Discrete Applied Mathematics 41 (1993) 245-261
North-Holland

Distributions on bicoloured binary trees arising from the principle of parsimony

M.A. Steel

Department of Mathematics and Statistics, Massey University, Palmerston North, New Zealand

Received 7 June 1989
Revised 14 December 1990

Abstract

Steel, M.A., Distributions on bicoloured binary trees arising from the principle of parsimony, Discrete Applied Mathematics 41 (1993) 245–261.

The distribution of binary trees with bicoloured endpoints under the taxonomic principle of parsimony is examined. Part one provides a constructive proof of an expression which describes the distribution of binary trees for a fixed colouring in terms of simple tree-related quantities. The result relies on Menger's theorem and an invariance property of binary trees. In part two a second invariance property gives the dual distribution, where the tree is fixed and the colourings vary. Applications to taxonomy, and the extension of results to r-colourings ($r>2$) are outlined briefly.

Keywords. Binary tree, forest, minimum-length tree, Menger's theorem.

1. Introduction

A main taxonomic method for reconstructing evolutionary trees from genetic and protein sequence data is the minimum-length tree method, based on the principle of *parsimony* (see, for example, Felsenstein [6]).

Essentially, one regards aligned molecular sequence data D of length c as sequences $D_1, \ldots, D_c$ of r-colourings of the taxa set (in applications $r=2,4,20$). Taking $\{1, \ldots, n\}$ as the taxa set, each binary tree, T, whose endpoints are labelled from the set $\{1, \ldots, n\}$ represents a hypothetical hierarchical relationship between the taxa, and each D_i induces a colouring of the endpoints of T. The *weight* of D_i on

Correspondence to: Dr. M.A. Steel, Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, 4800 Bielefeld, Germany.

T, denoted $w(D_i, T)$, is then the minimum number of edges of T which must be assigned differently-coloured ends so as to extend D_i to a colouring of all the vertices of T. This weight, and a minimal colouring extension can be found by an $O(n)$ algorithm due to Fitch [7], which has been rigorously justified by Hartigan [11]. The weight of D on T, denoted $w(D, T)$ is then defined as the sum of the weights $w(D_i, T)$ for $i = 1, \ldots, c$. This weight is regarded as an inverse measure of how well T fits the data D. The principle of parsimony is to select the binary tree(s) which best fits the data—that is which minimizes $w(D, T)$—to estimate the underlying evolutionary tree linking the species. Such trees are referred to as *minimum-length trees*.

In considering the significance of the weight of the minimum-length tree it is useful to consider

(a) how well, compared with other trees, a minimum-length tree fits given data, and

(b) how well the data, compared with other data sets, fits any binary tree; that is, how "tree-like" is the data?

Variations on the second question have been considered by Archie [1]; and Henderson, Hendy and Penny [12]; and Penny, Foulds and Hendy [15]. The last of these devised a simple test of whether different sets of data fit a common binary tree and applied this test in a study involving 11 taxa. Using a different approach the authors of [1] and [12] developed tests of the tree-likeness of data, using respectively simulation and computational methods.

For an analytical approach to questions (a) and (b) we consider respectively two dual distributions (in $k \geq 0$):

(a′) The number $N(D, c, k)$ of binary trees having weight k for fixed sequences D of length c.

(b′) The number $N^*(T, c, k)$ of sequences of length c having weight k for a fixed binary tree T.

Now $N(D, c, k)$ depends in a complex way on D, indeed calculating the smallest value of k for which $N(D, c, k) \neq 0$ is NP-complete in n (see, for example, Foulds and Graham [8]).

However the mean value, over k, of $N(D, c, k)$ can be readily calculated from the distributions $N(D_i, 1, k)$, giving a simple measure of how much better a minimum-length tree fits D than a "randomly-chosen" binary tree.

Furthermore, when $r = 2$, $N(D_i, 1, k)$ is described by an elegant expression, recently derived by Carter et al. [4]. Their proof is based on a lengthy, computer-assisted application of the multivariate Lagrange inversion formula (as described by Goulden and Jackson [9]). A simplified proof, based on a special case of this formula, has been given by Steel [16], but neither proof sheds light on why the expression is a product of tree-related quantities. In view of this, and as a first step towards extending the result to r-colourings, for $r > 2$, the authors of [4] ask for a structural proof of their theorem.

The first part of this paper provides such a proof, and shows how the result may

be used to compute the mean of the distribution N. We then describe the dual distribution, N^* when $c=1$, $r=2$, and apply this to calculate the asymptotic average weight of bicoloured sequences on their "best-fit" tree, as $c \to \infty$. The extension of these results when $r>2$ is discussed briefly.

2. Counting trees

We adopt the terminology and notation of Bondy and Murty [3]. A *binary tree* on a label set L is a tree consisting of zero or more unlabelled vertices of degree 3 and $|L|$ vertices of degree 1 (or degree 0 if $|L|=1$), each of which is assigned a distinct label from L. We refer to the degree-1 vertices as the *endpoints* of T and the degree-3 vertices as the *internal* vertices of T. For an *a/b colouring* (an ordered partition of L into two sets of size a and b), each binary tree T on L has an induced colouring of its endpoints. Let $f_k(a,b)$ be the number of binary trees of weight k (defined above) for a given a/b colouring. The main theorem from [4] gives a convenient expression for $f_k(a,b)$ in terms of other tree-based quantities, which also have simple expressions.

Specifically, let $b(n)$ denote the number of binary trees on $\{1,\ldots,n\}$ and $N(n,k)$ the number of forests consisting of exactly k rooted binary trees on a total of exactly n labels. Here a rooted binary tree is either a single labelled vertex or a tree with labelled endpoints obtained by subdividing an edge of a binary tree (the new vertex created by the edge subdivision is called a *root*). It is well known (see Constantinescu and Sankoff [5]) that

$$b(n) = (2n-5)!! = (2n-5)\times(2n-7)\times\cdots\times 3\times 1, \quad \text{for } n\geq 3,$$

while a standard argument, given in [4], shows:

$$N(n,k) = \begin{cases} \dfrac{(2n-k-1)!}{(n-k)!\,(k-1)!\,2^{n-k}}, & \text{if } 1\leq k\leq n, \\ 0, & \text{if } k>n. \end{cases}$$

We can now state the main result from [4].

The Bichromatic Binary Tree (BBT) Theorem.

$$f_k(a,b) = \frac{(k-1)!\,(2n-3k)N(a,k)N(b,k)b(n)}{b(n-k+2)}, \quad n=a+b.$$

We now proceed to a structural proof of this result, using Menger's theorem to express the weight of a bicolouring of the endpoints of a binary tree T in terms of the maximal size of certain sets of disjoint paths in T, and then showing how these paths decompose T into an appropriate forest, F. We begin by establishing a strong

invariance property required to enumerate the set of bicoloured binary trees which decompose into F.

3. Tree extensions

For $|L| \geq 1$, let $U(L)$ and $R(L)$ denote respectively the set of binary and rooted binary trees on L, written $U(n)$, $R(n)$ for $L = \{1, \ldots, n\}$. If $T \in R(L)$, $|L| \geq 2$, let $T^{\wedge}$ denote the unique binary tree from which T is obtained by an edge subdivision. Suppose $T_0 \in U(L_0)$, and for $i = 1, \ldots, r$, $T_i \in R(L_i)$ where $|L_0| \geq 2$, and $L_0, \ldots, L_r$ partitions $\{1, \ldots, n\}$. A tree $T \in U(n)$ is an *extension* of T_0 by $T_1, \ldots, T_r$ *of index* k if:

(a) T contains one subdivision of each of $T_0, T_1, \ldots, T_r$ as disjoint subtrees, and precisely k other internal vertices.

(b) When $|L_i| \geq 2$, any path in T joining vertices in (subdivisions of) T_0 and T_i passes through the root of (subdivided) T_i, for $i = 1, \ldots, r$.

This definition is illustrated (for $k = 1$, $r = 4$) in Fig. 1.

Let $\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)$ denote the set of extensions of T_0 by $T_1, \ldots, T_r$ of index k and let $\mathrm{Ext}(T_0; T_1, \ldots, T_r) = \bigcup_k \mathrm{Ext}_k(T_0; T_1, \ldots, T_r)$.

Lemma 3.1.

(1) $$|\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)| = \frac{\varepsilon_0 (k+r-1)!}{k!\, 2^k} \times \binom{2n-r-k-4}{r-k-1}$$

where ε_0 *is the number of edges of* T_0.

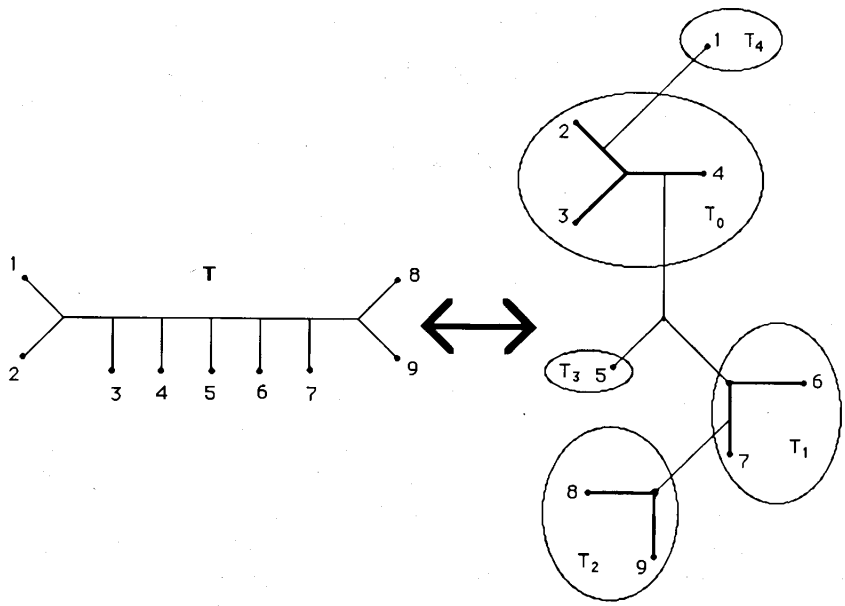


Fig. 1.

In particular, for fixed n, $|\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)|$ is independent of $T_1, \ldots, T_r$.

$$|\mathrm{Ext}(T_0; T_1, \ldots, T_{r-1})| = \frac{\varepsilon_0 b(n)}{b(n-r+2)}. \tag{2}$$

Proof. (1) We may suppose that for $\{T_0, \ldots, T_r\}$, $T_0, \ldots, T_s$ $(0 \le s \le r)$ each have at least two endpoints. Let V_1 denote the remaining $r-s$ labelled vertices. By Menon's theorem (Moon [14]) the number of trees having $s \ge 0$ labelled vertices $v_0, \ldots, v_s$ of degree $d_0, \ldots, d_s$, $r-s$ labelled vertices of degree 1, and k labelled vertices $w_1, \ldots, w_k$ of degree 3 is

$$\begin{cases} \dfrac{(k+r-1)!}{2^k \prod_{0 \le i \le s} (d_i - 1)!}, & \text{if } \sum\limits_{0 \le i \le s} d_i = r+s-k, \\ 0, & \text{otherwise.} \end{cases}$$

$\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)$ is then constructed from these trees by

(i) unlabelling the vertices $w_1, \ldots, w_k$, and

(ii) replacing each v_i by T_i for $i = 0, \ldots, s$, and choosing edge subdivisions of T_i so that each new vertex (in the subdivision of T_i) is one end of an edge formerly incident with v_i.

Let $\varepsilon_i = |E(T_i)|$. For $i = 0$, there are $\varepsilon_0 \times \cdots \times (\varepsilon_0 + d_0 - 1) = (\varepsilon_0 + d_0 - 1)_{d_0}$ possible choices for (ii), while for $i > 0$, the number of possible choices for (ii) is 1 if $d_i = 1$, and $\varepsilon_i \times \cdots \times (\varepsilon_i + d_i - 2) = (\varepsilon_i + d_i - 2)_{d_i - 1}$ if $d_i > 1$. Thus $|\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)|$ is the sum of

$$\frac{(k+r-1)!}{2^k \prod_{0 \le i \le s} (d_i - 1)!} \times \frac{1}{k!} \times (\varepsilon_0 + d_0 - 1)_{d_0} \times \prod_{i \ge 1 : d_i > 1} (\varepsilon_i + d_i - 2)_{d_i - 1}$$

over all positive choices of $d_0, \ldots, d_s$ for which $\sum_{0 \le i \le s} d_i = r+s-k$.

Now for $i = 0, \ldots, s$, let n_i denote the number of endpoints of T_i. Then $\varepsilon_0 = 2n_0 - 3$, and $\varepsilon_i = 2n_i - 2$, for $i > 0$. Then letting $x_i = d_i - 1$, the above four-term product can be written:

$$\frac{\varepsilon_0 (k+r-1)!}{k! \, 2^k} \times \prod_{0 \le i \le s : x_i \ge 1} \binom{2n_i - 3 + x_i}{x_i}.$$

This expression remains unchanged if the restriction $x_i \ge 1$ is modified to $x_i \ge 0$, by the convention $\binom{p}{0} = 1$, and $|\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)|$ is then the sum of this (modified) expression over all choices of $x_0, \ldots, x_s \ge 0$, for which $\sum_{0 \le i \le s} x_i = r-k-1$. Thus $|\mathrm{Ext}_k(T_0; T_1, \ldots, T_r)|$ equals $\varepsilon_0 (k+r-1)!/(k! \, 2^k)$ multiplied by the coefficient of x^{r-k-1} in

$$\prod_{i=0}^{s} \sum_{j \ge 0} \binom{(2n_i - 3) + j}{j} x^j = \prod_{i=0}^{s} (1-x)^{-((2n_i - 3) + 1)} = (1-x)^{-(2n - 2r - 2)}$$

which establishes the claim.

(2) By (1), $|\mathrm{Ext}(T_0; T_1, \ldots, T_{r-1})|$ is independent of $n_1, \ldots, n_{r-1}$. Thus we may assume $n_1 = n - n_0 - r + 2$, while $n_j = 1$ for $j > 1$. Then $|\mathrm{Ext}(T_0; T_1, \ldots, T_{r-1})| = |\mathrm{Ext}(T_0; T_1)| \times$

$(b(n)/b(n_0+n_1))$, since, by [5, Theorem 1], $b(n)/b(m)$ is the number of trees $T \in U(n)$ containing any given tree in $U(m)$. Finally, $\mathrm{Ext}(T_0; T_1) = \varepsilon_0$, giving the required result. □

4. Bicoloured trees

For a bicoloured binary tree (a binary tree whose endpoints have each been assigned one of two colours), a *proper path set* for T is a collection of disjoint paths, each connecting differently-coloured endpoints of T. A *maximal proper path set* for T is a proper path set of maximal cardinality amongst all such sets.

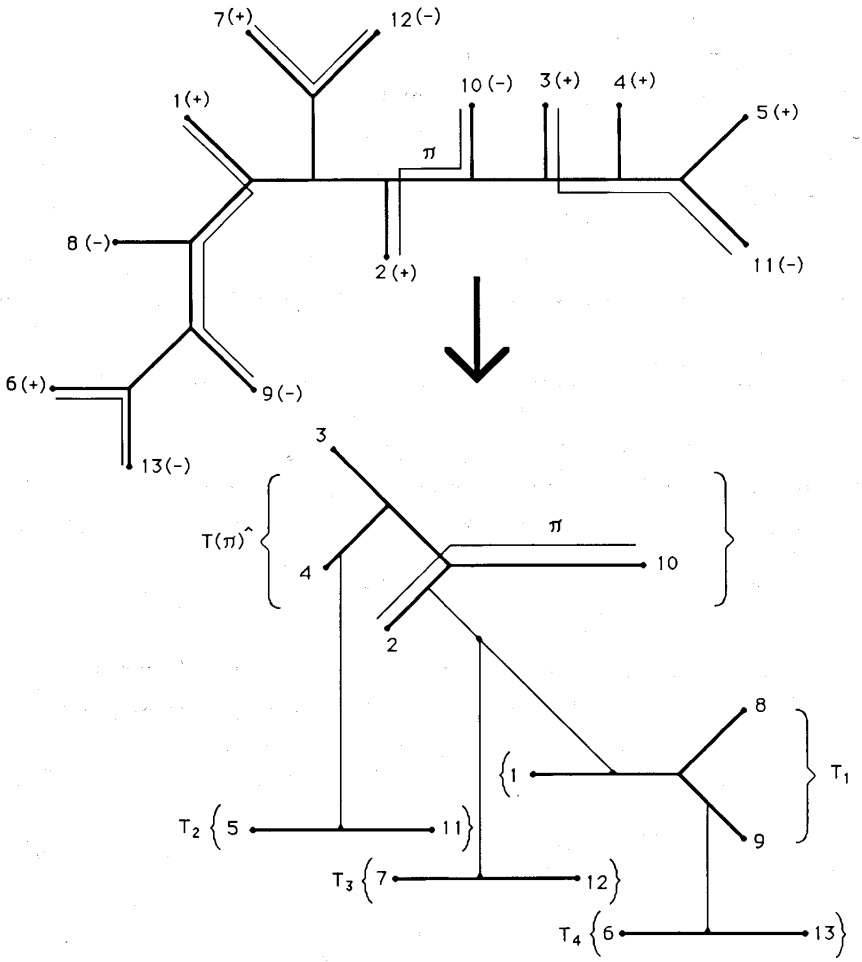


Fig. 2.

Theorem 4.1. (1) *Suppose a bicoloured binary tree T has a maximal proper path set Π of size k. Then each path $\pi \in \Pi$ defines a unique forest $F(T, \pi)$ of k rooted subtrees such that*:

(a) *π lies in exactly one tree, denoted $T(\pi)$, in $F(T, \pi)$.*

(b) *Deleting the root of any tree in $F(T, \pi)$ and its edges from that tree in $F(T, \pi)$ gives two oppositely-coloured, monochromatic subtrees.*

(c) *T is an extension of $T(\pi)^\wedge$ by $F(T, \pi) - \{T(\pi)\}$.*

(2) *Suppose a bicoloured binary tree T, and forest F of k rooted binary trees, including a tree T_0, satisfy* (b) *and* (c) *of* (1) *with $F(T, \pi)$ and $T(\pi)$ replaced by F and T_0, respectively. Then,*

(i) *T has a maximal proper path set of size k, and*

(ii) *for any such set Π_1, exactly one path in Π_1 lies in T_0.*

An example of this decomposition is given in Fig. 2, in which 1–7 are assigned one colour (+) and 8–13 are assigned another colour (−).

Proof. (1) We use induction on $|T|$, the number of endpoints of T. For $|T| \leq 4$ the result clearly holds, so suppose it holds for all $|T| < n$, $n \geq 5$, let $|T| = n$ and suppose Π is a maximal proper path set for T. Since T is binary and $n \geq 5$ there exist at least two internal vertices of T, each of which is adjacent to two endpoints. We denote these pairs of endpoints as $\{v_1, v_2\}$ and $\{w_1, w_2\}$, as indicated in Fig. 3(a). For $\pi \in \Pi$, distinguish two cases depending on the colour of v_1, v_2, w_1, w_2.

Case 1. One pair, say v_1, v_2, have the same colour. Then v_1 or v_2 (or both) does not lie on π. Deleting one such vertex (say v_2) and its incident edge gives a rooted binary tree T^1. Let $T_1 = (T^1)^\wedge$, a binary tree, with $|T_1| = n - 1$, as in Fig. 3(b). Let Π_1 be the maximal proper path set of size k obtained by restricting Π to T_1.

Case 2. Otherwise at least one pair is disjoint from π, for if not Π fails to be a *maximal* proper path set, since we could replace π by a (v_1, v_2)-path and a (w_1, w_2)-path. Thus we may suppose v_1 and v_2 are a pair not on π. Delete these vertices and their incident edges, together with the edge, e, adjacent to these edges to obtain a rooted binary tree T^2. Let $T_2 = (T^2)^\wedge$, a binary tree with $|T_2| = n - 2$, as in Fig. 3(c). At least one of the vertices v_1, v_2 lies on a path in Π. Deleting this path from Π, and restricting to T_2 gives a maximal proper path set Π_2 of size $k - 1$.

In both cases, restricting π to T_i distinguishes a path $\pi_i \in \Pi_i$. Applying the inductive hypothesis, there is a unique set $F(T_i, \pi_i)$ (of size k when $i = 1$, and $k - 1$ when $i = 2$) of rooted subtrees satisfying conditions (a)–(c) of the theorem for T_i and Π_i. In Case 1 suppose $v_1 \in V(T^1)$, where $T^1 \in F(T_1, \pi_1)$. Subdivide that edge of T^1 which is incident with v_1, and join the new vertex to v_2 by a new edge to obtain a binary tree T^2. In Case 2, T is an extension of T_2 by the rooted tree $T_{12} \in R(\{v_1, v_2\})$ (subdividing edge $e_{\alpha\beta}$, as indicated in Fig. 3(c), and joining the new vertex to the root of T_{12}). Thus taking $F(T, \pi) = F(T_1, \pi_1) \cup \{T^2\} - \{T^1\}$, in Case 1, and $F(T, \pi) = F(T_2, \pi_2) \cup \{T_{12}\}$, in Case 2, we satisfy conditions (a)–(c) of the theorem for T. Uniqueness of the trees in $F(T, \pi)$ follows similarly by induction.

(2) For each tree $T_i \in F$, choose any path which connects endpoints of T_i and crosses the root of T_i. These paths generate a proper path set, Π_0, of T of size k. Now for any proper path set Π' for T, replacing each tree T_i in F by a labelled vertex x_i defines a graph $G(\Pi')$ (with possible loops) on $\{x_0, \ldots, x_{k-1}\}$ as follows: join x_i and x_j if there is a path $\pi \in \Pi'$ having endpoints in T_i and T_j (thus a vertex can be joined to itself). Directing all edges of T not on T_0 away from T_0, and giving all loops of $G(\Pi')$ an arbitrary direction converts $G(\Pi')$ into a digraph $D(\Pi')$. Since T is a tree, $D(\Pi')$ has no cycles of length >1, and since Π' is a proper path set, conditions (b) and (c) imply that each vertex of $D(\Pi')$ has indegree equal to 0 or 1. Thus each component of $D(\Pi')$ consists of a tree directed away from some vertex x_i, together with at most one loop on x_i. Thus the number of arcs in $D(\Pi')$ is $\leq k$ with equality iff each component tree has an adjoined loop. But by construction $D(\Pi')$ has $|\Pi'|$ arcs. Thus $|\Pi'| \leq k$, so that the proper path set Π_0 constructed above is maximal, giving (i), while if Π' is maximal, (i.e., $|\Pi'| = k$) then x_0 has one loop, so that there is a unique path lying in T_0, establishing (ii). □

We now exploit the link provided by Menger's theorem between maximal proper path sets of size k and a/b colourings of weight k to establish our main result.

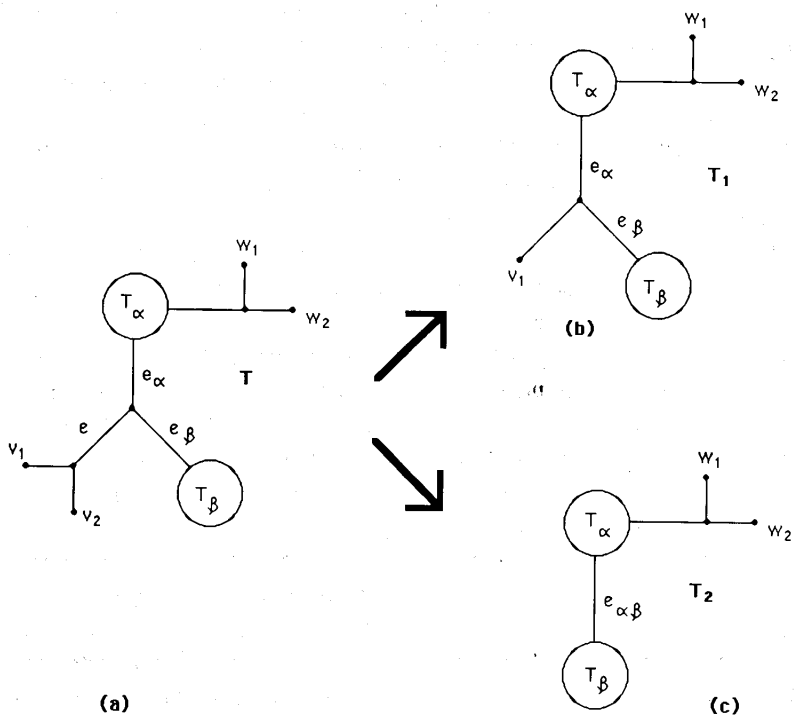


Fig. 3.

Assign one colour to $\{1,\dots,a\}$ and a different colour to $\{a+1,\dots,n\}$, $n=a+b$. Under this colouring, consider the set $F_k(a,b)$ of trees $T\in U(n)$ having at least one maximal proper path set, $\Pi(T)$, of size k. Let

$$F_k^*(a,b)=\{(T,\pi)\colon T\in F_k(a,b),\ \pi\in\Pi(T)\}.$$

Denote by $H(a,b,k)$ the collection of all forests on $\{1,\dots,n\}$, $n=a+b$, consisting of k rooted binary trees, such that for each tree deletion of the root partitions the labels on the endpoints into a subset of $\{1,\dots,a\}$, and a subset of $\{a+1,\dots,n\}$. Finally let

$$G_k(a,b)=\{(F,t,T)\colon F\in H(a,b,k),\ t\in F,\ T\in\mathrm{Ext}(t^{\wedge};F-\{t\})\}.$$

Theorem 4.2. (1) *There is bijection ψ from $F_k^*(a,b)$ onto $G_k(a,b)$.*

$$(2)\qquad f_k(a,b)=\frac{(k-1)!\,(2n-3k)N(a,k)N(b,k)b(n)}{b(n-k+2)},\quad n=a+b.$$

Proof. (1) For $(T,\pi)\in F_k^*(a,b)$ let $\psi(T,\pi)=(F,t,T)$ where $F=F(T,\pi)$ and $t=T(\pi)$, with $F(T,\pi)$ and $T(\pi)$ as defined by Theorem 4.1(1), which provides that ψ is a well-defined function from $F_k^*(a,b)$ into $G_k(a,b)$. Now if $\psi(T_1,\pi)=\psi(T_2,\pi')$ then $T_1=T_2$ by definition of ψ, and since π and π' both lie completely in the same component of F, then $\pi=\pi'$, so that ψ is one-to-one.

If $(F,t,T)\in G_k(a,b)$, then $T\in F_k(a,b)$ by Theorem 4.1(2)(i), and if $\pi(F,t,T)$ denotes the path defined by Theorem 4.1(2)(ii), with $\Pi_1=\Pi(T)$ and $T_0=t$ then by (1) of the same theorem, $\psi(T,\pi(F,t,T))=(F,t,T)$, so that ψ is onto, as required.

(2) By a suitable variation of Menger's theorem (see Harary [10, pp. 50–51]) it is easily shown that the weight of a bicoloured binary tree T is the size of a maximal proper path set for T. Thus, $f_k(a,b)=|F_k(a,b)|$, giving $|F_k^*(a,b)|=kf_k(a,b)$. Now,

$$\begin{aligned}|G_k(a,b)| &= \sum_{F\in H(a,b,k),\ t\in F}|\mathrm{Ext}(t^{\wedge};F-\{t\})|\\ &= \sum_{F\in H(a,b,k),\ t\in F}\frac{(2|t|-3)b(n)}{b(n-k+2)}\end{aligned}$$

by Lemma 3.1(2) where $|t|$ is the number of endpoints of t. Thus since $\sum_{t\in F}|t|=n$, we have

$$|G_k(a,b)|=\frac{(2n-3k)b(n)}{b(n-k+2)}\times|H(a,b,k)|.$$

Now $H(a,b,k)$ can be constructed as follows. Take a forest of k rooted binary trees on label set $\{1,\dots,a\}$, (in $N(a,k)$ ways) and a forest of k rooted binary trees on label set $\{a+1,\dots,n\}$, (in $N(b,k)$ ways), pair them up (in $k!$ ways) and make each pair of roots the ends of a new subdivided edge. In this way, $|H(a,b,k)|=k!N(a,k)N(b,k)$. The result now follows from the bijection in (1). □

5. Counting r-coloured trees, $r>2$

The proof of Lemma 3.1 (above) is similar (though the transition to part (2) is simpler) to the proof of [4, Theorem 2]. This result states:

$$f_{r-1}(a_1,\ldots,a_r)=\frac{b(n)}{b(n-r+2)}\times\prod_{1\le i\le r}N(a_i,1),$$

where $f_k(a_1,\ldots,a_r)$ is the number of binary trees of weight k for an $a_1/a_2/\cdots/a_r$ colouring, and $n=\sum_{1\le i\le r}a_i$. We firstly show how this result follows directly from Lemma 3.1, and then consider $f_r(a_1,\ldots,a_r)$.

Suppose $T_i\in U(L_i)$, $|L_i|\ge 2$, for $i=1,\ldots,r$, where $L_1,\ldots,L_r$ partition L. Let $\mathrm{Ext}(T_1,\ldots,T_r)$ denote the set of all trees in $U(n)$ containing subdivisions of $T_1,\ldots,T_r$ as disjoint subtrees. For an edge e_i of T_i let $T_i(e_i)$ denote the rooted binary tree obtained from T_i by subdividing e_i. Then $\mathrm{Ext}(T_1,\ldots,T_r)$ is the union over all choices of $\{e_2,\ldots,e_r\}$ of $\mathrm{Ext}(T_1;T_2(e_2),\ldots,T_r(e_r))$. As this union is disjoint, Lemma 3.1(2) gives:

$$|\mathrm{Ext}(T_1,\ldots,T_r)|=\frac{b(n)}{b(n-r+2)}\times\prod_{1\le i\le r}(2n_i-3)$$

where $n_i=|L_i|$, $n=\sum_i n_i$. Thus letting

$$\mathrm{Ext}(n_1,\ldots,n_r)=\bigcup_{\{T_1,\ldots,T_r:\,T_i\in U(L_i)\}}\mathrm{Ext}(T_1,\ldots,T_r)$$

we have

$$\begin{aligned}|\mathrm{Ext}(n_1,\ldots,n_r)|&=\frac{b(n)}{b(n-r+2)}\times\prod_{1\le i\le r}(2n_i-3)\times\prod_{1\le i\le r}b(n_i)\\&=\frac{b(n)}{b(n-r+2)}\times\prod_{1\le i\le r}N(n_i,1),\end{aligned}$$

since $N(n_i,1)=(2n_i-3)b(n_i)$.

Now trees of weight $r-1$ for an $a_1/a_2/\cdots/a_r$ colouring are precisely the trees obtained by taking all extensions of trees $T_1,\ldots,T_r$ where T_i has a_i endpoints, all of which are assigned the ith colour. Thus, $f_{r-1}(a_1,\ldots,a_r)=|\mathrm{Ext}(a_1,\ldots,a_r)|$, as required.

We now outline how these ideas may be extended to calculate $f_r(a_1,\ldots,a_r)$. We may suppose $a_i>1$ for each i, since if $a_1=1$ (say) then $f_r(a_1,\ldots,a_r)=(2n-5)f_{r-1}(a_2,\ldots,a_r)$. Let S_i denote the (disjoint) union of the sets $\mathrm{Ext}(T_1,\ldots,T_{r+1})$ over all collections $\{T_1,\ldots,T_{r+1}\}$ for which

(i) T_i and T_{r+1} have between them a_i endpoints, and

(ii) for $1<j\le r$, T_j has a_j endpoints.

Lemma 3.1(2) then gives:

$$(1)\qquad |S_i|=\frac{b(n)}{b(n-r+1)}\times N(a_i,2)\times\prod_{1\le j\le r:\,j\ne i}N(a_j,1)$$

$$= \frac{b(n)}{b(n-r+1)} \times \prod_{1 \le j \le r} N(a_j, 1),$$

since $N(p,2) = N(p,1)$ for $p \ge 2$.

Now regard the combined a_i endpoints of T_i and T_{r+1} as ith coloured (that is, assigned the ith colour), and the endpoints of T_j as jth coloured. Then $\bigcup_{1 \le i \le r} S_i$ consists precisely of those binary trees of weight r and $r-1$ for an $a_1/a_2/\cdots/a_r$ colouring, and a tree T in S_i has weight $r-1$ precisely if T_i and T_{r+1} each have a vertex incident with the same edge of T.

This observation has two immediate consequences. Let S_i^+ denote the subset of S_i of trees of weight r.

(2) $|S_i^+| = |S_i| - (2a_i - 3)\,|\mathrm{Ext}(a_1, \ldots, a_r)|$.

(3) $S_i^+ \cap S_j^+$ is the disjoint union over T^* of $\mathrm{Ext}(T^*; F)$ where F is a forest of $r-2$ rooted binary trees T_k, $k \ne i, j$, having a_k (kth coloured) endpoints, and T^* is a binary tree obtained by twice subdividing the edge of a binary tree with a_i (ith coloured) endpoints and making the two new vertices adjacent to the roots of two rooted binary trees having between them a total of a_j (jth coloured) endpoints, by introducing two new edges.

The number of such trees T^* is $2(2a_i - 3)\,b(a_i)N(a_j, 2)$ which equals $2N(a_i, 1)N(a_j, 1)$. Thus applying Lemma 3.1(2), gives

$$(4) \qquad |S_i^+ \cap S_j^+| = \frac{2(2(a_i + a_j) - 3)\,b(n)}{b(n-r+3)} \times \prod_{1 \le k \le r} N(a_k, 1).$$

Also by (3), $S_i^+ \cap S_j^+ \cap S_k^+ = \emptyset$ for i, j, k distinct, so that by the principle of inclusion and exclusion:

$$(5) \qquad f_r(a_1, \ldots, a_r) = \left| \bigcup_{1 \le i \le r} S_i^+ \right| = \sum_{1 \le i \le r} |S_i^+| - \sum_{\{i,j\}:\, i \ne j} |S_i^+ \cap S_j^+|.$$

Combining parts (1)-(5), and Lemma 3.1(2) gives

Theorem 5.1. *For $r \ge 2$ and $a_i \ge 2$ for $i = 1, \ldots, r$,*

$$f_r(a_1, \ldots, a_r) = \frac{(r-1)(4(n-r)^2 - 2n + r)\,b(n)}{b(n-r+3)} \times \prod_{1 \le i \le r} N(a_i, 1).$$

6. Application

Regarding aligned (binary-state) genetic sequence data D of length c as a sequence, $D_1, \ldots, D_c$, of bicolourings of $\{1, \ldots, n\}$, it is desirable to compare the weight of D on its minimum-length tree(s) with the average weight of D on all trees in $U(n)$. Let $\mu(D)$ denote this average, and for $1 \le i \le c$, let m_i denote the size of the smaller (or smallest equal) subset in the two-set partition of $\{1, \ldots, n\}$ induced by D_i.

Finally, set

$$f_D(a) = |\{i: m_i = a\}|, \qquad \mu_n(a) = \sum_{k \ge 0} \frac{k f_k(a, n-a)}{b(n)}.$$

Note that for any n, $\mu_n(1)=1$, while for fixed a, $\lim_{n\to\infty}\mu_n(a)=a$, by the BBT Theorem (and the expression for $N(n,k)$).

Proposition 6.1. $\mu(D)=\sum_{a=0}^{[n/2]} f_D(a)\,\mu_n(a)$.

Proof.

$$\begin{aligned}
\mu(D) &= b(n)^{-1} \sum_{\{T\in U(n)\}} w(D,T) \\
&= b(n)^{-1} \sum_{\{T\in U(n)\}} \sum_{i\geq 0} w(D_i,T) \\
&= b(n)^{-1} \sum_{a\geq 0} \sum_{k\geq 0} \sum_{\{T\in U(n)\}} \sum_{\{i:\, m_i=a,\, w(D_i,T)=k\}} k \\
&= b(n)^{-1} \sum_{a\geq 0} \sum_{k\geq 0} \sum_{\{i:\, m_i=a\}} k\times |\{T\in U(n):\, w(D_i,T)=k\}| \\
&= \sum_{a\geq 0} \sum_{k\geq 0} \frac{k f_k(a,n-a)}{b(n)} \times |\{i:\, m_i=a\}| \\
&= \sum_{a=0}^{[n/2]} f_D(a)\,\mu_n(a),
\end{aligned}$$

since $f_D(a)=0$ unless $0\leq 2a\leq n$, completing the proof. □

Thus $\mu(D)$ can be readily calculated, even for moderately large values of n, since $f_k(a,n-a)/b(n)$ and hence $\{\mu_n(a):\, 0\leq a\leq [n/2]\}$ can be efficiently calculated by the BBT Theorem. A further application of the BBT Theorem can be found in Steel, Hendy and Penny [17].

7. Counting colourings

We now determine the number of bicolourings of weight k of a binary tree T, denoted $f_k(T)$, and thereby derive the first two moments of the distribution N^*, for $r=2$. This is motivated by the desire to measure the extent to which genetic data is "tree-like"—that is whether the data can be fitted to a binary tree so as to induce a weight significantly less than random data.

Theorem 7.1. *For any* $T\in U(L)$, $n=|L|\geq 1$,

$$f_k(T)=\begin{cases} \left(\binom{n-k}{k}+\binom{n-k-1}{k}\right)2^k, & \text{if } 0\leq 2k<n, \\ 2^k, & \text{if } n=2k, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. We prove the theorem by induction on $n \geq 1$. The result holds when $n=1$ so suppose $n \geq 2$. Choose a rooted binary subtree of T on two endpoints v_1, v_2 of T, and represent T as in Fig. 4(a) (where T_2 and edge e exist only when $n>2$). Let T_1 be the binary tree obtained by deleting one of these vertices (say v_2) and its incident edge (as in Fig. 4(b)), and let T_2 be the binary tree (or empty set if $n=2$) obtained by deleting the whole rooted binary subtree on v_1, v_2, as in Fig. 4(c).

The colourings of T consist of two disjoint classes: C_1, those for which v_1 and v_2 have the same colouring, and C_2 the remainder. For a colouring of T of weight k, if the colouring lies in C_i, restricting the colouring to T_i gives a colouring of weight k for $i=1$, and $k-1$ for $i=2$. Conversely, each colouring of weight k of T_1 is the restriction of a unique colouring of weight k of T in C_1, whilst each colouring of weight $k-1$ of T_2 is the restriction of exactly two colourings of T of weight k in C_2. In this way,

$$f_k(T) = f_k(T_1) + 2f_{k-1}(T_2). \tag{1}$$

Since T_1 and T_2 have respectively $n-1$ and $n-2$ endpoints, it follows by induction that $f_k(T)$ depends only on n and k.

Let $P = P(x, y)$ be the ordinary generating function for $f_k(T(n))$, $T(n) \in U(n)$, where x marks $n \geq 1$ and y marks $k \geq 0$. From the recurrence (1), which applies (and for which T_1 and T_2 are binary trees) except when $n=2$ and the endpoints of T are differently coloured, or when $n=1$, we have $P - 2x - 2x^2y = xP + 2x^2yP$, so that $P(x, y) = (2x + 2x^2y)(1 - x - 2x^2y)^{-1}$. Extracting the coefficient of $x^n y^k$ gives the required result. □

Corollary 7.2. *Let $\mu(n, c)$, and $\sigma^2(n, c)$ denote the mean and variance for the weight of aligned binary-state sequence data, $D_1, \ldots, D_c$, to any tree in $U(n)$. Then,*

$$\mu(n, c) = \frac{c((3n-2) - (-2)^{1-n})}{9}$$

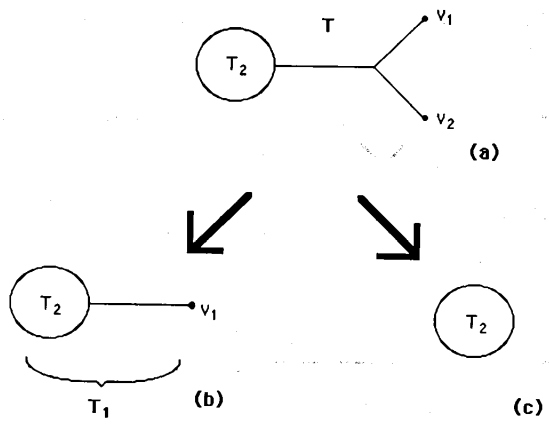


Fig. 4.

and

$$\sigma^2(n,c) = \frac{c(6n+2-(6n+1)(-2)^{1-n}-2^{2-2n})}{81}.$$

Proof. $\mu(n,1)$ and $\sigma^2(n,1)$ are the coefficients of x^n in respectively, $2^{-n}(\partial P/\partial y)\,|_{y=1}$ and $2^{-n}(\partial^2 P/\partial y^2)\,|_{y=1}+\mu(n)-\mu^2(n)$. The result when $c=1$ now follows from the expression for $P(x,y)$ given in the proof of part (2) of the previous theorem. Since the weight of a sequence is the sum of the weights of its component colourings, $\mu(n,c)=c\mu(n,1)$, and $\sigma^2(n,c)=c\sigma^2(n,1)$, as required. □

8. Remarks

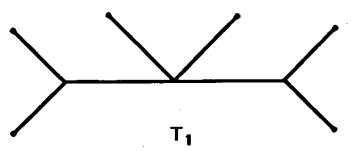
(1) The invariance of $f_k(T)$ to T does not generalize to trees having a given number of labelled endpoints and a given number of unlabelled vertices of degree ≥ 3, as the counterexample T_1, T_2 in Fig. 5 shows.

(2) The invariance of $f_k(T)$ to T also does not generalize to r-colourings of T, for $r>2$. For consider a "caterpillar" tree $J(n)\in U(n)$, $n\geq 1$, which has at most two internal vertices having the property that each is adjacent to at least two endpoints. Let $P_r(x,y)$ denote the ordinary generating function for the number of r-colourings of $J(n)$ of weight k, where x marks n and y marks k. In order to calculate $P_r(x,y)$, label the endpoints of $J(n)$ (for all $n\geq 1$) so that deleting from $J(n)$ any internal vertex and its incident edges partitions $\{1,\ldots,n\}$ into the sets $\{1,\ldots,i-1\}$, $\{i\}$, $\{i+1,\ldots,n\}$ for some i: $2\leq i\leq n-1$ (such a labelling for $J(9)$ is illustrated by the tree T in Fig. 1). For an r-colouring χ of $\{1,\ldots,n\}$ let

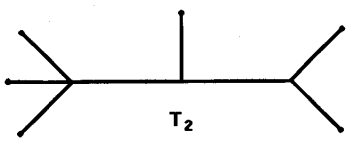
$$v(\chi) = \min\{j: j, j+1,\ldots,n \text{ are all differently coloured by } \chi\}.$$

By a simple application of [11, Theorem 2 (part 3)] of Hartigan, if $v(\chi)>1$, and χ' is the restriction of χ to $\{1,\ldots,v(\chi)-1\}$, we have

$$w(\chi, J(n)) = w(\chi', J(v(\chi)-1)) + n - v(\chi).$$



$f_k(T_1)$ = 1, k=0
8, k=1
19, k=2
4, k=3



$f_k(T_2)$ = 1, k=0
8, k=1
17, k=2
6, k=3

Fig. 5.

Otherwise, if $v(\chi)=1$, then $n \le r$ and all the endpoints of $J(n)$ are assigned different colours (in $n! \times \binom{r}{n}$ possible ways). Using these results we can extend the type of argument used in the proof of Theorem 7.1 to obtain the following:

$$P_r(x,y) = \frac{x \sum_{1 \le j \le r} \binom{r}{j} j! (xy)^{j-1}}{1 - x \sum_{1 \le j \le r} \binom{r-1}{j-1} j! (xy)^{j-1}}.$$

Let $\mu_r(n)$ denote the average weight of r-colourings on J_n. Then $r^n \mu_r(n)$ is the coefficient of x^n in

$$(\partial P_r(x,y)/\partial y)\,|_{y=1} = \frac{rx}{(1-rx)^2} \times \left[1 - \frac{r}{\sum_{1 \le j \le r} \binom{r}{j} j! x^{j-1}} \right].$$

Applying [2, Theorem 2] from Bender (with $B(z) = rz(1-rz)^{-2}$) gives

$$\lim_{n \to \infty} \frac{\mu_r(n)}{n} = 1 - \frac{r^r}{r! \sum_{0 \le j \le r-1} \frac{r^j}{j!}} \quad \text{for } r \ge 2.$$

However, using asymptotic methods similar to those employed by Meir, Moon and Mycielski [13, pp. 146–147] the average weight of r-colourings, averaged over $U(n)$ can be shown to exceed $\mu_r(n)$ for $r=3$ and 4 as $n \to \infty$.

(3) Let $\mu^*(n,c)$ denote the average weight of aligned binary-state sequence data, $D_1, \ldots, D_c$, on their minimum-length tree(s). By the invariance of $f_k(T)$ to T, and the weak law of large numbers, $\lim_{c \to \infty} \mu^*(n,c)/c = \mu(n,1)$. An exact expression for $\mu^*(n,c)$ is not known.

(4) Regarding $\mu(n,1)$ as the expected weight of a random bicolouring of a binary tree, we can compare $\mu(n,1)$ with the expected weight $\mu(T_n)$ of a random bicolouring of a star tree T_n consisting of n labelled endpoints all of which are adjacent to one unlabelled vertex. Such trees have been suggested by Thompson [18] as a suitable null hypothesis in testing evolutionary hypotheses (see for example [15]).

Clearly,

$$2^n \mu(T_n) = \sum_{0 \le k \le n} \binom{n}{k} \times \min\{k, n-k\}.$$

It can be shown (M.R. Carter, personal communication) that the summation term is $n(2^{n-1} - t_n)$ where

$$t_n = \begin{cases} \binom{2k}{k}, & \text{if } n = 2k+1, \\ (2 - (k+1)^{-1}) \binom{2k}{k}, & \text{if } n = 2k+2. \end{cases}$$

Thus, asymptotically in n, $\mu(T_n) \sim n/2$, compared with $\mu(n,1) \sim n/3$ for binary trees.

9. Conclusion

In this paper we have applied structural and inductive arguments to enumerate bicoloured binary trees and bicolourings of binary trees by weight. This is motivated by an attempt to understand the structure underlying the factorization of terms in the BBT Theorem, and the desire to find a corresponding expression for r-colourings. We have also obtained useful information about the biologically relevant distributions N and N^* described in the introduction. In the case of two colours the results give the mean $\mu(n, c)$ and variance $\sigma^2(n, c)$ of N^*, while the mean $\mu(D)$ of N can be readily found, by using the BBT Theorem.

Two problems remain. Firstly, it would be desirable (for $r=2$, say) to be able to readily calculate the variance of N. This would make the comparison of $\mu(D)$ with the weight of D on the minimum-length tree considerably more meaningful. The problem amounts to finding a suitable expression for $N(D, 2, k)$, when $r=2$.

A second problem is the enumeration of r-coloured trees by weight when $r>2$. This would allow, for example, the calculation of $\mu(D)$ for sequence data having more than two character states. [4, Theorem 2], together with Theorem 5.1 suggest that the product form of the BBT Theorem might carry over to r-coloured trees, however the authors of [4] have found that this is not so.

The proof of Theorem 5.1, while it might be extended to calculate $f_k(a_1, \ldots, a_r)$ for $k=r+1$, say, does not readily generalize to give a useful formula for general values of k. A structural proof along the lines of the BBT Theorem (using Lemma 3.1 and a suitable extension of Menger's theorem) could well hold more promise.

References

[1] J.W. Archie, A randomization test for phylogenetic information in systematic data, Syst. Zool. 38 (1989) 239–252.
[2] E.A. Bender, Asymptotic methods in enumeration, SIAM Rev. 16 (1974) 485–515.
[3] J.A. Bondy and U.S.R. Murty, Graph Theory with Applications (Macmillan, London, 1976).
[4] M. Carter, M.D. Hendy, D. Penny, L.A. Székely and N.C. Wormald, On the distribution of lengths of evolutionary trees, SIAM J. Discrete Math. 3 (1990) 38–47.
[5] M. Constantinescu and D. Sankoff, Tree enumeration modulo a consensus, J. Classification 3 (1986) 349–356.
[6] J. Felsenstein, Phylogenies from molecular sequences: Inference and reliability, Annual Rev. Genetics 22 (1988) 521–565.
[7] W.M. Fitch, Towards defining the course of evolution: Minimum change for a specific tree topology, Syst. Zool. 20 (1971) 406–416.
[8] L.R. Foulds and R.L. Graham, The Steiner problem in phylogeny is NP-complete, Adv. Appl. Math. 3 (1982) 43–49.
[9] I.P. Goulden and D.M. Jackson, Combinatorial Enumeration (Wiley, New York, 1983).
[10] F. Harary, Graph Theory (Addison-Wesley, Reading, MA, 1969).
[11] J.A. Hartigan, Minimum mutation fits to a given tree, Biometrics 29 (1973) 53–65.
[12] I.M. Henderson, M.D. Hendy and D. Penny, Influenza viruses, comets and the science of evolutionary trees, J. Theoret. Biol. 140 (1989) 289–303.

[13] A. Meir, J.W. Moon and J. Mycielski, Hereditarily finite sets and identity trees, J. Combin. Theory Ser. B 35 (1983) 142–155.

[14] J.W. Moon, Counting labelled trees, Canadian Mathematical Monographs 1 (Canad. Math. Congress, Montreal, 1970).

[15] D. Penny, L.R. Foulds and M.D. Hendy, Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences, Nature 297 (1982) 197–200.

[16] M.A. Steel, Distributions on bicoloured evolutionary trees, Bull. Austral. Math. Soc. 41 (1990) 159–160.

[17] M.A. Steel, M.D. Hendy and D. Penny, Significance of the length of the shortest tree, J. Classification 9 (1992) 71–90.

[18] E.A. Thompson, Human Evolutionary Trees (Cambridge University Press, Cambridge, 1975).