Journal of Classification 9:71-90 (1992)

Significance of the Length of the Shortest Tree

Michael A. Steel

Massey University

Michael D. Hendy

Massey University

David Penny

Massey University

Abstract: The distribution of lengths of phylogenetic trees under the taxonomic principle of parsimony is compared with the distribution obtained by randomizing the characters of the sequence data. This comparison allows us to define a measure of the extent to which sequence data contain significant hierarchical information. We show how to calculate this measure exactly for up to 10 taxa, and provide a good approximation for larger sets of taxa. The measure is applied to test sequences on 10 and 15 taxa.

Keywords: Binary trees; Parsimony; Minimum-length tree; Fitch's algorithm; Distributions.

Authors' Addresses: Michael A. Steel, Department of Mathematics, Massey University, Palmerston North, New Zealand; Michael D. Hendy, Department of Mathematics, Massey University, Palmerston North, New Zealand; and David Penny, Department of Botany and Zoology, Massey University, Palmerston North, New Zealand.

1. Introduction

Minimum-length trees are widely used for estimating phylogenetic relationships from aligned sequence data (Felsenstein 1988). These trees have their endpoints labelled with the taxa under study. The minimum-length method selects the tree(s) which require the fewest evolutionary events ("steps") on its edges, to account for the variation within the characters. Until recently, little attention was given to the question of how much better the optimal tree fits the data compared to other trees. If a large number of trees have length similar to the minimum-length tree, T, one might be less confident that T is the correct tree than if T is much shorter than all other trees. Indeed in such a case it could be doubted that the data are derived from an underlying tree-like process.

Related to this is the question of measuring how "tree-like" the data themselves are — by which we mean how much hierarchical information is implicit in the alignment of the characters. Some measures have been proposed, such as the consistency index of Kluge and Farris (1969), which in the case of two character states is equivalent to the reciprocal of the number of "steps" per character. However the significance of the values this measure takes is not clear, and it does not lead to a realistic test of tree-likeness. Further problems of the consistency index have been highlighted by Archie (1989).

A more meaningful measure of tree-likeness takes as a criterion the number of "steps" required to provide a tree-fit, and compares the original data with the data sets obtained after the original character state assignments have been randomly permuted within each character. This approach has recently been advocated by Archie (1989) who applied simulation to test for the presence of significant hierarchical information of 28 data sets. The purpose of this paper is to develop and apply analytical techniques to this problem.

One further, but minor difference between Archie's approach and ours is in the type of measures used. Archie considers indices such as the proportion, $\tau(M)$, of randomizations of the data M which give rise to minimum-length trees that require no more steps than M does on its minimum-length tree. For reasons of computational simplicity, we consider a related, but strictly different index, which is an upper bound on $\tau(M)$.

We define this index in Section 2, and develop the techniques required to calculate the index in Sections 3 and 4. The calculations are made on the partitions of the binary trees into their topological classes. A duality between certain trees in each class and certain colorings of a given representative from each class is exploited. The latter is easily enumerated by multiplying together the appropriate generating functions.

In Section 5 we apply these methods to sequence data on 10 taxa. Section 6 extends the method to larger numbers of taxa by invoking an approximation based on a recent theorem which counts a class of bicolored trees, leading to a further application in Section 7.

The results of this paper are complementary to those of Carter et al. (1990), where formulae for the numbers of trees of different lengths on a single column are derived. In Henderson, Hendy and Penny (1989) the results of this paper are applied to studying evolutionary models of some influenza viruses.

2. Definitions

For $n \geq 1$, let B_n denote the set of (unrooted) binary phylogenetic trees having n terminal vertices (*endpoints*) indexed from the set $\{1, \ldots, n\}$ of taxa, and having all other vertices unlabeled and of degree 3. We denote by $b(n)$ the number of such trees. It is well known (see for instance Harding 1971) that for $n \geq 3$, $b(n) = (2n-5)!! = (2n-5).(2n-7)....3.1$.

Let $C = \{A_1, \ldots, A_r\}$ be an alphabet of character states, which will be referred to as *colors*. For $r = 4$, C might correspond to the four nucleotide bases, while for $r = 2$ the bases may be grouped into purine and pyrimidine bases, or the colors may represent the presence and absence of some morphological characteristic. This latter case ($r = 2$) is the one we shall consider in applications, as the computations are simpler. However there is no fundamental barrier to applying these methods when $r > 2$, and many of the central results have been stated in the full generality of r colors.

Let M be an $n \times c$ array of character states (M_{ij}), $M_{ij} \in C$, representing aligned sequences (for example DNA), and let M_j be the j-th column of M, as in Table 4. Thus M_j is the character occuring at the j-th site in the aligned sequences. For $T \in B_n$ the *weight* of M_j on T, denoted $l(T,M_j)$ is the minimum number of edges of T which must be assigned differently colored endpoints in order to extend the coloring of the endpoints of T described by M_j to all the vertices of T. Fitch's algorithm (Fitch 1971) gives an efficient method ($O(n)$) for calculating $l(T,M_j)$, and a minimal weight coloring of the vertices of T. The length of M on T, denoted $l(T,M)$, is then defined as $\Sigma_{1 \leq j \leq c}\, l(T,M_j)$.

The principle of parsimony (Fitch 1971) regards length as an inverse measure of how well data fits a given tree, and thus selects the tree(s) T minimizing $l(T,M)$ to estimate the underlying evolutionary tree linking the taxa under study. How well such a tree fits M compared to other trees depends on the distribution of the values of $l(T,M)$ over B_n. Thus we define the polynomial:

$$F(M,x) = \Sigma_{s \geq 0} \mid \{T \in B_n : l(T,M) = s\} \mid x^s . \quad (1)$$

In particular we are interested in $l(M) = \min\{l(T,M) : T \in B_n\}$, which by definition is the smallest exponent of x in $F(M,x)$ with nonzero coefficient, (i.e., the length of the minimum-length tree).

To measure the significance of the extent to which M gives rise to a short minimum-length tree we need to measure how sensitive the lower tail of F is to randomizations of the columns of M. Given an element σ of the permutation group S_n on n elements, let $\sigma(M_j)$ be the character vector $(M_{\sigma(i),j})$. For $\sigma = (\sigma_1, \ldots, \sigma_c) \in S_n^c$ let $\sigma(M)$ be the $n \times c$ array whose j-th column is $\sigma_j(M_j)$. Permuting all the columns separately and averaging over all possible such σ destroys any hierarchical relationship between the columns, without altering the relative frequencies with which the various colors occur in each column. This gives us a second, finer distribution defined by the polynomial:

$$G(M,x) = \frac{1}{(n!)^c} \sum_{\sigma \in S_n^c} F(\sigma(M),x) . \quad (2)$$

Comparing the lower "tails" of F (representing the original data) and G (representing randomized columns) gives a measure of how well the columns are aligned to provide a low length fit to a common tree. For a polynomial $f(x)$ let $\langle x^t \rangle f(x)$ denote the sum of the coefficients of x^s in $f(x)$ for $0 \leq s \leq t$. We define the *randomized parsimony distribution index* of M, denoted $\pi(M)$, as

$$\pi(M) = \langle x^{l(M)} \rangle G(M,x) . \quad (3)$$

Thus $\pi(M)$ is the average (over the set of all randomizations, σ, of M) of the number of trees for which $\sigma(M)$ has length $\leq l(M)$.

In particular, if $\tau(M)$ denotes the proportion of randomizations, σ, of M for which $l(\sigma(M)) \leq l(M)$, (the type of statistic considered by Archie, 1989) we have:

$$\tau(M) \leq \pi(M)$$

Thus if $\pi(M) << 1$, few of the possible randomizations of the columns produce data which fit a binary tree as well as M, implying a strong correlation between the columns of M to produce a short tree. Conversely, if the columns of M do not contain any hierarchical information, we would expect $\pi(M)$ to be approximated by the average value over all randomizations, σ, of $\pi(\sigma(M))$. Now a simple, group-theoretic argument shows that for any M, the average value over all randomizations, σ, of $\tau(\sigma(M))$ lies between 0.5 and 1.

In view of the inequality $\pi(\sigma(M)) \geq \tau(\sigma(M))$, it follows that the average value of $\pi(\sigma(M))$ will be at least 0.5. Thus we describe the condition $\pi(M) \geq 0.5$ as the *big bang* model (Henderson, Hendy, and Penny 1989) as it captures the informal notion of there being no hierarchical relationship between the columns of M. This notion can be regarded as a null hypothesis against which to evaluate data M, as suggested in Thompson (1975).

Notice that it is not necessary to establish $l(M)$ in order to reject the big bang model, as a suitable upper bound provided by a heuristic algorithm may suffice to show that $\pi(M) << 1$. It is also worth noting that $\pi(M)$ is not a measure of how likely it is that the shortest tree is the true evolutionary tree linking the taxa. Instead it is a measure of how "tree-like" the data is, a concept which has been used to test biological hypotheses, as in Henderson et al. (1989).

3. Calculations (I)

We now show how to calculate $\pi(M)$.
First we list the elementary relationships between F and G.

(i) $F(M_j,x) = G(M_j,x)$ for $j = 1, \ldots, c$.
(ii) If $\sigma \in S_n^c$, $G(\sigma(M),x) = G(M,x)$.
(iii) If $\sigma^0 = (\sigma, \ldots, \sigma)$, $\sigma \in S_n$, $F(\sigma^0(M),x) = F(M,x)$.
(iv) $F(M,x) = x^k F(M',x)$ and $G(M,x) = x^k G(M',x)$, where M' is obtained from M by deleting the "singular" columns (having at most one color from C occurring more than once), $k = \Sigma_{1 \leq j \leq c} \partial(M_j)$, and $\partial(M_j) + 1$ is the number of colors in M_j if M_j is singular, $\partial(M_j) = 0$ otherwise.
(v) Let $\mu_F(M)$ denote the mean value of $l(T,M)$ averaged over B_n. That is, $\mu_F(M) = b(n)^{-1} \partial / \partial x |_{x=1} F(M,x)$. Let $\mu_G(M)$ denote the corresponding mean for the distribution described by G. Then $\mu_G(M) = \mu_F(M) = \Sigma_{1 \leq j \leq c} \mu_F(M_j)$.

By (iv), singular columns translate both the distributions F and G equally, so it is convenient to delete these columns from data, as in the application of our results below. Also, although the calculation of $l(M)$ is an NP-complete problem (Graham and Foulds 1982), $\mu_F(M)$ can be readily computed from (v) by using Theorem 6.2 (below) to calculate $\mu_F(M_j)$ for $j = 1, \ldots, c$. Despite the equality $\mu_F(M) = \mu_G(M)$ in (v), the variances $\sigma_F^2(M)$, $\sigma_G^2(M)$ of the distributions induced by F and G are not simply related, even when $r = 2$. For if M^k denotes k concatenations of M, then $\sigma_F^2(M^k) = k^2 \sigma_F^2(M)$, while $\sigma_G^2(M^k)$ can be shown to be approximately $k \sigma_G^2(M)$. Thus one can have $\sigma_F^2 >> \sigma_G^2$. Conversely, if M consists of one column for each of the 2^n bicolorings of $\{1, \ldots, n\}$ then $F(M,x) = b(n)x^w$, where

TABLE 1

The Number of Topological Classes of Binary Phylogenetic Trees for n≤18.
(Hendy et al. (1984))

n	≤5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\|\tau(n)\|$	1	2	2	4	6	11	18	37	66	135	265	552	1132	2410

$w = 2((3n-2)2^{n-1} + (-1)^n)/9$ (Steel 1990), so $\sigma_F^2(M^k) = 0$. But $\sigma_G^2(M) \neq 0$, giving $\sigma_G^2(M^k) \to \infty$ as $k \to \infty$.

Computing $G(M,x)$ directly by summing over all σ in S_n^c quickly becomes prohibitive as the length of the sequences, c, or the number of taxa, n, grows. We give a more convenient way to calculate $G(M,x)$ for moderate values of c and n and thereby derive approximations when n or c is large. We begin with the following observation.

The symmetric group S_n acts on B_n with $\sigma(T)$ being the tree obtained from T with endpoint v_i replaced by $v_{\sigma(i)}$, for $i = 1, \ldots, n$. Each orbit in B_n will be called a *topological class* (of order n). The collection of all such classes, $\tau(n)$, grows exponentially in size, indeed $|\tau(n)|$ is asymptotically proportional to $\lambda^n / n^{5/2}$, where $\lambda \simeq 2.48325$ (Otter 1948; Harding 1971). However $|\tau(n)|$ grows much more slowly than $b(n)$; for example $|\tau(10)| = 11$, while $b(10) = 2{,}027{,}025$. Values of $|\tau(n)|$, for $n \leq 18$ are given in Table 1.

For a topological class t of order n, and $T \in t$, we can calculate $|t|$ as:

$$|t| = |S_n : S(T)| = n! / |S(T)| \ , \tag{4}$$

where $S(T)$ is the subgroup of S_n which fixes T (Fraleigh 1982, Theorem 16.3). For binary trees we have $|S(T)| = 6^{k_3(T)}\, 2^{k_2(T) + k_1(T)}$, where $k_3(T)$, (resp. $k_2(T)$) is 1 precisely if T has a vertex (resp. edge) whose deletion breaks the tree into three (resp. two) topologically-equivalent rooted subtrees, and is 0 otherwise; $k_1(T)$ is the number of vertices v of T for which exactly two of the rooted subtrees obtained by deleting v are topologically-equivalent (Hendy, Little, and Penny 1984).

For $T \in B_n$, let $h_m(T,a_1,\ldots,a_r)$ denote the number of ways of coloring the endpoints of T so that precisely a_i endpoints are assigned color A_i for $i = 1,\ldots,r$, and so that the resulting coloring has weight m. Clearly the quantity $h_m(T,a_1,\ldots,a_r)$ is the same for all trees in a fixed topological class, though in general it varies across classes (however when $r = 2$ some topology-invariance results apply, described in Theorem 4.6, below). Let $H(T,a_1,\ldots,a_r) = \Sigma_{m\geq 0} h_m(T,a_1,\ldots,a_r)x^m$, and let $p(M;a_1,\ldots,a_r)$ denote the number of columns of M for which the number of occurrences of the various character states, arranged in nondecreasing frequency is $a_1,\ldots,a_r$.

Theorem 3.1.

$$G(M,x) = \Sigma_{t\in\tau(n)} |t| G(M,x;t)$$

where

$$G(M,x;t) = \Pi((n!)^{-1}\Pi_{1\leq j\leq r} a_j! H(T,a_1,\ldots,a_r)^{p(M;a_1,\ldots,a_r)}), \quad T \in t,$$

and the outer product is over all r-tuples $1 \leq a_1 \leq \cdots \leq a_r$.

Proof. By symmetry $|\{\sigma \in S_n^c : l(T,\sigma(M)) = s\}|$ depends only on the topological class t of T. Thus, for any $T_0 \in t$,

$$\sum_{\sigma\in S_n^c} |\{T \in t: l(T,\sigma(M)) = s\}| = |t|\,|\{\sigma \in S_n^c: l(T_0,\sigma(M)) = s\}|,$$

since both expressions enumerate the set of pairs

$$\{(T,\sigma(M)): T \in t,\ \sigma \in S_n^c \text{ and } l(T,\sigma(M)) = s\}.$$

Thus if we let $G^*(M,x;t) = \Sigma_{s\geq 0} G_s(t)x^s$, where

$$G_s(t) = (n!)^{-c} \,|\{\sigma \in S_n^c: l(T,\sigma(M)) = s\}|\, x^s$$

for any $T \in t$, we have

$$G(M,x) = \Sigma_{t\in\tau(n)} |t|\, G^*(M,x;t). \tag{5}$$

Furthermore $G^*(M,x;t) = \Pi_{1\leq j\leq c} G^*(M_j,x;t)$, and $G^*(M_j,x;t) = (n!)^{-1}\Pi_i a_i! H(T,a_1(j),\ldots,a_r(j))$ where $T \in t$, and M_j has $a_i(j)$ occurrences of the color A_i for $i = 1,\ldots,r$. Thus $G^*(M,x;t) = G(M,x;t)$ and the theorem follows from (5).

4. Calculations (II)

Theorem 3.1 reduces the calculation of $\pi(M)$ to that of evaluating $h_m(T,a_1,\ldots,a_r)$, for any $T \in t$. Let $\alpha = \alpha_n,\alpha_{n-1},\ldots,\alpha_m$, $1 \leq m \leq n$ (where

$\alpha_i \in C$) be a sequence of colors, and let T_α refer to a partial coloring of T with terminal vertex v_j colored α_j, for $j = m, \ldots, n$.

Define $H(T_\alpha, a_1, \ldots, a_r) = \Sigma_{i \geq 0} h_i x^i$, where h_i is the number of ways of assigning a_i endpoints of T color A_i given that vertex v_j is colored α_j for $j = m, \ldots, n$, and so that the resulting coloring has weight i on T. Clearly if α contains more than a_i occurrences of A_i for any i, then $H(T_\alpha, a_1, \ldots, a_r) = 0$.

Also, assigning color A_i to each nonterminal vertex of T, for $1 \leq i \leq r$, gives

$$h_i = 0 \text{ if } i > n - \max\{a_1, \ldots, a_r\} . \tag{6}$$

Regarding C as an ordered set, a permutation $\sigma \in S_r$ permutes the elements of C, and so we can define $\sigma(\alpha) = \sigma(\alpha_n), \ldots, \sigma(\alpha_m)$. Then we have the following result.

(Permutation) Lemma 4.1.

$$H(T_\alpha, a_1, \ldots, a_r) = H(T_{\sigma(\alpha)}, a_{\sigma(1)}, \ldots, a_{\sigma(r)}) , \tag{7}$$

$$H(T, a_1, \ldots, a_r) = H(T, a_{\sigma(1)}, \ldots, a_{\sigma(r)}) . \tag{8}$$

Considering all colors $\beta \in C$ that can be given to v_{m-1} gives a second elementary result.

(Extension) Lemma 4.2.

$$H(T_\alpha, a_1, \ldots, a_r) = \Sigma_{\beta \in C} H(T_{\alpha,\beta}, a_1, \ldots, a_r) . \tag{9}$$

We say T can be reduced at endpoint v_m if there is an edge e of T whose deletion partitions the endpoints into two subsets $\{v_1, \ldots, v_{m-1}\}$ and $\{v_m, \ldots, v_n\}$ with v_m adjacent to an endpoint of e. The subtree containing $v_1, \ldots, v_j, j \leq m - 1$, is denoted ${}^jT$.

We now present two recursions resulting from this decomposition which follow from a property of Fitch's algorithm (Fitch 1971). We first describe part of this algorithm. Given a coloring of the endpoints of a binary tree T, place a root vertex v_0 on the midpoint of any edge of T to give a rooted tree T^*, and direct all edges away from v_0. Assign to each internal vertex a nonempty set $S \subseteq \{A_1, \ldots, A_r\}$ recursively as follows: endpoints are assigned the set containing just their color, and for each vertex v incident with edges directed towards two vertices whose sets S_1, S_2 have already been chosen assign v the set $S(v)$, where

$$S(v) = \begin{cases} S_1 \cap S_2 & \text{if this set is nonempty,} \\ S_1 \cup S_2 & \text{if } S_1 \cap S_2 = \phi \ . \end{cases}$$

The following result is proved by Hartigan (1973).

(Fitch's) Lemma 4.3. *The weight of the coloring is the number of vertices of T^* (including v_0) for which $S(v)$ is defined by the second option in the above process.*

If T can be reduced at v_m then rooting T on an edge in ${}^{m-1}T$ and applying Fitch's lemma gives the following recursions.

(Reduction) Theorem 4.4. *Suppose T can be reduced at v_m, and $\alpha_{m+1}, \ldots, \alpha_n$ are all distinct.*

(i) *if $\alpha_m \in \{\alpha_{m+1}, \ldots, \alpha_n\}$ then*
$H(T_\alpha, a_1, \ldots, a_r) = x^{n-m-1} H({}^m T_{\alpha_m}, b_1, \ldots, b_r)$
where $b_i = a_i - 1$ if $A_i \in \{\alpha_{m+1}, \ldots, \alpha_n\}$, while $b_i = a_i$ otherwise.

(ii) *if $\alpha_m, \alpha_{m+1}, \ldots, \alpha_n$ are all distinct, and include all colors in C, then*
$H(T_\alpha, a_1, \ldots, a_r) = x^{n-m} H({}^{m-1}T, b_1, \ldots, b_r)$, *where $b_i = a_i - 1$.*

Because of (8) we need only calculate $H(T, a_1, \ldots, a_r)$ for $a_1 \le a_2 \le \cdots \le a_r$, the others being obtained by permutation.

We now specialize to the case $r = 2$, setting $a = a_1$, $b = a_2$, and referring to the two colors as A, B. By the Extension Lemma, (4.2),

$$H(T,a,b) = H(T_A,a,b) + H(T_B,a,b)\,, \tag{10}$$

$$H(T_A,a,b) = H(T_{AA},a,b) + H(T_{AB},a,b)\,. \tag{11}$$

By the Permutation Lemma (4.1),

$$H(T_B,a,b) = H(T_A,b,a) \tag{12}$$

Now if T can be reduced at v_{n-1} (that is, if v_n and v_{n-1} are incident with adjacent edges), the Reduction Theorem gives:

$$H(T_{AA},a,b) = H({}^{n-1}T_A, a-1, b)\,, \tag{13}$$

$$H(T_{AB},a,b) = xH({}^{n-2}T, a-1, b-1)\,. \tag{14}$$

These results allow a recursive method to calculate $H(T,a,b)$, and we now give two general applications.

First consider the class ${}^nZ$ (where $n \ge 1$) of linear ("caterpillar") trees, illustrated at the top of Table 2.

TABLE 2

Topological Classes of Trees

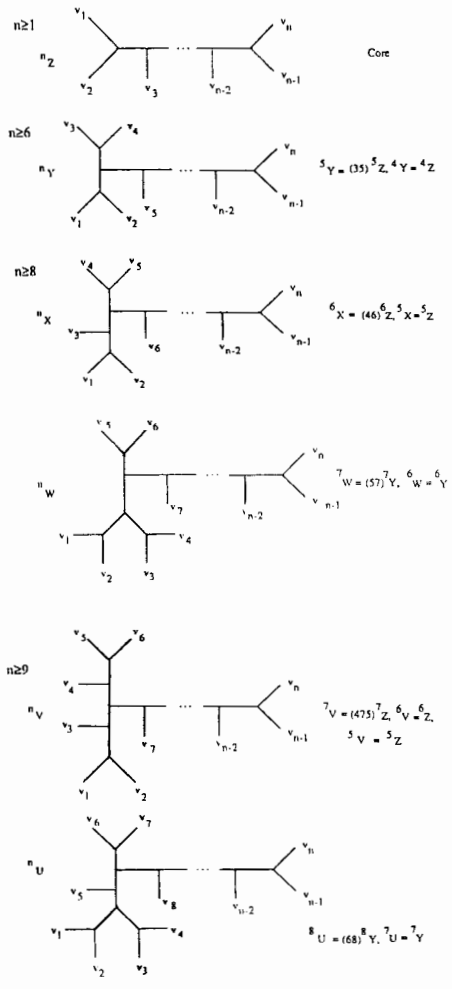
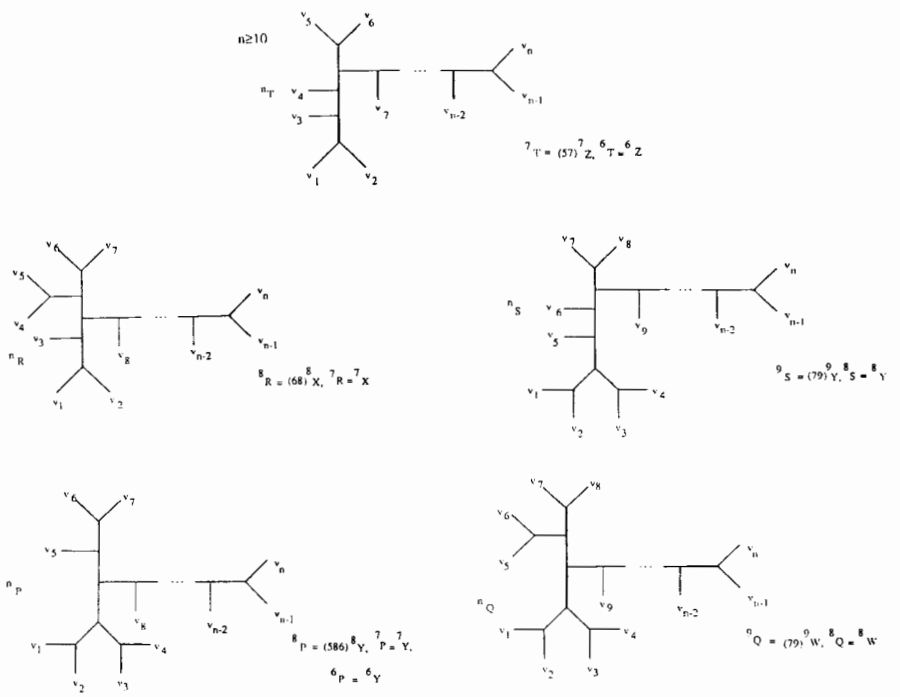


TABLE 2 (Continued)



Let $Z(u,v,x) = \Sigma_{a,b\geq 0,\, a+b\geq 1} H({}^{a+b}Z, a, b)u^a v^b$ (15)

Theorem 4.5.

$$Z(u,v,x) = \frac{(1-uv)}{P} - 1 ,$$

where $P = (1-u)(1-v) - uvx(2-u-v)$.

Proof. For $X = A, B, AB$, let $Z_X = Z_X(u,v,x)$ be defined as for $Z = Z(u,v,x)$ in (15) with ${}^{a+b}Z$ replaced by ${}^{a+b}Z_X$. Then for $n \geq 3$, ${}^nZ$ can be reduced at v_{n-1} and ${}^j({}^nZ) = {}^jZ$, so that by the recursions (10) to (14) we have:

$$Z = Z_A + Z_B ;$$

$$Z_A = uZ_A + uvxZ + u + uvx ;$$

$$Z_B = vZ_B + uvxZ + v + uvx .$$

Solving this system of linear equations gives $Z = Z(u,v,x)$, as required.

Generally, $h_m(T,a,b)$ depends on the topology of T. For example, $h_1(T,a,b)$ is the number of edges of T partitioning the labels of the endpoints of T into two sets of size a,b, and this can be zero on some trees and nonzero on others. We now present three results which are invariant to the topology of T.

Theorem 4.6. *For a tree T in any topological class of order n,*

(a) $h_m(T,m,m) = 2^m$ *for* $n = 2m$.

(b) $h_m(T,m,m+1) = (m+2)2^{m-1}$ *for* $n = 2m+1$.

(c) $\Sigma_{a,b} h_m(T,a,b) = ({}_{(n-m)}C_m + {}_{(n-m-1)}C_m)2^m$ *where* ${}_rC_k$ *denotes the binomial coefficient* $r!/k!(r-k)!$.

Proof. (a): For any $T \in B_n$ let $P(T)$ denote the collection of all the sets (including $\varnothing$) of edges of T which comprise disjoint paths joining endpoints of T. Then $P(T)$ forms a group under symmetric difference ∇.

Suppose $n = 2m$. Letting $\pi(1,i)$ denote the path joining endpoint v_1 and endpoint v_i, and $\Pi = \nabla_{1<i\leq n} \pi(1,i)$, we have $\Pi \in P(T)$ and Π has edges incident with every endpoint of T, so that Π has exactly m components. For any other set Π' of m disjoint paths, $\Pi \nabla \Pi'$ has no edge incident with any endpoint of T, thus since $\Pi \nabla \Pi' \in P(T)$, we have $\Pi \nabla \Pi' = \varnothing$, and so $\Pi' = \Pi$. Thus T has a unique set of m disjoint paths joining its endpoints. By Menger's theorem (Harary 1969, p. 50-51), the weight of a coloring of T is the maximal number of disjoint paths joining differently-colored endpoints of T. As there are 2 ways to color the endpoints of each path in Π in this way, $h_m(T,m,m) = 2^m$, which establishes (a).

For (b), relabel $T \in B_{2m+1}$ so that T can be reduced at v_{n-1}. Applying (8), $2h_m(T,m,m+1) = h_m(T,m,m+1) + h_m(T,m+1,m)$, which by (10) - (14) equals

$$h_m({}^{n-1}T_A,m,m) + h_m({}^{n-1}T_B,m,m) \qquad \text{(i)}$$

$$+ h_m({}^{n-1}T_A,m-1,m+1) + h_m({}^{n-1}T_B,m+1,m-1) \qquad \text{(ii)}$$

$$+ 2(h_{m-1}({}^{n-2}T,m-1,m) + h_{m-1}({}^{n-2}T,m,m-1)) . \qquad \text{(iii)}$$

Now by (10), (i) is $h_m(T,m,m)$, while (6) shows that (ii) is zero. By (8), (iii) is $4h_{m-1}({}^{n-2}T,m-1,m)$. Thus from part (a) we have

$$h_m(T,m,m+1) = 2^{m-1} + 2h_{m-1}({}^{n-2}T,m-1,m) ,$$

giving the inductive step for (b). Part (c) is proved by Steel (1990).

For $n < 6$ all the trees in B_n are in the topological class represented by ${}^nZ$. However for $n \geq 6$ other classes arise. We select representatives for these classes so that a recursion based on the Reduction Theorem connects representatives from each order. We can identify these families as trees growing along a single linear subtree, so that each successive tree J can be considered as an ${}^mZ$ branch attached to a "core" subtree ${}^{n-m}J$ which belongs to an earlier class.

Table 2 shows the 11 families of such trees, ${}^nZ, \ldots, {}^nP$, necessary to cover all topologies arising on up to 10 taxa. The core subtree ${}^{n-m}J$ has been written $\sigma^{n-m}J$, where σ is a permutation of a subset of $\{1, \ldots, n\}$, so that the labelings of the core subtrees are consistent with earlier trees in the hierarchy. For each tree other than ${}^nZ$ (dealt with already), ${}^nV$ and ${}^nP$, the tree J is formed by attaching a linear tree to the edge incident with v_r in $(r-2,r)\,{}^rJ$. The trees ${}^nV$ and ${}^nP$ are formed by attaching a linear tree to the edge incident with v_r in $(r-3,r,r-2)\,{}^rJ$. When $n = 10$, if the trees are ordered ${}^{10}Z, \ldots, {}^{10}P$, the size of the corresponding topological classes, t, calculated from (4), are given by $\log_2(10! / |t|) = 3,4,3,5,4,5,4,7,5,8,6$.

Applying the Extension and Permutation Lemmas, and the Reduction Theorem permits a recursive description of the polynomials $H(J,a,b)$, for the classes of trees J in Table 2. For $n \leq 10$, Table 3 lists the coefficients of these polynomials.

5. Application (I)

Table 4 lists sequence data M of length 56 for ten taxa. This data is a variation on that of Penny and Hendy (1986), converted to purine and pyrimidines, and with singular and constant columns deleted. A cat sequence has been added. The ape and sheep sequences from the original taxa set have been deleted so that there were no pairs of closely-related taxa. The nucleotides have been paired into purines and pyrimidines. A complete search over all trees on ten taxa shows that $l(M,T)$ ranges between 121 and 170, so that $l(M) = 121$. For this data we have the bicoloring frequencies:

$$p(M; a, 10-a) = 17, 19, 17, 3, \text{ for a} = 2, \ldots, 5.$$

Using Theorem 3.1 to calculate $G(M,x)$, Table 5 compares the cumulative coefficients of $F(M,x)$ and $G(M,x)$ in two ranges of interest along their lower tails. From the table we see $\pi(M) = 5.63 \times 10^{-7}$, suggesting the sequence data is highly tree-like. For these data, the minimal length tree has 121 changes in contrast to the 132 which would be expected for random data with the same bicolor column frequencies.

TABLE 3

Coefficients of H(T,a,b) for Topological Classes of Order 4-10.

Class	a	b	Coefficient of: x	x^2	x^3	x^4	x^5
Z	2	2	2	4			
	2	3	2	8			
	2	4	2	13			
	3	3	2	10	8		
	2	5	2	19			
	3	4	2	13	20		
	2	6	2	26			
	3	5	2	16	38		
	4	4	2	16	36	16	
	2	7	2	34			
	3	6	2	19	63		
	4	5	2	19	57	48	
	2	8	2	43			
	3	7	2	22	96		
	4	6	2	22	82	104	
	5	5	2	22	84	112	32
Y	2	4	3	12			
	3	3	0	12	8		
	2	5	3	18			
	3	4	1	14	20		
	2	6	3	25			
	3	5	1	19	36		
	4	4	2	12	40	16	
	2	7	3	33			
	3	6	1	23	60		
	4	5	2	16	60	48	
	2	8	3	42			
	3	7	1	27	92		
	4	6	2	18	90	100	
	5	5	2	22	76	120	32
X	2	6	3	25			
	3	5	2	16	38		
	4	4	0	18	36	16	
	2	7	3	33			
	3	6	2	21	61		
	4	5	1	18	59	48	
	2	8	3	42			
	3	7	2	25	93		
	4	6	1	23	82	104	
	5	5	2	16	90	112	32

TABLE 3 (Continued)

Class	a	b	Coefficient of: x	x^2	x^3	x^4	x^5
W	2	6	4	24			
	3	5	0	24	32		
	4	4	2	4	48	16	
	2	7	4	32			
	3	6	1	27	56		
	4	5	1	13	64	48	
	2	8	4	41			
	3	7	1	33	86		
	4	6	2	12	104	92	
	5	5	0	24	60	136	32
V	2	7	3	33			
	3	6	3	18	63		
	4	5	0	21	57	48	
	2	8	3	42			
	3	7	3	23	94		
	4	6	1	23	82	104	
	5	5	0	20	88	112	32
U	2	7	4	32			
	3	6	0	28	56		
	4	5	2	12	64	48	
	2	8	4	41			
	3	7	1	31	88		
	4	6	1	19	90	100	
	5	5	2	14	84	120	32
T	2	8	3	42			
	3	7	2	26	92		
	4	6	2	18	90	100	
	5	5	0	24	76	120	32
S	2	8	4	41			
	3	7	0	32	88		
	4	6	2	16	96	96	
	5	5	2	18	72	128	32
R	2	8	4	41			
	3	7	2	30	88		
	4	6	0	24	82	104	
	5	5	2	6	100	112	32
Q	2	8	5	40			
	3	7	0	40	80		
	4	6	2	8	120	80	
	5	5	0	20	40	160	32
P	2	8	4	41			
	3	7	2	30	88		
	4	6	1	19	90	100	
	5	5	0	16	84	120	32

Table 4. Ten mammalian sequences converted to 2-state characters. The sequences are basically those from Penny and Hendy (1986) converted to purines (R) or pyrimidines (Y) and then invariant columns and singletons eliminated. The first row contains the consensus sequence, only the differences from the consensus are shown in each column.

```
                    1         2         3         4         5
          12345678901234567890123456789012345678901234567890123456
          RRVRVVVRRRRVVVRVVVRRRVRRRVRRVVRVRVRVVRRRRRRVRVRVRRRVRRVV
 1 Monkey YV....RVV..RR.........V.R.V....V....V..V.....V..V.RVVR.
 2 Horse  .......VV....R...RV.VRV...VV.RV..RV..VVV.VV.......V....R
 3 Kangaroo ..R.R..V...R.R...R.V.....R..RRVR.....VVV....VR..VV.....R
 4 Rodent .V....R..VV...V.R.V..RV...V.....V...R...........V....V...
 5 Rabbit V....R.........R.....R..V..V.R..V........VV.V.......V...
 6 Dog    V...R...V.V...VRR..V....V.V...V..R.RR.VV....VR..V.V.....
 7 Pig    ...V.R.....R.V..RV...V...V...........V...........V.....
 8 Cat    .V..R..V..VR..VRR..VV....R......V..RR.V..VVRVR.RV.V.....
 9 Human  V..V..R..VVRR.........VV..V.........V..V.....V..V.RVVRR
10 Cow    ..R...R.....RV..RV.........R..R.RV....V.VVR...R..V.....
```

TABLE 5

Comparison of the Lower Tail Portion of F and G for the data of Table 4.

s	$\langle x^S \rangle F(M,x)^*$	$\langle x^S \rangle G(M,x)^{**}$
118	0	4.90×10^{-9}
119	0	2.47×10^{-8}
120	0	1.20×10^{-7}
121	1	$5.63 \times 10^{-7} = \pi(M)$
122	5	2.54×10^{-6}
123	17	1.11×10^{-5}
124	50	4.64×10^{-5}
---	---	---
131	3656	0.36
*** 132	5559	1.10

* Number of trees of length ≤ s.

** Expected number of trees of length ≤ s under the "big bang" model.

*** For random data, the expected length of the minimal tree is about 132, in contrast to the actual length of 121.

6. Approximation for n Large

As the number of taxa, n, increases beyond 10 it becomes increasingly difficult to apply Theorem 3.1 because of the number of topologies over which summation is required. However, in the case of two character states, a convenient approximation can be made. We begin with the following definitions.

For a topological class t of order n, let $f_m(a,b;t)$ (resp. $f_m(a,b)$) denote the number of trees in t (resp. in $b(n)$) which have weight m for the coloring of $\{1,\ldots,n\}$, in which $\{1,\ldots,a\}$ and $\{a+1,\ldots,a+b\}$ are assigned colors A and B respectively. The relationship between $f_m(a,b;t)$ and $h_m(T,a,b)$ for $T \in t$, is given by the following result.

Lemma 6.1. *For $T \in t$,*

$$\frac{f_m(a,b;t)}{|t|} = \frac{h_m(T,a,b)a!b!}{(a+b)!}$$

Proof. Consider the collection of pairs (T,χ) where $T \in t$ and χ is a bicoloring of $\{1,\ldots,n\}$ in which a (resp. b) labels are assigned character state A (resp. B), and so that the resulting bicoloring of T has weight m. This set can be enumerated in two ways: by counting the χ's for each T then summing over all trees in t to give $|t|\,h_m(T,a,b)$, or by counting the trees for each χ and then summing over all a/b bicolorings to give ${}_{(a+b)}C_a f_m(a,b;t)$.

We now describe an exact expression for $f_m(a,b)$. Let $N(k,m)$ denote the number of forests consisting of exactly m rooted binary trees (including the degenerate case of the root attached by an edge to a single vertex) on a total of exactly k labeled endpoints. Then from Carter et al. (1989),

$$N(k,m) = \begin{cases} {}_{(2k-m-1)}C_{(m-1)} \times b(k-m+2), \text{ if } k > m, \\ 1 \text{ if } k = m \\ 0, \text{ otherwise.} \end{cases}$$

With this definition we have the following result (proofs appear in Carter et al. (1990) and Steel (1990)).

(Bichromatic Binary Tree) Theorem 6.2

$$f_m(a,b) / b(n) = \begin{cases} (m-1)!(2n-3m)N(a,m)N(b,m) / b(n-m+2), \text{ if } a,b \geq m \\ 0, \text{ otherwise.} \end{cases}$$

where $n = a + b$.

We now apply Lemma 6.1 and Theorem 6.2 to approximate $\pi(M)$. The variation of $h_m(T,a,b)$ across different topology classes for T shows relatively little variation for most values of a,b,m, when $n = a + b$ is large. Indeed when $a = m$ and $b = m$ or $m + 1$ we have an equality across topologies (Theorem 4.2). We therefore introduce an approximation by replacing $h_m(T,a,b)$ with its average value as T varies across $b(n)$. Thus let

$$h_m(a,b) = b(n)^{-1}\Sigma_{T\in\, b(n)}\, h_m(T,a,b)\,.$$

Replacing $h_m(T,a,b)$ by $h_m(a,b)$, Lemma 6.1 shows that $H(T,a,b)$ becomes ${}_{(a+b)}C_a F(a,b)$ where $F(a,b) = b(n)^{-1}\Sigma_{m\geq 0} f_m^{(a,b)} x^m$. Thus from Theorem 3.1,

$$\pi(M) \simeq b(n)\, \langle x^{l(M)}\rangle \Pi_{1\leq a\leq b} F(a,b)^{p(M;a,b)}\,.$$

7. Application (II)

Applying the approximation to the data from Table 4, gives $\pi(M) \simeq 5.40 \times 10^{-7}$, which agrees well with the exact value $\pi(M) = 5.63 \times 10^{-7}$ from Theorem 3.1.

For the sequence data of length $c = 33$ on $n = 15$ genera of Berberidaceae in Penny and Hendy (1987) we have $l(M) = 56$, and bicoloring frequencies $p(M; a,15-a) = 7, 11, 5, 5, 3, 2$, for $a = 2, \ldots, 7$.

Applying the above approximation we find $\pi(M) = 4 \times 10^{-31}$, which again is highly significant.

8. Approximation for c Large

If $c >> b(n)$ we can approximate $\pi(M)$ by areas under normal-distribution curves.

For $0 \leq a_i \leq n$, let

$$\mu(a_1, \ldots, a_r;t) = \Pi_i a_i!\, \Sigma_{m>0} m h_m(T,a_1, \ldots, a_r) / n!$$

and

$$\sigma^2(a_1, \ldots, a_r;t) = \Pi_i a_i!\, \Sigma_{m>0} m^2 h_m(T,a_1,..,a_r) / n! - \mu^2(a_1,..,a_r;t)\,,$$

for any $T \in t$. Let

$$\mu(M,t) = \Sigma\, \mu(a_1,..,a_r;t)\, p\,(M; a_1, \ldots, a_r)\,,$$

$$\sigma^2(M,t) = \Sigma\, \sigma^2(a_1,..,a_r;t)\, p\,(M; a_1, \ldots, a_r)$$

where the summations are over all r-tuples $1 \leq a_1 \leq a_2 \leq \cdots \leq a_r$. Finally let $\phi(x)$ denote the area under the standard normal density curve to the left of x.

Theorem 8.1. *As* $c \to \infty$,

$$\pi(M) \simeq \Sigma_{t \in \tau(n)} \, |t| \, \phi(\lambda(M,t))$$

where $\lambda(M,t) = (l(M) - \mu(M,t)) / \sigma(M,t)$.

Proof. Let $Z_1(t), \ldots, Z_c(t)$ be independent random variables with $Z_j(t)$ assigned probability generating function $G^*(M_j,x;t)$ (defined in the proof of Theorem 3.1) and let $Z(M,t) = \Sigma_j Z_j(t)$. By definition, $Z_j(t)$ has mean $\mu(a_1(j), \ldots, a_r(j);t)$ and variance $\sigma^2(a_1(j), \ldots, a_r(j);t)$. As the $Z_j(t)$'s are independent and uniformly bounded, (though not identitically distributed), a suitable version of the central limit theorem (Bauer 1972, p.279, Example 3) shows that asymptotically $(Z(M,t) - \mu(M,t)) / \sigma(M,t)$ is normally distributed with mean 0 and variance 1. By the independence of $\{Z_1(t), \ldots, Z_c(t)\}$, $Z(M,t)$ has probability generating function $G^*(M,x;t)$. Thus, letting $P[\ ]$ denote the probability operator,

$$\langle x^{l(M)} \rangle \, G^*(M,x;t) = P[Z(M,t) \leq l(M)] =$$

$$P[(Z(M,t) - \mu(M,t)) / \sigma(M,t) \leq \lambda(M,t)] \sim \phi(l(M,t)) \,,$$

which together with (5) gives the theorem.

It is worth pointing out that although G is approximated well near its mean by a weighted sum of normal densities when $c \geq 30$, the tail of G is generally not well approximated, unless $c >> b(n)$ because of the rate at which the normal density function decays. This becomes increasingly important as n grows. For example, applying the approximation with $c = 56$ and $n = 10$ for the data in Table 4 gives an estimate for $\pi(M)$ many orders of magnitude too small. Using Theorem 8.1 to calculate $\pi(M)$ accurately will, for most practical purposes, limit n to being ≤ 6.

9. Summary

Comparing the tail of the distribution of trees according to their length on sequence data with the "randomized" distribution leads to a natural measure ($\pi(M)$) of tree-likeness. The purpose of this paper has been to develop the techniques required to calculate this measure, at least for the case of two character states. For the sequences considered, the very small value of $\pi(M)$ suggests that in both cases the minimal tree is considerably shorter than that expected from the "big bang" model, so that the data can be regarded as

strongly tree-like. It would be useful to extend Theorem 6.2 to deal with the case $r > 2$, at least to $r = 4$. Although the methods above can be used for this, the computations become extremely complex. One result in this direction appears in Carter et al. (1990). To extend the exact computations of this paper to larger values of n will also be difficult, because of the exponential growth of the number of topological classes of trees.

References

ARCHIE, J. W. (1989), "A Randomization Test for Phylogenetic Information in Systematic Data," *Systematic Zoology, 38(3)*, 239-252.

BAUER, H. (1972), *Probability Theory and Elements of Measure Theory*, New York: Holt, Rinehart, and Winston.

CARTER, M. R., HENDY, M.D., PENNY, D., SZÉKELY, L.A., and WORMALD, N.C. (1990), "On the Distribution of Lengths of Evolutionary trees," *SIAM Journal on Discrete Mathematics, 3*, 38-47.

FELSENSTEIN, J. (1988), "Phylogenies from Molecular Sequences: Inference and Reliability," *Annual Review of Genetics, 22*, 521-565.

FITCH, W. M. (1971), "Towards Defining the Course of Evolution: Minimum Change for a Specific Topology," *Systematic Zoology, 20*, 406-416.

FRALEIGH, J. B. (1982), *A First Course in Abstract Algebra* (3rd. ed.), Reading, MA: Addison-Wesley.

GRAHAM, R. L., and FOULDS, L. R. (1982), "Unlikelihood that Minimal Phylogenies for a Realistic Biological Study can be Constructed in Reasonable Computational Time," *Mathematical Biosciences, 60*, 133-142.

HARARY, F. (1969), *Graph Theory*, Reading, MA: Addison-Wesley.

HARDING, E. F. (1971), "The Probabilities of Rooted Tree-shapes Generated by Random Bifurcation," *Advances in Applied Probability, 3*, 44-77.

HARTIGAN, J. A. (1973), "Minimum Mutation Fits to a Given Tree," *Biometrics, 29*, 53-65.

HENDERSON, I. M., HENDY, M. D., and PENNY, D. (1989), "Influenza Viruses, Comets and the Science of Evolutionary Trees," *Journal of Theoretical Biology, 140*, 289-303.

HENDY, M. D., LITTLE, C. H. C., and PENNY, D. (1984), "Comparing Trees with Pendant Vertices Labelled," *SIAM Journal of Applied Mathematics, 44*, 1054-1065.

KLUGE, A. G. and FARRIS, J. S. (1969), "Quantitative Phyletics and the Evolution of Anurans," *Systematic Zoology, 18*, 1-13.

OTTER, R. (1948), "The Number of Trees," *Annals of Mathematics, 49*, 583-599.

PENNY, D., and HENDY, M. D. (1986), "Estimating the reliability of Evolutionary Trees," *Molecular Biology and Evolution, 3 (5)*, 403-417.

PENNY, D., and HENDY, M. D. (1987), "Turbo Tree: a Fast Algorithm for Minimal Trees," *CABIOS, 3 (3)*, 183-187.

STEEL, M. A. (1990), "Distributions on Bicolored Evolutionary Trees," *Bulletin of the Australian Mathematical Society, 41*, 159-160.

THOMPSON, E. A. (1975), *Human Evolutionary Trees*, Cambridge: Cambridge University Press.