121} = (H_{T_2})_{121}$. But this implies that $H_{T_1}$ and $H_{T_2}$ are not ordered equivalently, a contradiction.

Regarding part (2), we claim that every tree, leaf labeled by $\{1, \ldots, n\}$, has an associated positive edge weighting, such that the vectors $H_T = H(T)$ have no tied entries and are equivalently ordered. To construct such a family of edge weightings, one for each tree, let

$$q_e = \lambda_{i,j,e} = \begin{cases} \frac{1}{n} & \text{if } e \text{ is not incident with a leaf} \\ 2 & \text{if } e \text{ is incident with leaf } i = 1, \ldots, n \end{cases}$$

and then construct $H(T)$ by applying equation (4). Note that, by equation (4),

$$(H(T))_{ij} = -2 \left( \sum_{e \in \partial_i} 2^e + c_e(T) \right)$$

where $c_e(T)$ is some number in the interval $[0,1]$.

Thus, the vectors $H(T)$ are equivalently ordered, and so the vectors $\exp(H(T))$, are also equivalently ordered. By Theorem 1, there exist polynomials $p_T$, with non-negative coefficients, summing to 1, such that the vectors $p_T[\exp(H_T(T))]$ are all equal. Each polynomial $p_T$ can be written as $p_T(x) = \sum a_i x^i$, where $N$ is some positive integer, and the non-negative $a_i$ are dependent on $T$, and sum to 1. Thus, for tree $T$, consider the distribution $\partial_T$ which, for each site assigns $\mu_j = j$, with probability $a_j$. Thus,

$$s(T) = H^{-1} p_T[\exp(H_T(T))]$$

and since the vectors $p_T[\exp(H_T(T))]$ are the same for all $T$, it follows that $s(T)$ are the same for all $T$. Furthermore, if we select the uniform distribution at the root of $T^\mathcal{P}$, we can extend this to obtain that $f(T^\mathcal{P}, P, \partial_T)$ is the same for all $T$—that is, all trees induce the same sequence spectrum. This completes the proof of Theorem 3.

Open problem: Determine further conditions under which $T$ is uniquely determined by its sequence spectrum. For example, is $T$ uniquely determined under the generalized Cavender–Farris model, or more general models, when an unknown set of sites have $\mu_i = 0$, while, for the remaining set of sites, the $\mu_i$ values have a known distribution?

**APPENDIX**

Proof of Theorem 1.

**Definition.** A polynomial $q(x)$ is positive, if it has nonnegative coefficients and $q(1) = 1$, and the polynomial is not identically 1.

**Theorem** For any sequences $0 < x_1 < \ldots < x_n < 1$ and $0 < y_1 < \ldots < y_n < 1$, there are positive polynomials $p(x)$ and $r(x)$, such that

$$p(x_i) = r(y_i) \text{ for } i = 1, 2, \ldots, n.$$  

**Proof:** We prove a seemingly weaker statement:

**Theorem'** For any sequences $0 < x_1 < \ldots < x_n < 1$ and $0 < y_1 < \ldots < y_n < 1$, and any sign sequence $\delta_i = \pm 1$ (i = 1, 2, ..., n) there are positive polynomials $p(x)$ and $r(x)$, such that

$$\delta_i (p(x_i) - r(y_i)) \geq 0 \text{ for } i = 1, 2, \ldots, n.$$  

First we show that Theorem' implies the Theorem. For this purpose we state Theorem'":

**Theorem"** For any sequences $0 < x_1 < \ldots < x_n < 1$ and $0 < y_1 < \ldots < y_n < 1$, and any sign sequence $\delta_i = \pm 1$ or 0 (i = 1, 2, ..., n) there are positive polynomials $p(x)$ and $r(x)$, such that

$$\delta_i (p(x_i) - r(y_i)) \geq 0, \text{ if } \delta_i = \pm 1, \text{ and } (p(x_i) - r(y_i)) = 0, \text{ if } \delta_i = 0, \text{ for } i = 1, 2, \ldots, n.$$
CONSTRUCTION OF EVOLUTIONARY TREES

Clearly, Theorem° implies Theorem by selecting $\delta_i = 0$ for $i = 1, 2, \ldots, n$. We have to show that Theorem° implies Theorem°. We do it by induction on the number of $i$'s with $\delta_i = 0$ in the theorem, say, $m$. The case $m = 0$ is just Theorem°. The case $m = n$ is Theorem°, which is to be proved. Let us be given a sequence of $\delta_i$'s with $m + 1$ zeros in the sequence, assume that $\delta_j = 0$. Define

$$
\delta'_i = \begin{cases} 
1 & \text{if } i = j \\
\delta_i & \text{if } i \neq j
\end{cases}
$$

and

$$
\delta'_r = \begin{cases} 
-1 & \text{if } i = j \\
\delta_i & \text{if } i \neq j
\end{cases}
$$

By the hypothesis, there are $p'$ and $r'$ which satisfy Theorem° with $\delta'_r$, and there are $p''$ and $r''$ which satisfy Theorem° with $\delta'_r$. Since $p'(x_i) - r'(y_j)$ and $p''(x_i) - r''(y_j)$ have different signs, there is an $a$ with $0 \leq a \leq 1$, such that $a(p'(x_i) - r'(y_j)) + (1 - a)(p''(x_i) - r''(y_j)) = 0$. Take now $p = ap' + (1 - a)p''$ and $r = ar' + (1 - a)r''$. Obviously $p$ and $r$ are positive polynomials, $p(x_i) = r(y_j)$ by the choice of $a$; and in any $i, i \neq j$ with $\delta_i = 0$, $p'(x_i) = r'(y_j)$ and $p''(x_i) = r''(y_j)$ imply $p(x_i) = r(y_j)$. If $\delta_i = \pm 1$, then $\delta_i(p'(x_i) - r'(y_j)) \geq 0$ and $\delta_i(p''(x_i) - r''(y_j)) \geq 0$, hence $\delta_i(p(x_i) - r(y_j)) \geq 0$, showing that Theorem° holds with the sign sequence $\delta_i$.

We are left with the task of proving Theorem°. We are going to find the positive polynomials $p(x)$ and $r(x)$ in the following form:

$$
p(x) = \frac{\sum_{i=1}^{n}(1 + x_i^n)q_i}{\sum_{i=1}^{n}2^{q_i}}
$$

and

$$
r(x) = \frac{\sum_{i=1}^{n}(1 + x_i^n)y_i}{\sum_{i=1}^{n}2^{y_i}}
$$

with certain natural numbers $p_i, r_i, q_i$. Before giving the construction, we recall the facts that $1 + x \leq e^x$, $e^{3/4} > 2, 1 + 3x \geq (1 + x)^2$ for all $0 \leq x \leq 1$, and $1 + 2x > e^x$ for all $0 < x < 1/2$.

We define $p_i, r_i, q_i$ in this order for $i = 1, 2, \ldots, n$ recursively, such that we obey the rules below:

(i) if $\delta_i = +1$, then $(3/x_i)y_i \geq x_i^a > 3y_i \geq 3y_i \geq (3/y_i)x_i^a$,

(ii) if $\delta_i = -1$, then $3x_i^a < y_i \leq (3/y_i)x_i^a$;

(iii) $q_i$ is sufficiently large;

(iv) for all $1 \leq i \leq n$, with $i + j \leq n$, $(6y_i)(x_i+y_i)P_{i+j} < 1$ and $(6y_i)(x_i+y_i)P_{i+j} < 1$.

Notice that (iv) sets lower bounds for $p_i$ in terms of $r_{i+j}$, which was defined one step earlier, while (iii) defines $q_k$ in terms of $q_{k-1}$ and the $p$ and $r$ defined in the same step previous to $q_k$. Requirement (i) can be satisfied, since between $a$ and $a/x_1$, where $a/x_1$ is sufficiently small ($a$ and $a/y_1$, where $a/y_1$ is sufficiently small) we always find a member of the sequence $x_i^n(y_i^n)$. Observe that (iii) implies $q_1 < q_2 < \ldots < q_n$.

We are going to show, that evaluating $p(x_i)$ and $r(y_j)$, all other terms than the $k^{th}$ in the numerator of (9) [in the numerator of 10] are negligible compared to the $k^{th}$ term in the numerator of (9) [in the numerator of 10], and the comparison of the $k^{th}$ terms of the numerators of (9) and (10) shows the inequality required in Theorem°.

To substantiate our claims,

$$
\sum_{i=1}^{k-1}(1 + x_i^n)q_i \leq \sum_{i=1}^{k-1}2^{q_i} \leq 2^{k-1} \leq e^{3/4}q_{k-1} \leq [e^{(1/2)}q_{k-1}]^{3/4} \leq [(1 + x_i^n)q_i]^{3/4},
$$

and a similar estimation shows

$$
\sum_{i=1}^{k-1}(1 + y_i^n)q_i \leq \sum_{i=1}^{k-1}2^{y_i} \leq 2^{k-1} \leq e^{3/4}q_{k-1} \leq [e^{(1/2)}q_{k-1}]^{3/4} \leq [(1 + y_i^n)q_i]^{3/4}.
$$
To handle the terms after \( k \), observe that (i) and (iii) imply

\[
q_i \leq \max(2q_{i-1}x_{i}^{\gamma_i}, 2q_{i-1}y_{i}^{\gamma_i}) \leq \min((6x_i)q_{i-1}x_{i}^{\gamma_i}, (6/y_i)2q_{i-1}y_{i}^{\gamma_i}).
\]

(without loss of generality assume \( x_{i}^{\gamma_i} < y_{i}^{\gamma_i} \), which implies the first case of (i). Obviously \( 2/y_i \) settles the right term of the minimization and the inequality in the first case of (i) settles the left term of the minimization.) From here one has

\[
\sum_{i=k+1}^{n} (1 + x_{i}^{\gamma_i})^{\alpha_i} \leq \sum_{i=k+1}^{n} e^{x_{i}^{\gamma_i}} \leq \sum_{i=k+1}^{n} e^{q_i (x_{i}^{\gamma_i})} \leq 2e^{q_1 (x_{k}^{\gamma_i})} \leq 2e^k
\]

[every exponent is less than 1 by (iv)]

\[
en < 2^k
\]

[by the choice of \( q_1 \) in (ii)], and hence

\[
\sum_{i=k+1}^{n} (1 + x_{i}^{\gamma_i})^{\alpha_i} \leq [(1 + x_{k}^{\gamma_i})^{3/4}]
\]

since in (11) \( 2^k \) was among the estimated terms. A similar argument shows that

\[
\sum_{i=k+1}^{n} (1 + y_{i}^{\gamma_i})^{\alpha_i} \leq [(1 + y_{k}^{\gamma_i})^{3/4}]
\]

To finish the proof, we have to make sure, that \( (1 + x_{k}^{\gamma_i})^{3/4} \) and \( (1 + y_{k}^{\gamma_i})^{3/4} \) are large enough, i.e., the 3/4 power of them is negligible compared to the quantity itself. This follows from (11) and the formula after it, since these quantities are larger than the arbitrary \( 2^k \). Finally, we have to show that out of the dominant terms in \( p(x_i) \) and \( r(y_i) \), \( (1 + x_{k}^{\gamma_i})^{3/4} \) and \( (1 + y_{k}^{\gamma_i})^{3/4} \), the bigger term is the correct one, i.e., it is the term that is prescribed by the sign \( \lambda_k \). Since we have a complete symmetry, we may assume without loss of generality, that \( \delta_k = +1 \). By (i), we have

\[
(1 + y_{k}^{\gamma_i})^{2\alpha} \leq (1 + 3y_{k}^{\gamma_i})^{\theta} \leq (1 + x_{k}^{\gamma_i})^{\alpha}
\]

i.e.,

\[
(1 + y_{k}^{\gamma_i})^{2\alpha} \leq [(1 + x_{k}^{\gamma_i})^{1/2}]
\]

as required.

Proof of Theorem 1: We use induction on \( k \). For \( k = 2 \) this is just the theorem proved above. The inductive step from \( k - 1 \) to \( k \) is as follows: find the positive polynomials \( p_1', p_2', \ldots, p_{k-1} \) as required in the theorem. Define a sequence \( y_j : j = 1, 2, \ldots, n \) by \( y_j = p_i'((x_i)_j) \) and observe \( 0 < y_1 < y_2 < \ldots < y_n < 1 \). By the base case \( k = 2 \), there are two positive polynomials, \( h(x) \) and \( r(x) \), such that \( h(y_j) = r((x_i)_j) \) for \( j = 1, 2, \ldots, n \). Take \( p_i = h \circ p_i' \) for \( i = 1, 2, \ldots, k - 1 \) and \( p_k = r \). Since the functional composition of positive polynomials is a positive polynomial, the \( p_i \)'s have the required properties.

Remark. We constructed in the proof of Theorem 1 polynomials with rational coefficients, provided that the \( x_1, x_2, \ldots, x_k \) vectors had all rational coordinates. When defining the generalized Cavender-Farris model, we mentioned an alternative model (1), in which the \( \mu_i \)'s are well-defined but unknown numbers. This model admits a non-reconstructibility result analogous to Theorem 3(2), with a given number of sites, \( c \), for all \( n \)-leaf trees. Virtually the same proof goes through, since, by multiplying the \( q_e \) edge weights by \( 2n \cdot \ln 2 \), the vector \( \text{exp}(H(y,T)) \) has rational coordinates, and the proof is then easy to finish. Here we really need polynomials to obtain a \( c \) value, however, for Theorem 3(2) nonnegative power series would have sufficed.

REFERENCES


CONSTRUCTION OF EVOLUTIONARY TREES


Address reprint requests to:
Dr. M.A. Steel
Mathematics and Statistics Department
University of Canterbury
Christchurch, New Zealand

Received for publication March 15, 1994; accepted March 16, 1994.