Pergamon

*Bulletin of Mathematical Biology*, Vol. 57, No. 2, pp. 367–375, 1995
Elsevier Science Ltd

0092-8240/95 $9.50+0.00

0092-8240(94)00051-4

# FIVE SURPRISING PROPERTIES OF PARSIMONIOUSLY COLORED TREES

■ MIKE STEEL and MIKE CHARLESTON
Department of Mathematics,
Massey University,
Palmerston North,
New Zealand

(*Email: mas@math.canterbury.ac.nz*)

Trees with a coloration of their leaves have an induced "length" which forms the basis of the widely used maximum parsimony method for reconstructing evolutionary trees in biology. Here we describe five unexpected properties of this length function, including refinements of earlier results.

**1. Introduction.** A popular method for reconstructing phylogenetic trees from aligned sequence data is the selection of the tree of maximum parsimony (Felsenstein, 1988). In this method each site in the aligned sequences induces a positive valued "length" on each tree, and the maximum parsimony tree(s) is that which minimizes the sum of these lengths across the sites in the sequences. In order to study the statistical properties of this method it is useful to investigate aspects of this length function (Maddison and Slatkin, 1991; Goloboff, 1991; Steel, 1993). Here we describe five properties of this length function (four of which give new and exact formulae) which may, at first, seem counterintuitive.

We begin with some definitions. A *rooted binary tree* is a tree whose vertices are all of degree 1 or 3, and with one vertex of degree 2 called the *root* (for other graph theroretic terminology see Bondy and Murty, 1976). Vertices of degree 1 are called *leaves*. The two *descendants* of a vertex $v$ are those vertices $v_d$ which are adjacent to $v$ and such that $v$ lies on the path from $v_d$ to the root.

For any tree $T$ with each of its vertices assigned one color chosen from a set $S$ of colors, the *changing number* of this coloration is the number of edges which have different colors at their ends. If only the leaves of $T$ are colored, the *length* of this leaf coloration is the smallest value of the changing number across all colorations of the vertex set of $T$ that extend the leaf coloration (a coloration which has minimal changing number is called a *minimal coloration*). This length can be found in linear time by Fitch's algorithm (Fitch, 1971). This procedure also finds the set of colors which can be assigned to each vertex under at least

one minimal coloration (for further details see Hartigan (1973) and Erdös and Székely (1992)). To describe the algorithm for binary trees it is convenient to recall Fitch's original parsimony operation, $*$, which is a commutative non-associated binary operation defined on the nonempty subsets of a finite set. Specifically,

$$A * B = \begin{cases} A \cap B, & \text{if } A \cap B \neq \varnothing, \\ A \cup B, & \text{otherwise.} \end{cases}$$

Suppose a rooted binary tree $T$ has its leaves assigned colors from a set $X$. Fitch's alogrithm recursively colors the vertices of $T$ with nonempty *subsets* of $X$ starting with the leaves and progressing towards the root as follows: each leaf is assigned the singleton set consisting of its assigned color, while for each other vertex $v$, once the descendants of $v$ have been assigned subsets, say $A$ and $B$, then $v$ is assigned the set $A * B$. Eventually every vertex will be assigned a set, including the root, $\rho$, whose set we denote $X_\rho$. Hartigan (1973) showed that:

(1) the length of the leaf coloration is the number of times the empty intersection option for $*$ is taken in assigning sets to the vertices of $T$; and

(2) $X_\rho$ is the set of colors which can be assigned to the root in at least one minimal coloration.

**2. Results.** For a rooted binary tree $T$ consider the number of ways to color the leaves of $T$ either 0 or 1 so that the resulting leaf coloration has length $k$, and so that $X_\rho = \{0\}$, $\{1\}$ or $\{0, 1\}$. Call these numbers $N_0(k)$, $N_1(k)$ and $N_{01}(k)$, respectively.

It might be expected that these numbers should depend on the shape of $T$, and this is indeed the case if the leaves are $r$-colored (for $r > 2$); however, for bicolorings, the quantity depends, surprisingly, only on the number of leaves. We derive exact expressions for this and related quantities, thereby extending earlier results described in Steel (1993).

THEOREM 1. *For a rooted binary tree* T *with* n *leaves,*

$$N_0(k) = N_1(k) = \binom{n-k-1}{k} 2^k, \qquad N_{01}(k) = \binom{n-k-1}{k-1} 2^k.$$

*In particular, these quantities do not depend on the shape of* T.

*Proof.* It is convenient to work, initially, with binary trees, in which the root $\rho$ has degree 1. Thus, for $S = \{0\}$, $\{1\}$, $\{0, 1\}$, and a tree $T$ with a root vertex of degree 1, let $N_S(T, k)$ denote the number of ways to color the non-root leaves of $T$ so that the resulting leaf coloration has length $k$ and so that $X_\rho$ is $S$, where $X_\rho$ is the set of colors which the root vertex can take in at least one minimal coloration of $T$. By symmetry, this value is the same for $S = \{0\}$ or $S = \{1\}$.

Provided the number $n$ of non-root leaves of $T$ is at least 2, there is a pair $\{i, j\}$ of non-root leaves of $T$ which are adjacent to a common vertex $v_{ij}$. Delete leaves $i$ and $j$ and their incident edges from $T$, and let $T'$ be the resulting tree, in which $v_{ij}$ is a leaf. If $n>2$, delete from $T'$ the leaf vertex $v_{ij}$ and its incident edge, to obtain a second tree $T''$. The root leaf of $T$ provides a root leaf for $T'$ and $T''$.

We show that, for $n>2$,

$$N_S(T, k)=N_S(T', k)+2N_S(T'', k-1). \tag{1}$$

We distinguish the possible leaf colorations of $T$ of length $k$ for which $X_\rho=S$ into two types; those which color $i$ and $j$ the same, and those which assign different colors to $i$ and $j$. Note that each type 1 coloration corresponds to a unique leaf coloration of $T'$ (we color $v_{ij}$ the color shared by $i$ and $j$, and the remaining vertices are colored as in $T$) which, by Fitch's algorithm, has the same length as the original coloration, and for which $X_\rho=S$. This association is one-to-one. Also, each type 2 coloration of length $k$ gives a leaf coloration of $T''$ (by restriction) which, by Fitch's algorithm, is both of length $k-1$, and has $X_\rho=S$. This association is, however, two-to-one. Combining these correspondences gives equation (1).

Note that for $n=2$, $N_{\{0\}}(T, k)=N_{\{0\}}(T', k)$; however $N_{\{0,1\}}(T, k)\neq N_{\{0,1\}}(T', k)$. In fact $N_{\{0,1\}}(T, k)$ is 2 if $k=1$ and is 0 otherwise. For $n=1$, $N_{\{0\}}(T, k)$ is 1 if $k=0$, and is 0 otherwise. It is convenient to incorporate these boundary conditions into equation (1), by formally defining $N_S(T', k)$ and $N_S(T'', k)$ to be zero whenever $n=1$ or 2, respectively. Then, for all $n\geqslant 1$,

$$N_S(T, k)=N_S(T', k)+2N_S(T'', k-1)+d_S(T, k),$$

where
$$d_S(T, k)=\begin{cases} 2 \text{ if } n=2,\ S=\{0, 1\}, \text{ and } k=1; \\ 1 \text{ if } n=1,\ S=\{0\} \text{ or } \{1\}, \text{ and } k=0; \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

By induction on $n$, $N_S(T, k)$ depends only on $S$, $n$ and $k$, and not on the shape of $T$. Thus we can let $P_S(x, y$ be the ordinary generating function for $N_S(T, k)$ in which $y$ marks $k$ and $x$ marks the number of leaves of $T$ (for background refer to Goulden and Jackson, 1983). In this way, equation (2) can be rewritten:

$$P_S(x, y)=xP_S(x, y)+2x^2yP_S(x, y)+d_S(x, y), \tag{3}$$

where $d_S(x, y)$ is $x$ if $S=\{0\}$, and is $2x^2y$ if $S=\{0, 1\}$.

Solving equation (3) gives

$$P_S(x, y)=\frac{d_S(x, y)}{(1-x-2x^2y)} \tag{4}$$

and so $N_S(T, k)$ can then be routinely found as the coefficient of $x^n y^k$ in this expression.

Finally, although we have been dealing with trees with a degree 1 root, if we delete this root leaf and its incident edge, the resulting rooted binary tree $T^*$ has the same length and the same value of $X_\rho$, so that $N_S(T, k) = N_S(T^*, k)$, completing the proof.

COROLLARY 1. (Steel, 1989, 1993) *For a binary tree with* n *leaves, the number of leaf bicolorations of length* k *is*

$$\frac{(2n-3k)}{k}\binom{n-k-1}{k-1}2^k$$

*if* $k>0$, *and* 2 *if* $k=0$.

As a second corollary, we describe a number of related quantities which again, surprisingly, are independent of the shape of the tree. Given a rooted binary tree $T$, suppose we randomly assign characters (0 or 1) to the leaves with equal probability. Denote the probability that $X_\rho = \{0\}$, $\{1\}$ and $\{0, 1\}$ by $P_0(T)$, $P_1(T)$ *and* $P_{01}(T)$, respectively. Similarly, let $E_0(T)$, $E_1(T)$ and $E_{01}(T)$ denote the expected length of the coloration, given that $X_\rho = \{0\}, \{1\}$ and $\{0, 1\}$, respectively.

COROLLARY 2. *If* T *has* n *leaves,*

$$P_0(T) = P_1(T) = \frac{1}{3}\left(1-\left(\frac{-1}{2}\right)^n\right), \quad \textit{and} \quad P_{01}(T) = \frac{1}{3}\left(1+2\left(\frac{-1}{2}\right)^n\right) \qquad \text{(i)}$$

$$E_0(T) = E_1(T) = \frac{3n-4+(3n-2)(-0.5)^{n-1}}{9(1-(-0.5)^n)} \qquad \text{(ii)}$$

$$E_{01}(T) = \frac{3n+2-(6n-1)(-0.5)^n}{9(1+2(-0.5)^n)}.$$

*Proof.* These follow from equation (4) in the proof of Theorem 1: for (i) we simply put $y=1$ and extract the coefficient of $x^n$ and divide by $2^n$, the total number of leaf bicolorations. For (ii) note that

$$E_S(T) = \sum_k k\,\frac{N_S(T, k)}{2^n P_S(T)},$$

since the quotient term is the conditional probability that the length of a

random leaf bicoloration is $k$ given that $X_\rho = S$. We thus apply $\partial/\partial y|_{y=1}$ to equation (4), extract the coefficient of $x^n$, then apply to part (1) to obtain $E_S(T)$.

REMARK. From Corollary 2, the expected length $E(T)$ of an random leaf bicolaration of $T$ with $n$ leaves can be readily recovered. The result (as in Steel (1989, 1993)) is

$$E(T) = \frac{(3n-2-(-0.5)^{n-1})}{9}.$$

Goloboff (1991) considered a quantity ($S_T$) which is easily shown to be equal to $E(T)$ and for which he derived an extraordinarily complicated recursive formula. Independently, both Maddison and Slatkin (1991), and Archie and Felsenstein (1993) found a simpler expression for $E(T)$: however, once again this was a recursive formula.

Next, we demonstrate a simple but surprising example which shows that grouping the colors (something which occurs in phylogenetic analysis) can lead to different conclusions as to which is the most parsimonious color for the root of a tree. A biologically pertinent example is provided by mapping the set of the four DNA nucleotide {A, G, C, T} onto the set of {R, Y}, where R = {A, G} (adenosine and guanine, the purines) and Y = {C, T} (cytosine and thymine, the pyrimidines). The mapping $\phi$ is then defined by:

$$\phi A = \phi G = R, \qquad \phi C = \phi T = Y.$$

Given a tree $T$ whose leaves are labelled with nucleotides, Fitch's algorithm gives the set of possible (parsimonious) root colors $X_\rho \subseteq \{A, C, G, T\}$. Mapping the nucleotides at the leaves onto the purines and pyrimidines, and using the Fitch algorithm again, gives the corresponding set of (parsimonious) root colors $(\phi X)_\rho \subseteq \{R, Y\}$. For the example in Fig. 1 we have $(\phi X)_\rho \cap \phi(X_\rho) = \varnothing$. Thus, when the character states are coded as nucleotides all minimal colorations of the tree assign a pyrimidine (C) to the root, while when we code them as purine/pyrimidine, we always get a purine at the root!

Our next two examples concern the fully bifurcating tree, $T_k$, i.e. the rooted binary tree which has height $k$ and $n = 2^k$ leaves (see Fig. 2).

Suppose each leaf is assigned a color from $\{0, 1, \ldots, r-1\}$. We ask the question, what is the smallest proportion of leaves which must be colored 0 in a leaf coloration for which all the associated minimal colorations color the root 0? We determine this exactly for $r = 2$, and show that, somewhat surprisingly, this proportion can be made arbitrarily small for trees of sufficient height.

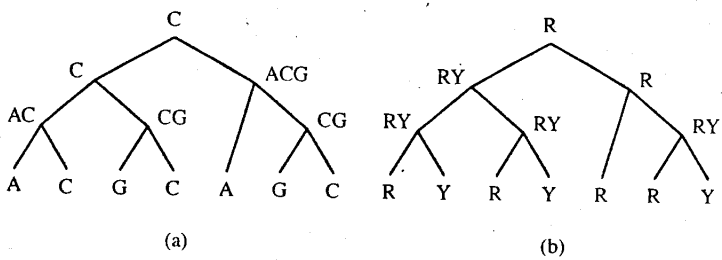


Figure 1. In (a) is shown a tree, each of whose leaves are assigned one of A, C or G, and the states of the remaining vertices are assigned according to Fitch's algorithm. In (b) the states A and G are replaced by R (purine), and C is replaced by Y (pyridine), and the states of the vertices are assigned with Fitch's algorithm once more. Note that (a) the root is assigned C, a pyrimidine, whereas in (b) the root is assigned purine.

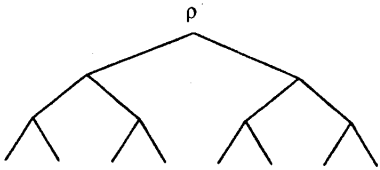


Figure 2. $T_3$ is a fully bifurcating tree of height 3 and with $2^3 = 8$ leaves.

THEOREM 2. *For a fully bifurcating tree of height* k, *the minimum number of leaves which need to be colored* 0 *in a leaf bicoloration for which* $X_\rho = \{0\}$ *equals the* $(k+1)$*th Fibonacci number.*

*Proof.* Let us first consider, more generally, leaf colorations of $T_k$ by elements of the set $\{0, 1, \ldots, r-1\}$. Provided $r \leqslant 2^k$, it is well defined to let $f_k^A$ denote the minimum number of leaves of $T_k$ which must be colored 0 so as to allow $X_\rho = A$, where $A \subseteq \{0, 1, \ldots, r-1\}$. Let $f_k = f_k\{0\}$. Clearly,

$$f_k^A = 0 \text{ if } 0 \notin A. \tag{i}$$

$$\text{if } 0 \in A \cap B, \text{ and } |A| = |B|, \text{ then } f_k^A = f_k^B. \tag{ii}$$

Also, by the standard decomposition of rooted binary trees (Goulden and Jackson, 1983), and Fitch's algorithm,

$$f_{k+1}^A = \min\{f_k^B + f_k^C : B * C = A\}, \tag{iii}$$

where $*$ is the parsimony operation, described earlier.

For $r=2$, (iii) becomes

$$f_{k+1}=\min\{f_k^{\{0,1\}}+f_k,\ 2f_k\},$$
$$f_{k+1}^{\{0,1\}}=\min\{2f_k^{\{0,1\}},\ f_k+f_k^{\{1\}}\}=\min\{2f_k^{\{0,1\}},\ f_k\} \tag{iv}$$

since $f_k^{\{1\}}=0$ by (i)). We first apply (iv) to prove by induction that, for all $k\geqslant 1$,

$$f_k^{\{0,1\}}\leqslant f_k \tag{I1}$$

$$f_k\leqslant 2f_k^{\{0,1\}}. \tag{I2}$$

These statements are true for $k=1$. Suppose they hold for $k=j$, then (iv) becomes

$$f_{j+1}=f_j^{\{0,1\}}+f_j,\quad f_{j+1}^{\{0,1\}}=f_j \tag{v}$$

by (I1) and (I2), respectively.

Thus, we see that $f_{j+1}^{\{0,1\}}\leqslant f_{j+1}$ so that (I1) hold for $k=j+1$. Applying (I1) to the equations in (v) gives $f_{j+1}=f_j^{\{0,1\}}+f_j\leqslant 2f_j=2f_{j+1}^{\{0,1\}}$ so that (I2) holds for $k=j+1$. Thus (I1) and (I2) hold for all $k$, and so therefore does (v). Combining the equations in (v) gives $f_k=f_{k-1}+f_{k-2}$, which together with the initial conditions, $f_1=2, f_2=3$, shows that $f_k$ is the $(k+1)$th Fibonacci number. That is,

$$f_{n-2}=\frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right).$$

Since $\lim_{k\to\infty} f_k/2^k=0$, the proportion of leaves which need to be assigned color 0 in order to force $X_\rho=\{0\}$ can be made arbitrarily small, as claimed.

We conjecture that for $r\geqslant 2$ colors,

$$f_k=\begin{cases} f_{k-p}+f_{k-p-1}, & \text{when } r=2p,\\ 2f_{k-p}, & \text{when } r=2p-1.\end{cases}$$

In our final example, consider the fully bifurcating tree $T$ with $n=2^k$ leaves, in which color 0 is assigned to the root. Randomly bicolor the remaining verticles of the tree, progressing from the root down towards the leaves, in the following way: once a vertex $v$ has been colored, assign a descendant of $v$ the same color with probability $1-p\geqslant\frac{1}{2}$, and different color with probability $p\leqslant\frac{1}{2}$ ($p$ fixed). In this way we generate a random leaf bicoloration (the resulting probability distribution on the leaf bicolorations is important in phylogenetic studies and was investigated for general trees, and $p$ not fixed, by Cavender (1978)). Now, some of these generated random lead bicolorations will be such that (when we apply Fitch's algorithm) they give $X_\rho=\{0\}$. Let us denote the probability of generating such a leaf bicoloration by $S(k)$. That is, $S(k)$ is the

probability that staring with state 0, a leaf coloration is generated for which all the resulting minimal colorations assign the root its true state, 0. Similarly let $D(k)$ and $U(k)$ denote the probabilities of generating the leaf colorations for which $X_\rho = \{1\}$ and $X_\rho = \{0, 1\}$, respectively.

We are interested in what happens to these values as $k \to \infty$. Intuitively, there are two opposing factors which make the limiting behaviour of these values unclear: as the height of the tree becomes large the probability that a leaf is colored 0 or 1 will tend to 0.5 (suggesting that $S(k)$ and $D(k)$ might approach a common value), however the number of leaves is growing exponentially, and each leaf is more likely to be 0 than 1 (though by a difference tending to 0) suggesting that $S(k)$ might tend to a value strictly larger than the limiting value of $D(k)$. Surprisingly, the question of which factor wins depends on whether $p \geqslant \frac{1}{8}$ or $p < \frac{1}{8}$. Let $s$, $d$ and $u$ denote the limiting values of $S(k)$, $D(k)$ and $U(k)$, respectively. A proof (and an appliction to analyse the statistical consistency of the maximum parsimony method of tree reconstruction) appears in Steel (1989) of the following.

THEOREM 3. *If* $\mathrm{p} \geqslant \frac{1}{8}$, *then* $\mathrm{s} = \mathrm{d} = \mathrm{u} = \frac{1}{3}$. *If* $\mathrm{p} < \frac{1}{8}$, *then*

$$\mathrm{s} = \frac{1}{2}\left(1 - 2\mathrm{x} + \frac{\sqrt{\Delta}}{1-2\mathrm{p}}\right), \quad \mathrm{d} = \frac{1}{2}\left(1 - 2\mathrm{x} - \frac{\sqrt{\Delta}}{1-2\mathrm{p}}\right), \quad \textit{and } \mathrm{u} = 2\mathrm{x},$$

*where* $\mathrm{x} = \mathrm{p}/(1-2\mathrm{p})$ *and* $\Delta = (1-6\mathrm{x})(1-2\mathrm{x})$.

Note that $s \to 1$ as $p \to 0$, and that the values when $p = \frac{1}{2}$ are also given by Corollary 2(i).

**3. Conclusion.** Motivated by a desire to better understand the properties of the maximum parsimony method for tree reconstruction, we have derived a number of exact and explicit formulae involving the length function. It turns out that these formulae have some surprising aspects, such as being independent of tree shape, or possessing unexpected asymptotic behaviour. Such formulae provide a more efficient and accurate way than computer-based simulation to estimate statistical properties of the parsimony method under the null model of random data; the results from Theorem 1 also provide a considerable simplification of previous published formulae for the expected length of a random leaf bicoloration (as described in the Remark following Corollary 2). The exact formulae apply only for leaf bicolorations. It would be useful to find similar results for leaf $r$-colorations, for $r \geqslant 2$, in particular for $r = 4$, however, the analogue of Theorem 1 would then have to take into account the shape of the tree.

# LITERATURE

Archie, J. W. and J. Felsenstein. 1993. The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theor. Pop. Biol.* **43**, 52–79.

Bondy, J. A. and U. S. R. Murty. 1976. *Graph Theory with Applications*. Macmillan, London.

Cavender, J. 1978. Taxonomy with confidence. *Math. Biosci.* **40**, 271–280.

Erdös, P. L. and L. A. Székely. 1992. Evolutionary trees: an integer multicommodity max-flow-min-cut theorm. *Adv. appl. Math.* **13**, 375–389.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Ann. Rev. Genet.* **22**, 521–565.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416.

Goloboff, P. A. 1991. Homoplasy and choice among cladograms. *Cladistics* **7**, 215–232.

Goulden, I. P. and D. M. Jackson. 1983. *Combinatorial Enumeration*. Wiley, New York.

Hartigan, J. A. 1973. Minimum mutation fits to a given tree. *Biometrics* **29**, 53–65.

Maddison, W. P. and M. Slatkin. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* **45**, 1184–1197.

Steel, M. A. 1989. Distributions in bicoloured evolutionary trees. Ph.D. thesis. Massey University, Palmerston North, New Zealand.

Steel, M. A. 1993. Distributions on bicolored binary trees arising from the principle of parsimony. *Discrete Appl. Math.* **41**, 245–261.

Received 1 April 1993
Accepted 31 August 1993