

A CENTRAL LIMIT THEOREM FOR THE PARSIMONY LENGTH OF TREES

MIKE STEEL,* *University of Canterbury*

LARRY GOLDSTEIN,** *University of Southern California*

MICHAEL S. WATERMAN,*** *University of Southern California*

Abstract

In phylogenetic analysis it is useful to study the distribution of the parsimony length of a tree under the null model, by which the leaves are independently assigned letters according to prescribed probabilities. Except in one special case, this distribution is difficult to describe exactly. Here we analyze this distribution by providing a recursive and readily computable description, establishing large deviation bounds for the parsimony length of a fixed tree on a single site and for the minimum length (maximum parsimony) tree over several sites. We also show that, under very general conditions, the former distribution converges asymptotically to the normal, thereby settling a recent conjecture. Furthermore, we show how the mean and variance of this distribution can be efficiently calculated. The proof of normality requires a number of new and recent results, as the parsimony length is not directly expressible as a sum of independent random variables, and so normality does not follow immediately from a standard central limit theorem.

TREES; DEPENDENT CENTRAL LIMIT THEOREM; PHYLOGENETIC ANALYSIS; PARSIMONY SCORE; LARGE DEVIATION BOUNDS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60F05

SECONDARY 92B10; 60C05

1. Introduction

Parsimony is a commonly used method to provide a numerical measure of how well given data fit an evolutionary tree. Cavalli-Sforza and Edwards (1967) suggested the use of parsimony for evolutionary studies, and Fitch (1971) provided an algorithm to calculate parsimony scores for molecular sequences. Suppose that homologous sequences for n species have been aligned. For each position of the alignment, we consider a tree T with n leaves corresponding to the n species. Each leaf is given the letter appearing at that position of the sequence of the

Received 26 September 1994; revision received 24 July 1995.

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

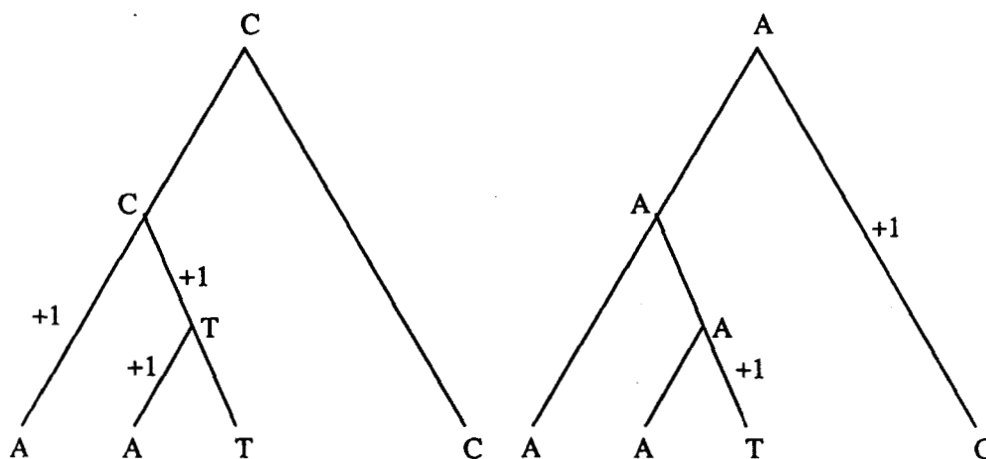
** Postal address: Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113, USA.

*** Postal address: Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, CA 90089-1113, USA.

The research of Larry Goldstein and Michael S. Waterman is supported in part by NSF grant DMS-9005833 and Larry Goldstein in part by NSF grant DMS-9505075. The research of Michael S. Waterman was also supported in part by NIH grant #6M 36230.

corresponding species. Letters are then assigned to all interior vertices of the tree, and the number of edges connecting vertices of different letters is the score of that assignment. The parsimony score $L(T)$ is the minimum score over all possible assignments. The usual procedure is to compute the score $L_i(T)$ associated with sequence position i , and to compute the total score $\sum L_i(T)$ by summing over all positions. We will study the one position problem in most of this paper.

For a simple example we look at two assignments for a tree with $n = 4$ leaves.



The assignment on the left has score $L(T) = 3$ while the assignment on the right has score $L(T) = 2$. Optimal assignments are not always unique. To conform to standard terminology in graph theory, we will refer to the letters as *colors* and to an assignment as a *coloration* of the tree. We now reformulate the problem in more precise terms.

Throughout this paper, by a *binary tree* we mean a tree $T = (V(T), E(T))$ that has labeled leaves of degree 1 and non-leaf vertices of degree 3. We will let n denote the number of leaves of T ; hence $|E(T)| = 2n - 3$. A *rooted binary tree* is a binary tree with a subdivided edge, the resulting newly created vertex of degree two being the *root* of the tree. For technical reasons we also consider an isolated leaf as a rooted binary tree. Given a (possibly rooted) binary tree T , and a coloration \mathcal{X}^* of $V(T)$ by a set of colors, the *changing number* of \mathcal{X}^* on T is the number of edges of T whose ends are assigned different colors by \mathcal{X}^* . Initially we are given a coloration of only the leaves of T . Given such a leaf coloration \mathcal{X} , the *length* of \mathcal{X} on T , denoted $L(T, \mathcal{X})$, or more briefly, $L(T)$, is the minimum changing number of any coloration \mathcal{X}^* of $V(T)$ that extends \mathcal{X} . Such a coloration \mathcal{X}^* is said to be a *minimal coloration* of T for \mathcal{X} . Note that a leaf coloration can have a large number of minimal colorations; indeed the number can grow exponentially with n (Steel 1993a).

A particularly efficient and useful way to calculate $L(T)$ is the forward version of Fitch's algorithm, which we now describe. If T is not already rooted, then define a root by choosing an arbitrary edge of T to subdivide. The value of $L(T)$ is not

dependent on the choice of root. Direct all of the edges of T away from the root, so that each non-leaf vertex has two 'children'. Now, to each vertex of T assign a pair (S, j) , where S is a non-empty set of colors, and j is a non-negative integer, according to the following recursive scheme:

to leaf i , assign the pair $(\{\mathcal{X}(i)\}, 0)$;

to a vertex whose children have been assigned (S_1, j_1) and (S_2, j_2) assign the pair $(S_1 * S_2, j)$ where:

$$(S_1 * S_2, j) = \begin{cases} (S_1 \cap S_2, j_1 + j_2), & \text{if } S_1 \cap S_2 \neq \emptyset, \\ (S_1 \cup S_2, j_1 + j_2 + 1), & \text{otherwise.} \end{cases}$$

Eventually, pairs will be assigned to all of the vertices, including the root v whose associated pair we denote $(S(T), J)$; we will call $S(T)$ the *root set* of T . Hartigan (1973) established the following result.

Lemma 1. $J = L(T)$ and $S(T) = \{\mathcal{X}^*(v) : \mathcal{X}^* \text{ is a minimal coloration of } T \text{ for } \mathcal{X}\}$.

After a parsimony score has been determined it must be evaluated. A natural procedure is to estimate the p -value of the score $L(T)$, i.e. estimate the probability of observing a value as large or larger than $L(T)$ when the colors at the leaves are randomly assigned. In Section 2 we prove large deviation bounds of the form $\mathbb{P}[L(T) - \mathbb{E}[L(T)] > \lambda\sqrt{n}] \leq e^{-\lambda^2/2}$, which hold for all trees with n leaves. We also establish a related bound for the tree that minimizes the total parsimony score over several positions. In Section 3 we prove a central limit theorem for $L(T)$. Some special cases have been considered previously (Moon and Steel 1993) but ours is the first general result. In particular, we allow the distribution of colors to vary from leaf to leaf, so that our results apply to sequences that exhibit variations in their base frequencies. In Section 4 we give recursions for computing the exact distribution of $L(T)$ in $O(n^2 4^c)$ steps, and for computing the mean and variance, $\mathbb{E}[L]$ and $V[L]$, in $O(n 4^c)$ steps, where the alphabet size (number of colors), fixed throughout, is denoted by c . In Section 5 we give an interesting example that shows that the normal limit may hold even when the vertex sets are asymptotically degenerate.

2. Large deviation bounds

We will consider first the single site model where each of the n leaves of the fixed tree T corresponds to the letter found in a given position in each of n (aligned) sequences. In this model, the leaves of a binary tree, T , are colored independently according to (possibly different) probability distributions. We will let π_i^α denote the probability that leaf i is assigned color α , π_i the probability distribution for leaf i , and $\pi = \{\pi_i\}$ be the collection of the leaf distributions.

In the special case where the leaves are bicolored according to the identical and uniform distribution, $\pi_i^\alpha = 0.5$ for all i and both α , the distribution of $L(T)$ depends only on n but not on T and has been determined exactly by Steel (1993b):

$$\mathbb{P}[L(T) = k] = \left[\binom{n-k}{k} + \binom{n-k-1}{k} \right] 2^{(k-n)}.$$

However, in general the exact distribution of $L(T)$ is complex, and the most one can hope for is either a recursive description (Section 3), an asymptotic expression as $n \rightarrow \infty$ (Theorem 3), or large deviation bounds for finite n (Theorem 1). Regarding large deviation bounds we have the following results, in which $\mathbb{E}[L] = \mathbb{E}[L(T)]$ and $V[L] = \text{Var}[L(T)]$.

Theorem 1. For any T and π , for all $\lambda > 0$,

- (1) $\mathbb{P}[L(T) - \mathbb{E}[L] < \lambda\sqrt{n}] \leq e^{-\lambda^2/2}$,
- (2) $\mathbb{P}[L(T) - \mathbb{E}[L] < -\lambda\sqrt{n}] \leq e^{-\lambda^2/2}$,

and for each $p > 0$

$$(3) \quad \mathbb{E}\left[\left|\frac{L(T) - \mathbb{E}[L]}{\sqrt{n}}\right|^p\right] \leq 2p \int_0^\infty \lambda^{p-1} e^{-\lambda^2/2} d\lambda.$$

For $p = 2$, this can be improved to

$$(4) \quad V[L]/n \leq \frac{1}{2}.$$

Proof. First we verify that the parsimony length L satisfies the Lipschitz condition

$$|L(X_1 \cdots X_{i-1}, X_i, X_{i+1}, \dots, X_n) - L(X_1, \dots, X_{i-1}, Y_i, X_{i+1}, \dots, X_n)| \leq 1$$

for all Y_i , where $L(X_1, \dots, X_n)$ is the parsimony length of T when leaf i is assigned color X_i , $i = 1, \dots, n$. By symmetry it suffices to establish

$$L(X_1, \dots, X_{i-1}, Y_i, X_{i+1}, \dots, X_n) \leq L(X_1, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_n) + 1.$$

But altering the color of leaf i from X_i to Y_i increases the length $L(X_1, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_n)$ by at most 1, and yields a coloration that extends $(X_1, \dots, X_{i-1}, Y_i, X_{i+1}, \dots, X_n)$, which, by minimality, has changing number no smaller than $L(X_1, \dots, X_{i-1}, Y_i, X_{i+1}, \dots, X_n)$.

Now, using the independence of X_1, X_2, \dots, X_n , Equations (1) and (2) follow by applying the Azuma–Hoeffding inequality as it appears in Theorem 4.2, p. 90, of Alon and Spencer (1992). Equations (1) and (2) imply $\mathbb{P}[|L - \mathbb{E}[L]| > \lambda\sqrt{n}] \leq 2e^{-\lambda^2/2}$, and (3) now follows from

$$\mathbb{E}[W^p] = p \int_0^\infty \lambda^{p-1} P(W > \lambda) d\lambda$$

for any $W \geq 0$.

The bound for $V[L(T)]$ follows immediately from Steele (1986), using the Lipschitz condition described above.

Again consider the case of n aligned sequences, each of length $k \geq 1$, where all sites are generated independently. Let X_{ij} , $i = 1, \dots, n$, $j = 1, \dots, k$, be the color

assigned to sequence i at position (site) j , and let $L^* = L^*(\{X_{ij}\})$ be the length of a minimum length (maximum parsimony) tree for this data, i.e.

$$L^* = \min_T \sum_{j=1}^k L(T, \mathcal{X}_j)$$

where $\mathcal{X}_j(i) = X_{ij}$. We then have the following result.

Theorem 2. Under the model described above, for $\lambda > 0$,

$$(5) \quad \mathbb{P}[L^* - \mathbb{E}[L^*] > \lambda\sqrt{nk}] \leq e^{-\lambda^2/2},$$

$$(6) \quad \mathbb{P}[L^* - \mathbb{E}[L^*] < -\lambda\sqrt{nk}] \leq e^{-\lambda^2/2}.$$

Proof. Suppose $X'_{ij} = X_{ij}$ except for one value (i_0, j_0) of (i, j) , and let L' be the length of the minimum length tree for $\{X'_{ij}\}$. As in the proof of Theorem 1, because the nk X_{ij} are independent, we need only verify $|L^* - L'| \leq 1$, which by symmetry follows from

$$(7) \quad L' \leq L^* + 1.$$

Suppose T is a minimum length (maximum parsimony) tree for $\{X_{ij}\}$. Then, for \mathcal{X}'_{j_0} (where $\mathcal{X}'_{j_0}(i) = \mathcal{X}_{i,j_0}$) which differs from \mathcal{X}_j in exactly one coordinate, we have that the length of T on $\{X'_{ij}\}$ is

$$\sum_{\substack{j=1, \dots, k \\ j \neq j_0}} L(T, \mathcal{X}_j) + L(T, \mathcal{X}'_{j_0}) \leq \sum_{\substack{j=1, \dots, k \\ j \neq j_0}} L(T, \mathcal{X}_j) + L(T, \mathcal{X}_{j_0}) + 1 = L^* + 1.$$

Now (7) follows by minimality of L' .

3. Central limit theorem

We turn now to the asymptotic behavior of $L(T)$. In the special 2-color case described earlier ($\pi_i^\alpha = 0.5$ for all i and both α), $L(T)$ was shown to be asymptotically normal (Moon and Steel 1993). In general, however, if no restrictions are placed on the distributions $\pi = \{\pi_i^\alpha\}$, then $L(T)$ need not be asymptotically normal; in fact, some condition is necessary to guarantee that the limiting distribution of L is not degenerate. In the case where for two colors, α and β , we have $\pi_i^\alpha = 1 - \pi_i^\beta = \varepsilon_n$, all leaves will be colored β with probability $(1 - \varepsilon_n)^n$, and L will converge in probability to 0 if $n\varepsilon_n \rightarrow 0$. Thus, in order to explore the asymptotic distribution of $L(T)$, we bound the π_i^α uniformly away from 0, i.e. we assume

$$(8) \quad \pi_i^\alpha > \varepsilon, \quad \text{for all } i, \alpha,$$

for some $\varepsilon > 0$ (independent of n).

A conjecture, which generalizes conjectures reported by Archie and Felsenstein (1993) and Moon and Steel (1993), is that $L(T)$ is asymptotically normal under condition (8). The following theorem shows that this is indeed so, and provides

order estimates for the growth in the mean and variance of the distribution. Note that, under condition (8), a quantity closely related to $L(T)$, namely the root set $S(T)$, can still be degenerate asymptotically, as the example in Section 5 shows. In the following theorem recall that n denotes the number of leaves of T .

Theorem 3. Under condition (8), the distribution of $(L(T) - \mathbb{E}[L])/\sqrt{V[L]}$ converges to the standard normal distribution $N(0, 1)$ as $n \rightarrow \infty$. Furthermore, both $\mathbb{E}[L]$ and $V[L]$ grow (approximately) linearly with n .

For completeness, and due to its central role, we now make precise as a separate proposition the last part of the claim in Theorem 3.

Proposition 1. For all π satisfying condition (8), there exists a $\delta > 0$, depending on ε , such that $[\delta, 1)$ contains $\mathbb{E}[L]/n$ and $V[L]/n$ for all binary trees T .

In order to prove this proposition and the theorem, we need to establish a number of preliminary results. The next three lemmas reflect purely combinatorial properties of binary trees.

Lemma 2. ('Lonely leaves lemma'.) The leaves of any binary tree T can be ordered l_1, l_2, \dots, l_n in such a way that, for at least $1 + n/3$ values of i , l_i and l_{i-1} are separated by no more than 3 edges.

Proof. For leaves i and j let $d(i, j)$ denote the number of edges of T separating i and j . For $n > 3$, delete from T all its leaves, and their incident edges, to obtain a tree T_1 , which is the subdivision of a unique binary tree T_2 , as in Figure 1. Note that an edge e of T_2 corresponds to a path in T_1 , and we denote by $X(e)$ the (possibly empty) set of leaves of T that are adjacent to any vertex in this path. We partition the edges of T_2 into four classes as follows:

- C_1 : edges incident with a leaf of T_2 .
- C_2 : edges not incident with a leaf of T_2 and with $|X(e)| = 0$.
- C_3 : edges not incident with a leaf of T_2 and with $|X(e)| = 1$.
- C_4 : edges not incident with a leaf of T_2 and with $|X(e)| \geq 2$.

For example, the tree in Figure 1 has $C_2 = \{e_2\}$, $C_3 = \{e_3\}$, $C_4 = \{e_4\}$. Note the sets $X(e)$ from cases (1), (3), and (4) partition the leaves of T , and case (3) covers precisely the 'lonely' leaves i for which $d(i, j) > 3$ for all leaves j . It is clear that we can relabel the leaves of T as l_1, \dots, l_n in such a way that, for any edge e in class (1) or (4), $d(l_i, l_{i-1}) \leq 3$ for at least $|X(e)| - 1$ values of i . Under this ordering, the number of leaves i for which $d(l_i, l_{i-1}) \leq 3$ satisfies

$$\sum_{e \in C_1 \cup C_4} (|X(e)| - 1) \geq \sum_{e \in C_1 \cup C_4} |X(e)|/2 \geq (n - |C_3|)/2.$$

Now, $|C_3|$ is bounded above by the number of edges of T_2 that are not incident with a leaf. Thus, if T_2 has k leaves, then $|C_3| \leq k - 3$, and furthermore, the number n of

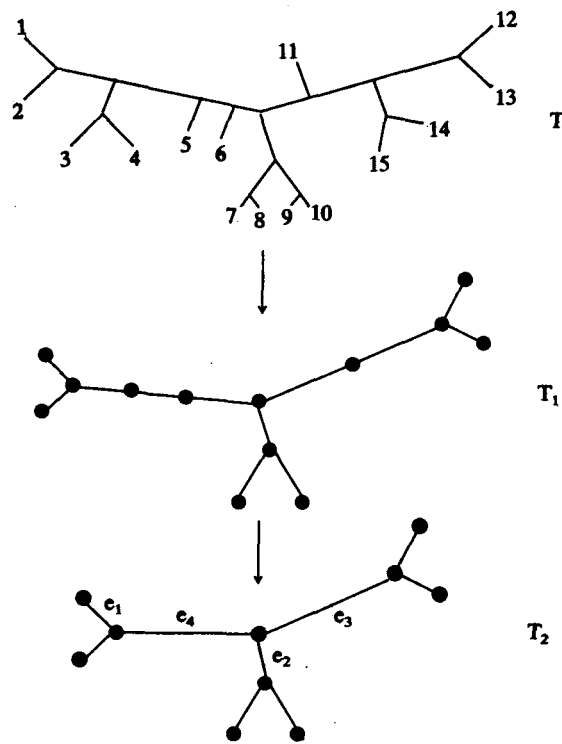


Figure 1. A derived tree T_2 with examples e_i of edges in class C_i

leaves of T is at least $2k + |C_3|$, thereby providing the inequality $n \geq 2(|C_3| + 3) + |C_3|$. Thus, $|C_3| \leq n/3 - 2$, and so $(n - |C_3|)/2 \geq n/3 + 1$. Combining this observation with the string of inequalities above establishes the Lemma, and shows that this bound is the best possible.

The following definition is illustrated in Figure 2.

Definition. Suppose T is a binary tree leaf-labeled by L . A leaf-covering forest

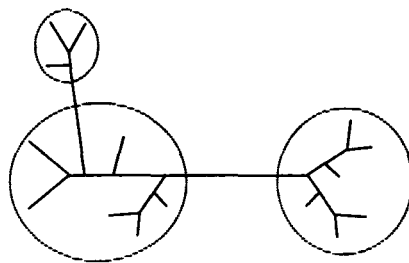


Figure 2. A leaf-covering forest (circled) for $k = 4$

for T is a collection of vertex disjoint subtrees of T whose leaf sets form a partition of the leaf set of T .

Lemma 3. ('Tree-chopping lemma'.) *Let T be an arbitrary binary tree and k an integer, $k \geq 2$. Then T has a leaf covering forest F with the following two properties.*

- (i) *The number of leaves in each tree in F is at most $2k - 2$.*
- (ii) *The number of leaves in each tree in F is at least k , except possibly for one tree.*

Proof. Select a leaf i of T , and let T' denote a subtree of T containing i . Direct all edges of T' away from i , thereby creating a partial order \leq on the vertices of T' with minimal element i . Thus, each vertex $v \in V[T']$ has a set of 'descendents' = $\{v' \in V[T'] : v \leq v'\}$, and we let $d(T', v)$ denote the number of leaves of T' which are descendents of v . The following algorithm constructs the required leaf covering forest F .

If T has fewer than k leaves, the lemma holds. Otherwise, there is a maximal (under \leq) vertex v of T , with $d(T, v) \geq k$. If $d(T, v) > 2k - 2$ and, if u, u' are the descendents of v , then $d(T, v) = d(T, u) + d(T, u')$, and $\max\{d(T, u), d(T, u')\} \geq k$, contradicting the maximality of v . Therefore $d(T, v) \leq 2k - 2$. Next we remove the tree consisting of v along with its descendent vertices and incident edges and place it in F . Then we inductively repeat the above procedure on the remaining tree until it has fewer than k leaves. Note that the subtrees removed can have vertices of degree 2.

Figure 2 shows a leaf-covering forest of T in case $k = 4$.

Lemma 4. *Suppose $F = \{T_1, \dots, T_r\}$ is a leaf-covering forest for T . Given a leaf coloration \mathcal{X} of T , let \mathcal{X}_i denote the restriction of \mathcal{X} to the leaves of T that lie in T_i , and let $\Delta = L(T) - \sum_{i=1}^r L(\mathcal{X}_i, T_i)$. Then $0 \leq \Delta \leq r - 1$.*

Remark. The bound $\Delta \leq 2r - 3$ is sufficient for proving Theorem 3, and follows immediately from the observation that the forest can have no more edges joining the r trees T_i than the number of edges of a binary tree with r leaves. The better bound established in the lemma might be useful in other contexts.

Proof of Lemma 4. Suppose T is any tree (possibly with degree 2 vertices) and F is a subforest of T , whose components collectively cover all T 's leaves. Then F determines a collection Q of subtrees of T , as follows. Let E' denote the set of edges of T that do not lie entirely in F , and let V' denote the vertices of T that are incident with an edge of E' . Let $Q = Q(T, F)$ denote the set of (leaf-overlapping) subtrees of (V', E') that have all their leaves, but no other vertices consisting of vertices from trees in F . An example of this construction is given in Figure 3, where the five trees in F are circled. Let n_t denote the number of leaves of $t \in Q$. We claim that

$$\sum_{t \in Q} (n_t - 1) = |F| - 1.$$

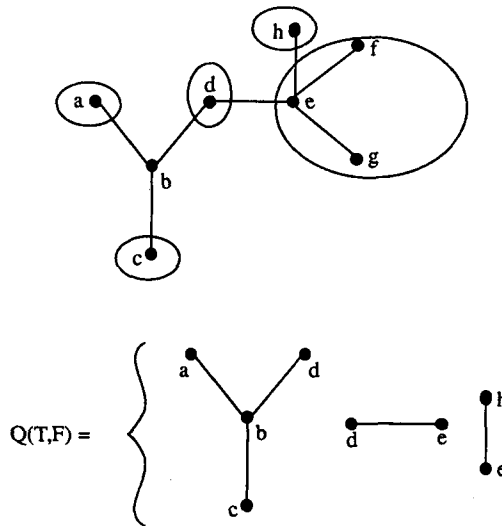


Figure 3. Subtrees obtained from a leaf covering forest of five trees (circled)

To establish this claim, we first note that we may, without loss of generality, assume that F consists entirely of isolated vertices.

Then, take any vertex v in F and direct all edges of T away from v . Then each vertex v' in $F - \{v\}$ has an incident edge $e_{v'}$ directed towards it, and $e_{v'}$ lies in precisely one of the trees $t_{v'}$ in Q , and v' is a leaf of $t_{v'}$. All but one of the leaves of $t_{v'}$ is identified in this way; thus we have a one-to-one mapping from $F - \{v\}$ to the one-vertex-deleted leaf sets of the trees in Q , and this establishes the claim.

Now let \mathcal{X}^* be a minimal coloration of T for \mathcal{X} . Let \mathcal{X}_i^* denote the restriction of \mathcal{X}^* to $V(T_i)$. Then \mathcal{X}_i^* extends \mathcal{X}_i and so $L(\mathcal{X}_i, T_i)$ is at most the changing number of \mathcal{X}_i^* on T_i , hence

$$\sum_{i=1}^r L(\mathcal{X}_i, T_i) \leq \text{changing number of } \mathcal{X} \text{ on } T = L(\mathcal{X}, T),$$

which shows that $0 \leq \Delta$.

To obtain an upper bound for Δ , let \mathcal{X}^i denote a minimal coloration of T_i for \mathcal{X}_i . Define a coloration \mathcal{X}' of the vertices of T that lie in trees from F , by setting $\mathcal{X}'(v) = \mathcal{X}^i(v)$ if $v \in V(T_i)$. Extend \mathcal{X}' to a coloration \mathcal{X}'' of $V(T)$ by coloring any vertices in T not covered by trees in F (and therefore lying in $Q(T, F)$) in such a way that all the non-leaf vertices in any component of $Q(T, F)$ are assigned the same color as one of the leaves of that component. The changing number of \mathcal{X}'' on T is the sum (over i) of the changing number of \mathcal{X}^i on T_i , plus the sum of the changing numbers of the restriction of \mathcal{X}'' to the components of $Q(T, F)$. This latter sum is no more than $\sum_{i \in Q} (n_i - 1)$. But this sum was shown above to equal $r - 1$, so that \mathcal{X}''

has changing number at most $\sum_{i=1}^r L(\mathcal{X}_i, T_i) + r - 1$, and this gives the required upper bound on $L(\mathcal{X}, T)$ because \mathcal{X}'' extends \mathcal{X} .

Proof of Proposition 1. The upper bounds are easily derived. However, the justification for the lower bounds, particularly for $V[L]$ is much more involved and occupies most of the proof. For any leaf coloration \mathcal{X} , consider the coloration \mathcal{X}^* of $V(T)$ that assigns the most frequently occurring leaf color to all the vertices of T . Then \mathcal{X}^* extends \mathcal{X} , and has changing number $n - a$, where a is the number of leaves colored with the most frequently occurring color. Because the most frequent color must occur at least n/c times, $L \leq (1 - 1/c)n$ and

$$\frac{\mathbb{E}[L]}{n} \leq 1 - \frac{1}{c} < 1,$$

where c is the number of colors.

To obtain a lower bound, consider any leaf i of T , and the edge e incident with i . Deleting e and i from T produces a rooted binary tree T' , rooted at a vertex v of degree 2, that was formerly incident with e . Let \mathcal{X}' denote the restriction of \mathcal{X} to T' . We see (Figure 4) that $L(T) = L(T') + D_i$, where D_i is a 0, 1 random variable, which equals 1 precisely if $\mathcal{X}(i) \notin S(T')$. Thus, $\mathbb{E}[L(T)] = \mathbb{E}[L(T')] + \mathbb{P}[D_i = 1]$. Now a modification of the Example in Section 5 shows that, in general, $\mathbb{P}[D_i = 1]$ can be arbitrarily close to 0. However, we can always find, in any binary tree, a leaf i that is separated from another leaf j by just two edges, and in this case we will show that $\mathbb{P}[D_i = 1]$ is bounded away from 0. Thus, represent T and T' as in Figure 4, and let T'' be the rooted tree obtained from T' by deleting leaf j , the root and its two incident edges (see Figure 4.) We have, for any $\alpha \neq \beta$,

$$\mathbb{P}[D_i = 1] = \mathbb{P}[\mathcal{X}(i) \notin S(T')] \geq \mathbb{P}[\mathcal{X}(i) = \alpha, S(T') = \{\beta\}],$$

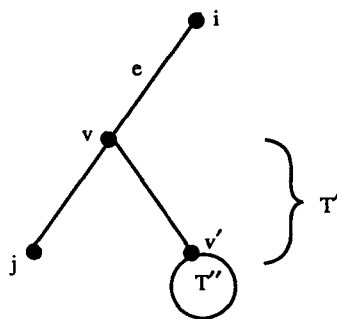


Figure 4. Tree decomposition

and by the independence of leaf colorations and condition (8),

$$\begin{aligned} \mathbb{P}[\mathcal{X}(i) = \alpha, S(T') = \{\beta\}] &= \mathbb{P}[\mathcal{X}(i) = \alpha] \times \mathbb{P}[S(T') = \{\beta\}] \\ &\geq \varepsilon \mathbb{P}[S(T') = \{\beta\}]. \end{aligned}$$

Combining these two inequalities we have, for any β ,

$$(9) \quad \mathbb{P}[D_i = 1] \geq \varepsilon \mathbb{P}[S(T') = \{\beta\}].$$

Now,

$$(10) \quad \begin{aligned} \mathbb{P}[S(T') = \{\beta\}] &\geq \mathbb{P}[\mathcal{X}(j) = \beta, \beta \in S(T'')] \\ &= \mathbb{P}[\mathcal{X}(j) = \beta] \times \mathbb{P}[\beta \in S(T'')] \geq \varepsilon \mathbb{P}[\beta \in S(T'')]. \end{aligned}$$

But $S(T'') \neq \emptyset$, so there is a color β for which $\mathbb{P}[\beta \in S(T'')] \geq 1/c$ where c is the number of colors. Thus, from (9) and (10), we have $\mathbb{P}[D_i = 1] \geq \varepsilon^2/c$ and so $\mathbb{E}[L(T)] \geq \mathbb{E}[L(T')] + \varepsilon^2/c$. Continuing in this way using the remaining tree with the root v removed, we will eventually, by induction, obtain a tree $T(2)$ with just 2 leaves, providing the inequality

$$\mathbb{E}[L(T)] \geq \frac{\varepsilon^2}{c} + \mathbb{E}[L(T')] \geq 2 \frac{\varepsilon^2}{c} + \mathbb{E}[L(T'')] \geq \dots \geq \frac{\varepsilon^2(n-2)}{c} + \mathbb{E}[K(T(2))]$$

and hence, $\mathbb{E}[L(T)]/n$ is uniformly bounded from below.

The upper bound $V[L]/n \leq \frac{1}{2}$ is given in Theorem 1. The remainder of the proof is devoted to obtaining a lower bound. First, apply Lemma 2 to order the leaves of T , l_1, \dots, l_n , so that for at least $1 + n/3$ values of i , l_i and l_{i-1} are separated by either 2 or 3 edges. Write $L(T)$ as a function $L(X_1, \dots, X_n)$ where $X_i = \mathcal{X}(l_i)$, and let $d_k = \mathbb{E}[L | \mathcal{F}_k] - \mathbb{E}[L | \mathcal{F}_{k-1}]$, for $k = 1, 2, \dots, n$, where \mathcal{F}_k is the sigma algebra generated by (X_1, \dots, X_k) . As d_k is a martingale difference sequence, the variates d_k are uncorrelated; now using that $L - \mathbb{E}[L] = \sum_k d_k$,

$$V[L] = \mathbb{E} \left[\sum_{k=1}^n d_k^2 \right].$$

Now, let T' denote the tree obtained from T by deleting leaf l_k and its incident edge. Applying Lemma 1 to T , taking as the root the vertex arising by subdividing this edge we have $L = L' + D_k$, where $L' = L(T')$ and $D_k = 1$ when $\mathcal{X}(l_k) \notin S(T')$, and is zero otherwise.

Since $\mathbb{E}[L' | \mathcal{F}_k] = \mathbb{E}[L' | \mathcal{F}_{k-1}]$ we have

$$(11) \quad d_k = \mathbb{E}[D_k | \mathcal{F}_k] - \mathbb{E}[D_k | \mathcal{F}_{k-1}].$$

Now, for any two colors α and α' ,

$$\begin{aligned}
 \mathbb{E}[d_k^2] &\cong \mathbb{E}[d_k^2 \mid X_{k-1} = \alpha, X_k = \alpha'] \times \mathbb{P}[X_{k-1} = \alpha, X_k = \alpha'] \\
 (12) \quad &\cong \varepsilon^2 \mathbb{E}[d_k^2 \mid X_{k-1} = \alpha, X_k = \alpha'] \\
 &\cong \varepsilon^2 \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_k = \alpha']^2,
 \end{aligned}$$

and using (11),

$$(13) \quad \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_k = \alpha'] = \mathbb{E}[D_k \mid X_{k-1} = \alpha, X_k = \alpha'] - \mathbb{E}[D_k \mid X_{k-1} = \alpha].$$

Letting $f_{\beta, \alpha} = \mathbb{P}[\beta \notin S(T') \mid X_{k-1} = \alpha]$, (13) can now be written

$$(14) \quad \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_k = \alpha'] = f_{\alpha', \alpha} - \sum_{\beta} \mathbb{E}[D_k \mid X_{k-1} = \alpha, X_k = \beta] \pi_k^{\beta}$$

$$(15) \quad = f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}.$$

Thus, from (12), (13) and (14), for all α, α' ,

$$(16) \quad \mathbb{E}[d_k^2] \cong \left[f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta} \right]^2 \varepsilon^2.$$

From Lemma 5, shown below, there is a $b > 0$ such that, when l_k and l_{k-1} are separated by either 2 or 3 edges, there exist α, α' such that $|f_{\alpha', \alpha} - f_{\alpha, \alpha}| > b$. By the triangle inequality, it follows that either $|f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}|$ or $|f_{\alpha, \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}|$ is at least $b/2$. Hence, by (16), and an appropriate choice of α, α' , $\mathbb{E}[d_k^2] \cong b^2 \varepsilon^2 / 4$, and so, by Lemma 2, $\text{Var}[L] \cong b^2 \varepsilon^2 (1 + n/3) / 4$, which provides the required lower bound. Therefore, the proof of the proposition is complete once we establish the following lemma.

Lemma 5. If leaves l_k and l_{k-1} are separated by either 2 or 3 edges, then α and α' can be chosen so that $|f_{\alpha, \alpha} - f_{\alpha', \alpha}| > b$ for some constant $b > 0$ independent of n .

Proof of Lemma 5. Let S_x be the root set of the tree in Figure 5, $S_x = (\{x\} * P) * Q$, where P, Q are random variables taking values in $2^{\mathcal{E}} - \emptyset$ (the non-

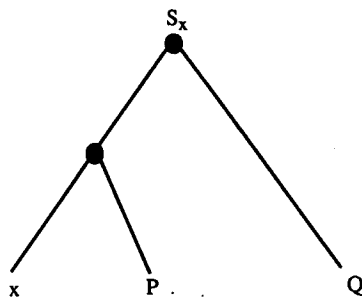


Figure 5. $S_x = (\{x\} * P) * Q$

and by the independence of leaf colorations and condition (8),

$$\begin{aligned} \mathbb{P}[\mathcal{X}(i) = \alpha, S(T') = \{\beta\}] &= \mathbb{P}[\mathcal{X}(i) = \alpha] \times \mathbb{P}[S(T') = \{\beta\}] \\ &\geq \varepsilon \mathbb{P}[S(T') = \{\beta\}]. \end{aligned}$$

Combining these two inequalities we have, for any β ,

$$(9) \quad \mathbb{P}[D_i = 1] \geq \varepsilon \mathbb{P}[S(T') = \{\beta\}].$$

Now,

$$(10) \quad \begin{aligned} \mathbb{P}[S(T') = \{\beta\}] &\geq \mathbb{P}[\mathcal{X}(j) = \beta, \beta \in S(T'')] \\ &= \mathbb{P}[\mathcal{X}(j) = \beta] \times \mathbb{P}[\beta \in S(T'')] \geq \varepsilon \mathbb{P}[\beta \in S(T'')]. \end{aligned}$$

But $S(T'') \neq \emptyset$, so there is a color β for which $\mathbb{P}[\beta \in S(T'')] \geq 1/c$ where c is the number of colors. Thus, from (9) and (10), we have $\mathbb{P}[D_i = 1] \geq \varepsilon^2/c$ and so $\mathbb{E}[L(T)] \geq \mathbb{E}[L(T')] + \varepsilon^2/c$. Continuing in this way using the remaining tree with the root v removed, we will eventually, by induction, obtain a tree $T(2)$ with just 2 leaves, providing the inequality

$$\mathbb{E}[L(T)] \geq \frac{\varepsilon^2}{c} + \mathbb{E}[L(T')] \geq 2 \frac{\varepsilon^2}{c} + \mathbb{E}[L(T'')] \geq \dots \geq \frac{\varepsilon^2(n-2)}{c} + \mathbb{E}[K(T(2))]$$

and hence, $\mathbb{E}[L(T)]/n$ is uniformly bounded from below.

The upper bound $V[L]/n \leq \frac{1}{2}$ is given in Theorem 1. The remainder of the proof is devoted to obtaining a lower bound. First, apply Lemma 2 to order the leaves of T , l_1, \dots, l_n , so that for at least $1 + n/3$ values of i , l_i and l_{i-1} are separated by either 2 or 3 edges. Write $L(T)$ as a function $L(X_1, \dots, X_n)$ where $X_i = \mathcal{X}(l_i)$, and let $d_k = \mathbb{E}[L | \mathcal{F}_k] - \mathbb{E}[L | \mathcal{F}_{k-1}]$, for $k = 1, 2, \dots, n$, where \mathcal{F}_k is the sigma algebra generated by (X_1, \dots, X_k) . As d_k is a martingale difference sequence, the variates d_k are uncorrelated; now using that $L - \mathbb{E}[L] = \sum_k d_k$,

$$V[L] = \mathbb{E} \left[\sum_{k=1}^n d_k^2 \right].$$

Now, let T' denote the tree obtained from T by deleting leaf l_k and its incident edge. Applying Lemma 1 to T , taking as the root the vertex arising by subdividing this edge we have $L = L' + D_k$, where $L' = L(T')$ and $D_k = 1$ when $\mathcal{X}(l_k) \notin S(T')$, and is zero otherwise.

Since $\mathbb{E}[L' | \mathcal{F}_k] = \mathbb{E}[L' | \mathcal{F}_{k-1}]$ we have

$$(11) \quad d_k = \mathbb{E}[D_k | \mathcal{F}_k] - \mathbb{E}[D_k | \mathcal{F}_{k-1}].$$

Now, for any two colors α and α' ,

$$\begin{aligned}
 \mathbb{E}[d_k^2] &\geq \mathbb{E}[d_k^2 \mid X_{k-1} = \alpha, X_k = \alpha'] \times \mathbb{P}[X_{k-1} = \alpha, X_k = \alpha'] \\
 (12) \quad &\geq \varepsilon^2 \mathbb{E}[d_k^2 \mid X_{k-1} = \alpha, X_k = \alpha'] \\
 &\geq \varepsilon^2 \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_{k-1} = \alpha']^2,
 \end{aligned}$$

and using (11),

$$(13) \quad \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_k = \alpha'] = \mathbb{E}[D_k \mid X_{k-1} = \alpha, X_k = \alpha'] - \mathbb{E}[D_k \mid X_{k-1} = \alpha].$$

Letting $f_{\beta, \alpha} = \mathbb{P}[\beta \notin S(T') \mid X_{k-1} = \alpha]$, (13) can now be written

$$(14) \quad \mathbb{E}[d_k \mid X_{k-1} = \alpha, X_k = \alpha'] = f_{\alpha', \alpha} - \sum_{\beta} \mathbb{E}[D_k \mid X_{k-1} = \alpha, X_k = \beta] \pi_k^{\beta}$$

$$(15) \quad = f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}.$$

Thus, from (12), (13) and (14), for all α, α' ,

$$(16) \quad \mathbb{E}[d_k^2] \geq \left[f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta} \right]^2 \varepsilon^2.$$

From Lemma 5, shown below, there is a $b > 0$ such that, when l_k and l_{k-1} are separated by either 2 or 3 edges, there exist α, α' such that $|f_{\alpha', \alpha} - f_{\alpha, \alpha}| > b$. By the triangle inequality, it follows that either $|f_{\alpha', \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}|$ or $|f_{\alpha, \alpha} - \sum_{\beta} f_{\beta, \alpha} \pi_k^{\beta}|$ is at least $b/2$. Hence, by (16), and an appropriate choice of α, α' , $\mathbb{E}[d_k^2] \geq b^2 \varepsilon^2 / 4$, and so, by Lemma 2, $\text{Var}[L] \geq b^2 \varepsilon^2 (1 + n/3) / 4$, which provides the required lower bound. Therefore, the proof of the proposition is complete once we establish the following lemma.

Lemma 5. *If leaves l_k and l_{k-1} are separated by either 2 or 3 edges, then α and α' can be chosen so that $|f_{\alpha, \alpha} - f_{\alpha', \alpha}| > b$ for some constant $b > 0$ independent of n .*

Proof of Lemma 5. Let S_x be the root set of the tree in Figure 5, $S_x = (\{x\} * P) * Q$, where P, Q are random variables taking values in $2^{\mathcal{E}} - \emptyset$ (the non-

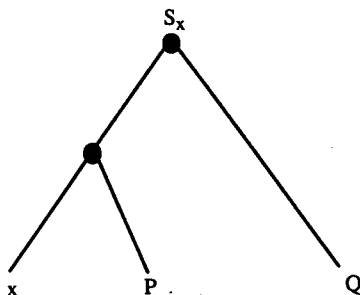


Figure 5. $S_x = (\{x\} * P) * Q$

empty subsets of the set of colors \mathcal{C} , and $x \in \mathcal{C}$. Let

$$\begin{aligned} E_{\alpha,\beta} &= \mathbb{P}[\alpha \notin S_\alpha] - \mathbb{P}[\beta \notin S_\alpha] \\ &= \mathbb{P}[\beta \in S_\alpha] - \mathbb{P}[\alpha \in S_\alpha]. \end{aligned}$$

Returning to the tree T , if l_k and l_{k-1} are separated by three edges, then taking P, Q to be the root sets of the two rooted subtrees whose roots are adjacent to the two internal vertices of the path connecting l_{k-1} and l_k , we have that $E_{\alpha,\beta} = f_{\alpha,\alpha} - f_{\beta,\alpha}$.

If l_k and l_{k-1} are separated by just two edges, then this same equation applies if we take P to be the root set of the rooted subtree between l_{k-1} and l_k , and if we take Q to be \mathcal{C} with probability 1. Thus to establish the lemma, it suffices to show that there exists $\eta = \eta(|\mathcal{C}|) > 0$, such that for any distribution on P, Q , there exists a pair $\alpha, \beta \in \mathcal{C}$ with $|E_{\alpha,\beta}| > \eta$. Let $x_p = \mathbb{P}[P = p]$, $p \in 2^\mathcal{C} - \emptyset$; $y_q = \mathbb{P}[Q = q]$, $q \in 2^\mathcal{C} - \emptyset$, and let $\psi := \sum_{\alpha,\beta:\alpha \neq \beta} E_{\alpha,\beta}^2 \geq 0$. Note that ψ is a continuous function of the $\{x_p\}$ and $\{y_q\}$ (because $\psi = \sum_{\alpha,\beta} (\sum_{p,q} \lambda_{pq} x_p y_q)^2$ for suitable coefficients $\lambda_{pq} = 0, \pm 1$) and that $\{x_p, y_q\}$ are constrained to lie in the closed set C :

$$C = \left\{ \begin{array}{l} x_p \geq 0, \quad y_p \geq 0, \\ \sum_p x_p = \sum_p y_p = 1. \end{array} \right.$$

By continuity of ψ and closure of C , to establish the claim it suffices to show that ψ is never 0 on C ; i.e. we wish to show that for any distribution on P, Q , we can find a pair α, β such that $E_{\alpha,\beta} \neq 0$. We suppose not to derive a contradiction. Thus, suppose for some distribution that $E_{\alpha,\beta} = 0$ for all α, β . In particular

$$(17) \quad \begin{aligned} E_{\alpha,\beta} + E_{\beta,\alpha} = 0 &\Rightarrow \underbrace{\mathbb{P}[\alpha \in S_\beta] - \mathbb{P}[\alpha \in S_\alpha]} + \underbrace{\mathbb{P}[\beta \in S_\alpha] - \mathbb{P}[\beta \in S_\beta]} = 0 \\ &\Rightarrow \Delta_{\alpha,\beta} + \Delta_{\beta,\alpha} = 0. \end{aligned}$$

Now, for any choice $P = p, Q = q$, we have

$$(18) \quad \alpha \in S_\beta \Rightarrow \alpha \in S_\alpha$$

(because if $\alpha \notin S_\alpha$ we must have $\alpha \notin p \cup q$, and $\alpha \notin p \cup q \Rightarrow \alpha \notin S_\beta$). Consequently $\Delta_{\alpha,\beta} \leq 0$, and hence $\Delta_{\beta,\alpha} \leq 0$ by symmetry. Thus, from (17), $\Delta_{\alpha,\beta} = 0$, and in order to obtain a contradiction it suffices to find α, β such that $\Delta_{\alpha,\beta} < 0$. We distinguish two cases:

- (I) there exists $\alpha : \mathbb{P}[\alpha \in P] > 0$ and $\mathbb{P}[\alpha \notin Q] > 0$;
- (II) for all $\alpha : \mathbb{P}[\alpha \in P] = 0$ or $\mathbb{P}[\alpha \notin Q] = 0$.

In case (I), select β so that $\mathbb{P}[\alpha \notin Q, \beta \in Q] > 0$. Let E be the event $\alpha \in P, \alpha \notin Q, \beta \in Q$. Then $\mathbb{P}[E] > 0$, and

$$\mathbb{P}[\alpha \in S_\alpha | E] = 1, \quad \mathbb{P}[\alpha \in S_\beta | E] = 0.$$

By implication (18), $\mathbb{P}[\alpha \in S_\alpha \mid P = p, Q = q] \geq \mathbb{P}[\alpha \in S_\beta \mid P = p, Q = q]$ for all p, q , hence $\Delta_{\alpha,\beta} < 0$, which gives the required contradiction.

For case (II) we consider the possible subcases:

- (i) there exists $\alpha, \beta: \mathbb{P}[\alpha \in P] = 0, \mathbb{P}[\beta \notin Q] = 0$;
- (ii) for all $\alpha: \mathbb{P}[\alpha \notin Q] = 0$;
- (iii) for all $\alpha: \mathbb{P}[\alpha \in P] = 0$; (impossible!)

In case (i), $\mathbb{P}[\alpha \in S_\beta] = 0$ and $\mathbb{P}[\beta \in S_\beta] = 1$, so $E_{\beta,\alpha} \neq 0$, the required contradiction.

In case (ii), we have $\mathbb{P}[Q = \mathcal{C}] = 1$, so select any $\alpha \in P$ for which $\mathbb{P}[\alpha \in P] > 0$ to obtain $E_{\alpha,\beta} \neq 0$ for any $\beta \neq \alpha$. Case (iii) cannot arise. This completes the proof.

Proof of Theorem 3. An outline of the proof is as follows. We use the tree-chopping lemma to construct a family of comparably-sized disjoint subtrees of T , the sum of whose intrinsic parsimony lengths approximates $L(T)$ via Lemma 4. It is important that in chopping up T the component subtrees grow in size sufficiently quickly, but not as rapidly as their number. In this way, we can apply a version of the central limit theorem, due to Liapunov, for double arrays of random variables, and verify its hypotheses using Proposition 1 and Theorem 1.

The required central limit theorem (Serfling 1980) states the following. For each n , let X_{n1}, \dots, X_{nr} be $r = r(n)$ independent random variables with finite p th moments for some $p > 2$. Let

$$(19) \quad A_n = \sum_j \mathbb{E}[X_{nj}]; \quad B_n = \sum_j \text{Var} [X_{nj}].$$

If

$$(20) \quad B_n^{-p/2} \sum_j \mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

then $W_n = (\sum_j X_{nj} - A_n) / \sqrt{B_n}$ converges to the standard normal distribution, written $W_n \rightarrow N(0, 1)$, as $n \rightarrow \infty$.

We apply this theorem as follows. Firstly, use Lemma 3 with $k(n) = \lfloor n^\gamma \rfloor$ where $\gamma > 0.5$ to construct a leaf-covering forest $F = \{T_1, \dots, T_r\}$ for T which satisfies the constraints prescribed by Lemma 3 for $k = k(n)$. The number r of trees in F has the property that r/n^η is contained in a fixed positive interval for all trees, where $\eta = 1 - \gamma$. Let $X_{nj} = L(T_j)$, as in Lemma 4. Consider the two quantities

$$Z_T = \frac{L(T) - \mathbb{E}[L]}{\sqrt{V[L]}}, \quad W_T = \frac{\sum_j X_{nj} - A_n}{\sqrt{B_n}},$$

where A_n and B_n are given by (19). Then T_j has at most $2k(n) - 2$ leaves, and (3) of Theorem 1 shows that $\mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] = O(n^{\eta p/2})$. Also by the lower bound on the

variance given by Proposition 1, we have $B_n \geq r\delta k(n) > c'n$ for some constant $c' > 0$ independent of n . Thus,

$$B_n^{-p/2} \sum_j \mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] = O((c'n)^{-p/2} \times n^\eta n^{p/2}) = O(n^{\eta(1-p/2)}),$$

which converges to 0 as $n \rightarrow \infty$ when $p > 2$. Thus, condition (20) is satisfied. Furthermore, for each n , X_{n_1}, \dots, X_{n_r} are independent. Thus, the central limit theorem described above applies, and shows that $W_T \rightarrow N(0, 1)$. Now,

$$(21) \quad Z_T = \sqrt{\frac{B_n}{V[L]}} W_T + \frac{\Delta - \mathbb{E}[\Delta]}{\sqrt{V[L]}}$$

where Δ is defined by Lemma 4. Again, invoking the lower bound for the variance (given by Proposition 1), this time for $V[L]$, and the upper bound from Lemma 4, $|\Delta| < r = O(n^\eta)$, we see that the second term in (21) converges in probability to 0. Regarding the first term, note that, from the definition of Δ , we have:

$$V[L] = B_n + \text{Var} [\Delta] + 2 \text{Cov} \left[\Delta, \sum_j X_{nj} \right].$$

Applying the Cauchy-Schwarz inequality, and again using the bound $|\Delta| < r = O(n^\eta)$, we have:

$$V[L] = B_n + O(n^{2\eta}) + O(n^\eta \sqrt{B_n}),$$

so that $B_n/V[L]$ converges to 1 as $n \rightarrow \infty$ because $B_n > c'n$. Thus, $\sqrt{B_n/V[L]}$ converges to 1 in probability and so, because $W_T \rightarrow N(0, 1)$, we can apply Slutsky's lemma (see, e.g., Durrett (1991)) and deduce that $Z_T \rightarrow N(0, 1)$, as required.

4. The exact distribution and its mean and variance

Application of Theorem 1 depends on knowledge of $\mathbb{E}[L]$ and $V[L]$, the mean and variance of $L = L(T)$. In this section, we present efficient recursions for calculating these quantities, given T and its leaf distribution π . First, however, we give an algorithm that is polynomial in n for computing the exact distribution of $L = L(T)$.

Let $\mathbb{P}[T, \pi, k]$ denote the probability that $L(T) = k$. To obtain a recursive formula, we need more generally to consider, for each non-empty subset X of colors, the quantity

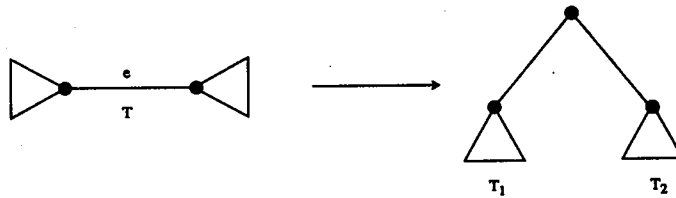
$$\mathbb{P}_X[T, \pi, k] = \mathbb{P}[L(T) = k \text{ and } S(T) = X].$$

Now consider the ordinary generating function:

$$F_X(T, \pi, x) = \sum_{k \geq 0} \mathbb{P}_X[T, \pi, k] x^k.$$

Subdivide an edge e of T , and let T_1, T_2 be the two rooted subtrees of T , whose

roots are adjacent to the root vertex on e . Let π_1, π_2 be the marginal distributions of π restricted to the leaves in T_1 and T_2 , respectively.



In view of Lemma 1 we have:

$$(22) \quad F_X(T, \pi, x) = \sum_{A \cap B = X} F_A(T_1, \pi^1, x) F_B(T_2, \pi^2, x) + \sum_{\substack{A \cap B = \emptyset \\ A \cup B = X}} x F_A(T_1, \pi^1, x) F_B(T_2, \pi^2, x),$$

because we add 0 or 1 to the sum of the lengths of T_1 and T_2 depending on whether $A \cap B \neq \emptyset$ or $A \cap B = \emptyset$, by the forward recursion version of Fitch's algorithm. The summation is over pairs A, B satisfying the stated conditions. Note that if T_1 has just one leaf i , then

$$F_A(T_1, \pi^1, x) = \begin{cases} \pi_i^1, & \text{if } A = \{i\}, \\ 0, & \text{if } |A| > 1. \end{cases}$$

Thus, one can efficiently calculate the polynomials $\{F_X(T, \pi, x) : X \neq \emptyset\}$ by starting from the leaves, working up to the root, and storing all the intermediate polynomials generated in the construction. Then $\mathbb{P}[T, \pi, k]$ is simply the coefficient of x^k in $\sum_{X \neq \emptyset} F_X(T, \pi, x)$.

Next we consider the complexity of this algorithm. For the subtree T_v below v , we must evaluate (22) $2^c - 1$ times, once for each $X \neq \emptyset$. The terms $F_A(T_1, \pi^1, x)$ and $F_B(T_2, \pi^2, x)$ can be multiplied in time proportional to $(\max L(T_1))(\max L(T_2))$ steps, which is less than or equal to the product of the number of leaves of the two trees. This product for tree T_v will be denoted by $n(v)$. The number of solutions to $A * B = X$ is

$$2^{|X|} + \sum_{i \geq 0} \binom{c - |X|}{i} 2^{c - |X| - i} = 2^{|X|} + 3^{c - |X|}.$$

Thus, $F_X(T, \pi, x)$ can be computed from $\{F_A(T_1, \pi^1, x), A \neq \emptyset\}$ and $\{F_B(T_2, \pi^2, x), B \neq \emptyset\}$ in $O((2^{|X|} + 3^{c - |X|})n(v))$ steps. Noting that $\sum_X 2^{|X|} + 3^{c - |X|} = 3^c + 4^c$, we see that $\{F_X(T_v, \pi, x), X \neq \emptyset\}$ can be obtained from the previous sets in $O(4^c n(v))$ steps. A straightforward inductive argument shows that, for any tree T with n leaves,

$$\sum_{\substack{v \in V[T] \\ \text{deg } v > 1}} n(v) \leq \binom{n}{2}.$$

(This upper bound is realized by a ‘caterpillar’ tree). Thus, $\{F_X(T_v, \pi, x), X \neq \emptyset\}$ can be computed in $O(4^c n^2)$ steps.

This recursive description allows the distribution of $L(T)$ to be effectively calculated, even when n is quite large (say 10^3). At each vertex, we must compute F_X for all non-empty subsets X . For c colors, this means $2^c - 1$ values of X . While this is not a problem for $c = 4$, it would be a problem for $c = 20$. In any case, this recursion appears to be of little help in determining what the limiting distribution is as $n \rightarrow \infty$.

Now, we describe a special algorithm designed to directly compute $\mathbb{E}[L] = \mathbb{E}[L(T)]$ and $V[L] = V[L(T)]$. First we compute $\mathbb{E}[L]$.

Subdivide an edge of T , root T at this newly created vertex, and direct all the edges of T away from this root. For any non-leaf vertex v of T , let T_v denote the subtree of T consisting of those vertices that are descendants of v . Note that T_v is a rooted binary tree, with v as its root. Let

$$P_X[T_v] = \mathbb{P}[S(T) = X].$$

Because v is not a leaf, it has two immediate descendent vertices v' and v'' , and then, by Lemma 1,

$$P_X[T_v] = \sum_{A \cup B = X} P_A(T_{v'}) P_B(T_{v''}).$$

Thus we can calculate the set $\Omega = \{P_X(T_v) : X \neq \emptyset, \text{deg}(v) > 1\}$, starting with the following initial conditions on the leaves:

$$P_X[i] = \begin{cases} \pi_i^\alpha, & \text{if } X = \{\alpha\}, \\ 0, & \text{if } |X| > 1. \end{cases}$$

Constructing Ω requires $O(n4^c)$ steps, as T has $O(n)$ vertices. Now, by Lemma 1,

$$\mathbb{E}[L] = \sum_{\text{deg}(v) > 1} \sum_{A \cup B = \emptyset} P_A(T_{v'}) P_B(T_{v''}),$$

so that $\mathbb{E}[L]$ can be calculated in a further $O(n3^c)$ steps from the set Ω constructed above; thus $\mathbb{E}[L]$ is computable in $O(n4^c)$ steps.

Recall that $L = L' + L'' + D$, where

$$D = \begin{cases} 1, & \text{if } S(T') \cap S(T'') = \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Now L' and L'' are independent and

$$(23) \quad V[L] = V[L'] + V[L''] + 2 \text{Cov}(D, L' + L'') + V[D].$$

The quantity $V[D]$ is handled next. Since D has values 0 and 1, $V[D] = \mathbb{P}[D = 1]\{1 - \mathbb{P}[D = 1]\}$, and

$$\mathbb{P}[D = 1] = \sum_{A \cap B = \emptyset} \mathbb{P}[S(T') = A] \mathbb{P}[S(T'') = B].$$

The probabilities in the above summand are computed in the algorithm for $\mathbb{E}[L]$. $V[D]$ is now computable in $O(3^c)$ additional steps because

$$\sum_i \binom{c}{i} 2^{c-i} = 3^c.$$

Now we consider the covariance term in (23):

$$\text{Cov}(D, L' + L'') = \mathbb{E}[DL'] + \mathbb{E}[DL''] - \mathbb{E}[D](\mathbb{E}[L'] + \mathbb{E}[L'']).$$

The terms not already considered are $\mathbb{E}[DL']$ and $\mathbb{E}[DL'']$. Now,

$$\begin{aligned} \mathbb{E}[DL'] &= \sum_k k \mathbb{P}[L' = k \cap D = 1] \\ &= \sum_{A \cap B = \emptyset} \sum_k k \mathbb{P}[L' = k \cap \{S(T') = A\} \cap \{S(T'') = B\}] \\ &= \sum_{A \cap B = \emptyset} \mathbb{P}[S(T'') = B] \sum_k k \mathbb{P}[L' = k \cap \{S(T') = A\}] \\ &= \sum_{A \cap B = \emptyset} \mathbb{P}[S(T'') = B] F(T', A), \end{aligned}$$

where $F(T', A)$ is defined by the last equation.

$$\begin{aligned} F(T, A) &= \sum_k k \mathbb{P}[\{L = k\} \cap \{S(T) = A\}] \\ &= \sum_{B \bullet C = A} \sum_{k_1, k_2} (k_1 + k_2 + \mathbb{I}(B \cap C = \emptyset)) \\ &\quad \times \mathbb{P}[\{L_1 = k_1\} \cap \{S(T') = B\}] \mathbb{P}[\{L_2 = k_2\} \cap \{S(T'') = C\}]. \end{aligned}$$

Breaking the sum into three parts, we obtain

$$\begin{aligned} F(T, A) &= \sum_{B \bullet C = A} \{F(T', B) \mathbb{P}[S(T'') = C] + F(T'', C) \mathbb{P}[S(T') = B] \\ &\quad + \mathbb{I}(B \cap C = \emptyset) \mathbb{P}[S(T') = B] \mathbb{P}[S(T'') = C]\}. \end{aligned}$$

Clearly this last equation allows us to recursively compute $\mathbb{E}[DL']$, and similarly $\mathbb{E}[DL'']$. Hence $\text{Cov}(D, L' + L'')$, and (23) for $V[L]$ can be computed. The number

of pairs B, C such that $B * C = A$ is $2^{|A|} + 3^{c-|A|}$, and therefore $V[L]$ can be computed in time $O(4^n)$.

We collect the results of this section in the following theorem.

Theorem 4. Let T be any binary tree with n leaves, colored from a set of c colors. Then the exact distribution of $L(T)$ may be computed in $O(n^2 4^c)$ steps, and the mean and variance of $L(T)$, $E[L]$ and $V[L]$, may be computed in $O(n 4^c)$ steps.

5. An example

Consider a ‘bush’ of height h , i.e. any binary tree T_h for which the path from any leaf to the root contains exactly h edges, and take the leaves of T_h to be colored from the set $\{\alpha, \beta\}$ with $\pi_i^\alpha = \pi$, where $1 > \pi > \frac{1}{2}$. Clearly, by Theorem 3, the normal limit holds in this case. However, in this section, we show that this is an example for which the vertex set $S(T_h)$ is degenerate as the size of the tree tends to infinity; in particular,

$$\lim_{h \rightarrow \infty} \mathbb{P}[S(T_h) = \{\alpha\}] = 1.$$

This behavior is in marked contrast to the case $\pi_i^\alpha = \frac{1}{2}$, where, from Charleston and Steel (1995),

$$\mathbb{P}[S(T_h) = \{\alpha\}] = \mathbb{P}[S(T_h) = \{\beta\}] = \frac{1}{3}[1 - (-0.5)^h],$$

and so

$$\lim_{h \rightarrow \infty} \mathbb{P}[S(T_h) = X] = \frac{1}{3} \text{ for } X = \{\alpha\}, \{\beta\}, \{\alpha, \beta\}.$$

Thus, the distribution of $S(T_h)$ can be asymptotically degenerate even when the leaf coloration distributions are i.i.d. and bounded away from zero (i.e. satisfy condition (8)).

In order to substantiate our claim, write $\mathbb{P}[S(T_h) = X]$ as $p_X(h)$. Deleting the root and its two incident edges from T_h gives two bushes of height $h - 1$, thereby providing the following system of simultaneous recursions:

$$\begin{aligned} p_{\{\alpha\}}(h) &= p_{\{\alpha\}}^2(h - 1) + 2p_{\{\alpha\}}(h - 1)p_{\{\alpha, \beta\}}(h - 1), \\ p_{\{\beta\}}(h) &= p_{\{\beta\}}^2(h - 1) + 2p_{\{\beta\}}(h - 1)p_{\{\alpha, \beta\}}(h - 1), \\ p_{\{\alpha, \beta\}}(h) &= p_{\{\alpha, \beta\}}^2(h - 1) + 2p_{\{\alpha\}}(h - 1)p_{\{\beta\}}(h - 1). \end{aligned}$$

Now, consider $D(h) = p_{\{\alpha\}}(h) - p_{\{\beta\}}(h)$. We have

$$D(h) = D(h - 1)[p_{\{\alpha\}}(h - 1) + p_{\{\beta\}}(h - 1) + 2p_{\{\alpha, \beta\}}(h - 1)].$$

But now, $p_{\{\alpha\}}(h-1) + p_{\{\beta\}}(h-1) = 1 - p_{\{\alpha,\beta\}}(h-1)$, and so,

$$(24) \quad D(h) = D(h-1)[1 + p_{\{\alpha,\beta\}}(h-1)].$$

Thus $1 \geq D(h) \geq D(h-1) \cdots \geq D(1) > 0$, and hence $D(h)$ converges to a positive limit. Hence, using (24), $p_{\{\alpha,\beta\}}(h)$ converges to zero. Therefore, the third equation of the recursion shows that $p_{\{\alpha\}}(h)p_{\{\beta\}}(h) \rightarrow 0$. Since $D(h)$ has a positive limit, this forces $p_{\{\beta\}}(h) \rightarrow 0$ and $p_{\{\alpha\}}(h) \rightarrow 1$.

6. Discussion

Theorems 1 and 3 apply to one position of n aligned homologous sequences. The common assumption in phylogenetic analysis is that positions are i.i.d. For aligned sequences, the criterion is the sum of parsimony scores over all positions. In this case, the central limit theorem for i.i.d. random variables applies. Our Theorem 3 shows that the individual terms making up this sum are themselves approximately normal, hence an excellent fit of this sum to the normal is expected. In contrast, Theorem 2 gives large deviation bounds for the tree that attains the minimum length (maximum parsimony) score over k positions of n aligned homologous sequences where we have no knowledge that such a central limit theorem applies.

We note finally that the asymptotic normal limit distribution does not hold in general for non-binary trees. For a simple counterexample, take the star tree, i.e. take the tree with $n+1$ vertices and n edges all of the form $\{v_0, v\}$ for a distinguished (center) vertex v_0 . For the star tree, the minimum coloration extending a given leaf coloration is the one in which the center vertex is colored the most frequent color; the length L is therefore the number of leaves not colored with the most frequent color. Hence, in the case of two colors α and β , for $k \leq n/2$, $L = k$ when there are k or $n-k$ leaves with color α . Assigning each color with equal probability at each leaf, $P[L = k] = \binom{n}{k} 2^{1-n}$ if $0 \leq k \leq n/2$ and 0 otherwise. As this probability decreases from its maximum value to 0 when increasing $[n/2]$ by 1 to $[n/2] + 1$, the distribution of L cannot converge to the normal.

Acknowledgements

We thank J. Michael Steele for describing the martingale-style approach for bounding variances from below, used in the proof of Proposition 1 (personal communication). We are extremely grateful to an anonymous referee for his extraordinary care and patience in reading and improving earlier versions of our manuscript, and for many insightful suggestions.

References

- ALON, N. AND SPENCER, J. H. (1992) *The Probabilistic Method*. Wiley, New York.
 ARCHIE, J. AND FELSENSTEIN, J. (1993) The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theoret. Pop. Biol.* **43**, 52-79.

- CARTER, M., HENDY, M., PENNY, D., SZEKELY, L. A. AND WORMALD, N. C. (1990) On the distribution of lengths of evolutionary trees. *SIAM J. Disc. Math.* **3**, 38–47.
- CAVALLI-SFORZA, L. L. AND EDWARDS, A. W. F. (1967) Phylogenetic analysis: Models and estimation procedures. *Amer. J. Hum. Genet.* **19**, 233–257.
- CHARLESTON, M. AND STEEL, M. A. (1995) Five surprising properties of parsimoniously colored trees. *Bull. Math. Biol.* **57**, 367–375.
- DURRETT, R. (1991) *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Belmont, CA.
- FITCH, W. M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416.
- HARTIGAN, J. A. (1973) Minimum mutation fits to a given tree. *Biometrics* **29**, 53–65.
- MOON, J. W. AND STEEL, M. A. (1993) A limiting distribution for parsimoniously bicolored trees. *Appl. Math. Lett.* **6**, 5–8.
- SERFLING, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- STEEL, M. A. (1993a) Decompositions of leaf-coloured binary trees. *Adv. Appl. Math.* **14**, 1–24.
- STEEL, M. A. (1993b) Distributions on bicoloured binary trees arising from the principle of parsimony. *Discr. Appl. Math.* **41**, 245–261.
- STEELE, J. M. (1986) An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14**, 753–758.